GÉRARD MARAL | MICHEL BOUSQUET
ZHILI SUN

SATELLITE COMMUNICATIONS SYSTEMS

SYSTEMS, TECHNIQUES AND TECHNOLOGY

SIXTH EDITION

WILEY

SATELLITE COMMUNICATIONS SYSTEMS

Sixth Edition

SATELLITE COMMUNICATIONS SYSTEMS

Systems, Techniques and Technology

Sixth Edition

GÉRARD MARAL

Ecole Nationale Supérieure des Télécommunications, Telecom Paris, Site of Toulouse, France

MICHEL BOUSQUET

(Retired), Ecole Nationale Supérieure de l'Aéronautique et de l'Espace (SUPAERO), Toulouse, France

ZHILI SUN

University of Surrey, UK

WILEY

This edition first published 2020

Edition History
John Wiley & Sons Ltd: 1e, 1986; 2e, 1993; 3e, 1998; 4e, 2002; 5e, 2009

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www .wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Maral, Gérard, author. | Bousquet, Michel, author. | Sun, Zhili, author.
Title: Satellite communications systems : systems, techniques and technology / Gérard Maral, Ecole Nationale Supérieure des Télécommunications, Site de Toulouse, France, Michel Bousquet, Ecole Nationale Supérieure de l'Aéronautique et de l'Espace (SUPAERO), Toulouse, France, Zhili Sun, University of Surrey, UK.
Other titles: Systémes de télecommunications par satellites. English
Description: Sixth edition. | Hoboken, N.J. : John Wiley & Sons, 2020. | Translation of: Systémes de télecommunications par satellites. | Includes bibliographical references and index.
Identifiers: LCCN 2019044958 (print) | LCCN 2019044959 (ebook) | ISBN 9781119382089 (hardback) | ISBN 9781119382072 (adobe pdf) | ISBN 9781119382126 (epub)
Subjects: LCSH: Artificial satellites in telecommunication.
Classification: LCC TK5104 .M3513 2020 (print) | LCC TK5104 (ebook) | DDC 621.382/5–dc23
LC record available at https://lccn.loc.gov/2019044958
LC ebook record available at https://lccn.loc.gov/2019044959

Cover Design: Wiley
Cover Image: © BlackJack3D/Getty Images

Set in 9/11pt, PalatinoLTStd by SPi Global, Chennai, India

Printed and bound by CPI Group (UK) Ltd, Croydon, CR0 4YY

10 9 8 7 6 5 4 3 2 1

CONTENTS

ACKNOWLEDGEMENT

Reproduction of figures extracted from the 1990 edition of the ITU (formerly CCIR) volumes (XVIIth CCIR Plenary Assembly, Düsseldorf 1990), the *Handbook on Satellite Communications* (ITU Geneva, 1988), and the ITU-R recommendations are made with the authorisation of the International Telecommunication Union (ITU) as copyright holder.

The choice of the excerpts reproduced remains the sole responsibility of the authors and does not involve in any way the ITU.

The complete ITU documentation can be obtained from:

ITU Publications
Place des Nations
1211 Geneva 20
Switzerland

Voice: +41 22 730 6141 (English)
+41 22 730 6142 (Français)
+41 22 730 6143 (Español)

Fax: +41 22 730 5194
Email: sales@itu.int

ACRONYMS

3GPP	3rd Generation Partnership Project
5G	5th generation mobile communications
AAL	ATM adaptation layer
ACI	adjacent channel interference
ACK	acknowledgement
ACM	adaptive coding and modulation
ACTS	advanced communications technology satellite
ADC	analogue to digital converter
ADM	adaptive delta modulation
ADPCM	adaptive pulse code modulation
ADSL	asymmetric digital subscriber line
AKM	apogee kick motor
ALC	automatic level control
AM	amplitude modulation
AMP	amplifier
AOCS	attitude and orbit control system
AOR	Atlantic ocean region
AOS	advanced orbiting systems
APD	avalanche photodetector
APSK	amplitude and phase shift keying
AR	available retio; also: axial ratio
ARQ	automatic repeat request
ARQ-GB(N)	automatic repeat request-go back n
ARQ-SR	automatic repeat request-selective repeat
ARQ-SW	automatic repeat request-stop and wait
ARTES	Advanced Research in Telecommunications Systems
ASCII	American Standard Code for Information Interchange
ASIC	application-specific integrated circuit
ASS	amateur satellite service
ATA	auto-tracking antenna
ATM	asynchronous transfer mode
AVBDC	absolute volume-based dynamic capacity
BAPTA	bearing and power transfer assembly
BAT	bouquet association table
BBS	baseband switch
BBP	baseband processor
BCH	Bose-Chaudhuri-Hocquenghem
BCR	battery charge regulator
BDR	battery discharge regulator
BEP	bit error probability
BER	bit error rate
BFN	beam forming network
BFSK	binary frequency shift keying
BGMP	border gateway multicast protocol
BGAN	broadband global area network
BGP	border gateway protocol
BHCA	busy hour call attempts
BHCR	busy hour call rate
BISDN	broadband ISDN
BIS	broadband interactive system
BITE	built-in test equipment
BOL	beginning of life
BPF	band pass filter
BPSK	binary phase shift keying
BSM	broadband satellite multimedia
BSS	broadcasting satellite service
BTP	burst time plan
C2P	connection control protocol
CAMP	channel amplifier
CAT	conditional access table
CBR	constant bit rate
CC	command and control
CCB	Common Core Booster
CCI	co-channel interference
CCIR	Comité Consultatif International des Radiocommunications (International Radio Consultative Committee); replaced by ITU-R
CCITT	Comité Consultatif International Télégraphe et Téléphone (International Telegraph and Telephone Consultative Committee); replaced by ITU-T
CCSDS	Consultative Committee for Space Data Systems

CDMA	code division multiple access
CEC	Commission of the European Communities
CELP	code excited linear prediction
CENELEC	Comité Européen pour la Normalisation en Electrotechnique (European Committee for Electrotechnical Standardisation)
CEPT	Conférence Européenne des Postes et des Télécommunications (European Conference of Post and Telecommunications)
CFM	companded frequency modulation
CIR	committed information rate
CIS	Commonwealth of Independent States
CLTU	command link transmission unit
CMOS	complementary metal oxide semiconductor
CNES	Centre National d'Etudes Spatiales (French space agency)
COFDM	coded orthogonal frequency division multiplexing
CoS	class of service
COTS	commercial off the shelf
CPS	chemical propulsion system
CRA	continuous rate assignment
CRC	cyclic redundancy check
CSC	common signal channel
CSMA	carrier sense multiple access
CTM	correction message table
CTU	central terminal unit
DAB	digital audio broadcasting
DAC	digital to analogue converter
DAMA	demand assignment multiple access
DARPA	Defense Advanced Research Project
dB	decibel
dBm	unit for expression of power level in dB with reference to 1 mW
dBmO	unit for expression of power level in dBm at a point of zero relative level (a point of a telephone channel where the 800 Hz test signal has a power of 1 mW)
DBS	direct broadcasting satellite
DC	direct current
DCE	data circuit terminating equipment
DCME	digital circuit multiplication equipment
DCS	digital cellular system (GSM At 1800 MHz)
DCT	discrete cosine transform
DCU	distribution control unit
DE	differentially encoded
DE-MPSK	differentially encoded M-ary phase shift keying
DES	data encryption standard
DM	delta modulation
DNS	domain name service (host name resolution protocol)
DOD	depth of discharge
DOF	degree of freedom
DSCP	differentiated service code point
DSI	digital speech interpolation
DSL	digital subscriber loop
DSM	digital storage medium
DSP	digital signal processing
DST	destination host address
DTE	data terminating equipment
DTH	direct to home
DTS	decoding timestamp
DVB	digital video broadcasting
DVB-S	DVB-Satellite
DVB-S2	DVB-Satellite 2nd generation
DVB-S2X	DVB-Satellite 2nd generation extension
DVB-RCS	DVB-Return Channel via Satellite
DVB-RCS2	DVB-Return Channel via Satellite 2nd generation
DVB-RCS2X	DVB-Return Channel via Satellite 2nd generation extension
EBU	European Broadcasting Union
ECN	explicit congestion notification
ECSS	European Cooperation for Space Standardization
EESS	Eutelsat earth station standards
EFS	error-free seconds
EIRP	effective isotropic radiated power (W)
EIT	event information table
EITA	electron-bombardment ion thruster assembly
ELSR	edge label switch router
EMC	electromagnetic compatibility
EMI	electromagnetic interference
EN	European standard
EOC	edge of coverage
EOL	end of life
EPC	electric power conditioner
EPIRB	emergency position indicating radio beam
ERC	European Radiocommunications Committee
ERO	European Radiocommunications Office (of the ERC)
ES	earth station; also: ETSI specification
ESA	European Space Agency
ESTEC	European Space Research and Technology Centre

ESS	earth exploration satellite service
ETR	Eastern Test Range
ETSI	European Telecommunications Standards Institute
EUTELSAT	European Telecommunications Satellite Organisation
FCA	free capacity assignment
FCC	Federal Communications Commission
FCS	frame check sequence
FCT	frame composition table
FDM	frequency division multiplex
FDMA	frequency division multiple access
FEC	forward error correction
FES	fixed earth station
FET	field effect transistor
FETA	field effect transistor amplifier
FFT	fast Fourier transform
FIFO	first in first out
FLS	forward link signalling
FM	frequency modulation
FMA	fixed-mount antenna
FMT	fade mitigation technique
FODA	FIFO ordered demand assignment
FPGA	field-programmable gate array
FS	fixed service
FSK	frequency shift keying
FSS	fixed satellite service
FTP	file transfer protocol
GaAs	gallium arsenide
GAN	global area network
GBN	go back N
GC	global coverage
GCE	ground communication equipment
GCS	ground control station
GDE	group delay equaliser
GEO	geostationary earth orbit
GMDSS	Global Maritime Distress and Safety System
GMSK	Gaussian-filtered minimum shift keying
GMT	Greenwich Mean Time
GOS	grade of service
GPRS	general packet radio service
GPS	global positioning system
GRE	generic routing encapsulation
GSE	generic stream encapsulation
GSM	global system for mobile communications
GSO	geostationary satellite orbit
GTO	geostationary transfer orbit
GW	satellite gateway
HDB3	high-density binary 3 code
HDLC	high-level data link control
HDTV	high-definition television
HEMT	high electron mobility transistor
HEC	header error check
HEO	highly elliptical orbit
HIO	highly inclined orbit
HLR	home location register
HPA	high-power amplifier
HPB	half-power beamwidth
HTML	hypertext markup language
HTS	high throughput satellite
HTTP	hypertext transfer protocol
IAT	interarrival time
IAU	international astronomical unit
IBO	input back-off
IBS	International Business Service
ICMP	Internet control message protocol
ICO	intermediate circular orbit
IGMP	internet group management protocol
IDC	intermediate rate digital carrier
IDR	intermediate data rate
IDU	indoor unit
IEEE	Institute of Electrical and Electronic Engineers
IESS	Intelsat earth station standards
IETF	Internet Engineering Task Force
IF	intermediate frequency
IFRB	International Frequency Registration Board
IGMP	Internet group management protocol
IHM	input hybrid matrix
ILS	International Launch Services
IM	intermodulation
IMP	intermodulation product
IMSI	international mobile subscriber identity
IMUX	input multiplexer
INIRIC	International Non-Ionising Radiation Committee
INMARSAT	International Maritime Satellite Organisation
INTELSAT	International Telecommunications Satellite Consortium
IOR	Indian Ocean region
IOT	in-orbit test
IP	Internet protocol
IPA	intermediate power amplifier
IPDR	IP packet duplicate ratio
IPER	IP packet error ratio
IPE	initial pointing error
IPSLBR	IP packet severe loss block ratio
IPLR	IP packet loss ratio
IPRR	IP packet reordered ratio
IPsec	IP security
IRD	integrated receiver decoder
ISDN	integrated services digital network
ISI	inter-symbol interference

ISL	intersatellite link
ISO	International Organisation for Standardisation
ISS	inter-satellite service
ITU	International Telecommunication Union
IVOD	interactive video on demand
IWU	Internet working unit
JPEG	Joint Photographic Expert Group
LAN	local area network
LAPB	link access protocol balanced
LDPC	low-density parity check
LES	land-earth station
LEO	low earth orbit
LH2	liquid hydrogen
LHCP	left-hand circular polarisation
LLC	logical link control
LLM	L band land mobile
LMSS	land mobile satellite service
LNA	low-noise amplifier
LNB	low-noise block
LO	local oscillator
LOS	line of sight
LOX	liquid oxygen
LPC	linear predictive coding
LPF	low-pass filter
LR	location register
LRB	liquid rocket booster
LRE	low-rate encoding
M-PSK	M-ary phase shift keying
MAC	medium access control
MAC	multiplexed analogue components; also: monitoring, alarm, and control
MAMA	multiple ALOHA multiple access
MAN	metropolitan area network
MCD	multicarrier demodulator
MCPC	multiple channels per carrier
MEO	medium earth orbit
MES	mobile earth station
MESFET	metal semiconductor field effect transistor
MF	multi frequency
MIC	microwave integrated circuit
MIFR	master international frequency register
MMDS	multipoint multichannel distribution system
MMIC	monolithic microwave integrated circuit
MMT	multicast map table
MNMC	mission and network management centre
MODEM	modulator/demodulator
MOS	mean opinion score
MPE	multi-protocol encapsulation
MP	measurement point
MPEG	Motion Picture Expert Group
MPLS	multi-protocol label switching
MPSK	M-ary phase shift keying
MS	mobile station; also management station
MSK	minimum shift keying
MSS	mobile satellite service
MTBF	mean time between failure
MTBO	mean time between outages
MTU	maximum transferable unit
MUX	multiplexer
NACK	no acknowledgment
NASA	National Aeronautics and Space Administration (USA)
NASDA	National Aeronautics and Space Development Agency (Japan)
NAT	network address translation
NCC	network control centre
NCS	network coordination station
NGSO	non-geostationary satellite orbit
NGEO	non-geostationary orbit
NH	Northern hemisphere
NIS	network information system
NIT	network information table
NMC	network management centre
NOAA	National Oceanic and Atmospheric Administration
NORM	NACK-oriented reliable multicast
NRZ	non-return to zero
NSE	network section ensemble
NTP	network time protocol
NVOD	near video on demand
OBC	on-board computer
OBO	output back-off
OBP	on-board processing
OBDH	on-board data handling
ODU	outdoor unit
OFDM	orthogonal frequency division multiplexing
OHM	output hybrid matrix
OICETS	optical inter-orbit communications engineering test satellite
OMUX	output multiplexer
OQPSK	offset QPSK
OSI	Open System Interconnection
OSPF	open shortest path first
OSR	optical solar reflector
PAM	payload assist module
PAT	program association table
PCM	pulse code modulation
PDF	probability density function
PDH	plesiochronous digital hierarchy
PDU	protocol data unit
PEP	performance enhancement protocol

PER	packet error rate
PFD	power flux density
PHEMT	pseudomorphic high electron mobility transistor
PHB	per-hop behaviour
PIA	percent IP service unavailability
PICH	pilot channel
PID	packet identifier
PIMP	passive intermodulation product
PIU	percent IP service unavailability
PKM	perigee kick motor
PL	physical layer
PLL	phase locked loop
PM	phase modulation
PMT	program map table
PN	pseudorandom number
PODA	priority oriented demand assignment
POR	Pacific Ocean region
PPP	point to point protocol
PRBS	pseudorandom binary sequence
PRMA	packet reservation multiple access
PSD	power spectral density
PSI	programme-specific information
PSK	phase shift keying
PSPDN	packet switched public data network
PSTN	public switched telephone network
PTA	programmed tracking antenna
PTS	presentation timestamp
PVA	perigee velocity augmentation
QEF	quasi-error-free
QoS	quality of service
QPSK	quaternary phase shift keying
RAAN	right ascension of the ascending node
RACH	random access channel
RADIUS	remote authentication dial in user service
RAM	random access memory
RBDC	rate-based dynamic capacity
RCS	return channel via satellite; also: reaction control system
RCVO	receive only
Rec	recommendation
Rep	report
RF	radio frequency
RFHMA	random frequency hopping multiple access
RFI	radio frequency interference
RHCP	right-hand circular polarisation
RIP	routing information protocol
RIPR	replicated IP packet ratio
RITA	RF ion thruster assembly
RMA	random multiple access
RMT	RCS map table
RR	Radio Regulations
RRC	Regional Radio Conference
RS	Reed Solomon (coding)
RSS	radiodetermination satellite service
RSVP	resource reservation protocol
RTCP	real-time transport control protocol
RTP	real-time transport protocol
RTT	round-trip times
RTU	remote terminal unit
Rx	receiver
S-ALOHA	slotted ALOHA protocol
SAP	service access point
SAW	surface acoustic wave
SC	suppressed carrier
SCADA	supervisory control and data acquisition
SCH	synchronisation channel
SCPC	single channel per carrier
SCT	superframe composition table
SDH	synchronous digital hierarchy
SDT	service description table
SEP	symbol error probability
SEU	single event upset
SFH	slow frequency hopping
SH	Southern hemisphere
SHF	super-high frequency (3–30 GHz)
SI	service information
SIA	Satellite Industry Association
SIM	subscriber identity module
SKW	satellite-keeping window
SLA	service-level agreement
SLC	satellite link control
SMAC	satellite medium access control
SMATV	satellite-based master antenna for TV distribution
SMS	satellite multi-services
SMTP	simple mail transfer protocol
SNG	satellite news gathering
SNMP	simple network management protocol
SNR	signal-to-noise ratio
SOC	state of charge
SORF	start of receive frame
SOS	space operation service
SOTF	start of transmit frame
SPT	satellite position table; also: stationary plasma thruster
SR	selective repeat
SRC	source host address
SRS	space research service
SS	satellite switched
SSB	solid support booster
SSMA	spread spectrum multiple access
SSO	sun-synchronous orbit
SSPA	solid state power amplifier

SS-TDMA	satellite-switched TDMA
ST	Sidereal time
STM	synchronous transport module
STS	space transportation system
SW	stop and wait
SWR	standing wave ratio
SYNC	synchronisation
TA	ETSI Technical Assembly
TACS	total access communication system
TCB	transmission control block
TBTP	terminal burst time plan
TC	telecommand; also: turbo-coding
TCP	transmission control protocol
TCT	time-slot composition table
TDM	time division multiplex
TDMA	time division multiple access
TDRS	tracking and data relay satellite
TDT	time and date table
TELNET	remote terminal application
TIA	Telecommunications Industry Association
TIM	terminal information messages
TIR	internal reflection
TM	telemetry
TS	transport stream
TR	technical report
TS	transport stream; also: technical specification
TTC	telemetry, tracking, and command
TTCM	telemetry, tracking, command, and monitoring
TTL	transistor–transistor logic; also: time to live
TWT	travelling wave tube
TWTA	travelling wave tube amplifier
Tx	transmitter
UDP	user datagram protocol
UHF	ultra-high frequency (300 MHz – 3 GHz)
UHTV	ultra high definition television
ULE	ultra-lightweight encapsulation; also: unidirectional lightweight encapsulation
USAT	ultra small aperture terminal
UT	universal time, user terminal
UW	unique word
VBR	variable bit rate
VBDC	volume-based dynamic capacity
VC	virtual channel; also: virtual container
VCI	virtual channel identifier
VEB	vehicle equipment bay
VHF	very-high-frequency (30–300 MHz)
VLSI	very large scale integration
VPI	virtual path identifier
VPN	virtual private network
VSAT	very small aperture terminal
WAN	wide area network
WARC	World Administrative Radio Conference
XPD	cross-polarisation discrimination
XPI	cross-polarisation isolation

NOTATIONS

a	orbit semi-major axis
A	azimuth angle (*also* attenuation, area, availability, traffic density, and carrier amplitude)
A_{eff}	effective aperture area of an antenna
A_{AG}	attenuation by atmospheric gases
A_{RAIN}	attenuation due to precipitation and clouds
A_{P}	attenuation of radio wave by rain for percentage p of an average year
B	bandwidth
b	voice channel bandwidth (3100 Hz from 300 to 3400 Hz)
B_{n}	noise measurement bandwidth at baseband (receiver output)
B_{N}	equivalent noise bandwidth of receiver
Bu	burstiness
c	velocity of light = 3×10^8 m s^{-1}
C	carrier power
C/N_0	carrier power-to-noise power spectral density ratio (W/Hz)
$(C/N_0)_{\text{U}}$	uplink carrier power-to-noise power spectral density ratio
$(C/N_0)_{\text{D}}$	downlink carrier power-to-noise power spectral density ratio
$(C/N_0)_{\text{IM}}$	carrier power-to-intermodulation noise power spectral density ratio
$(C/N_0)_{\text{I}}$	carrier power-to-interference noise power spectral density ratio
$(C/N_0)_{\text{I,U}}$	uplink carrier power-to-interference noise power spectral density ratio
$(C/N_0)_{\text{I,D}}$	downlink carrier power-to-interference noise power spectral density ratio
$(C/N_0)_{\text{T}}$	carrier power-to-noise power spectral density ratio for total link
D	diameter of a reflector antenna (*also* used as a subscript for downlink)
e	orbit eccentricity
E	elevation angle (also energy and electric field strength)
E_{b}	energy per information bit
E_{c}	energy per channel bit
f	frequency (Hz)
F_{c}	nominal carrier frequency
f_{d}	antenna focal length
f_{m}	frequency of a modulating sine wave
f_{max}	maximum frequency of the modulating baseband signal spectrum
f_{D}	downlink frequency
f_{U}	uplink frequency
F	noise figure
ΔF_{max}	peak frequency deviation of a frequency modulated carrier
f_{S}	sampling frequency
g	peak factor
G	power gain (also gravitational constant)
G_{sat}	gain at saturation
G_{R}	receiving antenna gain in direction of transmitter
G_{T}	transmitting antenna gain in direction of receiver
G_{Rmax}	maximum receiving antenna gain
G_{Tmax}	maximum transmitting antenna gain
G_{SR}	satellite repeater gain
G_{SRsat}	saturation gain of satellite repeater
G/T	gain to system noise temperature ratio of a receiving equipment
G_{CA}	channel amplifier
G_{FE}	front-end gain from satellite receiver input to satellite channel amplifier input
G_{ss}	small signal power gain
i	inclination of the orbital plane
k	Boltzmann's constant = 1.379×10^{-23} W KHz^{-1}
k_{FM}	FM modulation frequency deviation constant (MHz/V)
k_{PM}	PM phase deviation constant (rad/V)
K_{P}	AM/PM conversion coefficient
K_{T}	AM/PM transfer coefficient
l	earth station latitude

L	earth station-to-satellite relative longitude also loss in link budget calculations, and loading factor of FDM/FM multiplex also message length (bits)
L_A	atmospheric attenuation
L_e	effective path length of radio wave through rain (km)
L_{FRX}	receiver feeder loss
L_{FTX}	transmitter feeder loss
L_{FS}	free space loss
L_{POINT}	depointing loss
L_{POL}	antenna polarisation mismatch loss
L_R	receiving antenna depointing loss
L_T	transmitting antenna depointing loss
m	satellite mass
mc	power reduction associated with multicarrier operation
M	mass of the earth (kg) (*also* number of possible states of a digital signal)
N_0	noise power spectral density (W/Hz)
$(N_0)_U$	uplink noise power spectral density (W/Hz)
$(N_0)_D$	downlink noise power spectral density (W/Hz)
$(N_0)_T$	total link noise power spectral density (W/Hz)
$(N_0)_I$	interference power spectral density (W/Hz)
N	noise power (W) (*also* number of stations in a network)
p	pre-emphasis/companding improvement factor (*also* rainfall annual percentage)
p_w	rainfall worst month time percentage
P	power (*also* number of bursts in a TDMA frame)
P_b	information bit error rate
P_c	channel bit error rate
P_{HPA}	rated power of high power amplifier (W)
P_T	power fed to the antenna (W)
P_{Tx}	transmitter power (W)
P_R	received power (W)
P_{Rx}	power at receiver input (W)
P_{is}	input power in a single-carrier operation mode
$P_{o\,1}$	output power in a single-carrier operation mode
$(P_{i\,1})_{sat}$	input power in a single-carrier operation mode at saturation
$(P_{o\,1})_{sat}$	saturation output power in a single-carrier operation mode
$P_{i\,n}$	input power in a multiple-carrier operation mode (n carriers)
$P_{o\,n}$	output power in a multiple-carrier operation mode (n carriers)
$P_{IMX\,n}$	power of intermodulation product of order X at output of a nonlinear device in a multicarrier operation mode (n carriers)
Q	quality factor
r	distance between centre of mass (orbits)
R	slant range from earth station to satellite (km) (also symbol or bit rate)
R_b	information bit rate (s^{-1})
R_c	channel bit rate (s^{-1})
R_{call}	mean number of calls per unit time
R_E	earth radius = 6378 km
R_o	geostationary satellite altitude = 35 786 km
R_p	rainfall rate (mm/h) exceeded for time percentage p of a year
R_s	symbol (or signalling) rate (s^{-1})
S	user signal power (W)
S/N	signal-to-noise power ratio at user's end
T	period of revolution (orbits) (s) (*also* noise temperature (K))
T_A	antenna noise temperature (K)
T_{AMB}	ambient temperature (K)
T_b	information bit duration (s)
T_B	burst duration (s)
T_c	channel bit duration (s)
T_e	effective input noise temperature of a four-port element system (K)
T_E	mean sidereal day = 86 164.15
T_{eATT}	effective input noise temperature of an attenuator (K)
T_{eRx}	effective input noise temperature of a receiver
T_F	frame duration (s) (*also* feeder temperature)
T_m	effective medium temperature (K)
T_0	reference temperature (290 K)
T_{eRX}	effective input noise temperature of a receiver (K)
T_S	symbol duration (s)
T_{SKY}	clear key contribution to antenna noise temperature (K)
T_{GROUND}	ground contribution to antenna noise temperature (K)
U	subscript for uplink
v	true anomaly (orbits)
V_s	satellite velocity (m/s)
$V_{Lp/p}$	peak-to-peak luminance voltage (V)
$V_{Tp/p}$	peak-to-peak total video signal voltage (including synchronisation pulses)
V_{Nms}	root-mean-square noise voltage (V)
w	psophometric weighting factor

X	intermodulation product order (IMX)
α	angle from boresight of antenna
γ	vernal point
Γ	spectral efficiency (bit/s Hz)
δ	declination angle (*also* delay)
η	antenna aperture efficiency
λ	wavelength (=c/f) (*also* longitude, message generation rate [s^{-1}])
φ	latitude
τ	propagation time
θ_{3dB}	half-power beamwidth of an antenna
	wavelength = c/f
θ_R	receiving antenna pointing error
θ_T	transmit antenna pointing error
μ	= GM (G = gravitational constant, M = mass of earth; $G = 6.67\times10^{-11}$ m^3 kg^{-1} s^{-2}, $M = 5.974\times10^{24}$ kg; $\mu = GM = 3.986\times10^{14}$ m^3 s^{-2}
ρ	code rate
σ	Stefan-Boltzmann constant = 5.67×10^{-8} Wm^{-2} K^{-4}
ϕ	satellite-earth station angle from the earth's centre
Φ	power flux density (w/m^2)
Φ_{max}	maximum power flux density at transmit antenna boresight
Φ_{nom}	nominal power flux density at receive end required to build up a given power assuming maximum receive gain (no depointing)
Φ_{sat}	power flux density required to operate receive amplifier at saturation
ψ	polarisation angle
ω	argument of perigee
Ω	right ascension of the ascending node
Ω_E	angular velocity of rotation of the earth = $15.0469°$ h^{-1} = $4.17\times10^{-3}°$ s^{-1} = 7.292×10^{-5} rad s^{-1}

1 INTRODUCTION

This chapter provides introductions to the characteristics of satellite communication systems and technology development. It aims to satisfy the curiosity of an impatient reader and facilitate a deeper understanding by directing him or her to appropriate chapters without imposing the need to read the whole work from beginning to end.

1.1 BIRTH OF SATELLITE COMMUNICATIONS

Satellite communications are the outcome of research in the area of communications and space technologies whose objective is to achieve ever-increasing ranges and capacities with the lowest possible costs.

The World War II stimulated the expansion of two very distinct technologies – missiles and microwaves. The expertise eventually gained in the combined use of these two techniques opened up the era of satellite communications. The service provided in this way usefully complements that previously provided exclusively by terrestrial networks using radio and cables.

The space era started in 1957 with the launching of the first artificial satellite (Sputnik). Subsequent years have been marked by various experiments including the following: Christmas greetings from President Eisenhower broadcast by Score (1958), the reflecting satellite ECHO (1960), store-and-forward transmission by the Courier satellite (1960), powered relay satellites (Telstar and Relay in 1962), and the first geostationary satellite Syncom (1963).

In 1965, the first commercial geostationary satellite Intelsat I (or Early Bird) inaugurated the long series of Intelsats; in the same year, the first Soviet communications satellite of the Molniya series was launched.

1.2 DEVELOPMENT OF SATELLITE COMMUNICATIONS

The first satellites provided a low capacity at a relatively high cost; for example, Intelsat I weighed 68 kg at launch for a capacity of 480 telephone channels and an annual cost of $32 500 per channel at the time. This cost resulted from a combination of the cost of the launcher, that of the satellite, the short lifetime of the satellite (1.5 years), and its low capacity. The reduction in cost is the result of much effort, which has led to the production of reliable launchers that

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

can put heavier and heavier satellites into orbit (typically 5900 kg at launch in 1975, reaching 10 500 kg by Ariane 5 ECA and 13 000 kg by Delta IV in 2008). Today, Delta IV Heavy is capable of sending a payload of 28 790 kg to low earth orbit (LEO) and 14 220 kg to geostationary transfer orbit (GTO); SpaceX Falcon Heavy can send payload of 63 700 kg to LEO, 26 700 kg to GTO, and 3500 kg to Mars.

In addition, increasing expertise in microwave techniques has enabled realisation of contoured multibeam antennas whose beams adapt to the shape of continents, frequency reuse from one beam to the other, and incorporation of higher-power transmission amplifiers. Increased satellite capacity has led to a reduced cost per telephone channel in recent history and now is calculated as reduction of the cost per bit in the digital age.

In addition to the reduction in the cost of communication, the most outstanding feature is the variety of services offered by satellite communications systems. Originally these were designed to carry communications from one point to another, as with cables, and the extended coverage of the satellite was used to set up long-distance links; hence Early Bird enabled stations on opposite sides of the Atlantic Ocean to be connected. However, as a consequence of the limited performance of the satellite, it was necessary to use earth stations equipped with large antennas and therefore of high cost (around $10 million for a station equipped with a 30 m diameter antenna).

The increasing size and power of satellites has permitted a consequent reduction in the size of earth stations, and hence their cost, leading to an increase in number from thousands to millions. In this way it has been possible to exploit another feature of the satellite: its ability to collect or broadcast signals from or to several locations. Instead of transmitting signals from one point to another, transmission can be from a single transmitter to a large number of receivers distributed over a wide area; or, conversely, transmission can be from a large number of stations to a single central station, often called a *hub*. In this way, multipoint data-transmission networks and data-collection networks have been developed under the name *very small aperture terminal networks* (VSATs) [MAR-95]. Over 1 000 000 VSATs had been installed up to 2008 and about 6 000 000 in 2018.

For TV services, satellites are of paramount importance for satellite news gathering (SNG), for the exchange of programmes between broadcasters, and for distributing programmes to terrestrial broadcasting stations and cable heads, or directly to the individual consumer. The latter are commonly called *direct broadcasting by satellite* (DBS) systems, or direct-to-home (DTH) systems. A rapidly growing service is digital video broadcasting by satellite (DVB-S), developed in early 1991; the second generation (DVB-S2) has been standardised by the European Telecommunication Standard Institute (ETSI); and DVB-S2X as an extension of DVB-S2 was completed in 2014. These DBS systems operate with small earth stations having antennas with a diameter from 0.5 to 1 m.

In the past, customer stations were receive only (RCVO) stations. With the introduction of two-way communications stations, satellites are a key component in providing interactive TV and broadband Internet services, thanks to the implementation of the digital video broadcasting satellite return channel (DVB-RCS) standard for service providers' facilities that was started in 1999 and completed in 2008; DVB-RCS2 as the next generation of DVB-RCS, completed in 2009; and DVB-RCS2X as an extension of DVB-RCS2, which became an ETSI standard in 2014. It uses Transmission Control Protocol (TCP)/Internet Protocol (IP) to support Internet, multicast, and web-page caching services over satellite with the forward channel operating at several Mbps and enables satellites to provide broadband service applications for the end user, such as direct access and distribution services. IP-based triple-play services (telephony, Internet, and TV) are more and more popular. Satellites cannot compete with terrestrial asymmetric digital subscriber line (ADSL) or cable to deliver these services in high-density population areas. However, they complement nicely the terrestrial networks around cities and in rural areas where the distance to the telephone router is too large to allow delivery of the several Mbps required to run the service.

A further reduction in the size of the earth station antenna is exemplified in digital audio broadcasting (DAB) systems, with antennas on the order of 10 cm. The satellite transmits multiplexed digital audio programmes and supplements traditional Internet services by offering one-way broadcast of web-style content to the receivers.

Furthermore, satellites are effective in mobile communications. Since the end of the 1970s, INMARSAT satellites have been providing distress-signal services along with telephone and data communications services to ships and planes and, more recently, to portable earth stations (Mini M or satphone). Personal mobile communication using small handsets is available from constellations of non-geostationary satellites (such as Iridium and Globalstar), and geostationary satellites equipped with very large deployable antennas (typically 10–15 m and today can be more than 25 m), as with the Intelsat, Inmarsat, and Eutelsat satellites. The next step in bridging the gaps between fixed, mobile, and broadcasting radiocommunications services concerns satellite multimedia broadcast to fixed and mobile users. Satellite digital mobile broadcasting (SDMB) is based on hybrid integrated satellite–terrestrial systems to serve small hand-held terminals with interactivity.

Finally, high-throughput satellites (HTSs) have taken advantage of technology developments, further reducing the cost per bit over satellite with increased total capacity of a satellite from megabits per second (Mbps) to gigabits per second (Gbps), or even terabits per second (Tbps). Further, mega-LEO satellite constellations such as OneWeb will be able to have hundreds or even thousands of satellites delivering a total capacity of 7 Tbps.

1.3 CONFIGURATION OF A SATELLITE COMMUNICATIONS SYSTEM

Figure 1.1 gives an overview of a satellite communication system and illustrates its interfacing with terrestrial entities. The satellite system is composed of a space segment, a control segment, and a ground segment:

— The *space segment* contains one or several active and spare satellites organised into a constellation.
— The *control segment* consists of all ground facilities for the control and monitoring of the satellites, also named tracking, telemetry, and command (TTC) stations, and for the management of the traffic and the associated resources on board the satellite for communication networks.
— The *ground segment* consists of all the traffic earth stations. Depending on the type of service considered, these stations can be of different size, from a few centimetres to tens of metres.

Table 1.1 gives examples of traffic earth stations in connection with the types of service discussed in Section 1.7. Earth stations come in three classes, as illustrated in Figure 1.1: *user stations*, such as handsets, portables, mobile stations, and VSATs, which allow the customer direct access to the space segment; *interface stations*, known as *gateways*, which interconnect the space segment to a terrestrial network; and *service stations*, such as hub or feeder stations, which collect or distribute information from and to user stations via the space segment.

Communications between users are set up through *user terminals* – which consisted of equipment such as telephone sets, fax machines, and computers in the past and laptops and smartphones today – that are connected to the terrestrial network or to the user station (e.g. a VSAT), or are part of the user station (e.g. if the terminal is mobile).

The path from a source user terminal to a destination user terminal is called a *simplex connection*. There are two basic schemes: *single connection per carrier* (SCPC), where the modulated carrier supports one connection only; and *multiple connections per carrier* (MCPC), where the

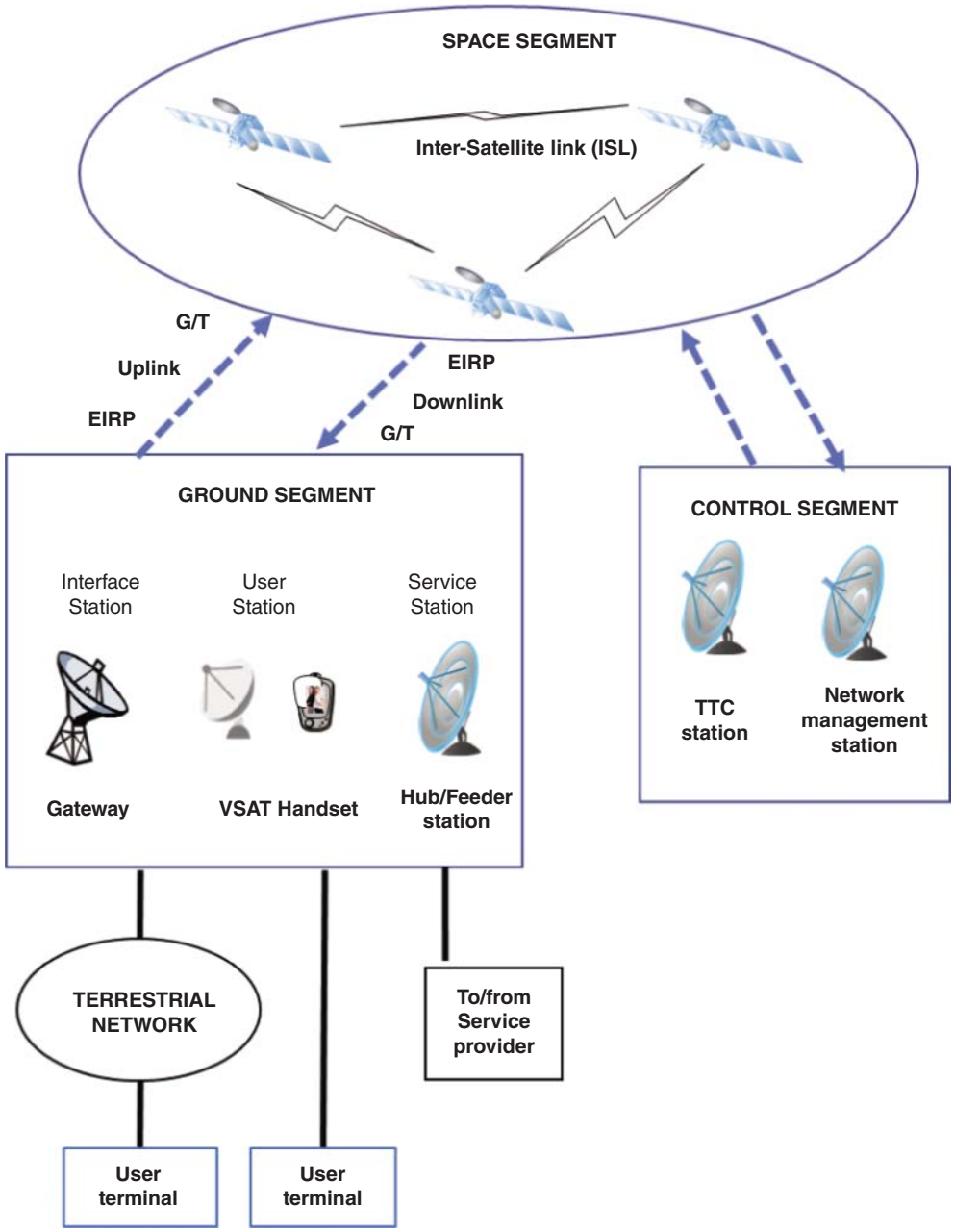


Figure 1.1 Satellite communications system interfacing with terrestrial entities.

Table 1.1 Services from different types of traffic and earth stations

Type of service	Type of earth station	Typical size (m)
Point-to-point	Gateway, hub	2–10
	VSAT	1–2
Broadcast/multicast	Feeder station	1–5
	VSAT	0.5–1.0
Collect	VSAT	0.1–1.0
	Hub	2–10
Mobile	Handset, portable, mobile	0.1–0.5
	Gateway	2–10

modulated carrier supports several time- or frequency-multiplexed connections. Interactivity between two users requires a duplex connection between their respective terminals, i.e. two simplex connections, each in one direction. Each user terminal should then be capable of sending and receiving information.

A connection between a service provider and a user goes through a hub (for collecting services) or a feeder station (e.g. for broadcasting services). A connection from a gateway, hub, or feeder station to a user terminal is called a *forward* connection. The reverse connection is the *return* connection. Both forward and return connections entail an uplink and a downlink, and possibly one or more intersatellite links (ISLs).

1.3.1 Communications links

A link between transmitting equipment and receiving equipment consists of a radio or optical modulated carrier. The performance of the transmitting equipment is measured by its *effective isotropic radiated power* (EIRP), which is the power fed to the antenna multiplied by the gain of the antenna in the considered direction. The performance of the receiving equipment is measured by G/T, the ratio of the antenna receive gain, G, in the considered direction and the system noise temperature, T; G/T is called the receiver's *figure of merit*. These concepts are detailed in Chapter 5.

The types of link shown in Figure 1.1 are:

— *Uplinks* from the earth stations to the satellites
— *Downlinks* from the satellites to the earth stations
— *Intersatellite links* between the satellites

Uplinks and downlinks consist of radio frequency modulated carriers, while ISLs can be either radio frequency or optical. Some large-capacity data-relay satellites also use optical links with their ground stations. Carriers are modulated by baseband signals conveying information for communications purposes.

The link performance can be measured by the ratio of the received carrier power, C, to the noise power spectral density, N_0, and is denoted as the C/N_0 ratio, expressed in hertz (Hz). The values of C/N_0 for the links that participate in the connection between the end terminals determine the quality of service, specified in terms of *bit error rate* (BER) for digital communications.

Another parameter of importance for the design of a link is the bandwidth, B, occupied by the carrier. This bandwidth depends on the information data rate, the channel coding rate (forward error correction, [FEC]), and the type of modulation used to modulate the carrier. For satellite

links, the trade-off between required carrier power and occupied bandwidth is paramount to the cost-effective design of the link. This is an important aspect of satellite communications, as power impacts both satellite mass and earth station size, and bandwidth is constrained by regulations.

According to the Shannon-Hartley theorem, the maximum rate at which information can be transmitted over a communication channel of a specified bandwidth in the presence of noise can be calculated as the following:

$$R = B \log_2(1 + S/N)$$

where R is the maximum rate, B is the bandwidth, S is the signal power, and N is the noise power.

Moreover, a service provider that rents satellite transponder capacity from the satellite operator is charged according to the highest share of either power or bandwidth resource available from the satellite transponder. The service provider's revenue is based on the number of established connections, so the objective is to maximise the throughput of the considered link while keeping a balanced share of power and bandwidth usage. This is discussed in Chapter 4.

In a satellite system, several stations transmit their carriers to a given satellite, and therefore the satellite acts as a network node. The techniques used to organise access to the satellite by the carriers are called *multiple-access techniques* (Chapter 6).

1.3.2 The space segment

A satellite consists of the *payload* and the *platform*. The payload consists of the receiving and transmitting antennas and all the electronic equipment that supports the transmissions of the carriers. The two types of payload organisation are illustrated in Figure 1.2. Some experiments have also been carried out for IP routers on board satellites.

Figure 1.2a shows a *transparent* payload (sometimes called a *bent pipe* type) where carrier power is amplified and frequency is downconverted. Power gain is on the order of 100–130 dB required to raise the power level of the received carrier from a few tens of picowatts to the power level of the carrier fed to the transmit antenna (a few watts to a few tens of watts). Frequency conversion is required to increase isolation between the receiving input and the transmitting output. Due to technology power limitations, the overall satellite payload bandwidth is split into several sub-bands, and the carriers in each sub-band are amplified by a dedicated power amplifier. The amplifying chain associated with each sub-band is called a *satellite channel*, or transponder. The bandwidth splitting is achieved using a set of filters called the *input multiplexer* (IMUX). The amplified carriers are recombined in the *output multiplexer* (OMUX).

The transparent payload in Figure 1.2a belongs to a single-beam satellite where each transmit and receive antenna generates one beam only. One could also consider multiple-beam antennas. The payload would then have as many inputs/outputs as upbeams/downbeams. Routing of carriers from one upbeam to a given downbeam implies either routing through different satellite channels; *transponder hopping*, depending on the selected uplink frequency; or *on-board switching with transparent on-board processing*. These techniques are presented in Chapter 7.

Figure 1.2b shows a multiple-beam *regenerative* payload where the uplink carriers are demodulated. The availability of the baseband signals allows *on-board processing* and routing of information from upbeam to downbeam through *on-board switching at baseband*. The frequency conversion is achieved by modulating on-board-generated carriers at downlink frequency. The modulated carriers are then amplified and delivered to the destination downbeam.

Figure 1.3 illustrates a multiple-beam satellite antenna and its associated coverage areas. Each beam defines a *beam coverage area*, also called a *footprint*, on the earth surface. The aggregate beam coverage areas define the *multibeam antenna coverage area*. A given satellite may have several multiple-beam antennas, and their combined coverage defines the *satellite coverage area*.

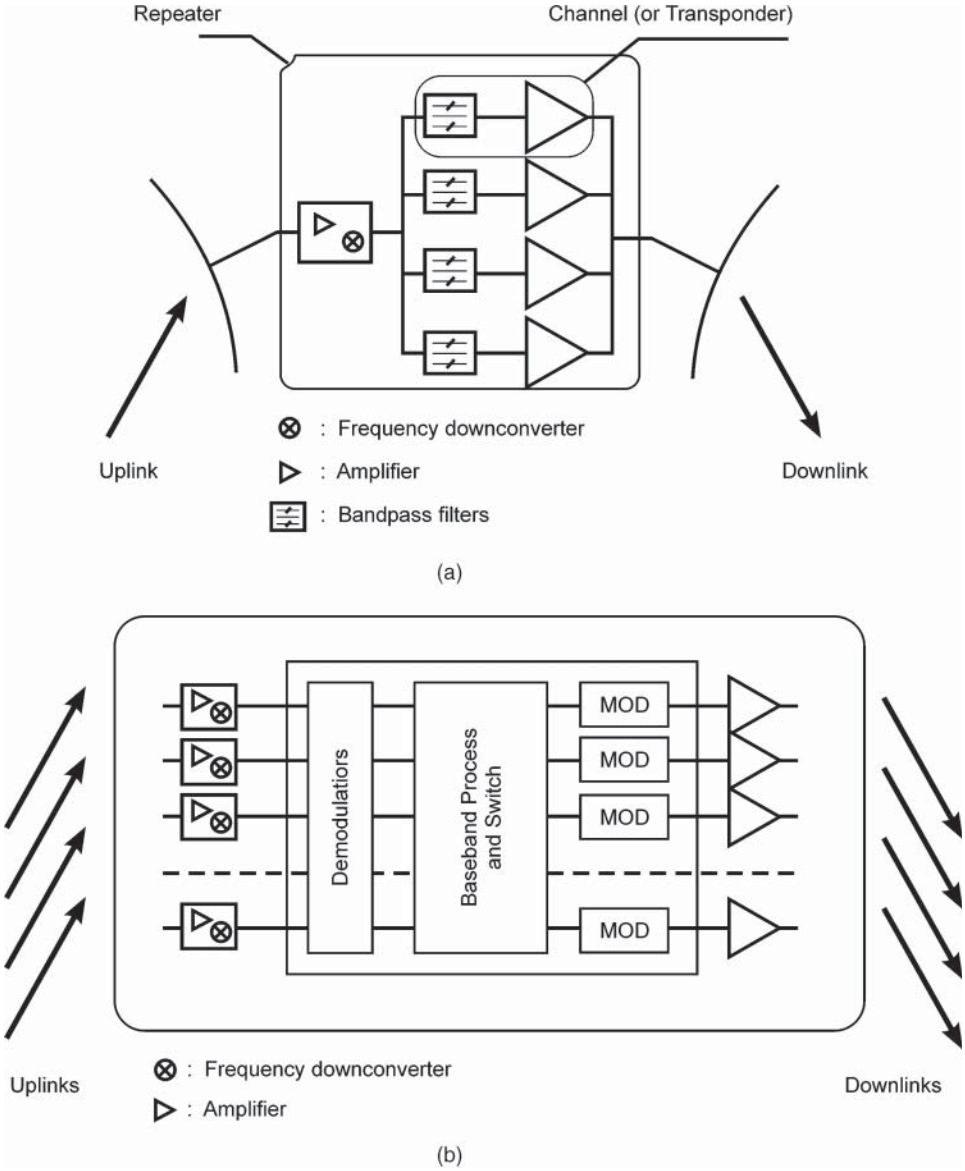


Figure 1.2 Payload organisation: (a) transparent; (b) regenerative.

Bandwidth reuse based on multibeam antennas is a key technology to achieve high capacity for HTS to reduce the cost per bit for information delivery. The available bandwidth can be divided into three or four sub-bands (also called three- or four-colour techniques, according to arrangement of the spot beams) so that different sub-bands (colours) can be allocated to different spot beams; adjacent spot beams use different sub-bands (colours) to avoid interference between adjacent spot beams. Figure 1.3b shows an example of four-colour reuse.

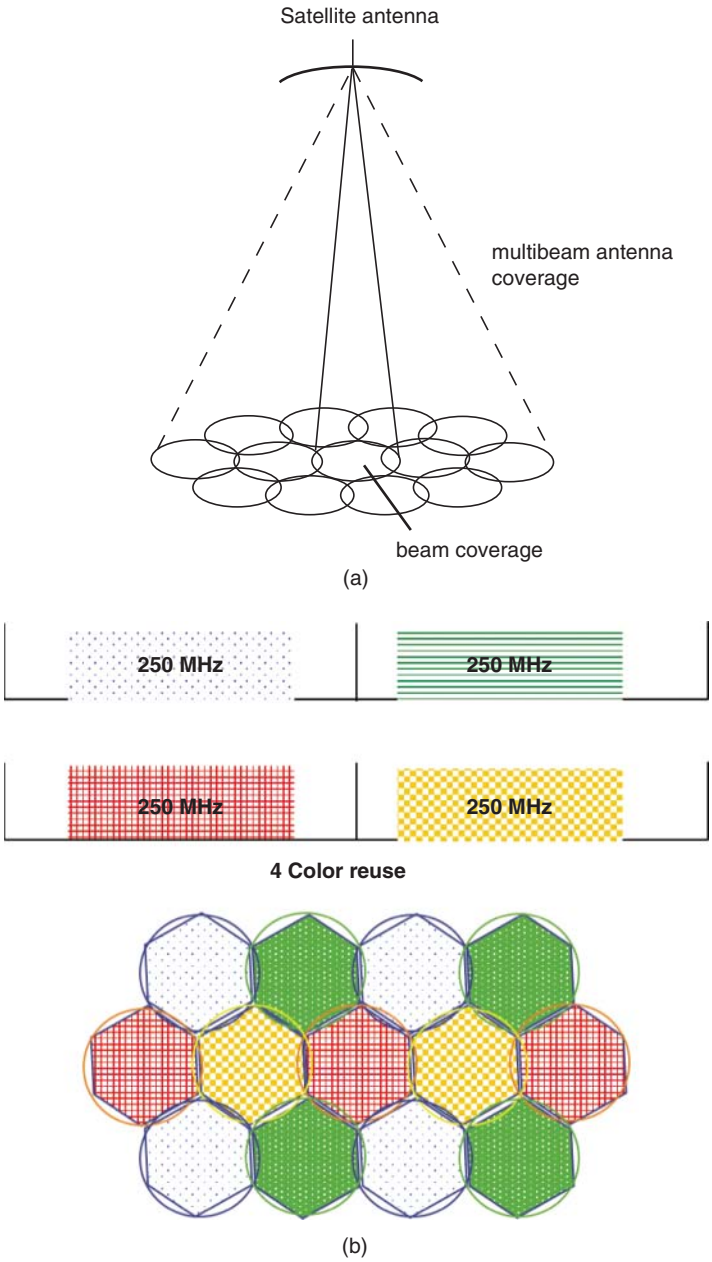


Figure 1.3 (a) Multiple-beam satellite antenna and associated coverage area; (b) example of four-colour reuse.

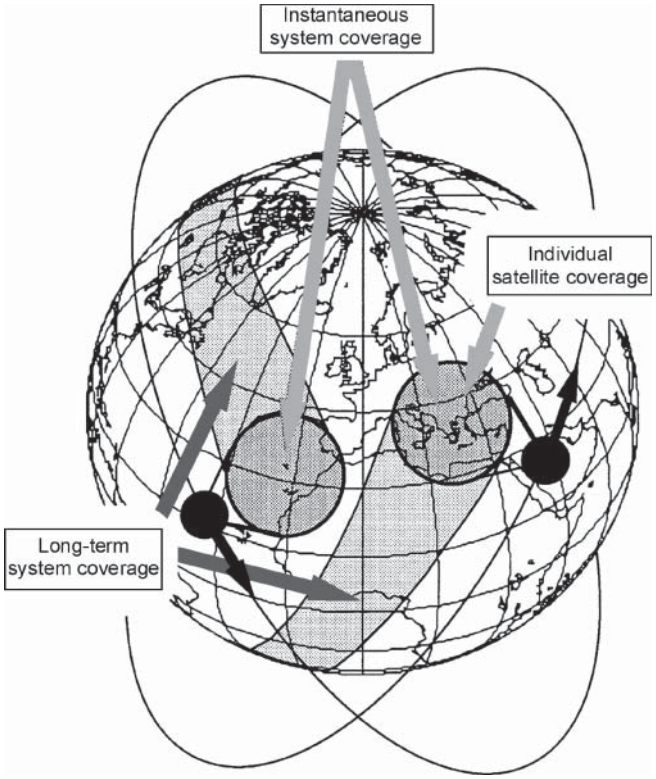


Figure 1.4 Types of coverage.

Figure 1.4 illustrates the concept of instantaneous system coverage and long-term coverage. *Instantaneous system coverage* consists of the aggregation at a given time of the coverage areas of the individual satellites participating in the constellation. *Long-term coverage* is the area on the earth scanned over time by the antennas of the satellites in the constellation.

The coverage area should encompass the *service zone*, which corresponds to the geographical region where the stations are installed. For real-time services, the instantaneous system coverage should at any time have a footprint covering the service zone, while for non-real-time (store-and-forward) services, it should have long-term coverage of the service zone.

For LEO or medium earth orbit (MEO) satellites, a large number of satellites are needed to provide continuous global coverages. In cases of LEO, the Iridium Next (constellation of second-generation Iridium satellites) has 66 satellites plus 6 spares; OneWeb plans to have 648 satellites plus 252 spares; and Starlink by SpaceX plans to have 4425 satellites plus some spares. In case of MEO, O3b has 20 satellites including 3 on-orbit spares, all operating in equatorial orbit.

The platform consists of all the subsystems that permit the payload to operate. Table 1.2 lists these subsystems and indicates their respective main functions and characteristics.

The detailed architecture and technology of the payload equipment are explained in Chapter 9. The architecture and technologies of the platform are considered in Chapter 10. The operations of orbit injection and the various types of launcher are the subject of Chapter 11. The space environment and its effects on the satellite are presented in Chapter 12.

Table 1.2 Platform subsystems

Subsystem	Principal functions	Characteristics
Attitude and orbit control system (AOCS)	Attitude stabilisation, orbit determination	Accuracy
Propulsion	Provision of velocity increments	Specific impulse, mass of propellant
Electric power supply	Provision of electrical energy	Power, voltage stability
Telemetry, tracking, and command (TTC)	Exchange of housekeeping information	Number of channels, security of communications
Thermal control	Temperature maintenance	Dissipation capability
Structure	Equipment support	Rigidity, lightness

To ensure service with a specified availability, a satellite communication system must make use of several satellites in order to provide redundancy. A satellite can cease to be available due to a failure or because it has reached the end of its lifetime. In this respect, it is necessary to distinguish between the reliability and the lifetime of a satellite. *Reliability* is a measure of the probability of a breakdown and depends on the reliability of the equipment and any schemes to provide redundancy. The *lifetime* is conditioned by the ability to maintain the satellite on station in the nominal attitude, and depends on the quantity of fuel available for the propulsion system and attitude and orbit control system (AOCS). In a system, provision is generally made for an operational satellite, a backup satellite in orbit, and a backup satellite on the ground. The reliability of the system involves not only the reliability of each of the satellites but also the reliability of launching. An approach to these problems is treated in Chapter 13.

1.3.3 The ground segment

The ground segment consists of all the earth stations; these are most often connected to the end user's terminal by a terrestrial network or, in the case of small station VSATs, directly connected to the end user's terminal. Stations are distinguished by their size, which varies according to the volume of traffic to be carried on the satellite link and the type of traffic (telephone, television, data, or multimedia Internet services). In the past, the largest were equipped with antennas 30 m diameter (Standard A of the Intelsat network). The smallest have 0.6 m antennas (receiving stations from DBSs) or even smaller (0.1 m) antennas (mobile stations, portable stations, or handsets). Some stations both transmit and receive. Others are RCVO stations; this is the case, for example, with receiving stations for a broadcasting satellite system or a distribution system for television or data signals. Figure 1.5 shows the typical architecture of an earth station for both transmission and reception. Chapter 5 introduces the characteristic parameters of the earth station that appear in the link budget calculations. Chapter 3 presents the characteristics of signals supplied to earth stations by the user terminal, either directly or through a terrestrial network; the signal processing at the station (such as source coding and compression, multiplexing, channel coding, scrambling, and encryption); and transmission and reception (including modulation and demodulation). Chapter 7 explains the concepts of satellite networks, as satellite communication systems are integrated more closely with terrestrial networks to provide broadband multimedia services as well as mobile networks services. Chapter 8 treats the organisation and equipment of earth stations.

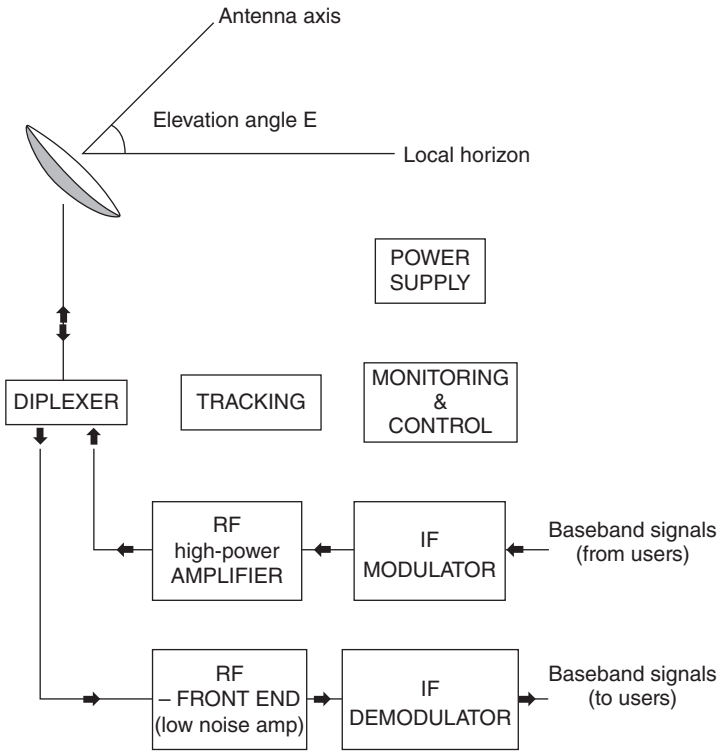


Figure 1.5 The organisation of an earth station. RF = radio frequency, IF = intermediate frequency.

1.4 TYPES OF ORBIT

The *orbit* is the trajectory followed by the satellite. The trajectory is within a plane and shaped like an ellipse with a maximum extension at the apogee and a minimum at the perigee. The satellite moves more slowly in its trajectory as the distance from the earth increases, according to the laws of physics. Chapter 2 provides a definition of the orbital parameters.

The most favourable orbits are as follows:

— *Elliptical orbits* inclined at an angle of 64° with respect to the equatorial plane. This type of orbit is particularly stable with respect to irregularities in terrestrial gravitational potential and, owing to its inclination, enables the satellite to cover regions of high latitude for a large fraction of the orbital period as it passes to the apogee. This type of orbit has been adopted by Russia for the satellites of the Molniya system with a period of 12 hours. Figure 1.6 shows the geometry of the orbit. The satellite remains above the regions located under the apogee for a time interval on the order of eight hours. Continuous coverage can be ensured with three phased satellites in different orbits. Several studies relate to elliptical orbits with a period of 24 hours (Tundra orbits) or a multiple of 24 hours. These orbits are particularly useful for satellite systems for communication with mobiles, where the masking effects caused by surrounding obstacles such as buildings and trees and multiple-path

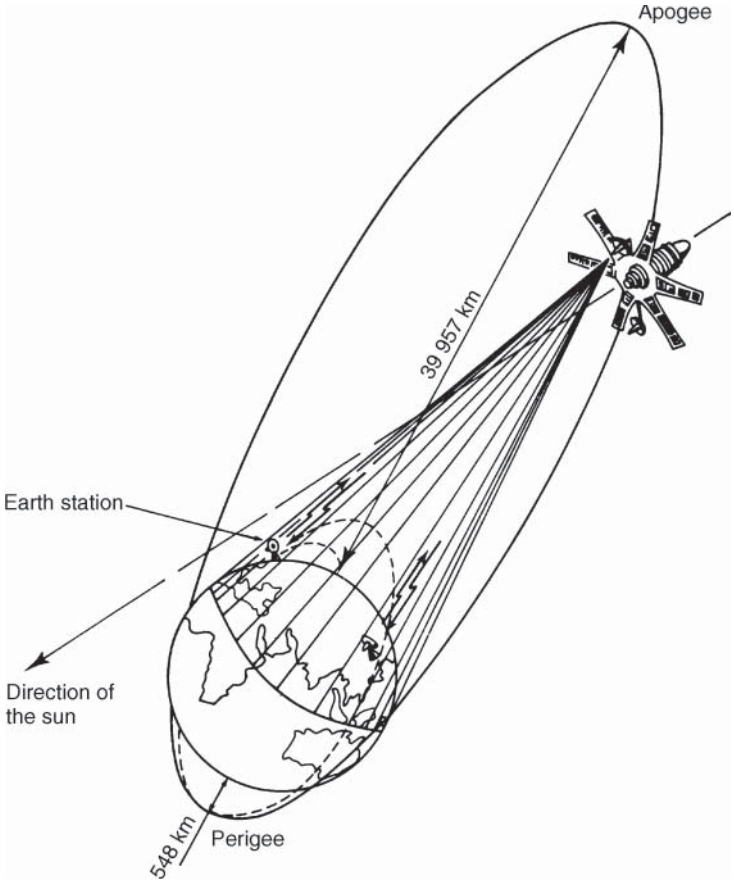


Figure 1.6 The orbit of a Molniya satellite.

effects are pronounced at low elevation angles (say, less than 30°). In fact, inclined elliptic orbits can provide the possibility of links at medium latitudes when the satellite is close to the apogee with elevation angles close to 90°; these favourable conditions cannot be provided at the same latitudes by geostationary satellites. In the late 1980s, the European Space Agency (ESA) studied the use of elliptical highly inclined orbits (HEOs) for DAB and mobile communications in the framework of its Archimedes programme. The concept became reality at the end of the 1990s with the Sirius system delivering satellite digital audio radio services to millions of subscribers (mainly automobiles) in the United States using three satellites in HEO Tundra-like orbits [AKT-08]. Both Molnya and Tundra orbits provide users with higher elevation angles than geostationary earth orbit (GEO) orbit at high latitude.

— *Circular LEOs:* The altitude of the satellite is constant and equal to several hundreds of kilometres. The period is on the order of one and a half hours. With near 90° inclination, this type of orbit guarantees worldwide long-term coverage as a result of the combined motion of the satellite and earth rotation, as shown in Figure 1.7. This is the reason for choosing this type of orbit for observation satellites (for example, the SPOT satellite: altitude

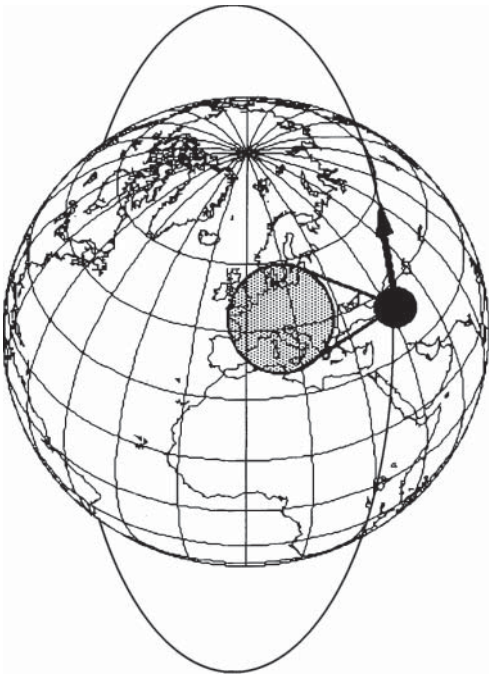


Figure 1.7 Circular polar low earth orbit (LEO).

830 km, orbit inclination 98.7°, period 101 minutes). One can envisage the establishment of store-and-forward communications if the satellite is equipped with a means of storing information. A constellation of several tens of satellites in low-altitude (e.g. IRIDIUM with 66 satellites at 780 km) circular orbits can provide worldwide real-time communication (see Figure 1.8). Non-polar orbits with less than 90° inclination can also be envisaged. For instance, the GLOBALSTAR constellation incorporates 48 satellites at 1414 km with 52° orbit inclination.

— *Circular MEOs*, also called intermediate circular orbits (ICOs), have an altitude of about 10 000 km and an inclination of about 50°. The period is six hours. With constellations of about 10–15 satellites, continuous coverage of the world is guaranteed, allowing worldwide real-time communications. A planned system of this kind was the ICO system (which emerged from Project 21 of INMARSAT but was not implemented) with a constellation of 10 satellites in two planes at 45° inclination. O3b is a special case of a MEO circular orbit satellite constellation with altitude at 8063 km and 20 satellites. Each satellite has 12 steerable Ka band antennas of which 2 are for gateways and 10 are for user terminals (see Figure 1.9).

— *Circular orbits with zero inclination (equatorial orbits)*: The most popular is the geostationary satellite orbit; the satellite orbits around the earth in the equatorial plane according to the earth's rotation at an altitude of 35 786 km. The period is equal to that of the rotation of the earth. The satellite thus appears as a point fixed in the sky and ensures continuous operation as a radio relay in real time for the area of visibility of the satellite (43% of the earth's surface).

— *Hybrid systems*: Some systems may include combinations of circular and elliptical orbits. Studies have been carried out to determine how to combine satellites from different orbits to achieve communication and network objectives, though these have been used in navigation satellite systems.

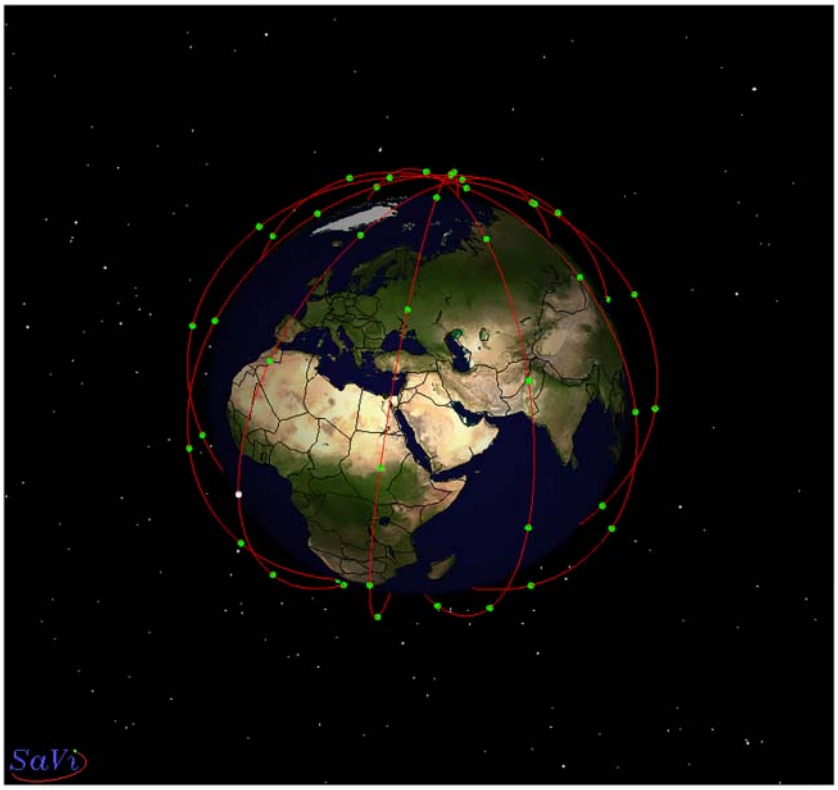


Figure 1.8 Illustration of Iridium as an example of a low earth orbit (LEO) satellite constellation.

The choice of orbit depends on the nature of the mission, the acceptable interference, and the performance of the launchers:

— *The extent and latitude of the area to be covered*: Contrary to widespread opinion, the altitude of the satellite is not a determining factor in the link budget for a given earth coverage. Chapter 5 shows that the propagation attenuation varies as the inverse square of the distance, and this favours a satellite following a low orbit on account of its low altitude; however, this disregards the fact that the area to be covered is then seen through a larger solid angle. The result is a reduction in the gain of the satellite antenna, which offsets the distance advantage. A satellite following a low orbit provides only limited earth coverage at a given time and limited time at a given location. Unless low-gain antennas (on the order of a few dB) that provide low directivity and hence almost omnidirectional radiation are installed, earth stations must be equipped with satellite-tracking devices, which increases the cost. The geostationary satellite thus appears to be particularly useful for continuous coverage of extensive regions. However, it does not permit coverage of the polar regions, which are accessible by satellites in inclined elliptical orbits or polar orbits.
— *The elevation angle*: A satellite in an inclined or polar elliptical orbit can appear overhead at certain times, which enables communication to be established in urban areas without encountering the obstacles that large buildings constitute for elevation angles between 0°

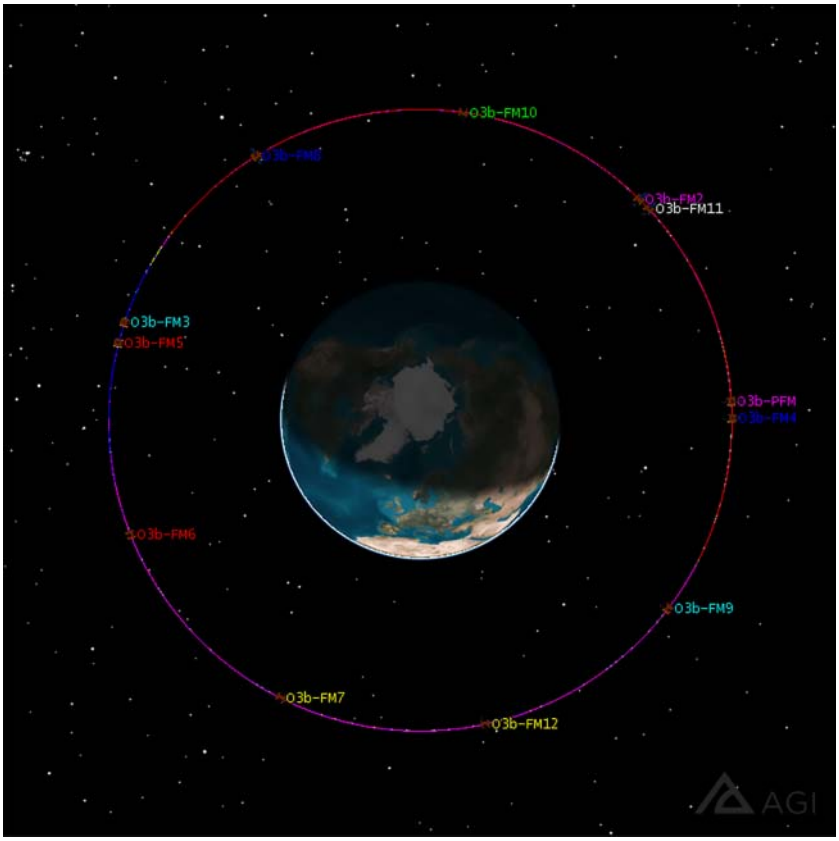


Figure 1.9 Illustration of O3B as an example of a medium earth orbit (MEO) satellite constellation.

and approximately 70°. With a geostationary satellite, the angle of elevation decreases as the difference in latitude or longitude between the earth station and the satellite increases.

— *Transmission duration and delay*: A geostationary satellite provides a continuous relay for stations within visibility, but the propagation time of the radio waves from one station to the other is on the order of 0.25 seconds. This requires the use of echo control devices on telephone channels or special protocols for data transmission. For the Internet, performance enhancement protocols (PEPs) have been introduced for efficient utilisation of satellite link resources. A satellite moving in a low orbit confers a reduced propagation time. The transmission time is thus low between stations that are close and simultaneously visible to the satellite, but it can become long (several hours) for distant stations if only store-and-forward transmission is considered. For large mega-LEO satellite constellations, complicated dynamic routing mechanisms are needed, and the satellites in the constellations must be managed.

— *Interference*: Geostationary satellites occupy fixed positions in the sky with respect to the stations with which they communicate. Protection against interference between systems is ensured by planning the frequency bands and orbital positions. The small orbital spacing between adjacent satellites operating at the same frequencies leads to an increase in the level of interference, and this impedes the installation of new satellites. Different systems could use different frequencies, but this is restricted by the limited number of frequency bands

assigned for space radiocommunications by the International Telecommunication Union (ITU) Radio Regulations (RR). In this context, one can refer to an *orbit-spectrum* resource that is limited. With orbiting satellites, the geometry of each system changes with time, and the relative geometries of one system with respect to another are variable and difficult to synchronise. The probability of interference is thus high.

— *The performance of launchers*: The mass that can be launched decreases as the altitude increases.

The geostationary satellite is certainly the most popular. At the present time there are around 600 geostationary satellites in operation within the 360° of the whole orbital arc. Some parts of this orbital arc, however, tend to be highly congested (for example, above the American continent and Europe). Figure 1.10 illustrates the satellite orbit altitudes (LEO/MEO/GEO) and coverage areas.

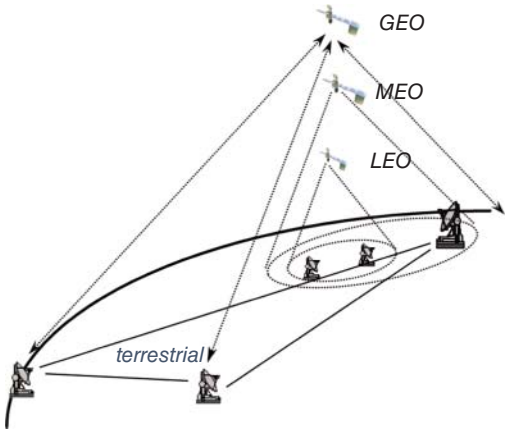


Figure 1.10 Illustration of orbit altitudes and coverages.

1.5 RADIO REGULATIONS

The latest publication is the 2016 edition of the ITU Radio Regulations Articles [ITU-16], freely available from the ITU publications on the Internet. Radio regulations are necessary to ensure an efficient and economical use of the radio-frequency spectrum by all communications systems, both terrestrial and satellite. While so doing, the sovereign right of each state to regulate its telecommunications must be preserved. It is the role of the ITU to promote, coordinate, and harmonise the efforts of its members to fulfil these possibly conflicting objectives. Further studies have been carried out for global satellite communications [ITU-12].

1.5.1 The ITU organisation

The ITU, a United Nations organ, operates under a convention adopted by its member administrations. The ITU publishes the Radio Regulations (RR), which are reviewed by the delegates from ITU member administrations at periodic World Radio Conferences/Regional Radio Conferences (WRCs/RRCs).

From 1947 to 1993, technical and operational matters were administered by two committees: the CCIR (Comité Consultatif International des Radiocommunications) and the CCITT (Comité Consultatif International Télégraphique et Téléphonique). The International Frequency Registration Board (IFRB) was responsible for the examination of frequency-use documentation submitted to the ITU by its member administrations, in compliance with the RR, and for maintaining the Master International Frequency Register (MIFR).

Since 1994, the ITU has been reorganised into three sectors:

— The Radiocommunications Sector (ITU-R) deals with all regulatory and technical matters that were previously handled, respectively, by the IFRB and the CCIR.
— The Telecommunication Standardisation Sector (ITU-T) continues the work of the CCITT, and those studies by the CCIR dealing with the interconnection of radiocommunications systems with public networks.
— The Development Sector (ITU-D) acts as a forum and an advisory structure for the harmonious development of communications in the world.

The abundant and useful technical literature previously published in the form of reports and recommendations by the CCIR and the CCITT has now been reorganised in the form of ITU-R and ITU-T series recommendations.

1.5.2 Space radiocommunications services

The RR refer to the following space radiocommunications services, defined as transmission or reception of radio waves for specific telecommunications applications [ITU-16]:

— *Fixed-satellite service (FSS)*: A radiocommunication service between earth stations at given positions, when one or more satellites are used. The given position may be a specified fixed point or any fixed point within specified areas. In some cases this service includes satellite-to-satellite links, which may also be operated in the inter-satellite service (ISS). The FSS may also include feeder links for other space radiocommunication services.
— *Mobile satellite service (MSS)*: A radiocommunication service between mobile earth stations and one or more space stations, or between space stations used by this service; or between mobile earth stations by means of one or more space stations. This service may also include feeder links necessary for its operation.
— *Broadcasting satellite service (BSS)*: A radiocommunication service in which signals transmitted or retransmitted by space stations are intended for direct reception by the general public. In the BSS, the term *direct reception* encompasses both individual reception and community reception.
— *Earth exploration satellite service (EES)*: A radiocommunication service between earth stations and one or more space stations, which may include links between space stations, in which: information relating to the characteristics of the earth and its natural phenomena, including data relating to the state of the environment, is obtained from active sensors or passive sensors on earth satellites. Similar information is collected from airborne or earth-based platforms; such information may be distributed to earth stations within the system concerned, and platform interrogation may be included. This service may also include feeder links necessary for its operation.
— *Space research service (SRS)*: A radiocommunication service in which spacecraft or other objects in space are used for scientific or technological research purposes.

— *Space operation service (SOS)*: A radiocommunication service concerned exclusively with the operation of spacecraft, in particular space tracking, space telemetry, and space telecommand. These functions will normally be provided within the service in which the space station is operating.
— *Radiodetermination satellite service (RSS)*: *A* radiocommunication service for the purpose of radiodetermination involving the use of one or more space stations. This service may also include feeder links necessary for its own operation.
— *Inter-satellite service (ISS)*: A radiocommunication service providing links between artificial satellites.
— *Amateur satellite service (ASS)*: A radiocommunication service using space stations on earth satellites for the same purposes as those of the amateur service.

The main services for satellite communications are FSS, MSS, and BSS. Now all shift from the traditional fixed voice and data services toward mobile IP-based broadband multimedia Internet services; and from basic channel-based standard TV services to HD, 4K and even 8K TV on-demand services.

1.5.3 Frequency allocation

Frequency bands are allocated to the various radiocommunications services to allow compatible use. The allocated bands can be either exclusive for a given service or shared among several services. *Allocations* refer to the following division of the world into three regions (refer to Figure 1.11):

— *Region 1:* Includes the area limited on the east by line A (lines A, B, and C are defined shortly) and on the west by line B, excluding any of the territory of the Islamic Republic of Iran that lies between these limits. It also includes the whole of the territory of Armenia, Azerbaijan, the Russian Federation, Georgia, Kazakhstan, Mongolia, Uzbekistan, Kyrgyzstan, Tajikistan, Turkmenistan, Turkey, and Ukraine and the area to the north of the Russian Federation that lies between lines A and C.
— *Region 2:* Includes the area limited on the east by line B and on the west by line C.
— *Region 3:* Includes the area limited on the east by line C and on the west by line A, except any of the territory of Armenia, Azerbaijan, the Russian Federation, Georgia, Kazakhstan, Mongolia, Uzbekistan, Kyrgyzstan, Tajikistan, Turkmenistan, Turkey, and Ukraine and the area to the north of the Russian Federation. It also includes that part of the territory of the Islamic Republic of Iran lying outside of those limits.

The lines A, B, and C are defined as follows:

— *Line A:* Extends from the North Pole along meridian 40° East of Greenwich to parallel 40° North; thence by great circle arc to the intersection of meridian 60° East and the Tropic of Cancer; thence along the meridian 60° East to the South Pole.
— *Line B:* Extends from the North Pole along meridian 10° West of Greenwich to its intersection with parallel 72° North; thence by great circle arc to the intersection of meridian 50° West and parallel 40° North; thence by great circle arc to the intersection of meridian 20° West and parallel 10° South; thence along meridian 20° West to the South Pole.
— *Line C:* Extends from the North Pole by great circle arc to the intersection of parallel 65° 30′ North with the international boundary in the Bering Strait; thence by great circle arc to the intersection of meridian 165° East of Greenwich and parallel 50° North; thence by great circle

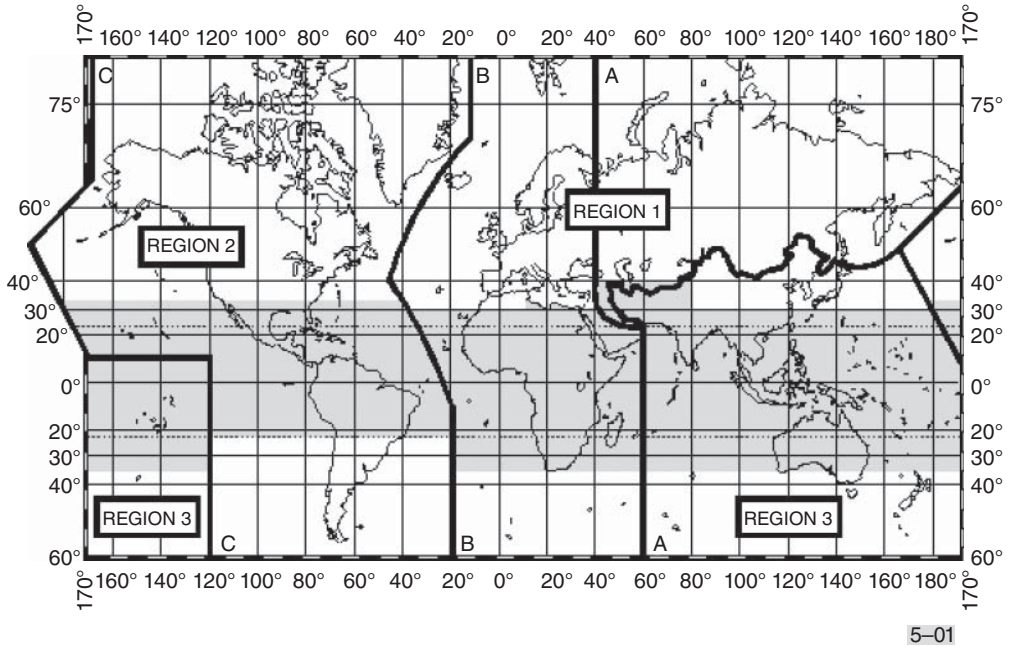


Figure 1.11 Map of regions and areas for frequency allocation in the ITU Radio Regulations (RR) [ITU-16].

arc to the intersection of meridian 170° West and parallel 10° North; thence along parallel 10° North to its intersection with meridian 120° West; thence along meridian 120° West to the South Pole.

For example, the FSS makes use of the following bands:

— Around 6 GHz for the uplink and around 4 GHz for the downlink (systems described as 6/4 GHz or C band). These bands are occupied by the oldest systems (such as Intelsat, American domestic systems, etc.) and tend to be saturated.
— Around 8 GHz for the uplink and around 7 GHz for the downlink (systems described as 8/7 GHz or X band). These bands are reserved, by agreement between administrations, for government use.
— Around 14 GHz for the uplink and around 12 GHz for the downlink (systems described as 14/12 GHz or Ku band). This corresponds to current operational developments (such as Eutelsat, etc.).
— Around 30 GHz for the uplink and around 20 GHz for the downlink (systems described as 30/20 GHz or Ka band).

Since 2010, large numbers of satellites have been launched and many more operational satellites have been planned on Ka band to exploit the benefit of its available large bandwidth. In combination with multispot beams and bandwidth-reuse technologies in Ka band, the capacity of each satellite has been increased significantly with 10–100 fold increases known as HTS.

The bands above 30 GHz will be used eventually in accordance with developing requirements and technology. Table 1.3 summarises this discussion.

Table 1.3 Frequency allocations

Radiocommunications service	Typical frequency bands for uplink/downlink (GHz)	Usual terminology
Fixed-satellite service (FSS)	6/4	C band
	8/7	X band
	14/12–11	Ku band
	30/20	Ka band
	50/40	V band
Mobile satellite service (MSS)	1.6/15	L band
	30/20	Ka band
Broadcasting satellite service (BSS)	2/2.2	S band
	12	Ku band
	2.6/2.5	S band

The MSS makes use of the following bands:

— VHF (very high frequency, 137–138 MHz downlink, and 148–150 MHz uplink) and UHF (ultra-high frequency, 400–401 MHz downlink, and 454–460 MHz uplink). These bands are for non-geostationary systems only.
— About 1.6 GHz for uplinks and 1.5 GHz for downlinks, mostly used by geostationary systems such as INMARSAT; and 1610–1626.5 MHz for the uplink of non-geostationary systems such as GLOBALSTAR.

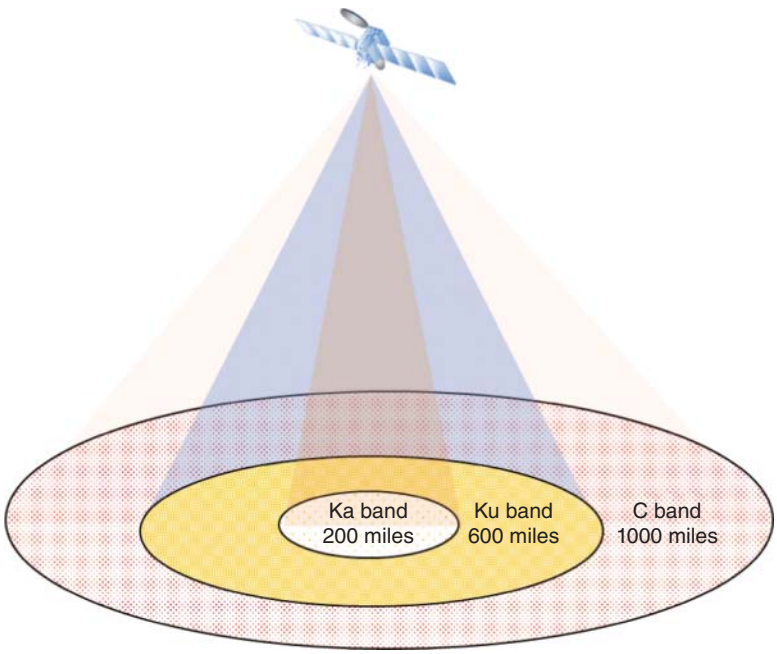


Figure 1.12 Relationship between coverage and frequency bands.

— About 2.2 GHz for downlinks and 2 GHz for uplinks for the satellite component of IMT2000 (International Mobile Telecommunications).
— About 2.6 GHz for uplinks and 2.5 GHz for downlinks.
— Frequency bands have also been allocated at higher frequencies such as Ka band.

The BSS makes use of downlinks at about 12 GHz. The uplink is operated in the FSS bands and is called a *feeder link*. Table 1.3 summarises the main frequency allocation and indicates the correspondence with some usual terminology.

Figure 1.12 shows the relationship between coverage and frequency bands. It can be seen that the higher the frequency band, the smaller the spot beam size.

1.6 TECHNOLOGY TRENDS

Figure 1.13 shows the developments since the start of the satellite communication era.

The start of commercial satellite telecommunications can be traced back to the commissioning of Intelsat I (Early Bird) in 1965. Until the beginning of the 1970s, the services provided were telephone and television (TV) signal transmission between continents. Satellites were designed to complement submarine cables and played essentially the role of telephone trunk connections. The goal of increased capacity has led rapidly to the institution of multibeam satellites and the reuse of frequencies first by orthogonal polarisation and subsequently by angular separation (see Chapter 5).

Communication techniques (see Chapter 4) have changed from analogue to digital. The second-generation DVB-S2, although backward compatible with DVB-S, has made use of the many novel technologies developed in recent years, including modulation techniques of 8 phase shift keying (8PSK) and 16 and 32 amplitude and phase shift keying (16-APSK and 32-APSK) in addition to quadrature phase shift keying (QPSK); efficient FEC with new low-density parity check (LDPC) codes; adaptive coding and modulations (ACMs); and performance close to the Shannon limit. This makes DVB-S2 30% more efficient than DVB-S. Furthermore, DVB-S2X (the extension of DVB-S2) has made efficient gains up to 51% compared to DVB-S2, with higher modulation schemes (including 64/128/256 APSK) and smaller roll-off factors (including 5%, 10%, and 15%).

DVB-RCS can provide up to a 20 Mbps forward link to the user terminal and a 5 Mbps return link from the user terminal, which is comparable to ADSL technology. DVB-RCS2 improved the performance of DVB-RCS by 30%. Multiple access to the satellite (see Chapter 6) was resolved by frequency division multiple access (FDMA). The increasing demand for a large number of low-capacity links, such as for national requirements or for communication with ships, led in 1980 to the introduction of demand assignment (see Chapter 6), first using FDMA with single channel per carrier/frequency modulation (SCPC/FM) or PSK and subsequently using time division multiple access/phase shift keying (TDMA/PSK) in order to profit from the flexibility of digital techniques (see Chapter 4).

Simultaneously, the progress of antenna technology (see Chapter 9) enabled the beams to conform to the coverage of the service area. In this way, the performance of the link was improved while reducing interference between systems.

Multibeam satellites emerged, with interconnection between beams achieved by transponder hopping or on-board switching using satellite-switched time division multiple access (SS-TDMA). Scanning or hopping beams have been implemented in connection with on-board processing on modern satellites.

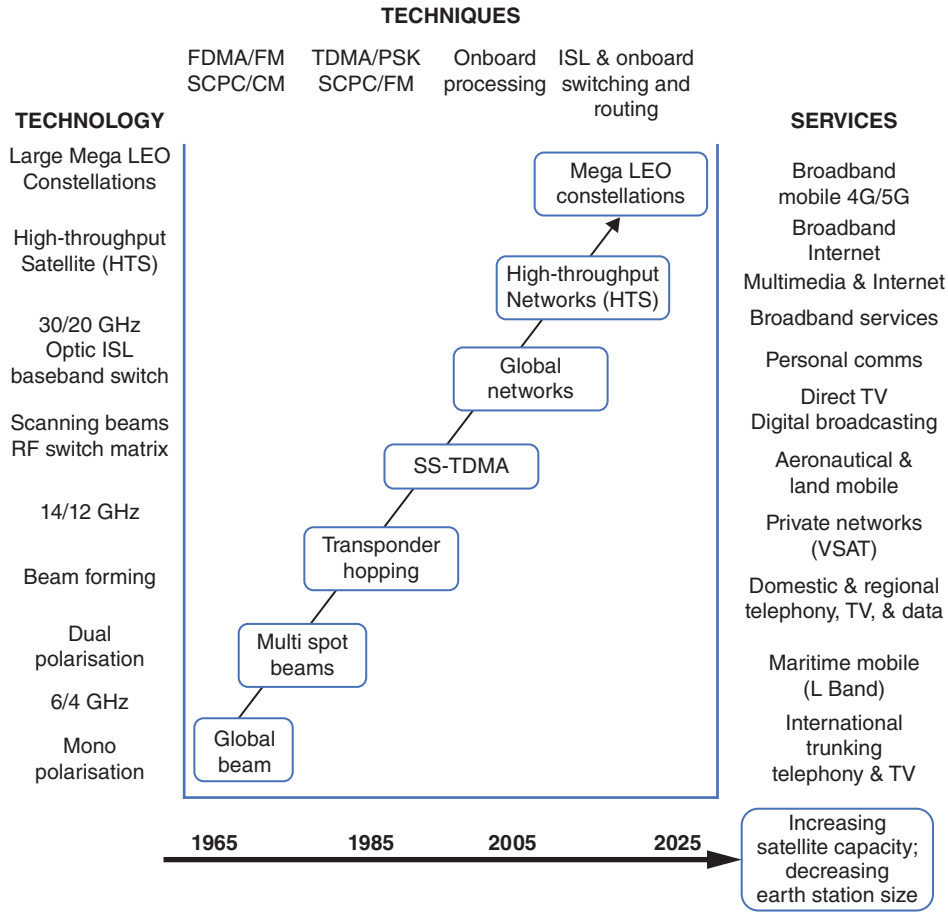


Figure 1.13 The evolution of satellite communication technologies.

Multiple-beam antennas today may produce hundreds of beams. This provides a twofold advantage: the link budget is improved to small user terminals, thanks to the high satellite antenna gain obtained with very narrow beams; and capacity is increased by reusing the frequency band allocated to the system many times.

Flexible interconnectivity between beams is required more than ever and may be achieved at different network layers by transparent or regenerative on-board processing. Regenerative payloads take advantage of the availability of baseband signals thanks to carrier demodulation. This is discussed in Chapters 7 and 9. Inter-satellite links were developed for civilian applications in the framework of multi-satellite constellations, such as IRIDIUM for mobile applications, and eventually will develop for geostationary satellites (Chapters 5 and 7). The use of higher frequencies (Ka band at 30/20 GHz) enables the emergence of broadband services and development of HTS, thanks to the large amount of bandwidth currently available, despite the propagation problems caused by rain effects (Chapter 5).

1.7 SERVICES

Initially designed as *trunks* that duplicate long-distance terrestrial links, satellite links have rapidly conquered specific markets. A satellite telecommunication system has three properties that are not found in terrestrial networks, or are found only to a lesser extent:

— The possibility of broadcasting with large coverage
— Wide bandwidth
— Rapid setup and ease of reconfiguration

With these properties, satellites can provide telecommunication services everywhere in the world. It is also possible to support mobile communication including services to airplanes, cruise ships, and high-speed trains.

It is possible to combine cellular networks and fibre for most users (including back-haul support for cellular networks such as 3G/4G and 5G mobile communications). Providing broadcasting services for large populations on a global scale has been one of the main advantages of satellite systems.

Further, it is possible to communicate in places with hostile terrain or a poorly developed terrestrial infrastructure; as well as in niche markets where obtaining the right of way for laying fibre is difficult or unduly expensive, such as remote rural areas, islands, and oil rigs. Finally, satellite service is very important when rapid deployment is critical, such as disaster relief and rescue services and government/military communication systems, as well as scientific explorations. Figure 1.14 illustrates typical satellite services and applications [SUN-14].

The preceding sections describe the state of technical development and show the development of the ground segment with respect to reduction in the size of stations and decreasing station cost. Initially, satellite systems contained a small number of earth stations: several stations per country, equipped with large 15–30 m diameter antennas collecting traffic from an extensive area by means of a ground network. Subsequently, the number of earth stations has increased, with

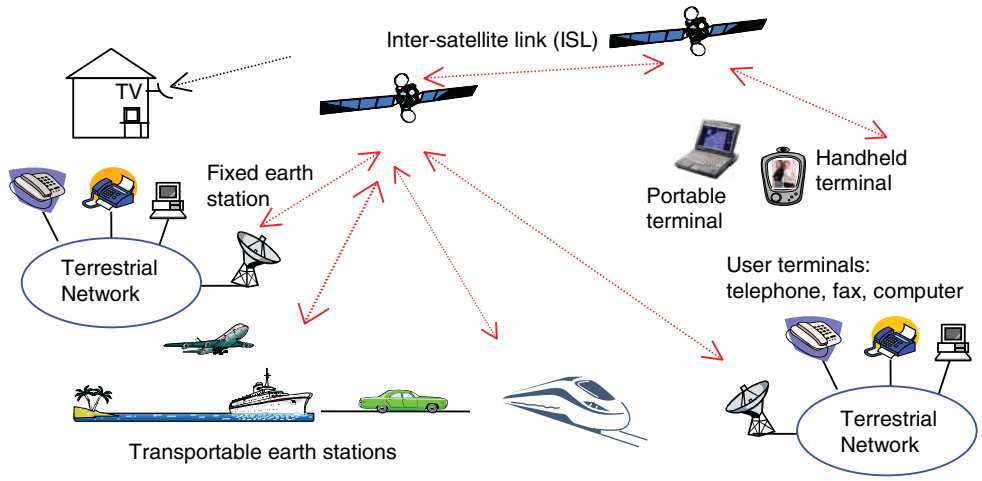


Figure 1.14 Illustration of typical satellite services and applications.

a reduction in size (1–4 m antennas) and greater geographical dispersion. The stations are closer to the user, and may be transportable or mobile. The potential of the services offered by satellite telecommunications has thus diversified:

— *Trunking telephony and television programme exchange*: This is a continuation of the original service. The traffic concerned is part of a country's international traffic. It is collected and distributed by the ground network on a scale appropriate to the particular country. Examples are Intelsat and Eutelsat (broadband connectivity for Internet and digital broadcasting services). The earth stations are equipped with 15–30 m diameter antennas.
— *Multiservice systems:* Telephone and data for user groups that are geographically dispersed. Each group shares an earth station and accesses it through a ground network whose extent is limited to one district of a town or an industrial area. Today these are mainly used for broadband Internet services. The earth stations are equipped with 3–10 m diameter antennas.
— *VSAT systems:* Low-capacity data transmission (uni- or bidirectional), television, or digital sound programme broadcasting [MAR-95]. Most often, the user is directly connected to the station. VSATs are equipped with antennas 0.6–1.2 m in diameter. The introduction of Ka band will allow even smaller antennas (ultra-small aperture terminal [USATs]) to provide even larger capacity for data transmission, allowing multimedia interactivity, data-intensive business applications, residential and commercial Internet connections, two-way videoconferencing, distance learning, and telemedicine.
— *Digital audio, video, and data broadcasting:* The emergence of standards for compression, such as the MPEG (Motion Picture Expert Group) standard for video, has triggered the implementation of digital services to small earth stations installed at the user's premises with antennas on the order of a few tens of centimetres. For *television*, such services using the DVB-S/S2/S2X standards have all become digital and can support HD TV, 4K, and even 8K TV. For *sound*, several systems incorporating on-board processing have been launched in such a way as to allow FDMA access by several broadcasters on the uplink and time-division multiplexing (TDM) on a single downlink carrier of the sound programmes. This approach avoids the delivery of programmes to a single feeder earth station, and allows operation of the satellite payload at full power; thus combining flexibility and efficient use of the satellite. The ability of the user terminal to process digital data paves the way for satellite distribution of files on demand through the *Internet*, with a terrestrial request channel or even a satellite-based channel such as DVB-RCS or DVB-RCS2. This anticipates broadband multimedia satellite services [TS-15].
— *Mobile and personal communications:* Despite the proliferation of cellular and terrestrial personal communication services around the world, there are still vast geographic areas not covered by any wireless terrestrial communications. These areas are open fields for mobile and personal satellite communications, and they are key markets for the operators of geostationary satellites, such as INMARSAT, and of non-geostationary satellite constellations, such as IRIDIUM and GLOBALSTAR. The next step in bridging the gaps between fixed, mobile, and broadcasting services concerns satellite multimedia broadcast to fixed and mobile users. Smart overlay broadcast networks based on hybrid satellite–terrestrial mobile systems will efficiently provide end users with a full range of entertainment services with interactivity [WER-07]. Studies of internetworking between satellite and terrestrial networks have been carried out by the ETSI [ETSI-13] and the 3rd Generation Partnership Project (3GPP) [3GPP-17].
— *Multimedia services:* These services aggregate different media, such as text, data, audio, graphics, fixed or slow-scan pictures, and video in a common digital format to offer potential for online services, teleworking, distance learning, interactive television, telemedicine, etc. Interactivity is therefore an embedded feature. This requires increased bandwidth compared to

conventional services, such as telephony, and has triggered the concept of an *information superhighway*. Satellites complement terrestrial, high-capacity fibre, and cable-based networks with the following characteristics: use of Ka band, multibeam antennas, wideband transponders (typically 125 MHz), on-board processing and switching, a large range of service rates (from tens of kbps to hundreds of Mbps), and quasi-error-free transmission (typically 10^{-11} BER).

1.8 THE WAY FORWARD

In the last 50 years, since the first commercial satellite, Early Bird, started its operation in 1965, the satellite telecommunications landscape has changed significantly. Advances in satellite technology have enabled satellite telecommunications providers to expand service offerings. The mix of satellite telecommunications is continuously evolving. Point-to-point trunking for analogue voice and television was the sole service initially provided by satellites; today, telecommunications satellites also provide digital audio and video broadcasting, mobile communications, on-demand narrowband data services, and broadband multimedia and Internet services. The mix of service offerings will continue to change significantly in the future.

Satellite services can be characterised as either satellite relay applications or end-user applications (fixed or mobile). For *satellite relay applications*, a content provider or carrier leases capacity from a satellite operator, or uses its own satellite system to transmit content to and from terrestrial ground stations where the content is routed to the end user. Relay applications accounted for around $10 billion in 2000. *End-user satellite applications* provide information directly to individual customers via consumer devices such as small antennas (less than earth station) and hand-held satellite user terminals. End-user applications accounted for about $25 billion in 2000.

It was reported by the Satellite Industry Association (SIA) on 11 June 2008 that the worldwide market in 2007 was $123 billion; average annual growth was 11.5% from 2002–2007 and jumped to 16% in 2007; satellite services grew 18% in 2007 to $37.9 billion, of which TV accounted for three quarters; launch was $3.2 billion, up 19%; ground equipment was $34.3 billion, up 19%; and satellite manufacture was $11.6 billion, dipping slightly (reflecting a larger number of microsatellites). In June 2017, SIA reported satellite industry indicators [SIA-17] that 2016 global revenue was $260.5B, of which satellite service was $127.7B; ground equipment 113.4B, satellite manufacture $13.9B and launch $5.5B.

The DVB-S2 standard was published in March 2005 and it was quickly adopted by industry. It is reported by the DVB forum that major broadcasters in Europe have started to use DVB-S2 in conjunction with MPEG-4 for high definition television (HDTV) services; examples include BSkyB in UK and Ireland, Premiere in Germany, and Sky in Italy. It has also been deployed in America, Asia, and Africa. DVB-S2X was completed as an extension to the DVB-S2 by ETSI in 2014.

There were also many initiatives for satellites in 2008 to deliver a range of multimedia services targeting fixed terminals at Ka band (Telesat Anik F2 multispot Ka band, Eutelsat Ka-Sat) [FEN-08]; broadband mobile terminals on board planes, trains, and ships at Ku or Ka band (satellite-on-the-move communications) [GIA-08]; and fixed and mobile users with hybrid terrestrial/satellite systems at S band [SUE-08, CHU-08]. Other initiatives include the provision of air traffic management services [WER-07]. According to the 2017 SIA report on satellite service revenue, satellite TV accounted for $97.8B; satellite radio $4.6B; satellite broadband $1.9B; fixed transponder agreement and managed services $12.4B and $5.5B, respectively; mobile $3.4B; and earth observation $1.8B.

Numerous new technologies are under development in response to the tremendous demand for emerging global telecommunications applications. Improved technology leads to the production of individual satellites that are more powerful and capable than earlier models. With

larger satellites (up to 10 000 kg) able to carry additional transponders and more powerful solar arrays and batteries, these designs will provide a higher power supply (up to 20 kW) to support a greater number of transponders (up to 150). New platform designs allowing additional capacity for station-keeping propellant and the adoption of new types of thrusters are contributing to increased service life of up to 20 years for geostationary satellites. This translates into increased capacity from satellites with more transponders, longer lives, and the ability to transmit more data through increasing rates of data compression.

In recent years, satellite services have grown by 5% per year, including revenue from Ku and Ka band satellite FSS capacity provided by MSS operators to provide services to maritime, airborne, and other mobile applications. FSSs has decreased by 3% per year due to decreased transponder agreement revenue, although revenue for managed services has grown 12% driven primarily by HTS capacity on the supply side and in-flight services on the demand side.

As an example, ViaSat-1 launched on 19 October 2011, it had the world's highest-capacity communications satellite with a total capacity of more than 140 Gbps – more than all the satellites covering North America combined, at the time of its launch. ViaSat-2 launched on 2 June 2017 with a capacity of 300 Gbps. The main technologies included multiple spot beams, spectrum reuse, high-gain spot beams, and a high-gain antenna. The main services included broadband access, data relay, mobile communications, and broadcasting, including 4K TV.

In addition to HTS, MEO, and LEO, satellite orbits have also had rapid development: typical examples are O3B, OneWeb, and Starlink, in addition to the Iridium Next. Table 1.4 shows some examples of next-generation LEO Mega constellations, and Table 1.5 shows some frequency and optical wavelength allocations for ISL.

Table 1.4 Examples of next-generation LEO mega constellation

	Iridium Next	LeoSat	OneWeb	Starlink	Hongyun project
Number of satellites	66	108	648 (+1972)	4425 +7518	156
Orbit altitude	781 km	1400 km	1200 km	1200 and 340 km	1000 km
Signal transmission frequency	L band Ka band	Ka band	Ku band (V band)	Ku band Ka band V band	Ka band
Capacity per satellite	N/A	11.6 Gbps	N/A	N/A	4 Gbps
Data speed	128 kbps 1.5 Mbps 8 Mbps	50 Mbps–1.6 Gbps 5.2 Gbps	50 Mbps	Gigabit per second	40 Mbps
Transmission latency	N/A	<20 ms	N/A	~25 ms	N/A
Year of operation	2015	2022	2019	2019	2024
Supporting enterprises	Iridium Inc.	LeoSat	Qualcomm, Virgin Group, Airbus, etc.	SpaceX	China Aerospace Science and Industry Corporation (CASIC)

Table 1.5 Frequency or optical wavelength allocations for intersatellite links (ISLs)

RF or laser	Frequency band or wavelength range	Available bandwidths or technologies
Microwave	22.55–23.55 GHz	1 000 MHz
	24.45–24.75 GHz (zones 1 and 3)	300 MHz
	25.25–27.50 GHz	2 250 MHz
mm wave	32–33 GHz	1 000 MHz
	54.25–58.20 GHz	3 950 MHz
	59–64 GHz	5 000 MHz
	65–71 GHz	6 000 MHz
	116–134 GHz	18 000 MHz
	116–134 GHz	18 000 MHz
	170–182 GHz	12 000 MHz
THz	0.3–30 THz	To be specified
Laser	10.6 μm	CO_2 lasers
	1.06 μm	Nd:YAG lasers
	0.532 μm	Nd:YAG lasers
	0.8–0.9 μm	Al GaAs lasers

REFERENCES

[3GPP-17] 3rd Generation Partnership Project. (2017). Study on new radio (NR) to support non-terrestrial networks. Technical report 38.811 (V0.2.1). 3GPP.

[AKT-08] Akturan, R. (2008). An overview of the Sirius satellite radio system. *International Journal of Satellite Communications* **26** (5): 349–358.

[SIA-17] Bryce Space and Technology. (2017). State of the satellite industry report. Satellite Industry Association (SIA).

[CHU-08] Chuberre, N. et al. (2008). Hybrid satellite and terrestrial infrastructure for mobile broadcast services delivery: an outlook to the 'Unlimited Mobile TV' system performance. *International Journal of Satellite Communications* **26** (5): 405–426.

[ETSI-13] ETSI. (2013) satellite earth stations and systems (SES); combined satellite and terrestrial network scenarios. TR 103 124 (V.1.1.1)

[TS-15] ETSI. 2015. Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); QoS functional architecture. TS 102 462 (V1.2.1).

[FEN-08] H. Fenech. (2008). The Ka-Sat satellite system. 14th Ka and Broadband Communications Conference, Matera, Italy, Sept 18-22.

[GIA-08] Giambene, G. and Kota, S. (2007). Special issue on satellite networks for mobile service. *Space Communications Journal* **21** (1): 2.

[ITU-12] ITU. (2012). Regulation of global broadband satellite communications. Broadband series.

[MAR-95] Maral, G. (1995). *VSAT Networks*. Wiley.

[SUE-08] Suenaga, M. (2008). Satellite digital multimedia mobile broadcasting (S-DMB) system. *International Journal of Satellite Communications* **26** (5): 381–390.

[SUN-14] Sun, Z. (2014). *Satellite Networking: Principles and Protocols*. Wiley.

[WER-07] Werner, M. and Scalise, S. (2008). Special issue on air traffic management by satellite. *Space Communications Journal* **21** (3): 4.

[ITU-16] ITU. (2016). Radio Regulations.

2 ORBITS AND RELATED ISSUES

This chapter examines various aspects of the satellite's motion around the earth; these include Keplerian orbits, orbit parameters, perturbations, eclipses, and the geometric relationships between satellites and earth stations. Such aspects will be used in Chapter 5 in relation to radio-frequency link performance and Chapters 7–12 dealing with the operation of earth stations and the launching and operation of the satellite.

2.1 KEPLERIAN ORBITS

These orbits are named after Johannes Kepler (a German mathematician, astronomer, and astrology, 27 December 1571–15 November 1630), who established, at the start of the seventeenth century, that the trajectories of planets around the sun were ellipses and not combinations of circular movements as had been thought since the time of Pythagoras (a Greek philosophy, around 570–4950 BC. Keplerian movement is the relative movement of two point bodies under the sole influence of their Newtonian attractions.

2.1.1 Kepler's laws

These laws arise from observation by Kepler of the movement of the planets around the sun:

(a) The planets move in a plane; the orbits described are ellipses with the sun at one focus (1602).
(b) The vector from the sun to the planet sweeps equal areas in equal times (the law of areas, 1605).
(c) The ratio of the square of the period T of revolution of a planet around the sun to the cube of the semi-major axis a of the ellipse is the same for all planets (1618).

2.1.2 Newton's law

Sir Isaac Newton (an English mathematician, astronomer, theologian, author, and physicist, 25 December 1642–20 March 1726) extended the work of Kepler and, in 1667, discovered the

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

universal law of gravitation. This law states that two bodies of mass m and M attract each other with a force that is proportional to their masses and inversely proportional to the square of the distance r between them:

$$F = GM\,m/r^2 \tag{2.1}$$

where G is a constant, called the *universal gravitation constant*, and $G = 6.672 \times 10^{-11}$ m^3 kg^{-1} s^{-2}.

As the mass of the earth $M = 5.974 \times 10^{24}$ kg, the product GM has a value $\mu = GM = 3.986 \times 10^{14}$ m^3 s^{-2}.

From the universal law of gravitation and using the work of Galileo Galilei (an Italian polymath, 15 December 1564–8 January 1642), a contemporary of Kepler, Newton proved Kepler's laws mathematically and identified the assumptions (the problem of two spherical and homogeneous bodies). He also modified these laws by introducing the concept of orbit perturbations to take actual movements into account.

2.1.3 Relative movement of two point bodies

The movement of satellites around the earth observes Kepler's laws to a first approximation. The proof results from Newton's law and the following assumptions:

— The mass m of the satellite is small with respect to the mass M of the earth, which is assumed to be spherical and homogeneous.
— Movement occurs in free space; the only bodies present are the satellite and the earth.

The actual movement must take into account the fact that the earth is neither spherical nor homogeneous, the attraction of the sun and moon, and other perturbing forces.

2.1.3.1 Keplerian potential

Kepler's laws can be explained by treating the relative movement of two bodies by applying Newton's law. It is convenient to consider the body of greater mass to be fixed, with the other moving around it (as the force of attraction is the same for the two bodies, the resulting acceleration is much greater for the body of low mass than for the higher mass).

Consider an orthogonal coordinate system as illustrated in Figure 2.1 whose origin is at the centre of the earth and whose z axis coincides with the line of the poles (assumed fixed in space). The satellite SL of mass m ($m \ll M$)is at a distance r from the centre of the earth O (**r** is the vector O–SL).

The force of gravitation **F** acting on the satellite can be written:

$$\mathbf{F} = -GMm\,\mathbf{r}/r^3(\mathrm{N}) \tag{2.2}$$

(**F** is a vector centred on SL along SL–O)

This force always applies to the centre of gravity of the two bodies and, in particular, to the centre of the earth O. It is a central force. It derives from a potential gradient U such that $U = GM/r = \mu/r$. The attraction force per unit mass is given by:

$$F/m = \mathrm{d}/\mathrm{d}r[\mu/\mathrm{r}] = \mathrm{grad}\ U\ \ (\mathrm{ms}^{-2}) \tag{2.3}$$

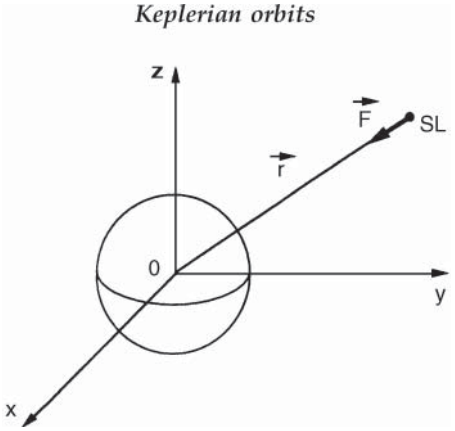


Figure 2.1 Geocentric coordinate system.

2.1.3.2 The angular momentum of the system

The angular momentum **H** of the system with respect to the point O can be written:

$$\mathbf{H} = \mathbf{r} \wedge m\mathbf{V} \ \text{(Nm s)} \tag{2.4}$$

where **V** is the velocity vector of the satellite. The momentum theorem states that the vector differential with respect to time of the instantaneous angular momentum is equal to the moment **M** of the external forces about the origin of the angular momentum:

$$d\mathbf{H}/\mathrm{d}t = \mathbf{M} \ \text{(Nm)} \tag{2.5}$$

In the system under consideration, the only external force **F** passes through the origin. The moment **M** is therefore zero; hence $\mathrm{d}\mathbf{H}/\mathrm{d}t$ is equal to zero. The result is that the angular momentum **H** is of constant magnitude, direction, and sign.

As the angular momentum is always perpendicular to **r** and **V**, movement of the satellite occurs in a plane that passes through the centre of the earth and has a fixed orientation in space perpendicular to the angular momentum vector.

In this plane, the satellite is identified by its polar coordinates **r** and θ (Figure 2.2). Hence:

$$\mathbf{H} = \mathbf{r} \wedge m\mathbf{V} = \mathbf{r} \wedge m(\mathbf{V}_\mathrm{R} + \mathbf{V}_\mathrm{T}) = \mathbf{r} \wedge m\mathbf{V}_\mathrm{R} + \mathbf{r} \wedge m\mathbf{V}_\mathrm{T}$$

As $\mathbf{V}_\mathrm{R}$ passes through O, the vector product $\mathbf{r} \wedge m\mathbf{V}_\mathrm{R} = 0$.

Hence $\mathbf{H} = \mathbf{r} \wedge m\mathbf{V}_\mathrm{T}$: that is, $|\mathbf{H}| = H = r \times mr\mathrm{d}\theta/\mathrm{d}t$. From which:

$$H = mr^2 \, \mathrm{d}\theta/\mathrm{d}t = C \ \text{(Nm s)} \tag{2.6}$$

where C is constant as the angular momentum is constant.

The expression $r^2 \, d\theta/\mathrm{d}t$ represents twice the area swept by the radius vector r during $\mathrm{d}t$. This area is thus constant, and Kepler's area law is verified.

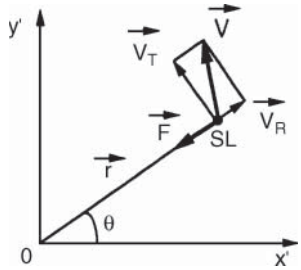


Figure 2.2 Location of the satellite SL in polar coordinates.

2.1.3.3 Equation of motion

The satellite describes a trajectory such that at every point it is in equilibrium between the inertial force $m\gamma$ and the force of attraction (Figure 2.2):

$$\mathbf{F} = m\gamma = -\mu m\, \mathbf{r}/r^3(\mathrm{N})$$

Equilibrium of the radial components of forces leads to:

$$\mathrm{d}^2r/\mathrm{d}t^2 - r(\mathrm{d}\theta/\mathrm{d}t)^2 = -\mu/r^2 \quad (\mathrm{ms}^{-2}) \tag{2.7}$$

where $\mathrm{d}^2r/\mathrm{d}t^2$ represents the variation of radial velocity and $r(\mathrm{d}\theta/\mathrm{d}t)^2$ is the centripetal acceleration.

Taking into account Eq. (2.6), $r^2\mathrm{d}\theta/\mathrm{d}t - H/m$ gives:

$$\mathrm{d}^2r/\mathrm{d}t^2 - H^2/m^2r^3 = -\mu/r^2 \quad (\mathrm{ms}^{-2}) \tag{2.8}$$

The equation of the orbit is obtained by eliminating time from this equation. The derivative of r with respect to time is put in the form:

$$\mathrm{d}r/\mathrm{d}t = (\mathrm{d}r/\mathrm{d}\theta)(\mathrm{d}\theta/\mathrm{d}t)$$

with $\mathrm{d}r/\mathrm{d}\theta = -(1/\rho^2)\mathrm{d}\rho/\mathrm{d}\theta$, by putting $\rho = 1/r$ and

$$\mathrm{d}\theta/\mathrm{d}t = H/mr^2 \tag{2.9}$$

Hence:

$$\mathrm{d}r/\mathrm{d}t = -(H/m)\mathrm{d}\rho/\mathrm{d}\theta \quad \text{and} \quad \mathrm{d}^2r/\mathrm{d}t^2 = -(H^2/m^2)\rho^2(\mathrm{d}^2\rho/\mathrm{d}\theta^2)$$

Equation (2.8) becomes:

$$\mathrm{d}^2\rho/\mathrm{d}\theta^2 + \rho = \mu m^2/H^2 \tag{2.10}$$

This equation is integrated to give $\rho = \rho_0\cos(\theta = \theta_0) + \mu m^2/H^2$ (ρ_0 and θ_0 are the constants of integration), which, replacing ρ with $1/r$, can be written:

$$r = (H^2/\mu m^2)/[1 + \rho_0\,(H^2/\mu m^2)\cos(\theta - \theta_0)]$$

Hence:

$$r = p/[1 + e\,\cos(\theta - \theta_0)](\mathrm{m}) \tag{2.11}$$

with $p = (H^2/\mu m^2)$ and $e = (\rho_0 \mathrm{H}^2/\mu m^2)$.

This is the equation in polar coordinates of a conic section with focus at the origin O, radius vector r, and argument θ with respect to an axis making an angle θ_0 relative to the axis of symmetry of the conic section.

2.1.3.4 Trajectories

In Eq. (2.11), letting the value $(\theta - \theta_0) = 0$, we get $r_0 = p/(1+e)$, from which $e = (p/r_0) - 1 = (H^2/\mu m^2 r_0) - 1$ with Eq. (2.6) $H = m r_0 V_0$, since the velocity V_0 is perpendicular to the minimum radius vector.

The quantity e can thus be written:

$$e = (r_0 V_0^2/\mu) - 1 \tag{2.12}$$

The type of conic section depends on the value of e:

For $e = 0$, $V_0 = \sqrt{(\mu/r_0)}$, and the trajectory is a circle.
For $e < 1$, $V_0 < \sqrt{(2\,\mu/r_0)}$, and the trajectory is an ellipse.
For $e = 1$, $V_0 = \sqrt{(2\,\mu/r_0)}$, and the trajectory is a parabola.
For $e > 1$, $V_0 > \sqrt{(2\,\mu/r_0)}$, and the trajectory is a hyperbola.

Only values of $e < 1$ correspond to a closed trajectory around the earth and are thus of use for communications satellites. Values of $e \geq 1$ correspond to trajectories that lead to the satellite freeing itself from terrestrial attraction (becoming space probes).

2.1.3.5 Energy of the satellite in the trajectory

The concept of the energy of the satellite in the trajectory is introduced by setting the variation of potential energy between the current point on the trajectory and the point chosen as the origin (such as the extremity of the minimum length radius vector) equal to the variation of kinetic energy between these two points:

$$(1/2)m(V^2 - V_0^2) = m\mu[(1/r) - (1/r_0)], \quad \text{(J)}$$

$$\text{Hence: } (V_0^2/2) - \mu/r_0 = (V^2/2) - \mu/r = E_0 \tag{2.13}$$

E_0 is a constant that is equal, for unit mass ($m = 1$), to the sum of the kinetic energy $V^2/2$ and the potential energy $-\mu/\mathrm{r}$ (equal to the potential μ/r with a change of sign). This sum is the total energy of the system.

2.1.4 Orbital parameters

The orbits of communications satellites are thus, in general, ellipses defined in the orbital plane by the Eq. (2.11) as:

$$r = p/[1 + e\ \cos(\theta - \theta_0)] \text{ with } e < 1$$

2.1.4.1 Shape parameters: semi-major axis and eccentricity

Referring to Figure 2.3, the radius vector r is maximum for $\theta - \theta_0 = \pi$ and corresponds to the apogee of the orbit:

$$r_a = p/(1 - e) \quad \text{(m)} \tag{2.14}$$

The radius vector r_p corresponding to the perigee of the orbit is that of minimum length $r_0 (r_p = r_0 = p/(1+e))$.

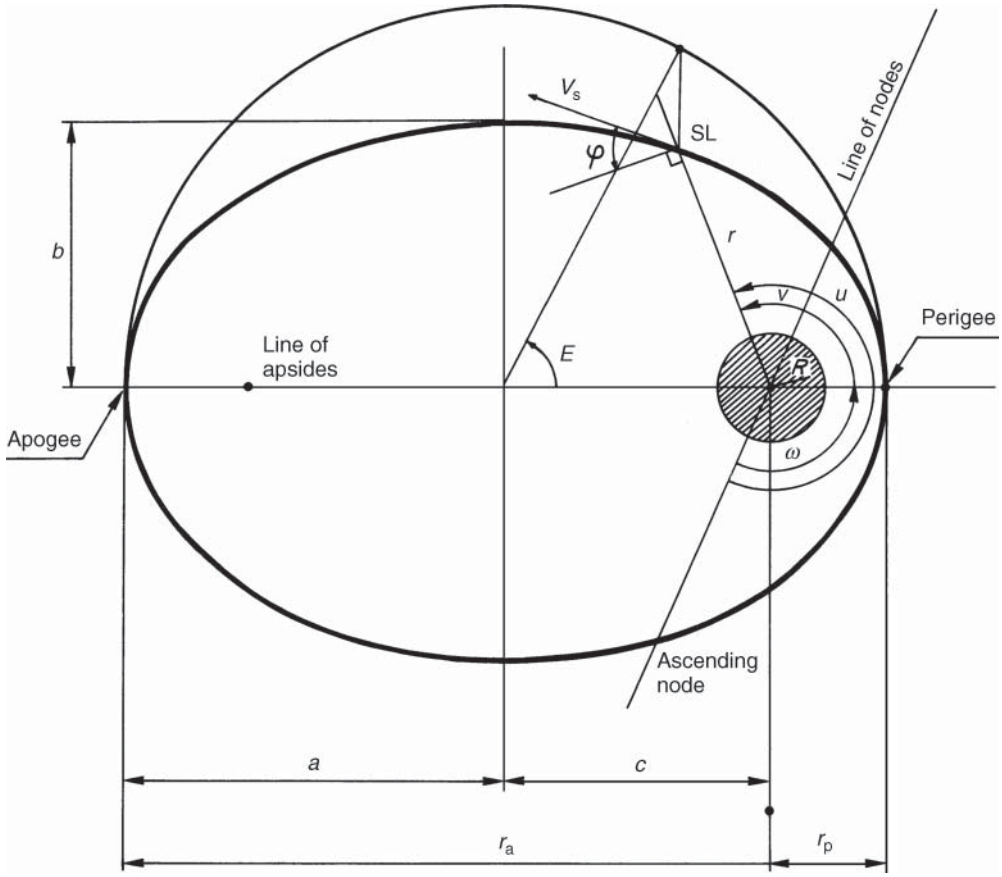


Figure 2.3 Parameters that define the form of the orbit: semi-major axis a, semi-minor axis $b = a\sqrt{(1-e^2)}$, $c = \sqrt{(a^2 - b^2)}$, eccentricity $e = c/a$, distances from the centre of the earth to the perigee, $r_p = a(1-e)$, and to the apogee, $r_a = a(1+e)$. $e = \sqrt{(1 - a^2/b^2)}$; $a^2/b^2 = (1-e^2)$; $e = (r_a - r_p)/(2a)$; $a = (r_p + r_a)/2$; $b = \sqrt{(r_p\, r_a)}$.

The sum $r_p + r_a$ represents the *major axis* of the ellipse of length $2a$; we get:

$$a = \frac{1}{2}(r_p + r_a) = p/(1 - e^2) \quad \text{(m)} \tag{2.15}$$

from which:

$$H^2/\mu m^2 = p = a(1 - e^2).$$

By putting $\theta - \theta_0$ equal to v, the equation of the ellipse becomes:

$$r = a(1 - e^2)/(1 + e\ \cos\ v) \quad \text{(m)} \tag{2.16}$$

The *eccentricity* (e) and *semi-major axis* (a) parameters appear, and these define the shape of the orbit.

The eccentricity e can be written:

$$e = (r_a - r_p)/(r_a + r_p) \tag{2.17a}$$

Also:

$$r_p = a(1-e) \quad \text{(m)} \tag{2.17b}$$

$$r_a = a(1+e) \quad \text{(m)} \tag{2..17c}$$

2.1.4.2 *Energy and velocity of the satellite*

From Eq. (2.13), the energy for unit mass $E_0 = (V_0^2/2) - \mu/r_0$ can be written as $E_0 = (H^2 - 2\mu r_0)/2r_0^2$ since $H = mr_0V_0$ ($= r_0V_0$ for unit mass, $m = 1$).

From Eq. (2.15), the semi-major axis can be written in the form:

$$a = \mu r_0^2/(2\mu r_0 - H^2) \quad \text{(m)}$$

and hence the *energy* E_0 has a value:

$$E_0 = -\mu/2a \tag{2.18}$$

Introducing the expression for E_0 into Eq. (2.13) gives $(V^2/2) - \mu/r = -\mu/2a$, which leads to an expression for the *velocity* V of the satellite:

$$V = \sqrt{\mu[(2/r) - (1/a)]} \quad (\text{ms}^{-1}) \tag{2.19a}$$

where $\mu = GM = 3.986 \times 10^{14}$ m^3 s^{-2} and r is the distance from the satellite to the centre of the earth. In the case of a circular orbit ($r = a$), the velocity is constant:

$$V = \sqrt{(\mu/a)} \quad (\text{ms}^{-1}) \tag{2.19b}$$

2.1.4.3 *Period of the orbit*

The duration of rotation of the satellite in the orbit, or *period* (T), is related to the area Σ of the ellipse by the law of areas with Eq. (2.6), which leads to $\Sigma = (H/m)(T/2)$. From Eq. (2.15), $H/m = \sqrt{[a\mu(1-e^2)]}$. The area of the ellipse is also given by $\pi ab = \pi a^2\sqrt{(1-e^2)}$. Therefore, $\sqrt{[a\mu(1-e^2)]}(T/2) = \pi a^2\sqrt{(1-e^2)}$.

Hence:

$$T = 2\pi\sqrt{(a^3/\mu)} \quad \text{(s)} \tag{2.20}$$

Some examples of values of the period T and velocity V for a circular orbit as a function of satellite altitude are given in Table 2.1 (the radius of the earth, R_E, is taken as 6378 km).

Table 2.1 Altitude, radius, period, and velocity for some circular orbits (earth radius $R_E = 6378$ km)

Altitude (km)	Radius (km)	Period (s)	Velocity (m s^{-1})
200	6 578	5 309	7 784
290	6 668	5 419	7 732
800	7 178	6 052	7 450
20 000	26 378	42 636	3 887
35 786	42 164	86 164	3 075

2.1.4.4 Position of the satellite in the orbit – anomalies

In the plane of the orbit, using the notation of Figure 2.3, the equation of the orbit in polar coordinates is given by Eq. (2.16):

$$r = a(1 - e^2)/(1 + e\ \cos\ v) \quad \text{(m)}$$

True anomaly (v). The position of the satellite is determined by the angle v, called the true anomaly, an angle counted positively in the direction of movement of the satellite from 0° to 360°, between the direction of the perigee and the direction of the satellite.

Eccentric anomaly (E). The position of the satellite can also be defined by the eccentric anomaly E, which is the argument of the image in the mapping that transforms the elliptical trajectory into its principal circle (see Figure 2.3).

From Figure 2.3, we have c = *ae* and c = *a cos E* + *r cos v* together (Eq. (2.16)); the true anomaly v is related to the eccentric anomaly E by:

$$\cos\ v = (\cos\ E - e)/(1 - e\ \cos\ E) \tag{2.21a}$$

and by:

$$\tan\ (v/2) = \sqrt{[(1 + e)/(1 - e)]}\ \tan\ (E/2) \tag{2.21b}$$

Conversely, the eccentric anomaly E is related to the true anomaly v by:

$$\tan\ (E/2) = \sqrt{[(1 - e)/(1 + e)]}\ \tan\ (v/2) \tag{2.21c}$$

and by:

$$\cos\ E = (\cos\ v + e)/(1 + e\ \cos\ v) \tag{2.21d}$$

Finally, the following relation avoids singularities in the calculations:

$$\tan[(v - E)/2] = (A\ \sin\ E)/(1 - A\ \cos\ E) = (A\ \sin\ v)/(1 + A\ \cos\ v) \tag{2.21e}$$

with:

$$A = e/[1 + \sqrt{(1 - e^2)}]$$

The distance r of the satellite from the centre of the earth can be written:

$$r = a(1 - e\ \cos\ E) \quad \text{(m)} \tag{2.22}$$

Mean movement (n). It is permissible to define the mean movement of the satellite n as the mean angular velocity of the satellite of period T in its orbit:

$$n = 2\pi/T \quad \text{(rad/s)} \tag{2.23}$$

Mean anomaly (M). The position of the satellite can thus be defined by the mean anomaly M that would be the true anomaly of a satellite in a circular orbit of the same period T. The mean anomaly is expressed as a function of time t by:

$$M = (2\pi/T)(t - t_p) = n(t - t_p) \quad \text{(rad)} \tag{2.24}$$

where t_p is the instant of passing through the perigee. The mean anomaly is related to the eccentric anomaly by Kepler's equation:

$$M = E - e\ \sin\ E \quad \text{(rad)} \tag{2.25}$$

2.1.4.5 Position of the orbital plane in space

The position of the orbital plane in space is specified by means of two parameters: the *inclination i* and the *right ascension of the ascending node* Ω. These parameters are defined, as shown in Figure 2.4, with respect to a coordinate system whose origin is the centre of mass of the earth, whose Oz axis is in the direction of terrestrial angular momentum (the axis of rotation normal to the equatorial plane), whose O*x* axis (normal to Oz) in the equatorial plane is oriented in the direction of a reference defined shortly, and whose O*y* axis in the equatorial plane is such that the coordinate system is regular.

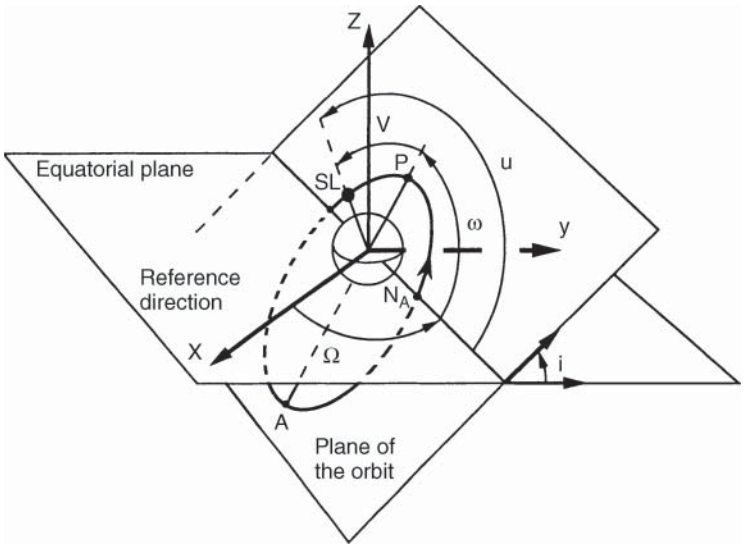


Figure 2.4 Positioning of the orbit in space: Ω = right ascension of the ascending node, ω = argument of the perigee, v = true anomaly, u = nodal angular elongation. The axis X identifies with the direction of the vernal point γ.

Inclination of the plane of the orbit (i). This is the angle at the ascending node (N_A), counted positively in the forward direction between 0° and 180°, between the normal (directed towards the east) to the line of nodes in the equatorial plane and the normal (in the direction of the velocity) to the line of nodes in the orbital plane. This is also the angle at the centre of the coordinate system between the angular momentum **H** of the orbit and the Oz axis (the direction of the pole). For an inclination less than 90°, the satellite rotates eastward in the same direction as the earth (this is called a *direct orbit*, *prograde*, or *non-retrograde orbit*). For an inclination greater than 90°, the satellite rotates westward in the opposite direction of the earth (this is called a *retrograde* orbit). For 90°, it is called *polar orbit*.

Right ascension of the ascending node (Ω). The right ascension of the ascending node (RAAN) is the angle taken positively from 0° to 360° in the forward direction, between the reference direction and that of the ascending node of the orbit (the intersection of the orbit with the plane of the equator, the satellite crossing this plane from south to north).

The reference direction (axis X in Figure 2.4) is given by the line of intersection of the equatorial plane and the plane of the ecliptic, oriented positively towards the sun (see Section 2.1.5.2 and Figure 2.5). With the Keplerian assumptions for the orbit of the earth around the sun, this line

(which is contained in the equatorial plane) maintains a fixed orientation in space with time and passes through the sun at the spring equinox, thereby defining the axis X as the direction of the vernal point γ.

In reality, the irregularities of terrestrial rotation (Section 2.1.5.3) cause the direction of intersection of the planes to vary somewhat; the coordinate system defined in this way is therefore not inertial and does not permit orbital motions to be integrated. Consequently, specific axes are defined: for example, the position of the coordinate system at a particular date. The date usually adopted is noon on 1 January 2000. On this date, the track of the line considered on the celestial sphere (a sphere of infinite radius centred on the earth) defines the point γ_{2000}.

The Veis coordinate system is also used; in it, the Oz axis is the axis from the centre of the earth to the north pole and the Ox axis is the projection at the date considered on the equatorial plane of the intersection of the equatorial plane and the plane of the ecliptic at 00.00 hours on 1 January 1950. This projection defines the pseudovernal point γ_{50} (so called since it is not always within the plane of the ecliptic). The Veis coordinate system has the advantage of permitting simple transformation, by a single rotation, to the terrestrial coordinate system in which the earth stations are located.

2.1.4.6 Location of the orbit in its plane

The orientation of the orbit in its plane is defined by the *argument of the perigee* ω. This is the angle, taken positively from 0° to 360° in the direction of motion of the satellite, between the direction of the ascending node and the direction of the perigee (Figure 2.4).

2.1.4.7 Conclusion

A knowledge of the five parameters (a, e, i, Ω, and ω) completely defines the trajectory of the satellite in space. The motion of the satellite in this trajectory can be defined by one of the anomalies (v, E, or M).

The *nodal angular elongation* u can also be used to define the position of the satellite in its orbit. This is the angle taken positively in the direction of motion from 0° to 360° between the direction of the ascending node and the direction of the satellite: $u = \omega + v$ (Figure 2.4). This parameter is useful in the case of a circular orbit where the perigee is unknown.

2.1.5 The earth's orbit

2.1.5.1 The earth

In the Keplerian hypotheses, the earth is assumed to be a spherical and homogeneous body. The real earth differs from this primarily by a flattening at the poles. The terrestrial surface is equivalent, on a first approximation, to that of an ellipsoid of revolution about the line of the poles whose parameters depend on the model chosen. The International Astronomical Union has, since 1976, recommended a value of 6378.144 km (mean equatorial radius R_E) for the semi-major axis and for the oblateness $A = (a - b)/a$, the value 1/298.257 (b is the semi-minor axis).

2.1.5.2 Motion of the earth about the sun

The earth rotates around the sun (Figure 2.5) with a period of approximately 365.25 days following an ellipse of eccentricity 0.01673 and semi-major axis 149 597 870 km, which defines

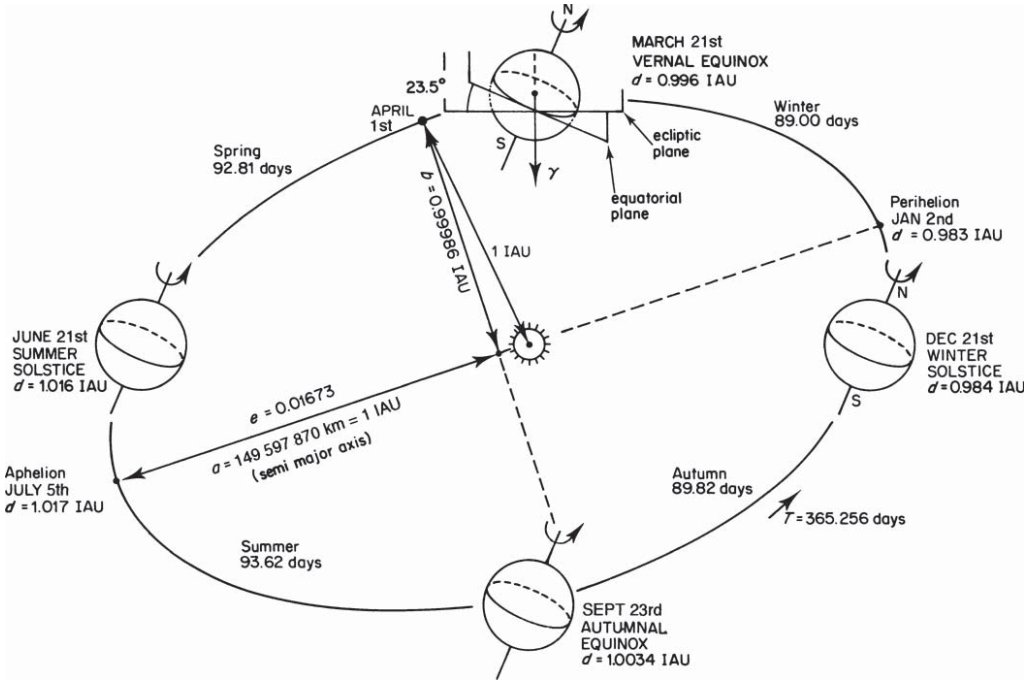


Figure 2.5 Orbit of the earth round the sun.

the astronomical unit of distance (AU). Around 2 January, the earth is nearest to the sun (the perihelion), while around 5 July it is at its aphelion (around 152 100 000 km).

The plane of the orbit is called the *plane of the ecliptic*. The plane of the ecliptic makes an angle of 23.44° (the obliquity of the ecliptic, which decreases around 47″ per century) with the mean equatorial plane.

The apparent movement of the sun around the earth with respect to the equatorial plane is represented by a variation of the declination of the sun (the angle between the direction of the sun and the equatorial plane, see Section 2.1.5.4). The declination varies during the year between +23.44° (at the summer solstice) and −23.44° (at the winter solstice). The declination is zero at the equinoxes. The direction of the sun at the spring equinox defines the vernal point or γ point on the celestial sphere (the geocentric sphere of infinite radius). The sun passes through it from the southern hemisphere to the northern hemisphere, and the declination is zero becoming positive.

The relation between the declination of the sun δ and the date is obtained by considering the apparent movement of the sun about the earth in an orbit of ellipticity e equal to 0.01673, inclined at the equator with obliquity ε. Hence (Figure 2.6):

$$\sin\delta = \sin\varepsilon \ \sin\ u \tag{2.26}$$

with $\sin\varepsilon = \sin 23.44° = 0.39795$ and u, the nodal elongation of the sun, equal to the sum of the true anomaly of the sun and the argument of the perigee ω_{SUN}. The argument of the perigee of the orbit representing the apparent movement of the sun about the earth remains more or less constant through the years if the precession of the equinoxes is neglected and has a value around 280°.

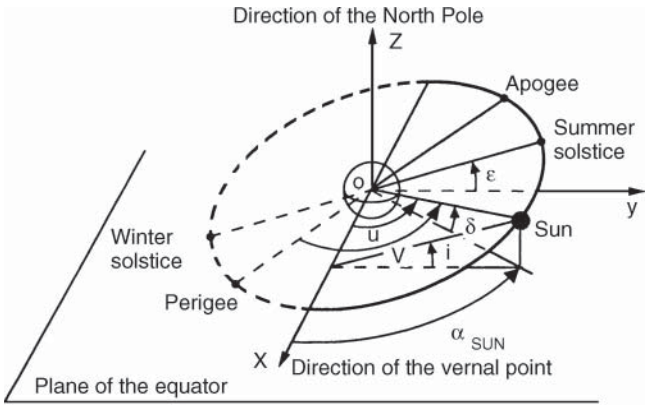


Figure 2.6 Apparent movement of the sun about the earth.

The true anomaly of the sun is expressed as a function of its eccentric anomaly E_{SUN} by means of Eqs. (2.21a) and (2.21b), and the eccentric anomaly as a function of the mean anomaly M_{SUN} by Kepler's Eq. (2.25). The mean anomaly is related to time by $M_{SUN} = n_{SUN}(t - t_0)$, with n_{SUN} the mean movement of the sun such that:

$$n_{SUN} = 2\pi/365.25 \text{ rad/day} = 360°/365.25 = 0.985\,626°/day$$

and t_0 the date of passing through the perihelion (about 2 January). Variation of declination with date is shown in Figure 2.7.

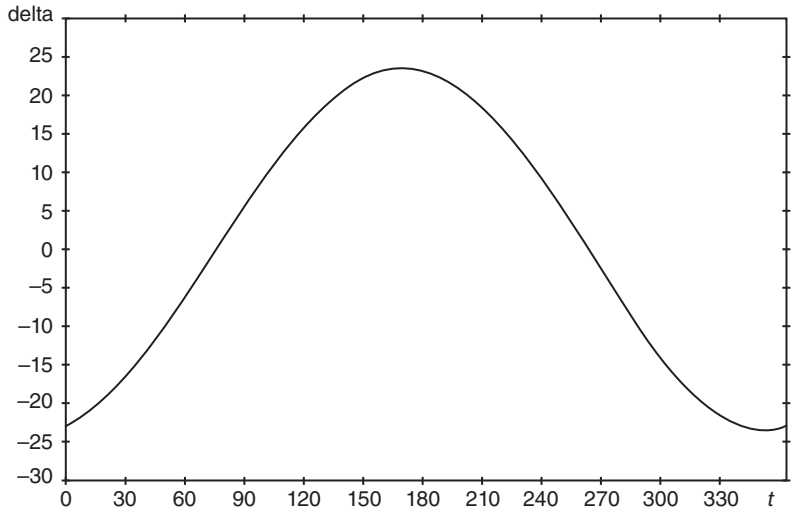


Figure 2.7 Variation of the declination of the sun in the course of a year (time t in days and declination delta in degrees, where $t = 0$ on 2 January).

2.1.5.3 Rotation of the earth

The axis of rotation of the earth cuts its surface at the poles. The pole moves slightly with time (within a circle of about 20 m diameter) with respect to the surface of the earth. The axis of rotation also moves in space. Its movement is a combination of periodic terms of limited amplitude (less than 20 seconds of arc), nutation, other non-periodic terms that have a cumulative effect, and precession. Precession relates to the angular momentum of the earth describing a cone, in 25 770 years, about the pole of the ecliptic (the axis normal to the plane of the ecliptic). These movements cause a westward motion of the vernal point γ along the ecliptic at the rate of approximately 50″ per year. Elimination of the periodic terms permits definition of the *mean equator*; this term is also applied with the same definition to various elements (coordinates, planes, times, etc.) that are affected by the irregularities of the rotation of the earth.

2.1.5.4 Terrestrial, equatorial, and temporal coordinates

Terrestrial coordinates. The terrestrial coordinates of a location are defined by:

— The *geographical longitude* λ: The angle, in the equatorial plane, between the origin meridian and the meridian of the location, taken positively towards the east from 0° to 360°, as recommended since 1982 by the International Astronomical Union (it should be noted that this convention is not universal).
— The *geographical latitude* φ: The angle between the vertical at the location and the plane of the equator, expressed in degrees from −90° (south pole) to +90° (north pole). The meridian of the location is the intersection of the half plane passing through the line of poles containing the location and the terrestrial surface. The origin meridian of longitude is the international meridian called Greenwich.

If the earth is considered to be spherical, the vertical (the perpendicular to the local horizontal plane) of any location passes through the centre of the earth. The flattening of the earth causes

the vertical of a location of latitude other than 0° or 90° to no longer pass exactly through the centre of the earth. Hence, the *geographical latitude* φ is different from the *geocentric latitude* φ' (the angle between the geocentric direction of the location and the equatorial plane). These two quantities are related by [PRI-93, p. 133]:

$$R_E^2 \tan \varphi' = b^2 \tan \varphi \tag{2.27a}$$

where b is the semi-major axis of the ellipsoid and R_E the mean equatorial radius. The distance R_C from the location to the centre of the earth is given by an approximation to the equation of the ellipse:

$$R_C = R_E(1 - A \sin^2 \varphi') \tag{2.27b}$$

where A is the oblateness of the earth (see Section 2.1.5.1).

Equatorial coordinates. The equatorial coordinates of a direction having its origin at the centre of the earth (the geocentric direction) are defined by:

— The *right ascension* α: The angle in the equatorial plane from the direction of the vernal point γ to the intersection of the meridian plane of the considered direction with the equatorial plane (this meridian plane contains the considered direction and the line of poles, and is perpendicular to the equatorial plane). This angle is taken positively in the direct direction (that of the rotation of the earth).
— The *declination* δ: The angle in the meridian plane of the considered direction between the equatorial plane and the considered direction. This angle is taken positively towards the north pole.

Hour (angle) coordinates. The local hour coordinates of a direction are defined by:

— The *hour angle* H: Taken positively in the equatorial plane in the reverse direction (westward) from the meridian plane of the observer's location to the meridian plane of the considered direction.
— The *declination* δ: Defined earlier.

H is most often measured in hours (1 h = 15°, 1 min = 15′, 1 s = 15″ and conversely 1° = 4 min, 1′ = 4 *s*, 1″ = 0.066 s). As the earth rotates in the direct direction (eastward), a fixed direction in space thus sees an increasing hour angle in time while its declination δ remains constant.

The coordinates just defined are geocentric coordinates. One can also define *topocentric coordinates*: that is, the coordinates of the direction of a point in space from a particular location on the surface of the earth. These coordinates are defined by using the plane parallel to the equator passing through the location as a reference plane. Topocentric coordinates differ from geocentric coordinates because of the *parallax*: the angle through which the terrestrial radius of the location is seen from the point in space considered.

Ecliptic coordinates are the *celestial longitude* and the *celestial latitude* of a direction; the reference plane is the ecliptic instead of the equatorial plane.

Finally, the *horizontal coordinates* of a direction, used by astronomers, are the *azimuth,* the angle taken in the horizontal plane of the location in the reverse direction from the south towards the projection of the direction; and the *height,* the angle between the direction and the horizontal plane. The *zenithal distance,* equal to the 90° complement of the height, is also used. To define the pointing direction of earth station antennas, it is customary to reckon the *azimuth* from the *north* and to call the height the *elevation*. Various expressions permit conversion from one set of coordinates to another; they can be found in most books on astronomy.

Sidereal time. The hour angle of the point γ is called the *local sidereal time* (LST). For a fixed direction, H-LST is constant and is such that (Figure 2.8a):

$$H = \text{LST-}\alpha \text{ (degrees or h min s)} \tag{2.28}$$

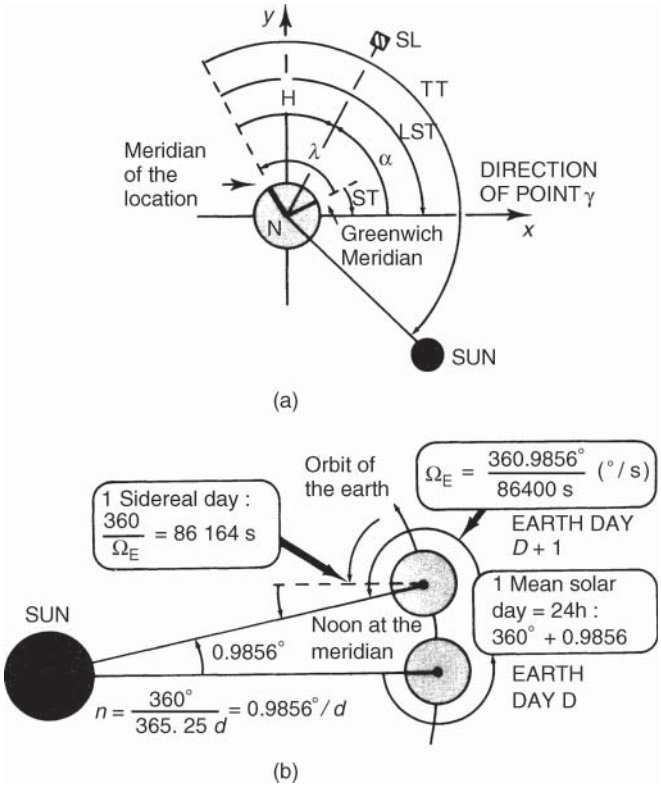


Figure 2.8 Space and time reference: (a) definition of angles;(b) sidereal day and mean solar day.

The *sidereal time* (ST) is the LST of the international meridian. If l is the geographical longitude of the location (positive towards the east):

$$\mathrm{ST} = \mathrm{LST} - \lambda \text{ (degrees or h min s)} \tag{2.29}$$

The sidereal time (of the Greenwich meridian) increases with time. Neglecting perturbations caused by variations of the fundamental planes (an error less than a hundredth of a degree) gives:

$$\mathrm{ST} = \mathrm{ST}_0 + \Omega_E t \tag{2.30}$$

where ST_0 is the sidereal time (of Greenwich) at 00.00 hour universal time (UT) on 1 January of each year (for example 100.776° for 2009, see Table 2.2). Ω_E is the velocity of rotation of the earth = 15.04169°/h = 4.178×10^{-3}°/s.

Solar time. True local solar time (TT) is the hour angle of the centre of the sun. True solar time has a value of 0 hour when the sun passes through the meridian of the location. Mean solar time (MT) is the solar time corrected for the periodic variations ΔE associated with the irregularities of the movement of the earth. Hence:

$$\mathrm{MT} = \mathrm{TT} - \Delta E \tag{2.31}$$

where ΔE is called the *time equation*. The time equation takes into account the relative movement of the sun in an elliptical orbit with respect to the earth and the effect of the obliquity of the

Table 2.2 Sidereal time on 1 January at 00.00 UT; calculated for the period from 2000–2020

Year	2000	2001	2002	2003	2004	2005	2006
ST0 (deg)	99.968	100.714	100.476	100.237	99.999	100.746	100.507
Year	2017	2008	2009	2010	2011	2012	2013
ST0 (deg)	100.268	100.029	100.776	100.538	100.299	100.060	100.807
Year	2014	2015	2016	2017	2018	2019	2020
ST0 (deg)	100.568	100.330	100.091	100.838	100.599	100.361	100.122

ecliptic. An approximate expression giving the time equation ΔE is:

$$\Delta E = 460 \sin n_{SUN} t - 592 \ \sin \ 2(\omega_{SUN} + n_{SUN} t)(\text{s}) \tag{2.32}$$

where t is the time in days from passing through the perihelion (around 2 January). The maximum value of ΔE is 4/15 h or 16 min ($n_{SUN} = 0.9856$ deg./day, $\omega_{SUN} = 280°$).

2.1.5.5 Time References

Sidereal day. The various solar and sidereal times are, in spite of their names, angles. Successive returns of a fixed star, or the point γ, to the meridian of a location define a timescale in *true sidereal days*. After elimination of the periodic terms, the *mean sidereal day* is obtained. This mean sidereal day defines the period T_E of rotation of the earth and has a value of 23 h 56 min 4.1 s or 86 164.1 s. The angular velocity of the earth is $\Omega_E = 360°/86\,164.1\ \text{s} = 4.17807 \times 10^{-3}°/\text{s} = 7.292 \times 10^{-5}\ \text{rad s}^{-1}$.

Solar day. Successive returns of the sun to the meridian of a location provide a timescale in *true solar days* and, by elimination of the periodic terms, in *mean solar days* of duration 24 hours or 86 400 seconds.

The sidereal day and the solar day differ because of the rotation of the earth around the sun, which has a mean value of 0.9856° per day (Figure 2.8b). A time interval measured in sidereal time must be multiplied by 86 164.1/86 400 or 0.997 269 6 to obtain a measurement in mean time. Conversely, a time interval measured in mean time must be multiplied by 86 400/86 164.1 or 1.002 737 9 to obtain a measurement in sidereal time.

Civil time, universal time, and sidereal time. Civil time is mean solar time increased by 12 hours (the civil day starts 12 hours later than the mean solar day). To define a time that is independent of location, the civil time at Greenwich, or *universal time* (incorrectly called Greenwich Mean Time [GMT]), is used.

Sidereal time (ST) can be determined from universal time (UT) using formulas that differ according to the selected time reference. The following formula is consistent with the time reference noted J_{2000} (noon on 1 January 2000; see Section 2.1.4.5):

$$\begin{aligned}\text{ST(seconds)} = \text{UT(seconds)} \times 1.0027379 + 24110.54841 \\ + 8640184.812866 \times T + 0.093140 \times T^2 - 6.2 \times 10^{-6} \times T^3\end{aligned} \tag{2.33}$$

where T = D/36525 is the number of Julian centuries (1 Julian century = 36 525 days) elapsed since the reference time J_{2000} until 12.00 hours UT of the considered date.

D is the number of mean solar days elapsed since 1 January 2000 at 00.00 hour UT to the considered date ST expressed in degrees.

The Julian calendar starts at noon on 1 January 4713 BC and constitutes a system of time representation where a century is equal to 36 525 days. The Julian day that starts at noon on 1 January 2000 is numbered $JD_0 = 2451\,545$.

Taking into account that the solar day starts at 00.00 hour UT (at Greenwich), a time shift of half a day is to be introduced with respect to the Julian calendar. The value of D, at the date of interest, is given by:

$$\begin{aligned} D = &(\text{day number in year} - 1.5) \\ &+ 365\ (\text{considered year} - 2000) \\ &+ \text{number of leap years fully completed since 2000, inclusive.} \end{aligned}$$

The following rules decide which years are leap years:

1. Every year divisible by 4 is a leap year.
2. But every year divisible by 100 is not a leap year.
3. Unless the year is also divisible by 400, in which case it is still a leap year.

This means 2004 is a leap year (rule 1), and 1900 is not a leap year (rule 2), while year 2000 is a leap year (rule 3). Here are some examples of leap years: 2000, 2004, 2008, 2012, 2016, 2020, 2024, etc.

Example 2.1 Equation (2.33) is used to calculate the value of ST_0 used in Eq. (2.30). For instance, on 1 January 2018 (day number = 1), there have been five leap years since 2000, and we can calculate using the formula as follows:

$$D = (1 - 1.5) + 365 \times (2018 - 2000) + 5 = 6574.5$$

2000 and 2004 are leap years; so are 2008, 2012, and 2016, but 2018 is not fully completed at the considered date, so the number of fully completed leap years is only 5.

$$\text{Julian day JD} = JD_0 + D = 2451545 + 6574.5 = 2458119.5$$

Then, $T = D/36522 = 6574.5/36525 = 0.18000000000$ Julian century.
From which:

$$\begin{aligned} ST_0\ (\text{seconds}) &= 0.0 + 24\,110.54841 + 8\,640\,184.812866\ T + 0.093104\ T^2 - 6.2 \times 10^{-6}\ T^3 \\ &= 1\,579\,344\ \text{s} \\ &= 6.706616\ \text{hours (modulo 24 hour)} \\ &= 6.706616 \times 15\ \text{degree/hour} \\ &= 100.599\ \text{degree} \end{aligned}$$

The results for 2000–2020 are indicated in Table 2.2.

Legal time and official time. Most countries, in accordance with their region of longitude, use a time, *legal time*, which is derived from universal time by correction of an integer number of hours. In some cases, the correction is a multiple of half hours, or some other specific correction

applies. Finally, economic considerations lead to a correction of legal time according to the season (summer time), which gives the *official time*.

2.1.6 Earth–satellite geometry

2.1.6.1 The satellite track

The satellite track on the surface of the earth is the locus of the point of intersection of the earth centre–satellite vector with the surface of the earth. The track takes into account the movement of the surface of the earth with respect to the actual displacement (as a function of the true anomaly) of the earth centre–satellite vector. The equation of the track of an orbit of fixed inclination and ellipticity is obtained from the following procedure.

From Figure 2.9a, the coordinates (λ_{SL},φ) of the satellite SL in an earth-centred, fixed reference frame are related by the following equation (the longitude is taken with respect to a reference meridian):

$$\tan\,\varphi = \tan\,i\,\sin\,(\lambda_{SL} - \lambda_N) \tag{2.34}$$

where φ is the latitude of the satellite, λ_{SL} is the longitude with respect to a reference meridian (fixed earth), λ_N is the longitude of the ascending node with respect to the reference meridian (fixed earth), and i is the inclination of the orbit.

The arc N–SL is the track of the satellite (on the fixed earth). The arc N–SL subtends an angle u (the nodal angular elongation) such that:

$$\sin\varphi = \sin\,i\sin u \tag{2.35a}$$

$$\tan(\lambda_{SL} - \lambda_N) = \tan u\cos i \tag{2.35b}$$

$$\cos u = \cos\varphi\cos(\lambda_{SL} - \lambda_N) \tag{2.35c}$$

The point on the track of highest latitude is called the *vertex*, and the longitude λ_V of this differs from that of the node by $\pi/2$: $\lambda_V = \lambda_N + 90°$.

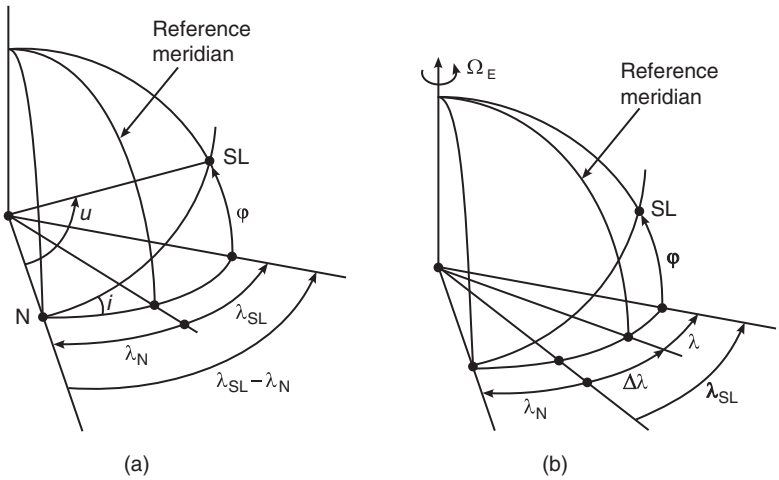


Figure 2.9 Definition of the track of a satellite: (a) fixed earth; (b) rotating earth.

It is necessary to take into account the rotation of the reference meridian during the satellite motion. Let Δt be the time elapsed since the passage of the satellite through the reference meridian (with a fixed earth) until the present time:

$$\Delta t = t_S - t_0$$

where t_S is the time elapsed since the passage of the satellite through the perigee and t_0 is the time of passage through the origin position of the reference meridian, measured from the time of passage through the perigee. As a consequence of the rotation of the earth, the displacement $\Delta\lambda$ of the reference meridian towards the east during time At has a value $\Delta\lambda = \Omega_E \Delta t$, where Ω_E is the angular velocity of the earth (see Figure 2.9b):

$$\Delta\lambda = \Omega_E \Delta t = \Omega_E (t_S - t_0) \tag{2.36a}$$

From Eq. (2.24), $t_S = M_s/n + t_P$; $t_0 = M_0/n + t_P$, where n is the mean movement of the satellite; and M_S and M_0 are, respectively, the mean anomaly of the satellite at time t_S and t_0. Hence:

$$\Delta\lambda = \Omega_E \Delta t = \Omega_E (t_S - t_0) = \Omega_E (M_S - M_0)/n \tag{2.36b}$$

The relative longitude λ of the satellite with respect to the rotating reference meridian can thus be written:

$$\lambda = \lambda_{SL} - \Delta\lambda \tag{2.37}$$

Longitude of the track relative to the meridian of the ascending node. If the origin position of the reference meridian is the ascending node, the longitude λ_N of the node is zero (i.e. $\lambda_N = 0$). Equations (2.34) and (2.35b) become:

$$\tan\varphi = \tan i \sin\lambda_{SL}$$

$$\tan\lambda_{SL} = \tan u \cos i \quad \text{or} \quad \cos\lambda_{SL} = \frac{\cos u}{\cos\varphi}$$

Furthermore, M_0 is equal to the mean anomaly M_N of the ascending node. The relative longitude with respect to the satellite meridian on passing through the ascending node can thus be expressed:

$$\lambda = \lambda_{SL} - \Delta\lambda = \text{arc sin}[(\tan\varphi)/(\tan i)] - M(\Omega_E/n) + M_N(\Omega_E/n) \tag{2.38a}$$

or:

$$\lambda = \lambda_{SL} - \Delta\lambda = \text{arc tan}[(\tan u)(\cos i)] - M(\Omega_E/n) + M_N(\Omega_E/n) \tag{2.38b}$$

or:

$$\lambda = \lambda_{SL} - \Delta\lambda = \text{arc cos}[(\cos u)/(\cos\varphi)] - M(\Omega_E/n) + M_N(\Omega_E/n) \tag{2.38c}$$

M and M_N, the mean anomalies of the satellite and the ascending node at considered time t, are calculated from the eccentric anomalies by using Kepler's Eq. (2.25).

The latitude φ of the satellite can be eliminated from Eq. (2.38a) by using Eq. (2.34). Also, $u = \omega + v$ (modulo 2π). This gives:

$$\lambda = \text{arc sin}\{\sin(\omega + v)\cos i[1 - \sin^2 i \sin^2(\omega + v)]^{-1/2}\}$$
$$- [(\Omega_E/n)(E - e\ \sin E) - (\Omega_E/n)(E_N - e\ \sin E_N)] \tag{2.39a}$$

or:

$$\lambda = \text{arc tan}[\tan(\omega + v)(\cos i)] - [(\Omega_E/n)(E - e\sin E) - (\Omega_E/n)(E_E - e\sin E_N)] \tag{2.39b}$$

where E is the eccentric anomaly of the satellite at time t and E_N is that on passing through the ascending node.

Latitude of the satellite. The latitude φ is not modified by the rotation of the earth and therefore does not depend on the choice of the origin position of the reference meridian. It can be written from (2.35a):

$$\varphi = \text{arc sin}[\sin i \sin(\omega + v)] \tag{2.40}$$

λ and φ can be expressed as a function of only one of the parameters E or v by using one of Eqs. (2.21a)–(2.21e).

For certain orbits, the mean movement n of the satellite is chosen to be equal to a multiple of the rank m of the angular velocity Ω_E of the earth. In the absence of perturbations, the track is thus unique (that is, the satellite passes through the same points again after m revolutions) and fixed with respect to the earth. In practice, it is necessary to take into account the precession of the plane of the orbit (the drift of the RAAN under the effect of the asymmetry of the terrestrial potential; see Section 2.2.1).

2.1.6.2 Satellite distance

Distance of the satellite from a point on the earth. The coordinates of the satellite in Figure 2.10 are φ for the latitude (the centre angle is TOA, T being the sub-satellite point) and λ for the longitude with respect to a reference meridian. Those of the point P considered are l for the latitude (centre angle POB) with ψ for the longitude with respect to the same reference meridian. For clarity of the figure, only the difference in longitude $L = \psi - \lambda$ from that of point P to that of the satellite (centre angle AOB) is represented. The centre angle BOT has a value ζ, and the centre angle POT (in the plane of the earth centre, the satellite SL, and point P) has a value φ. Let R be the distance from the satellite to point P, r the distance from the satellite to the earth centre, and R_E the radius of the earth.

Consider the triangle OPS (S is used instead of SL for simplicity). This gives $R^2 = R_E^2 + r^2 - 2R_E r \cos\phi$, and hence:

$$R = \sqrt{(R_E^2 + r^2 - 2R_E r \cos\phi)} \tag{2.41}$$

It remains to evaluate $\cos\phi$.

In the spherical triangle TPB, the cosine law gives:

$$\cos\phi = \cos\zeta \cos l + \sin\zeta \sin l \cos\text{PBT}$$

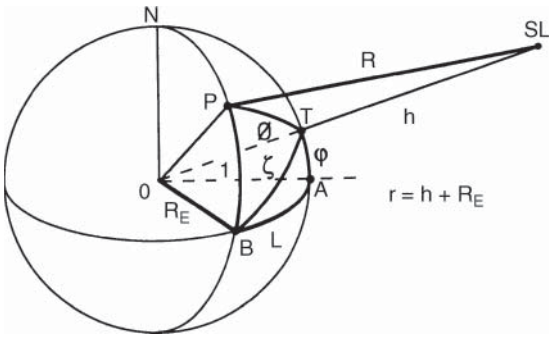


Figure 2.10 Earth–satellite geometry.

The sine rule in the spherical triangle TAB gives:

$$(\sin \text{TAB})/(\sin \zeta) = (\sin \text{TAB})/(\sin \varphi), \text{with TAB} = \pi/2 \text{ and TAB} = (\pi/2) - \text{PBT}$$

Hence, $\sin\zeta \cos \text{PBT} = \sin \varphi$.

Furthermore, in the triangle TAB, $\cos\zeta = \cos L \cos \varphi$, from which:

$$\cos \phi = \cos L \cos \varphi \cos l + \sin \varphi \sin l \tag{2.42}$$

The proposed equations assume that the earth is spherical and of radius R_E (the mean equatorial radius). For a more precise calculation, it would be convenient to define the actual radius from the reference ellipsoid (see Section 2.1.5.1) and use the geocentric latitude φ' of the point that is obtained from the geographic latitude φ by Eq. (2.27a).

Satellite altitude. The altitude h of the satellite corresponds to its distance from the sub-satellite point (the distance SL–T in Figure 2.10). This gives:

$$h = r - R_E \tag{2.43}$$

2.1.6.3 Satellite location – elevation and azimuth

Two angles are necessary to locate the satellite from the point P on the surface of the earth. It is customary to use the elevation and azimuth angles.

Elevation angle. The elevation angle is the angle between the horizon at the point considered and the satellite, measured in the plane containing the point considered, the satellite, and the centre of the earth. This is the angle E in Figure 2.11, which represents the triangle OPS from Figure 2.10. It follows, by considering the right angle OP'S (formed by extending the segment OP) and noting that the angle PSP' is equal to E:

$$\cos E = (r/R) \sin \phi \quad \text{hence } E = \arccos[(r/R) \sin \phi] \tag{2.44a}$$

The distance r is that of the satellite from the centre of the earth, R is that of the satellite from the point P that is calculated using Eq. (2.41), and $\sin \phi$ is obtained from Eq. (2.42) using $\sin \phi = \sqrt{(1 - \cos^2 \phi)}$.

The radius of the earth R_E can be introduced with:

$$\tan E = [\cos \phi - (R_E/r)]/\sin \phi \tag{2.44b}$$

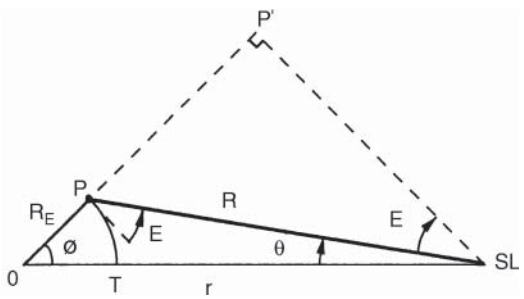


Figure 2.11 Elevation and nadir angles.

Another form is possible:

$$\sin E = [\cos\ \phi - (R_E/r)]/(R/r) \tag{2.44c}$$

with, from Eq. (2.41), $(R/r) = \sqrt{[1 + (R_E/r)^2 - 2(R_E/r)\cos\phi]}$.

Azimuth angle. The azimuth angle A is the angle measured in the horizontal plane of the location between the direction of geographic north and the intersection of the plane containing the satellite and the centre of the earth (the plane OPS). This angle varies between 0° and 360° as a function of the relative positions of the satellite and the point considered. It is the angle NPT in the spherical triangle of the same name in Figure 2.10. This gives:

$$(\sin\ \mathrm{NPT})/[\sin(90 - \varphi)] = (\sin \mathrm{PNT})/(\sin\phi)$$

In the triangle NBA, (sin BNA)/(sin L) = (sin BAN)/(sin AON) = 1.

From which, since the angle BNA is equal to the angle PNT (i.e. BNA = PNT):

$$\sin \mathrm{NPT} = (\sin L \cos\varphi)/\sin\phi$$

The calculation gives an angle less than $\pi/2$, although in the case of the figure, the azimuth is greater than $\pi/2$. This arises from the symmetry properties of the sine function. The result of the calculation is taken as an intermediate parameter, called a, in the determination of the azimuth ($a < \pi/2$). Thus:

$$\sin a = (\sin L \cos\varphi)/\sin\varphi \tag{2.45}$$

and hence:

$$a = \mathrm{arc}\ \sin[(\sin L \cos\varphi)/\sin\phi]\ \ (\text{with: } \phi > 0, L > 0)$$

The true azimuth A is obtained from a in accordance with the position of the sub-satellite point T with respect to the point P. The various cases are summarised in Table 2.3.

Table 2.3 Determination of the azimuth A

Position of the sub-satellite point T with respect to point P	Relation between A and a
South-east	$A = 180° - a$
North-east	$A = a$
South-west	$A = 180° + a$
North-west	$A = 360° - a$

2.1.6.4 Nadir angle

In Figure 2.11, the angle at the satellite SL between the direction of the centre of the earth O and the direction of the point P is called the *nadir angle* θ. In the triangle OPS:

$$\sin\theta = (\sin\phi)R_E/R \quad \text{hence: } \theta = \mathrm{arc}\ \sin[R_E(\sin\phi)/\mathrm{R}] \tag{2.46a}$$

or, if the elevation angle E is taken into consideration:

$$\phi + \theta + E = \pi/2 \tag{2.46b}$$

which, with $\sin\theta = [\sin(\pi - \theta - \phi)]/r$, leads to:

$$\sin\theta = (\cos E)R_E/r \quad \text{hence: } \theta = \mathrm{arc}\ \sin[R_E(\cos E)/r] \tag{2.46c}$$

2.1.6.5 Coverage at a given elevation angle

The *coverage zone* can be specified as the region of the earth where the satellite is seen with a minimum elevation angle E. The contour of the coverage zone is defined by a set of ground locations determined by their geographical coordinates and hence known by their relative longitude L and latitude l values. The relationship between L and l is as follows from (2.42):

$$L = \text{arc cos}[(\cos\phi - \sin\varphi \sin l)/\cos\varphi \cos l] \tag{2.47}$$

where $\phi = \pi/2 - E - \text{arc sin}[R_E(\cos E)/r]$ from Eqs. (2.46a) and (2.46c).

The longitudinal extent of the coverage zone with respect to the sub-satellite point T is obtained by putting $l = \varphi$ in the previous expression. The latitudinal extent is equal to ϕ.

2.1.6.6 Propagation time – the Doppler effect

Propagation time. The trajectory of radio waves on a link between an earth station and a satellite at distance R requires a propagation time τ equal to:

$$\tau = R/c \quad \text{(s)} \tag{2.48}$$

where c is the velocity of light (3×10^8 m s^{-1}).

Variation of relative distance – the Doppler effect. When the satellite moves with respect to the earth, the relative distance R from the satellite to a point on the surface of the earth varies. A range variation rate $\mathrm{d}R/\mathrm{d}t = V_r$ can be defined in accordance with the point considered ($V_r = V\cos\zeta$, where ζ is the angle between the direction of the point considered and the velocity V of the satellite).

This positive or negative range variation rate causes, at the receiver, an apparent increase or decrease, respectively, in the frequency of the radio wave transmitted on the link (the Doppler effect). The phenomenon occurs, of course, on both the uplink and the downlink. The shift Δf_d in the frequency f of the wave on the link can be written:

$$\Delta f_d = V_r f/c = V \cos\zeta (f/c) \quad \text{(Hz)} \tag{2.49}$$

where:

$c = 3 \times 10^8$ = velocity of light (m/s)

f = frequency of the transmitted wave (Hz)

V_r = range variation rate (m/s)

The geometry of the system changes with the movement of the satellite with respect to the point considered; the apparent velocity of the satellite varies with time and thus involves a variation of the Doppler shift.

An important parameter that will affect the performance of the automatic frequency control of the receiver system is the rate of variation of frequency $\mathrm{d}(\Delta f_d)/\mathrm{d}t$:

$$\mathrm{d}(\Delta f_d)/\mathrm{d}t = \mathrm{d}/\mathrm{d}t(V_r)f/c \quad \text{(Hz/s)} \tag{2.50}$$

Detailed calculations are given in [VIL-91]. On an equatorial circular orbit, the maximum value of the Doppler shift (when the satellite appears or disappears at the horizon) can be estimated by (CCIR-Rep 214) and ITU-R S.730 [ITUR-92a]:

$$\Delta f_d \cong \pm 1.54 \times 10^{-6} f\, m(\text{Hz})$$

where m is the number of revolutions per day of the satellite with respect to a fixed point on the earth (the period T of the orbit is equal to $24/(m+1)$ hours) and f is the frequency. For $m = 0$, the period is 24 hours, the satellite remains fixed with respect to the earth (a geostationary satellite; see Section 2.2.5), and the Doppler shift is theoretically zero. For $m = 3$, the period T has a value of 6 hours (for an altitude around 11 000 km), and the Doppler shift is on the order of 18 kHz at 6 GHz.

For an elliptical orbit with high eccentricity ($e > 0.6$), when the altitude of the satellite is large with respect to the radius of the earth, i.e. near apogee, the variation of distance R is nearly equal to the variation of radial distance r. The velocity V_r can be written:

$$dr/dt = (dr/d\theta)(d\theta/dt),$$

with $d\theta/dt = H/(mr^2)$ from Eq. (2.9) and $dr/d\theta$, which is determined from Eq. (2.10). This gives (with $v = \theta$):

$$V_r = dr/dt = e\sqrt{\mu}/\sqrt{[a(1-e^2)]}\sin v \quad \text{(m/s)} \tag{2.51}$$

where e is the eccentricity, μ is GM, a is the semi-major axis, and v is the true anomaly of the satellite. The radial velocity is maximum for $v = 90°$.

Apart from the problems posed by tracking variations of incident signal frequency at the receiver, variations of relative distance lead to problems of synchronisation between the signals originating from different earth stations (see Section 6.5.4). Distance variations also cause variations of propagation time on the link (CCIR-Rep 383) [ITUR-92b] [ITUT-03].

2.1.7 Eclipses of the sun

An eclipse of the sun occurs for a satellite when it passes into the conical shadow region of the earth or moon. The occurrence and duration of these eclipses depend on the characteristics of the satellite orbit. The consequences of the eclipse on the satellite are of two types. On the one hand, the electrical power supply system of the satellite, which includes photovoltaic cells to convert solar energy into electrical energy, must make use of an alternative energy source. On the other hand, as the satellite is no longer heated by the sun, the thermal equilibrium of the satellite is greatly modified and the temperature tends to decrease rapidly.

2.1.7.1 Eclipses of the sun by the moon

The orbit of the moon around the earth, with a semi-major axis of 384 400 km and a period of 27 days, has an inclination of 5.14° with respect to the ecliptic. The RAAN on the ecliptic is also affected by a precession in the retrograde direction of period 18.6 years. The relative movement of an artificial earth satellite and the natural satellite is thus complex, and determination of the dates at which the artificial satellite is aligned with the sun–moon direction cannot easily be formulated for the general case. Examples will be given for the orbit of geostationary satellites (Section 2.2.5.6).

Eclipses by the moon are infrequent, are most often of short duration, and most often do not totally obscure the solar disc. They do not generally constrain the satellite design and operation unless they precede or follow an eclipse of the sun by the earth that extends the total time during which the satellite is in the dark.

2.1.7.2 *Eclipses of the sun by the earth*

The sun's rays are assumed to be parallel, and this corresponds to a sun assumed to be a point at infinite distance. The relationship between the declination δ of the sun and the latitude l of the satellite for there to be an eclipse is as follows (refer to Figure 2.6):

$$-\delta - \text{arc sin}(R_E/r) < \text{latitude of the satellite} < -\delta + \text{arc sin}(R_E/r)$$

The centre of the eclipse corresponds to a value of the nodal angular elongation u of the satellite (equal to the sum of the argument of the perigee ω and the true anomaly v of the satellite) that fulfils:

$$\alpha_{SUN} + \pi = \Omega + \arctan(\tan u \cos i)$$

where α_{SUN} is the right ascension of the sun and Ω is the RAAN of the satellite orbit.

The duration of the eclipse varies as a function of the distance r and the inclination i of the satellite orbit with respect to the declination of the sun. The longest durations are observed when the declination of the sun is equal to the inclination of the orbit.

2.1.8 Sun–satellite conjunction

A sun–satellite conjunction occurs when the sun is aligned with the satellite as seen from an earth station and leads to a large increase of the station antenna noise temperature. This occurs for a station situated on the track of the satellite when the following two conditions are satisfied:

— The latitude of the satellite is equal to the declination of the sun (angle δ, referring to Figure 2.6).
— The hour angle (or longitude) of the satellite is equal to the hour angle (or longitude) of the sun (α_{SUN}).

The conditions for occurrence and duration of the conjunction with the sun are discussed in Section 2.2 for particular types of orbit. The effect of sun–satellite conjunction at the earth station is examined in Chapter 8.

2.2 USEFUL ORBITS FOR SATELLITE COMMUNICATION

In principle, the plane of the orbit can have any orientation, and the orbit can have any form. The orbital parameters are determined by the initial conditions as the satellite is injected into orbit. With the Keplerian assumptions, these orbital parameters, and hence the shape and orientation of the orbit in space, remain constant with time. It will be seen in the following sections that, under the effect of various perturbations, the orbital parameters change with time. Hence, if it is required to maintain the satellite in a particular orbit, orbit control operations are necessary. The cost of these operations can be minimised by choosing particular values for certain orbit parameters in accordance with the constraints imposed by the telecommunications mission.

Systems based on polar or non-polar circular orbits have been proposed to provide world-wide communications services using low earth orbits (LEOs) or medium earth orbits (MEOs)

(Section 1.4). These systems entail constellations of several satellites, increasing in number with decreasing altitude of the orbit. The lower the altitude, the smaller the path loss in the link budget and the smaller the propagation delay. Moreover, satellites are viewed from the user with a high elevation angle, whatever the user's location. This makes such constellations attractive for personal mobile communications.

Inclined elliptical orbits are most useful for providing regional communications services to regions below the apogee of the orbit. In these regions, the satellite is viewed with a near-zenith elevation angle. This is of interest for mobile communications, but the high altitude of the apogee introduces a large path loss and delay. Only a few satellites are required.

Geostationary satellite systems (non-retrograde circular orbit in the equatorial plane at an altitude of 35 785 km) provide large coverage of the earth with a single satellite, or nearly worldwide coverage (polar regions excepted) with as few as three satellites.

2.2.1 Elliptical orbits with non-zero inclination

In an elliptical orbit, the velocity of the satellite is not constant. This velocity, given by Eq. (2.16), is maximum at the perigee and minimum at the apogee. Hence, for a given period, the satellite remains in the vicinity of the apogee for a longer time than in the vicinity of the perigee, and this effect increases as the eccentricity of the orbit increases. The satellite is thus visible to stations situated under the apogee for a large part of the orbital period, and this permits communication links of long duration to be established.

To establish repetitive satellite communication links, it is useful for the satellite to return systematically to an apogee above the same region. The period of orbits of this kind is thus a submultiple of the time taken by the earth to perform one rotation with respect to the line of nodes of the orbit. On the basis of the Keplerian hypotheses, the line of nodes is fixed in space, and this duration is equal to a sidereal day. In practice, it is necessary to take into account the rotation of the line of nodes (the drift of the RAAN) due to the effect of perturbations (see Section 2.3.2.3). The period of the orbit must thus be a submultiple of the time T_{EN} taken by the earth to turn through an angle equal to $(360° + \Delta\lambda)$, where $\Delta\lambda$ is the drift of the ascending node during time T_{EN} (Figure 2.8). This drift depends on the inclination, the eccentricity, and the semi-major axis of the orbit.

With an orbit of non-zero inclination, the satellite passes over regions situated on each side of the equator and possibly the polar regions, if the inclination of the orbit is close to 90°. By orientating the apsidal line (the line from the perigee to the apogee) in the vicinity of the perpendicular to the line of nodes (the argument of the perigee ω is close to 90° or 270°), the satellite at the apogee systematically returns above the regions of a given hemisphere. It is thus possible to establish links with stations located at high latitudes (see Figure 2.4).

The apogee of the orbit is permanently situated above the same hemisphere if there is no rotation of the orbit in its plane: that is, if the drift of the argument of the perigee is zero. This is the case with the Keplerian hypotheses. In reality, various perturbations cause the orbital parameters to vary. By choosing an inclination of 63.45°, the drift of the argument of the perigee becomes zero (see Section 2.3.2.3).

Although the satellite remains for several hours in the vicinity of the apogee, it does move with respect to the earth, and, after a time dependent on the position of the station, the elevation angle of the satellite as seen from the earth station decreases below some acceptable values. To establish permanent links, it is thus necessary to provide several suitably phased satellites in similar orbits that are spaced around the earth (with different right ascensions of the ascending node and, for example, regularly distributed between 0 and 2π) in such a way that the satellite moving away from the apogee is replaced by another satellite in the same region of the sky

as seen from the stations. In this way, the problems of satellite acquisition and tracking by the stations are simplified. The problem of switching the links from one satellite to another remains; the link frequencies of the various satellites can be different in order to avoid interference.

Different types of orbit can be envisaged. In the following sections, Molniya (12 hours) and Tundra (24 hours) orbits [BOU-90] and an interesting concept called LOOPUS are discussed.

2.2.1.1 *Molniya orbits*

These orbits take their name from the communications system installed by the Soviet Union to service territories situated in the northern hemisphere at high latitudes (see Figure 1.6). The period T of the orbit is equal to $T_{EN}/2$ or about 12 hours. The characteristics of an example orbit of this type are given in Table 2.4.

Table 2.4 Example of a Molniya orbit

Period (T) (half sidereal day)	12 h (11 h 58 min 2 s)
Semi-major axis (a)	26 556 km
Inclination (i)	63.4°
Eccentricity (e)	0.6 to 0.75
Perigee altitude h_p (e.g. $e = 0.71$)	$a(1-e) - R_E$ (1250 km)
Apogee altitude h_a (e.g. $e = 0.71$)	$a(1+e) - R_e$ (39 105 km)

The equation of the track is determined as indicated in Section 2.1.6 by considering that the angular velocity of the earth Ω_E is approximately equal to $n/2$, where n is the mean movement of the satellite. The relative longitude with respect to the satellite meridian as it passes through the ascending node is thus expressed from Eqs. (2.38a), (2.38b), and (2.38c) by:

$$\lambda = \lambda_{SL} - \Delta\lambda = \text{arc sin}[(\tan\varphi)/(\tan i)] - (M/2) + (M_N/2) \tag{2.52a}$$

or:

$$\lambda = \lambda_{SL} - \Delta\lambda = \text{arc tan}[(\tan u)/(\cos i)] - (M/2) + (M_N/2) \tag{2.52b}$$

or:

$$\lambda = \lambda_{SL} - \Delta\lambda = \text{arc cos}[(\cos u)/(\cos \varphi)] - (M/2) + (M_N/2) \tag{2.52c}$$

The latitude of the track is given by (2.40) as:

$$\varphi = \text{arc sin}[\sin i \sin u]$$

M and M_N, the mean anomalies of the satellite and the ascending node, are calculated from the eccentric anomalies by using Kepler's Eq. (2.25). The nodal angular elongation u is equal to $\omega + v$, where the true anomaly v is deduced from the eccentric anomaly E by means of Eqs. (2.21a)–(2.21e).

The track of the satellite on the surface of the earth is illustrated in Figure 2.12 for an argument ω of perigee equal to 270°. The satellite at the apogee passes successively on each orbit above two points separated by 180° in longitude. The apogee is situated above regions of latitude 63° (the latitude of the vertex is equal to the value of the inclination, and the apogee coincides with

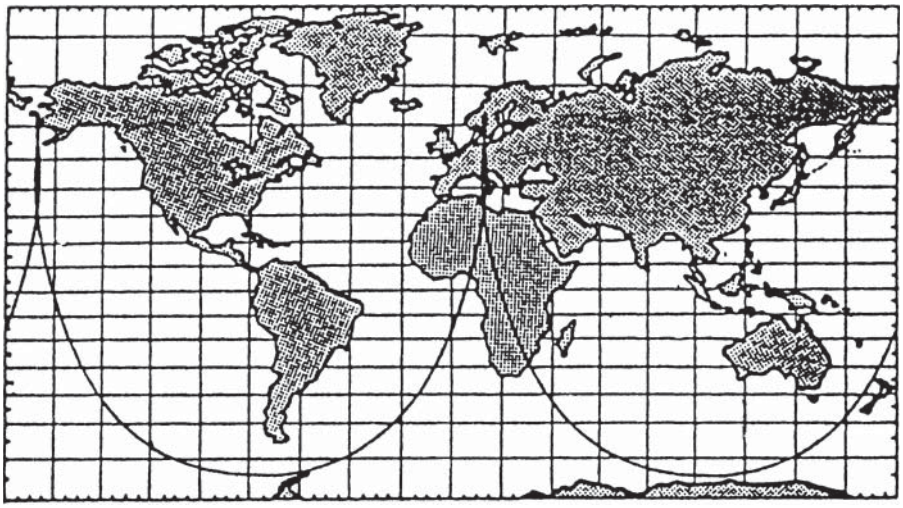


Figure 2.12 The track of a Molniya orbit ($\omega = 270°$). Source: after [ASH-88]. Reproduced by permission of the Institution of Electrical Engineers.

the vertex of the track when the argument of the perigee is equal to 270°). The large ellipticity of the orbit results in a transit time for the part of the orbit situated in the northern hemisphere greater than that in the southern hemisphere. As the argument of the perigee is 270°, the true anomaly v on crossing the equatorial plane has a value $v = 90$. The eccentric anomaly E_N of the corresponding ascending node thus has a value of 45° (calculated from Eq. (2.21d) with $e = 0:71$). Knowing the mean movement of the satellite $n = 2\pi/T$, Eqs. (2.24) and (2.25) give a transit time t_N from the perigee to the ascending node equal to 32 minutes. Hence the satellite remains for $2t_N$, or around 1 hour, in the southern hemisphere and for $T–2t_N$, on the order of 11 hours, in the northern hemisphere. The satellite thus remains for several hours in the vicinity of the apogee and hence is visible from the regions situated beneath it.

The value of inclination that makes the drift of the argument of the perigee (and thus that of the apogee) equal to zero is 63.45° (see Section 2.3.2.3). A value different from this leads to a drift that is not zero but remains small for values of inclination that do not deviate too greatly from the nominal value. By way of example, for an inclination $i = 65°$, which is a variation of 1.55°, the drift of the argument of the perigee has a value of around 6.5° per year.

When the argument of the perigee is different from 270°, the latitude of the satellite at the apogee is no longer the maximum latitude of the track. The variation of satellite velocity in the orbit is no longer symmetrical with respect to the point of maximum latitude, and the track of the satellite on the ground loses its symmetry with respect to the meridian passing through this point.

2.2.1.2 Tundra orbits

The period T of the orbit is equal to T_E, which is nearly 24 hours. The characteristics of an example orbit of this type are given in Table 2.5. An example of the track of the satellite on the surface of the earth is given in Figure 2.13 for an argument of the perigee equal to 270°.

The equation of the track is determined by considering that the angular velocity of the earth Ω_E is very little different from the mean movement n of the satellite, and hence $\Omega_E/n = 1$. The vertex

Table 2.5 Example of a Tundra orbit

Period (T) (half sidereal day)	24 h (23 h 56 min 4 s)
Semi-major axis (a)	42 164 km
Inclination (i)	63.4°
Eccentricity (e)	0.25 to 0.4
Perigee altitude h_p (e.g. $e = 0.25$)	$a(1-e) - R_E$ (25 231 km)
Apogee altitude h_a (e.g. $e = 0.25$)	$a(1+e) - R_E$ (46 340 km)

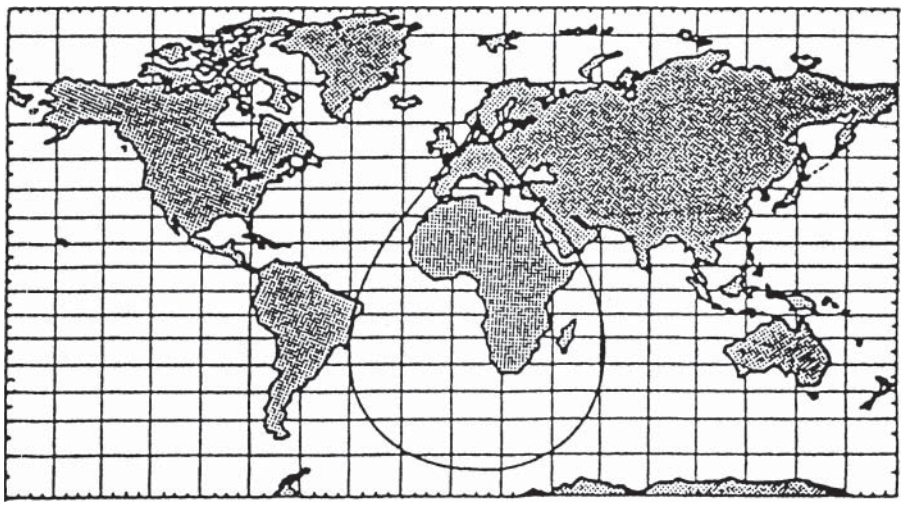


Figure 2.13 The track of a Tundra orbit ($\omega = 270°$). Source: after [ASH-88]. Reproduced by permission of the Institution of Electrical Engineers.

is chosen as the origin of the reference meridian, since regions of high latitude are those for which the orbit is of interest; hence, in Eq. (2.36), M_0 identifies with the mean anomaly of the vertex M_V. With respect to the vertex, the longitude λ_N of the node has a value $\lambda_N = -\pi/2$. Equations (2.34) and (2.35a-c) become:

$$\tan \varphi = \tan i \cos \lambda_s$$

$$\text{cotan } \lambda_s = -\tan u \cos i$$

Also:

$$\Delta\lambda = M(\Omega_E/n) - M_0(\Omega_E/n) \cong M - M_V$$

where M_V is the mean anomaly of the vertex.

The relative longitude with respect to the satellite meridian on passing through the vertex can thus be expressed:

$$\lambda = \lambda_S - \Delta\lambda = \text{arc cos}[(\tan \varphi)/(\tan i)] - M + M_V \tag{2.53a}$$

or:

$$\lambda = \text{arc cos}\{[\sin(\omega + v)](\cos i)/\sqrt{[1 - \sin^2 i \sin^2(\omega + v)]}\}$$
$$- (E - e \sin E) + (E_V - e \sin E_V) \tag{2.53b}$$

or again:

$$\lambda = -\text{arc cotan}[(\tan u)(\cos i)] - (E - e\sin E) + (E_V - e\sin E_V) \tag{2.53c}$$

where E is the eccentric anomaly of the satellite and E_V that of the vertex.

The latitude of the track is given by:

$$\varphi = \text{arc sin}[\sin i \sin(\omega + v)] \tag{2.39}$$

For $\omega = 270°$ and $i = 63.4°$, the variations of λ and φ are given in Figure 2.14 for various eccentricities. According to the value of the eccentricity, the loop above the northern hemisphere is accentuated to a greater or lesser extent. For an eccentricity equal to 0, the track has the form of a figure eight with loops of the same size and symmetrical with respect to the equator (see Section 2.2.3). When the eccentricity increases, the upper loop decreases, and the crossover point of the track is displaced towards the north. This loop disappears for a value of eccentricity on the order of 0.42. The transit time of the loop represents a substantial part of the period of the orbit and varies with the eccentricity.

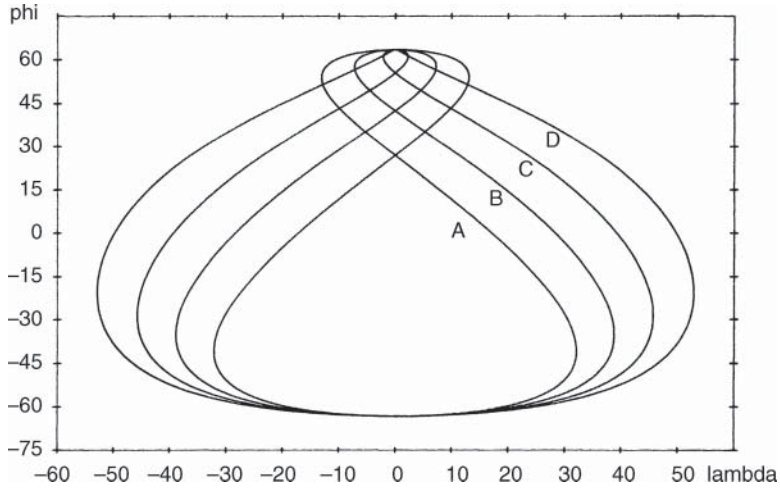


Figure 2.14 Tracks of a Tundra orbit ($i = 63.4°$, $\omega = 270°$) for various values of eccentricity: (a) $e = 0.15$; (b) $e = 0.25$; (c) $e = 0.35$; (d) $e = 0.45$.

The position of the loop can be displaced towards the west or east with respect to the point of maximum latitude by changing the value of the argument of the perigee v and the eccentricity e. Some examples are presented in Figure 2.15, where the various tracks represent different values of the argument of the perigee and the eccentricity.

2.2.1.3 Visibility of the satellite

The elevation angle and the time of visibility are two important parameters to be considered in the choice of orbit type. The ideal would be to have the satellite permanently at the zenith of the earth stations. For an operational system, variations of pointing angle with respect to the zenith are permissible either because the stations are equipped with a tracking system or because the antennas used have a large beam width. Limitations in the minimum values of elevation angle

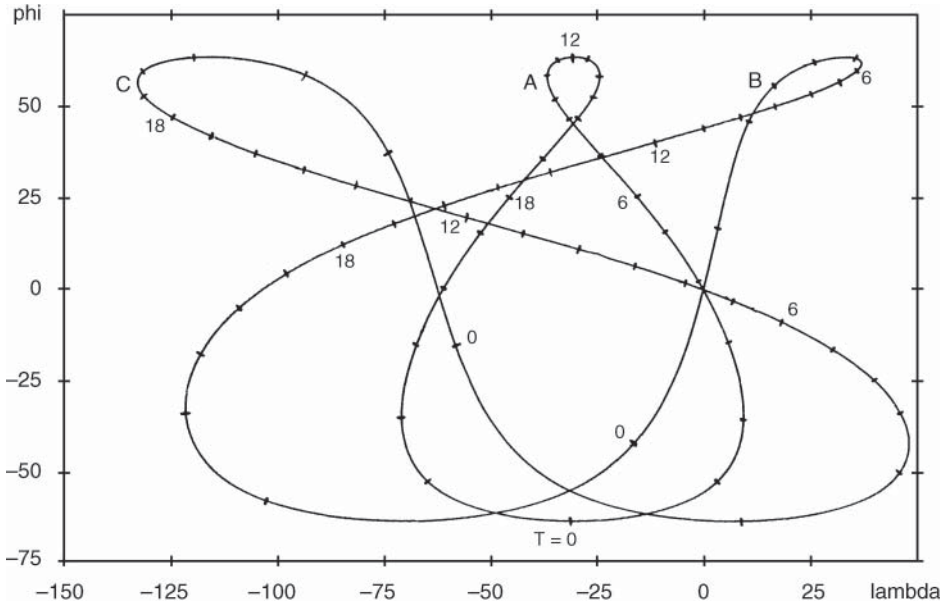


Figure 2.15 Variation of the track of a Tundra orbit ($i = 63.4°$) as a function of e and ω: (a) $e = 0:25$, $\omega = 270°$; (b) $e = 0.6$, $\omega = 315°$; (c) $e = 0.6$, $\omega = 202.5°$.

are due to the increase of antenna noise temperature and to problems of obstacles blocking the radio frequency link (particularly for mobiles). Specification of a permissible range of variation of pointing angle with respect to the zenith for a satellite on a given orbit and at a particular point on the orbit results in the definition of a geographical region within which the satellite is visible with an elevation angle greater than a fixed minimum value. This region enlarges when the satellite is further from the earth. But, within these regions, stations do not all see the satellite (which moves with respect to the earth) for the same duration. Stations situated on the track of the satellite, and particularly in the vicinity of the apogee, see the satellite for longer than the others. The time of visibility varies according to the considered location.

Orbits whose tracks contain loops are particularly useful since, for regions situated under the loop, the satellite moves, during entry to the loop and during the time within the loop, in the same region of sky as seen at a high elevation angle from the earth station.

Continuous visibility. To ensure continuous coverage of these regions, a system with several satellites is required such that, for any station in the region, when the tracked satellite disappears below the minimum elevation angle, it is replaced by another satellite that is visible at an elevation angle greater than this fixed value. The orbits of these satellites are generally similar as far as the form (a, e) and inclination (i) parameters are concerned, but the values of right ascension differ since the orbits of these satellites must be in different planes to take into account the rotation of the earth with respect to the plane of the orbits. This is explained in Figure 2.16: the station situated on the meridian M(1) acquires satellite S_1 at A, at which point the satellite is considered to be active. The station follows S_1 during its trajectory on the orbital arc AB of length 2ρ. At B, the elevation angle becomes less than the required minimum, and the satellite becomes inactive. The meridian M has turned during this interval and is now in position 2. By an appropriate phasing of its orbit, satellite S_2 enters the arc BC at this time, the station can acquire it with an elevation angle greater than the required minimum, and S_2 is thus active. It is necessary for

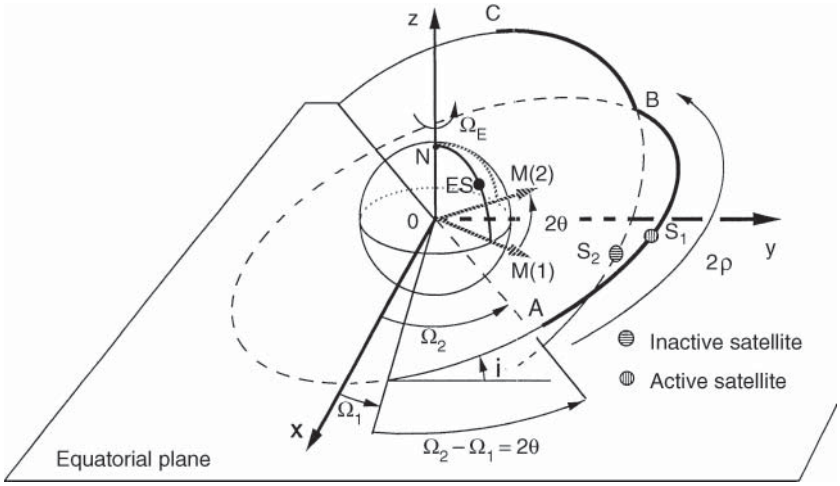


Figure 2.16 Satellites in orbits of different right ascension that are successively visible from an earth station at an elevation angle greater than a given value.

the right ascension Ω_2 of the orbit of S_2 to be offset with respect to Ω_1 by an amount 2θ through which the meridian of the station has turned during the motion of the satellite from A to B (the orbital planes are assumed to be fixed in space).

The number of satellites necessary to continuously cover a given geographical region depends on the fixed minimum elevation angle and the characteristics of the orbit.

Orbits of the Molniya type. With orbits of the Molniya type (period = 12 hours), a visibility duration of more than 8 hours is possible with large elevation angles in regions situated under the apogee (Figure 2.17). A system with three satellites on orbits of right ascension differing by 120° thus permits continuous visibility to be ensured in these regions. Figure 2.18 shows the orbits of these satellites seen by a fixed observer located in space away from the earth.

Orbits of the Tundra type. With an orbit of the Tundra type (period = 24 hours), a visibility duration of more than 12 hours is possible with high elevation angles; hence two satellites on orbits with right ascension differing by 180° are sufficient. The form of the track is given by the curve in Figure 2.15. Figure 2.19 shows the region within which the active satellite (or the two in the system) is seen with an elevation angle greater than 55°. Typical parameters of the orbit are as follows: $a = 42\,164$ km, $e = 0.35$, $i = 63.4°$, $\Omega = 270°$ (and 90°).

Figure 2.20 illustrates the apparent trajectory of the satellites seen by a distant observer rotating with the earth. The second satellite takes over from the previous one in the useful part of the trajectory on both sides of the apogee.

LOOPUS orbits. In a system with several satellites, one of the problems encountered by the earth stations is that of repointing the antenna during the handover from one satellite to the other. With orbits whose track contains a loop, it is possible to use only the loop as the useful part of the track of the trajectory; the satellite leaving the loop is replaced by another that enters it. Switching from one satellite to the other is thus performed at the crossover point of the track; at this instant, the two satellites are seen from the earth station in exactly the same direction. It is not, therefore, necessary to repoint the antenna. This principle, called LOOPUS, is described in [DON-84]. To achieve continuous coverage of the region situated under the loop, the transit time of the loop

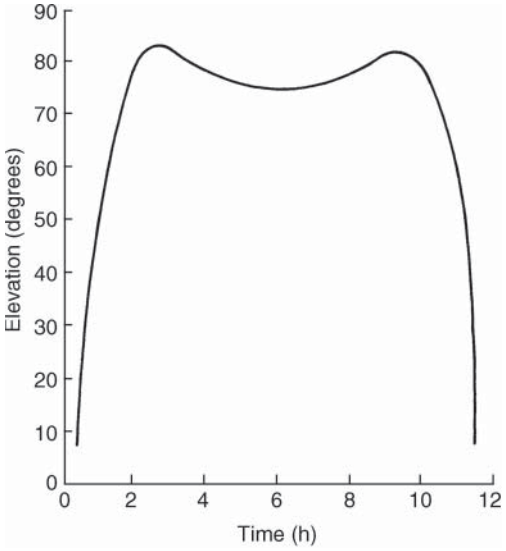


Figure 2.17 Example of the duration of visibility for an earth station in a region under the apogee in relation to the elevation angle for a Molniya orbit.

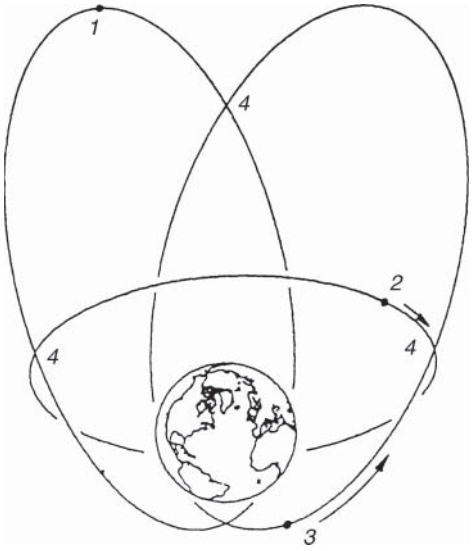


Figure 2.18 Three Molniya orbits whose right ascensions differ by 120° viewed from a fixed point in space. Source: from [DON-84]. Reproduced by permission of P. Dondl.

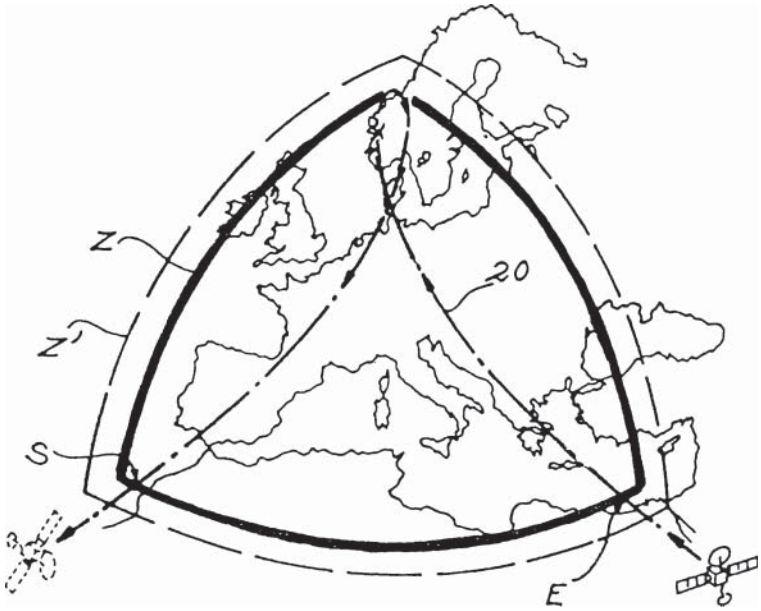


Figure 2.19 Zone of coverage with an elevation angle greater than 55° (two satellites in Tundra orbits). Source: after [ROU-88]. Reproduced by permission of the Institution of Electrical Engineers.

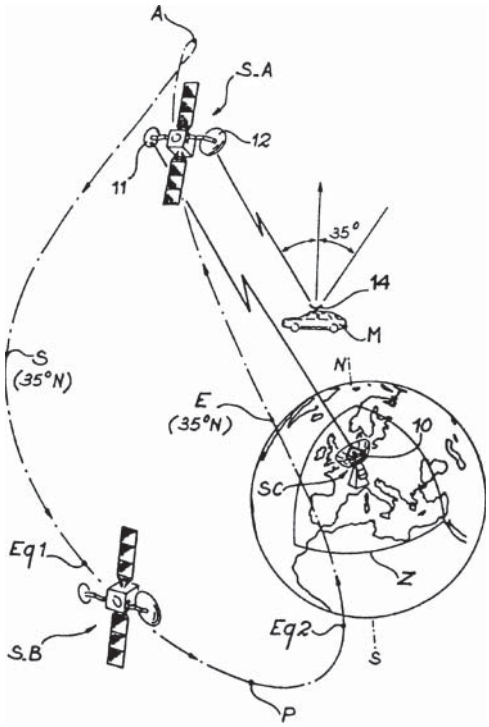


Figure 2.20 Apparent trajectory of satellites seen by an observer in space rotating with the earth.

must be a submultiple of the period of the orbit, and the number of satellites necessary is equal to the rank of the submultiple. The coverage can be extended to one part of the hemisphere by increasing the number of satellites in orbits regularly spaced about the globe.

To illustrate the concept, consider a system using an orbit of the Molniya type (12 hours) with the following parameters: $a = 26\,562$ km, $e = 0.72$, $i = 63.4°$, $\omega = 270°$. The track of this orbit on the ground contains a loop whose transit time is eight hours. Three satellites thus permit continuous coverage of the region below the loop.

In Figure 2.21, the useful arc CC′ corresponds to the loop of the track. For points C and C′ to coincide on the track, it is necessary for the meridian passing through C at the initial instant to be at C′ at the same time as the satellite. The earth has thus turned by an amount equal to the projection 2ρ on the equatorial plane of the variation of the true anomaly 2σ.

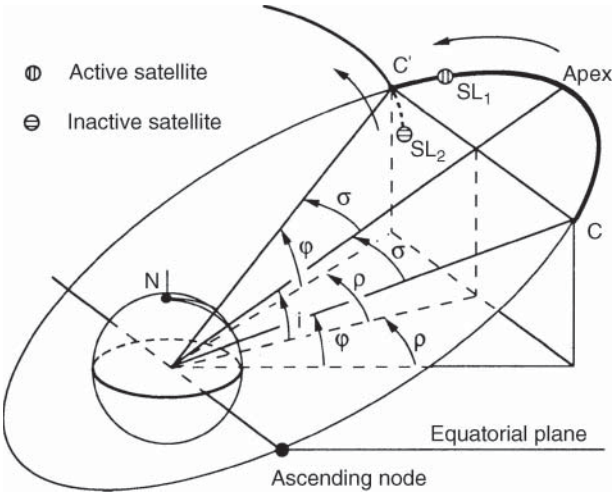


Figure 2.21 LOOPUS orbits.

Hence:

$$2\rho = 360°/(24\text{ h}/8\text{ h}) = 120°$$

The variation of the anomaly σ is such that $\tan\sigma = \cos i \tan\rho$, and hence $\sigma = 32:8°$.

The latitude φ of the crossover point (C or C′) is given by:

$$\varphi = \text{arc sin}(\sin i \cos\sigma)$$

and has a value $\varphi = 45°$.

The orbit of the succeeding satellite is such that the arc corresponding to CC′ is reproduced from the point C′. The difference in right ascension of the nodes is thus equal to 2ρ or 120°.

Example 2.2 Variation of the pointing angle of stations remains limited. Consider, for example, a station situated at the crossover point of the track. When a satellite arrives in the loop, it is at the zenith of the station. Subsequently, the elevation angle decreases during the four hours of movement of the satellite towards the north to the apogee. The variation ΔE of the elevation angle between the passage of the satellite through the crossover point B (at

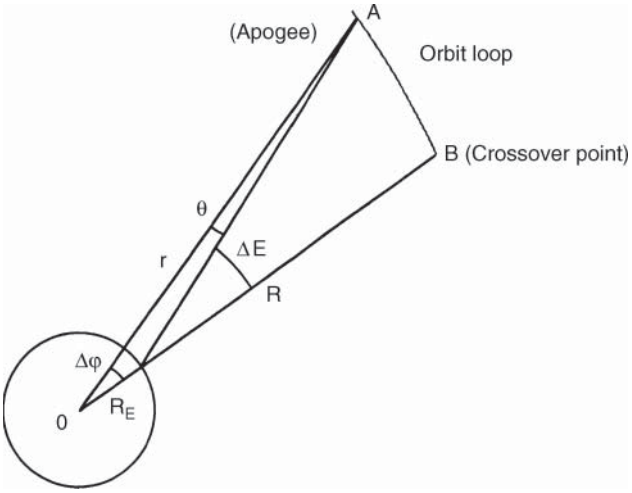


Figure 2.22 Variation of pointing angle during passage through a loop.

the zenith) and the apogee A of the orbit is calculated from Figure 2.22:

$$\mathrm{OA} = a(1+e) = r$$

$$\Delta E = \theta + \Delta\varphi$$

$$\Delta\varphi = (\text{latitude of the apogee} = i) - (\text{latitude of B} = \varphi = 45°)$$

$$(\sin\Delta\varphi)/R = (\sin\Delta E)/R_\mathrm{E}$$

$$R = \sqrt{(R_\mathrm{E}^2 + r^2 - 2R_\mathrm{E} r\cos\Delta\varphi)}$$

In the system considered with $e = 0.72$, $\omega = 270°$, and $i = 63.4°$, the variation of the pointing angle with respect to the zenith is on the order of 20° with a minimum elevation angle of 70° for a station situated at the crossover point of the track. The elevation angle then increases for four hours to become equal to 90° again at the instant when the satellite leaves the loop.

2.2.1.4 *Advantages of high-inclination elliptic orbits*

Large elevation angle. The main application of inclined elliptic orbits is to ensure coverage of regions at high latitude under a large elevation angle with satellites whose apparent movement with respect to the earth is small. A high elevation angle is particularly required in applications that include systems for communication with mobiles. Blocking of the radio frequency link due to occultation of the satellite by buildings and trees is minimised. Multiple paths caused by successive reflections by various obstacles are also reduced in comparison with systems operating with low elevation angles (geostationary satellite systems, for example). Tracking of the satellite is facilitated by the small apparent movement and the long duration of visibility. It is even possible to use antennas whose 3 dB beam width is a few tens of degrees with fixed pointing towards the zenith; this permits the complexity and cost of the terminal to be reduced while retaining sufficient gain. Finally, the noise captured by the earth station antenna from the ground or due to interference from other terrestrial radio systems is minimised due to the high elevation angle.

This applies to all signal attenuation and noise-generation effects (atmospheric gases, rain, etc.) associated with the slant path through the atmosphere. These advantages have led Russia to use these orbits for a long time in order to provide coverage of high-latitude territories; their use is also of interest for satellite systems providing communication with mobiles. The concept is used today by the Sirius satellite radio system [AKT-08]. The use of inclined elliptical orbits coupled with multiple modes of transmission diversity provides notable advantages for broadcast services to vehicles. Indeed, Sirius uses three satellites in Tundra-type orbits of eccentricity 0.2684 separated by 120° in the RAAN, which results in separation from the other two by eight hours in ground track. The satellite phasing and orbit ellipticity ensure that there are always two active satellites visible from the North American service area (fixed apogee longitude of 96°) with average elevation angle higher than 60°. As the two active satellites transmit the same signal with a four-second delay, the spatial diversity (satellites seen in different directions) and temporal diversity increase the chance that the user will receive a signal that does not experience blockage or foliage attenuation.

Sun–satellite eclipses and conjunction. In a communication system that operates satellites on elliptic orbits of high inclination, the operational part of the orbit is situated on each side of the apogee, most often coincident with the vertex of the orbit. If the inclination of the orbits is 63.4°, the maximum latitude of the satellite is 63.4° and its minimum value during the operational phase depends on the extent of the active part on each side of the vertex. This extent is reduced when the number of satellites is large. In this case, the latitude of the satellite remains high during the operational phase, and eclipses are not frequent during this phase (Section 2.1.7). This is confirmed for the Molniya and Tundra systems with three satellites. With a two-satellite Tundra system, eclipse occurrence depends on the values chosen for the RAAN. A similar analysis can be performed in connection with solar interference.

2.2.1.5 Disadvantages of high-inclination elliptic orbits

Traffic switching between satellites. More than one satellite in orbit is necessary in order to provide continuous service over a given geographical region, and this adversely affects the cost of the space segment. Furthermore, it is necessary periodically to hand over the traffic from one satellite to another. These special procedures cause an operational load at the control centre and reduce capacity during switching; it may be necessary to have two antennas at each earth station to point at the two satellites simultaneously and thus transfer the traffic from one satellite to the other without interruption of service.

Variation of distance. The variation of distance between the service region and the satellite during the time of activity of the latter is greater with orbits of the Molniya type than with orbits of the Tundra type. This distance variation has the following consequences:

— Variation of propagation time (52 ms variation for Molniya orbits).
— Doppler effect (14 kHz for Molniya orbits and 6 kHz for Tundra orbits in the L Band (1.6 GHz) [ASH-88]).
— Variation of received carrier level (4.4 dB for Molniya orbits) on the up- and downlinks.
— Modification of the coverage of the satellite antennas. Figure 2.23 shows the coverage obtained at the apogee and at the point of switching from one satellite to the other; Europe is seen through an angle of 4.9° at the apogee, which changes to 8.4° at the crossover point with a Molniya orbit (3.6° to 4.6° for a Tundra orbit). The antenna aperture can be optimised so that the decrease of gain at the edge of coverage at the crossover point with respect to the gain at the apogee is compensated by the reduction of free space losses.

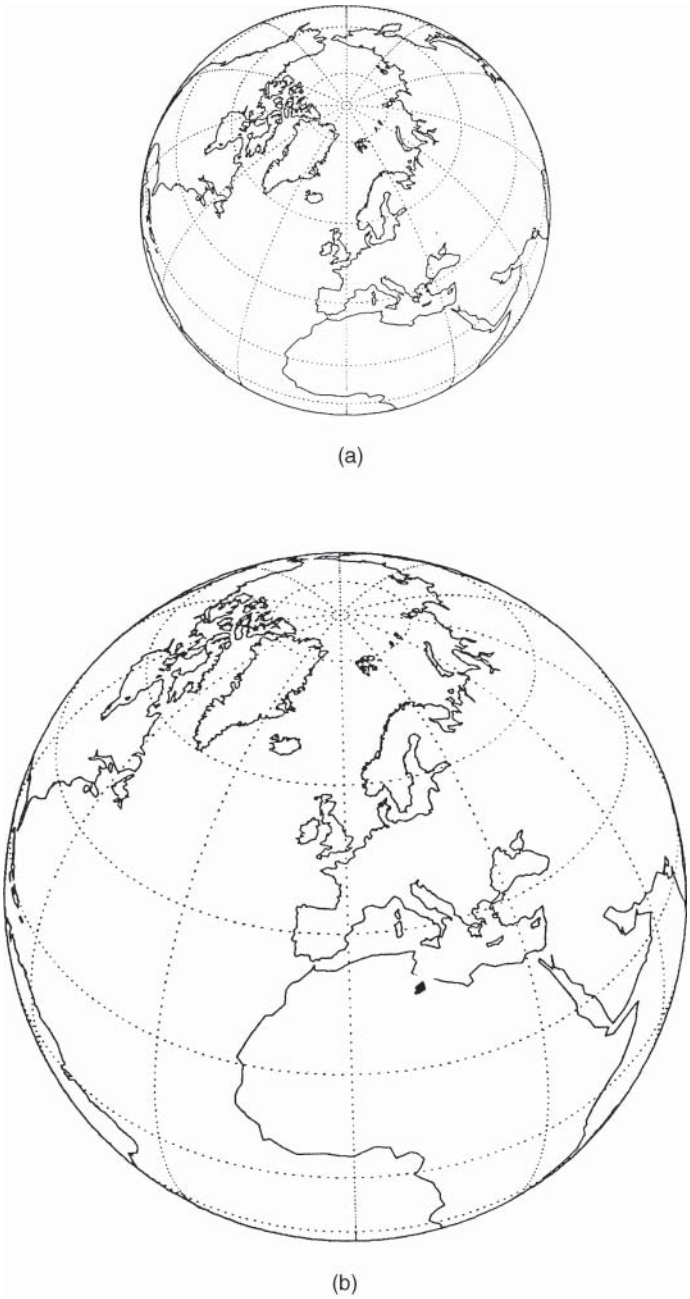


Figure 2.23 Variation of coverage (Molniya orbit) between (a) the apogee (visual angle 16.1°) and (b) the point of switching from one satellite to the other (visual angle 24.7°). Source: from [DON-84]. Reproduced by permission of P. Dondl.

Radiation. Molniya orbits are characterised by a perigee at an altitude of around 1200 km, which means the satellite crosses the Van Allen belts twice per orbit (their altitude is on the order of 20 000 km; see Chapter 11); in these belts are high-energy radiations that degrade the semiconductor components (such as solar cells and transistors) used in the satellite. Tundra orbits have the advantage of reducing the duration of crossing these bands.

Perturbations of orbit. For elliptic orbits of low altitude at the perigee, the satellite is strongly subjected to the effects of the asymmetry of the terrestrial potential, and this leads to perturbations of the orbit that must be controlled.

2.2.2 Geosynchronous elliptic orbits with zero inclination

The inclination is equal to zero. The period of the orbit is equal to one sidereal day (there is no longer a drift of the ascending node). The mean movement n of the satellite is equal to Ω_E, the angular velocity of the earth. The track of the satellite remains in the equatorial plane and becomes a periodic oscillation (period T_E) about the point of longitude λ_P representing the satellite at the perigee.

The longitude of the sub-satellite point with respect to that of the perigee can be written:

$$\Lambda = \lambda - \lambda_P = v - \Omega t = v - M = \text{arc cos}[(\cos E - e)(1 - e\cos E)^{-1}] - (E - e\sin E) \tag{2.54}$$

The maximum longitudinal shift Λ_{max} is obtained for $d(\lambda - \lambda_P)/dE = 0$, which leads to [BIE-66]:

$$\cos E_m = [1 \pm (1 - e^2)^{1/4}]e^{-1} \tag{2.55}$$

Figure 2.24 gives the *maximum longitudinal shift* Λ_{max} and the time t necessary to reach this point as a function of the eccentricity e.

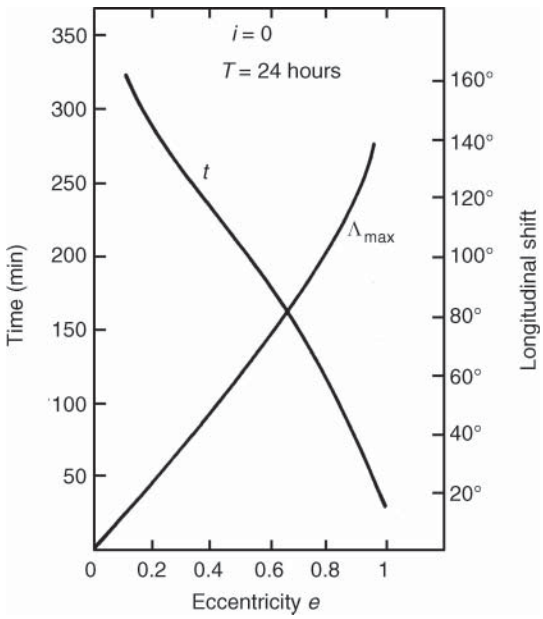


Figure 2.24 Maximum longitudinal shift of a synchronous equatorial satellite and the time t after which this shift is achieved as a function of the eccentricity of the orbit.

For an eccentricity less than 0.4:

$$\Lambda_{max} \cong 2e(\text{rad}) = 114e \quad \text{(degrees)} \tag{2.56}$$

If the eccentricity is small (10^{-3}), the maximum is reached after six hours. This time decreases for large values of eccentricity.

2.2.3 Geosynchronous circular orbits with non-zero inclination

The eccentricity is equal to zero. The period of the orbit is not very different from a sidereal day (the difference comes from the effect of the drift of the ascending node). The mean movement n of the satellite is thus not very different from Ω_E, the angular velocity of the earth. The nodal angular elongation u has a value $u = nt_{NS} \cong \Omega_E t_{NS}$, where t_{NS} is the time elapsed from the passage of the satellite at the ascending node N to the present position of the satellite S. Movement of the satellite in its orbit is at constant angular velocity. On the other hand, the projection of this movement on the equatorial plane is not at constant velocity. There is thus an apparent movement of the satellite with respect to the reference meridian on the surface of the earth (that of the satellite on passing through the nodes).

The projection of the satellite orbit on the equatorial plane is illustrated in Figure 2.25 by the dotted curve. The projection of point A (the position of a fictitious satellite rotating at the velocity of the reference meridian in the plane of the equator) perpendicularly to the line of nodes cuts this curve at point B. The coordinates of B, in a geocentric reference in the plane of the equator such that O_x is along the line of nodes and O_y is orthogonal to O_x, are:

$$x_B = R_E \cos \Omega_E t \text{ and } y_B = R_E \sin \Omega_E t \cos i.$$

Hence: $\tan(\zeta t) = y_B / x_B = \cos i \tan(\Omega_E t)$, from which $\Omega_E t = \text{arc tan}[(1/\cos i)\tan(\zeta t)]$.

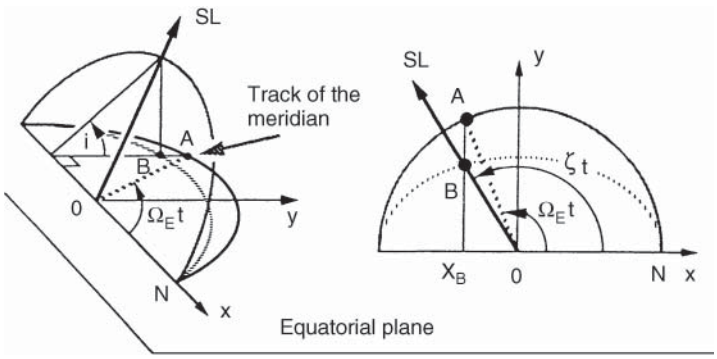


Figure 2.25 Projection of the orbit of a synchronous satellite on to a circular orbit of non-zero inclination.

Differentiating gives $\Omega_E = d(\Omega_E t)/dt$:

$$\Omega_E = \zeta[1 + \tan^2(\zeta t)]/\{\cos i + [\tan^2(\zeta t)]/\cos i\} \tag{2.57}$$

In the vicinity of the nodes, ζt tends to 0, from which $\Omega_E \cong \zeta / \cos i$. Hence $\zeta \cong \Omega_E \cos i$; the angular velocity of the satellite meridian is less than that of the reference meridian, and the satellite drifts towards the west.

In the vicinity of the point of maximum latitude, ςt tends to $\pi/2$, from which $\Omega_E \cong \cos i$. Hence $\varsigma \cong \Omega_E/\cos i$; the angular velocity of the satellite meridian is greater than that of the reference meridian, and the satellite drifts towards the east.

Taking the satellite meridian on passing through the ascending node as a reference, the relative longitude is calculated from Eq. (2.38):

$$\lambda = \lambda_{SL} - \Delta\lambda = \text{arc}\ \sin[(\tan\varphi)/(\tan i)] - \Omega_E t_{NS} \tag{2.58a}$$

Hence:

$$\lambda = \text{arc}\ \sin\{(\cos i \sin u)/\sqrt{(1 - \sin^2 i \sin^2 u)}\} - u \tag{2.58b}$$

and:

$$\lambda = \text{arc}\ \tan[(\tan u)(\cos i)] - u \tag{2.58c}$$

The latitude φ has a value from (2.40):

$$\varphi = \text{arc}\ \sin(\sin i \sin u)$$

The track of the satellite as u varies from 0° to 360° is represented in Figure 2.26 for various values of inclination i. The maximum latitude φ_m attained (the vertex) is equal to the value of inclination i of the orbit. The associated longitude λ is zero (with respect to the reference meridian).

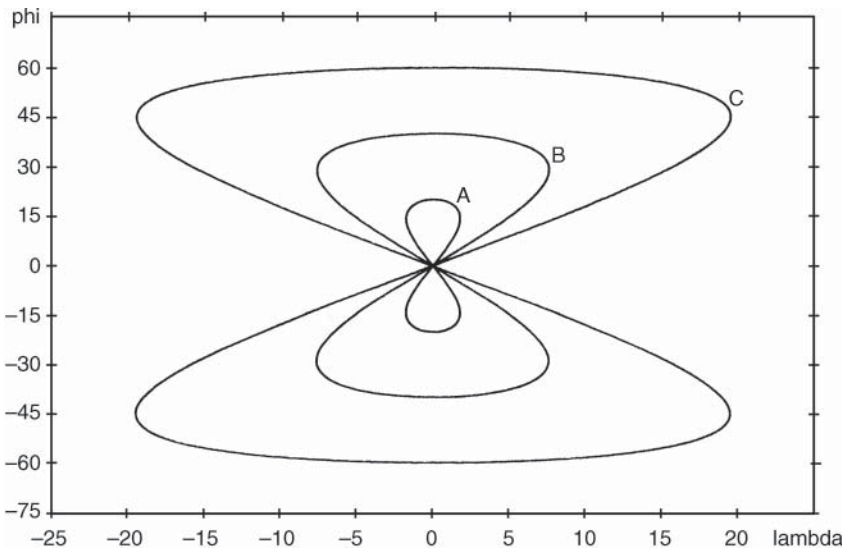


Figure 2.26 Tracks of circular synchronous orbits for various values of inclination: (a) $i = 20°$; (b) $i = 40°$; (c) $i = 60°$.

The maximum displacement in longitude λ_{max} with respect to the reference meridian is obtained for $d\lambda/du = 0$, which leads to:

$$\tan u = 1/\sqrt{(\cos i)} \quad \text{or} \quad \sin u = 1/\sqrt{(1 + \cos i)} = 1/(\sqrt{2})\cos(i/2)$$

from which:

$$\lambda_{max} = \text{arc}\ \tan[\sqrt{(\cos i)}] - u = \text{arc}\ \cos[1/\sqrt{(1 + \cos i)}] - \text{arc}\ \sin[1/(1 + \cos i)]$$

Hence:

$$\lambda_{max} = \text{arc cos}[1/(\sqrt{2})\cos(i/2)] - \text{arc sin}[1/(\sqrt{2})\cos(i/2) \tag{2.59a}$$

or:

$$\lambda_{max} = \text{arc sin}[(1-\cos i)/(1+\cos i)] = \text{arc sin}[\sin^2(i/2)/\cos^2(i/2)] \tag{2.59b}$$

The associated latitude φ_m at the maximum displacement λ_{max} can be written:

$$\varphi_m = \text{arc sin}(\sin i \sin u) = \text{arc sin}\{(\sin i)/[(\sqrt{2})\cos(i/2)]\}$$

Hence:

$$\varphi_m = (\sqrt{2})\sin(i/2) \tag{2.60}$$

For small i:

$$\lambda_{max} = i^2/4 \quad \text{and} \quad \varphi_m = i/\sqrt{2} \quad \text{(rad)} \tag{2.61}$$

2.2.4 Sun-synchronous circular orbits with zero inclination

A communications satellite must be visible from the regions concerned during the periods when it is desired to provide a communications service; this can vary from a few hours to 24 hours per day. When the service is not continuous, it is desirable that the intervals during which the service is available repeat each day at the same time. A satellite following a sun-synchronous equatorial orbit can cover a given geographical region at the same local time each day. The duration of uninterrupted service that such a satellite can provide in a given region on the terrestrial surface is a function of its altitude and the latitude of the receiver. Table 2.6 shows some typical visibility durations (CCIR-Rep 215) [ITUR-90]. Such orbits could be considered for satellite broadcasting systems.

Table 2.6 Duration of visibility for satellites in a geostationary orbit or a sun-synchronous circular equatorial orbit (non-retrograde)

Approximate period (h)	Altitude (km)	Number of transits per day above a given point	Approximate duration of visibility above the horizon on each transit (h)			
			At the equator	At ±15° latitude	At ±30° latitude	At ±45° latitude
24*	35 786	Stationary [(24/h) – 1]	Continuous	Continuous	Continuous	Continuous
12	20 240†	1	10.1	10.0	9.9	9.3
8	13 940†	2	4.8	4.7	4.6	4.2
6	10 390†	3	3.0	2.9	2.8	2.5
3	4190†	7	1.0	1.0	0.9	0.6

*Exact period = 23 h 56 min 4 s.
†Approximate value.

2.2.5 Geostationary satellite orbits

A particular case of the preceding section, the circular orbit ($e = 0$) in the equatorial plane ($i = 0$), is geosynchronous. The angular velocity of the satellite is the same as that of the earth ($n = \Omega_E$)

and in the same direction (direct orbit). The track of the satellite is reduced to a point on the equator; the satellite remains permanently on the vertical at this point. To a terrestrial observer, the satellite appears fixed in the sky. Table 2.7 shows the characteristics of orbit.

Table 2.7 Characteristics of the nominal Keplerian orbit of a geostationary satellite

Semi-major axis	$a = r$	42 164.2 km
Satellite velocity	$V_S = \sqrt{(a^3/\mu)}$	3075 m s^{-1}
Satellite altitude	R_0	35 786.1 km
Mean equatorial radius	R_E	6378.1 km
Ratio	R_0/R_E	6.614

The semi-major axis a of the orbit is such that:

$$2\pi\sqrt{(a^3/\mu)} = T_E = 1 \text{ sidereal day} = 86\,164.1 \text{ s}$$

2.2.5.1 *Distance of the satellite from an earth station*

For a geostationary satellite, Eq. (2.41) can be put into the form:

$$R^2 = R_E^2 + r^2 - 2R_E r\cos\phi \text{ with } r = R_E + R_0$$

Hence:

$$R^2 = R_0^2 + 2R_E(R_0 + R_E)(1 - \cos\phi) \tag{2.62}$$

The values of R_0 and R_E give $R_E/R_0 = 0.178$, and this gives:

$$(R/R_0)^2 = 1 + 0.42(1 - \cos\phi) \tag{2.63}$$

$$\cos\phi = \cos L\cos l \tag{2.64}$$

where l is the *latitude* of the station and L is the *relative longitude* of the satellite with respect to the station. The variation of $(R/R_0)^2$ as a function of l is given, for various values of L, by the curves in Figure 2.27. The maximum value of $(R/R_0)^2$ is 1.356. When R^2 is replaced by R_0^2, the maximum error is 1.3 dB.

2.2.5.2 *Elevation and azimuth angle*

From an earth station whose position is defined by its latitude l and its relative longitude L with respect to the satellite, the elevation angle E at which the satellite is seen is obtained from Eqs. (2.44) with $r = R_E + R_0$. Hence, for example:

$$E = \arctan\{[\cos\phi - (R_E/(R_E + R_0))]/\sqrt{(1 - \cos^2\phi)}\}$$

where the angle ϕ is given by Eq. (2.64).

Figure 8.12 indicates the value of elevation angle E as a function of the earth station position relative to that of the satellite. The *azimuth angle* A is obtained from the intermediate parameter a defined from Eq. (2.45) with $\varphi = 0$. Hence:

$$a = \text{arc } \sin[\sin L/\sin\phi] \text{ (with } \phi > 0, L > 0) \tag{2.65}$$

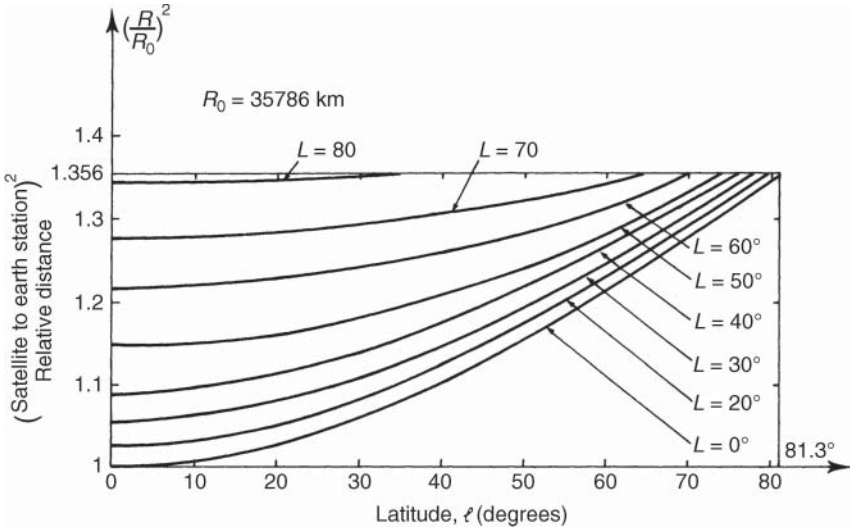


Figure 2.27 Variation of the square of the ratio of the station–satellite distance R to its nominal altitude R_0 as a function of the satellite–station latitude l and relative longitude L.

Table 2.8 Determination of the azimuth A

Hemisphere of station	Position of satellite with respect to station	Relation between A and a
Northern	East	$A = 180° - a$
Northern	West	$A = 180° + a$
Southern	East	$A = a$
Southern	West	$A = 360° - a$

The true azimuth A is obtained from a as a function of the position of the station with respect to the satellite by using Table 2.8 and Figure 8.11.

2.2.5.3 Nadir angle and maximum coverage

From Eq. (2.46), the nadir angle θ is equal to arc sin$[R_E(\cos E)/r]$. The maximum geographical coverage is given by the portion of the earth included in a cone that is tangential to the surface of the earth and has the satellite at its vertex.

The limiting elevation angle is thus $E = 0°$. The vertex angle of the cone, the angle at which the earth is seen from the geostationary satellite, is:

$$2\theta_{max} = 2\text{arc sin}[R_E/(R_0 + R_E)] = 17.4° \tag{2.66}$$

The maximum latitude l_{max}, or the maximum deviation in longitude L_{max} with respect to the satellite, corresponds to the value of ϕ_{max} given by Eq. (2.46b), where $E = 0°$ and $\theta = \theta_{max} = 8.7°$; that is, $\phi_{max} = l_{max} = L_{max} = 81.3°$.

2.2.5.4 *Propagation time*

The distance between two ground stations via the satellite varies between

$$2R_{\max}(L = 0°, l = 81.3°) = 83\,352.60 \text{ km and } 2R_0 = 71\,572.2 \text{ km}.$$

The propagation time is greater than 0.238 seconds and can reach 0.278 seconds.

2.2.5.5 *Eclipses of the sun by the earth*

Knowledge of the duration and periodicity of eclipses is important in the case of satellites that use solar cells as a source of energy. Furthermore, an eclipse causes a thermal shock that should be taken into account in the design of the satellite.

Duration of an eclipse. The movement of the earth around the sun is represented in Figure 2.5. Figure 2.28 shows the apparent movement of the sun with respect to the equatorial plane. The orbit of the satellite is perpendicular to the plane of the figure. At the solstices, the satellite is always illuminated; but in the vicinity of the equinoxes, it could pass in the earth's shadow. Considering, as a first approximation, that the sun is a point at infinity, this shadow is a cylinder that is tangential to the earth. On the day of the equinox, the eclipse has a maximum duration $d_{\max}$ determined from Figure 2.29 such that:

$$d_{\max} = (17.4°/360°) \times (23 \text{ h} \times 60 \text{ min} + 56 \text{ min}) = 69.4 \text{ min}$$

In reality, the sun has an apparent diameter of 0.5° as seen from the earth, and there is a cone of shadow where the eclipse is total and a region of penumbra where the eclipse is partial (see Figure 10.42). The penumbra has a width equal to the apparent diameter of the sun: that is, 0.5°. In its orbit, the satellite moves 1° in four minutes; also, the total duration of the eclipse is equal to 71.5 minutes, of which 2 minutes are penumbra at the start and finish.

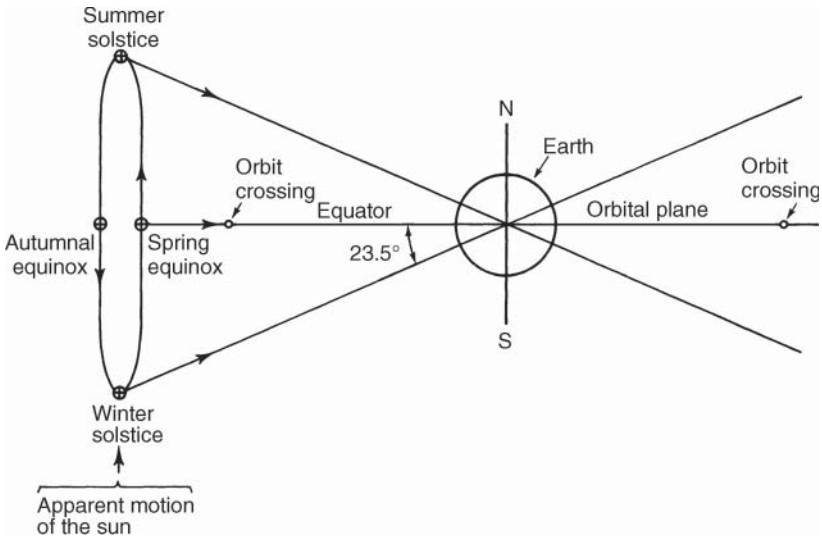


Figure 2.28 Apparent movement of the sun with respect to the orbit of geostationary satellites.

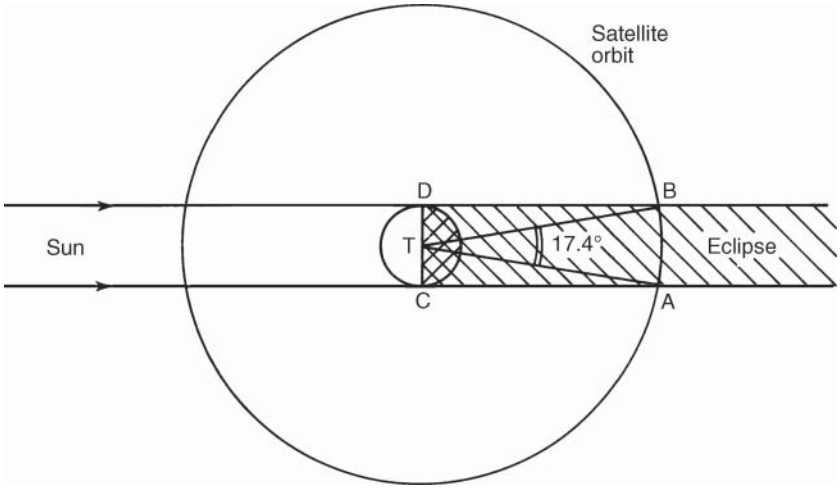


Figure 2.29 Eclipses at the equinoxes (figure in the plane of the equator).

Useful Orbits for Satellite Communication

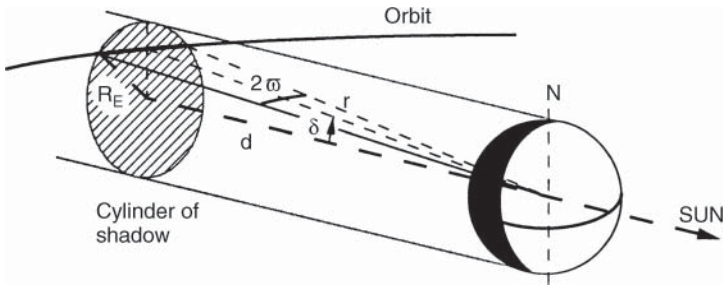


Figure 2.30 Geometry of the shadow region and the orbit out of equinox.

To evaluate the duration of the eclipse other than at the equinox, consider Figure 2.30, where the arc $2\overline{\omega}$ is the arc of the orbit contained within the cylindrical shadow of radius R_E and δ_{SUN} is the declination of the sun. This gives:

$$\cos\overline{\omega}\cos\delta_{SUN} = d/r \quad \text{and} \quad d^2 + R_E^2 = r^2$$

from which

$$\cos\overline{\omega} = \sqrt{[1-(R_E/r)^2]}/\cos\delta_{SUN} = 0.9885/\cos\delta_{SUN} \tag{2.67}$$

First and last day of an eclipse. The first day of the eclipse before the spring equinox corresponds to the relative position of the sun such that the cone of the earth's shadow is tangent to the satellite orbit.

Figure 2.31 illustrates the situation before the autumn equinox as the declination of the sun decreases. The value of $\overline{\omega}$ is thus zero ($\cos\overline{\omega} = 1$), and the declination of the sun δ_0 is such that

$$\cos\delta_0 = \sqrt{[1-(R_E/r)^2]} \quad \text{or} \quad \sin\delta_0 = R_E/r$$

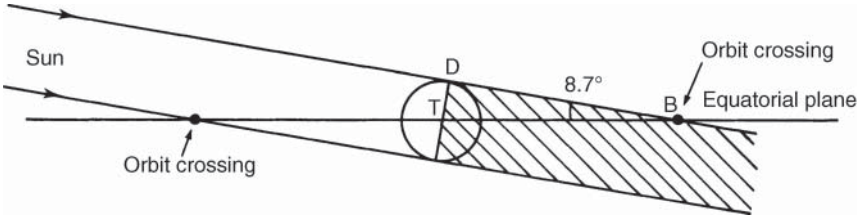


Figure 2.31 First day of the eclipse before the autumn equinox (the plane of the equator is perpendicular to the plane of the figure).

and hence $\delta_0 = \arcsin R_E/r = 8.7°$ (the triangle BDT is the same as the corresponding triangle in Figure 2.29).

The last day of the eclipse at the autumn equinox would correspond to a figure that is symmetrical with respect to the equatorial plane. A similar situation occurs again at the spring equinox. When the declination of the sun is greater than the absolute value of δ_0, the shadow does not intercept the orbit, and there is no eclipse.

The first and last days of the eclipse seasons are obtained by determining the dates at which the declination of the sun has a value $\delta_0 = \pm 8.7°$: that is, $\sin \delta_0 = \pm R_E/r = \pm 0.15128$. From Eq. (2.26), the values of nodal angular elongation of the sun are $u = \arcsin[\pm 0.151\ 28/\sin \overline{\omega}]$: that is, $u = \pm 22.34°$ and $u = \pm 22.34° + 180°$.

The true anomalies of the sun before and after the vernal and autumn equinoxes are deduced from $v = u + \overline{\omega}_{SUN}$. Equations (2.21) and (2.25) enable the values of the eccentric anomalies and hence those of the associated mean anomalies to be calculated: $M_1 = 54.17°$ and $M_2 = 98.57°$, $M_3 = 232.34°$, and $M_4 = 282.31°$.

The related dates of passing through the perigee are given by $t = M/n_{SUN}$, where $n_0 = 2\pi/365.25$ is the mean movement of the sun in radians per day. This gives $t_1 = 54$ d 23 h, $t_2 = 99$ d 23.5 h, $t_3 = 240$d 19 h, $t_4 = 286$ d 10 h. Since passage through the perigee occurs between 2 and 3 January, the dates of the start and end of the periods of eclipse are 26 February and 12 April for the spring equinox and 31 August and 16 October for the autumn equinox.

Furthermore, the dates of the spring and autumn equinoxes with respect to passing through the perigee are 77 d 8 h and 263 d 18.5 h respectively: that is, 21 March and 23 September. The eclipse season thus extends over about 22 days before and after the spring equinox and 23 days before and after the autumn equinox.

The daily duration of the eclipse is calculated from Eq. (2.67) by knowing the value of the declination d at the considered date. The total duration of the eclipse is $8\overline{\omega}$ minutes, $\overline{\omega}$ being expressed in degrees. Figure 2.32 gives the daily duration of the eclipse, assuming a cylindrical shadow and a circular earth orbit about the sun.

Time of an eclipse. At half the daily duration, the satellite crosses the plane orthogonal to the equatorial plane formed by the sun and the axis of the earth. It is thus midnight true solar time at the longitude of the satellite.

The eclipse starts at $24\overline{\omega}/360$ (h) before true solar midnight and ends at $24\overline{\omega}/360$ (h) after midnight. The mean solar time from which legal time is defined is obtained by adding the time equation ΔE given by Eq. (2.32). The time equation varies from +12 to 0 minutes during the spring equinox eclipse season and from 0 to −15 minutes during the autumn equinox eclipse season.

Example 2.3 This example calculates the time of the start of the eclipse at the autumn equinox. At the equinox, the total duration of the eclipse is 71.5 minutes.

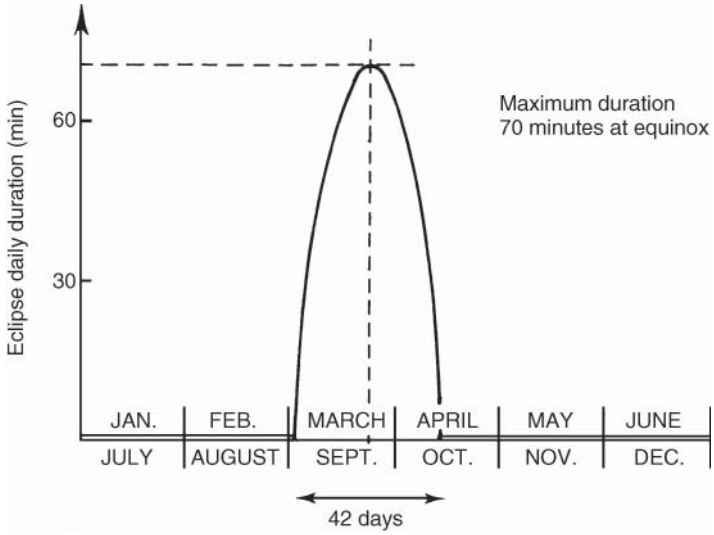


Figure 2.32 Daily duration of eclipses as a function of the date (simplified assumptions).

The true solar time of the start of the eclipse (for the meridian of the satellite) is:

$$TL = 12\text{h} - (71.5/2) = 11\text{ h } 24.25\text{ min}(\text{true solar time is } 23\text{ h } 24.25\text{ min}).$$

The time equation on 23 September gives:

$$\Delta E = 460 \sin n_{\text{SUN}} t - 592 \sin 2(\omega_{\text{SUN}} + n_{\text{SUN}} t) = -453\text{s} = -2.5\text{ min}$$

with:

$$t = 263\text{ d } 18.5\text{ h}, n_{\text{SUN}} = 360/362.5 = 0.985626^\circ/\text{day}, \omega_{\text{SUN}} = 280^\circ.$$

The mean solar time of the start of the eclipse is thus:

$$\text{TM} = \text{TL} + \Delta E = 11\text{ h}(24.25 - 2.5)\text{min} = 11\text{ h } 16.75\text{ min}$$

For a satellite at longitude λ, the universal time UT will have this value:

$$\text{UT} = 11\text{ h } 16.75\text{ min} - 12\text{ h} + \lambda/15.$$

With, for example, $\lambda = 19^\circ$ W, the universal time has a value of 0 h 33 min at the start of the eclipse. The legal time of the service region usually differs with respect to UT by an integer number of hours; hence, for France (summer time), for example:

$$\text{Time of the start of the eclipse} = \text{UT} + 2 = 02\text{ h } 33\text{min}.$$

Operation during an eclipse. If a satellite uses solar energy as a source of power, and if the satellite must provide continuous service, it is necessary to provide an energy store that permits normal operation at the equinoxes for about 70 minutes.

Another solution is to use the backup satellite, if one exists. This is satisfactory if the two satellites are sufficiently far apart in longitude so that one is always illuminated when the other is in shadow. The separation of the two satellites must be greater than 17.4°. There are, however, two disadvantages:

— The change of satellite involves a reorientation of the antennas on the ground and hence an interruption of service unless two antennas, or one antenna with electronic pointing, are provided.
— The coverage is that which is common to the two satellites.

For certain types of satellite, such as satellites for direct television broadcasting, it is conceivable that service would not be provided, since eclipses always occur at night when customers are, presumably, asleep. They occur later at night when the satellite is further to the west of the region to be covered; a shift of 15° in the longitude of the satellite towards the west with respect to the longitude of the service region corresponds to an eclipse occurring at 01 h 00 true solar time of the service region: that is, around 02 h 00 or 03 h 00 in legal time.

2.2.5.6 Eclipses of the sun by the moon

In addition to eclipses due to the earth, the solar disc as seen by a geostationary satellite can be partially or totally obscured by the moon. Compared with those due to the earth, eclipses due to the moon are of irregular occurrence and extent. The number of eclipses per year due to the moon for a given orbital position varies from zero to four with a mean of two. Eclipses can occur twice in a period of 24 hours. The duration of an eclipse varies from several minutes to more than two hours with a mean of around 40 minutes (CCIR-Rep 802). Figure 2.33 shows an example of the occurrences, duration, and depth of eclipses of the sun by the moon for a satellite at 31° W over a period of 12 years. The figure shows that about 20 eclipses with a depth (that is, the percentage of the solar disc that is shadowed by the moon) larger than 40% occur in that period;

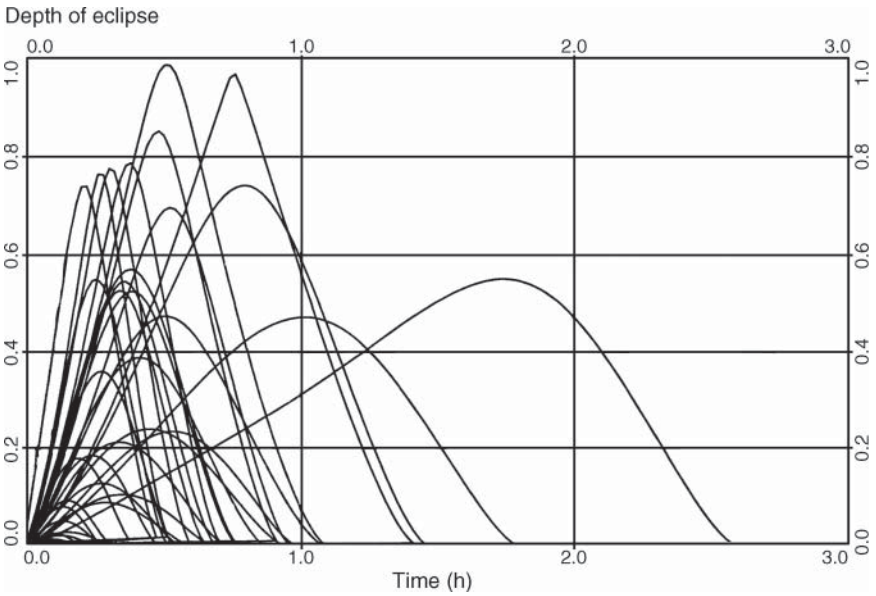


Figure 2.33 Eclipses of the sun by the moon. Depth of eclipse as a function of its duration for a geostationary satellite at longitude 31°W from January 1999 to December 2010.

among them, some last over one hour and one lasts over two hours. A few are over 90% depth. The satellite is liable to suffer from problems associated with excessive discharge of the batteries and a large fall in the temperature of certain parts if an eclipse of the sun due to the moon of significant duration and depth occurs just before or after an eclipse of the sun by the earth. This may occur as the occurrence of eclipses of the sun by the moon spreads over the year, and some may occur in the equinoctial season. The number and occurrence of these eclipses depend a lot on the orbital position of the satellite and also on the date of launch and mission duration. This investigation is part of the preliminary system mission analysis to identify the possible impact on the satellite.

2.2.5.7 Conjunction of the sun and the satellite

A conjunction of the sun and the satellite occurs when the axis of the antenna beam from a ground station pointing towards the satellite passes through the sun. This implies that the declination of the sun is equal to the angle that the radiating axis of the antenna makes with the equatorial plane. Whatever the position of the station on the surface of the earth, this angle has a maximum value of 8.7° (see Section 2.2.5.3). Sun–satellite conjunction thus occurs at times of the year that are close to the equinoxes, as follows:

— Before the spring equinox and after the autumn equinox for a station in the northern hemisphere.
— After the spring equinox and before the autumn equinox for a station in the southern hemisphere.

Date of conjunctions. Considering a nadir angle θ of 8.7° and an infinitely narrow beam width, the dates of sun–satellite conjunction for a station at the limit of visibility in the northern hemisphere are those corresponding to a declination of the sun equal to 8.7°. These dates were determined in Section 2.2.5.5 for the start of the eclipse season before the spring equinox and the end of the season after the autumn equinox; they are around 26 February and 16 October.

If the nadir angle is less than 8.7°, the dates approach the respective equinoxes. The date and time of maximum conjunction can be calculated from the angles of elevation E (obtained using Eq. (2.44a)) and azimuth A (Eq. (2.65)) under which the satellite is seen from the station. For greater accuracy, the oblateness of the earth is taken into account by replacing the equatorial radius R_E with R_C, the distance from the station to the centre of the ellipsoid (see Eq. (2.27b)), in the calculation of the station–satellite distance R from Eq. (2.41) and by replacing the geographic latitude l with the geocentric latitude l' (see Eq. (2.27a)) in Eq. (2.64) for calculation of the angle ϕ.

The coordinates δ (declination) and H (the hour angle of the satellite) are obtained from the elevation and azimuth by using the formulae for conversion from horizontal coordinates to hour coordinates, modified to take into account the non-astronomical definition of the azimuth angle.

This gives:

$$\sin\delta = \sin l' \sin E + \cos l' \cos E \cos A \tag{2.68a}$$

and

$$\cos H = (\cos l' \sin E - \sin l' \sin\delta \cos A)/\cos\delta \tag{2.68b}$$

Conjunction with the sun occurs on the day when the declination of the sun δ_{SUN} is equal to the declination δ of the satellite. The declination of the sun is related to its right ascension α_{SUN} by $\delta_{SUN} = \tan\varepsilon \sin\alpha_{SUN}$ (see Figure 2.6). Two values of right ascension of the sun correspond

to a given declination. One value in the vicinity of the spring equinox (before the equinox for a station in the northern hemisphere) corresponds to:

$$\alpha_{SUN} = \text{arc sin}[\tan \delta_{SUN} / \tan \varepsilon]$$

The other value in the vicinity of the autumn equinox is such that:

$$\alpha_{SUN} = 180° - \text{arc sin}[\tan \delta_{SUN} / \tan \varepsilon]$$

The relation between the date and the corresponding value of right ascension of the sun is given by Eqs. (2.28) and (2.31) given that Greenwich civil time, or universal time UT, is the mean solar time increased by 12 hours. Mean solar time differs from the true solar time and can be calculated using Eq. (2.31). True solar time is the hour angle of the sun that is related by Eq. (2.28) to sidereal time ST through the right ascension of the sun α_{SUN}. Finally, this gives for date t at 0 hour UT (α_{SUN} and ST in hours):

$$\alpha_{SUN} = ST + 12 - \Delta E \tag{2.69}$$

The sidereal time ST (in hours) for the date JD at 0 hours is obtained using Eq. (2.33). By simplifying Eq. (2.33), one can write:

$$ST = (1/3600)[24110.6 + 8640184.812866 \times T]$$

where T is the number of Julian centuries between the date JD at 0 hours and 1 January 2000 at 12 hours (adding or subtracting multiples of 24 hours so that $0 < ST < 24$ hours).

To obtain JD from α_{SUN}, the calculation is more difficult since the time equation ΔE depends on the date. It is necessary to proceed by iterations starting from an initial date obtained by consulting astronomical tables or by determining the approximate date when the sun has the declination δ_{SUN} in Figure 2.7.

Time of conjunction. To obtain the time of conjunction with the sun on day JD, the hour angle H of the satellite must be set equal to the hour angle of the sun at the earth station. The LST is obtained by adding the value of right ascension of the sun on day JD:

$$LST = H + \alpha_{SUN}$$

and the sidereal time on subtracting the longitude east of the station:

$$ST = LST - \lambda$$

The time SU of the conjunction measured in sidereal units of time is then obtained by subtracting the sidereal time at 0 hours on day JD:

$$SU = ST - ST(\text{JD at 0 h})$$

It is necessary to convert sidereal units of time SU into universal time UT (see Section 2.1.5.5):

$$UT = SU \times 0.9972696$$

The universal time UT (of Greenwich) of the maximum conjunction with the sun is thus obtained. The local time at the station will take into account the corresponding time zone and, if necessary, the date (summer or winter time).

Number of days of interference. Conjunction is defined by alignment of earth station, satellite, and sun. If the antenna beam is assumed to be infinitely narrow, conjunction corresponds to the only situation where the earth station suffers interference. If this beam has an equivalent aperture θ_i, interference will occur on several successive days around the initial date defined by the value of nadir angle θ such that the declination of the sun remains between $\beta - \theta_i/2$ and $\theta + \theta_i/2$. The declination of the sun around the equinoxes varies approximately $0:4°$ per day. The number N_i of consecutive days of satellite–sun interference thus has a value:

$$N_i = 2.5\theta_i \text{ days} \tag{2.70}$$

where θ_i is the equivalent aperture of the antenna beam in degrees. By way of example, if $\theta_i = 2°$, the interference occurs for five consecutive days: that is, two days before and two days after the nominal date.

Duration of the interference. The duration of interference with the sun is determined by noticing that the apparent daily movement of the sun around the earth has a value of 0.25° per minute. The duration Δt_i of the interference is thus:

$$\Delta t_i = 4\theta_i \text{ min} \tag{2.71}$$

Taking $\theta_i = 2°$, the duration of the interference is equal to eight minutes.

During this interference, the antenna noise temperature increases abruptly. The value of this increase and the method of determining the angular diameter θ_i of the zone of solar interference are discussed in Chapter 8.

2.3 PERTURBATIONS OF ORBITS

Movement of a satellite in its orbit is determined by the forces acting on the centre of mass. With the Keplerian hypotheses, only the attraction of a central, spherical, and homogeneous body defines a conservative field of forces (Eq. (2.3)). The trajectory obtained is a plane, fixed in space and characterised by a set of constant orbital parameters. These orbital parameters can be obtained from the position and velocity vectors of the satellite by a geometric transformation. In the case of a perturbed orbit, the orbital parameters are no longer constant but are a function of the date for which the transformation is applied. Extrapolation of the orbit could be made by numerical integration of the equation of motion after taking into account the various perturbations.

Perturbations of the orbit are the result of various forces that are exerted on the satellite other than the force of attraction of the central, spherical, and homogeneous body. These forces consist mainly of:

— The non-spherical components of terrestrial attraction
— The attraction of the sun and the moon
— Solar radiation pressure
— Aerodynamic drag
— Motor thrust

The first two contributions are gravitational forces from perturbing potentials. In contrast, the other forces do not depend on the mass of the satellite and are not conservative; they are due to exchanges of the amount of movement at the surface of the satellite and depend on the aspect and geometry of the satellite, and therefore provide the possibility of control.

2.3.1 The nature of perturbations

2.3.1.1 Asymmetry of the terrestrial potential

The earth is not a spherical homogeneous body. The terrestrial potential at a point in space depends not only on the distance r to the centre of mass but also on the latitude and longitude of the point concerned and the time. This is due to the irregularities of the rotation of the earth and the mass distribution (caused by oceanic and terrestrial tides: that is, movement of the surface of the oceans and the earth's crust under the effect of lunar attraction and internal geophysical phenomena). With the choice of a reference tied to the earth's crust, a simplified expansion of the static part (using mean coefficients) of the terrestrial potential is as follows:

$$u = (\mu/r)\left[1 - \sum_{n=2}^{\infty}(R_E/r)^n J_n P_n(\sin\ \varphi) + \sum_{n=2}^{\infty}\sum_{q=1}^{\infty}(R_E/r)^n J_{\mathrm{nq}} P_{\mathrm{nq}}(\sin\ \varphi)(\cos\ q(\lambda - \lambda_{\mathrm{nq}}))\right] \tag{2.72}$$

where:

μ = 3.986×10^{14} m^3 s^{-2}, the gravitational constant of the earth
r = distance of the point considered with respect to the centre of the earth
R_{E} = 6378.14 km, the mean terrestrial equatorial radius
φ, λ = the latitude and longitude of the point considered
J_{n} = zonal harmonics
J_{nq} = tesseral harmonics
P_{n} = Legendre polynomial of order n:

$$P_{\mathrm{n}}(x) = [1/(2^n n!)]d^n/dx^n[(x^2 - 1)^n]$$

P_{nq} = the associated Legendre function:

$$P_{nq}(x) = (1 - x^2)^{q/2}\mathrm{d}^q/\mathrm{dx}^q P_n(x)$$

The J_{n} and J_{nq} terms are constants that are characteristic of the distribution of the mass of the earth. The J_{n} terms are the zonal harmonics reflecting the potential dependence on the latitude. The J_2 term, due to the flattening of the earth (about 20 km) dominates all the other terms. The J_{nq} terms are the tesseral harmonics ($n \neq q$) reflecting the combined dependence on the latitude and longitude, or the sectorial harmonics ($n = q$) as a function of the longitude. The dominant term J_{22} is characteristic of the ellipticity of the equator (a difference of 150 m between the semi-minor and semi-major axes).

The values of the coefficients are given by various models (with many formulations for the expansion of the potential) such as those developed by the Goddard Space Flight Centre and the GRIM models of the Groupe de Recherche en Géodésie Spatiale and the Deutsches Geodätisches Forschungsinstitut. Some numerical values of the coefficients are as follows (GEM4 model):

$$J_2 = 1.0827 \times 10^{-3};\ J_{22} = 1.083 \times 10^{-6};\quad \lambda_{22} = -14.91^\circ$$

The order of magnitude of the coefficients J_{n} and J_{nq} for $n > 2$ is given by $10^{-5}/n^2$ [KAU-66]. The perturbing potential is given by:

$$U_p = U - \mu/r \tag{2.73}$$

For a geostationary satellite, the ratio R_E/r is small and the latitude φ is close to 0. To a first approximation, by limiting the expansion to order 2, this gives:

$$U \approx \mu/r[1 + (R_E/r)^2\{J_2/2 + 3J_{22} \cos 2(\lambda - \lambda_{22})\}] \tag{2.74}$$

with $\lambda_{22} = -14.91°$: that is, 15° longitude west.

2.3.1.2 *Attraction of the moon and the sun*

The moon and the sun each create a gravitational potential whose expression is of the form:

$$U_P = \mu_P\{1/\Delta - [(\mathbf{r}_P \cdot \mathbf{r})/|\mathbf{r}_P|^3]\} \text{ with } \Delta^2 = |\mathbf{r}_P - \mathbf{r}| \tag{2.75}$$

where $\mathbf{r}$ is the vector from the centre of the earth to the satellite, $\mathbf{r}_P$ is the vector from the centre of the earth to the perturbing body, and $\mu_P = GM_P$ (M_P = mass of the perturbing body) is the attraction constant of the perturbing body (the moon or sun). For the moon, $\mu_P = 4.8999 \times 10^{12}\ m^3\ s^{-2}$; for the sun, $\mu_P = 1.345 \times 10^{20}\ m^3\ s^{-2}$.

2.3.1.3 *Solar radiation pressure*

A surface element dS with normal $\mathbf{n}$ oriented in the direction of the sun and making an angle θ with the unit vector $\mathbf{u}$ directed towards the latter is subjected to the following pressure:

$$d\mathbf{F}/dS = -(W/c)[(1 + \rho)(\cos\theta)^2\mathbf{n} + (1 - \rho)(\cos\theta)\mathbf{n} \wedge (\mathbf{u} \wedge \mathbf{n})] \tag{2.76}$$

where ρ is the reflectivity of the surface (the ratio of the reflected and incident fluxes), W is the solar flux (power received per unit surface area), and c is the velocity of light ($W/c = 4.51 \times 10^{-6}\ N\,m^{-2}$ at 1 IAU).

If the element dS is totally reflecting ($\rho = 1$), the pressure is normal to the surface:

$$d\mathbf{F}/dS = (2W/c)(\cos\theta)^2\mathbf{n}$$

If the element dS is totally absorbent ($\rho = 0$), the radiation pressure divides into a normal component $(dF/dS)_N = (W/c)(\cos\theta)^2$ and a tangential component $(dF/dS)_T = -(W/c) \times (\cos\theta)^2 \sin\theta$.

A satellite of apparent surface S_a in the direction of the sun and reflectivity ρ equal to 0.5 (a typical value) is subjected to a perturbing force:

$$F_P = -1.5(W/c)S_a \quad (N) \tag{2.77}$$

If the satellite is of mass m, the acceleration due to radiation pressure is:

$$\Gamma = 6.77 \times 10^{-6} S_a/m \quad (m/s^2)$$

The solar panels constitute practically the whole of the apparent surface of the satellite. With communications satellites of low power (1 kW), the solar panels are not extensive and the ratio S_a/m is on the order of $2 \times 10^{-2}\ m^2\ kg^{-1}$. This was the case, for example, for Intelsat V, for which $S_a = 18\ m^2$ and $m = 1000$ kg; thus $S_a/m = 1.8 \times 10^{-2}\ m^2\ kg^{-1}$. With these satellites, the acceleration due to radiation pressure is on the order of $10^{-7}\ m\,s^{-2}$, and its effect is limited.

For satellites of high electrical power on which very extensive solar panels are mounted (a surface of 100 m^2 for a mass of 1000 kg, for example), the ratio S_a/m is on the order of 10^{-1};

the acceleration due to radiation pressure must then be taken into account in calculating perturbations.

The main effect of solar radiation pressure is to modify the eccentricity of the orbit that evolves with a period of one year (Section 2.3.3.5). For satellites in a low orbit, it is also necessary to take into account the radiation pressure of the solar flux reradiated from the surface of the earth (albedo) whose effect can be significant (20%) with respect to that of the direct solar flux.

2.3.1.4 Aerodynamic drag

In spite of the low value of atmospheric density encountered at the altitudes of satellites, their high velocity means that perturbations due to aerodynamic drag are very significant at low altitude (200–400 km) and are negligible only above about 3000 km. The aerodynamic force is exerted on the satellite in the opposite direction to its velocity and is of the form:

$$F_{AD} = -0.5\rho_A C_D A_e V^2 \tag{2.78}$$

where ρ_A is the density of the atmosphere, C_D is the coefficient of aerodynamic drag, A_e is the equivalent surface area of the satellite perpendicular to the velocity, and V is the velocity of the satellite with respect to the atmosphere.

The density of the atmosphere depends on the altitude (the variation is exponential), the latitude, the time, solar activity, and so on. Various models have been developed (e.g. [JAC-77; HED-87]). The coefficient of aerodynamic drag is a function of the form and nature of the surface. The velocity with respect to the atmosphere differs from the velocity of the satellite in an inertial reference since the atmosphere has some velocity as a consequence of dragging by terrestrial rotation and the phenomena of wind.

If the satellite is of mass m, the acceleration due to aerodynamic drag is:

$$\Gamma_{AD} = -0.5\rho_A C_D V^2 A_e/m \quad (\mathrm{m/s^2}) \tag{2.79}$$

The main effect of atmospheric friction is a decrease of the semi-major axis of the orbit due to a reduction of the energy of the orbit. A circular orbit remains as such, but its altitude reduces, whereas the velocity of the satellite increases. For an elliptical orbit, the braking occurs principally at the perigee. The altitude of the apogee decreases, the altitude of the perigee remains almost constant, the eccentricity decreases, and the orbit tends to become circular. By way of example, for an elliptical orbit with perigee altitude 200 km and apogee altitude 36 000 km (a transfer orbit; see Chapter 11), the reduction of altitude of the apogee is around 5 km on each orbit.

2.3.2 The effect of perturbations; orbit perturbation

2.3.2.1 Osculatory parameters

The actual movement of the satellite is obtained from the fact that the satellite is in equilibrium between the inertial force $m\,\mathrm{d}^2r/\mathrm{d}t^2$ and the various forces that are exerted on it. The latter include:

— The forces of attraction due to the potential of a spherical and homogeneous earth
— Forces due to the various perturbing potentials
— Non-conservative perturbing forces

Hence:

$$m\,\mathrm{d}^2\mathbf{r}/\mathrm{d}t^2 = m\mu(\mathbf{r}/r^3) + (m/r)\mathbf{r}\,\mathrm{d}U_p/\mathrm{d}r + \mathbf{f}_p \tag{2.80}$$

It is thus possible to determine the position and velocity of the satellite at each instant by integration in a geocentric reference frame. Using a geometric transformation defined by the Keplerian hypotheses, it is possible, on given data, to obtain the six orbital parameters that are characteristic of the movement of a satellite. Unlike the Keplerian orbit, where the parameters are constants, the parameters are functions of time for a perturbed orbit.

These parameters, determined for the current date t, are called *osculatory parameters*. The osculatory elements for the date t are the orbital elements of Keplerian movement that would describe the satellite motion if the perturbations were cancelled from the date t. The trajectory defined in this way is called the *osculatory ellipse*. This nomenclature is inappropriate, as the curvature of the osculatory ellipse, although tangential to the trajectory on date t, is *not* the actual curvature.

Using the osculatory parameters is a convenient way of representing the satellite orbit within a limited time frame when subjected to perturbations.

2.3.2.2 *Variation of the orbital parameters*

The variation of the orbital parameters ($\mathrm{d}a/\mathrm{d}t$, $\mathrm{d}e/\mathrm{d}t$, $\mathrm{d}i/\mathrm{d}t$, $\mathrm{d}\Omega/\mathrm{d}t$, $\mathrm{d}\Omega/\mathrm{d}t$, and $\mathrm{d}M/\mathrm{d}t$) is obtained from the components of the perturbing acceleration in an orthogonal coordinate system centred on the satellite by means of Gauss's equations.

If the perturbing accelerating field is due to a potential (only forces of gravitational origin are present), the system of differential equations can be put in a particular form as a function of the partial derivatives of the perturbing potentials with respect to the orbital parameters; these are Lagrange's equations. Integration of Lagrange's equations [KAU-66] produces the orbital parameters as the sum of a mean parameter, periodic terms with short and long periods (with respect to the period of the orbit), and, for some, a secular term (that is an increasing function of time).

2.3.2.3 *Long-term progression*

Perturbations of the terrestrial potential cause long-term progression that affects the ω (argument of the perigee), Ω (right ascension), and M (mean anomaly) parameters. These secular terms are a function of the even zonal harmonics, particularly J_2. This gives:

$$\mathrm{d}\omega/\mathrm{d}t = (3/4)n_0 A J_2[5\cos^2 i - 1] \tag{2.81a}$$

$$d\Omega/\mathrm{d}t = -(3/2)n_0 A J_2 \cos i \tag{2.81b}$$

$$\mathrm{d}M/\mathrm{d}t = n_0[1 + 3/4A(1-e)^{1/2}J_2\,(3\cos^2 i - 1)] \tag{2.81c}$$

where:

$$A = R_\mathrm{E}^2/a^2(1-e^2)^2$$

R_E = radius of the earth
e, a = eccentricity and semi-major axis of the satellite orbit
i = inclination of the orbit
n_0 = mean movement of the satellite = $2\pi/\mathrm{T} = \sqrt{(\mu/a^3)}$

For example, for an elliptical orbit of inclination 7° whose perigee altitude is 200 km and apogee altitude is 36 000 km (a transfer orbit; see Chapter 11), the drift of the argument of the perigee $d\omega/dt$ has a value of 0.817°/day. For a circular orbit of altitude 290 km and inclination 28° (the parking orbit of the Space Transportation Service [STS]; see Chapter 11), the drift of the RAAN (the nodal regression) has a value $d\Omega/dt = 25°$/day. This nodal regression is zero for polar orbits ($i = 90°$).

Selection of the value of certain orbital parameters enables the drift of another parameter to be fixed at a particular value. Hence, to make the drift of the argument of the perigee $d\omega/dt$ zero, it is acceptable to choose the value of inclination that makes the term $5(\cos^2 i) - 1$ equal to zero in (2.81a): that is, $i = 63.4°$. There will no longer be a rotation of the perigee–apogee line in the plane of the orbit, and the apogee will remain permanently above the same hemisphere. This is the reason that led to the choice of 63.4° for the inclination of the Molniya and Tundra orbits (see Section 2.2).

By choosing a pair of particular values for a and i, it is possible to obtain an orbit for which the RAAN varies each day by a quantity equal to the mean variation of the right ascension of the sun: that is, $d\Omega/dt = 360°/365\,\text{day} = 0.9856°$/day in (2.81b). For a circular orbit, the condition can be written as $-6530a^{-7/2} \cos i = 0.986$, where a is expressed in 10^3 km.

The angle between the line of nodes of the orbits and the mean direction of the sun obtained in this way remains constant throughout the year. The conditions for illumination are thus always identical from one orbit to another, with fluctuations due to the variation of the declination of the sun and the time equation. The satellite is said to be *sun-synchronous*. If the orbit is also phased – that is, the period is a submultiple of the sidereal day or an integer number of days – the satellite passes over the same points again with a period equal to the number of days concerned. These orbits are thus particularly well suited to earth observation missions. For example, the orbit of the Spot satellites has an altitude of 822 km ($a = 7200$ km), an inclination of 98 : 7°, and a period of 101.3 minutes; it returns over the same point after 26 days.

2.3.3 Perturbations of the orbit of geostationary satellites

The orbit of geostationary satellites has been defined in Section 2.2.5 as a non-retrograde circular orbit ($e = 0$) in the plane of the equator ($i = 0$) whose period of revolution is equal to the period of rotation of the earth ($T = 86\,164$: 1 second); this gives rise to a semi-major axis a_k, calculated using the Keplerian hypotheses, of 42 164.2 km.

If a satellite is placed in an orbit defined in this way, it is observed that, due to the effect of perturbations, the parameters of the orbit do not remain constant as Kepler's equations would predict. The apparent movement of the satellite with respect to a rotation coordinate system related to the earth is as follows:

— Displacement in the east–west plane with respect to the nominal position defined by the longitude of the satellite station (predicted to be fixed, since the velocity of rotation of the satellite has been made the same as that of the earth). A modification of the radial distance is associated with this displacement.
— Displacement in the north–south direction with respect to the equatorial plane.

Examination of the orbit parameters after several weeks shows that the values of the semi-major axis, eccentricity, and inclination of the orbit are no longer equal to the initial values. The satellite is no longer perfectly geostationary.

2.3.3.1 Modified orbit parameters

Conventional parameters are not well suited to characterisation of the orbit of a quasigeostationary satellite; when the inclination tends to zero, the position of the ascending node becomes indeterminate. The same applies to the position of the perigee when the eccentricity tends to zero. It is thus logical to characterise the simultaneous progression of i and Ω and the simultaneous progression of e and $(\Omega + \Omega)$.

This is obtained by introducing:

— The inclination vector **i** with components:

$$i_x = i \cos \Omega$$

$$i_y = i \sin \Omega$$

— The eccentricity vector **e** with components:

$$e_x = e \cos (\omega + \Omega)$$

$$e_y = e \sin (\omega + \Omega)$$

The inclination vector is represented in Figure 2.34 by a vector along the line of nodes directed towards the ascending node and of modulus equal to the inclination. The eccentricity vector is represented in Figure 2.34 by a vector along the line of apsides directed towards the perigee and of modulus equal to the eccentricity.

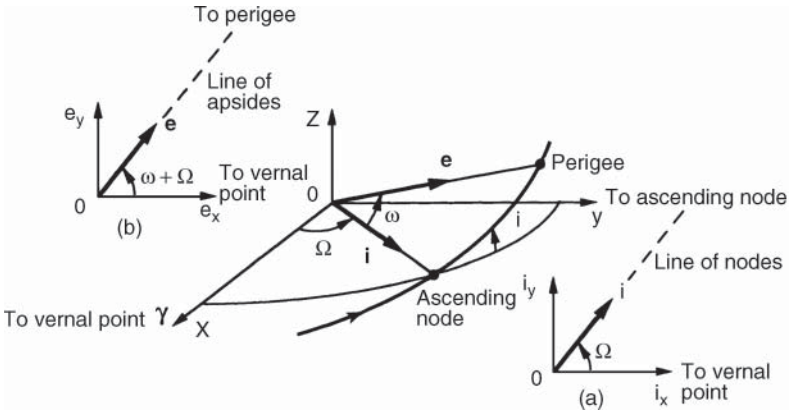


Figure 2.34 Inclination and eccentricity vectors for characterising the orbit of a quasigeostationary satellite.

The angle $(\Omega + \omega)$ is the sum of two angles in planes that are, in principle, different but close since the inclination remains small. Other definitions of the inclination vector can be used, such as, for example, that of a vector along the axis of the angular momentum of the orbit of modulus equal to the inclination.

Furthermore, rather than using the mean anomaly, the position of the satellite in the orbit is characterised by the mean longitude λ_m or the true longitude λ_v, given by:

$$\lambda_m = \omega + \Omega + M - \mathrm{ST} \tag{2.82a}$$

$$\lambda_v = \omega + \Omega + v - \mathrm{ST} \tag{2.82b}$$

where M and v are the mean and true anomalies respectively, and ST is the sidereal time of the Greenwich meridian.

As a consequence of the rotation of the earth, the sidereal time of the Greenwich meridian increases by $(360° + 0.9856°)/24 = 15.041\,69°$ per hour (see Eq. (2.30)). As the eccentricity is small, the relation between the mean anomaly and the true anomaly is given by:

$$\lambda_v = \lambda_m + 2e(\sin M) \tag{2.83}$$

The true longitude of the satellite thus oscillates about the mean longitude with an amplitude of $2e$ in the course of the day (see Section 2.2.2).

In conclusion, the orbital parameters adapted to the quasigeostationary satellite are:

$$a, e_x, e_y, i_x, i_y \quad \text{and} \quad \lambda_m.$$

2.3.3.2 *Semi-major axis of a geosynchronous circular orbit*

The semi-major axis of the perturbed geosynchronous circular orbit is different from the semi-major axis a_K calculated using the Keplerian hypotheses. Lagrange's equation leads to an expression for the drift of the mean longitude:

$$d\lambda_m/dt = -(2/na)(dU_p/dr)_{r=a} + n - \Omega_E \tag{2.84}$$

where $\Omega_E = 4.178 \times 10^{-3}°/s$ is the angular velocity of rotation of the earth.

For the satellite to be geostationary, it is necessary for the drift of the mean longitude $d\lambda_m/dt$ to be zero. The value of the semi-major axis a_s corresponding to a geosynchronous orbit is then obtained from:

$$\begin{aligned} a_s &= a_k + 2J_2 a_k (R_E/a_k)^2 + \cdots \\ &= a_k + 2.09 \text{ km} = 42166.3 \text{ km} \end{aligned}$$

where a_k is the semi-major axis of the Keplerian orbit.

Cancellation of a long-term drift due to lunar–solar attraction leads to a further modification of the semi-major axis that has this final value:

$$a_s = 42165.8 \text{ km}$$

Equation (2.84) enables the derivative of the mean longitude $d\lambda_m/dt$ to be related to the variation Δa of the semi-major axis with respect to its value a_s corresponding to the synchronous orbit. This gives:

$$d\lambda_m/dt = -(3/2)(n_s/a_s)(\Delta a) = k_\lambda (a - a_s) \tag{2.85}$$

where n_s equals the angular velocity of the earth $\Omega_E = 4.178 \times 10^{-3}°/\text{second}$, $a_s = 42\,165.8$ km, and therefore $k_\lambda = -0.0128°/\text{day km}$. Note that $a > a_s$, the mean longitude of the satellite decreases with time ($d\lambda_m/dt$ is negative as k_λ is negative), which makes sense: the velocity of the satellite is reduced as the radius of the orbit augments; and then, the satellite angular velocity being less than that of the earth, the longitude of the satellite decreases.

2.3.3.3 *Progression of the longitude of the satellite*

For a quasigeostationary satellite, the terrestrial perturbing potential is approximated by Eq. (2.74):

$$U_p = \mu/r[(R_E/r)^2\{J_2/2 + 3J_{22}\cos 2(\lambda - \lambda_{22})\}]$$

This potential creates a tangential acceleration Γ_T such that:

$$\Gamma_T = -(1/r)\mathrm{d}U_p/\mathrm{d}\lambda = (\mu/r^2)(R_E/r)^2 6J_{22}\sin 2(\lambda - \lambda_{22}) \tag{2.86}$$

This acceleration causes a variation of the velocity V_{SL} of the satellite in the orbit:

$$\Gamma_T = \mathrm{d}V_{SL}/\mathrm{d}t = \mathrm{d}t = \mathrm{d}/\mathrm{d}t[r\omega_{SL}] = [\mathrm{d}r/\mathrm{d}t]\omega_{SL} + r(\mathrm{d}\omega_{SL}/\mathrm{d}t) \tag{2.87}$$

where ω_{SL} is the angular velocity of the satellite.

As the orbit is quasi-circular, it is permissible to consider $r \approx a_s$ the semi-major axis of the geosynchronous circular orbit and $\omega_{SL} \approx n_S = \sqrt{(\mu/a_s^3)}$, which leads to:

$$\mathrm{d}\omega_{SL}/\omega_{SL} \approx -(3/2)(\mathrm{d}r/r) \quad \text{for } r = a_s \tag{2.88}$$

Combining Eqs. (2.87) and (2.88), this gives:

$$\mathrm{d}r/\mathrm{d}t(\text{for } r = a_s) \approx -(2/\omega_{SL})\Gamma_T \tag{2.89}$$

The longitudinal acceleration experienced by the satellite can thus be written:

$$\mathrm{d}^2\lambda/\mathrm{d}t^2 = \mathrm{d}\omega_{SL}/\mathrm{d}t \approx (3/a_s)\Gamma_T \approx D\sin 2(\lambda - \lambda_{22}) \tag{2.90}$$

where $D = 18n_S^2(R_E/a_S)^2 J_{22} = 3\times 10^{-5}\mathrm{rad/(day)^2} = 4\times 10^{-15}\mathrm{rad/s^2}$.

The longitudinal acceleration thus depends on the longitude of the satellite. This acceleration varies sinusoidally with respect to longitude $\lambda_{22} = -14.91°$ and is zero for $\lambda = \lambda_{22} + k\pi/2$; in this way, four equilibrium points are defined. Two of these points are points of stable equilibrium (that is, if the satellite is displaced from the equilibrium position, it tends to return to it); the other two are points of unstable equilibrium.

By putting $\Lambda = \lambda - \lambda_{22} \pm 90°$, the longitude of the satellite with respect to the closest point of stable equilibrium, the movement of the satellite about the point of equilibrium is governed by the equation

$$\mathrm{d}^2\Lambda/\mathrm{d}t^2 = -D\sin 2\Lambda$$

The longitude drift $\mathrm{d}\Lambda/\mathrm{d}t$ as a function of longitude with respect to the point of stable equilibrium is thus of the form

$$(d\Lambda/\mathrm{d}t)^2 - D\cos 2\Lambda = \text{constant} \tag{2.91}$$

The curves in Figure 2.35 show the variation of drift $\mathrm{d}\Lambda/\mathrm{d}t$ as a function of longitude Λ about a stable equilibrium point. The figures in parentheses give the period of the oscillatory movement of the satellite with respect to the point of stable equilibrium; it is at least two years. It can also be observed that, for excessive values of initial drift at the point of stable equilibrium, the natural acceleration will not cancel this drift before the satellite arrives in the vicinity of the adjoining point of unstable equilibrium. The satellite thus overshoots the point of unstable equilibrium and is attracted to the next stable equilibrium point, where the same process is repeated. The drift is never cancelled, and the satellite thus rotates perpetually with respect to the earth.

The results presented here are obtained by neglecting terms with an order greater than 2 in the expansion of the perturbing potential. Figure 2.36 shows the actual longitudinal acceleration as a function of the longitude of the satellite station (CCIR-Rep 843). The longitudes of the positions of stable equilibrium are approximately 102° longitude west and 76° longitude east; those of the two positions of unstable equilibrium are 11° longitude west and 164° longitude east.

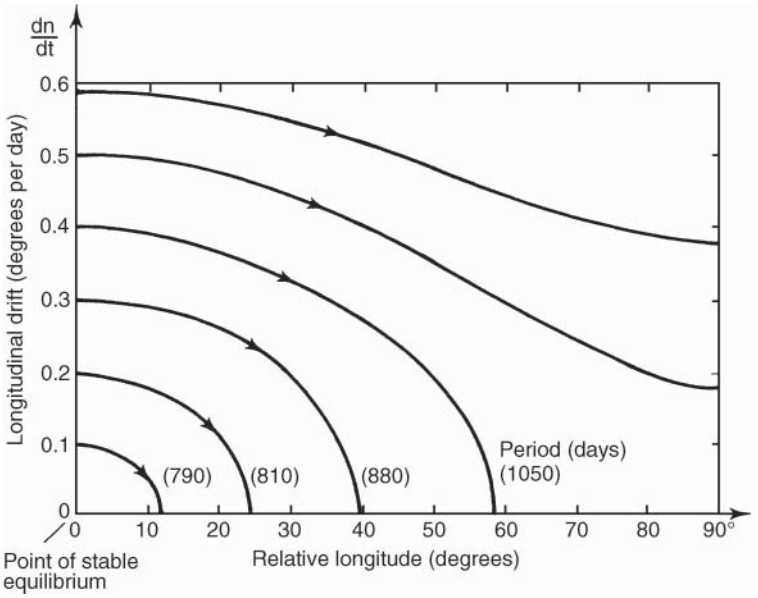


Figure 2.35 Evolution of the longitude drift as a function of the longitude with respect to a point of stable equilibrium.

2.3.3.4 Progression of the inclination

The effect of the attraction of the moon and the sun can be seen in Figure 2.37, where the O_y axis is in the equatorial plane perpendicular to the direction of the vernal point (the O_x axis is not shown and is towards the front of the figure), and the Oz axis is the polar axis (see Figure 2.6). At the summer solstice, the sun (in the plane of the ecliptic) is above the equatorial plane. The plane of the moon's orbit makes an angle of 5.14° with the plane of the ecliptic. The track of the moon's orbit in Figure 2.37 is within the region defined by two lines making an angle of ±5.14° with the ecliptic; the track is a function of the value of the RAAN of the lunar orbit, which varies through 360° in 18.6 years. On the date considered, the sun and moon are assumed to be on the right of the figure (which corresponds to a new moon on the earth).

When the satellite is on the right of the figure, it is more strongly attracted by the sun and the moon than when it is on the left, since the distance is less. The earth–satellite system behaves as if there were a (net) perturbing force δF acting in one direction on half of the orbit and another one acting in the opposite direction on the other half, as indicated in Figure 2.37. The same result is obtained when the moon is in the left part of the figure (full moon) after a lunar half period (13 days). The direction and magnitude of the perturbing force remain the same (except for variation of the earth–sun distance) if the sun is below the plane of the ecliptic on the left (winter solstice) for both positions of the moon. When the moon or the sun is in the equatorial plane, the component of the perturbing force normal to the plane of the orbit caused by the body concerned is zero.

The component of the perturbing force in the plane of the orbit affects the semi-major axis and eccentricity of the orbit. The long-term effect is cancelled by adjusting the value of the semi-major axis a_s to 42 165 : 8 km (see Section 2.3.3.2). The component of the force perpendicular to the plane of the orbit affects the inclination vector of the orbit. The effect of the sun is maximum at the

Figure 2.36 Longitudinal acceleration as a function of station longitude.

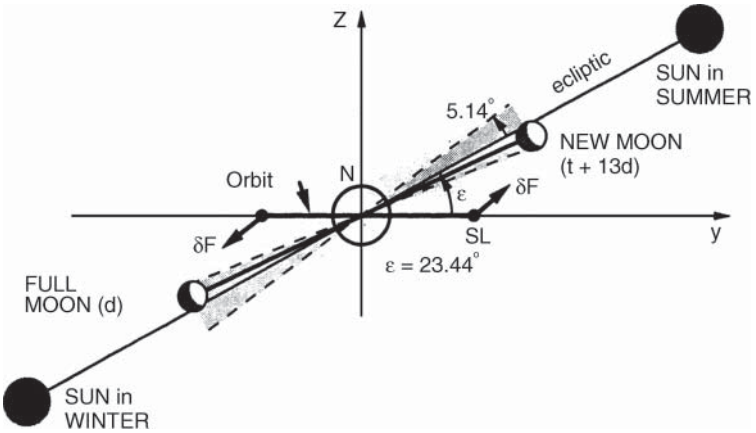


Figure 2.37 Attraction of the moon and sun on the orbit of geostationary satellites.

solstices and zero at the equinoxes; this leads to a mean drift of the inclination vector of 0.27° per year. The component of the perturbation normal to the plane of the orbit due to the moon is maximum twice per lunar period and passes through zero between. The effect leads to a mean drift of the inclination vector between 0.48°/year and 0.68°/year as a function of the value of the RAAN of the lunar orbit within its period of 18.6 years.

The combined effects of lunar–solar attraction on the inclination vector of the orbit of a quasi-geostationary satellite show the following principal effects:

— An oscillation of period 13.66 days and amplitude 0.0035°
— An oscillation of period 182.65 days and amplitude 0.023°
— A long-term progression

The components of the long-term progression are given by:

$$di_x/dt = H = (-3.6 \sin \Omega_M) \times 10^{-4} \text{ degrees/day}$$

$$di_y/dt = K = (23.4 + 2.7 \sin \Omega_M) \times 10^{-4} \text{ degrees/day}$$

where Ω_M is the right ascension of the ascending node of the lunar orbit for the period concerned and is given by:

$$\Omega_M(\text{degrees}) = 12.111 - 0.052954\ T\ (T = \text{days since} 1/1/1950)$$

Figure 2.38 shows the long-term progression of the inclination vector on a given date. The direction of drift Ω_D and the value of the derivative $\Delta i = \Delta t$ between the initial date t_0 and date t are such that:

$$\cos\ \Omega_D = H/\sqrt{(H^2 + K^2)}\ \ \Delta i/\Delta t = \sqrt{(H^2 + K^2)} \tag{2.92}$$

As a function of the epoch within the 18.6-year period of Ω_M:

— Ω_D varies from 81.8 to 98.9 degrees.
— $\Delta i/\Delta t$ varies from 0.75 to 0.95°/year.

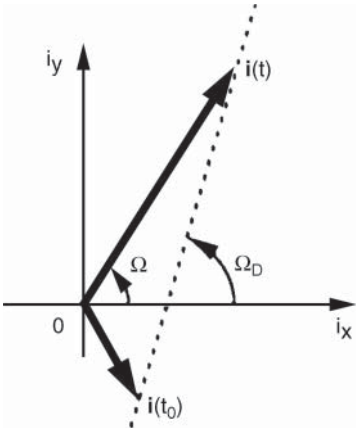


Figure 2.38 Long-term progression of the inclination vector.

The zonal terms of terrestrial potential also affect the progression of the inclination vector and cause a recession of the RAAN Ω equal to 4.9°/year.

The various contributions cause the extremity of the inclination vector on average to describe a circle in 54 years about a point with coordinates $i_x = -7.4°$ and $i_y = 0°$. This point constitutes a point of stable equilibrium for the long-term drift of the plane of the orbit corresponding to an inclination i equal to 7.4° and a RAAN Ω equal to 0°.

2.3.3.5 *Progression of the eccentricity*

Solar radiation pressure creates a force that acts in the direction of the velocity of the satellite on one half of the orbit and in the opposite direction on the other half. In this way, a circular orbit tends to become elliptical (Figure 2.39a). The apsidal line of the orbit is perpendicular to the direction of the sun.

The ellipticity of the orbit does not increase permanently. With the movement of the earth about the sun, the ellipse deforms continuously, and the eccentricity may remain limited to a maximum value.

The progression of the eccentricity and the argument of the perigee is represented by that of the eccentricity vector obtained from Lagrange's equations involving e, ω, and Ω to a first order in e

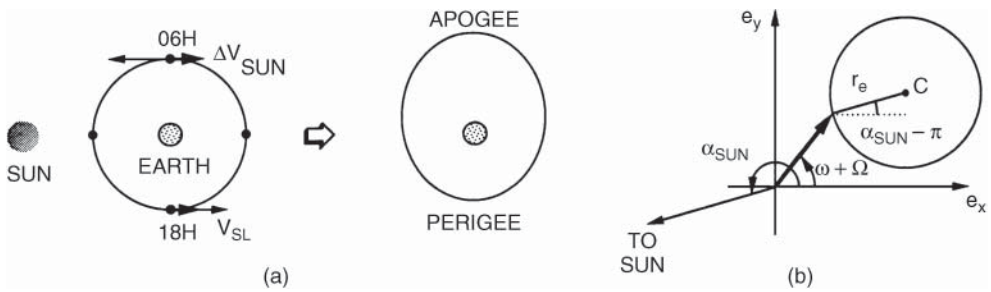


Figure 2.39 Effect of solar radiation pressure on the eccentricity of the orbit: (a) deformation of the orbit; (b) circle of natural eccentricity.

and i by considering that the perturbing acceleration derives from a pseudopotential. Assuming that the apparent orbit of the sun is circular and equatorial and neglecting short period terms (one-day period), calculation leads to:

$$de_x/dt = -(3/2)(C/n_s a_s)\ \sin\alpha_{SUN}$$
$$de_Y/dt = (3/2)(C/n_s a_s)\ \cos\alpha_{SUN}$$

where:

α_{SUN} = right ascension of the sun
n_s = mean movement of a geosynchronous satellite (rad/s) equal to the angular velocity of the earth $\Omega_E = 4{:}178 \times 10^{-3}°/s = 7{:}292 \times 10^{-5}°/h$
a_s = semi-major axis of the geosynchronous orbit = $42\,165.8 \times 10^3$ m, $C = (1+\rho)(S_a/m)(W/c)$, with
S_a = apparent surface area of the satellite in the direction of the sun (m^2), m = mass of the satellite (kg)
ρ = coefficient of reflectivity $\approx 0{:}5$
W/c = $4.51 \times 10^{-6}\,N\,m^{-2}$

Assuming C to be constant, the extremity of the eccentricity vector thus describes, in one year, a circle of radius r_e whose centre has coordinates (Figure 2.39b):

$$C_X = e_x(t_0) - r_e\ \cos\ \alpha_{SUN}(t_0)$$
$$C_Y = e_y(t_0) - r_e\ \sin\ \alpha_{SUN}(t_0)$$

This circle is called the *circle of natural eccentricity*. Its radius has the value

$$r_e = (3/2)C/n_s a_s \Omega_{SUN}$$

with:

$$\Omega_{SUN} = \text{mean angular velocity of the sun} = d\alpha_{SUN}/dt = 0.9856°/\text{day}$$

Hence, $r_e = 1.105 \times 10^{-2}(1+\rho)(S_a/m)$, with $\rho \approx 0.5$, S_a in m^2, and m in kg.

The radius of the circle of natural eccentricity is on the order of 5×10^{-4} for a satellite of 2 kW $S_a = 30\,m^2$ and 1000 kg in orbit. The eccentricity vector is such that the radius vector from the centre of the circle of natural eccentricity to the extremity of the eccentricity vector is directed towards the sun.

Finally, lunar–solar attraction also causes a perturbation with a period of about one month and amplitude on the order of $3{:}5 \times 10^{-5}$ that is superimposed on the progression of the eccentricity vector under the effect of radiation pressure.

2.3.4 Orbit corrections: station keeping of geostationary satellites

As a consequence of perturbations, the orbit parameters of geostationary satellites differ from the nominal parameters. The orbit is characterised by an inclination i, an eccentricity e, and a longitude drift $d\lambda/dt$, which are small but not zero. The effect of these parameters on the position of the satellite is first analysed in order to determine the station-keeping requirements. Correction procedures are then presented.

2.3.4.1 *Position and velocity of the satellite*

In a geocentric rotating reference, the spherical coordinates of the satellite are the radius r, the declination or latitude φ, and the longitude λ. Since the eccentricity e and the inclination i are small, these coordinates are related to the orbital parameters by:

$$\begin{aligned} r &= a_s + \Delta a - a_s e \cos \upsilon = a_s + \Delta a - a_s e \cos(\alpha_{SL} - (\omega + \Omega)) \\ &= a_s + \Delta a - a_s(e_x \cos \alpha_{SL} + e_y \sin \alpha_{SL}) \end{aligned} \tag{2.93a}$$

with Δa, the difference between the actual half axis and the synchronous half axis, equal to $-(2/3)(a_s/n_s)\mathrm{d}\lambda_m/\mathrm{d}t$ (cf. Eq. (2.85)) and α_{SL}, the right ascension of the satellite, equal to $\upsilon + \Omega + \Omega$ since i is small:

$$\begin{aligned} \lambda &= \lambda_m + 2e \sin M = \lambda_m + 2e \sin[\alpha_{SL} - (\omega + \Omega)] \\ \lambda &= \lambda_m + 2e_x \sin \alpha_{SL} - 2e_y \cos \alpha_{SL} \end{aligned} \tag{2.93b}$$

$$\begin{aligned} \varphi &= \text{arc sin}[\sin(\omega + \upsilon) \sin i] \\ &\approx i \sin(\omega + \upsilon) = i \sin(\alpha_{SL} - \Omega) \\ &= i_x \sin \alpha_{SL} - i_y \cos \alpha_{SL} \end{aligned} \tag{2.93c}$$

The velocity of the satellite in a geocentric inertial reference can be resolved into the components V_N perpendicular to the orbital plane towards the north, V_R in the earth–satellite direction, and V_T perpendicular to the radius vector in the plane of the orbit in the direction of the velocity.

Neglecting the variation of the satellite angular velocity with small inclination and eccentricity with respect to the angular velocity of rotation of the earth, Ω_E, the derivative $\mathrm{d}\alpha_{SL}/\mathrm{d}t$ is nearly equal to Ω_E. This gives, replacing $a_S\ \mathrm{d}\alpha_{SL}/\mathrm{d}t$ with $V_S = a_S\ \Omega_E$, the synchronous satellite velocity ($V_S = 3075\ \mathrm{m\ s^{-1}}$):

$$\begin{aligned} V_R &= a_s \mathrm{d}r/dt \\ &= -a_s(-e_x \sin \alpha_{SL} d\alpha_{SL}/dt + e_y \cos \alpha_{SL} d\alpha_{SL}/t) \\ &= V_S(e_x \sin \alpha_{SL} - e_y \cos \alpha_{SL}) \end{aligned} \tag{2.94a}$$

$$\begin{aligned} V_T &= a_s(\cos i\ \mathrm{d}\lambda/\mathrm{d}t) + a_s\Omega_E \\ &= a_s(\mathrm{d}\lambda/\mathrm{d}t) + a_s\Omega_E \\ &= a_s(\mathrm{d}\lambda_m/\mathrm{d}t) + V_S[1 + 2(e_x \cos \alpha_{SL} + e_y \sin \alpha_{SL})] \end{aligned} \tag{2.94b}$$

$$\begin{aligned} V_N &= a_s\ d\varphi/dt \\ &= V_S(i_x \cos \alpha_{SL} + i_y \sin \alpha_{SL}) \end{aligned} \tag{2.94c}$$

2.3.4.2 *Effect of non-zero eccentricity and inclination*

The non-zero eccentricity leads to an oscillation of the longitude of the satellite about the mean longitude of its station. This is illustrated in Figure 2.40a, where the successive positions of two satellites are represented; one is on a circular orbit of period one sidereal day, and the other is on an elliptical orbit of the same period. The difference in longitude $\Delta\lambda$ is obtained from Eq. (2.83)

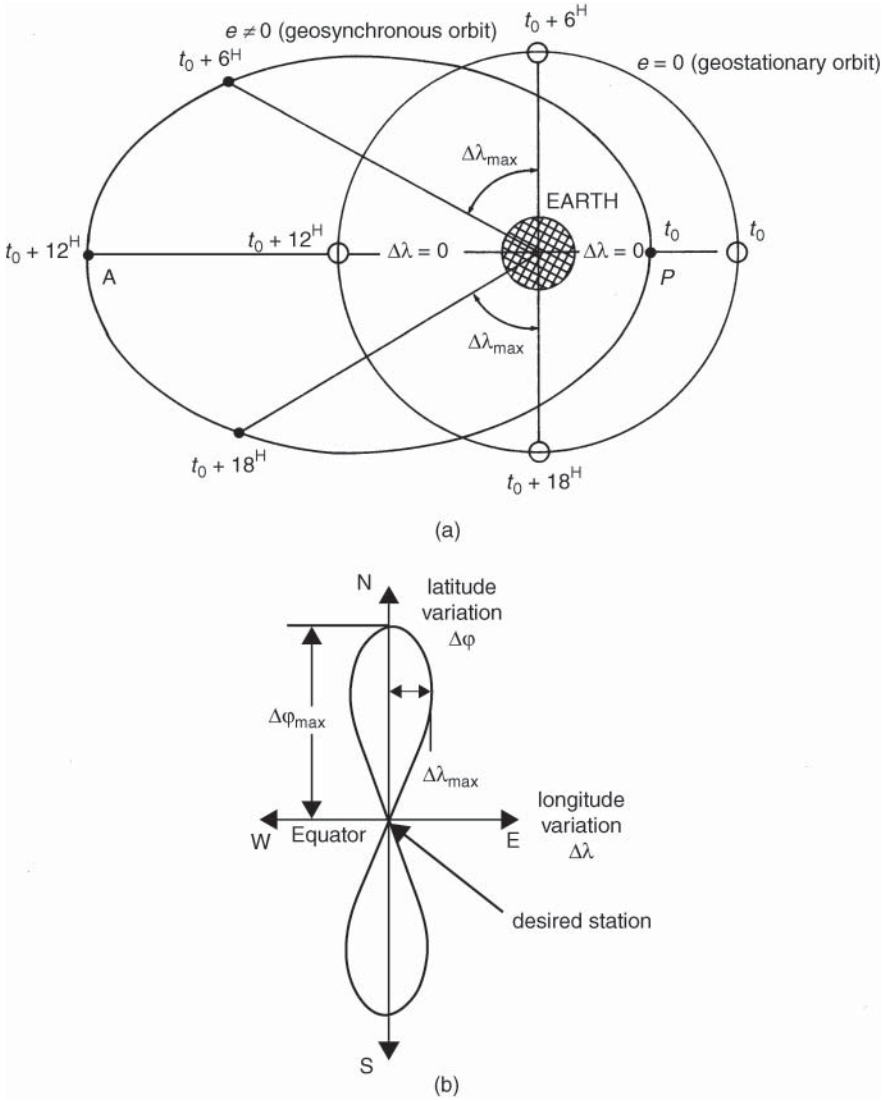


Figure 2.40 The effect of (a) non-zero eccentricity and (b) non-zero inclination.

and has a value $\Delta\lambda = \lambda_v - \lambda_m = 2e \sin M$. The maximum difference in longitude $\Delta\lambda_{max}$ thus has a value of $2e$ radians: that is, $2\pi e/180 = 114e$ degrees (see Section 2.2.2).

The non-zero inclination causes an apparent daily movement, as shown in Figure 2.40b, of the satellite with respect to the equator and the longitude of the station in the form of a figure eight (see Section 2.2.3). The amplitude of the latitude variation $\Delta\varphi_{max}$ is equal to the value of the inclination i. The maximum longitude variation $\Delta\lambda_{max}$ has a value of $4:36 \times 10^{-3} i^2$ (degrees), and the related latitude φ_m is 0.707 (see Eq. (2.60)). This maximum longitude shift is reached at the end of time t such that $(2\pi t/T) = 1/[\sqrt{2} \cos (i/2)]$: for i small, $\Delta\lambda_{max}$ is negligible (e.g. i = 1°,

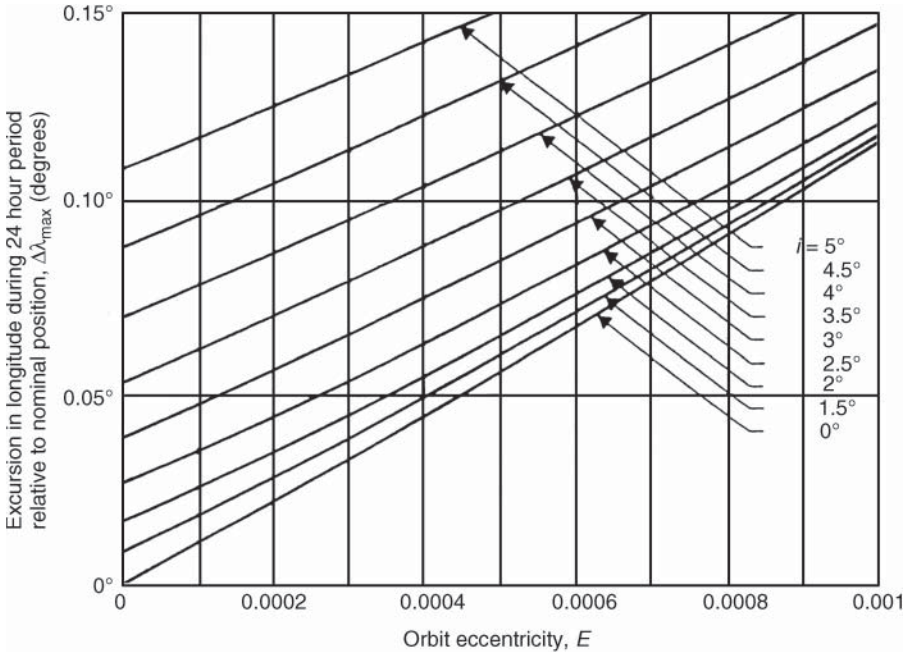


Figure 2.41 Daily variation of longitude due to the effect of residual eccentricity e and inclination i (the peak-to-peak variation is equal to twice the value given in the figure). Source: from (CCIR-Rep 556-4). Reproduced by permission of the ITU.

$\Delta\lambda_{max} = 4.36 \times 10^{-3}$ degrees). The maximum daily variation of longitude due to eccentricity and inclination of orbit is illustrated in Figure 2.41.

2.3.4.3 *The station-keeping box*

To fulfil its mission, the satellite must remain stationary with respect to the earth and occupy a well-defined position on the equator. However, the combined effect of oscillations of the period of 24 hours due to inclination and eccentricity and the long-term drift of the mean longitude leads to an apparent movement of the satellite with respect to its nominal position. Figure 2.42 shows the relative movement of the satellite with respect to its nominal position for an orbit of semi-major axis 42 164.57 km, eccentricity 2×10^{-4}, and inclination 0.058°. As it is, in practice, impossible to maintain the satellite absolutely immobile with respect to the earth, a *station-keeping box* is defined.

The station-keeping box represents the maximum permitted values of the excursions of the satellite in longitude and latitude. It can be represented as a pyramidal solid angle, whose vertex is at the centre of the earth, within which the satellite must remain at all times. The station-keeping box is defined by the two half-angles at the vertex, one within the plane of the equator (E–W width), and the other in the plane of the satellite meridian (N—S width). The maximum value of the residual eccentricity determines the variations of the radial distance (i.e. $2ae$, from Eq. (2.16)). Figure 2.43 shows the volume available for relative displacement of the satellite with respect to its original central position for a window with a typical specification of $\pm 0:05°$ in longitude and latitude and 4×10^{-4} eccentricity.

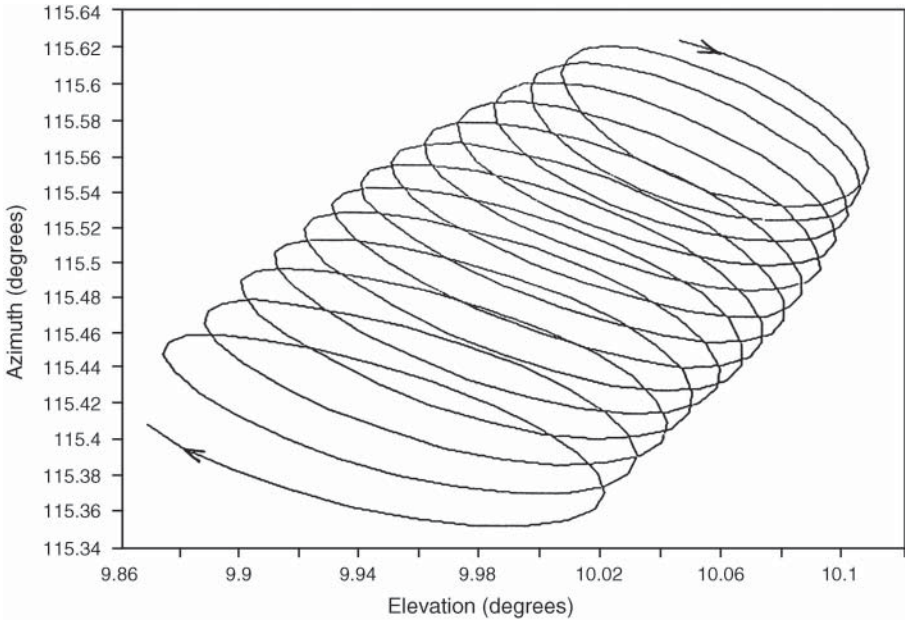


Figure 2.42 Apparent movement of a satellite due to the combined effect of non-zero eccentricity $= 2 \times 10^{-4}$ and inclination = 0.058° (during the 14 days between station-keeping manoeuvres).

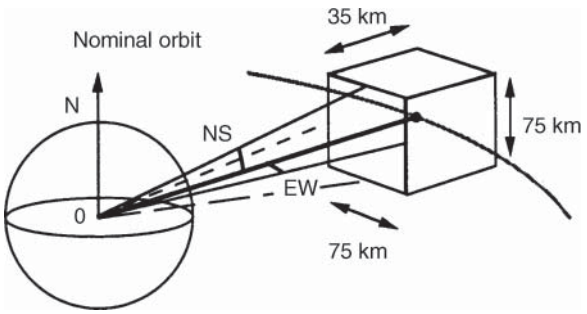


Figure 2.43 Station-keeping box (±0.05° in longitude and latitude, $e = 4 \times 10^{-4}$).

The objective of station keeping is to control the progression of the orbital parameters under the effect of perturbations by applying periodic orbit corrections in the most economic manner so that the satellite remains within the box.

The dimensions of this box are fixed by the mission. They are determined by the following considerations:

— As the dimensions become smaller, the ground station antenna pointing and tracking systems become simpler.
— When the beam width of the ground station antennas is large or when the station is mounted on a vehicle (an aircraft, a boat, or a lorry) that contains a pointing system to take its movement into account, a box of fairly large dimensions is acceptable.

— Geostationary satellites equipped with narrow-beam antennas pointing towards specific sites on the earth require more and more precise station keeping as the beams become narrower. This precision also permits the use of ground station antennas with fixed pointing.
— The adoption of a strict station-keeping tolerance for satellites permits better utilisation of the orbit of geostationary satellites and the radio-frequency spectrum (ITU-R Rec. S.484) [ITUR-92c].

The Radio Communication Regulations [ITU-16] impose a station-keeping accuracy of ±0.1° in longitude for fixed and broadcast service satellites. A tolerance of ±0.5° in longitude is permitted for satellites that do not use the frequency bands allocated to fixed or broadcast satellite services.

2.3.4.4 Effect of orbit corrections

Orbit corrections are achieved by applying *velocity increments* ΔV to the satellite at a point in the orbit. These velocity increments are the result of forces acting in particular directions on the centre of mass of the satellite for sufficiently short periods (compared with the period of the orbit) for them to be considered as impulses. The impulse applied can be radial, tangential, or normal to the orbit in accordance with the definitions given earlier for the velocity at a point in the orbit defined by r, λ, and φ. The impulse does not change the values of r, λ, and φ instantaneously but modifies the component of velocity concerned by a quantity ΔV. The effect of this velocity increment on the orbit parameters is determined from Eqs. (2.94a)–(2.94c). It can be shown that a normal impulse modifies the inclination, a radial impulse modifies the longitude and the eccentricity, and a tangential impulse modifies the drift and the eccentricity.

Actuators are, therefore, mounted on the satellite and are capable of producing forces perpendicular to the orbit to control the inclination and tangential forces (parallel to the velocity). There is no need to generate radial thrusts since a modification of the longitude is obtained from a drift created by the tangential impulses, which also permits the eccentricity to be controlled at lower cost.

The actuators thus permit independent control of movements out of the plane of the orbit (*north–south station keeping*) by control of the inclination and movements in the plane of the orbit (*east–west station keeping*) by control of the drift and, if required, the eccentricity.

There could, however, be coupling due to inaccuracy of the satellite attitude control and bias of the actuator mountings; this could, for example, cause a thrust that should be oriented perpendicularly to the orbit not to be perpendicular in practice. A component acting in the plane of the orbit is thus generated. The most commonly used actuators generate thrusts by burning chemicals called *propellants*. The quantity of propellant consumed is related to the velocity increment provided (see Section 10.3.1).

2.3.4.5 North–south station keeping

North–south station keeping is achieved by thrusts acting perpendicularly to the plane of the orbit, thereby modifying its inclination. Only the long-term drift of the inclination vector is corrected, since the amplitude of periodic perturbations (2×10^{-2} degree) remains less than the normal size (0.1°) of the window.

The optimum procedure is to induce a modification of the inclination vector in the opposite direction from that of the drift Ω_D defined by Eq. (2.92). This conditions the value of right ascension of the point of the orbit where the manoeuvre is performed. In Figure 2.44a, the exterior circle represents the north–south width of the window. The maximum permitted value of inclination

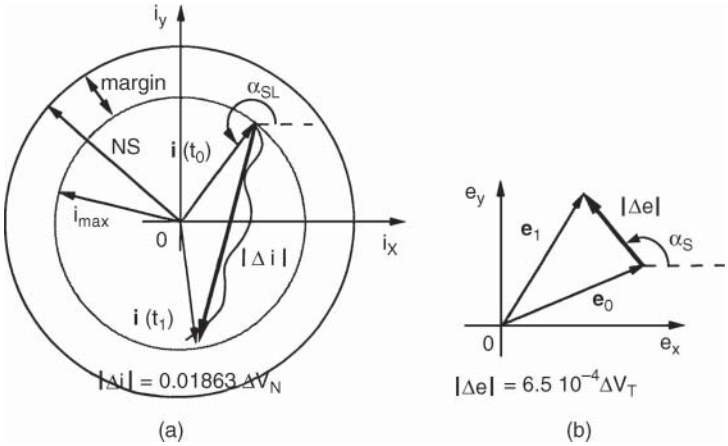


Figure 2.44 Modification of (a) inclination and (b) eccentricity.

is represented by the interior circle, which differs from the previous one by a margin calculated to take into account measurement and orbit restoration inaccuracies.

Modification of the i_x and i_y components of the inclination vector as a function of the value of the normal velocity increment ΔV_N at a point of right ascension α_{SL} is given by:

$$\Delta i_x = \Delta V_N \cos \alpha_{SL} / V_S$$
$$\Delta i_y = \Delta V_N \sin \alpha_{SL} / V_S \tag{2.95}$$

where V_S is the velocity of the satellite in the orbit and is equal to 3075 m s^{-1}. The relation between the modulus of the inclination correction $|\Delta i|$ and the modulus of the normal velocity increment V_N is thus:

$$|\Delta i| = 0.01863 |\Delta V_N| \tag{2.96}$$

where

$$|\Delta V_N| = 53.7 |\Delta i| \quad (\Delta V : \text{m/s}, \Delta i : \text{degrees})$$

As discussed in Section 2.3.3.4, the $\Delta i = \Delta t$ calculated from Eq. (2.92) varies from 0.75 to 0.95°/year depending on the considered year. Therefore the velocity increments necessary to compensate for the corresponding inclination drift each year vary between $53.7 \times 0.75 = 40$ m s^{-1} and $53.7 \times 0.95 = 51$ m s^{-1} depending on the year. The total amount of ΔV for the duration of the mission is to be calculated depending on the launch date in order to determine how much propellant is required for north–south station keeping (see Section 10.3). That total is around 665–695 m s^{-1} for a 15-year mission life.

In the case where only the long-term drift is corrected, the cost of north–south control is independent of the number of manoeuvres performed.

Various strategies are possible:

— One strategy is to allow the inclination vector to drift up to the maximum permitted value and then to apply a correction in the opposite direction from that of the drift such that the inclination vector returns to the opposite position of the permitted region. This strategy minimises the number of manoeuvres but requires large velocity increments, which may lead to problems of coupling with east–west station keeping.

— It may be useful, from the point of view of operational load at the satellite control centre, to coordinate north–south manoeuvres with those of east–west station keeping, whose period of recurrence is different and shorter. In this case, corrections are made before the inclination reaches the limiting value.
— With the previous two strategies, for the corrections to be optimum, the values of right ascension of the points where they are performed are imposed by the position of the drift vector at the end of the cycle. Depending on the techniques used for attitude control during the correction, it can be that the manoeuvres are not permitted in certain sections of the orbit (the earth–sun–satellite geometry may affect the accuracy of attitude measurement, for example). The corrections cannot, therefore, be made in an optimum manner during certain periods of the year when the value of right ascension falls in the prohibited regions. The strategy thus consists of positioning the extremity of the inclination vector just before entry into the critical period, so that the greatest area of the permitted region is available to the inclination vector drift.

It should be noted that the right ascension of the point where the correction is performed does not necessarily correspond to a node of the orbit. This position is obligatory only if it is wished to make the inclination zero, and this generally is not the case for the strategies described. However, the general direction of the drift tends to lead the inclination vector to the y axis (Ω close to 90°) in the reference frame of Figure 2.4. Control of the drift is, therefore, performed globally by compensating for rotation of the plane of the orbit by means of thrusts directed towards the south with Ω close to 90° or thrusts directed towards the north with Ω close to 270°. The time of day depends on the season. In summer, the thrust to the south is performed towards midday and in winter towards midnight; in spring and autumn, it is performed in the evening and morning, respectively. The thrust towards the north is performed with an offset of 12 hours.

Finally, in order to reduce the cost of station keeping, it is possible not to provide inclination control by allowing a large value for the maximum permitted inclination: for example, 3°. At the start of the satellite's life, an inclination equal to the maximum permitted value and a RAAN of the orbit are imposed such that the initial inclination vector is parallel to and opposed to the mean direction of the natural drift for a period corresponding to the lifetime of the satellite. The inclination decreases for about half of the lifetime, passes through zero, and then increases until it reaches the maximum value that determines the end of the operational life of the satellite. As the mean annual drift is on the order of 0.85°, the lifetime could, for example, be chosen as around seven years.

The principal consequences of a non-zero inclination are a north–south oscillation of the satellite (see Section 2.2.3) as seen from earth stations and a displacement of the coverage of the satellite antennas. This displacement can be compensated for by using a steerable antenna or acting on the satellite attitude control (*Comsat manoeuvre*) [ATI-90].

A similar strategy can be used when a satellite is put into orbit before the operational requirements become effective (as a consequence of an excessively early launch slot reservation). At launch, the satellite is put into a parking orbit whose inclination is chosen to be such that, as a consequence of natural drift, the inclination is zero on the date when the satellite is put into service.

2.3.4.6 East–west station keeping

East–west station keeping is provided by thrusts acting tangentially to the orbit. It is divided into control of drift (maintenance of mean longitude) and, if necessary, control of eccentricity.

Maintenance of longitude consists of compensating for longitudinal drift due to the ellipticity of the equator, and hence the value depends on the orbital position of the satellite. Control of eccentricity consists of maintaining the modulus of the eccentricity less than the maximum permitted eccentricity.

An isolated tangential impulse modifies both the semi-major axis, and hence the drift, and the eccentricity of the orbit. From Eq. (2.19b), the modification of the semi-major axis Δa as a function of the value of the tangential velocity increment ΔV_T has the value

$$\Delta a = -(2/\Omega_E)\Delta V_T \text{ (m)} \tag{2.97}$$

where $\Omega_E = 7:292 \times 10^{-5}$ rad s^{-1} is the angular velocity of rotation of the earth, and ΔV_T is expressed in m/s.

From Eq. (2.85), the modification $\Delta d = \Delta\lambda_m/dt$ of the drift as a function of Δa or ΔV_T has the value

$$\Delta d = -(3\Omega_E/2a_s)\Delta a \tag{2.98a}$$

where a_s is the semi-major axis of the orbit. Therefore:

$$\Delta d = (3/a_s)\Delta V_T = (3\Omega_E/V_S)\Delta V_T \text{ (degree/day)} \tag{2.98b}$$

The modification of the e_x and e_y components of the eccentricity vector as a function of the value of the tangential velocity increment ΔV_T at a point of right ascension α_{SL} is given by (Figure 2.44b):

$$\begin{aligned} \Delta e_x &= 2\Delta V_T(\cos\alpha_{SL}/V_S) \\ \Delta e_y &= 2\Delta V_T(\sin\alpha_{SL}/V_S) \end{aligned} \tag{2.99a}$$

where V_S is the velocity of the satellite in the orbit equal to 3075 m s^{-1}. The relation between the modulus of the eccentricity correction $|\Delta e|$ and the modulus of the tangential velocity increment $|\Delta V_T|$ is then:

$$\begin{aligned} |\Delta e| &= 2\,|\Delta V_T|/V_S = 6.5 \times 10^{-4}\,|\Delta V_T| \\ |\Delta V_T| &= 1537.5\,|\Delta e| \quad \text{(m/s)} \end{aligned} \tag{2.99b}$$

2.3.4.6.1 Modification of satellite drift

A tangential velocity increment $\Delta V_T = 1$ m s^{-1} leads to a decrease of the semi-major axis $\Delta a = -27.4$ km, a positive drift increment $\Delta d = 7.11 \times 10^{-8}$ rad s^{-1} $= 4.08 \times 10^{-6}$ degrees/ 0.352°/ day, and a modification of the eccentricity $|\Delta e| = 0.65 \times 10^{-3}$.

The mean longitude is not instantaneously modified by the applied velocity increment; it changes progressively as a consequence of the combined effect of drift and eccentricity. Figure 2.45 illustrates the modification of the orbit and the variation of longitude as a function of time for an initially geostationary satellite to which a thrust and hence a velocity increment ΔV have been applied directed towards the east. Following the impulse, the longitude increases slightly towards the east and then decreases continuously towards the west. The new final orbit (f) is an elliptical orbit with an apogee altitude above the initial orbit and the perigee located where the impulse was applied.

Modification of the value of the drift alone, without changing the eccentricity, or the inverse, can be achieved by applying two thrusts in opposite directions separated by half a sidereal day:

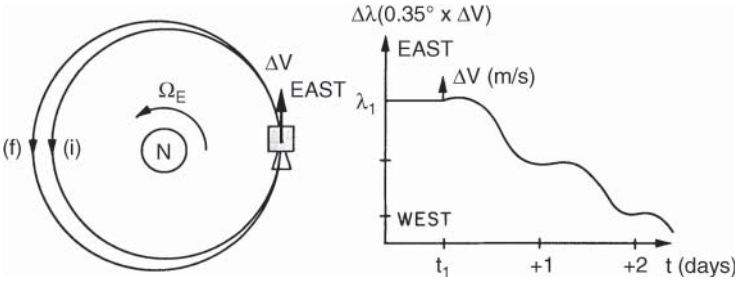


Figure 2.45 The effect of an impulse tangential to the orbit.

that is, at two points of right ascension α_{SL} and $\alpha_{SL} + \pi$. This gives:

$$\Delta d = +(3\Omega_E/V_S)(\Delta V_{T1} + \Delta V_{T2})$$
$$\Delta e = (2/V_S)(\Delta V_{T1} - \Delta V_{T2})(\cos\alpha_{SL} + \sin\alpha_{SL})$$

The strategies used to control the natural drift depend on the size of the radius of the circle of natural eccentricity with respect to the maximum permitted eccentricity.

2.3.4.6.2 Control of natural drift

If the radius r_e of the circle of natural eccentricity is less than the permitted eccentricity, only the progression of drift is controlled. To obtain an eccentricity of the orbit that is always less than the limiting value, it is necessary to locate the centre of the circle of eccentricity at the origin of the reference frame and to orient the eccentricity vector in the direction of the sun. This means a non-zero eccentricity equal to the radius of eccentricity must be imposed on the orbit on injection of the satellite into orbit; the perigee of this orbit must be in the direction of the sun. The orientation of the orbit follows the direction of the sun with the rotation of the earth about the sun, and the resulting – called 'natural' – eccentricity, with value e_n equal to r_e, remains constant and hence less than the limiting value.

The strategy used to control drift alone depends on the position of the satellite with respect to the stable points. Movement in the equatorial plane due to asymmetry of the terrestrial potential obeys Eq. (2.91):

$$(d\Lambda/dt)^2 - D\cos 2\Lambda = \text{constant}$$

where $\Lambda = \lambda - \lambda_{22} \pm 90°$ is the longitude of the satellite measured with respect to the closest stable equilibrium point and $D = 18n_S^2(R_E/a_s)^2J_{22} = 4 \times 10^{-15}\text{rad/s}^2$.

The satellite must be maintained at the nominal longitude Λ_N, measured with respect to the position of the nearest stable equilibrium point, while tolerating a small maximum deviation $\varepsilon/2$ on each side of Λ_N. This deviation $\pm\varepsilon/2$ is determined from the dimensions in longitude of the station-keeping window by deducting a margin that is intended to absorb orbit restitution errors, manoeuvre inaccuracies, and short-period oscillations.

If Δ is the longitude measured from $\Lambda_N(\Delta = \Lambda - \Lambda_N)$, this gives:

$$\cos 2\Lambda = \cos\ 2(\Lambda_N + \Delta) = \cos 2\Lambda_N - 2\Delta\sin 2\Lambda_N$$

and Eq. (2.91) can be written:

$$(d\Delta/dt)^2 + 2D\Delta\sin 2\Lambda_N = \text{constant} \qquad (2.100)$$

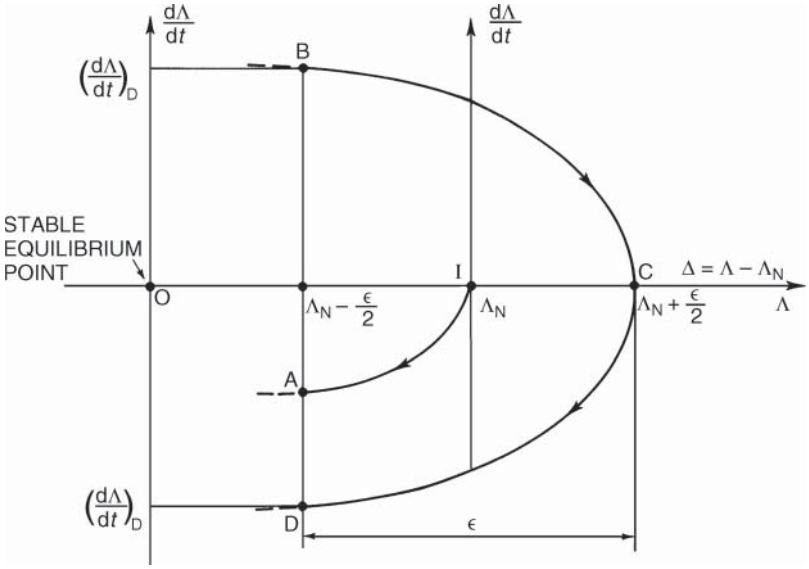


Figure 2.46 Strategy for maintaining longitude away from a point of equilibrium.

The curve representing the drift $d\Delta/dt$ as a function of Δ is a parabola defined by the nominal longitude Λ_N and the initial conditions. If the satellite is initially at point I (nominal position) with zero drift (Figure 2.46), it moves towards the point of stable equilibrium by following the parabola from Eq. (2.100) that has its vertex at I.

Strategy of 'the inclined plane'. When the longitude reaches the value $\Lambda = \Lambda_N - \varepsilon/2$ (point A, the west limit), a velocity increment is applied to the satellite to cause it to describe, at the point representing the satellite, a parabola that has its vertex at $\Lambda = \Lambda_N + \varepsilon/2$ (point C), the eastern limit of movement. From C, the satellite moves nearer to the point of stable equilibrium to reach point D ($\Lambda = \Lambda_N - \varepsilon/2$, the western limit). By applying a suitable velocity impulse ΔV_T at this point, the drift $(d\Lambda/dt)_D$ changes sign, and the representative point in Figure 2.46 changes from D to B. Then the cycle starts again.

East–west station keeping thus consists of causing the satellite to describe the curve BCD in Figure 2.46 at the point representing the satellite.

The velocity increment to be applied. In Eq. (2.100), the constant is calculated from a location at point $C(\Delta = \varepsilon/2;\ d\Lambda/dt = 0)$. Hence:

$$\text{Constant} = 2D(\varepsilon/2)\sin 2\Lambda_N$$

Equation (2.100) can be written:

$$(d\Lambda/dt)^2 = -2D(\Delta - \varepsilon/2)\sin 2\,\Lambda_N \tag{2.101}$$

At the point $D(\Delta = -\varepsilon/2)$:

$$(d\Lambda/dt)_D^2 = 2D_\varepsilon \sin 2\Lambda_N$$

from which:

$$(d\Lambda/dt)_D = -\sqrt{(2D_\varepsilon \sin 2\Lambda_N)} \tag{2.102}$$

To move from point D to point B, it is necessary to impose a variation of drift $d\Lambda/dt$ equal to $-2(d\Lambda/dt)_D$. Eq. (2.98b) enables the corresponding velocity increment ΔV_T to be calculated:

$$\Delta V_T = (V_S/3\Omega_E)\Delta d \ (\Delta V_T \text{ and } V_S \text{ in m/s}; \Omega_E \text{ and } \Delta d \text{ in rad/s})$$

Hence, for $\Delta d = -2(d\Lambda/dt)_D$

$$\Delta V_T = (V_S/3\Omega_E)2\sqrt{(2D_\varepsilon \sin\ 2\Lambda_N)}$$

Finally:

$$\Delta V_T = 2.5\sqrt{(\varepsilon \sin\ 2\Lambda_N)} \text{ (m/s; rad)} \tag{2.103}$$

As the satellite always drifts in the same direction, this velocity increment is always applied in the same direction (that of the natural drift).

Correction periods. The period of application of the velocity impulses is equal to the duration T to evolve along the parabolic path in Figure 2.46 from point B to D. This period T is equal to twice the time calculated by integrating Eq. (2.101) from $\varepsilon/2$ to $-\varepsilon/2$:

$$T = [(2\sqrt{2})/\sqrt{D}]\sqrt{(\varepsilon/\sin 2\Lambda_N)} = 516\sqrt{(\varepsilon/\sin 2\Lambda_N)} \text{ (days)} \tag{2.104}$$

where ε is in radians.

The velocity impulse and cycle duration depend on the position of the satellite with respect to the stable equilibrium point and the longitudinal dimension of the window.

Annual velocity impulse. The velocity impulse to be applied per year is $\Delta V_{T\,year} = \Delta V_T(365/T)$,

$$\Delta V_{T\ year} = 1.77 \sin 2\Lambda_N \text{ (m/s)} \tag{2.105}$$

It depends on the longitude of the point about which the satellite is maintained and not on the total window width (if only the long-term drift is corrected and not the short-term variations). At maximum, it is on the order of $2\,m\,s^{-1}$.

The preceding calculations are based on the equations in Section 2.3.3.3, which were obtained by neglecting, in particular, terms of order greater than 2 in the expansion of the terrestrial potential. When the actual acceleration to which the satellite is subjected is considered (Figure 2.36), the required annual velocity increment is given by Figure 2.47.

The 'return to centre' strategy. Equation (2.105) and Figure 2.47 indicate that the annual velocity increment depends on the longitude of the position in the nominal orbit of the satellite. In the vicinity of a point of equilibrium, this velocity increment is very small. In practice, it is not the drift due to asymmetry of the terrestrial potential that must be compensated for, but the east–west component of the velocity increments induced by corrections perpendicular to the equatorial plane (north–south control). The strategy adopted in this case consists of locating the satellite at the centre of the window to perform the north–south correction in order to provide the maximum margin in longitude. A correction in longitude is then performed before the satellite reaches the limit of the window in order to return it to the centre of the window. A second correction in longitude cancels the drift caused by the first.

Example 2.4 Figure 2.48a illustrates the variations of longitude of a satellite with the 'inclined plane' strategy. The nominal position is at longitude 49° east (Indian Ocean) and $\Lambda_N = |\lambda - \lambda_{22} - 90°| = 26°$. The natural drift of the longitude is directed towards the east. To guarantee a window of ±0.5° ($\varepsilon = 1° = 0.017$ rad), the correction period must be 75 days, and the annual velocity increment is $1.5\,m\,s^{-1}$. Velocity increments are applied when the satellite approaches the eastern edge of the window in such a way as to impose a drift that becomes zero before it reaches the western edge of the window.

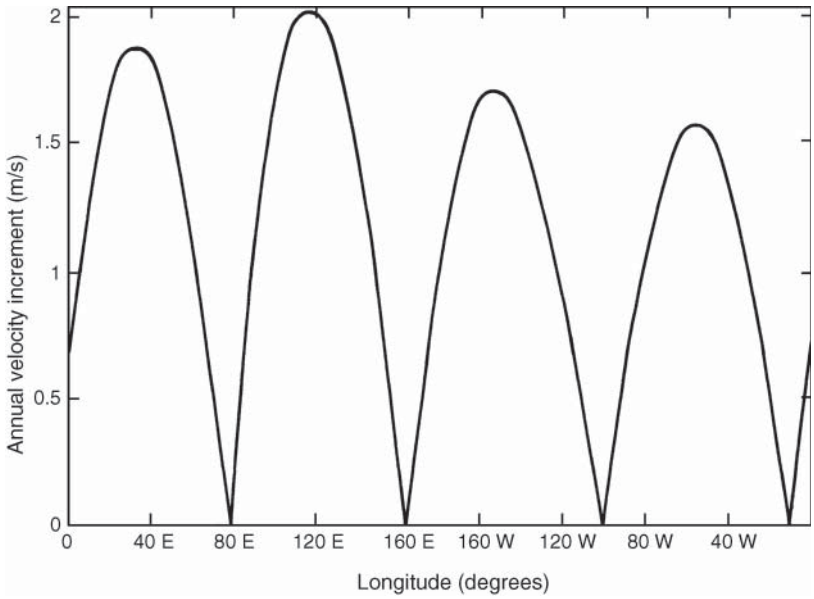


Figure 2.47 Annual velocity increment required to control east–west drift as a function of satellite longitude.

Figure 2.48b shows the 'return to centre' strategy. The longitude of the satellite is 11.5° west, which is near a point of unstable equilibrium. The effect of a north–south correction on the initial longitude drift can be observed. An east–west correction creates a drift that returns the satellite to the centre of the window. A second east–west correction cancels this drift.

2.3.4.6.3 Control of eccentricity

The preceding strategies are applicable when the radius of eccentricity r_e of the orbit is sufficiently small for the eccentricity of the orbit not to be controlled. When this is not the case, for satellites with large solar panels, the natural eccentricity ($e_n = r_e$) of the orbit is such that the induced non-controlled movement in longitude ($\Delta\lambda = 2e_n \sin M$) would occupy an excessively large part of the window.

It is then necessary to prevent the eccentricity from exceeding a value e_{max} such that $2\Delta\lambda_{max} = 4e_{max}$ is the part of the box allocated to oscillation of longitude under the effect of eccentricity. This part is determined from the dimension in longitude of the station-keeping window by deducting (i) a band that permits variation of the mean longitude under the effect of long-term drift during the expected time between two east–west manoeuvres, (ii) a margin to take into account the inaccuracies of manoeuvring and orbit restoration, and (iii) the east–west effects induced by north–south control.

The strategy consists of locating the eccentricity vector ($\mathbf{e}_0$) on the boundary circle that defines the region of maximum permitted variation of eccentricity (centre 0, radius e_{max}) at the start; the right ascension of the sun is $(\alpha_{SUN})_0$ (Figure 2.49). The extremity of the eccentricity vector varies on a circle of radius r_e and centre C such that the radius vector from the centre C to the extremity of the eccentricity vector **e** remains parallel to the direction of the sun. When the eccentricity vector again reaches the boundary circle ($\mathbf{e}_1$), the right ascension of the sun is

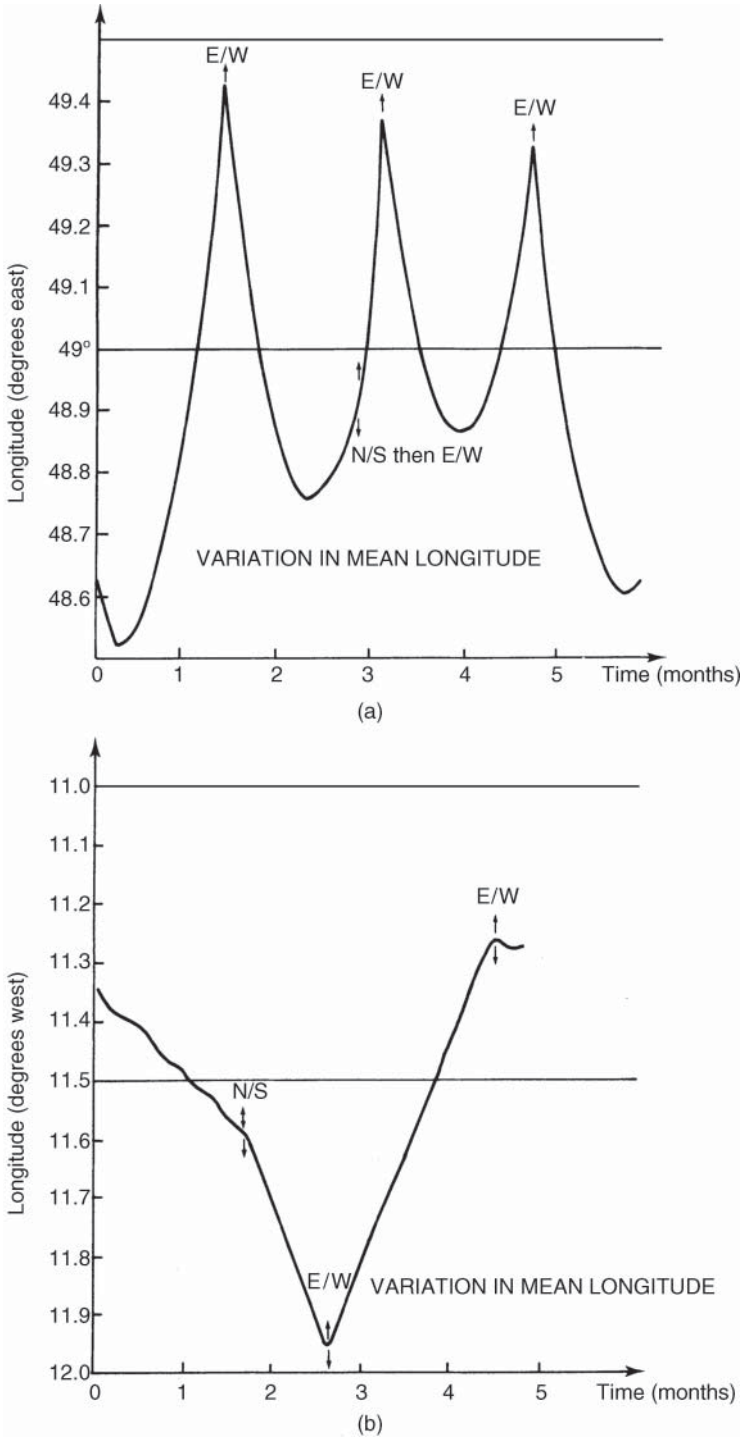


Figure 2.48 Evolution and control of longitude: (a) satellite far from a point of equilibrium; (b) satellite close to a point of unstable equilibrium.

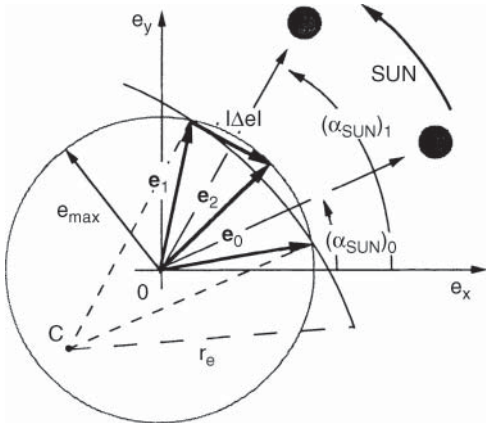


Figure 2.49 Strategy for controlling eccentricity.

$(\alpha_{SUN})_1$. A velocity increment is then applied that puts the extremity of the eccentricity vector on the boundary circle in the same position ($\mathbf{e}_2$) with respect to the sun as at the start of the cycle.

The station-keeping strategy cycle is thus determined by the duration between corrections, and the cost is high when the maximum permissible eccentricity is small. It is possible to combine eccentricity control manoeuvres with those of drift control.

2.3.4.7 Operational aspects: correction cycles

The control strategies aim to provide satellite station keeping while minimising the amount of propellant consumed to generate the required velocity increments. It is also necessary to take into account constraints imposed by operational aspects (such as obligations to personnel) and security. Hence repetitive correction cycles with a period that is a multiple of a week are well suited.

A typical example is a 14-day cycle that includes north–south and east–west corrections. The organisation of the cycle is as follows:

1. Inclination correction at the start of the cycle.
2. Measurement and restoration of the orbit (about two days of measurements are necessary to determine the orbit with sufficient accuracy to calculate the corrections to be performed).
3. Eccentricity or drift corrections.
4. Measurement and restoration of the orbit; verification of the result of the corrections.
5. Natural progression of the orbit.
6. Measurement and restoration of the orbit at the end of the cycle in preparation for the following cycle.

2.3.4.8 Overall cost of station keeping

As long as the amplitude of periodic perturbations is compatible with the size of the station-keeping box, the effect of these perturbations is not corrected and does not affect the station-keeping budget.

By correcting only long-term drifts, the budget is on the order of:

— 43–48 m s^{-1} per year for north–south control (inclination correction)
— 1–5 m s^{-1} per year for east–west control (longitude drift and eccentricity corrections)

The actual total cost depends on:

— The date of the start of station keeping
— The longitude of the station
— The S_a/m ratio of the satellite (the ratio of satellite of apparent surface S_a in the direction of the sun and satellite mass m)
— The dimensions of the window

2.3.4.9 Termination of station keeping at the end of life

Satellite station keeping is possible by means of propellants (indispensable for the operation of thrusters), which are stored in reservoirs. When the propellants are consumed, station keeping is no longer provided, and the satellite drifts under the effect of the various perturbations. In particular, it adopts an oscillatory movement in longitude about the point of stable equilibrium (see Section 2.3.3.3), which causes it to sweep a portion of space close to the orbit of geostationary satellites. Although small, the associated probability of collision is not zero (around 10^{-6} per year [HEC-81]).

Consequently, a special procedure is adopted that aims to remove satellites from geostationary orbit at the end of their lifetime (ITU-R Rec. S.1003) [ITUR-10]. Manoeuvres are performed using a small quantity of propellant that is reserved for this purpose before complete exhaustion of the reservoirs. These manoeuvres place the satellite in an orbit of higher altitude (about 150 km: that is, a ΔV requirement of 5.4 m s^{-1}) than that of geostationary satellites (an orbit of lower altitude is undesirable because of the possible danger of collision during operations to install geostationary satellites in orbit). This operation requires a quantity of propellant on the order of 7 kg for a satellite of 3000 kg at the end of its life. This quantity of propellant represents approximately six weeks of normal station keeping and hence potential availability of the satellite.

A major difficulty lies in estimating the quantity of propellant remaining in the reservoirs at a given instant. This estimate is made from the variation of the pressurising gas in the propellant reservoirs and by integration of the operating time of the thrusters during the lifetime of the satellite. The error is large and can be on the order of the quantity of propellant required to provide satellite station keeping for a few months.

2.3.4.10 Measurement and orbit parameter estimation

Determination of the position of a geostationary satellite depends on distance and angular measurement.

Distance measurement. Measurement of satellite distance depends on measurement of the propagation time of an electromagnetic wave between the ground and the satellite. The distance d between a transmitter and a receiver is deduced from a measurement of the phase shift $\Delta\Phi$ between the transmitted and received waves:

$$\Delta\Phi = 2\pi f(d/c) \qquad \text{(rad)}$$

In practice, the phase shift between a sinusoidal signal of frequency f that modulates the command carrier and the same signal after retransmission by the satellite in the form of modulation of the telemetry carrier is measured (see Section 10.5).

Angular measurement. Several procedures are possible:

— *By measuring the antenna pointing angles*: The radiation pattern of the antenna receiving the telemetry signal from the satellite is used. The antenna direction is controlled in such a way that the satellite is on the axis of the main lobe of the antenna. The accuracy of measurement is on the order of 0.005° to 1° according to the mechanical characteristics of the antenna.

— *By interferometry*: Two stations A and B separated by a distance L, called the *base*, receive a signal from the satellite: for example, the telemetry carrier of frequency f (Figure 2.50). The trajectory difference Δd between the satellite and each of the stations A and B gives rise to a propagation time difference $\Delta t = \Delta d/c$ and is measured by a phase shift $\Delta\Phi = 2\pi f \times \Delta t$ between the received signals:

$$\Delta\Phi = (2\pi L \cos E)/\lambda$$

The value of the elevation angle E is deduced from this expression, and the satellite is situated on a cone of axis AB and half angle E at the vertex. The combination of two bases enables the satellite to be located on the common generating line of the two cones. The accuracy of the measurement is on the order of 0.01°. It is not sufficient to determine the final orbit, but this type of measurement is useful during the launch into orbit phase.

Estimation of the orbit parameters is made by means of a series of distance or angle measurements. Various methods are possible in accordance with the number of operational stations. For

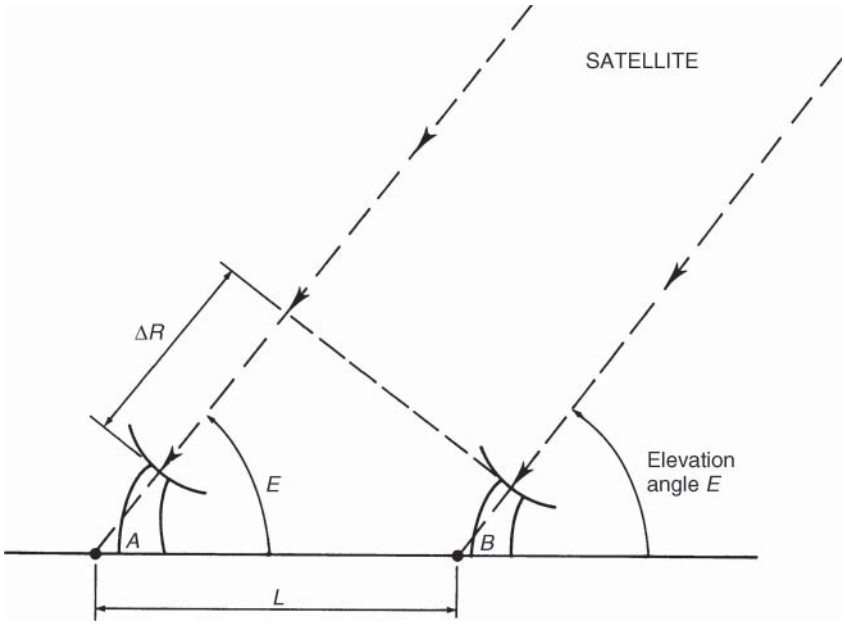


Figure 2.50 Angle measurement by interferometry.

geostationary satellites, it is current practice to use only one measurement station and combine a series of distance and angle measurements made at different times. The typical accuracy for the estimation of orbit parameters is as follows:

— *Semi-major axis*: 60 m
— *Eccentricity*: 10^{-5}
— *Inclination*: 3×10^{-3} degrees
— *Longitude*: 2×10^{-3} degrees

2.4 CONCLUSION

The aspects of orbit geometry presented in this chapter are fundamental to comprehension of the design and modes of operation of artificial satellites. In the context of satellite communications systems, they determine the launching and orbit-control procedures, the design of the platform (including attitude control, thermal control, the electrical power supply system, the propulsion system, etc.), and the characteristics of the radio-frequency links (such as path losses, propagation times, antenna pointing, and satellite–sun conjunction).

REFERENCES

[AKT-08] Akturan, R. (2008). An overview of the Sirius satellite radio system. *International Journal of Satellite Communications* **26** (5): 349–358.

[ASH-88] Ashton, C.J. (1988). Archimedes: Land mobile communications from highly inclined satellite orbits. In: *Fourth International Conference on Satellite Systems for Mobile Communications and Navigation*, 133–132. IET.

[ATI-90] Atia, A., Day, S., and Westerlund, L. (1990). Communications satellite operation in inclined orbit: 'the Comsat Maneuver'. In: *13th International Communication Satellite Systems Conference, Los Angeles, March*, 452–455. AIAA.

[BIE-66] Bielkowicz, P. (1966). Ground tracks of earth-period satellites. *AIAA Journal* **4** (12): 2190–2195.

[BOU-90] Bousquet, M. and Maral, G. (1990). Orbital aspects and useful relations from earth satellite geometry in the frame of future mobile satellite systems. In: *13th AIAA International Communication Satellite Systems Conference and Exhibit, Los Angeles, CA, Technical Papers. Part 2 (A90-25601 09-32)*, 783–789. Washington, DC: AIAA.

[DON-84] Dondl, P. (1984). LOOPUS opens a new dimension in satellite communications. *International Journal of Satellite Communications* **2** (4): 241–250.

[HEC-81] Hechler, M. and Van Der Ha, J.C. (1981). Probability of collisions in the geostationary ring. *Journal of Spacecraft* **18** (4): 361–366.

[HED-87] Hedin, A.E. (1987). MSIS-86 Thermospheric model. *Journal of Geophysical Research* **92**: 4649–4662.

[ITU-16] ITU. (2016). Radio regulations.

[ITUR-90] ITU-R. (1990). Systems for the broadcasting satellite service (sound and television). Report BO.215-7.

[ITUR-92a] ITU-R. (1992). Compensation of the effects of switching discontinuities for voice band data and of doppler frequency-shifts in the fixed-satellite service. S.730.

[ITUR-92b] ITU-R. (1992). The effect of transmission delay in fixed satellite service. CCIR report 383-4.

[ITUR-92c] ITU-R. (1992). Station-keeping in longitude of geostationary satellites in the fixed-satellite service. Recommendation S.484-3.

[ITUR-10] ITU-R. (2010). Environmental protection of the geostationary-satellite orbit. Recommendation S.1003-2.

[ITUT-03] ITU-T. (2003). One-way transmission time. G114.

[JAC-77] L. Jacchia et al. (1977). Thermospheric temperature, density and composition: new models. Smithsonian Astrophysical Observatory Special Report No. 375.

[KAU-66] Kaula, W. (1966). *Theory of Satellite Geodesy*. Waltham: Blaisdell.

[PRI-93] Pritchard, W., Suyderhoud, H., and Nelson, R. (1993). *Satellite Communication Systems Engineering*, 2e. Prentice Hall.

[ROU-88] Rouffet, D., Dulck, J.F., Larregola, R., and Mariet, G. (1988). SYCOMORES: a new concept for land mobile satellite communications. In: IEEE Conference on Satellite Mobile Communications, Brighton, Sept., 138–142. IEEE.

[VIL-91] Vilar, E. and Austin, J. (1991). Analysis and correction techniques of Doppler shift for nongeosynchronous communication satellites. *International Journal of Satellite Communications* **9** (2): 122–136.

3 BASEBAND DIGITAL SIGNALS, PACKET NETWORKS, AND QUALITY OF SERVICE (QOS)

In this chapter, the term *signal* relates to the voltage conveying information (such as voice, sound, video, or data) from one user terminal to another. Such a signal, called a *baseband signal*, conditions the quality of service (QoS) as perceived by the user. In order to access the radio-frequency channel for routing via the satellite, the baseband signal modulates a radio-frequency carrier. Some processing prior to modulation may be desirable.

The baseband signal considered here is *digital* (the voltage takes discrete values, of which there are a finite number). It can convey information from a single source, or from several sources (known as a *composite* baseband signal, produced through multiplexing of signals from the individual sources).

This chapter discusses the types of baseband signal associated with the considered service and its related quality. The QoS is discussed here in relation to performance objectives, service availability, and delay. Performance is measured by *bit error rate* (BER) for digital signals. The *availability* is the fraction of time during which the service is provided with the desired performance, and the *delay* is the latency from transmission of information to its reception. Delay builds up from propagation delay and network delay.

Baseband digital signals are transferred across networks at the physical level in the form of bits or bytes streams. Packets are transferred at the link or network level in the form of blocks of bytes called *frames* or *packets*. Switching can be at the link level or network level. QoS of packet networks is considered at the packet level, particularly for the Internet and Internet Protocol (IP) packets.

Chapter 4 presents digital communications techniques used to convey digital information signals thanks to modulated carriers.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

3.1 BASEBAND SIGNALS

The following baseband signals are considered:

— Telephone signal
— Sound programme
— Television
— Data

These are the most common baseband signals. Historically, some signals, such as telex and facsimile transmitted on telephone channels, have different characteristics than voice signals.

In general, today's telecommunication services consist of one or more media components of the following [ETSI-07]:

— *Speech*. Voice telecommunication.
— *Audio*. Telecommunication of sound in general.
— *Video*. Telecommunication of full motion pictures, and of stills.
— *Data*. Telecommunication of information-files (text, graphics, etc.).
— *Multimedia (MM)*. A combination of two or more of the previous components (speech, audio, video, data), with a temporal relationship (e.g. synchronisation) between at least two components.

3.1.1 Digital telephone signal

Digital telephony is achieved using voice *encoders*. Various techniques are used that can be categorised into waveform encoding and vocoders [FRE-91].

3.1.1.1 Waveform encoding

Waveform encoding entails three processes:

— Sampling
— Quantisation
— Source encoding

The most popular encoding techniques are pulse code modulation (PCM), delta modulation (DM), and adaptive differential pulse code modulation (ADPCM).

3.1.1.1.1 Pulse code modulation

Sampling is performed at the rate $f_s = 8\,\text{kHz}$, slightly higher than the Nyquist rate (equal to twice the maximum frequency $f_{max} = 3400\,\text{Hz}$ of the spectrum of telephony signals). Quantisation transforms each voltage sample in the sample output into a finite number M of discrete levels ($M = 2^8 = 256$ in Europe). This introduces an error, called *quantisation noise*.

Quantisation can be uniform or non-uniform. Uniform quantisation corresponds to equal quantization steps. Non-uniform quantisation adapts the magnitude of each step according to a compression law that depends on the amplitude distribution of the samples, so as to maintain a

constant signal-to-quantisation-noise power ratio for all sample amplitudes. Two compression laws are commonly used: μ-law and A-law, specified in G711 [ITUT-88a; SCH-80].

Source encoding aims at generating a bit stream that reflects the sequence of quantised samples. The corresponding encoded bit rate is $R_b = mf_s$, where $m = \log_2 M$ represents the number of bits per sample. With $M = 2^8 = 256$, the resulting bit rate is $R_b = 64\,\text{kbit}\,\text{s}^{-1}$.

3.1.1.1.2 Delta modulation

Sampling is performed at the rate $f_s = 16$ or 32 kHz, which is more than twice the Nyquist rate. Quantisation applies to the difference between two successive samples. This difference is encoded in the form of one bit. Therefore, the resulting bit rate is $R_b = 16$ or $32\,\text{kbit}\,\text{s}^{-1}$, depending on the sampling rate. The high value of the sampling rate results from the small information content of the one-bit quantisation process.

3.1.1.1.3 Adaptive differential PCM

A difference signal is generated by taking an estimate of the input signal obtained from the previous input sample and subtracting it from the input signal itself. This difference signal is quantised using four bits per sample. The estimator is adaptive as it takes into account the local waveform shape. This resulting bit rate is $R_b = 16$–$64\,\text{kbit}\,\text{s}^{-1}$.

3.1.1.2 Vocoders

Vocoders assume a given speech-production mechanism and transmit the parameters of the mechanism. The *linear predictive coding* (LPC) technique assumes that the speech-production mechanism can be modelled by a filter. The coefficients of the filter are periodically updated by statistical optimisation over a given number of samples. The period conditions the duration of the frame (10–50 ms) within which the coefficients are transmitted at a data rate R_b as low as 2.4–$4.8\,\text{kbit}\,\text{s}^{-1}$.

3.1.1.3 Digital telephony multiplex

Figure 3.1 illustrates the principle of time division multiplexing (TDM) and demultiplexing. For multiplexing digital telephone channels, two hierarchies (ITU-T Rec. G702 and G704) [ITUT-88b; ITUT-88c] are widespread – the European hierarchy of the European Conference on Post and Telecommunications (CEPT) and the T-carrier hierarchy used in Japan and North America (United States and Canada). Table 3.1 summarises the characteristics of these multiplexing techniques

3.1.1.3.1 The CEPT hierarchy

The CEPT hierarchy is based on a frame of 256 binary elements. The frame duration is 125 microseconds. The bit rate is $2.048\,\text{Mbit}\,\text{s}^{-1}$. The multiplex capacity is 30 telephone channels, 16 bits per frame being used for signalling and the frame synchronisation signal. The highest capacities are obtained by successive multiplexing of multiplexes of equal capacity. In this way, a multiplexing hierarchy is established that contains several levels; each level is constructed by

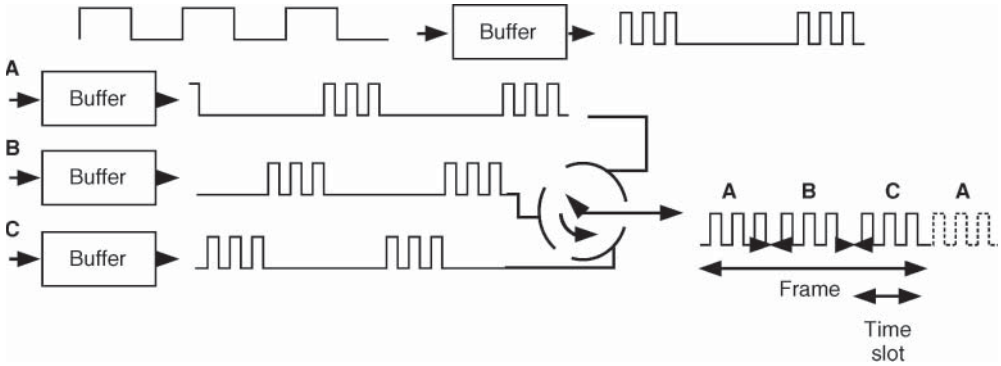


Figure 3.1 Time division multiplexing and demultiplexing.

Table 3.1 Characteristics of CEPT and T-carrier multiplexes

	CEPT		United States/Canada		Japan	
Hierarchy level	Throughput (Mbit s^{-1})	Capacity (channels)	Throughput (Mbit s^{-1})	Capacity (channels)	Throughput (Mbit s^{-1})	Capacity (channels)
1	2048	30	1544	24	1544	24
2	8448	120	6312	96	6312	%
3	34 368	480	44 736	672	32 064	480
4	139 264	1920	274 176	4032	97 728	1440
5	557 056	7680			400 352	5760

multiplexing four multiplexes with a capacity equal to the capacity of the immediately lower hierarchy level.

3.1.1.3.2 The T-carrier hierarchy

This hierarchy is based on a frame of 192 bits obtained by multiplexing 24 samples, each of 8 bits, to which one frame alignment bit is added. Each frame thus contains 193 bits. The frame duration is 125 ms. The bit rate is 1.544 Mbit s^{-1}. The multiplex capacity is 24 channels (23 and 1 for signalling). The multiplexing hierarchy differs between Japan and North America.

3.1.1.3.3 Digital speech interpolation

Techniques for digital speech concentration such as *digital speech interpolation* (DSI) take account of the activity factor of telephone channels in order to reduce the number of satellite channels (called *bearer channels*) required to transmit a given number of terrestrial telephone channels. The speech-interpolation technique is based on the fact that in a normal telephone conversation, each participant monopolises the telephone channel for only around half the time. As the silences between syllables, words, and phrases increases, so does the unoccupied time. Hence the activity factor τ of a circuit is about 40%. By making use of the actual activity of the telephone channels, several users can be permitted to share the same bearer channel [CAM-76].

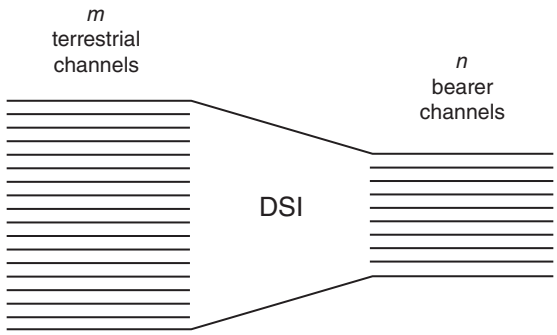


Figure 3.2 Digital speech interpolation (DSI).

Figure 3.2 shows this principle. The gain of the digital concentrator is given by the ratio m/n. In the Intelsat/Eutelsat system, 240 terrestrial telephone channels require only 127 bearer channels plus 1 assignment channel, and the gain is $240/127 = 1.9$. This gain assumes bearer channels at 64 kbit s^{-1}. By adding a low rate encoder (LRE) to the digital speech concentrator, the gain can be further increased. For example, with encoding at 32 kbit s^{-1}, a gain increase by a factor of 2 can be obtained. These techniques are used in *digital circuit multiplication equipment* (DCME). More details on the operation of this equipment are given in Section 8.6.

3.1.1.3.4 Synchronisation between networks

The previous digital hierarchies, called *plesiochronous* digital hierarchies (PDHs) (ITU-T Rec. G702) [ITUT-88b], imply multiplexing the signals at given nominal service rates in a series of stages within the hierarchy to achieve the final bit rate at which the multiplexed signal is sent over the carrier. At the receiving end, the extraction of constituent signals in the aggregate bit stream necessitates going through the whole process of demultiplexing down to the appropriate lower level.

Due to movement of the satellite in its orbit, even if it is a geostationary satellite for which this movement is small but not zero (see Chapter 2), a Doppler effect is observed and the received binary rate is not always equal to the transmitted binary rate. Furthermore, the terrestrial networks in Figure 3.1, when digital, do not always have strictly synchronous clocks. To compensate for these variations, buffer memories are provided at the station–network interfaces. In choosing the size of these memories, account must be taken of the station-keeping specifications and *plesiochronism* between digital interfaces. Plesiochronism exists when the clocks of each network have an accuracy on the order of $\pm 10^{-11}$. This leads to a frame slip for a multiplex frame of 125 microseconds once every 72 days.

3.1.1.3.5 Synchronous digital hierarchy

A synchronous digital hierarchy (SDH) compensates for the need to demultiplex the entire PDH multiplex at the receiving earth station to extract a constituent signal at a given hierarchical level [ITUT-10; ITUT-17]. Moreover, it offers additional features in terms of information content management. With an SDH, the service rate signal is directly mapped in an appropriately sized container, itself mapped along with an overhead signal into a virtual container (VC). Virtual containers can be thought of as a limited set of payload structures, which have specific size and

structural relationships with each other, and which have the ability to carry several tributaries with different capacities and formats. Virtual containers can match the bit rates of the PDH; for instance, the VC11 and VC12 levels correspond respectively to the 1.544 and 2.048 Mbit s^{-1} PDH tributaries. SDH offers several advantages over PDH [ITUT-96]:

— Individual tributaries are more easily identified and extracted from the multiplex, thus reducing the complexity and improving the reliability of multiplexing and cross-connect equipment.
— Integer bit-rate multiplication factors between levels of the multiplex hierarchy, with byte-interleaved multiplexing, allow the easier formation of higher-order multiplex signals.
— Worldwide international standards eliminate different and costly international internetworking arrangements.
— Enhanced network management functions facilitate the cooperation of network operators.

The first hierarchical level of the SDH is STM-1, at a bit rate of 155.52 Mbps, and is not common over satellite links due to power and bandwidth limitations. Sub-STM-1 operation is more appropriate.

3.1.2 Sound signals

A high-quality radio sound programme occupies a band from 40 to 15 kHz. The test signal is a pure sinusoid at a frequency of 1 kHz. Its power relative to the zero reference level for an impedance of 600 Ω is 1 mW or 0 dBm0s (the *s* suffix indicates that the value relates to the sound programme test signal). The mean power of a sound programme is −3.4 dBm0, and the peak power (exceeded for a fraction less than 10^{-5} of the time) is equal to 12 dBm0.

For the emission of digitally encoded audio signals, the analogue sound programme must go through an analogue-to-digital converter. This implies sampling, quantisation, and source encoding. Two encoding techniques, PCM and adaptive delta modulation (ADM), are considered. Some formats have been defined using sampling rates of 32, 44.1, or 48 kHz (S/PDIF, AES/EBU, MUSICAM, etc.). In particular, MUSICAM is a popular standard for digital audio compression based on the division of the audio-frequency band into 30 subbands. MUSICAM proposes various compression ratios (from 4 to 12) starting from a 48 kHz sampling rate and a 16-bit quantisation. MPEG-1 audio was finalised in 1992 and MPEG-2 audio in 1994; the original MUSICAM algorithm is not used anymore.

The ITU-R recommends that for 15 kHz band audio signals broadcast from satellites, where PCM encoding is employed, the sampling frequency should be 32 kHz with 14 bits per sample [ITUR-86].

Where satellites are used for broadcasting of audio programmes to mobile receivers (digital audio broadcasting [DAB]), orthogonal frequency division multiplexing (OFDM) can be used. This transmits data by dividing the stream into several parallel bit streams, each at a lower bit rate, and then uses these substreams to modulate several carriers. OFDM time-domain waveforms are chosen such that mutual orthogonality is ensured, even though the spectra from several subcarriers may overlap. Coded orthogonal frequency division multiplexing (COFDM) is able to combat the time dispersion due to multipath, frequently encountered over mobile satellite channels [SAR-94].

3.1.3 Television signals

Transmission of colour television started using analogue techniques in the mid-1900s. Various non-compatible standards were available: NTSC (Japan, United States, Canada, Mexico, some

South American countries, and Asia), PAL (Europe, except France, Australia, some South American countries, and some African countries), and SECAM (France, countries of the former USSR, Eastern countries and, some African countries). The multiplexed analogue components (MAC) standard was proposed in the 1980s for satellite broadcasting (direct broadcast television) but was never developed into a successful commercial service. Fully digital techniques based on video compression were developed in the 1990s, resulting in the widely recognised MPEG standards. The TV signal has a baseband signal of a few Mbit/s, and hence the transmission of digital television is possible without requiring a huge amount of radio-frequency spectrum (see Chapter 4).

At the same time, standards for digital video broadcasting (DVB), in particular the satellite version (DVB-S), have been adopted, making satellite broadcasting (direct broadcast television) a successful commercial service.

Publication of ITU-R Recommendation BT.709-1 in November 1993 signalled the development of HDTV. BT.709-6 is the latest edition of the HDTV standard, as of June 2015 [ITUR-15].

In 2012, the ITU-R published Recommendation ITU-R BT.2020, also known as the Ultra High Definition Television (UHDTV) standard, or 4K resolution TV, or simply 4K TV. It has been in rapid development since then. The latest version of the recommendation was published in October 2015. Now the satellite versions, DVB-S2 and DVB-S2x, have been well developed to support HDTV as well as UHDTV [DVB-17].

3.1.3.1 Luminance and chrominance components

Television signals contain three components: the *luminance* signal, which represents the image in black and white; the *chrominance* signal, which represents the colour; and the *sound*. Figure 3.3 shows how the luminance and chrominance signals are generated: the television camera produces three voltages E_R, E_B, and E_G representing the red (R), blue (B), and green (G) components

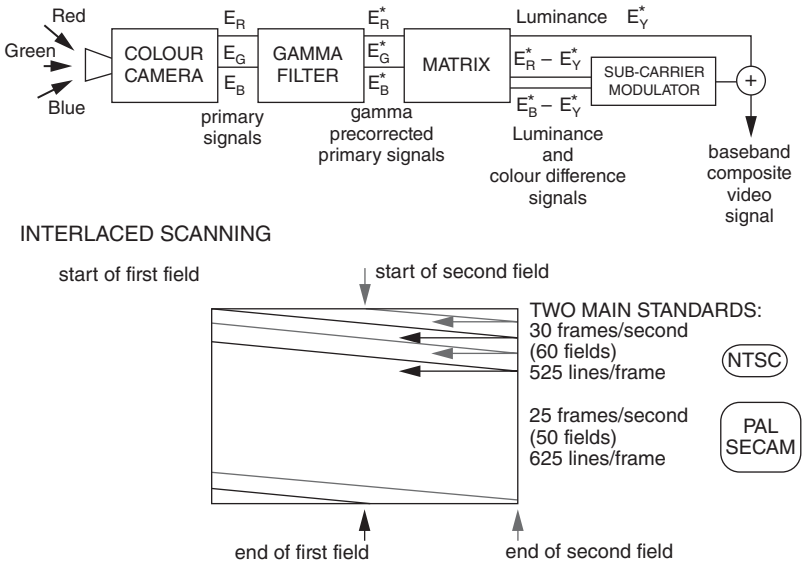


Figure 3.3 Generation of television signals.

of the colour at one point of the scanned image (525 lines per frame and 60 fields per second – that is, 30 images per second – in the NTSC standard; 625 lines per frame and 50 fields per second – that is, 25 images per second – in the PAL and SECAM standards). These signals are filtered by the gamma filters in order to compensate for the nonlinear response of the receiving cathode-ray tube and are then combined to generate the luminance signal E_{Y^*} defined by

$$E_{Y*} = 0.3E_{R*} + 0.59E_{G*} + 0.11E_{B*} \tag{3.1}$$

and the two components of the chrominance signal $E_{R^*}-E_{Y^*}$ and $E_{B^*}-E_{Y^*}$, which contain the information required to reconstruct the components of the original colour signal.

3.1.3.2 NTSC, PAL, and SECAM colour television signals

The composite video signal is formed by summing the E_{Y^*} signal and a subcarrier modulated by the two components of the chrominance signal (Figure 3.3). A subcarrier modulated by the sound is added to the composite video signal. The modulation techniques depend on the particular system. Figure 3.4 shows the spectrum of a television signal.

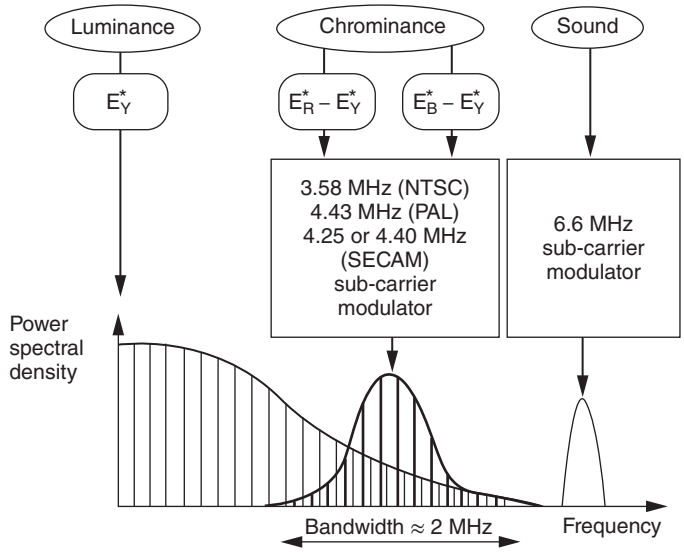


Figure 3.4 The spectrum of the composite television signal.

The composite video signal has the advantage of compatibility with a monochrome video signal but also some disadvantages:

— The receiver is unable to separate completely the luminance and chrominance components. This results in cross-colour and cross-luminance effects, whereby rapid variations of luminance are interpreted by the receiver as variations in colour and striations are introduced into the image.
— The sound does not have the quality of digital sound to which the viewer is now accustomed (with digital audio recording or compact disc).

Moreover, television broadcasters have been pushing toward the promotion of high-definition television, delivering television pictures with a subjective quality similar to that of a projected 35 mm theatrical film. This implies a wider aspect ratio (16/9 instead of 4/3), higher resolution (1125 lines per frame and 60 fields per second or 1250 lines per frame and 50 fields per second), and superior audio sound, similar to compact disc quality. These features make the signal bandwidth larger, typically 30 MHz.

To offset the inconveniences of the composite video signal and to promote high-definition television, digital standards have been proposed.

3.1.3.3 Compressed digital television signal

DVB uses MPEG-2 compression for video and either MP2 (MPEG-1 audio layer 2) or AC3 (Dolby Digital 2.0 or 5.1) for audio. Audio bitrates used are usually in the 192–256 kbit s^{-1} range for MP2 and 192–448 kbit s^{-1} for AC3.

An MPEG-2 encoder has two output options: elementary streams and programme (system) streams. *Elementary streams* display one audio (.mp2) and one video (.m2v or .mpv) file. *Programme streams* comprise a single file containing both audio and video (usually an.mpg file). In the latter format, the encoder divides audio and video into packets of a common size (the size can vary).

Each packet of such a stream (known as a packetised elementary stream PES) has an 8-byte header that consists of a 3-byte start code – one byte for the stream ID, two bytes to indicate the length of the packet, and two *timestamps*: the decoding timestamp (DTS) and the presentation timestamp (PTS). The DTS indicates when a packet has to be decoded and the PTS when the decoded packet has to be sent to the decoder output.

These timestamps allow for bidirectional encoding (*b-frames*), which requires certain frames to be decoded out of order. (As an example, b-frames reference previous and future frames and, in order to decode, both referenced frames have to be available. So if frame N references frames $N-1$ and $N+1$, and N is a b-frame, the decoder has to decode the frames in the order $N-1$, $N+1$, N and send them to the output in the order N_1, N, N þ 1 [ISO/IEC-18].)

To transmit the various components (video, sound, and data) of one or several TV programmes, a specific data structure has been defined: *transport streams* (TSs). A transport stream can convey multiple TV chains (the full set is called a *bouquet*), each encoded at different bit rates and having different timestamps (in contrast, a programme stream allows for only one video stream).

The TS packets are all 188 bytes. The first four bytes are used by a header, which contains a transport error indicator, packet identification, some scrambling information (for scrambled TV channels), a continuity counter (which allows the decoder to determine if a packet has been omitted, repeated, or transmitted out of sequence), and some more specific fields that can be used for applications dealing with TS streams. To have a common clock (ticking at 27 MHz), the adaptation field is periodically used to insert a global timestamp (known as a *program clock reference* [PCR]).

In order to identify which packets belong to which TV programme, additional information is needed: the program-specific information (PSI) is used to tell a decoder which packets belong together (video, audio, and additional data such as subtitles, teletext, etc.). Based on the PSI, a decoder can extract the packet identifiers (PIDs) specific to a certain TV channel and decode only those relevant to this channel.

Digital broadcasting has different distribution and transmission requirements, as shown in Figure 3.5 (FEC stands for *forward error correction*). Broadcasters produce transport streams that contain several television programmes. Transport streams have no protection against errors, and

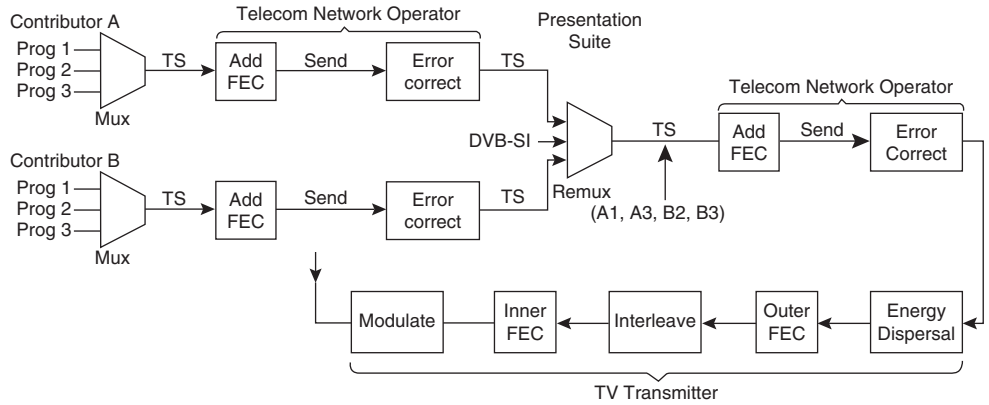


Figure 3.5 Functional blocks of programme multiplexing and transmission system.

in compressed data the effect of errors may be severe. Transport streams need to be delivered quasi-error-free (QEF) (i.e. typically 10^{-11} BER) to transmitters, satellite uplinks, and cable head ends. This task is normally entrusted to telecommunications network operators, which use an additional layer of error correction as needed. This layer should be transparent to the destination.

A particular transmitter or cable operator may not want all of the programmes in a transport stream. Several transport streams may be received, and a selection of channels may be made and encoded into a single output transport stream using a re-multiplexer. The configuration may change dynamically. Service information (DVB-SI) that is encoded into the transport stream is metadata describing the transmission, including details of programmes carried on other multiplexes and services such as teletext [ETSI-16; ETSI-09].

Broadcasting requires some form of standardised FEC so that receivers can handle it. The addition of error correction increases the bit rate as far as the transmitter or cable is concerned.

3.1.4 Data and multimedia signals

Data is becoming the most common vehicle for information transfer related to a large variety of services, including voice telephony, video, and computer-generated information exchange. One of the most appealing aspects of data transmission is the ability to combine, onto a single transmission, support data generated by several individual sources, resulting in a single data stream called *aggregate traffic*. This is paramount for transfer of multimedia traffic integrating voice, video, and application data. Traffic like this often displays a reduced bit-rate variance compared to the traffic generated by the individual sources, due to the embedded statistical multiplexing. A network operator uses this opportunity to dimension its links (and particularly satellite links) with a capacity that is less than the sum of the peak bit rate of the individual sources. This dimensioning considers both the burstiness of the traffic and the multiplexing techniques in use.

The traffic is typically transported by packets. The data structure (transport stream) developed for conveying TV programme components could be used to carry any type of data, taking advantage of the well-recognised standard and mass-market production of equipment. For example, MPEG-2 transport stream (MPEG-TS) packets are used with DVB-S data transmission. As for video programmes, MPEG-2 transport stream (MPEG-TS) packets have a fixed size of 188 bytes, of which 4 bytes are used for the packet header and 144 bytes for the payload. The header consists of a synchronisation (sync) byte, a PID, a transport error indication, and a field of adaptation options.

The asynchronous transfer mode (ATM) data format is used in high-data-rate, terrestrial networks. The ATM packet, called an *ATM cell*, is a fixed size: 53 bytes, of which 5 bytes are used for the header and 48 bytes for the payload. The header consists of a virtual channel identifier (VCI), a virtual path identifier (VPI), a payload type, and priority and header error check (HEC) fields. As an MPEG-TS option, ATM cells are used in the DVB return channel via satellite (DVB-RCS).

3.2 PERFORMANCE OBJECTIVES

The performance objectives have been established by ITU recommendations. The quality of the baseband signal is defined depending on the considered service at the user–terminal interface or at the interface between a satellite network and a terrestrial network. The *baseband signal to baseband noise power ratio* (S/N) is the basic parameter for analogue signals. The *bit error rate* (BER) is the basic parameter for digital signals.

This book considers only digital signals and techniques. For digital signals, the BER conditions other types of objectives, such as errored seconds or error-free seconds, used to give a better appreciation of the quality of the delivered service [ITUR-05a; ITUT-88d].

3.2.1 Telephone

ITU-R stipulates that the BER must not exceed [ITUR-94]:

— One part in 10^6, 10-minute mean value for more than 20% of any month
— One part in 10^4, 1-minute mean value for more than 0.3% of any month
— One part in 10^3, 1-second mean value for more than 0.05% of any month

3.2.2 Sound

When a sound programme is transmitted by satellite in digital form, the performance objective is stipulated in terms of BER. Errors have the effect of generating audible clicks. To limit their frequency to about one per hour, a BER on the order of 10^{-9} is required [CCIR-90].

3.2.3 Television

In the DVB-S standard, the performance objectives are to achieve QEF transmission of BER from 10^{-10} to 10^{-11} after the Reed-Solomon inner code and BER of 2×10^{-4} after the Viterbi outer code [ETSI-97]. This imposes the ratio of energy per information bit to power spectral density of noise, E_b/N_0, to be equal to or less than 4.5 dB for the inner code rate of 1/2, less than 5.0 dB for 2/3, 5.5 dB for 3/4, 6.0 dB for 5/6, and 6.4 for 7/8. Margins of 0.8 dB for modem implementation and 0.36 dB for the noise bandwidth may be added.

3.2.4 Data

ITU-R stipulates that the BER for satellite transmission of data at 64 kbit s^{-1} at a frequency below 15 GHz on a link that is part of an integrated services digital network (ISDN) must not exceed [ITUR-05a]:

— 10^{-7} during more than 10% of any month
— 10^{-6} during more than 2% of any month
— 10^{-3} during more than 0.03% of any month

For higher bit rates, refer to [ITUT-02].

Packet-transfer performance parameters can be estimated on the basis of observations of ATM cell transfers (Table 3.2):

— Cell transfer delay (CTD) applies to successfully transferred cells and is the time between the occurrences of two corresponding cell-transfer points. It consists of two parameters: mean CTD and cell delay variation (CDV).
— Cell loss ratio (CLR) is the ratio of total lost cells to total transmitted cells in a population of interest.
— Cell error ratio (CER) is the ratio of total errored cells to the total of successfully transferred cells, plus tagged cells and errored cells in a population of interest. Successfully transferred cells, tagged cells, and errored cells contained in severely errored cell blocks are excluded from the calculation of CER.
— Cell misinsertion rate (CMR) is the total number of misinserted cells observed during a specified time interval divided by the duration of the time interval (or the number of misinserted cells per connection second). Misinserted cells and time intervals associated with severely errored cell blocks are excluded from the calculation of CMR.
— Severely errored cell block ratio (SECBR) is the ratio of severely errored cell blocks to total cell blocks in a population of interest.

The values in Table 3.2 are provisional; they need not be met by networks until they are revised (up or down) based on real operational experience. The objectives apply to public B-ISDNs and are believed to be achievable on 27 500 km hypothetical reference connections.

3.3 AVAILABILITY OBJECTIVES

Availability is the fraction of time during which a service conforming to the specifications is provided. It is affected by both equipment breakdown and propagation phenomena.

The ITU-R stipulates that the unavailability for telephony must not exceed [ITUR-05b]:

— 0.2% of a year in the case of breakdown (interruption of the service must be less than 18 h y^{-1})
— 0.2% of any month if the service interruption is due to propagation

The effects of propagation on the quality of the link are examined in Chapter 5. Breakdowns may involve both earth station and satellite equipment. For earth stations, service interruption due to conjunction of the station, the satellite, and the sun is regarded as a breakdown.

From a dependability point of view [ITUT-00], a portion of an international B-ISDN ATM semi-permanent connection has the following properties of availability:

— The fraction of time during which it is in a down state (i.e. unable to support a transaction) should be as low as possible.
— Once a transaction has been established, it should have a low probability of being terminated (because of insufficient data-transfer performance) or prematurely released (due to the failure of a network component) before the intended end of transaction.

Table 3.2 Provisional QoS class definitions and network performance objectives

	CTD Upper bound on the mean CTD	Two-point CDV Upper bound on the difference between the upper and lower 10^{-8} quantiles of CTD	CLR_{0+1} Upper bound on the cell loss probability	CLR_0 Upper bound on the cell loss probability	CER Upper bound on the cell error probability	CMR Upper bound on the mean CMR	SECBR Upper bound on the SECB probability
Default objectives	—	—	—	—	4×10^{-6}[a]	1/day[b]	10^{-4}[c]
QoS class 1 (stringent class)	400 ms[d,e]	3 ms[f]	3×10^{-7}[g]	—	Default	Default	Default
QoS class 2 (tolerant class)	U[h]	U	10^{-5}	—	Default	Default	Default
QoS class 3 (bi-level class)	U	U	U	10^{-5}	Default	Default	Default
QoS class 4 (U class)	U	U	U	U	U	U	U
QoS class 5 (stringent bi-level class)	400 ms[d]	6 ms[f,i]	—	3×10^{-7}	Default	Default	Default

[a] In the near future, networks may be able to commit to a CER of 4×10^{-7}. This is a subject for further study.
[b] Some network phenomena have been observed that tend to increase the CMR as the cell rate of the virtual connection increases. More complete analyses of these phenomena may ultimately suggest a larger CMR objective for high-bit-rate connections.
[c] The SECBR is sensitive to short (i.e. 2–9-second) interruptions in the cell stream. They will give rise to many SECBs and may make the SECBR objective difficult to meet.
[d] See [ITUT-03] for further guidance on the delay requirements of some applications.
[e] Some applications may require performance similar to QoS class 1 but do not require a CTD commitment. These applications can make use of QoS class 1, but the need for a new QoS class is a subject for further study.
[f] Two-point CDV applies when there are no more than nine ATM nodes in the connection with 34–45 Mbit s^{-1} output links and all other ATM nodes are opera ting at 150 Mbit s^{-1} or higher. Two-point CDV will generally increase as transport rates decrease. High-bit-rate DBR connections may need and may receive less CDV. This is for further study.
[g] In the near future, networks may be able to commit to a CLR for QoS class 1 of 10^{-8}. This is a subject for further study.
[h] *UI* means *unspecified* or *unbounded*: the ITU-T establishes no objective for this parameter, and any default objective can be ignored.
[i] It is not certain that the applications that choose QoS class 5 require a 6 ms bound for CDV or that achieving it will be economically justifiable. This objective requires further study.

Availability of a semi-permanent connection portion is defined as the fraction of time during which the portion is able to support a transaction. Conversely, unavailability of a portion is the fraction of time during which the portion is unable to support a transaction (i.e. it is in the down state). A common availability model, used in [ITUT-00], applies to any semi-permanent connection type. The model uses two states corresponding to the ability or inability of the network to sustain a connection in the available state. Transitions between the states of the model are governed by the occurrence of patterns of severely errored seconds. The availability is from the network perspective, where performance is characterised independently of user behaviour.

[ITUT-00] defines two availability performance parameters:

— Availability ratio (AR) applies to semi-permanent connection portions. The AR is defined as the proportion of scheduled service time that the connection portion is in the available state. The AR is calculated by dividing the total service available time by the duration of the scheduled service time. During the scheduled service time, the user may or may not transmit cells.
— Mean time between outages (MTBO) applies to semi-permanent connection portions. The MTBO is defined as the average duration of continuous periods of available time. Where scheduled service times are not contiguous, they are concatenated in calculating MTBO.

As far as a satellite is concerned, it is necessary to consider its reliability, which is determined by breakdowns of on-board equipment, breaks during an eclipse if the only source of power for on-board equipment is solar power, and the lifetime of the satellite. In general, an operational satellite, a back-up satellite in orbit, and a back-up satellite on the ground are provided. Availability depends also on the reliability of launchers, which are indispensable for replacement of satellites at the end of their life. The approach to these problems is treated in more detail in Chapter 13.

3.4 DELAY

Delay builds up from sending user terminal to destination user terminal on account of:

— Delay in the terrestrial network, if any
— Propagation delay over satellite links
— Baseband-signal processing time
— Protocol-induced delay

3.4.1 Delay in the terrestrial network

This delay, including switching and propagation time, can be estimated from the following formula:

$$t_{\mathrm{TN}}(\mathrm{ms}) = 12 + 0.004 \times \mathrm{distance\ (km)} \tag{3.2}$$

3.4.2 Propagation delay over satellite links

Over any radio-frequency up- or downlink and radio-frequency or optical intersatellite link, the propagation delay is given by:

$$t_{\mathrm{SL}} = R/c \quad \mathrm{(s)}$$

where R is the range from the transmission equipment to the receiving equipment, and c is the speed of light ($c = 3 \times 10^8$ m s^{-1}). The overall propagation delay builds up from the delays in the individual links that constitute the overall link from earth station to earth station.

For a geostationary satellite with no intersatellite link, the minimum and maximum values of the overall delay, often called the *hop delay*, are calculated as follows. The minimum value is when both end earth stations are located at the sub-satellite point:

$$R_U = R_D = R_0 = 35786 \text{ km}$$

where R_0 is the geostationary satellite altitude. Thus the overall delay is 238 ms. The maximum value is when both end earth stations are located at the edge of coverage with elevation angle 0°:

$$R_U = R_D = (R_0 + R_E) \cos (17.4^\circ/2)$$

where R_E is the earth radius ($R_E = 6378$ km). Thus the overall delay is 278 ms.

3.4.3 Baseband-signal processing time

This delay, which results from baseband-signal processing in the earth stations and on board regenerative satellites (there is no processing on board transparent satellites), depends on the type of process. There may be delay in the source encoders as a result of information compression; in the implementation of the multiplexing, demodulation, and decoding processes; and in the buffering associated with switching and multiple access.

3.4.4 Protocol-induced delay

Error-free delivery of data packets implies the use of automatic repeat request (ARQ) protocols (e.g. Transmission Control Protocol [TCP] congestion and flow controls) for retransmission of unacknowledged messages. Another cause of delay is temporary congestion, which affects mostly data transfers under the 'best effort' type of service.

For telephony services, the ITU-T stipulates that the transmission delay between subscribers must not exceed 400 ms. It recommends the use of echo suppressors or echo cancellers when this time is between 150 and 400 ms.

For connections between subscribers established with geostationary satellites, this leads to the following requirements:

— Installation of echo suppressors or cancellers (see Chapter 8).
— Avoidance of *double hops* (links through two satellites without an intersatellite link). If the system contains an intersatellite link, the propagation time t_{ISL} between the two satellites must remain less than 90 ms. The orbital separation between the two satellites for propagation time $t_{ISL} = R_{ISL}/c$ is given by the following expression:

$$\theta = 2 \text{ arc sin } ct_{ISL}/2(R_E + R_0) \tag{3.3}$$

With $R_E = 6378$ km, $R_0 = 35\,786$ km, and $t_{ISL} < 90$ ms, this gives $\vartheta < 37°$.

3.5 IP PACKET TRANSFER QOS AND NETWORK PERFORMANCE

Today, information services and applications have moved toward all-IP solutions, including telephony, TV broadcasting, mobile communications, Internet of Thing (IoT), etc. Therefore, it is important to cover these aspects and recent new developments and related standards. As this book is focused on communications systems, techniques, and technology, we only cover this topic here at an introductory level. Chapter 7 provides more detail about satellite networking and related principles and protocols. Further details can also be found in the textbook *Satellite Networking – Principles and Protocols* [SUN-14].

3.5.1 Definition of QoS in the ETSI and ITU-T standards

Quality of service (QoS) is a well-studied topic in academics and industries as well as by standardisation bodies. Here, we provide the explanation as defined in the ETSI standard [ETSI-15; ETSI-03] and ITU-T standards [ITUT-92]:

> Quality of Service (QoS) – IETF definition: The ability to segment traffic or differentiate between traffic types in order for the network to treat certain traffic differently from others. QoS encompasses both the service *categorization* and the overall performance of the network for each category. It also refers to the capability of a network to provide better service to selected network traffic over various technologies and IP-routed networks that may use any or all of the underlying technologies.

> Quality of Service (QoS) – ITU definition: QoS is defined as the collective effect of service performances which determines the degree of satisfaction of a user of a service. It is characterized by the combined aspects of performance factors applicable to all services, such as: service operability performance; service accessibility performance; service retainability performance; service integrity performance and other factors specific to each service.

Related to QoS are some of its parameters, which can be specified or monitored to ensure QoS. In service levels of QoS, it considers the end-to-end QoS capabilities of the network that enable it to deliver a service needed by a specific mix of network traffic. A service-level agreement (SLA) is used to describe the agreement between a service provider (SP) and its subscriber (or between an SP and an access network operator), characterised by the choice of one data-transfer capability and the allocation attribute related to this transfer capability.

For satellite networks, the ETSI has recommended that the broadband satellite multimedia (BSM) QoS architecture should support the following service requirements [ETSI-15]:

- Compatibility with end-to-end IP network QoS parameters, services, and mechanisms within integrated networks.
- Satisfying the QoS requirements of IP flows across the integrated network as determined by SLAs.
- Support for control of both relative QoS and guaranteed QoS.
- The use of BSM traffic classes to define the QoS properties for the transport of IP packets across the BSM subnetwork.
- The mapping of the QoS attributes of IP packets to and from satellite independent – service access point (SI-SAP) QoS attributes at the BSM subnetwork edges.

3.5.2 IP packet transfer performance parameters

A *network section ensemble* (NSE) refers to any connected subset of networks together with all of the exchange links (ELs) that interconnect them. Pairs of distinct NSEs are connected by ELs. The term *NSE* can also be used to represent the entire end-to-end IP network. NSEs are delimited by measurement points (MPs).

The performance of any given NSE is measurable relative to any given unidirectional end-to-end IP service. The *ingress MPs* are the set of MPs crossed by packets from that service as they go into that NSE. The *egress MPs* are the set of MPs crossed by packets from that service as they leave that NSE.

The Internet Protocol packet transfer reference events (IPREs) is defined in [ITUT-16] to refer to a specified end-to-end IP service. An IP packet-transfer event occurs when: (i) an IP packet crosses a MP, (ii) standard IP procedures applied to the packet verify that the header checksum is valid, and (iii) the source and destination address fields within the IP packet header represent the IP addresses of the expected source host address (SRC) and destination host address (DST).

Four types of IP packet-transfer events are defined:

- *IP packet entry event into a host*. Occurs when an IP packet crosses an MP entering a host (NS router or DST) from the attached EL.
- *IP packet exit event from a host*. Occurs when an IP packet crosses an MP exiting a host (NS router or SRC) into the attached EL.
- *IP packet ingress event into a basic section or NSE*. Occurs when an IP packet crosses an ingress MP into a basic section or an NSE.
- *IP packet egress event from a basic section or NSE*. Occurs when an IP packet crosses an egress MP out of a basic section or an NSE.

[ITUT-16] has provided a recommendation on a set of IP packet information transfer performance parameters on the basis of observations made at MP for qualifying the applicability of performance parameters to sets of packet in IP networks. The performance parameters are defined for IP packets from source (SRC) to destination (DST). The measurement can be carried out in the MPs within the networks and edge of the networks as well as at the SRC and DST of the packets.

3.5.2.1 Packet flow

A *packet flow* is the set of packets associated with a given connection or connectionless stream having the same SRC, DST, class of service, and session identification (e.g. port numbers from a higher-layer protocol). Other documents may use the terms *microflow* or *subflow* when referring to packet streams with this degree of classification. A packet flow is the most common example of a population of interest.

IPv6 packets have an additional field for the source host to label sequences of packets that should receive some special treatment in IPv6 routers. This field is called the *flow label* and, in combination with the source address, uniquely defines a packet flow.

3.5.2.2 IP packet-transfer delay

IP packet-transfer delay (IPTD) is defined for all successful and errored packet outcomes across a basic section or an NSE. IPTD is the time, $(t_2 - t_1)$ between the occurrence of two corresponding

IP packet reference events, ingress event IPRE-1 at time t_1 and egress event IPRE-2 at time t_2, where $(t_2 > t_1)$ and $(t_2 - t_1) \leq \text{Tmax}$. If the packet is fragmented within the NSE, t_2 is the time of the final corresponding egress event. The end-to-end IPTD is the one-way delay between the MP at the SRC and DST.

Measurements of IP packet delay consist of the following parameters:

- *Mean IPTD*. The arithmetic average of IPTDs for a population of interest.
- *Minimum IPTD*. The smallest value of IPTD among all IPTDs of a population of interest. This includes propagation delay and queuing delays common to all packets. Therefore, this parameter may not represent the theoretical minimum delay of the path between MP.
- Median IPTD: The 50th percentile of the frequency distribution of IPTDs from a population of interest. The median is the middle value once the transfer delays have been rank-ordered. To obtain this middle value when the population contains an even number of values, the mean of the two central values is used.
- *End-to-end two-point IP packet delay variation (PDV)*: The variations in IPTD are also important. Streaming applications might use information about the total range of IP delay variation to avoid buffer underflow and overflow. Extreme variations in IP delay will cause TCP retransmission timer thresholds to grow and may also cause packet retransmissions to be delayed or cause packets to be retransmitted unnecessarily. The end-to-end two-point IP PDV is defined based on the observations of corresponding IP packet arrivals at ingress and egress MP (e.g. MPDST, MPSRC). These observations characterise the variability in the pattern of IP packet arrival events at the egress MP and the pattern of corresponding events at the ingress MP with respect to a reference delay.

The two-point PDV (v_k) for an IP packet k between SRC and DST is the difference between the absolute IPTD (x_k) of packet k and a defined reference IPTD, $d_{1,2}$, between those same MPs: $v_k = x_k - d_{1,2}$.

The reference IPTD, $d_{1,2}$, between SRC and DST is the absolute IPTD experienced by a selected IP packet between those two MPs.

Positive values of two-point IP packet delay variation (IPDV) correspond to IPTDs greater than those experienced by the reference IP packet; negative values of two-point PDV correspond to IPTDs less than those experienced by the reference IP packet. The distribution of two-point PDVs is identical to the distribution of absolute IPTDs displaced by a constant value equal to $d_{1,2}$.

3.5.2.3 IP packet error ratio (IPER)

IP packet error ratio (IPER) is the ratio of total errored IP packet outcomes to the total of successful IP packet transfer outcomes plus errored IP packet outcomes in a population of interest.

3.5.2.4 IP packet loss ratio (IPLR)

IP packet loss ratio (IPLR) is the ratio of total lost IP packet outcomes to total transmitted IP packets in a population of interest.

3.5.2.5 Spurious IP packet rate

Spurious IP packet rate at an egress MP is the total number of spurious IP packets observed at that egress MP during a specified time interval divided by the time interval duration (equivalently, the number of spurious IP packets per service-second)1.

3.5.2.6 IP packet reordered ratio (IPRR)

IP packet reordered ratio (IPRR) is the ratio of the total reordered packet outcomes to the total of successful IP packet transfer outcomes in a population of interest.

3.5.2.7 IP packet severe loss block ratio (IPSLBR)

IP packet severe loss block ratio (IPSLBR) is the ratio of the IP packet severe loss block outcomes to total blocks in a population of interest.

3.5.2.8 IP packet duplicate ratio (IPDR)

IP packet duplicate ratio (IPDR) is the ratio of total duplicate IP packet outcomes to the total of successful IP packet transfer outcomes minus the duplicate IP packet outcomes in a population of interest.

3.5.2.9 Replicated IP packet ratio (RIPR)

Replicated IP packet ratio (RIPR) is the ratio of total replicated IP packet outcomes to the total of successful IP packet transfer outcomes minus the duplicate IP packet outcomes in a population of interest.

3.5.3 IP service availability parameters

3.5.3.1 Percent IP service unavailability (PIU)

Percent IP service unavailability (PIU) is the percentage of total scheduled IP service time (the percentage of Tav intervals) that is (are) categorised as unavailable using the IP service availability function.

3.5.3.2 Percent IP service availability (PIA)

Percent IP service availability (PIA) is the percentage of total scheduled IP service time (the percentage of Tav intervals) that is (are) categorised as available using the IP service availability function. PIU and PIA are related as:

$$\text{PIU} = 100\text{–PIA}$$

Because the IPLR typically increases with increasing offered load from SRC to DST, the likelihood of exceeding the threshold increases with increasing offered load. Therefore, PIA values are likely to be smaller when the demand for capacity between SRC and DST is higher.

3.5.4 IP network QoS class

The latest revision of [ITUT-11] was published on 14 December 2011. It provides IP network QoS class definitions and network performance objectives as shown in Table 3.3.

Table 3.3 IP network QoS class definitions and network performance objectives [SUN-14; ITUT-11]

Network performance parameter	Nature of network performance objective	QoS classes					
		Class 0	Class 1	Class 2	Class 3	Class 4	Class 5 Unspecified
IPTD	Upper bound of the mean IPTD[a]	100 ms	400 ms	100 ms	400 ms	1 s	U
IPDV	Upper bound of the $1–10^{-3}$ quantile of IPTD minus the minimum IPTD[b]	50 ms[c]	50 ms[c]	U	U	U	U
IPLR	Upper bound of the packet loss probability	1×10^{-3}[d]	1×10^{-3}[d]	1×10^{-3}	1×10^{-3}	1×10^{-3}	U
IPER	Upper bound	1×10^{-4}[e]					U

General notes: The objectives apply to public IP networks. The objectives are believed to be achievable on common IP network implementations. The network providers' commitment to the user is to attempt to deliver packets in a way that achieves each of the applicable objectives. The vast majority of IP paths advertising conformance with Recommendation ITU-T Y.1541 [ITUT-11] should meet those objectives. For some parameters, performance on shorter and/or less-complex paths may be significantly better.
An evaluation interval of one minute is provisionally suggested for IPTD, IPDV, and IPLR, and in all cases the interval is be reported.
Individual network providers may choose to offer performance commitments better than these objectives.
U means *unspecified* or *unbounded*. When performance relative to a particular parameter is identified as being U, the ITU-T establishes no objective for this parameter, and any default Y.1541 objective can be ignored. When the objective for a parameter is set to U, performance with respect to that parameter may, at times, be arbitrarily poor.
All values are provisional, and they need not be met by networks until they are revised (up or down) based on real operational experience.
[a]Very long propagation times will prevent low end-to-end delay objectives from being met. In these and some other circumstances, the IPTD objectives in Classes 0 and 2 will not always be achievable. Every network provider will encounter these circumstances, and the range of IPTD objectives in Table 1 provides achievable QoS classes as alternatives. The delay objectives of a class do not preclude a network provider from offering services with shorter delay commitments. According to the definition of IPTD in Recommendation ITU-T Y.1541 [ITUT-11], packet insertion time is included in the IPTD objective. This Recommendation suggests a maximum packet information field of 1500 bytes for evaluating these objectives.
[b]The definition and nature of the IPDV objective is under study. See Appendix II of Recommendation ITU-T Y.1541 [ITUT-11] for more details.
[c]This value is dependent on the capacity of internetwork links. Smaller variations are possible when all capacities are higher than the primary rate (T1 or E1) or when competing packet information fields are smaller than 1500 bytes (see Appendix IV of Recommendation ITU-T Y.1541 [ITUT-11]).
[d]The Class 0 and 1 objectives for IPLR are partly based on studies showing that high-quality voice applications and voice codecs will be essentially unaffected by a 10^{-3} IPLR.
[e]This value ensures that packet loss is the dominant source of defects presented to upper layers, and is feasible with IP transport on ATM.

3.6 CONCLUSION

This chapter has presented techniques used for generation of baseband information signals intended for modulation of the transmitted radio-frequency carrier. It has also introduced the concept of QoS in terms of performance objectives (BER) and availability objectives. The techniques that enable these baseband digital signals to be conveyed through modulation and coding are presented in Chapter 4, along with the useful relations that express the delivered QoS as a function of the radio-frequency link performance C/N_0 where C is the carrier power and N_0 is the power spectral density of noise. For packet networks, all the QoS and performance measurements are carried out IP packets at the IP network level. All of these have been well developed with new techniques and standards for both satellite and terrestrial networks. Development has also been synchronised between satellite and terrestrial networks as well as mobile 4G/5G networks in recent years.

REFERENCES

[CAM-76] Campanella, S.J. (1976). Digital speech interpolation. *COMSAT Technical Review* **6** (1): 127–157, Spring.

[CCIR-90] CCIR. (1990). Digital sound-programme transmission impairments and methods of protection against them. Report 648.

[DVB-17] Digital Video Broadcasting Project. (2017). Digital video broadcasting (DVB); specification for the use of video and audio coding in broadcast and broadband applications. DVB Document A001.

[ETSI-97] ETSI. (1997). Digital video broadcasting (DVB); framing structure, channel coding and modulation for 11/12 GHz satellite services. EN 300 421 (V1.1.2).

[ETSI-03] ETSI. (2003). Satellite earth stations and systems (SES); broadband satellite multimedia; IP interworking over satellite; performance, availability and quality of service. TR 102 157 (V1.1.1).

[ETSI-07] ETSI. (2007). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); services and architectures. TR 101 984 (V1.2.1).

[ETSI-09] ETSI. (2009). Digital video broadcasting (DVB); guidelines on implementation and usage of service information (SI). TR 101 211.

[ETSI-15] ETSI. (2015). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); QoS functional architecture. TS 102 462 (V1.2.1).

[ETSI-16] ETSI. (2016). Digital video broadcasting (DVB); specification for service information (SI) in DVB systems. EN 300 468 (V1.15.1).

[FRE-91] Freeman, R. (1998). *Telecommunications Transmission Handbook*, 4e. Wiley.

[ISO/IEC-18] ISO/IEC. (2018). Information technology -- generic coding of moving pictures and associated audio information -- part 1: systems. 13818-1.

[ITUR-94] ITU. (1994). Allowable bit error ratios at the output of the hypothetical reference digital path for systems in the fixed-satellite service using pulse-code modulation for telephony. Recommendation S.522.

[ITUR-86] ITU-R (1986) Digital PCM coding for the emission of high-quality sound signals in satellite broadcasting (15 kHz nominal bandwidth). Recommendation BO.651.

[ITUR-05a] ITU-R. (2005). Allowable error performance for a satellite hypothetical reference digital path in the fixed-satellite service operating below 15 GHz when forming part of an international connection in an integrated services digital network. Recommendation S.614.

[ITUR-05b] ITU-R (2005) Availability objectives for a hypothetical reference circuits and hypothetical reference digital paths when used for telephony using pulse code modulation, or as part of an integrated services digital network hypothetical reference connection, in the fixed-satellite service operating below 15 GHz. Recommendation S.579–6.

[ITUR-15] ITU-R (2015) Parameter values for the HDTV standards for production and international programme exchange. Recommendation BT.709–6.

[ITUT-88a] ITU-T. (1988). Pulse code modulation (PCM) of voice frequencies. Recommendation G.711.
[ITUT-88b] ITU-T. (1988). Digital hierarchy bit rates. Recommendation G.702.
[ITUT-88c] ITU-T. (1988). Synchronous frame structures used at 1544, 6312, 2048, 8448 and 44 736 kbit/s hierarchical levels. Recommendation G.704.
[ITUT-88d] ITU-T. (1988). General quality of service parameters for communication via public data networks. Recommendation X.140.
[ITUT-96] ITU-T. (1996). Characteristics of a flexible multiplexer in a synchronous digital hierarchy environment. Recommendation G.785.
[ITUT-00] ITU-T. (2000). B-ISDN semi-permanent connection availability. Recommendation I.357.
[ITUT-02] ITU-T. (2002). End-to-end error performance parameters and objectives for international, constant bit-rate digital paths and connections. Recommendation G.826.
ITU-T. (2003). One-way transmission time. Recommendation G.114.
[ITUT-10] ITU-T. (2010). Terms and definitions for synchronous digital hierarchy (SDH) networks. Recommendation G.780.
[ITUT-11] ITU-T. (2011). Network performance objectives for IP-based services. Recommendation Y.1541.
[ITUT-16] ITU-T. (2016). Internet protocol data communication service – IP packet transfer and availability performance parameters. Recommendation Y.1540.
[ITUT-17] ITU-T. (2017). Synchronization layer functions. Recommendation G.781.
[SAR-94] Sari, H. and Jeanclaude, I. (1994). An analysis of orthogonal frequency-division multiplexing for mobile radio applications. In: *1994 IEEE VTC'94, Stockholm, Sweden, June 1994*, 1635–1639. IEEE.
[SCH-80] Schwartz, M. (1980). *Information, Transmission, Modulation and Noise*. McGraw-Hill.
[SUN-14] Sun, Z. (2014). *Satellite Networking: Principles and Protocols*, 2e. Wiley.

4 DIGITAL COMMUNICATIONS TECHNIQUES

This chapter examines the techniques that enable baseband signals to be carried over a distance, adapting the characteristics of the transmitted signals to the constraints of the satellite communications channel. Basically, these constraints are power and bandwidth. It is shown how those resources can be traded-off against each other. This is an important aspect of satellite communications as power impacts both satellite mass and earth station (ES) size, and bandwidth is constrained by regulations. The objective is to achieve an optimum trade-off, giving the maximum capacity at minimum system cost.

Figure 4.1 is an excerpt from Figure 1.1 representing the basic communications functions in the transmitting and receiving earth stations. Similar functions are implemented on board a regenerative satellite to communicate with earth stations or other satellites over intersatellite links.

Figure 4.2 provides some insight into Figure 4.1 for digital transmission [BOU-87]. Source encoding and time division multiplexing (TDM) are discussed in Section 3.1.1. This chapter discusses the following typical functions implemented for transmission of digital signals:

— Baseband processing or formatting
— Digital modulation and demodulation
— Channel coding and decoding

Channel coding is presented after modulation, as the dependency of bit error rate (BER) upon the ratio of energy per bit to noise power spectral density (E_c/N_0) should be discussed first. The chapter also covers the power–bandwidth trade-off due to channel coding and the established digital transmission standards (DVB-S, DVB-S2, and DVB-S2x), and concludes with examples.

The typical performance of a digital communications system is measured by the BER, as introduced in Chapter 3. This chapter indicates how this performance depends on the radio-frequency link performance expressed by the ratio of the received carrier power, C, to the noise power spectral density, N_0. Chapter 5 introduces the concepts of effective isotropic radiated power (EIRP) and G/T, which appear in Figure 4.1, and their impact on the C/N_0 ratio.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

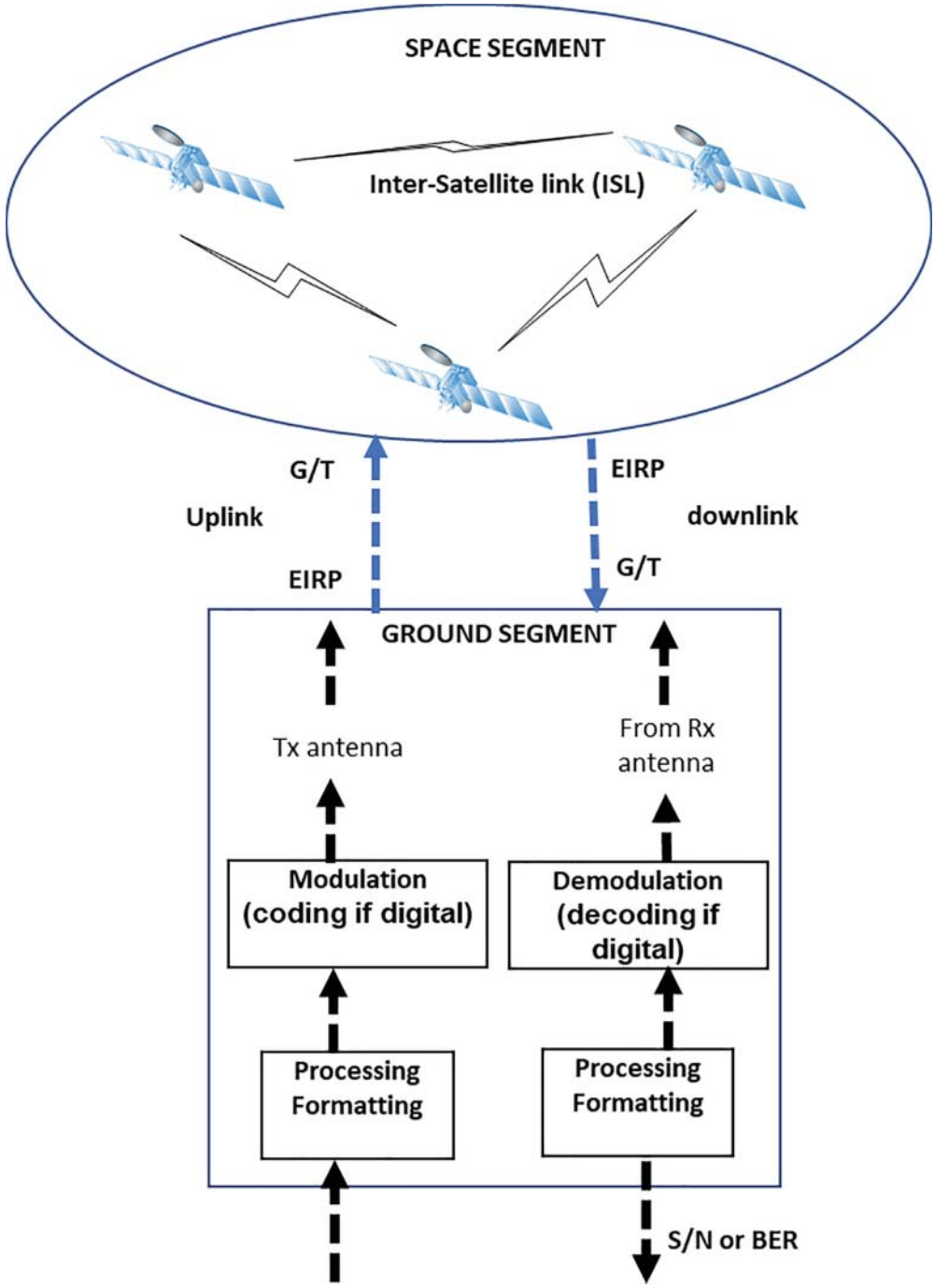


Figure 4.1 Basic communications functions in an earth station.

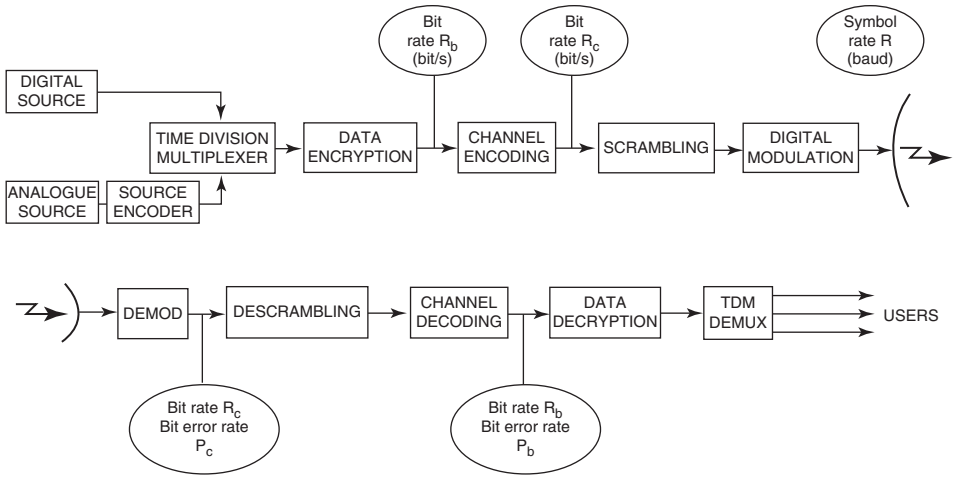


Figure 4.2 The elements of a digital transmission system.

4.1 BASEBAND FORMATTING

4.1.1 Encryption

Encryption is used when it is wished to prevent unauthorised users exploiting, or tampering with, transmitted messages. It consists of performing an algorithmic operation, in real time, bit by bit on the binary stream. The set of parameters that defines the transformation is called the *key*. Although the use of encryption is often associated with military communications, commercial satellite systems are increasingly induced by their customers to propose encrypted links for commercial and administrative networks. In fact, due to the extended coverage of satellites and easy access to them by small stations, eavesdropping, and message falsification are potentially within the reach of a large number of agents of reduced means.

Figure 4.3 illustrates the principle of encrypted transmission. The encryption and decryption units operate with a key provided by the key-generation units. Acquisition of a common key implies a secure method of key distribution.

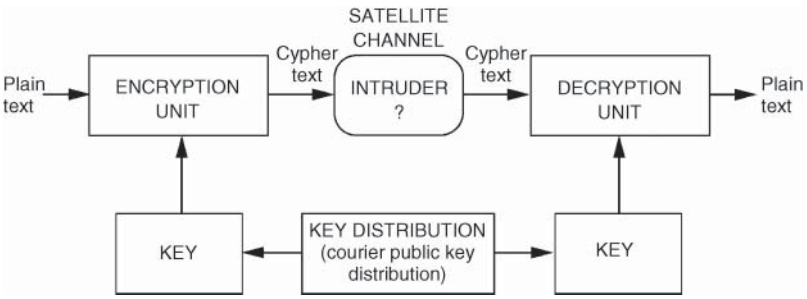


Figure 4.3 The principle of encrypted transmission.

Encryption consists of confidentiality (avoiding exploitation of the message by unauthorised people) and authenticity (providing protection against modification of the message by an intruder). Two techniques are used [TOR-81]:

— *Online encryption (stream ciphering):* Each bit of the original binary stream (the plain text) is combined using a simple operation (for example, modulo 2 addition) with each bit of a generated binary stream (the keystream) produced by a key device. The key device could be, for example, a pseudorandom sequence generator whose structure is defined by the key.
— *Encryption by block (block ciphering):* The original binary stream is transformed into an encrypted stream block by block according to logic defined by the key.

4.1.2 Scrambling

The International Telecommunication Union (ITU) recommends the use of energy-dispersion techniques (ITU-R Rec. S.446) [ITUR-93] in order to limit interference between radiocommunications systems sharing the same frequency bands. In digital transmission, when the bit stream is random, the carrier energy is spread throughout the spectrum of the modulating signal. By limiting the transmitted EIRP of the satellite, one can remain below the limit on surface power density at ground level. In contrast, if the bit stream contains a repeated fixed pattern, lines appear in the spectrum of the modulated carrier and their amplitude can lead to the limit on surface power density at ground level being exceeded. The principle of energy dispersion is to generate a modulating bit stream that has random properties regardless of the structure of the information bit stream. This operation, which is performed at the transmitter before modulation, is called *scrambling*. On reception, the inverse operation, performed after demodulation, is called *descrambling*.

Figure 4.4 shows an example of scrambler and descrambler realisation. In the scrambler, each incoming information bit is combined by modulo 2 addition with each bit generated by a pseudorandom sequence generator. The pseudorandom sequence generator consists of a shift register with various feedback paths. The descrambler contains the same pseudorandom sequence generator and, by virtue of the properties of modulo 2 addition, the combination by modulo 2 addition of the bits of the demodulated binary stream with those of the random sequence provides recovery of the information content. This implies synchronism of the two pseudorandom sequence generators. The arrangement of Figure 4.4 automatically ensures synchronism; after r bits transmitted without error, the r stages of the scrambling and descrambling shift registers are in the same state. However, an error in one bit produces as many errors in an interval of r bits as there are non-zero coefficients a_i in the feedback paths. An additional advantage provided by scrambling is suppression of sequences of logical 0s or 1s that, in non-return-to-zero level (NRZ-L) coding, can lead to a loss of synchronisation of the bit-timing recovery circuit and introduce detection errors at the demodulator output as a result of a timing error in the instant of decision. An example of implementation of scrambling is that used in the DVB-S standard (see Section 4.7).

4.2 DIGITAL MODULATION

Figure 4.5 shows the principle of a modulator. It consists of:

— A symbol generator
— An encoder or mapper
— A signal (carrier) generator

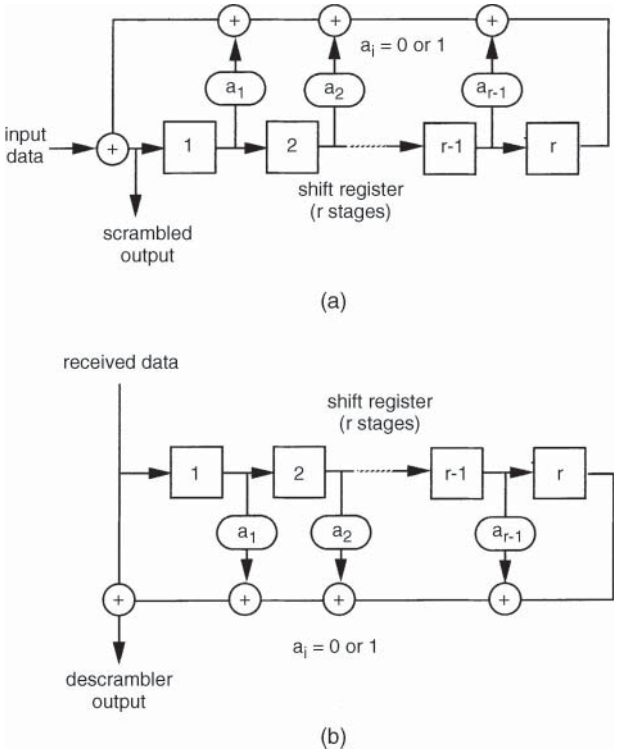


Figure 4.4 (a) A scrambler; (b) a descrambler.

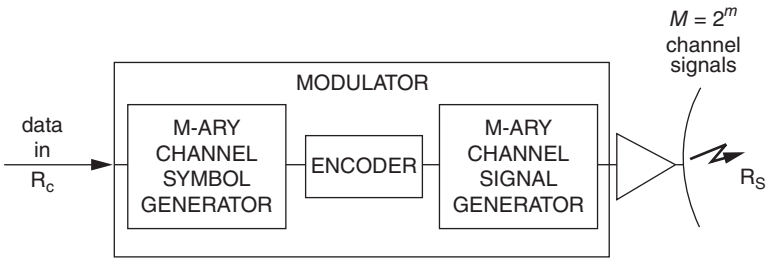


Figure 4.5 The principle of a modulator for digital transmission.

The symbol generator generates symbols with M states, where $M = 2^m$, from m consecutive bits of the input bit stream. The encoder establishes a correspondence between the M states of these symbols and M possible states of the transmitted carrier. Two types of correspondence can be considered:

— *Direct mapping:* One state of the symbol defines one state of the carrier;
— *Encoding of transitions (differential encoding):* One state of the symbol defines a transition between two consecutive states of the carrier.

For a bit rate R_c (bit/s) at the modulator input, the signalling rate R_s at the modulator output (the number of changes of state of the carrier per second) is given by:

$$R_s = R_c/m = R_c/\log_2 M \text{ (baud)} \tag{4.1}$$

Phase modulation, or *phase shift keying* (PSK), is particularly well suited to satellite links. In fact, it has the advantage of a constant envelope; and, in comparison with frequency shift keying (FSK), it provides better spectral efficiency (number of bits/s transmitted per unit of radio-frequency bandwidth – see Section 4.2.7). Depending on the number m of bits per symbol, different M-ary phase shift keying (MPSK) modulations are considered:

— The simplest form is basic two-state modulation ($M = 2$), called binary phase shift keying (BPSK) with standard direct mapping. When differential encoding is considered, it is called differentially encoded BPSK (DE-BPSK). It is of interest because it enables a simplified demodulation process (differential demodulation, see Section 4.2.6.1).
— If two consecutive bits are grouped to define the symbol, a four state modulation ($M = 4$) is defined, called quadrature phase shift keying (QPSK) with direct mapping. Differentially encoded QPSK (DE-QPSK) could be envisioned, but it is not used in practice (except for the specific case of $\pi/4$ QPSK; see Section 4.2.3.2) as differential demodulation displays significant performance degradation compared to standard coherent demodulation when M is larger than 2.
— Higher-order modulations ($M = 8$, 8PSK; $M = 16$, 16PSK; etc.) are obtained with m = 3, 4, etc. bits per symbol. As the order of the modulation increases, the spectral efficiency increases as the number of bits per symbol. On the other hand, higher-order modulations require more energy per bit (E_b) to get the same BER at the output of the demodulator (see Section 4.2.6).

With a modulation of high order (M equal to or larger than 16), better performance is achieved by considering hybrid amplitude and phase shift keying (APSK). States of the carrier correspond to given values of carrier phase and carrier amplitude (two values for 16APSK, three values for 32APSK).

4.2.1 Two-state modulation– BPSK and DE-BPSK

Figure 4.6 shows the structure of a two-state phase modulator. There is no symbol generator since the binary symbol identifies with the input bit. Let b_k be the logical value of a bit at the modulator input in the time interval $[kT_c, (k+1)T_c]$. The encoder transforms the input bit b_k into a modulating bit with logical value m_k such that:

— For direct mapping (BPSK): $m_k = b_k$.
— For differential encoding (DE-BPSK): $m_k = b_k \oplus m_{k-1}$, where $\oplus$ represents the *exclusive OR* logical operation.

The channel signal generator is controlled by the bit m_k, which is represented in the time interval $[kT_c, (k+1)T_c]$ by a voltage $v(kT_c) = \pm V$. The carrier of frequency $f_c = \omega_c/2\pi$ can be expressed during this interval:

$$C(t) = \sqrt{2C}\cos(\omega_c t + \theta_k) = v(kT_c)A\cos(\omega_c t) \quad \text{(V)} \tag{4.2}$$

where C is the modulated carrier power and A the reference carrier amplitude, where $\theta_k = \overline{m}_k\pi$ and $\overline{m}_k$ is the logical complement of m_k. $\theta_k = 0$ if $m_k = 1$ and $\theta_k = \pi$ if $m_k = 0$. During this time

interval, the carrier thus exhibits a constant phase state, either 0 or π. The last term of Eq. (4.2) shows that this phase modulation can be regarded as suppressed carrier amplitude modulation with two amplitude states ±V (notice that the envelope remains constant). This modulation can be realised simply, as shown in Figure 4.6, by multiplication of the reference carrier by the voltage $v(t)$. Table 4.1 shows the relationship between b_k and the carrier phase for the two types of encoding.

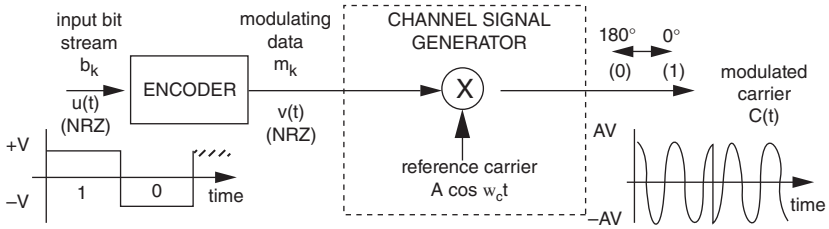


Figure 4.6 Two-state phase modulator (BPSK).

Table 4.1 Relationship between the bit and the carrier phase in BPSK

(a) Direct encoding

b_k	Phase
0	π
1	0

(b) Differential encoding

	Previous state		Present state		
b_k	m_{k-1}	Phase	m_k	Phase	
0	0	π	0	π	No phase change
	1	0	1	0	
1	0	π	1	0	Phase change
	1	0	0	π	

4.2.2 Four-state modulation – QPSK

Figure 4.7 shows the configuration of a four-state phase modulator. The symbol generator is a serial–parallel converter that generates two binary streams A_k and B_k, each of bit rate $R_c/2$, from the input stream of bit rate R_c. The symbol A_kB_k is a *dibit* that occupies the time interval $[kT_s, (k+1)T_s]$ equal to $T_s = 2T_c$ or the duration of two bits. The mapper, or encoder, transforms dibit A_kB_k into dibit I_kQ_k. As discussed earlier, only direct mapping is typically considered:

$$I_k = A_k \qquad Q_k = B_k \tag{4.3}$$

The signal generator superposes two carriers in quadrature. These two carriers are amplitude modulated (with suppressed carrier) by bits I_k and Q_k, which are represented in the time interval

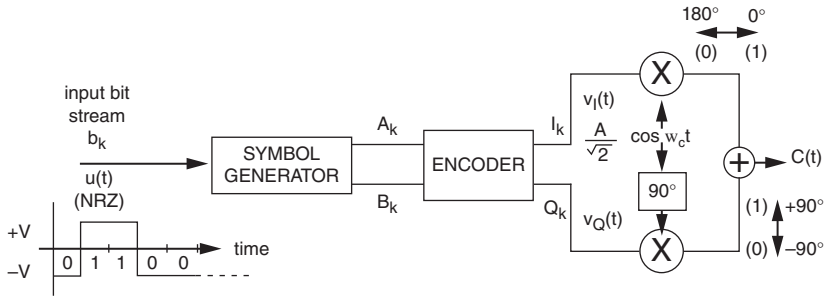


Figure 4.7 Four-state phase modulator (QPSK).

$[kT_s, (k+1)T_s]$ by voltages $v_1(kT_s) = \pm V$ and $v_Q(kT_s) = \pm V$. The expression for the carrier during the interval $[kT_s, (k+1)\ T_s]$ is:

$$C(t) = v_1(kT_s)\frac{A}{\sqrt{2}}\cos(\omega_c t) - v_Q(kT_s)\frac{A}{\sqrt{2}}\sin(\omega_c t)$$
$$= AV\ \cos(\omega_c t + \theta_k) \qquad \text{(V)} \tag{4.4}$$

where $\theta_k = 45°, 135°, 225°$, or $315°$ according to the values of the voltages $v_I(kT_s)$ and $v_Q(kT_s)$. In Figure 4.8, it can be seen that the carrier can take one of four phase states, each state being associated with one value of the symbol I_kQ_k. In general, two phase states separated by 90 are associated with two dibits I_kQ_k that differ by a single bit (Gray code). Hence an error at the receiver in recognising the phase between two adjacent phases leads to an error in a single bit. Table 4.2 shows the correspondence between dibit A_kB_k and the carrier phase.

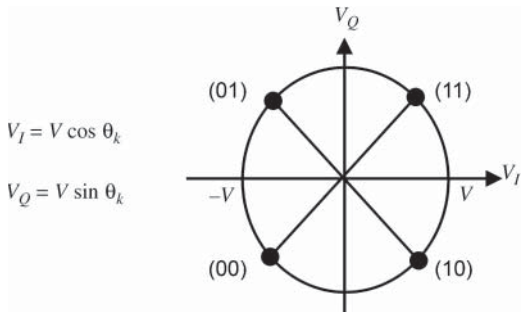


Figure 4.8 QPSK constellation.

4.2.3 Variants of QPSK

In QPSK modulation, the voltages that modulate the two carriers in quadrature change simultaneously, and the carrier can be subjected to a phase change of 180°. In a satellite link that includes filters, large phase shifts cause amplitude modulation of the carrier. The nonlinearity of the channel transforms these amplitude variations into phase variations (see Chapter 9) that degrade the performance of the demodulator. Several variants of QPSK modulation have been proposed to limit the amplitude of the phase shift.

Table 4.2 Relationship between the bit pair A_kB_k and the carrier phase in QPSK (direct encoding)

A_kB_K	Phase
00	$5\pi/4$
01	$3\pi/4$
10	$7\pi/4$
11	$\pi/4$

Furthermore, baseband pulse shaping can be introduced to limit the spectrum width of the modulating baseband waveform using a filter that smoothes the abrupt time variation of the NRZ rectangular-shaped pulse voltage associated with bits in Figure 4.7.

The most popular variants of QPSK are:

— Offset quadrature phase shift keying (OQPSK)
— $\pi/4$ QPSK
— Minimum shift keying (MSK)

4.2.3.1 Offset QPSK

With OQPSK, also called staggered quadrature phase shift keying (SQPSK), the I_k and Q_k modulating bit streams are offset by half a symbol duration, i.e. $T_s/2 = T_c$, the duration of a bit. The phase of the carrier changes every bit period but only ±90° or 0°. This avoids the possible 180° phase shift associated with the simultaneous change in the two bits in the modulating dibit with QPSK. It results in a reduced envelope variation when the modulated carrier is filtered.

The International Maritime Satellite Organisation (INMARSAT) aeronautical service uses aviation quadrature phase shift keying (A-QPSK). It is equivalent to OQPSK but replaces the baseband NRZ modulating voltage of bits I_k and Q_k in Figure 4.7 by the response to an impulse of amplitude V of a raised cosine pulse filter with transfer function H(*f*) given by [PRO-01, p. 546]:

$$H(f) = \begin{cases} T_s & \text{for } 0 \le |f| \le \frac{1-\alpha}{2T_s} \\ \frac{T_s}{2}\left\{1 + \cos\left[\frac{\pi T_s}{\alpha}\left(|f| - \frac{1-\alpha}{2T_s}\right)\right]\right\} & \text{for } \frac{1-\alpha}{2T_s} \le |f| \le \frac{1+\alpha}{2T_s} \\ 0 & \text{for } |f| > \frac{1+\alpha}{2T_s} \end{cases} \tag{4.5}$$

where α is the roll-off factor. The selected roll-off factor of A-QPSK is $\alpha = 1$.

4.2.3.2 π/4 QPSK

This modulation scheme is another approach to avoiding 180° instantaneous phase shifts. It uses differential encoding. The modulating data I_k and Q_k at time k are determined from the incoming dibits A_kB_k and the previous modulating dibits $I_{k-1}Q_{k-1}$, according to the following transform:

$$\begin{pmatrix} I_k \\ Q_k \end{pmatrix} = \begin{pmatrix} \cos\theta_k & -\sin\theta_k \\ \sin\theta_k & \cos\theta_k \end{pmatrix} \begin{pmatrix} I_{k-1} \\ Q_{k-1} \end{pmatrix}$$

where $\theta_k = \pi/4; 3\pi/4; -3\pi/4; -\pi/4$, depending on the incoming dibit A_kB_k (11, 01, 00, 10). For example, at instant $k-1$, the carrier phase is one of the four phases of constellation 1 (Figure 4.9a).

At instant k, the carrier takes one of the possible phase values of constellation 2 (Figure 4.9b). As the two constellations are shifted with respect to each other by $\pi/4$, the possible phase changes are $\pm\pi/4$ or $\pm 3\pi/4$, as illustrated by the lines in Figure 4.9c indicating the possible phase transitions starting from a particular phase at instant $k-1$.

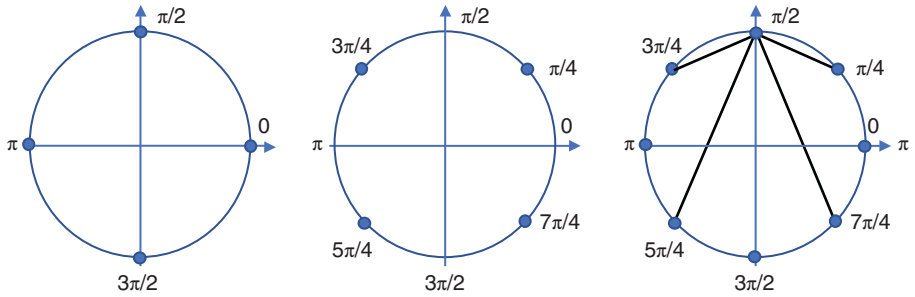


Figure 4.9 QPSK modulation: (a) constellation 1; (b) constellation 2; (c) possible phase transitions on successive symbols.

4.2.3.3 Minimum shift keying

This MSK modulation scheme is a special case of OQPSK, replacing the NRZ rectangular-shaped pulse voltage associated with bits I_k and Q_k in Figure 4.7 by the response to an impulse of amplitude V of a pulse filter with impulse response $h(t)$ given as follows [GRO-76]:

— For the I_k stream

$$h(t) = \begin{cases} \cos(\pi t/2T) & \text{for } 0 \le t \le T \\ 0 & \text{otherwise} \end{cases}$$

— For the Q_k stream

$$h(t) = \begin{cases} \sin(\pi t/2T) & \text{for } 0 \le t \le T \\ 0 & \text{otherwise} \end{cases}$$

The carrier phase increases by $\pm\Delta\omega T_c$ during each bit duration T_c with $\Delta\omega = \pi/2T_c$, as shown in Figure 4.10. The phase varies linearly with time during each bit period T_c and is equal to an integer multiple of $\pi/2$ at the end of each bit period. This phase variation translates into a constant frequency during each bit period, which makes the modulation equivalent to a FSK modulation with two frequencies, $f_c - 1/4T_c$ and $f_c + 1/4T_c$, depending on the incoming bit, where f_c is the reference carrier frequency.

To avoid the sharp changes in phase slopes at the end of each bit interval in Figure 4.10, Gaussian-filtered minimum shift keying (GMSK) uses premodulation baseband filtering with a Gaussian-shaped frequency response. This results in the carrier experiencing $\pi/2$ phase steps smoothed in order to reduce its spectrum width. This technique is used when bandwidth saving is at a premium. The inconvenience is the generated intersymbol interference (ISI), as each data bit influences the carrier phase during a time period exceeding the bit duration (typically 3 T_c).

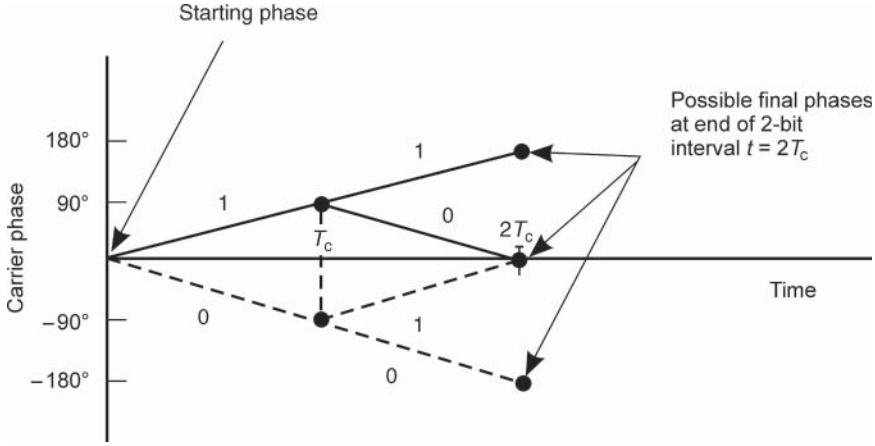


Figure 4.10 Phase of an MSK-modulated carrier.

4.2.4 Higher-order PSK and APSK

When combining three consecutive bits to define the symbol, 8PSK modulation ($M = 8$) is defined (Figure 4.11b). Higher-order PSK modulation can be considered. However, as the order of the modulation increases, the phase difference between carrier states decreases, and it is required to increase the amplitude of the carrier to maintain the same distance between carrier plots so as to obtain the same BER at output of the demodulator (see Section 4.2.6.3).

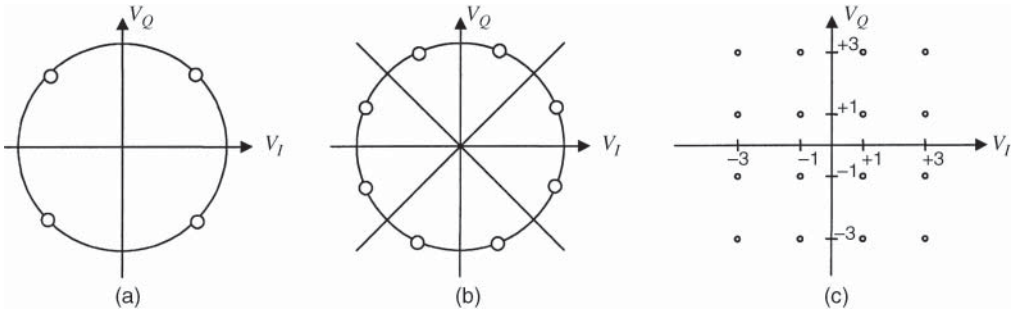


Figure 4.11 Some alphabet modulation schemes: (a) QPSK; (b) 8PSK; (c) 16QAM amplitude, and phase modulation where $v_1(kT_s)$ and $v_Q(kT_s)$ take values in the domain $\{\pm 1, \pm 3\}$.

Based on the configuration of the quadrature phase modulator where the radio-frequency signal generator superposes two carriers in quadrature, one could envision modulating the amplitude of each carrier using two positive and two negative signal voltages (4-ary amplitude symbol). This would result in 16-quadrature amplitude modulation (16QAM), where carrier plots are organised in squares (Figure 4.11c). This modulation does not display a constant envelope as there are three possible values of the carrier amplitude that makes the modulation quite sensitive to nonlinearity of the satellite channel.

Keeping the same number of carrier states ($M = 16$), it is possible to reduce the number of amplitude values to two (so as to reduce the nonlinear impairments) by distributing the

carrier plots on two concentric circles. This modulation is called 16APSK (see Figure 4.12a). If quasi-linear channels are available (thanks to the use of on-board linearisers; see Chapter 9), a three-level amplitude keyed carrier could be considered. In combination with different phase values, this results in the 32APSK modulation (Figure 4.12b).

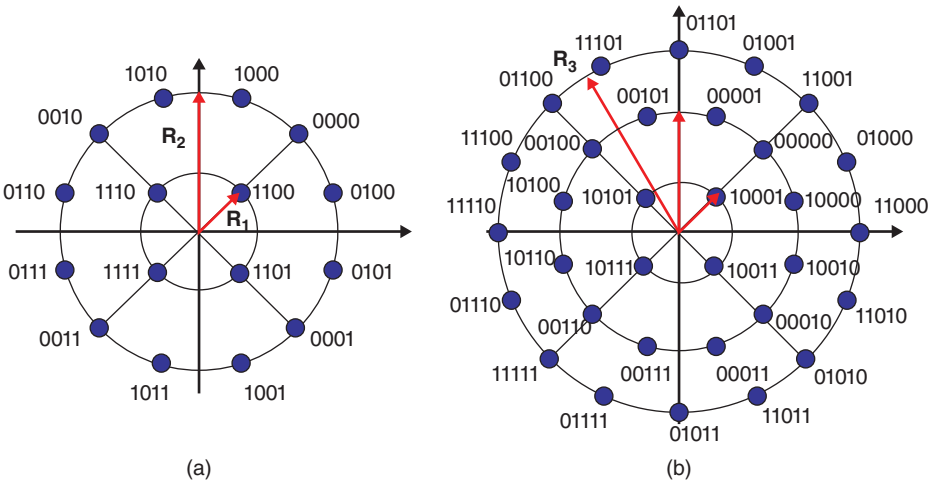


Figure 4.12 (a) 16APSK and (b) 32APSK modulations.

4.2.5 Spectrum of unfiltered modulated carriers

Figure 4.13 shows the shape of the power spectral density (W/Hz) of modulated carriers according to the modulation schemes, as a function of the normalised frequency. The vertical axis displays the relative level in decibels (dB) of the power density with respect to the maximum value at the unmodulated carrier frequency f_c. The normalised frequency on the horizontal axis is defined as the frequency difference between the considered frequency f and the unmodulated carrier frequency f_c with respect to the bit rate modulating the carrier, R_c.

The displayed spectra correspond to unfiltered modulated carriers. There are two concerns: the width of the main lobe of the spectrum of the modulated carrier, which conditions the required bandwidth; and the spectral decay of side lobes with frequency, which conditions interference to adjacent carriers. In this respect, QPSK ($M = 4$, $m = 2$) outperforms BPSK ($M = 2$, $m = 1$) in terms of spectrum width. As a general rule, the width of the main lobe decreases as a function of the value of the number m of bits per symbol, which translates into higher spectral efficiency. MSK shows a faster side lobe decay, at the expense of a larger main lobe width compared to QPSK.

In practice, filtering is implemented at the transmitter and the receiver side, in order to limit the interference to adjacent out-of-band carriers. The effect of filtering is discussed in Section 4.2.7.

4.2.6 Demodulation

The role of the demodulator is to identify the phase (or phase shift) of the received carrier and to deduce from it the value of the bits of the transmitted binary stream. Demodulation can be:

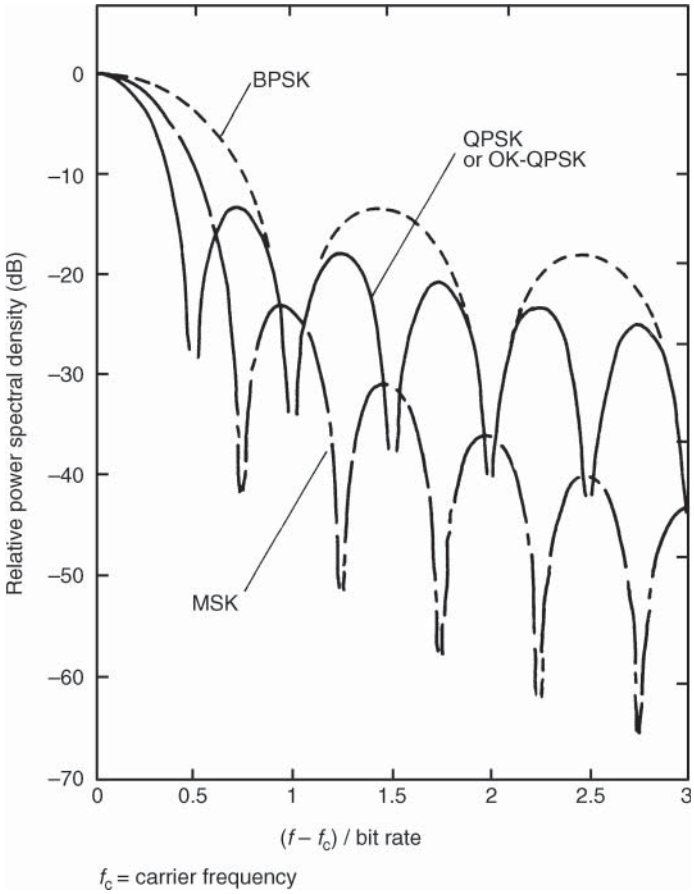


Figure 4.13 The spectrum of digital carriers.

— *Coherent:* The demodulator makes use of a local sinusoidal reference signal having the same frequency and phase as the modulated wave at the transmitter. The demodulator interprets the phase of the received carrier by comparing it with the phase of the reference signal. Coherent demodulation enables the binary stream to be reconstructed for both cases of transmission encoding – direct (BPSK and QPSK) and differential (DE-BPSK and DE-QPSK).

— *Differential:* The demodulator compares the phase of the received carrier for the duration of transmission of a symbol and its phase for the duration of the preceding symbol. The demodulator thus detects phase changes. The transmitted information can be recovered only if it is contained in phase changes; differential demodulation is *always* associated with differential encoding on transmission. This type of modulation and demodulation is identified as differential demodulation (D-BPSK).

The structure of BPSK and QPSK demodulators is examined in the following sections, and then the performance of the various types of modulation and demodulation is compared.

4.2.6.1 Coherent and differential demodulators

(a) *Coherent demodulation of BPSK modulation (Figure* 4.14*a):* The received modulated carrier is proportional to $\cos(\omega_c t + \theta_k)$. It is multiplied by the reference carrier $\cos \omega_c t$ delivered at the output of the carrier recovery circuit. The result is proportional $\cos(2\omega_c t + \theta_k) + \cos\theta_k$. The low-pass filter eliminates the component at frequency $2f_c = 2\omega_c/2\pi$ and outputs a voltage proportional to $\cos\theta_k$, which is positive or negative depending on whether $\theta_k = \pi$ or 0. The current bit value is decided by comparing the voltage and the zero threshold of the detector at the bit timing recovered by the bit timing recovery circuit.

(b) *Differential demodulation of DE-BPSK (Figure* 4.14*b):* The received DE-BPSK carrier is fed into a delay line (with a delay equal to the duration of one bit) and is multiplied by the delayed output of the line. The result of the multiplication, $\cos(\omega_c t + \theta_k)\cos(\omega_c t + \theta_{k-1})$, is filtered by a low-pass filter, the output of which is $\frac{1}{2}\cos(\theta_k - \theta_{k-1})$; the value of m_k is deduced from the sign of $\cos(\theta_k - \theta_{k-1})$.

(c) *Coherent demodulation of QPSK (Figure* 4.15*):* The demodulator is an extension of coherent demodulation of a BPSK carrier to channels in phase and quadrature.

4.2.6.2 Symbol and bit error probabilities

Carrier phase (or phase shift) changes under the influence of noise lead to errors in identification of the received symbols, hence the received bits. *Symbol error probability* (SEP) is the probability of a symbol being detected in error; *bit error probability* (BEP) is the probability of a bit being detected in error.

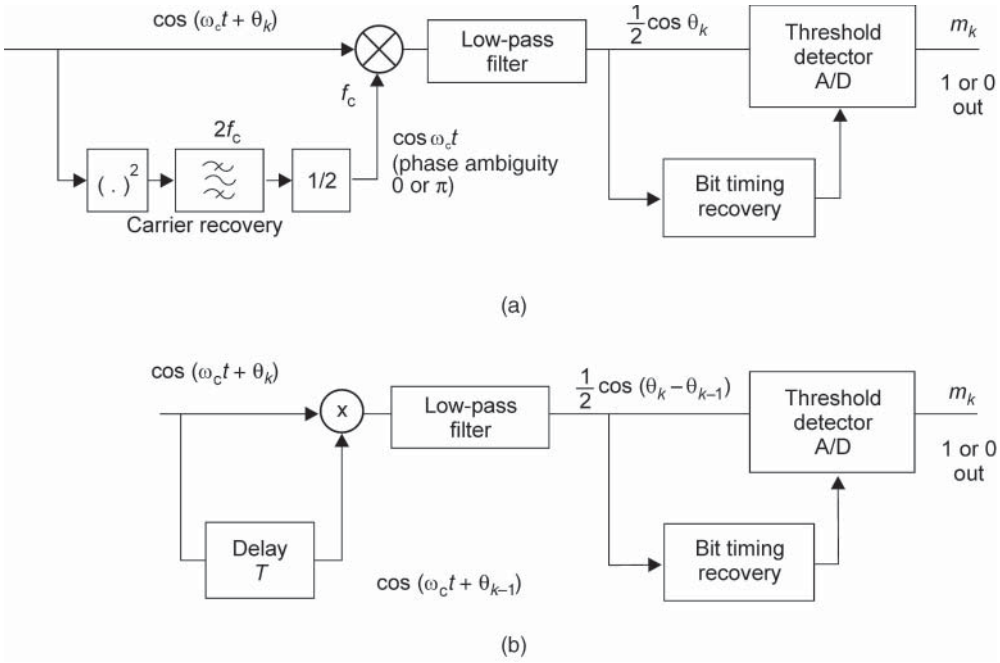


Figure 4.14 Coherent demodulator for (a) BPSK and (b) DE-BPSK.

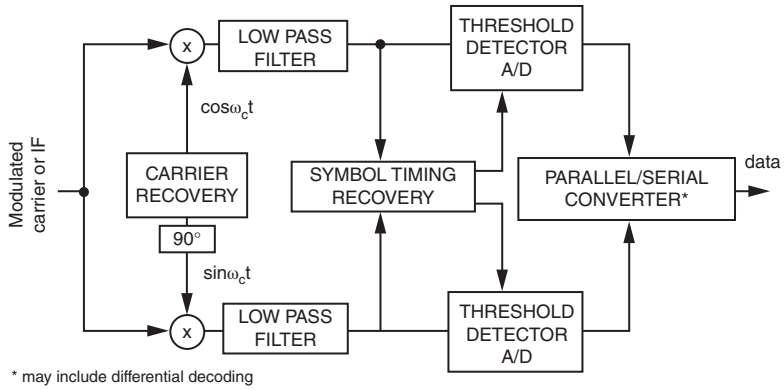


Figure 4.15 The structure of a coherent demodulator for four-state phase modulation.

For two-state modulation, the symbol identifies to the bit. Hence the SEP represents the BEP:

$$\text{BEP} = \text{SEP} \tag{4.6}$$

For four-state modulation, where association of the symbols $I_k Q_k$ with the phase states follows a Gray code, the BEP is given by:

$$\text{BEP} = \text{SEP}/2 \tag{4.7}$$

More generally:

$$\text{BEP} = \text{SEP}/\log_2 M \quad \text{for} \quad M \geq 2 \tag{4.8}$$

Table 4.3 gives the expressions for the BEPs for the demodulators considered previously [PRO-01]. Figure 4.16 shows the corresponding BEP curves. The function erfc is the complementary error function defined by:

$$\text{erfc}\,(x) = (2/\sqrt{\pi}) \int_x^\infty e^{-u^2} du \tag{4.9}$$

Table 4.3 Expressions for bit error probabilities (BEP)

Type of modulation–demodulation	Bit error probability
Coherent demodulation:	
Direct encoding:	
BPSK	$(1/2)\text{efrc}\sqrt{(E_c/N_0)}$
QPSK	$(1/2)\text{efrc}\sqrt{(E_c/N_0)}$
Differential encoding:	
DE-BPSK	$\text{efrc}\sqrt{(E_c/N_0)}$
DE-QPSK	$\text{efrc}\sqrt{(E_c/N_0)}$
Differential demodulation (differential encoding only):	
D-BPSK	$(1/2)\exp(-E_c/N_0)$

A convenient approximation for erfc $\sqrt{(E_c/N_0)}$ is $(1/\sqrt{\pi})\dfrac{\exp(-E_c/N_0)}{\sqrt{(-E_c/N_0)}}$, when $\frac{E_c}{N_0} \geq 4 (= 6\text{ dB})$.

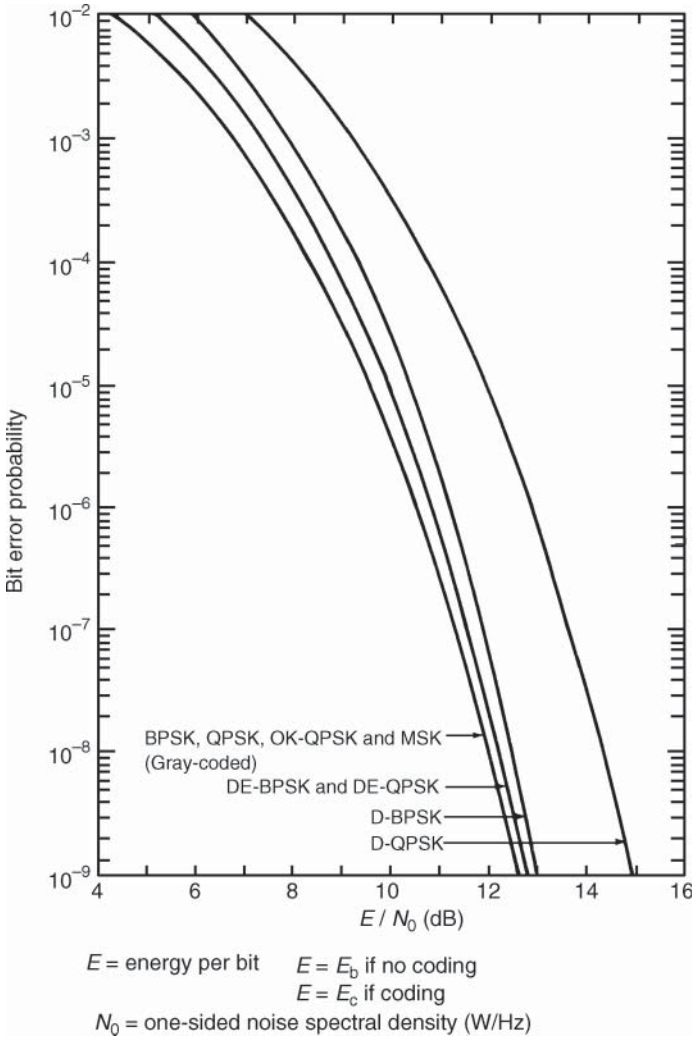


Figure 4.16 Theoretical bit error probability (BEP).

The ratio E_c/N_0 arises in the expression for error probability, where E_c is the energy per channel bit. This is the product of the power of the received carrier for the duration of one bit, namely $E_c = CT_c = C/R_c$. Hence:

$$E_c/N_0 = (C/R_c)/N_0 = (C/N_0)/R_c \tag{4.10}$$

For a transparent satellite link, the value of C/N_0 used is the overall link performance $(C/N_0)_T$ as derived in Chapter 5. For a regenerative satellite link, the value of C/N_0 used is that of either the uplink or the downlink.

4.2.6.3 Bit error rate (BER)

The BER measures the performance of the demodulator by counting the number of bits in error, n, in a stream of N received bits:

$$\text{BER} = n/N$$

The BER constitutes an estimate of the BEP. A level of confidence is associated with this estimate, as follows:

$$\text{BEP} = \text{BER} \pm k(\sqrt{n})/N$$

A 63% level of confidence is obtained for $k = 1$ and a 95% level of confidence for $k = 2$. For instance, if $n = 100$ errors are observed within a run of $N = 10^5$ bits, the BEP is $10^{-3} \pm 10^4$ with a 63% confidence level.

Performance objectives stipulate a given BER (Section 3.2), from which the required value of E_c/N_0 is determined. Table 4.4 indicates the theoretical values of E_c/N_0 necessary to achieve a given BEP for each type of modulation and demodulation. The figures in brackets indicate the difference between the value of E_c/N_0 for the modulation/demodulation type considered and the value obtained with BPSK or QPSK.

It can be seen that differentially encoded modulation requires a higher E/N_0 value, which translates into a higher C/N_0 requirement to achieve a given BEP. However, as the information is conveyed in the phase shift between two successive signals, there is no need to recover the exact reference phase of the carrier. With coherent demodulation, this avoids the need to resolve the reference carrier phase ambiguity introduced by the carrier recovery circuit as a result of taking the frequency at the output of the squaring device in Figure 4.14a and dividing it by 2. With direct encoding, the ambiguity is resolved by preamble insertion of a known bit sequence (the preamble) at the transmit side and preamble detection at the receive side, which adds to the complexity of the transmission scheme. Differential demodulation avoids the need for recovery of the reference carrier, which makes the demodulator simple but degrades its performance with respect to coherent demodulation.

As a result of the demodulator implementation, the value of BER is higher than the theoretical BEP, as given by the values in Table 4.4. In order to obtain the required BER, one has to increase by some quantity the value of E_c/N_0, obtained from the expressions in Table 4.4. This quantity is the demodulator *implementation degradation*. Depending on the technology and the considered

Table 4.4 Theoretical values of E_c/N_0 to achieve a given bit error probability (E_c = energy per channel bit, N_0 = noise spectral density)

BEP	BPSK QPSK (dB)	DE-BPSK (Δ) DE-QPSK	D-BPSK (Δ)	D-QPSK (Δ)
10^{-3}	6.8	7.4 dB (0.6 dB)	7.9 dB (1.1 dB)	9.2 dB (2.4 dB)
10^{-4}	8.4	8.8 dB (0.4 dB)	9.3 dB (0.9 dB)	10.7 dB (2.3 dB)
10^{-5}	9.6	9.9 dB (0.3 dB)	10.3 dB (0.7 dB)	11.9 dB (2.3 dB)
10^{-6}	10.5	10.8 dB (0.3 dB)	11.2 dB (0.7 dB)	12.8 dB (2.3 dB)
10^{-7}	11.3	11.5 dB (0.2 dB)	11.9 dB (0.6 dB)	13.6 dB (2.3 dB)
10^{-8}	12.0	12.2 dB (0.2 dB)	12.5 dB (0.5 dB)	14.3 dB (2.3 dB)
10^{-9}	12.6	12.8 dB (0.2 dB)	13.0 dB (0.4 dB)	14.9 dB (2.3 dB)

Δ = difference in E_c/N_0 relative to BPSK and Q-PSK.

BER, the degradation ranges from 0.5 dB for simple BPSK modulation to a few dB for high-order modulations (M = 16 or 32) that are sensitive to channel nonlinearity and synchronisation errors.

4.2.7 Modulation spectral efficiency

Modulation spectral efficiency can be defined as the ratio of the transmitted bit rate R_c to the bandwidth occupied by the carrier. The bandwidth occupied by the carrier depends on the spectrum of the modulated carrier and the filtering it undergoes.

The spectra displayed in Figure 4.13 correspond to unfiltered modulated carriers. In practice, filtering is implemented at the transmitter and the receiver side, in order to limit the interference to adjacent out-of-band carriers. However, such filtering introduces ISI [PRO-01], pp. 536–537], which degrades the BER performance compared to the earlier theoretical results. For rectangular pulse shaping, as used in BPSK and QPSK, ISI-free transmission can be achieved using a rectangular bandpass filter (a 'brick wall' filter) that corresponds to the minimum required carrier bandwidth. This bandwidth (the Nyquist bandwidth) is equal to $1/T_s$ the inverse of the signal or symbol duration T_s. This is not implemented in practice, as a filter with sharp transitions in the frequency domain is not realisable. Also the slow decay in the time domain response translates any timing error by the clock recovery circuit of the receiver into erroneous decisions in the detection process and these are detrimental to the BER performance.

ISI-free transmissions can be achieved with specific filters displaying smoother transitions in the frequency domain, such as the raised cosine filter introduced by Eq. (4.5). A larger bandwidth is then required, depending on the value of the *roll-off factor* α. For a raised cosine filter with roll-off factor α, the bandwidth B occupied by the carrier is then:

$$B = (1 + \alpha)/T_s \quad \text{(Hz)} \tag{4.11a}$$

Therefore the spectral efficiency Γ for an M-ary modulation scheme is:

$$\Gamma = \frac{R_c}{B} = R_c T_s/(1 + \alpha) = \log_2 M/(1 + \alpha) \quad (\text{bit s}^{-1}\,\text{Hz}^{-1}) \tag{4.11b}$$

where $m = \log_2 M$ is the number of bits per symbol.

— With a roll-off factor $\alpha = 0.35$, the required bandwidth is $1.35/T_s$, and the spectral efficiency is $\Gamma = 0.7$ bit s^{-1} Hz^{-1} for BPSK, $\Gamma = 1.5$ bit s^{-1} Hz^{-1} for QPSK, $\Gamma = 2.2$ for 8PSK, etc.

Filtering is implemented at both ends of the link. If the channel is linear, theory indicates that the overall filtering should be split evenly between the transmitter and the receiver filters. A satellite channel is usually nonlinear, as a result of the nonlinear characteristics of the power amplifiers in the earth stations and on board the satellite. The previous filtering approach no longer provides ISI-free transmission. Furthermore, the nonlinear channel introduces spectrum spreading of the filtered modulated carrier, which increases the adjacent channel interference (ACI). Backing off the amplifier operating point is a way to reduce the spectrum spreading at the expense of carrier power reduction.

The noise power introduced into the receiver is given by $N = N_0 B_N$, where B_N is the receiver *noise bandwidth*. For a Nyquist filter, the noise bandwidth is equal to $1/T_s$, independent of the roll-off factor α. In practice, the noise bandwidth of the implemented filter exhibits some dependency on roll-off. It is then convenient to consider that the noise bandwidth of the receiver is equal to the carrier bandwidth, i.e. B_N = B, as given by Eq. (4.11a).

4.3 CHANNEL CODING

Figure 4.17 illustrates the principle of channel encoding. It has the objective of adding, to the information bits, redundant bits, which are used at the receiver to detect and correct errors [PRO-01]. This technique is called forward error correction (FEC). The code rate is defined as:

$$\rho = n/(n+r) \tag{4.12a}$$

where r is the number of redundant bits added for n information bits.

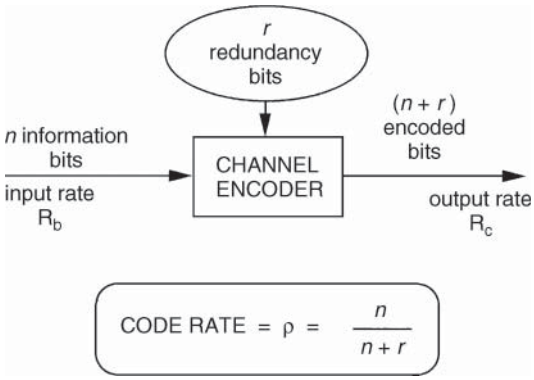


Figure 4.17 The principle of channel encoding.

The bit rate at the encoder input is R_b. At the output, it is greater and is equal to R_c. Hence:

$$R_c = R_b/\rho \qquad \text{(bit/s)} \tag{4.12b}$$

4.3.1 Block encoding and convolutional encoding

Two encoding techniques are used:

— *Block encoding:* The encoder associates r bits of redundancy with each block of n information bits; each block is coded independently of the others. The code bits are generated by linear combination of the information bits of the corresponding block. Cyclic codes are most used, particularly the codes of Reed–Solomon (RS) and Bose, Chaudhari, and Hocquenghem (BCH) for which every code word is a multiple of a generating polynomial.
— *Convolutional encoding*: $(n+r)$ Bits are generated by the encoder from the $(N-1)$ preceding packets of n bits of information; the product $N(n+r)$ defines the *constraint length* of the code. The encoder consists of shift registers and adders of the 'exclusive OR' type.

The choice between block encoding and convolutional encoding is dictated by the types of error that are expected at the output of the demodulator. The distribution of errors depends on the nature of noise and propagation impairments encountered on the satellite link:

- Under stable propagation conditions and Gaussian noise, errors occur randomly and convolutional encoding is most commonly used.

- Under fading conditions, errors occur mostly in bursts; compared with convolutional coding, block encoding is less sensitive to bursts of errors, so block encoding is preferred under fading conditions.

4.3.2 Channel decoding

With FEC, the decoder uses the redundancy introduced at the encoder in order to detect and correct errors. Various possibilities are available for decoding block and convolutional codes. With block cyclic codes, one of the conventional methods uses the calculation and processing of syndromes resulting from division of the received block by the generating polynomial; this is zero if the transmission is error free. For convolutional codes, the best performance is obtained with the Viterbi decoding algorithm [VIT-79].

At the decoder input, the bit rate is R_c and the bit error probability is $(BEP)_{in}$. At the output, the information rate is again R_b, which was that at the encoder input. Because of the error correction provided by the decoder, the bit error probability $(BEP)_{out}$ is lower than at the input. Figure 4.18 depicts an example of the relation between $(BEP)_{out}$ and $(BEP)_{in}$. The value of $(BEP)_{in}$ is given as a function of E_c/N_0, according to the modulation/demodulation type, by one of the curves of Figure 4.16. By combining this curve with the curve of Figure 4.18 for the encoding/decoding scheme considered, the curves of Figure 4.19 can be established. These curves establish the performance of the modulation and encoding system.

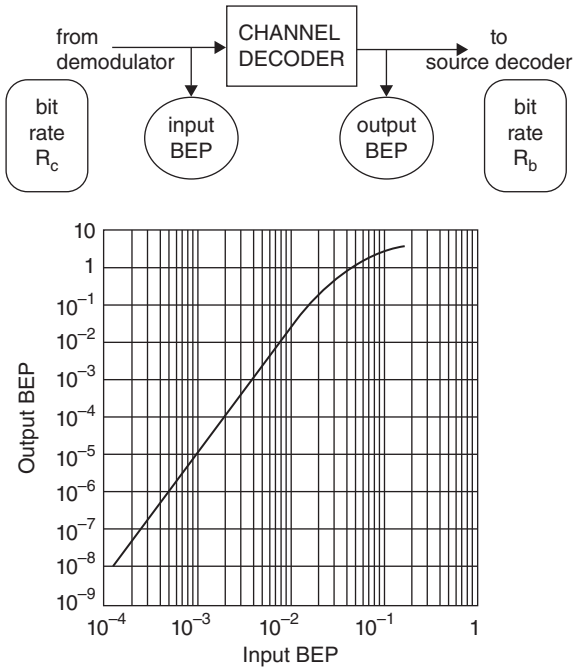


Figure 4.18 The relationship between the bit error probability $(BEP)_{out}$ at the output of an error-correcting decoder and the bit error probability at the input $(BEP)_{in}$.

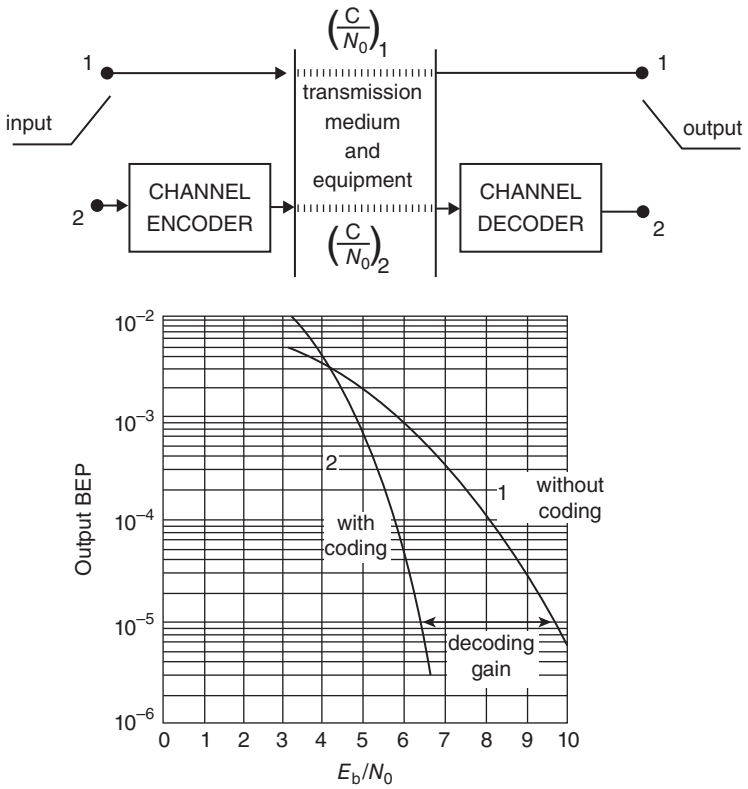


Figure 4.19 The performance of a modulation and coding system; the definition of decoding gain.

Notice that the BEP is expressed as a function of E_b/N_0, where E_b represents the energy per information bit; that is, the amount of power accumulated from the carrier over the duration of the considered information bit. As the carrier power is C, and the duration of the information bit is $T_b = 1/R_b$, where R_b is the information bit rate, then E_b is equal to C/R_b.

E_b/N_0 relates to E_c/N_0, where E_c is the energy per coded bit modulating the carrier (bits at output of channel encoder and input to channel decoder) as follows:

$$E_b/N_0 = E_c/N_0 - 10\log\rho \text{ (dB)} \tag{4.13}$$

where ρ is the code rate defined by Eq. (4.12a). The *decoding gain* G_{cod} is defined as the difference in decibels (dB) at the considered value of BEP between the required values of E_b/N_0 with and without coding, assuming equal information bit rate R_b. Table 4.5 indicates typical values of decoding gain for a BEP equal to 10^{-6} considering standard Viterbi decoding of a convolutionally encoded bit stream. The use of turbo codes, as per the iterative design of the decoder [BER-93], brings larger values of the decoding gain. Another efficient FEC scheme makes use of low-density parity check (LDPC) block codes. Combined with an outer coder and interleaving in a concatenated coding structure (see Section 4.3.3), the performances are about 0.7–1 dB from the Shannon channel capacity limit (BER $< 10^{-5}$ for $E_b/N_0 = 0.7$ dB).

Table 4.5 Typical values of decoding gain

Code rate ρ	E_b/N_0 required for BEP = 10^{-6} (dB)	Decoding gain (dB)
1	10.5	0
7/8	6.9	3.6
3/4	5.9	4.6
2/3	5.5	5.0
1/2	5.0	5.5

4.3.3 Concatenated encoding

Block encoding and convolutional encoding can be combined in a *concatenated encoding scheme* (Figure 4.20). The layout incorporates an outer block encoder, followed by an inner encoder. At the receiving end, the inner decoder corrects errors at the output of the demodulator. The outer decoder is able to correct the occasional bursts of errors generated by the inner decoder's decoding algorithm, which produces such bursts of errors whenever the number of errors in the incoming bit stream oversteps the correcting capability of the algorithm. The performance of concatenated encoding is improved while using simple outer coders by implementing interleaving and deinterleaving between the outer and inner coders.

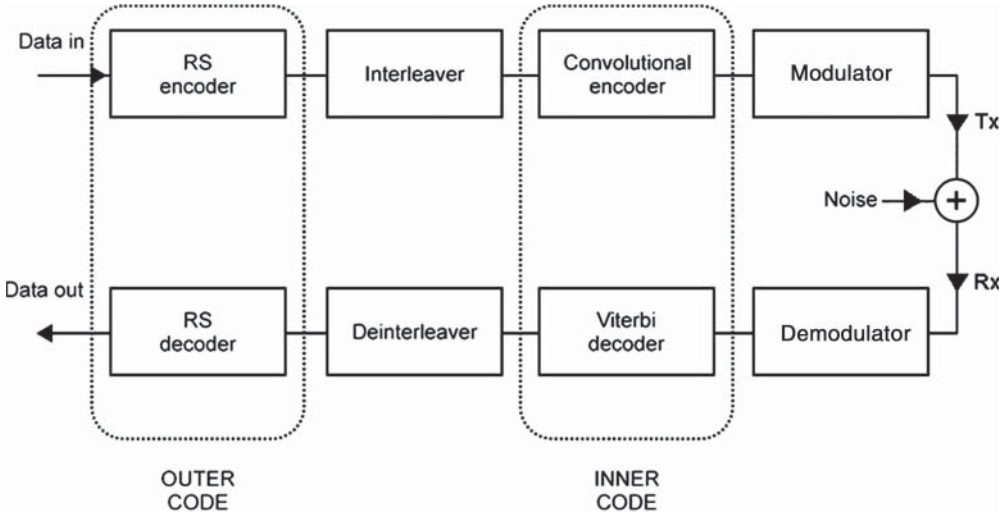


Figure 4.20 Concatenated encoding scheme as used for the DVB-S standard.

Concatenated encoding has been retained for the DVB-S standard [ETSI-97]. The outer block encoder is an RS (204, 188) encoder where 16 redundancy bytes are added to each input block of 188 bytes, yielding output blocks of 204 bytes (code rate $\rho = 188/204$). The inner encoder is a convolutional code with five different code rates (1/2, 2/3, 3/4, 5/6, and 7/8). The DVB-S2 standard also makes use of concatenated encoding, but using two block codes: BCH for the outer code with adaptive coding and modulation (ACM) range of 18 dB and LDPC for the inner

code with codes rates of 1/4, 1/3, 3/5, 4/5, 8/9, and 9/10 in addition to DVB-S (see Section 4.8) [ETSI-14]. Furthermore, DVB-S2x has further innovation with more modulation and coding up to 256 APSK [ETSI-15a].

4.3.4 Interleaving

Interleaving is a way to improve the performance of convolutional encoding with respect to bursts of errors. It consists of ordering the encoded bits before transmission so that bursts of errors are randomised when the bits are reordered before decoding at the receiving end. Interleaving is also used with concatenated coding to distribute bursts of errors as the output of the inner decoder over different code blocks (see Section 4.3.3). Two interleaving techniques are used:

— *Block interleaving* (Figure 4.21a): Bits are organised in blocks of N bits that are sequentially laid down in the B rows of an (N, B) memory array and read out for transmission from the N columns in blocks of B bits. A burst of errors spanning N bits affects only one bit in each transmitted block. This technique introduces a delay approximately equal to $2NB$ bits duration.
— *Convolutional interleaving* (Figure 4.21b): Bits are organised in blocks of N bits. The ith bit ($i = 1, 2, \ldots, N$) in each block is delayed by $(i-1)NJ$ time units through an $(i-1)J$ stage shift register clocked once every N bit times, where $J = B/N$. A time unit thus corresponds to the transmission of a block of N bits. The output bits are serialised for transmission. At the receiving end, groups of N bits are reblocked, and the ith bit in each block is delayed by $(N-i)NJ$ time units through an $(N-i)J$ stage shift register. This technique introduces a constant delay of $(N-i)J$ time units, equal to $(N-i)J = (N-i)B$ bit duration. The delay is therefore about half the delay introduced by an (N, B) block interleaver.

4.4 CHANNEL CODING AND THE POWER–BANDWIDTH TRADE-OFF

4.4.1 Coding with variable bandwidth

Coding allows bandwidth to be exchanged for power, so the link performance can be optimised with respect to cost. This is paramount in the design of a link. Consider a satellite link that conveys an information bit rate $R_b = 2.048$ Mbit s^{-1} using BPSK with spectral efficiency $\Gamma = 0.7$ bit s^{-1} Hz^{-1}. The objective bit error rate is BER = 10^{-6}.

(a) Without coding ($\rho = 1$)

$$\text{The transmitted bit rate is } R_c = R_b = 2.048\ \text{Mbit/}s$$

$$\text{The bandwidth used is } \quad B_{\text{nocod}} = R_c/\Gamma = 2.048/0.7 = 2.9\ \text{MHz}$$

The theoretical required value for E_b/N_0 (not taking into account implementation degradation) is given in Table 4.5 and equals $(E_b/N_0) = 10.5$ dB. The required value of C/N_0 is:

$$(C/N_0)_{\text{nocod}} = (E_b/N_0)_{\text{nocod}}\, R_{\text{b}} = 10.5\ \text{dB} + 63.1\ \text{d Bbit/s} = 73.6\ \text{dBHz}$$

This corresponds to the required value for the overall (from transmit earth station to receive earth station) link performance $(C/N_0)_{\text{T}}$ for a transparent satellite (with no on-board demodulation), or the relevant uplink $(C/N_0)_{\text{U}}$ or downlink $(C/N_0)_{\text{D}}$ performance for a regenerative satellite.

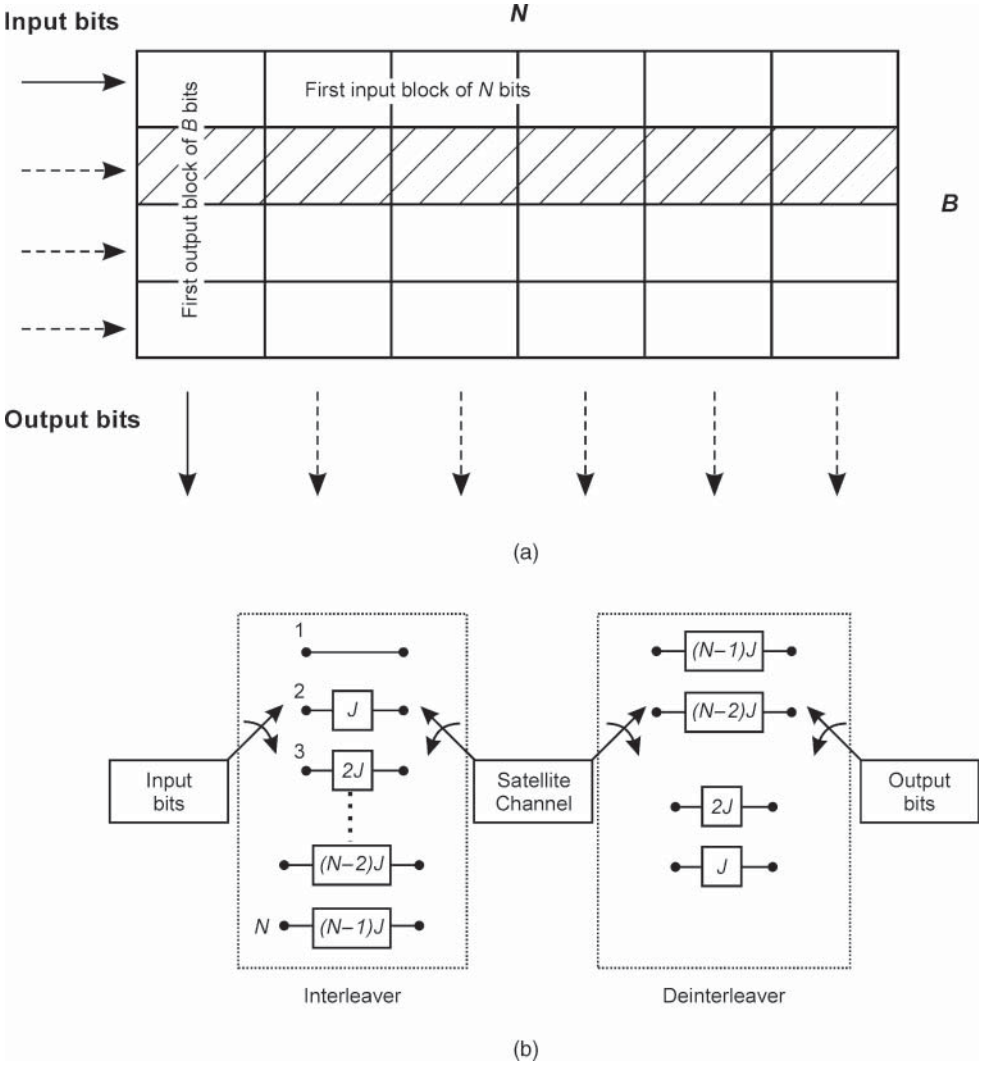


Figure 4.21 Interleaving techniques: (a) block interleaving (shaded boxes represent bits that would be errorred without interleaving during a burst of N consecutive errors); (b) convolutional interleaving.

(b) With coding ($\rho < 1$)

Assume for instance	$\rho = 7/8$
The transmitted bit rate is	$R_c = R_b/\rho = 2.048/(7/8) = 2.34$ Mbit/s
The bandwidth used is	$B_{cod} = R_c/\Gamma = 2.34/0.7 = 3.34$ MHz

The theoretical required value for E_b/N_0 (not taking into account implementation degradation) depends on the code rate ρ as indicated by Table 4.5. With $\rho = 7/8$ the required E_b/N_0 is $(E_b/N_0)_{cod} = 6.9$ dB. The required value of C/N_0 is

$$(C/N_0)_{cod} = (E_b/N_0)_{cod}\, R_b = 6.9 \text{ dB} + 63.1 \text{ dBbit/s} = 70 \text{ dBHz}$$

Table 4.6 Impact of coding for variable bandwidth: object BER = 10^{-6}, BPSK modulation

Code rate ρ	Typical required E_b/N_0 (dB)	Required C/N_0 (dB Hz)	Required bandwidth (MHz)
1	10.5	73.6	2.9
7/8	6.9	70.0	3.3
3/4	5.9	69.0	3.9
2/3	5.5	68.6	4.4
1/2	5.0	68.1	5.9

Table 4.6 displays these results along with those obtained for different code rates. Depending on the selected code rate, one observes a reduction in the required value of C/N_0; this corresponds to a lower power requirement and an expansion in the required bandwidth. The reduction $\Delta C/N_0$ is equal to the decoding gain:

$$\Delta C/N_0 = (E_b/N_0)_{nocod} - (E_b/N_0)_{cod} = G_{cod} \quad \text{(dB)}$$

The reduction in the required E_b/N_0, which translates to an equal reduction in the required carrier power C, is paid for by an increase in the required bandwidth used on the satellite link. In fact, it is necessary to transmit a bit rate R_c that is greater than the information bit rate R_b, and, according to Eq. (4.12b), the bandwidth used is $B = R_c\Gamma = R_b/\rho\Gamma$: The bandwidth expansion is:

$$\Delta B = 10 \log B_{cod} - 10 \log B_{nocod} = -10 \log \rho \quad \text{(dB)}$$

Figure 4.22 illustrates the variation of the requirements upon C/N_0 and B as a function of the code rate ρ. As the code rate decreases, less power is demanded but more bandwidth is required.

4.4.2 Coding with constant bandwidth

Coding with constant bandwidth is performed when a given bandwidth is allocated to a given link. Coding is introduced without changing the carrier bandwidth B and therefore at a constant transmitted rate R_c. Consequently, the information bit rate R_b must be reduced.

Without coding, the transmitted bit rate R_c is constrained by the allocated bandwidth B. Assuming $B_{nocod} = 2.9$ MHz, the transmitted bit rate is $R_c = (R_b)_{nocod} = \Gamma B = 2.048$ Mbit s^{-1}. The required value of C/N_0 is given by:

$$(C/N_0)_{nocod} = (E_b/N_0)_{nocod}\,(R_b)_{nocod} = 10.5\,\text{dB} + 63.1\text{dBbit/s} = 73.6\,\text{dBHz}$$

With coding, the transmitted bit rate R_c remains constant whatever the code rate ρ, and the information bit rate R_b varies as $(R_b)_{cod} = \rho R_c$. The required value of C/N_0 is:

$$(C/N_0)_{cod} = (E_b/N_0)_{cod}\,(R_b)_{cod}$$

The reduction $\Delta C/N_0$ is given by:

$$\begin{aligned}\Delta C/N_0 &= (C/N_0)_{nocod} - (C/N_0)_{cod} \\ &= [(E_b/N_0)_{nocod} - (E_b/N_0)_{cod}] - 10 \log \rho \\ &= G_{cod} - 10 \log \rho\,(\text{dB})\end{aligned}$$

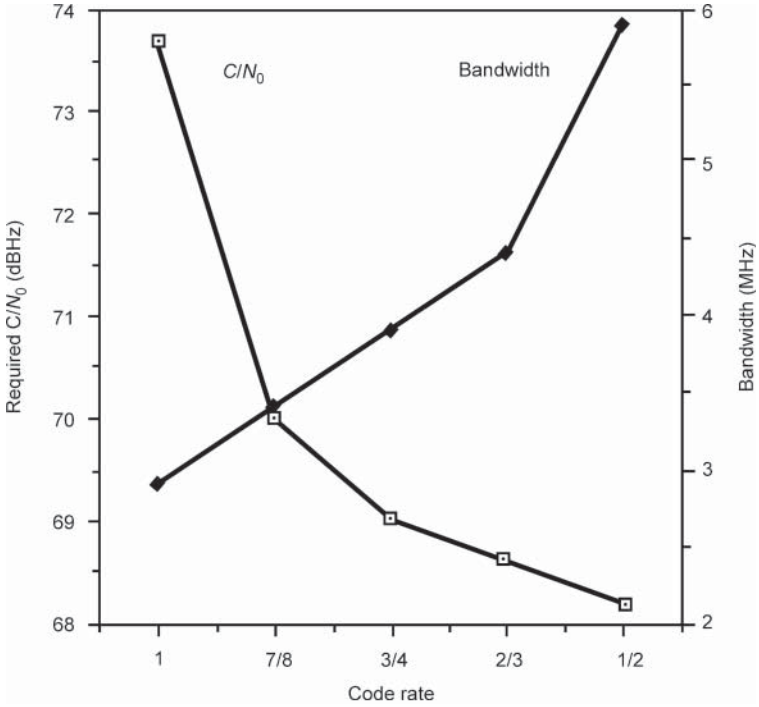


Figure 4.22 As the code rate decreases, less power is demanded but more bandwidth is required.

The reduction is equal to the decoding gain plus $-10 \log \rho$ (a positive gain in dB) that results from the reduction in the information bit rate.

Table 4.7 displays calculated values of the C/N_0 reduction $\Delta C/N_0$ for different code rates. These values do not take into account the implementation degradation. Depending on the selected code rate, one observes a reduction in the required value of C/N_0.

Table 4.7 Impact of coding for constant bandwidth: objective BER = 10^{-6}, BPSK modulation

Code rate ρ	Typical decoding gain (dB)	$-10 \log \rho$ (dB)	C/N_0 reduction $\Delta(C/N_0)$
1	0.0	0.0	0.0
7/8	3.6	0.6	4.2
3/4	4.6	1.3	5.9
2/3	5.0	1.8	6.8
1/2	5.5	3.0	8.5
1/3	6.0	4.8	10.8

Comparing Tables 4.6 and 4.7, one observes a higher reduction in C/N_0 for the constant-bandwidth case, as a result of reducing the information bit rate. This C/N_0 reduction can be used, for instance, to combat temporary link degradation due to rain, at the expense of a temporary

information capacity reduction on the considered link. Another application is adjusting the power requirement on the link to specific requirements for the transmit and receive equipment.

Example 4.1 Downlink coding with on-board regeneration

The concept of regenerative payload was introduced in Chapter 1. The availability of binary digits on board the satellite offers several opportunities. The payload implementation is discussed in Chapter 9. As the uplink and the downlink can use different modulation and coding formats, error-correcting coding can be used on either the up- or downlink. For the downlink, the encoder is located on board the satellite and is activated by telecommand (an order sent from an earth station). The link thus benefits from the decoding gain but, on the other hand, the transmission rate increases by a factor equal to the inverse of the coding ratio. This implies that the downlink is limited in power but not in bandwidth. If the link is limited in bandwidth, the transmission rate must be kept constant and, consequently, the information rate must be reduced (and, therefore, the capacity of that link). This reduction of throughput provides a margin on $(C/N_0)_D$ that is added to that provided by the decoding gain as explained in the previous section.

Let $(C/N_0)_1$ and $(C/N_0)_2$ be the values of $(C/N_0)_D$ without and with coding respectively. Hence:

$$(C/N_0)_1 = (E_b/N_0)_1 R_{b1}$$

where R_{b1}, the information rate, is equal to the rate R_c that modulates the carrier,

$$(C/N_0)_2 = (E_b/N_0)_2 R_{b2}$$

where

$$R_{b2} = \rho R_c$$

The margin realised here (in dB) is thus equal to

$$\begin{aligned}\text{Margin} &= \Delta(C/N_0)_D \\ &= (C/N_0)_1 - (C/N_0)_2 \\ &= [(E_b/N_0)_1 - (E_b/N_0)_2] - 10 \log \rho \\ &= \text{Decoding gain} + \text{gain provided by rate reduction.}\end{aligned}$$

For example, consider the use of a code with coding rate $\rho = 1/3$ and a decoding gain of 5 dB; at constant bandwidth, a margin on the required value of $(C/N_0)_D$ of 10 dB is obtained. The price to be paid is a reduction of two-thirds in the information rate and hence a reduction in the capacity of this downlink. This margin of 10 dB can be used, for example, to compensate for temporary degradation due to rain for links at 20 GHz (see Section 5.8).

4.4.3 Conclusion

Figure 4.23 shows the information bit rate R_b as a function of C/N_0 at a constant BER. Each curve in Figure 4.23 corresponds to a given transmission scheme, one with no coding and another with FEC. The segments *ab* and *cd* correspond to power-limited link operation. Any increase in R_b requires an increase in C/N_0. Once $R_b = R_{bmax}$, the whole available bandwidth B_a is utilised, and any increase in C/N_0 generates power margin but no further increase in the information bit rate; the link is now bandwidth limited.

Going from *a* to *c* illustrates implementing coding with variable bandwidth at a constant information bit rate. The reduction in C/N_0 from the no-coding scheme to the coding scheme is equal

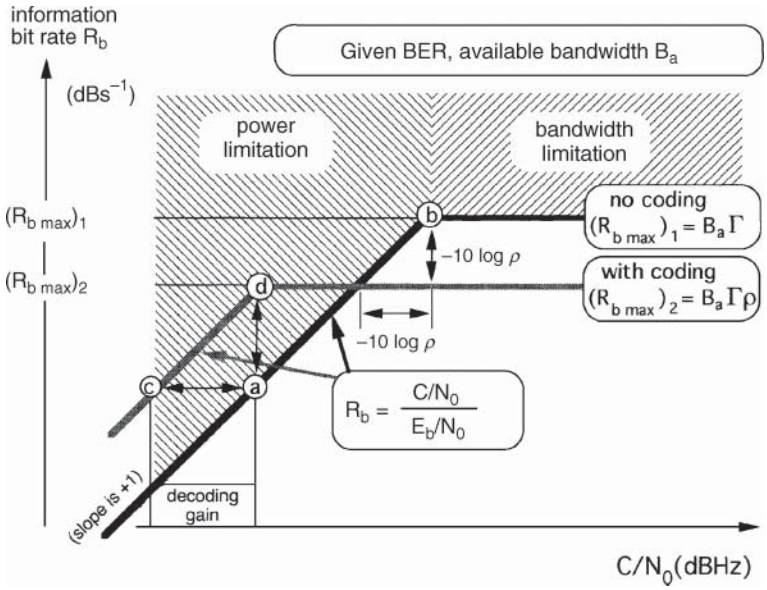


Figure 4.23 Information rate as a function of C/N_0 at constant bit error rate.

to the decoding gain G_{cod}. The utilised bandwidth expands from R_b/Γ to $R_b/\Gamma\rho$. The carrier power is reduced, whereas the occupied bandwidth is increased. This means exchanging power for bandwidth.

Going from *b* to *d* illustrates implementing coding with constant bandwidth. The reduction in C/N_0 from the no-coding scheme to the coding scheme is now equal to the sum of the decoding gain and $-100 \log \rho$.

Finally, going from *a* to *d* illustrates a case where the information bit rate can be increased at constant C/N_0, under the condition that the uncoded link is not bandwidth limited.

4.5 CODED MODULATION

The previous sections consider modulation and error correction in the transmitter as two separate processes. As the transmitter encoder adds redundancy bits, the transmitted bit rate R_c is higher than the information bit rate R_b, and this requires a larger carrier bandwidth. Using conventional QPSK, transmission of high bit rates such as 140 and 155 Mbit s^{-1} over typical 72 MHz satellite transponders is not feasible. A higher spectral efficiency could be obtained using modulations with larger alphabet such as 8PSK and 16QAM (see Figure 4.11) and 16APSK and 32APSK (see Figure 4.12). The transmitted symbol s_k at instant kT_s is a complex element from the signal constellation, $s_k = v_I(kT_s) + jv_Q(kT_s)$. The multilevel/phase modulated transmitted carrier $\omega_C t$ in the time interval $[KT_s, (k+1)\,T_s)]$ is $C(t) = v_I(kT_s)\cos(\omega_C t) - v_Q\,(KT_s)\sin(\omega_C t)$. Equivalently $C(t) = A\cos(\omega_C t + \theta_k)$, given $s_k = V\exp(j\theta_k)$, and $\theta_k = (2m_k + 1)\pi/M$, where $m_k = 0, \ldots, (M-1)$.

However, these modulation schemes require a higher E_b/N_0 for the required BER compared to QPSK, and hence more power on the link. This could be compensated for by using FEC coding. However, if the code is selected independently of the modulation, the overall power-bandwidth trade-off does not show a significant advantage over uncoded QPSK. Moreover, modulations

with large alphabet, particularly QAM modulation, suffer from the nonlinear characteristic of the satellite channel.

Coded modulation is a technique where FEC and modulation, instead of being performed in two separate steps, are merged into one process. Redundancy is achieved not by adding redundant bits as in the schemes described earlier, but by expanding the alphabet of the modulation with respect to common schemes such as BPSK and QPSK. Thus, to transmit n information bits per symbol duration T_s, a modulation based on an enlarged alphabet of $M = 2^m = 2^{n+1}$ symbols is used. Therefore $n = m - 1$ bits are transmitted per symbol instead of m and this technique results in a modulated carrier with slightly less spectral efficiency than M-PSK modulation, but a significant reduction in E_b/N_0 for the required BER. For instance, coded 8PSK can offer up to 6 dB reduction in E_b/N_0 compared to uncoded QPSK, for the same theoretical spectral efficiency (2 bit s^{-1} Hz^{-1}) [UNG-82].

Coded modulation conveys a sequence $\{s_k\}$ where s_k is a symbol from an M-ary alphabet at instant kT_s. All sequences are part of a specific set designed so that the minimum distance between all pairs of two sequences, called the *free distance* d_{free}, is as large as possible in order to reduce the error probability; d_{free} is defined by:

$$d_{free}^2 = \min_{\{s_k\} \neq \{s'_k\}} \left[\sum_k d^2(s_k, s'_k) \right]$$

where $d(s_k, s'_k)$ is the Euclidean distance between symbols s_k and s'_k.

Best performance in terms of asymptotic coding gain $G_{cod}(\infty)$ (coding gain when $E_b/N_0 \to \infty$) is achieved with maximum d_{free} and the smallest average number N_{free} of sequences at this distance.

The *asymptotic coding gain* is usually calculated with reference to an uncoded modulation that transmits the same average number of information bits per symbol duration Ts. Denoting by d_{unc} the minimum distance between all pairs of two symbols of the uncoded modulation, the asymptotic coding gain is given by:

$$G_{cod}(\infty) = 10 \log \left(\frac{d_{free}^2 / E_{cod}}{d_{unc}^2 / E_{unc}} \right) \quad \text{(dB)}$$

where E_{cod} and E_{unc} are the average signal energies of the coded and the uncoded schemes, respectively. The ratio E_{unc}/E_{cod} represents the loss due to the alphabet expansion caused by the enlarged redundant signal set of the coded modulation [FOR-84] and is less than or equal to 1. With PSK schemes, it is equal to 1 as all signals are of equal energy on a circle.

When N_{free} increases, the coding gain reduces. A rule of thumb states that every factor of 2 increase in N_{free} reduces the coding gain by about 0.2 dB.

There are two main classes of coded modulation:

— *Trellis-coded modulation* (TCM), where convolutional encoding is implemented
— *Block-coded modulation* (BCM), using block encoding

Further implementations include *multilevel trellis-coded modulation* (MLTCM) and *trellis-coded modulation using a multidimensional signal set* (multi-D TCM).

4.5.1 Trellis-coded modulation

With TCM, the set of sequences $\{s_k\}$ represents all the allowed paths of a trellis. In such a trellis, the nodes represent the encoder states and a branch between two states corresponds to one symbol.

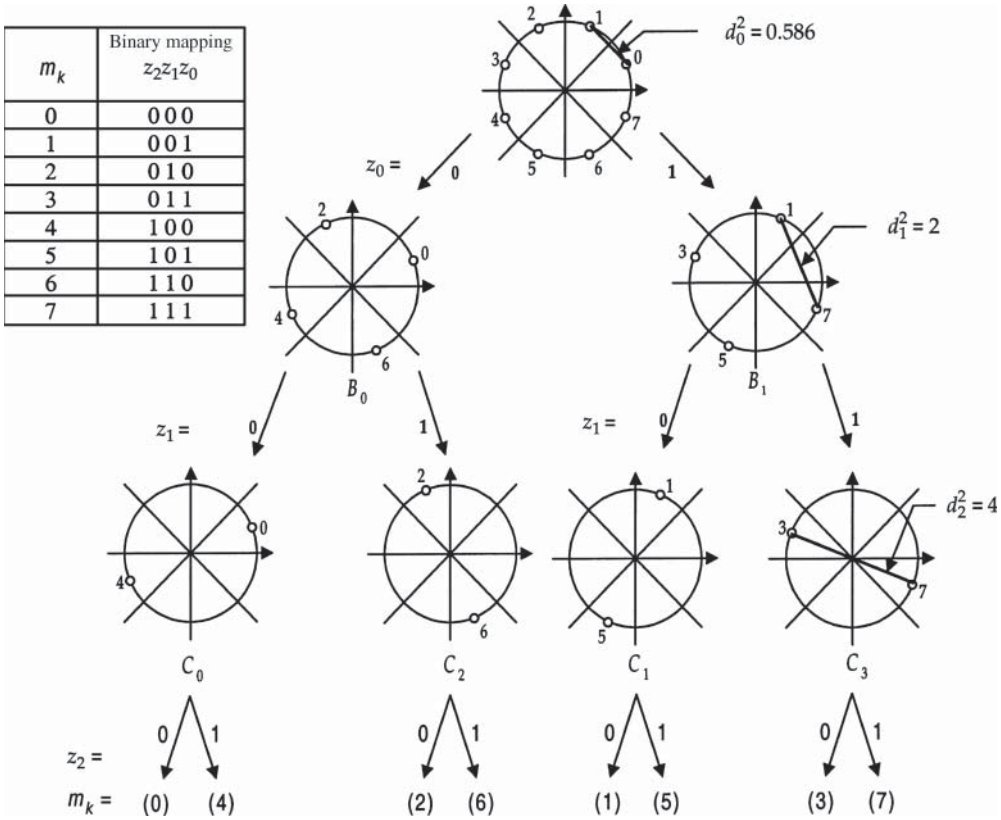


Figure 4.24 Set partitioning of an 8PSK signal constellation that maximises intraset distances d_i.

Symbols are assigned to each branch of the trellis according to rules proposed by Ungerboeck [UNG-87]. The first step consists of partitioning the signal constellation. Figure 4.24 considers 8PSK ($m = n + 1 = 3$) coded modulation, an appropriate choice for the satellite channel; the partitioning entails smaller subsets with maximally increasing intrasubset distances $d_{i+1} \geq d_i$. Each partition is two-way.

The second step is to assign M-ary symbols or subsets obtained from the set partitioning to each branch of the trellis in a way that maximises the free distance d_{free}. Figure 4.25 shows the corresponding trellis, where σ_0, σ_1, σ_2, and σ_3 represent the four possible encoder states. With convolutional encoding, the number of encoder states is 2^{k-1}, where K is the constraint length of the code. Each figure associated with a branch identifies a symbol in the 8PSK constellation. Branches originating from or merging in a given state are labelled with symbols from the first level (B_0 and B_1 in Figure 4.24) of the two-way partitioning tree. For example, the alphabet {0, 2, 4, 6} is associated with branches that originate from states σ_0 or σ_2, and branches merging in states σ_0 or σ_2. Parallel branches (branches originating from and terminating at the same state) are associated with symbols from the second step of the set partitioning (i.e. C_i, $i = 0$ to 3, in Figure 4.24).

The *trellis distance* d_{tr} is the minimal distance between two paths originating from and merging in a given state, apart from parallel branches (and therefore incorporating more than one branch). In Figure 4.25, the path indicated by a dashed line (branches 2, 1, 2) is at the trellis distance from

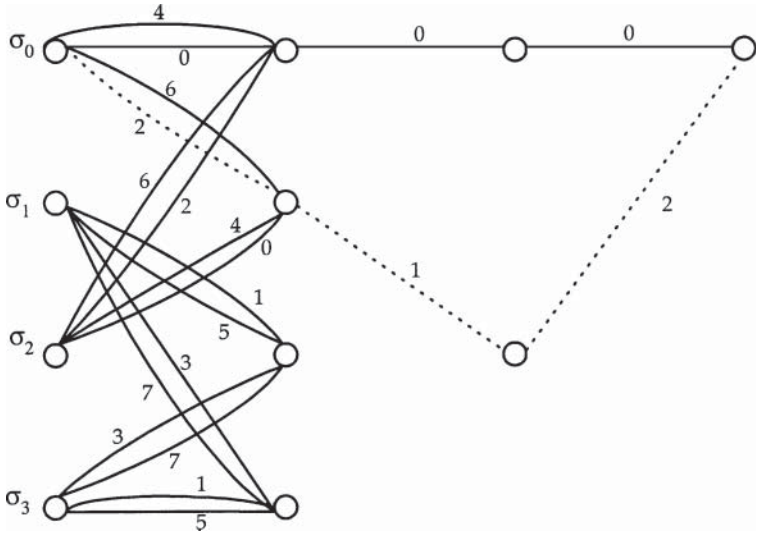


Figure 4.25 Trellis associated with the four-state 8PSK TCM of Figure 4.24.

the path (branches 0, 0, 0). According to the set partitioning of Figure 4.24, $d_{tr}^2 = d_1^2 + d_0^2 + d_1^2 = 4.586$. The distance d_{tr} is greater than the distance d_2 between two parallel branches, since $d_2^2 = 4.0$. Thus $d_{free}^2 = \min(d_2^2, d_{tr}^2) = 4$; d_{free}^2 is twice the square of the minimal distance of uncoded QPSK (where $d_{unc}^2 = 2$), and the asymptotic coding gain is $\gamma_{dB} = 3$ dB.

To evaluate the performance upper bound for practical values of E_b/N_0 (which are less than those pertaining to the asymptotic coding gain), one must compute the distance distribution of the code, which is not linear [BIG-84 ; ZEH-87].

The configuration of a TCM encoder is illustrated in Figure 4.26, where $\tilde{n}$ information bits (from b_1 to $b_{\tilde{n}}$) are encoded using a binary convolutional encoder and $n - \tilde{n}$ information bits (from $b_{\tilde{n}+1}$ to b_n) are left uncoded. The code rate of the convolutional encoder is $\tilde{n}/(\tilde{n}+1)$ and the code rate of the TCM is $n/(n+1)$.

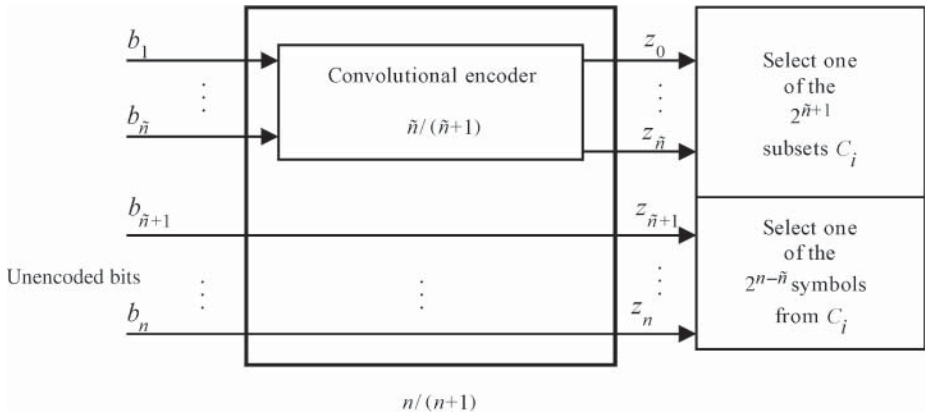


Figure 4.26 The usual implementation of a TCM encoder.

The set partitioning tree of Figure 4.24 is labelled by the encoder outputs z_0 to $z_{\tilde{n}}$ (here $\tilde{n} = 1$). These $\tilde{n} + 1 = 2$ output bits determine one of the $2^{\tilde{n}+1} = 2^2 = 4$ subsets up to the corresponding level $\tilde{n} + 1 = 2$ in the set partitioning tree. The $\tilde{n} - n = 1$ uncoded bits ($z_{\tilde{n}+1}$ to z_n i.e.z_3) determine one of the $2^{n-\tilde{n}} = 2$ symbols of this subset and condition the parallel branches in the trellis. The mapping between the 8PSK symbols and the encoder outputs z_0 to z_n is given as a table inset to Figure 4.24. The identification of the subset of the last level is made more secure using error protection provided by the convolutional encoding.

In the four-state trellis of Figure 4.25:

— $\tilde{n} = 1$
  The number of partitioning levels is 2.
  The binary encoder has code rate $\tilde{n}/(\tilde{n} + 1) = \frac{1}{2}$.
— $n = 2$
  The parallel branches are determined by the uncoded bit z_2.
  The TCM rate is $n/(n + 1) = 2/3$.

If one assumes that the spectral efficiency of 8PSK is 3 bit s^{-1} Hz^{-1} (maximum theoretical value), the overall spectral efficiency is 2 bit s^{-1} Hz^{-1}. Compared to uncoded QPSK, this TCM scheme has equal theoretical spectral efficiency, but offers a potential power saving of 3 dB (as d^2_{free} is twice that of QPSK, resulting in 3 dB asymptotic coding gain).

4.5.2 Block-coded modulation

With BCM, the transmitted set of sequences $\{s_k\}$ of L symbols of an M-ary signal set is obtained from binary block encoders. The bits that differentiate the nearest symbols in the mapping process are protected by the most powerful error-correcting code. Figure 4.27 takes an 8PSK modulation, where $M = 2^m = 8$ ($m = n + 1 = 3$), and illustrates a multilevel construction associated with the set partitioning tree of Figure 4.24 [IMA-77; POT-89]. This construction forms a binary array of $n + 1$ rows and L columns. Each column labels a symbol from an 8-ary signal set. A sequence of bits z_i at the ith level of the 8PSK partitioning tree is a code word of C_i. Hence the modulation entails a set of $n + 1$ binary block codes C_i, each with its own code rate $(\tilde{n}_i/L)$, and with minimal Hamming distance $\delta_i \geq \delta_{i+1}$, i = 0, … , n, in such a way that bits z_i are better protected than bits z_{i+1}. The normalised information bit rate, in bits/T_s (where T_s is the symbol duration), of the coded modulation is:

$$\eta = \sum_i \tilde{n}_i / L \quad (\text{bits}/T_s)$$

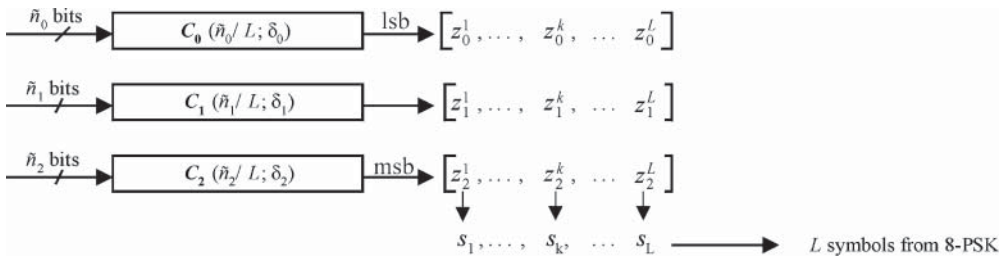


Figure 4.27 A block-coded 8PSK with a multilevel construction (lsb = least significant bit, msb = most significant bit).

According to the set partitioning distances of Figure 4.24, the free distance for the 8PSK BCM is:

$$d^2_{\text{free}} = \min\{\delta_0 \times d_0^2; \delta_1 \times d_1^2; \delta_2 \times d_2^2\}$$

4.5.3 Decoding coded modulation

Decoding is based on the maximum likelihood (ML) where the Euclidean distances between the received noisy sequence and the set of allowed sequences in the trellis are computed.

TCM is soft-decoded using the Viterbi algorithm [VIT-79], which implements an ML trellis search technique [FOR-73]. This technique identifies the closest sequence to the observed one. Generally, the complexity grows exponentially with the number of states in the trellis. Commonly used binary convolutional encoders are at rate 1/2. Higher-rate codes are derived by a puncturing technique [CAI-79], applied to the rate 1/2 code. The trellis of the latter is used to decode punctured codes.

Decoding BCM involves distance computation with each of the $2\sum_i \tilde{n}i$ code words. Since the trellis structure of BCM is not straightforward, ML decoding appears to be prohibitively complex and grows exponentially with $\sum_i \tilde{n}i$, which is usually large. An alternative approach is to apply ML to each constituent code C_i separately, in a cascaded structure, as shown in Figure 4.28, where $\{y_k\}$ is the received code word corrupted by noise. This multistage decoding [CAL-89] is made possible by the multilevel construction of the code. However, the decoder suffers from stage-to-stage error propagation. For values of E_b/N_0 high enough (typically 7 dB), the decoder performance nears ML decoding. For lower E_b/N_0, the number of nearest neighbours using the suboptimal strategy becomes larger than the number computed theoretically in the case of ML, and this results in a larger degradation [SAY-86].

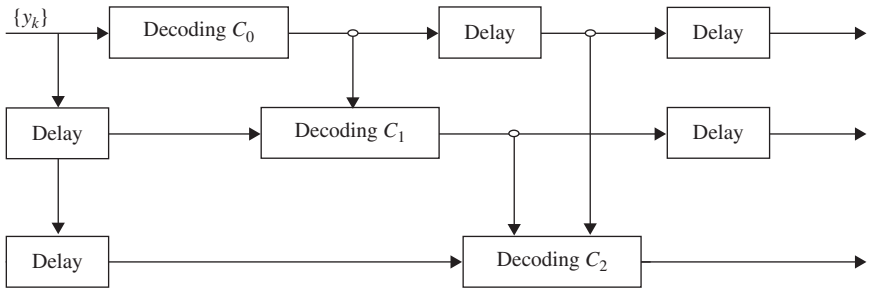


Figure 4.28 Multistage decoding for block-coded 8PSK.

4.5.4 Multilevel trellis-coded modulation

The multilevel construction of BCM allows simple multistage decoding with satisfactory performance. Hence multilevel construction appears of practical interest to implement efficient codes. This construction also offers the possibility of using available binary coding and decoding circuits (pragmatic codes). On the other hand, the trellis structure is well adapted to ML Viterbi soft decoding. The use of convolutional codes with the multilevel construction benefits from both advantages.

MLTCM achieves significant practical decoding gain (2–3 dB at BER = 10^{-5}) with appealing spectral efficiency (about 2 bit s^{-1} Hz^{-1}) and low decoding complexity [WU-92; KAS-90]. The

research area is still open in MLTCM, especially in improving the multistage decoding process through the choice of codes and the associated soft-decoding circuits with a proper trade-off between performance and complexity.

A class of coding/decoding technique, called *turbo-codes* as per the iterative design of the decoder [BER-93] (also known as *parallel concatenation codes*), performs near channel capacity (BER $< 10^{-5}$ for $E_b/N_0 = 0.7$ dB) with simple constituent codes. The iterative decoding process of turbo-code decoding fits well in the multistage decoding of MLTCM [ISA-00], which then benefits from the high decoding gain of turbo-codes. Turbo TCMs are also considered [BEN-96].

4.5.5 TCM using a multidimensional signal set

Coded modulation is also considered with a multidimensional signal set (multi-D TCM). A candidate scheme for satellite communications is the one generated from L 8PSK signal sets and is denoted $L \times 8$PSK. The multi-D signal is obtained by sending L consecutive signals of the 8PSK signal set. The structure of multi-D TCM is the same as depicted in Figure 4.26. The encoder of overall rate $n/(n+1)$ is combined with an $L \times M$PSK signal set, and thus the average number of information bits per 2D symbol is n/L bits per symbol. Partitioning the signal set requires particular attention; efficient methods are described in [PIE-90; WEI-89]. In addition, the mapping process involves modulo-M adders and depends tightly on the set partitioning.

Multi-D 8PSK TCM displays advantages compared to conventional 2D 8PSK TCM [PIE-90]:

— Flexibility in achieving a variety of fractional average number of information bits per symbol duration T_s.
— Insensitivity to discrete phase rotations of the signal set.
— Suitability for use as inner codes in a concatenated coding scheme due to their symbol-oriented nature.
— Higher decoding speed as the decoder decodes n bits at each decision step of the algorithm, given that the encoder rate in a multi-D TCM is larger than for 2D TCM (with n up to 15 for some multi-D codes).

Multi-D 8PSK trellis codes have been investigated for high-rate telemetry as inner codes in a concatenated scheme and have been proposed for the satellite news-gathering (SNG) service in connection with provision of digital video broadcasting (DVB) services.

4.5.6 Performance of coded modulations

Figure 4.29 presents the BER versus $E_b = N_0$ obtained from simulations for different types of coded modulations with spectral efficiency equal to that of uncoded QPSK, shown here for reference. The illustrated modulations of similar complexity are four-state trellis-coded 8PSK modulation (TCM), block-coded 8PSK modulation (BCM), multilevel trellis-coded 8PSK modulation (MLTCM), and six-dimensional 8PSK trellis-coded modulation (6D TCM). The performance of MLTCM considers interstage interleaving, which prevents stage-to-stage propagation of errors in the decoding process of Figure 4.28.

The decoding gain at BER $= 10^{-5}$ is between 2.5 and 3.5 dB with respect to uncoded QPSK. This provides potential power savings on the satellite link. Multilevel and multidimensional coded modulations can also be thought of as a means to provide increased spectral efficiency (up to 20% more) with reduced or no power saving.

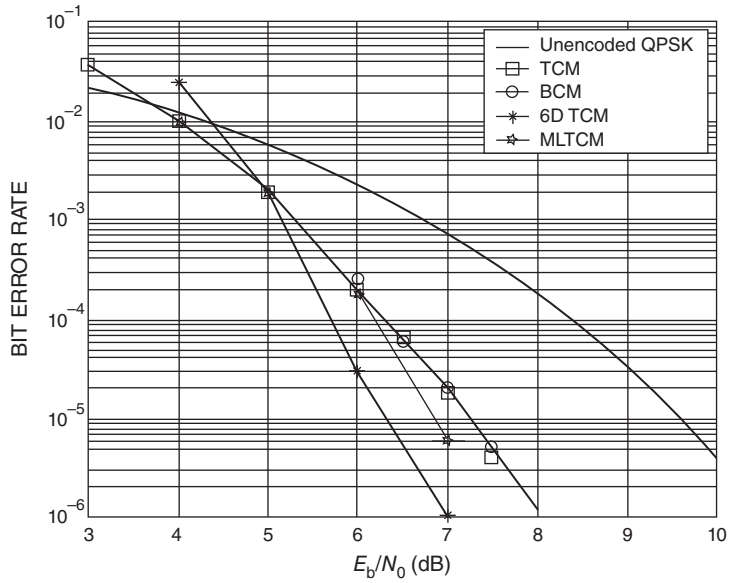


Figure 4.29 Performance of modulations with equal spectral efficiency. For a given BER, the difference between the required value of E_b/N_0 for unencoded QPSK and the required value for the coded modulation scheme (the decoding gain) indicates potential power saving at constant bandwidth utilisation.

4.6 END-TO-END ERROR CONTROL

The previous techniques for error control offer quasi-error-free (QEF) transmission (BER $< 10^{-10}$) at the expense of power or bandwidth. QEF transmission can also be achieved by using a different technique based on end-to-end error control, implying retransmission of information identified as being corrupted at the receiving end at the expense of a variable delivery delay. This is called automatic repeat request (ARQ). Due to the variable delay, this technique applies particularly to data packet transmission. The decoder detects errors but does not correct them: a retransmission request is sent to the transmitter. It is, therefore, necessary to provide a return channel. This can be a satellite or terrestrial channel. The use of error-detecting codes requires the ability to control the source throughput and results in a variable delivery delay. These disadvantages are compensated for by the simplicity of decoder realisation, the possibility of adapting to varying error statistics, and low error rates.

Three basic techniques are employed [BHA-81; MAR-95] (see Figure 4.30):

— Retransmission with stop and wait or reception acknowledgement (Stop-and-wait ARQ [ARQ-SW])
— Continuous retransmission (Go-Back-N ARQ [ARQ-GB])
— Selective retransmission (Selective-repeat ARQ [ARQ-SR])

The performance is measured in terms of efficiency, expressed as the ratio of the mean number of information bits transmitted in a given time interval to the total number of bits that could be transmitted during the same time.

Consider a digital satellite link with a capacity of $R = 48$ kbit s^{-1}. The round-trip return time is taken to be $T_{RT} = 600$ ms. The bit error rate is BER $= 10^{-4}$. Transmission is in blocks of $n = 1000$ bits.

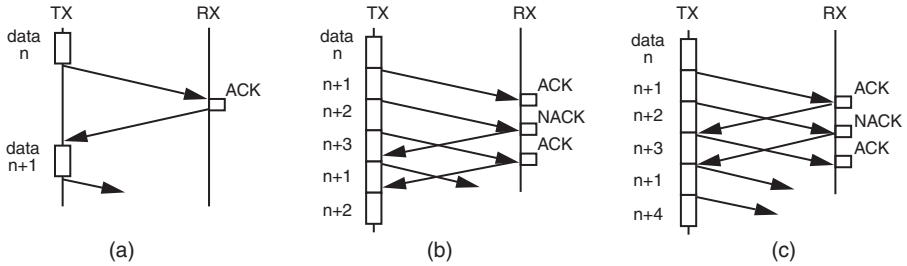


Figure 4.30 Error detection with retransmission: (a) Stop-and-wait ARQ; (b) Go-Back-N ARQ; (c) Selective-repeat ARQ.

The block error probability is $P_B = 1-(1-\text{BEP})^n = 1-\exp(-n\text{BEP})$ for $n\text{BEP} \ll 1$, hence $P_B = 0.1$. From [MAR-95], assuming that any error is detected:

— Efficiency in ARQ-SW: $\eta = n(1-P_B)/RT_{RT} = 0.03$
— Efficiency in ARQ-GB (N): $\eta = n(1-P_B)/[n(1-P_B) + RT_{RT}P_B] = 0.2$
— Efficiency in ARQ-SR: $\eta = 1-P_B = 0.9$

The increase in efficiency from one technique to another is accompanied by an increase in the complexity of the equipment.

4.7 DIGITAL VIDEO BROADCASTING VIA SATELLITE (DVB-S)

The European Telecommunications Institute (ETSI) is a nonprofit organisation that creates standards for different areas of telecommunications. A standardised radio interface enables a mass market for consumer reception devices. Having in mind the many past issues resulting from the different analogue TV standards and its multiple variations, most of the actors (broadcasters, service providers, operators, equipment and chip manufacturers, etc.) worked together at the end of the 1980s to define a DVB standard. This standard has been broken down into different versions depending on the specific properties of the transmission channel that conditions the physical layer (PL) characteristics: DVB-T for terrestrial digital TV, DVB-C for cable, and DVB-S for satellite. Later standards have been introduced: DVB-RCS for the return channel, DVB-S2 (the second generation of DVB-S), DVB-H for handheld terminals, DVB-SH for satellite handheld terminals, etc.

The ETSI documents are published as the following four categories:

(i) *Technical report (TR)*: Typically a set of guidelines for the implementation of a more normative specification or standard. It is approved by the ETSI Technical Committee that proposes the document.
(ii) *Technical specification (TS)*: A document that can contain normative text, i.e. mandatory text such as 'shall'; approved by the ETSI Technical Committee that proposes the document; generally forming a stepping stone to a more stable document(s).
(iii) *ETSI specification (ES)*: A document approved by the entire ETSI membership, not just the Technical Committee proposing it. It is a more stable document than either a TR or a TS.
(iv) *European standard (EN)*: The highest-ranking ETSI publication approved by the national standards organisations of Europe. It can be and is often included in European and national legislation.

This section provides a brief introduction to the DVB-S system based on [ETSI-97]. The DVB-S system provides direct-to-home (DTH) services for consumer integrated receiver decoders (IRD), as well as collective antenna systems (satellite master antenna television [SMATV]) and cable television head-end stations. The overview covers the physical layer that comprises adaptation, framing, coding, interleaving, and modulation, and discusses error performance requirements to achieve quality of service (QoS) targets.

Although the DVB-S standard was designed initially for satellite digital television services, the physical layer of DVB-S can carry streams of packetised data of any kind. Mass-market production, and the availability of different equipment and related building blocks, makes the standard appealing for a lot of applications other than the transmission of TV signals, such as Internet traffic.

4.7.1 Transmission system

The transmission system consists of the functional block of equipment to transport baseband TV signals in the format of the Motion Picture Expert Group (MPEG-2) transport stream over the satellite channel. The transmission system carries out the following processes on the data stream:

— Transport multiplex adaptation and randomisation for energy dispersal
— Outer coding, i.e. Reed–Solomon (RS)
— Convolutional interleaving
— Inner coding (i.e. punctured convolutional code)
— Baseband shaping for modulation
— Modulation

Digital satellite TV services have to be delivered to home terminals with rather small antennas (around 0.6 m) that translate typically into a power-limited downlink. To achieve a high power efficiency without excessively penalising the spectrum efficiency, DVB-S uses QPSK modulation and the concatenation of convolutional and RS codes. The convolutional code can be configured flexibly, allowing the optimisation of the system performance for a given satellite transponder bandwidth.

DVB-S is directly compatible with MPEG-2-coded TV signals (defined by ISO/IEC DIS 13818-1). The modem transmission frame is synchronous with MPEG-2 multiplex transport packets. If the received signal is above the considered threshold for the carrier-to-noise power ratio, C/N, the FEC technique can provide a QEF quality target. The QEF means BER less than 10^{-10}–10^{-11} at the input of the MPEG-2 demultiplexer.

4.7.1.1 Input stream scrambling

The DVB-S input stream is the MPEG-2 transport stream (MPEG-TS) from the transport multiplexer. The packet length of the MPEG-TS is 188 bytes. This includes one sync-word byte (i.e. 47_{HEX}). The processing order at the transmitting side starts from the most significant bit (MSB).

In order to comply with ITU-R Radio Regulations and to ensure adequate binary transitions, the data of the input MPEG-2 multiplex is randomised in accordance with the configuration depicted in Figure 4.31.

The polynomial for the pseudorandom binary sequence (PRBS) generator is defined as:

$$1 + X^{14} + X^{15}$$

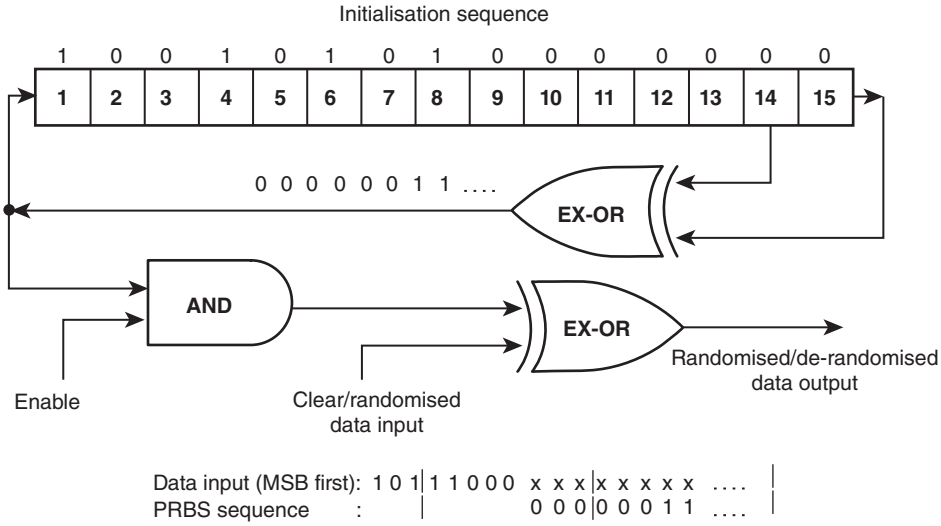


Figure 4.31 Randomiser/derandomiser schematic diagram.

Loading the sequence 100101010000000 into the PRBS registers is initiated at the start of every eight transport packets. To provide an initialisation signal for the descrambler, the MPEG-2 sync byte of the first transport packet in a group of eight packets is bit-wise inverted from 47_{HEX} to $B8_{HEX}$. This process is referred to as *transport multiplex adaptation*.

The first bit at the output of the PRBS generator is applied to the first bit (i.e. MSB) of the first byte following the inverted MPEG-2 sync byte (i.e. $B8_{HEX}$). To aid other synchronisation functions, during the MPEG-2 sync bytes of the subsequent seven transport packets, the PRBS generation continues but its output is disabled, leaving these bytes unrandomised. Thus, the period of the PRBS sequence is 1503 bytes.

The randomisation process is also active when the modulator input bit stream is non-existent, or when it is non-compliant with the MPEG-2 transport stream format (i.e. 1 sync byte + 187 packet bytes). This is to avoid the emission of an unmodulated carrier from the modulator, as energy concentrated on the carrier frequency can cause interference with neighbouring satellites.

4.7.1.2 Reed–Solomon (RS) outer coding, interleaving, and framing

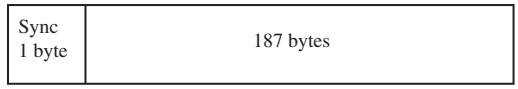
The framing organisation is based on the input packet structure shown in Figure 4.32a. The Reed–Solomon RS(204, 188, T = 8) shortened code, from the original RS(255, 239, T = 8) code, is applied to each randomised transport packet (188 bytes) of Figure 4.32b to generate an error-protected packet (see Figure 4.32c). T is the number of bytes that can be corrected in RS error-protected packet. Reed–Solomon is applied to the packet sync byte, either non-inverted (i.e. 47_{HEX}) or inverted (i.e. $B8_{HEX}$).

The code generator polynomial is:

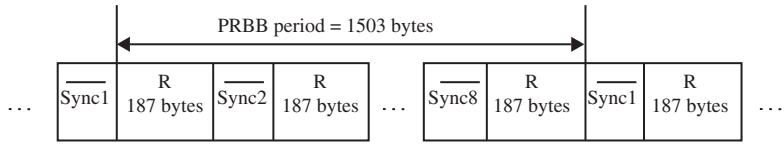
$$g(x) = (x + \lambda^0)(x + \lambda^1)(x + \lambda^2) \ \ldots \ (x + \lambda^{15}), \text{where } \lambda = 02_{HEX}$$

The field generator polynomial is:

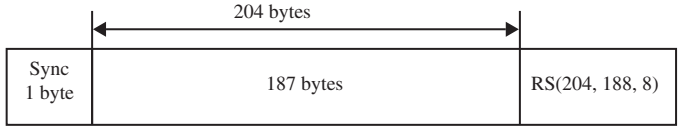
$$p(x) = x^8 + x^4 + x^3 + x^2 + 1$$



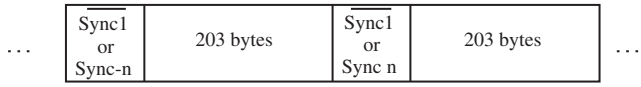
(a) MPEG-2 transport multiplexing (MUX) packet



(b) Randomised transport packets: Sync bytes and randomised sequence R.



(c) Reed-Solomon RS(204, 188, T=8) error protected packet



(d) Interleaved frames with interleaving depth of I = 12 bytes

Note: $\overline{\text{Sync1}}$ is not randomised complemented sync byte;
Sync-n is not randomised sync byte, where n = 2, 3, ..., 8.

Figure 4.32 Framing structure.

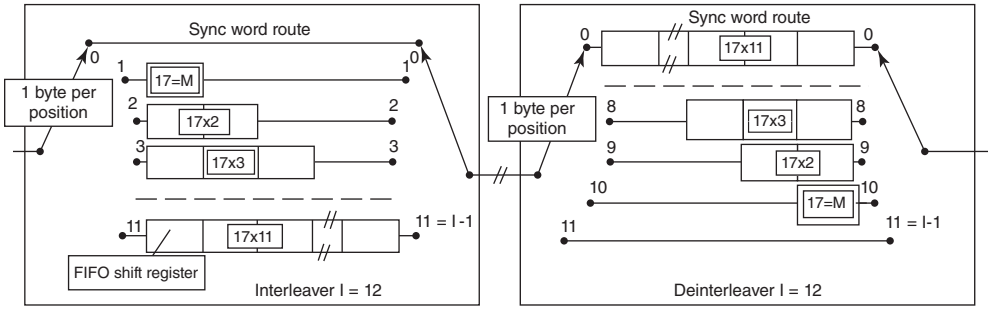


Figure 4.33 Convolutional interleaver and deinterleaver.

The shortened Reed–Solomon code is implemented by adding 51 bytes, all set to zero, before the information bytes at the input of a (255, 239) encoder. After the RS coding procedure, these null bytes are discarded.

Following the conceptual scheme of Figure 4.33, convolutional interleaving with depth I = 12 is applied to the error protected packets (see Figure 4.32c). This produces an interleaved frame (see Figure 4.32d).

The interleaved frame consists of overlapping error-protected packets and is delimited by inverted or non-inverted MPEG-2 sync bytes (preserving the periodicity of 204 bytes).

The interleaver consists of I = 12 branches, cyclically connected to the input byte stream by the input switch. Each branch is a first-in–first-out (FIFO) shift register, with depth (M_j) cells (where $M = 17 = N/I$, $N = 204$ [error protected frame length], $I = 12$ [interleaving depth], and j is the branch index). The cells of the FIFO contain 1 byte, and the input and output switches are synchronised. For synchronisation purposes, the sync bytes and the inverted sync bytes are always routed in branch 0 of the interleaver (corresponding to a null delay).

The deinterleaver is similar, in principle, to the interleaver, but the branch indexes are reversed (i.e. j = 0 corresponds to the largest delay). The deinterleaver synchronisation is carried out by routeing the first recognised sync byte in the 0 branch.

4.7.1.3 Inner convolutional coding

DVB-S allows for a range of punctured convolutional codes, based on a rate 1/2 convolutional code with constraint length K = 7. This allows selection of the most appropriate level of error correction for a given service data rate. It allows convolutional coding with code rates of 1/2, 2/3, 3/4, 5/6, and 7/8. Table 4.8 gives the definition of punctured convolutional code.

Table 4.8 Punctured code definition. Original code: K = 7; G_1 (X) = 171_{OCT}; G_2 (Y) = 133_{OCT}

	P				
Code rate	X	Y	I	Q	d_{free}
1/2	1	1	X_1	Y_1	10
2/3	1 0	1 1	X_1 Y_2 Y_3	Y_1 X_3 Y_4	6
3/4	1 0 1	1 1 0	X_1 Y_2	Y_1 X_3	5
5/6	1 0 1 0 1	1 1 0 1 0	X_1 Y_2 Y_4	Y_1 X_3 X_5	4
7/8	1 0 0 0 1 0 1	1 1 1 1 0 1 0	X_1 Y_2 Y_4 Y_6	Y_1 Y_3 X_5 X_7	3

1 = transmitted bit.
0 = non-transmitted bit.

4.7.1.4 Baseband shaping and modulation

DVB-S employs conventional Gray-coded QPSK modulation with direct mapping (no differential coding). Prior to modulation, the I and Q signals (mathematically represented by a succession of Dirac delta functions spaced by the symbol duration $Ts = 1/Rs$, with appropriate sign) are square-root-raised-cosine filtered. The value of 0.35 is selected as the roll-off factor (α).

4.7.2 Error performance requirements

Table 4.9 gives the modem BER versus E_b/N_0 performance requirements. The figures of E_b/N_0 refer to the useful bit rate before RS coding and include a modem implementation margin of 0.8 dB and the noise bandwidth increase due to the outer code (10 log 188/204 = 0.36 dB). QEF means less than one uncorrected error event per hour, corresponding to BER = 10^{-10}–10^{-11} at the input of the MPEG-2 demultiplexer.

Table 4.9 BER versus E_b/N_0 performance requirements

Inner code rate	Required E_b/N_0 for BER = 2×10^{-4} after Viterbi QEF after Reed–Solomon
1/2	4.5
2/3	5.0
3/4	5.5
5/6	6.0
7/8	6.4

4.8 SECOND GENERATION DVB-S (DVB-S2)

The DVB-S standard uses QPSK modulation and concatenated convolutional and Reed–Solomon (RS) channel coding. It has been adopted by most satellite operators worldwide for television and data broadcasting services. Digital satellite transmission technology has evolved significantly in several areas since the first development of the DVB-S standard in 1994. The first major release of the DVB-S standard was published by ETSI in August 1997 [ETSI-97] to support digital and high definition television (HDTV) broadcasting services over satellite. In November 2014, the second generation DVB-S (DVB-S2) was published with improvement on framing structure, channel coding, and modulation systems for broadcasting, interactive services (IS), news gathering, and other broadband satellite applications, as Part 1: DVB-S2 of the standard [ETSI-14]; an extension to DVB-S2 was published in February 2015 as Part 2: DVB-S2X [ETSI-15b]. Without going into too many details of the standard, this section provides a brief summary of DVB-S2 and DVB-S2X new technology, transmission system architecture, and performance.

4.8.1 New technology in DVB-S2

DVB-S2 makes use of the new developments in technology and future applications of broadband satellite applications. The main features can be summarised as the following:

— New channel coding schemes to achieve a capacity gain on the order of 30%.
— Variable coding and modulation (VCM) to provide different levels of error protection to different service components (e.g. SDTV and HDTV, audio, multimedia).
— Extended flexibility to cope with other input data formats (such as multiple transport streams or generic data formats in addition to the single MPEG transport stream [MPEG-TS] in DVB-S) without significant complexity increase.

In the case of interactive and point-to-point applications, the VCM functionality is combined with the use of return channels to achieve ACM. This technique provides dynamic link adaptation to propagation conditions, targeting each individual receiving terminal. ACM systems promise satellite capacity gains of more than 30%. Such gains are achieved by informing the satellite uplink station of the channel condition (e.g. the value of carrier power to-noise and interference power ratio, $C/(N+I)$) of each receiving terminal via the satellite or terrestrial return channels.

DVB-S2 makes use of new technology in the following functions:

— Stream adapter, suitable for operation with single and multiple input streams of various formats (packetised or continuous).
— FEC based on LDPC codes concatenated with BCH codes, allowing QEF operation at about 0.7–1 dB from the Shannon limit.
— Wide range of code rates (from 1/4 up to 9/10).
— Four constellations (QPSK, 8PSK, 16APSK, 32APSK), ranging in spectrum efficiency from 2 to 5 bit s^{-1} Hz^{-1}, optimised for operation over nonlinear transponders.
— Three spectrum shapes with roll-off factors 0.35, 0.25, and 0.20.
— ACM functionality, optimising channel coding, and modulation on a frame-by-frame basis.

DVB-S2 has also been designed to support a wide range of broadband satellite applications, including:

— *Broadcast services* (*BS*): Digital multi-programme television (TV) and HDTV for primary and secondary distribution in the fixed-satellite service (FSS) and broadcast satellite service (BSS) bands. BS has two modes: non-backward-compatible broadcast services (NBC-BS) allows exploitation of the full benefit of the DVB-S2 but is not compatible with DVB-S; backward-compatible broadcast services (BC-BS) is backward-compatible with DVB-S to give time for migration from DVB-S to DVB-S2.
— *Interactive services* (*IS*): Data services including Internet access for providing interactive services to consumer IRD and to personal computers, where DVB-S2's forward path supersedes the current DVB-S for interactive systems. The return path can be implemented using various DVB interactive systems, such as DVB-RCS (ETSI-09), DVBRCP (ETS-300-801), DVB-RCG (EN-301-195), and DVB-RCC (ES-200-800).
— *Digital TV contribution and satellite news gathering* (*DTVC/DSNG*): Temporary and occasional transmission with short notice of television or sound for broadcasting purposes, using portable or transportable uplink earth stations. DTVC applications by satellite consist of point-to-point or point-to-multipoint transmissions, connecting fixed or transportable uplink and receiving stations. They are not intended for reception by the general public.
— *Professional services* (*PS*): Data content distribution/trunking and other professional applications for point-to-point or point-to-multipoint, including interactive services to professional head-ends, which redistribute services over other media. Services may be transported in (single or multiple) generic stream (GS) format.

Digital transmissions via satellite are affected by power and bandwidth limitations. DVB-S2 helps to overcome these limits by making use of transmission modes (FEC coding and modulations), giving different trade-offs between power and spectrum efficiency.

For some specific applications (e.g. broadcasting), modulation techniques, such as QPSK and 8PSK with their quasi-constant envelope, are appropriate for operation with saturated satellite power amplifiers (in a single carrier per transponder configuration). When higher power margins are available, spectrum efficiency can be further increased to reduce bit delivery cost. In these cases, 16APSK and 32APSK can also operate in single carrier mode close to satellite high power amplifier (HPA) saturation if linearisation by predistortion techniques is implemented.

DVB-S2 is compatible with MPEG-2 and MPEG-4 coded TV services (ISO/IEC 13818-1), with a transport stream packet multiplex. All service components are TDM on a single digital carrier.

4.8.2 Transmission system architecture

The DVB-S2 system consists of a number of functional blocks of equipment performing the adaptation of the baseband digital signals from the output of one or more MPEG transport stream multiplexers (ISO/IEC 13818-1) or one or more generic data sources to the satellite channel characteristics. Data services may be transported in transport stream format according to (EN-301-192) (e.g. using multi-protocol encapsulation [MPE]) or GS format.

DVB-S2 provides a QEF quality target of 'less than one uncorrected error event per transmission hour at the level of a 5 Mbit s^{-1} single TV service decoder', approximately corresponding to a transport stream packet error ratio (PER) of less than 10^{-7} before de-multiplexer.

Figure 4.34 illustrates the following function blocks in a DVB-S2 system:

— *Mode adaptation* is application dependent. It provides the following function blocks:
  — Input stream interfacing.
  — Input stream synchronisation (optional).
  — Null-packet deletion (for ACM and transport stream input format only). CRC-8 coding for error detection at packet level in the receiver (for packetised input streams only).
  — Merging of input streams (for multiple input stream modes only) and slicing into data fields.
  — Appending a baseband header in front of the data field, to notify the receiver of the input stream format and mode adaptation type. Note that the MPEG multiplex transport packets may be asynchronously mapped to the baseband frames (BB frames).
— *Stream adaptation* has two functions:
  — Padding to complete a BB frame.
  — BB frame scrambling.
— *FEC encoding* is carried out by two coding functions and one interleaving function:
  — BCH outer codes.
  — LDPC inner codes (rates 1/4, 1/3, 2/5, 1/2, 3/5, 2/3, 3/4, 4/5, 5/6, 8/9, 9/10).
  — Interleaving applied to FEC coded bits for 8PSK, 16APSK, and 32APSK.
— *Mapping* maps the bit stream of the FEC into QPSK, 8PSK, 16APSK, and 32APSK constellations depending on the application area. Gray mapping of constellations is used for QPSK and 8PSK.
— *Physical layer* (*PL*) *framing* is used for synchronisation with the FEC frames, to provide the following functions:
  — PL scrambling for energy dispersal.
  — Dummy PLFRAME insertion for when no useful data is ready to be sent on the channel.
  — PL signalling and pilot symbol insertion (optional).
— *Baseband filtering and quadrature modulation* shapes the signal spectrum (squared-root raised cosine, roll-off factor $\alpha = 0.35$, 0.25, or 0.20) and generates the RF signal.

4.8.3 Error performance

The error performance is described to meet the QEF requirements. Table 4.10 illustrates the error performance provided in the DVB-S2 standard, as a function of the ratio of the average energy per transmitted symbol, E_s, to the noise power spectral density N_0 (E_s/N_0, expressed in dB). The performance is obtained by computer simulation, assuming perfect carrier and synchronisation recovery, no phase noise.

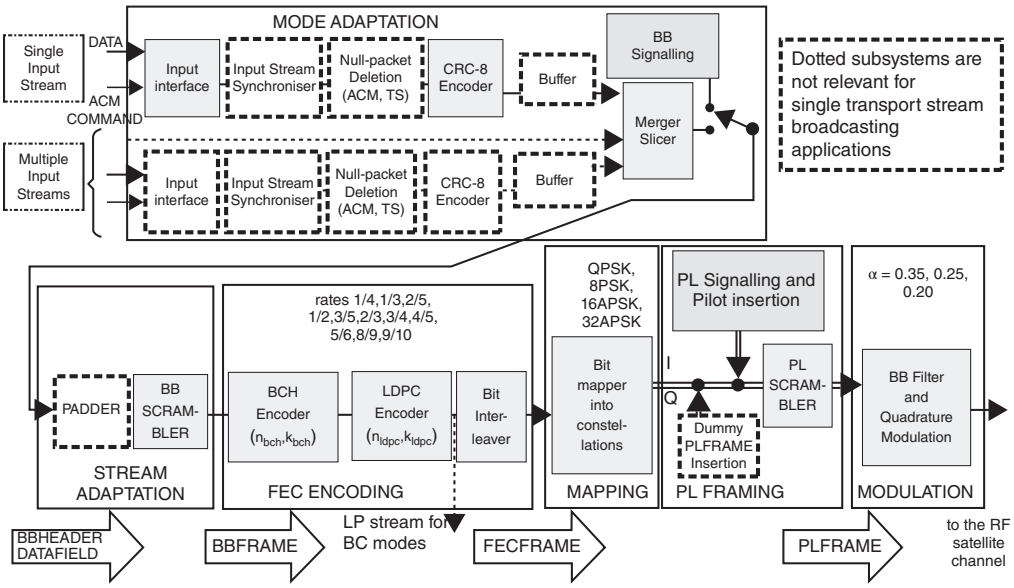


Figure 4.34 Functional block diagram of the DVB-S2 system.

Table 4.10 E_s/N_0 quasi-error-free performance (PER = 10^{-7})

Mode	Spectral efficiency	Ideal E_s/N_0 (dB) for FECFRAME length 64 800
QPSK 1/4	0.49	−2.35
QPSK 1/3	0.66	−1.24
QPSK 2/5	0.79	−0.30
QPSK 1/2	0.99	1.00
QPSK 3/5	1.19	2.23
QPSK 2/3	1.32	3.10
QPSK 3/4	1.49	4.03
QPSK 4/5	1.59	4.68
QPSK 5/6	1.65	5.18
QPSK 8/9	1.77	6.20
QPSK 9/10	1.79	6.42
8PSK 3/5	1.78	5.50
8PSK 2/3	1.98	6.62
8PSK 3/4	2.23	7.91
8PSK 5/6	2.48	9.35
8PSK 8/9	2.65	10.69
8PSK 9/10	2.68	10.98
16APSK 2/3	2.64	8.97
16APSK 3/4	2.97	10.21
16APSK 4/5	3.17	11.03
16APSK 5/6	3.30	11.61
16APSK 8/9	3.52	12.89
16APSK 9/10	3.57	13.13
32APSK 3/4	3.70	12.73
32APSK 4/5	3.95	13.64
32APSK 5/6	4.12	14.28
32APSK 8/9	4.40	15.69
32APSK 9/10	4.45	16.05

NOTE: Given the system spectral efficiency Γ_{tot}, the ratio between the energy per information bit and single-sided noise power spectral density $E_b/N_0 = E_s/N_0 - 10\log_{10}(\Gamma_{tot})$.

PER is the ratio between the useful MPEG transport stream packets (188 bytes) correctly received and affected by errors, after FEC.

The standard also suggests that, for short FEC frames (FECFRAME), an additional degradation of 0.2–0.3 dB has to be taken into account; for calculating link budgets, specific satellite channel impairments should be taken into account. Spectral efficiencies (per unit symbol rate) are computed for normal FEC frame length and no pilots.

4.8.4 FEC encoding

Here we provide an introduction to the forward error encoding scheme, which consists of the BCH multiple error correction binary block code as the outer coding method and the LDPC as the inner coding method recommended by the ETSI in the DVB-S2 standard [ETSI-14; ETSI-15b].

In DVB-S2, this is considered a subsystem performing outer coding (BCH), inner coding (LDPC) and bit interleaving. The input stream is composed of baseband frames (BBFRAMEs) and the output stream of FEC frames (FECFRAMEs).

Each BBFRAME (K_{bch} bits) is processed by the FEC coding subsystem to generate a FECFRAME (n_{ldpc} bits). The parity check bits (BCHFEC) of the systematic BCH outer code are appended after the BBFRAME, and the parity check bits (LDPCFEC) of the inner LDPC encoder are appended after the BCHFEC field, as shown in Figure 4.35.

Table 4.11 gives the FEC coding parameters for the normal FECFRAME (n_{ldpc} = 64 800 bits) and Table 4.12 for the short FECFRAME (n_{ldpc} = 16 200 bits).

4.8.4.1 Outer encoding (BCH)

A t-error correcting BCH (N_{bch}, K_{bch}) code is applied to each BBFRAME (K_{bch}) to generate an error-protected packet. The BCH code parameters for n_{ldpc} = 64 800 are given in Table 4.11 and for n_{ldpc} = 16 200 in Table 4.12.

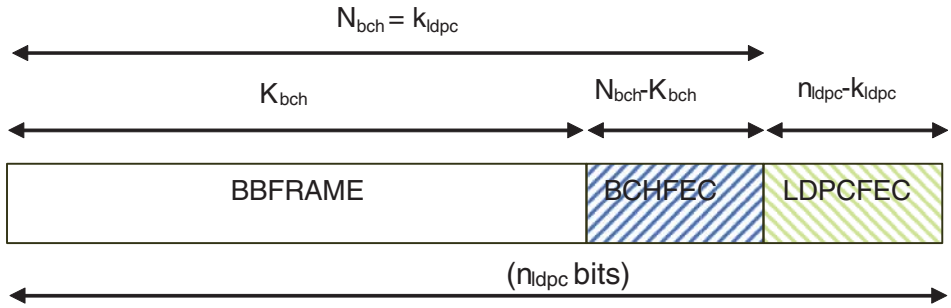


Figure 4.35 Format of data before bit interleaving (nldpc = 64 800 bits for normal FECFRAME, nldpc = 16 200 bits for short FECFRAME); K_{bch} is the number of bits of BCH uncoded block; N_{bch} number of bits of BCH coded block; k_{ldpc} number of bits of LDPC uncoded block; n_{ldpc} number of bits of LDPC coded block.

Table 4.11 Coding parameters (for normal FECFRAME n_{ldpc} = 64 800)

LDPC code	BCH uncoded block K_{bch}	BCH coded block N_{bch} LDPC uncoded block k_{ldpc}	BCH t-error correction	LDPC coded block n_{ldpc}
1/4	16 008	16 200	12	64 800
1/3	21 408	21 600	12	64 800
2/5	25 728	25 920	12	64 800
1/2	32 208	32 400	12	64 800
3/5	38 688	38 880	12	64 800
2/3	43 040	43 200	10	64 800
3/4	48 408	48 600	12	64 800
4/5	51 648	51 840	12	64 800
5/6	53 840	54 000	10	64 800
8/9	57 472	57 600	8	64 800
9/10	58 192	58 320	8	64 800

Table 4.12 Coding parameters (for short FECFRAME $n_{ldpc} = 16\,200$)

LDPC code identifier	BCH uncoded block K_{bch}	BCH-coded block N_{bch} LDPC uncoded block k_{ldpc}	BCH t-error correction	Effective LDPC rate $k_{ldpc}/16200$	LDPC coded block n_{ldpc}
1/4	3 072	3 240	12	1/5	16 200
1/3	5 232	5 400	12	1/3	16 200
2/5	6 312	6 480	12	2/5	16 200
1/2	7 032	7 200	12	4/9	16 200
3/5	9 552	9 720	12	3/5	16 200
2/3	10 632	10 800	12	2/3	16 200
3/4	11 712	11 880	12	11/15	16 200
4/5	12 432	12 600	12	7/9	16 200
5/6	13 152	13 320	12	37/45	16 200
8/9	14 232	14 400	12	8/9	16 200
9/10	NA	NA	NA	NA	NA

The generator polynomial of the t error correcting BCH encoder is obtained by multiplying the first t polynomials in Table 4.13 for $n_{ldpc} = 64\,800$ and in Table 4.14 for $n_{ldpc} = 16\,200$.

BCH encoding of information bits $m = (m_{k_{bch}-1}, m_{k_{bch}-2}, \ldots, m_1, m_0)$ onto a code word

$$c = (m_{k_{bch}-1}, m_{k_{bch}-2}, \ldots, m_1, m_0, d_{n_{bch}-k_{bch}-1}, d_{n_{bch}-k_{bch}-2}, \ldots, d_1, d_0)$$

is achieved as follows:

- Multiply the message polynomial $m(x) = m_{k_{bch}-1}x^{k_{bch}-1} + m_{k_{bch}-2}x^{k_{bch}-2} + \ldots + m_1 x + m_0$ by $x^{n_{bch}-k_{bch}}$.
- Divide $x^{n_{bch}-k_{bch}}\, m(x)$ by $g(x)$, the generator polynomial. Let $d(x) = d_{n_{bch}-k_{bch}-1}x^{n_{bch}-k_{bch}-1} + \ldots + d_1 x + d_0$ be the remainder.
- Set the code word polynomial $c(x) = x^{n_{bch}-k_{bch}}m(x) + d(x)$.

4.8.4.2 Inner encoding (LDPC)

The LDPC encoder systematically encodes an information block of size k_{ldpc}, $i = (i_0, i_1, \ldots, i_{k_{ldpc}-1})$ onto a code word of size n_{ldpc}, $c = (i_0, i_1, \ldots, i_{k_{ldpc}-1}, p_0, p_1, \ldots p_{n_{ldpc}-k_{ldpc}-1})$. The transmission of the

Table 4.13 BCH polynomials (for normal FECFRAME $n_{ldpc} = 64\,800$)

$g_1(x)$	$1 + x^2 + x^3 + x^5 + x^{16}$
$g_2(x)$	$1 + x + x^4 + x^5 + x^6 + x^8 + x^{16}$
$g_3(x)$	$1 + x^2 + x^3 + x^4 + x^5 + x^7 + x^8 + x^9 + x^{10} + x^{11} + x^{16}$
$g_4(x)$	$1 + x^2 + x^4 + x^6 + x^9 + x^{11} + x^{12} + x^{14} + x^{16}$
$g_5(x)$	$1 + x + x^2 + x^3 + x^5 + x^8 + x^9 + x^{10} + x^{11} + x^{12} + x^{16}$
$g_6(x)$	$1 + x^2 + x^4 + x^5 + x^7 + x^8 + x^9 + x^{10} + x^{12} + x^{13} + x^{14} + x^{15} + x^{16}$
$g_7(x)$	$1 + x^2 + x^5 + x^6 + x^8 + x^9 + x^{10} + x^{11} + x^{13} + x^{15} + x^{16}$
$g_8(x)$	$1 + x + x^2 + x^5 + x^6 + x^8 + x^9 + x^{12} + x^{13} + x^{14} + x^{16}$
$g_9(x)$	$1 + x^5 + x^7 + x^9 + x^{10} + x^{11} + x^{16}$
$g_{10}(x)$	$1 + x + x^2 + x^5 + x^7 + x^8 + x^{10} + x^{12} + x^{13} + x^{14} + x^{16}$
$g_{11}(x)$	$1 + x^2 + x^3 + x^5 + x^9 + x^{11} + x^{12} + x^{13} + x^{16}$
$g_{12}(x)$	$1 + x + x^5 + x^6 + x^7 + x^9 + x^{11} + x^{12} + x^{16}$

Table 4.14 BCH polynomials (for short FECFRAME $n_{ldpc} = 16\,200$)

$g_1(x)$	$1 + x + x^3 + x^5 + x^{14}$
$g_2(x)$	$1 + x^6 + x^8 + x^{11} + x^{14}$
$g_3(x)$	$1 + x + x^2 + x^6 + x^9 + x^{10} + x^{14}$
$g_4(x)$	$1 + x^4 + x^7 + x^8 + x^{10} + x^{12} + x^{14}$
$g_5(x)$	$1 + x^2 + x^4 + x^6 + x^8 + x^9 + x^{11} + x^{13} + x^{14}$
$g_6(x)$	$1 + x^3 + x^7 + x^8 + x^9 + x^{13} + x^{14}$
$g_7(x)$	$1 + x^2 + x^5 + x^6 + x^7 + x^{10} + x^{11} + x^{13} + x^{14}$
$g_8(x)$	$1 + x^5 + x^8 + x^9 + x^{10} + x^{11} + x^{14}$
$g_9(x)$	$1 + x + x^2 + x^3 + x^9 + x^{10} + x^{14}$
$g_{10}(x)$	$1 + x^3 + x^6 + x^9 + x^{11} + x^{12} + x^{14}$
$g_{11}(x)$	$1 + x^4 + x^{11} + x^{12} + x^{14}$
$g_{12}(x)$	$1 + x + x^2 + x^3 + x^5 + x^6 + x^7 + x^8 + x^{10} + x^{13} + x^{14}$

code word starts in the given order from i_0 and ends with $p_{n_{ldpc}-k_{ldpc}-1}$. LDPC code parameters are (n_{ldpc}, k_{ldpc}).

During inner coding for a normal FECFRAME, the task of the encoder is to determine $n_{ldpc} - k_{ldpc}$ parity bits $(p_0, p_1, \ldots, p_{n_{ldpc}-k_{ldpc}-1})$ for every block of k_{ldpc} information bits, $(i_0, i_1, \ldots, i_{k_{ldpc}-1})$. The procedure is as follows:

- Initialize $p_0 = p_1 = p_2 = \ldots = p_{n_{ldpc}-k_{ldpc}-1} = 0$.
- Accumulate the first information bit, i_0, at parity bit addresses. For example, for rate 2/3:

$p_0 = p_0 \oplus i_0$	$p_{2767} = p_{2767} \oplus i_0$
$p_{10491} = p_{10491} \oplus i_0$	$p_{240} = p_{240} \oplus i_0$
$p_{16043} = p_{16043} \oplus i_0$	$p_{18673} = p_{18673} \oplus i_0$
$p_{506} = p_{506} \oplus i_0$	$p_{9279} = p_{9279} \oplus i_0$
$p_{12826} = p_{12826} \oplus i_0$	$p_{10579} = p_{10579} \oplus i_0$
$p_{8065} = p_{8065} \oplus i_0$	$p_{20928} = p_{20928} \oplus i_0$
	$p_{8226} = p_{8226} \oplus i_0$

- For the next 359 information bits, i_m, $m = 1, 2, \ldots, 359$ accumulate i_m at parity bit addresses $\{x + m \bmod 360 \times q\} \bmod (n_{ldpc} - k_{ldpc})$ where x denotes the address of the parity bit accumulator corresponding to the first bit i_0, and q is a code rate dependent constant specified in Table 4.15. Continuing with the example, $q = 60$ for rate 2/3. So for example for information bit i_1, the following operations are performed:

$p_{60} = p_{60} \oplus i_1$	$p_{2827} = p_{2827} \oplus i_1$
$p_{10551} = p_{10551} \oplus i_1$	$p_{300} = p_{300} \oplus i_1$
$p_{16103} = p_{16103} \oplus i_1$	$p_{18733} = p_{18733} \oplus i_1$
$p_{566} = p_{566} \oplus i_1$	$p_{9339} = p_{9339} \oplus i_1$
$p_{12886} = p_{12886} \oplus i_1$	$p_{10639} = p_{10639} \oplus i_1$
$p_{8125} = p_{8125} \oplus i_1$	$p_{20988} = p_{20988} \oplus i_1$
$p_{8286} = p_{8286} \oplus i_1$	

Table 4.15 q values for normal frames and short frames

Code rate	q	
	Normal frames	Short frames
1/4	135	36
1/3	120	30
2/5	108	27
1/2	90	25
3/5	72	18
2/3	60	15
3/4	45	12
4/5	36	10
5/6	30	8
8/9	20	5
9/10	18	NA

- For the 361st information bit, i_{360} is the addresses of the parity bit accumulators. In a similar manner, the addresses of the parity bit accumulators for the following 359 information bits $i_m, m = 361,\ 362, \ldots, 719$ are obtained using the formula $\{x + (m \bmod 360) \times q\} \bmod (n_{ldpc} - k_{ldpc})$ where x denotes the address of the parity bit accumulator corresponding to the information bit i_{360}.
- In a similar manner, for every group of 360, new information bits are used to find the addresses of the parity bit accumulators.

After all of the information bits are exhausted, the final parity bits are obtained as follows:

- Sequentially perform the following operations starting with $i = 1$:

$$p_i = p_i \oplus p_{i-1},\ \ i = 1, 2, \ldots, n_{ldpc} - k_{ldpc} - 1$$

- Final content of p_i, $i = 0, 1, .., n_{ldpc} - k_{ldpc} - 1$ is equal to the parity bit p_i.

For inner coding for a short FECFRAME, k_{ldpc} BCH encoded bits are systematically encoded to generate n_{ldpc} bits.

4.9 NEW FEATURES OF DVB-S2X

The publication of an extension to DVB-S2 (DVB-S2X) marked the successful completion of the DVB-S2 extension [ETSI-15a]; [ETSI-15c]. DVB-S2X is an extension of DVB-S2 to provide additional technologies and features with improved performance and features for the core applications of DVB-S2, including DTH, very small aperture terminal (VSAT) and DSNG, as well as new services and applications including mobile applications. Following the success of DVB-S2, further technology developments and new application requirements have enabled significant improvement on higher spectral efficiency for the carrier to noise ratios (C/N) typical for professional applications such as contribution links and IP-trunking; and have also allowed operations

at very low C/N, down to −10 dB for mobile applications including maritime, aeronautical, trains, emergencies, disaster relief services, etc.

DVB-S2X introduced LDPC FEC as inner coding using BCH FEC as outer coding schemes. Further improved features include:

— Smaller roll-off options of 5% and 10% (in addition to 20%, 25%, and 35% in DVB-S2)
— A finer gradation and extension of the number of modulation and coding modes
— New constellation options for linear and nonlinear channels
— Additional scrambling options for critical co-channel interference situations
— Channel bonding of up to three channels
— Very low signal-to-noise ratio (SNR) operation support down to −10 dB SNR
— Super-frame option

The capability to extend the C/N range to very low values down to −10 dB with additional framing, coding, and modulation options enables new satellite services for mobile applications at sea, in the air, and on the roof of vehicles, as well as very small directive antennas for mobile and portable user terminals.

DVB-S2X will be able to allow advanced techniques that support future broadband interactive networks as well as better integration of satellites into global broadband network infrastructure including 4G/5G mobile communication networks. It will allow future smooth evolution towards UHD (4K TV) as well as 8K TV; furthermore, it will provide better support to professional and DSNG applications with high-efficiency modulation schemes and C/N values with high gain improvement.

4.10 CONCLUSION

This section concludes this chapter by giving examples from digital transmission of telephony and television broadcasting.

4.10.1 Digital transmission of telephony

Let R_b be the bit rate associated with a telephone channel. It is assumed for simplicity that transmission of n telephone channels at a rate R_b also requires transmission of signalling occupying 5% of the multiplex capacity. This multiplex is transmitted after encoding at a rate R_c, and this rate modulates the QPSK carrier that thus occupies a bandwidth $B = 36$ MHz, corresponding to the typical bandwidth of a satellite channel.

Hence:

— Bit rate of one telephone channel: R_b (bit/s)
— Multiplex capacity: R (bit/s)
— Number of telephone channels: $n = R/(1.05\, R_b)$
— Bit rate of modulating binary stream: $R_c = R/\rho$, where ρ is the code rate
— Bandwidth used: $B = R_c/\Gamma$, where Γ is the spectral efficiency of QPSK modulation ($\Gamma = 1.5\,\text{bit}\,\text{s}^{-1}\,\text{Hz}^{-1}$)

In total, the number of telephone channels is given by:

$$n = B_\rho \Gamma/(0.15\, R_b)$$

Table 4.16 Required values of C/N_0 and C/N in accordance with the capacity of a digital telephone multiplex of 64 kbit s^{-1} telephone channels with carrier occupying a bandwidth of 36 MHz

Coding ratio ρ	Number of telephone channels n	C/N_0 (dB Hz)	C/N (dB)
1	804	87.8	12.3
7/8	703	83.6	8.1
3/4	603	82.0	6.4
2/3	536	81.1	5.5
1/2	402	79.3	3.8

Furthermore, the required values of C/N_0 and C/N are given by:

$$C/N_0 = (E_b/N_0)R = (E_b/N_0) \times \rho\,\Gamma B \quad \text{(Hz)}$$
$$C/N = (C/N_0) \times 1/B$$

where the value of E_b/N_0 is obtained from Table 4.5 in accordance with the chosen coding scheme. Table 4.16 indicates the results for $R_b = 64$ kbit s^{-1}.

The results from Table 4.16 are shown in Figure 4.36, represented as TDM/QPSK curves. The figure extends the comparison to other transmission schemes: an analogue frequency modulation scheme with frequency multiplexing of analogue telephone channels (frequency division multiplex [FDM]/FM), companded frequency modulation (FDM/CFM) offering a near twofold increase in capacity, and a scheme with digital speech interpolation (DSI/TDM/QPSK), a technique that takes advantage of the telephony voice activity factor (Section 3.1.1) and is discussed in Section 8.6.2. With digital transmission, an additional factor of 2 in capacity is obtained by combining low rate encoding (LRE), 32 kbit s^{-1} instead of 64 kbit s^{-1}, and DSI [CAM-76]. This is implemented using standard digital circuit multiplication equipment (DCME), which is presented in Section 8.6.3.

4.10.2 Digital broadcasting of television

Transmissions of digital multi-programme TV services use satellites in both the FSS and the BSS bands. Satellite channel bandwidth of 27 or 36 MHz is typical.

Consider broadcasting MPEG-2-coded television using the DVB-S standard [ETSI-97]. The carrier conveys a TDM of several television programmes. QPSK modulation and concatenated coding (Sections 4.3.3 and 4.7.1.2) is based on convolutional code with code rate ρ and a (204, 188) RS code. Spectral efficiency of the QPSK modulation is 1.56 bit s^{-1} Hz^{-1}, and the corresponding transmitted bit rate, given the used bandwidth $B = 27$ MHz, is equal to:

$$R_c = 1.56 \times 27\ \text{MHz} = 42.1\ \text{M bit/s}$$

The information bit rate should take into account the respective code rate of the inner convolution code (code rate ρ) and outer RS code (code rate 188/204), i.e.

$$R_b = (\rho \times 188/204)R_c\ \text{(bit/s)}$$

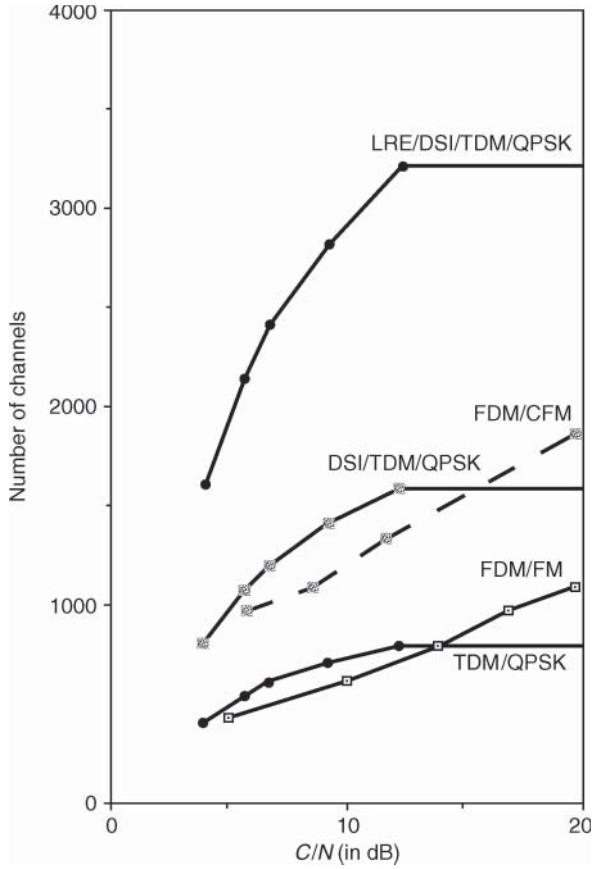


Figure 4.36 Comparison of analogue and digital transmission of a telephone multiplex by a carrier occupying a bandwidth of 36 MHz. TDM/QPSK: source coding at 64 kbit s^{-1}, digital time division multiplexing, four-state phase modulation with direct encoding, and coherent demodulation. FDM/FM: frequency division multiplexing, frequency modulation. DSI/TDM/QPSK: source coding 64 kbit s^{-1}, digital speech interpolation, digital time division multiplexing, four-state phase modulation with direct encoding and coherent demodulation. FDM/CFM: companding, frequency division multiplexing, frequency modulation. LRE/DSI/TDM/QPSK: source coding at 32 kbit s^{-1}, digital speech interpolation, digital time division multiplexing, four-state phase modulation with direct encoding and coherent demodulation.

For illustration, consider these two values of ρ:

$$\text{when } \rho = 7/8, \text{ then } R_b = 34 \text{ M bit/s}$$

$$\text{when } \rho = 1/2, \text{ then } R_b = 19.4 \text{ M bit/s}$$

The MPEG-2 format allows the broadcaster to select in a flexible way the compression ratio according to the programme content and the subjective quality of the programme as perceived by customers. Typically, the resulting programme information bit rate requires from 1.5 to 6 Mbit s^{-1}. Assuming that a television programme requires about 3.8 Mbit s^{-1}, the satellite can broadcast from five to nine television programmes, depending on the selected code rate ρ.

The MPEG-2 decoder requires QEF transmission (i.e. a BER of about 10^{-10}–10^{-11}) at the output of the outer RS decoder. This corresponds to a BER at the output of the inner (Viterbi) decoder no higher than 2×10^{-4}, as stated in [ETSI-97]. The theoretical required value for E_b/N_0 (not taking into account implementation degradation) without inner decoding would be $(E_b/N_0)_{nocod} = 7.6$ dB. The decoding gain G_{cod} provided by the inner convolutional coding scheme, typically $G_{cod} = 2$ dB for $\rho = 7/8$ and $G_{cod} = 3.9$ dB for $\rho = 1/2$, means the required E_b/N_0 value, $(E_b/N_0)_{cod}$, is:

$$(E_b/N_0)_{cod} = (E_b/N_0)_{nocod} - G_{cod} = \begin{cases} 7.6 - 2 = 5.6 \text{ dB for } \rho & = 7/8 \\ 7.6 - 3.9 = 3.7 \text{ dB for } \rho & = 1/2 \end{cases}$$

In practice, an implementation degradation (about 0.8 dB, see Table 4.9) should be taken into account. The required value for C/N_0 is:

$$C/N_0(\text{dBHz}) = (E_b/N_0)_{cod}\,(\text{dB}) + 10 \log R_b$$
$$= \begin{cases} 6.4 + 10 \log 34\,\text{M bit/s} & = 81.7 \text{ dBHz for } \rho = 7/8 \\ 4.5 + 10 \log 19.4\,\text{Mbit/}s & = 77.4 \text{ dBHz for } \rho = 1/2 \end{cases}$$

This illustrates the impact of coding with constant bandwidth and variable code rate, which translates into power reduction at the expense of capacity reduction depending on the selected code rate, as discussed in Section 4.4.

All DVB technologies use the same MPEG-2 transport streams but with different transmission technologies. This chapter has presented techniques for DVB over satellite and standards including the original standard for digital satellite broadcasting, DVB-S, and its succeeding generation, DVB-S2 and DVB-S2X.

Chapter 5 discusses how to set up the required value of the carrier power-to-noise power spectral density, C/N_0, whose required value is conditioned, as shown in this chapter, by the required baseband signal quality, E_b/N_0, and the information bit rate R_b.

REFERENCES

[BEN-96] Benedetto, S., Divsalar, D., Montorsi, G., and Pollara, F. (1996). Parallel concatenated trellis coded modulation. In: *Proceedings of the International Conference on Communications*, 974–978. IEEE.

[BER-93] Berrou, C., Glavieux, A., and Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding: turbo-codes (1). In: *Proceedings of the IEEE International Conference on Communications, Geneva, Switzerland, May*, 1064–1070. IEEE.

[BHA-81] Bhargava, V.K., Hacoun, D., Matyas, R., and Nuspl, P. (1981). *Digital Communications by Satellite*. Wiley.

[BIG-84] Biglieri, E. (1984). High-level modulation and coding for nonlinear satellite channels. *IEEE Transactions on Communications* **32**: 616–626.

[BOU-87] Bousquet, M. and Maral, G. (1987). Digital communications: satellite systems. *Systems and Control Encyclopedia*: 1050–1057.

[CAI-79] Cain, J.B., Clark, G.C. Jr., and Geist, J.M. (1979). Punctured convolutional codes of rate (n–1)/n and simplified maximum likelihood decoding. *IEEE Transactions on Information Theory* **25**: 97–100.

[CAL-89] Calderbank, A.R. (1989). Multilevel codes and multistage decoding. *IEEE Transactions on Communications* **37**: 222–229.

[CAM-76] Campanella, S.J. (1976). Digital Speech Interpolation. *COMSAT Technical Review* **6** (1): 127–157.

[ETSI-97] ETSI. (1997). Digital video broadcasting (DVB); framing structure, channel coding and modulation for 11/12 GHz satellite services. EN 300 421 (V1.1.2).
[ETSI-09] ETSI. (2009). Digital video broadcasting (DVB); interaction channel for satellite distribution systems. EN 301 790 (V1.5.1).
[ETSI-14] ETSI. (2014). Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications; part 1: DVB-S2. EN 302 307-1 (V1.4.1)
[ETSI-15a] ETSI. (2015). Digital video broadcasting (DVB); second generation frame structure, channel coding and modulation systems for broadcasting interactive services, news gathering and other broadband satellite applications; part 2: DVB-S2 extensions (DVB-S2X). EN 302 307-2 (V1.1.1).
[ETSI-15b] ETSI. (2015). Digital video broadcasting (DVB); implementation guidelines for the second generation system for broadcasting, interactive services, news gathering and other broadband satellite applications; part 1: DVB-S2. TR 102 376-1 (V1.2.1).
[ETSI-15c] ETSI. (2015). Digital video broadcasting (DVB); implementation guidelines for the second generation system for broadcasting, interactive services, news gathering and other broadband satellite applications; part 2: S2 extensions (DVB-S2X). TR 102 376-2 (V1.1.1).
[FOR-73] Forney, G.D. Jr. (1973). The Viterbi algorithm. *IEEE Proceedings* **61** (3): 268–278.
[FOR-84] Forney, G.D. Jr., Gallager, R.G., Lang, G.R. et al. (1984). Efficient modulation for band-limited channels. *IEEE Journal on Selected Areas in Communications* **2** (5): 632–647.
[GRO-76] Gronomeyer, S. and McBride, A. (1976). MSK and offset QPSK modulation. *IEEE Transactions on Communications* **24** (8): 809–820.
[IMA-77] Imai, H. and Hirakawa, S. (1977). A new multilevel coding method using error correcting codes. *IEEE Transactions on Information Theory* **23**: 371–377.
[ISA-00] Isaka, M. and Imai, H. (2000). Design and iterative decoding of multilevel modulation codes. In: *Proceedings of the 2nd International Symposium on Turbo Codes and Related Topics, Sept*, 193–196. IEEE.
[ITUR-93] ITU-R (1993) Carrier energy dispersal for systems employing angle modulation by analogue signals or digital modulation in the fixed-satellite service. Recommendation S.446.
[KAS-90] Kasami, T., Takata, T., Fujiwara, T., and Lin, S. (1990). A concatenated coded modulation scheme for error control. *IEEE Transactions on Communications* **38**: 752–763.
[MAR-95] Maral, G. (1995). *VSAT Networks*. Wiley.
[PIE-90] Pietrobon, S.S., Deng, R.H., Lafanechere, A. et al. (1990). Trellis coded multidimensional phase modulation. *IEEE Transactions on Information Theory* **36**: 63–89.
[POT-89] Pottie, G.J. and Taylor, D.P. (1989). Multilevel codes based on partitioning. *IEEE Transactions on Information Theory* **35**: 87–98.
[PRO-01] Proakis, J.G. (2001). *Digital Communications*, 4e. McGraw-Hill.
[SAY-86] Sayegh, S.I. (1986). A class of optimum block codes in signal space. *IEEE Transactions on Communications* **34**: 1043–1045.
[TOR-81] Torrier, D.J. (1981). *Principles of Military Communications Systems*. Artech House.
[UNG-82] Ungerboeck, G. (1982). Channel coding with multilevel/phase signals. *IEEE Transactions on Information Theory* **28**: 55–67.
[UNG-87] Ungerboeck, G. (1987). Trellis-coded modulation with redundant signal sets, Parts I and II. *IEEE Communications Magazine* **25**: 5–20.
[VIT-79] Viterbi, A.J. and Omura, J.K. (1979). *Principles of Digital Communication and Coding*. New York: McGraw-Hill.
[WEI-89] Wei, L.F. (1989). Rotationally invariant trellis-coded modulations with multidimensional MPSK. *IEEE Journal on Selected Areas in Communications* **7**: 1281–1295.
[WU-92] Wu, J., Costello, D.J. Jr, and Perez, L.C. (1992). On multilevel trellis M-PSK codes. Presentation at the IEEE International Symposium on Information Theory.
[ZEH-87] Zehavi, E. and Wolf, J.K. (1987). On performance evaluation of trellis codes. *IEEE Transactions on Information Theory* **33**: 196–202.

5 UPLINK, DOWNLINK, AND OVERALL LINK PERFORMANCE; INTERSATELLITE LINKS

This chapter presents the performance evaluations as shown in Figure 1.1 of Chapter 1 and Figure 4.1 of Chapter 4:

— *Uplinks* from earth stations (ESs) to satellites
— *Downlinks* from satellites to earth stations
— *Intersatellite links* (ISLs) between satellites

Uplinks and downlinks consist of radio-frequency modulated carriers, while ISLs can be either radio frequency or optical. Carriers are modulated by baseband signals conveying information for communications purposes. Connections between end users entail an uplink and a downlink, and possibly one or several ISLs.

The performance of the individual links that participate in the connection between the end terminals conditions the quality of service (QoS) for the connection between end users, specified in terms of bit error rate (BER) for digital communications. Chapter 4 has shown how the BER conditions the required value of the ratio of the energy per information bit to the power spectral density of noise (E_b/N_0) and how this impacts the value of the link performance evaluated as the ratio of the received carrier power, C, to the noise power spectral density, N_0, quoted as the ratio C/N_0, expressed in hertz (Hz). This chapter discusses the parameters that impact link performance C/N_0 and provides the means to evaluate the performance of an individual link given the transmit and receive equipment in the link or to dimension the transmit and receive equipment in order to achieve a given link performance.

We first consider individual link performance, providing the tools to evaluate the carrier-power budget and the noise-contribution budget. Then we introduce the concept of link performance for the overall link from origin to destination station, for connections supported by transparent satellites and regenerative satellites.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

Towards end of the chapter, we will explain the link performance of multibeam antenna coverage and discuss its advantages and disadvantages. With the development of the technology, it is considered a major breakthrough for development of high-throughput satellites (HTSs) today. Performance of ISLs will also be explained related to the links between geostationary earth orbit (GEO) and GEO satellites as well as GEO and low earth orbit (LEO) satellites. These will form a key technology for future mega-LEO/medium earth orbit (MEO) satellite constellations.

5.1 CONFIGURATION OF A LINK

Figure 5.1 represents the elements participating in a link. The transmit equipment consists of transmitter T_X connected by a feeder to the transmit antenna with gain G_T in the direction of the receiver. The power P_T radiated by the transmit equipment in the direction of the receiving equipment and the performance of the transmit equipment are measured by its *effective isotropic radiated power* (EIRP), which is defined as:

$$\text{EIRP} = P_T G_T \ \text{(W)} \tag{5.1}$$

On its way, the radiated power suffers from path loss L.

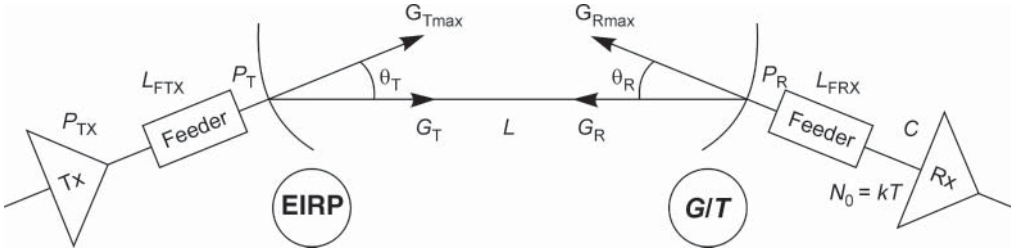


Figure 5.1 Configuration of a link.

The receiving equipment consists of the receive antenna with gain G_R in the direction of the transmit equipment, connected by a feeder to the receiver R_X. At the receiver input, the power of the modulated carrier is C and all sources of noise in the link contribute to the system noise temperature T. This system noise temperature conditions the noise power spectral density N_0, and therefore the link performance C/N_0 can be calculated at the receiver input. The receiving equipment performance is measured by its figure of merit, G/T, where G represents the overall receiving equipment gain.

The following sections present definitions of the relevant parameters that condition the link performance and provide useful equations that permit the calculation of C/N_0.

5.2 ANTENNA PARAMETERS

5.2.1 Gain

The *gain* of an antenna is the ratio of the power radiated (or received) per unit solid angle by the antenna in a given direction to the power radiated (or received) per unit solid angle by an isotropic antenna fed with the same power. The gain is maximum in the direction of maximum radiation (the electromagnetic axis of the antenna, also called the *boresight*) and has a value given by:

$$G_{\max} = (4\pi/\lambda^2)A_{\text{eff}} \tag{5.2}$$

where $\lambda = c/f$, c is the velocity of light $= 3 \times 10^8$ m s^{-1}, and f is the frequency of the electromagnetic wave. A_{eff} is the effective aperture area of the antenna. For an antenna with a circular aperture or reflector of diameter D and geometric surface $A = \pi D^2/4$, $A_{\text{eff}} = \eta A$, where η is the efficiency of the antenna. Hence:

$$G_{\max} = \eta(\pi D/\lambda)^2 = \eta(\pi D f/c)^2 \tag{5.3}$$

Expressed in dBi (the gain relative to an isotropic antenna), the actual maximum antenna gain is:

$$G_{\max\,\text{dBi}} = 10 \log \eta(\pi D/\lambda)^2 = 10 \log \eta(\pi D f/c)^2 \text{ (dBi)}$$

The efficiency η of the antenna is the product of several factors that take into account the illumination law, spill-over loss, surface impairments, ohmic and impedance mismatch losses, and so on:

$$\eta = \eta_{\text{i}} \times \eta_{\text{s}} \times \eta_{\text{f}} \times \eta_{\text{z}} \ \ldots \tag{5.4a}$$

The *illumination efficiency* η_{i} specifies the illumination law of the reflector with respect to uniform illumination. Uniform illumination ($\eta_{\text{i}} = 1$) leads to a high level of secondary lobes. A compromise is achieved by attenuating the illumination at the reflector boundaries (aperture edge taper). In the case of a Cassegrain antenna (see Section 8.3.4.3), the best compromise is obtained for an illumination attenuation at the boundaries of 10–12 dB that leads to an illumination efficiency η_{i} on the order of 91%.

The *spill-over efficiency* η_{s} is defined as the ratio of the energy radiated by the primary source that is intercepted by the reflector to the total energy radiated by the primary source. The difference constitutes the spill-over energy. The larger the angle under which the reflector is viewed from the source, the greater the spill-over efficiency. However, for a given source radiation pattern, the illumination level at the boundaries becomes less with large values of view angle, and the illumination efficiency collapses. A compromise leads to a spill-over efficiency on the order of 80%.

The *surface finish efficiency* η_{f} takes into account the effect of surface roughness on the gain of the antenna. The actual parabolic profile differs from the theoretical one. In practice, a compromise must be found between the effect on the antenna characteristics and the cost of fabrication. The effect on the on-axis gain is of the form:

$$\eta_{\text{f}} = \Delta G = \exp[-B(4\pi\varepsilon/\lambda)^2]$$

where ε is the root mean square (rms) surface error, i.e. the deviation between the actual and theoretical profiles measured perpendicularly to the concave face; and B is a factor, less than or equal to 1, whose value depends on the radius of curvature of the reflector. This factor increases as the radius of curvature decreases. For parabolic antennas of focal distance f, it varies as a function of the ratio f/D, where D is the diameter of the antenna. With $f/D = 0.7$, B is on the order of 0.9 considering ε on the order of $\lambda/30$; the surface finish efficiency η_{f} is on the order of 85%.

The other losses, including ohmic and impedance mismatch losses, are of less importance. In total, the overall efficiency η, the product of the individual efficiencies, is typically between 55% and 75%.

Figure 5.2 gives values of $G_{\max}$ in dBi as a function of diameter for different frequencies. It shows a reference case corresponding to a 1 m antenna at frequency 12 GHz. The corresponding gain is $G_{\max} = 40$ dBi. It is easy to derive other cases from this reference case: for instance, dividing the frequency by 2 ($f = 6$ GHz) reduces the gain by 6 dB, so $G_{\max} = 34$ dBi. Keeping frequency constant ($f = 12$ GHz) and increasing the size of the antenna by a factor of 2 ($D = 2$ m) increases the gain by 6 dB ($G_{\max} = 46$ dBi).

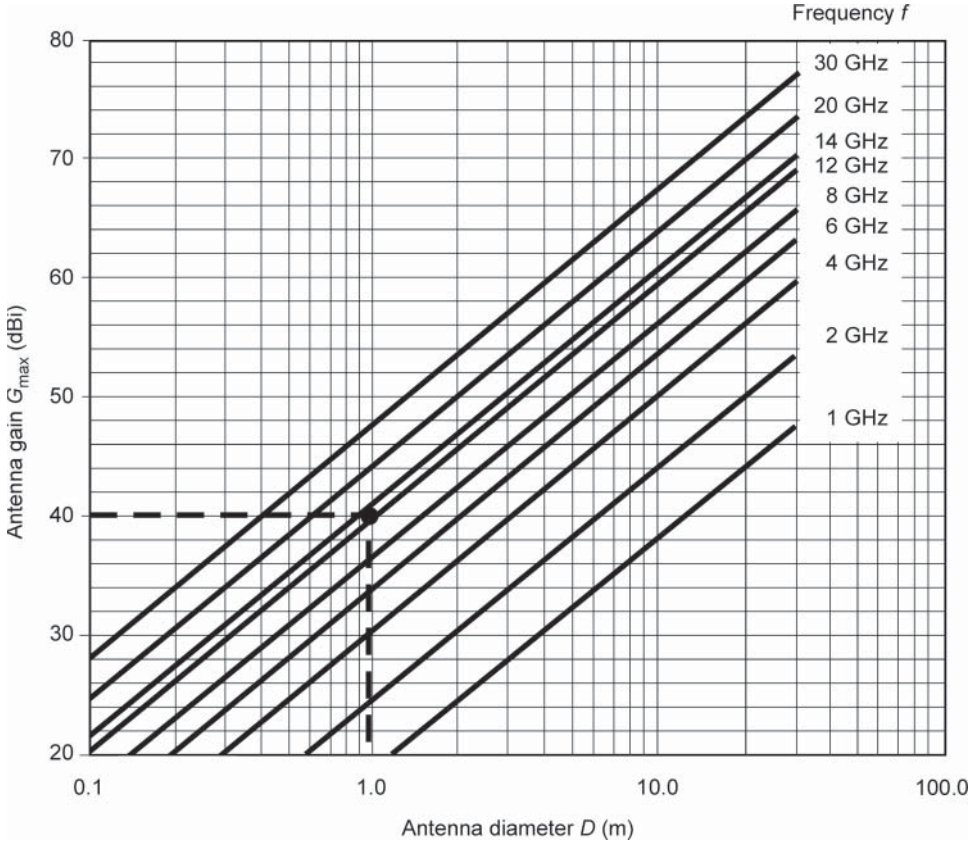


Figure 5.2 Maximum antenna gain as a function of diameter for different frequencies at $\eta = 0.6$. A 1 m antenna at 12 GHz has a gain of 40 dBi.

5.2.2 Radiation pattern and angular beamwidth

The radiation pattern indicates the variations of gain with direction. For an antenna with a circular aperture or reflector, this pattern has rotational symmetry and is completely represented within a plane in polar coordinate form (Figure 5.3a) or Cartesian coordinate form (Figure 5.3b). The main lobe contains the direction of maximum radiation. Side lobes should be kept to a minimum.

The angular beamwidth is the angle defined by the directions corresponding to a given gain fallout with respect to the maximum value. The 3 dB beamwidth, indicated in Figure 5.3a by $\theta_{3\,\mathrm{dB}}$, is often used. The 3 dB beamwidth corresponds to the angle between the directions in which the gain falls to half its maximum value. The 3 dB beamwidth is related to the ratio λ/D by a coefficient whose value depends on the chosen illumination law. For uniform illumination, the coefficient has a value of 58.5°. With non-uniform illumination laws, which lead to attenuation at the reflector boundaries, the 3 dB beamwidth increases and the value of the coefficient depends on the particular characteristics of the law. The value commonly used is 70°, which leads to the following expression:

$$\theta_{3\mathrm{dB}} = \frac{70\lambda}{D} = \frac{70c}{fD} \text{ (degrees)} \tag{5.4b}$$

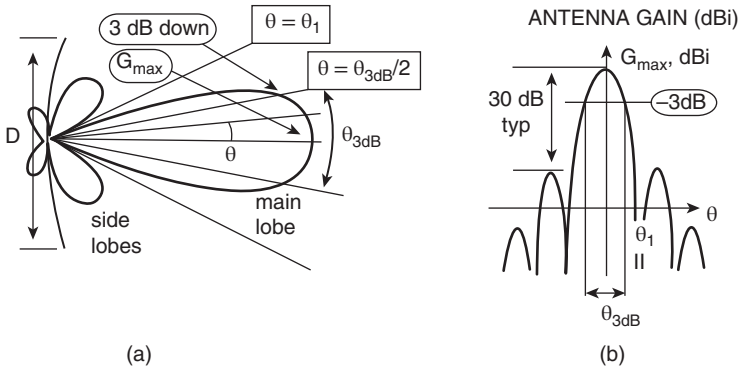


Figure 5.3 Antenna radiation pattern: (a) polar representation; (b) Cartesian representation.

In a direction θ with respect to the boresight, the value of gain is given by:

$$G(\theta)_{\text{dBi}} = G_{\text{max,dBi}} - 12(\theta/\theta_{3\text{ dB}})^2(\text{dBi}) \tag{5.5}$$

This expression is valid only for sufficiently small angles (θ between 0 and $\theta_{3\text{ dB}}/2$).

Combining Eqs. (5.3) and (5.4b), it can be seen that the maximum gain of an antenna is a function of the 3 dB beamwidth, and this relation is independent of frequency:

$$G_{\text{max}} = \eta(\pi Df/c)^2 = \eta(\pi 70/\theta_{3\text{ dB}})^2 \tag{5.6}$$

If a value of antenna efficiency $\eta = 0.6$ is considered, this gives:

$$G_{\text{max}} = 29\,000/(\theta_{3\text{ dB}})^2 \tag{5.7}$$

in which $\theta_{3\text{ dB}}$ is expressed in degrees.

Figure 5.4 shows the relationship between 3 dB beamwidth and maximum gain for three values of antenna efficiency. The gain is expressed in dBi and the 3 dB beamwidth in degrees.

$$G_{\text{max d Bi}} = 44.6 - 20\log\theta_{3\text{ dB}}\ (\text{dBi})$$

$$\theta_{3\text{ dB}} = 170/10^{\frac{G_{\text{max dBi}}}{20}}\ (\text{degrees})$$

By differentiating Eq. (5.5) with respect to θ, we obtain:

$$\frac{dG(\theta)}{d\theta} = -\frac{24\theta}{\theta^2_{3\text{ dB}}}\ (\text{dB/deg})$$

This allows us to calculate the gain fallout ΔG in dB at angle θ degrees from the boresight, for a depointing angle $\Delta\theta$ degrees about the θ direction:

$$\Delta G = -\frac{24\theta}{\theta^2_{3\text{ dB}}}\Delta\theta\ (\text{dB}) \tag{5.8}$$

The gain fall-out is maximum at the edge of 3 dB beamwidth ($\theta = \frac{1}{2}\theta_{3\text{ dB}}$ in Eq. (5.8)) and is equal to:

$$\Delta G = -\frac{12\Delta\theta}{\theta_{3\text{ dB}}}(\text{dB}) \tag{5.9}$$

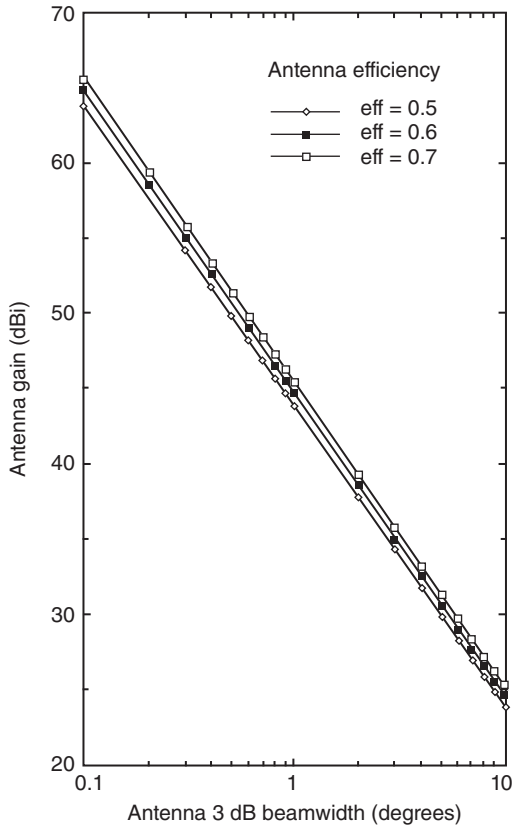


Figure 5.4 Antenna gain in the direction of maximum radiation as a function of the angular beamwidth $\theta_{3\,\mathrm{dB}}$ for three values of efficiency ($\eta = 0.5$, $\eta = 0.6$, and $\eta = 0.7$).

5.2.3 Polarisation

The wave radiated by an antenna consists of an electric field component and a magnetic field component. These two components are orthogonal and perpendicular to the direction of propagation of the wave; they vary at the frequency of the wave. By convention, the polarisation of the wave is defined by the direction of the electric field. In general, the direction of the electric field is not fixed; i.e. during one period, the projection of the extremity of the vector representing the electric field onto a plane perpendicular to the direction of propagation of the wave describes an ellipse; the polarisation is said to be elliptical (Figure 5.5).

Polarisation is characterised by the following parameters:

— *Direction of rotation* (with respect to the direction of propagation): right-hand (clockwise) or left-hand (counter-clockwise).
— *Axial ratio* (AR): AR = $E_{\max}/E_{\min}$, which is the ratio of the major and minor axes of the ellipse. When the ellipse is a circle (axial ratio = 1 = 0 dB), the polarisation is said to be circular. When the ellipse reduces to one axis (infinite axial ratio: the electric field maintains a fixed direction), the polarisation is said to be linear.
— *Inclination* τ of the ellipse.

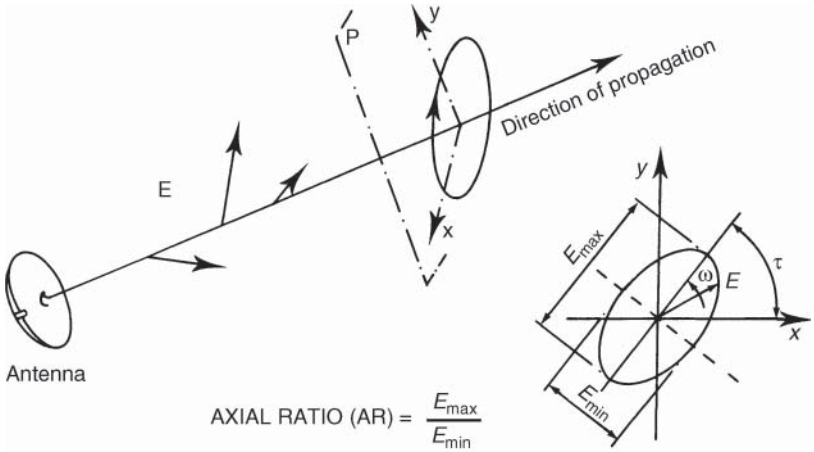


Figure 5.5 Characterisation of the polarisation of an electromagnetic wave.

Two waves are in orthogonal polarisation if their electric fields describe identical ellipses in opposite directions. In particular, the following can be obtained:

— Two orthogonal circular polarisations described as right-hand circular and left-hand circular (the direction of rotation is for an observer looking in the direction of propagation).
— Two orthogonal linear polarisations described as horizontal and vertical (relative to a local reference).

An antenna designed to transmit or receive a wave of given polarisation can neither transmit nor receive in the orthogonal polarisation. This property enables two simultaneous links to be established at the same frequency between the same two locations; this is described as frequency reuse by orthogonal polarisation. To achieve this, either two polarised antennas must be provided at each end or, preferably, one antenna that operates with the two specified polarisations may be used. This practice must, however, take into account imperfections of the antennas and the possible depolarisation of the waves by the transmission medium (Section 5.7.1.2). These effects lead to mutual interference of the two links.

This situation is illustrated in Figure 5.6, which relates to the case of two orthogonal linear polarisations (but the illustration is equally valid for any two orthogonal polarisations). Let a and b be the amplitudes, assumed to be equal, of the electric field of the two waves transmitted simultaneously with linear polarisation, a_C and b_C be the amplitudes received with the same polarisation, and a_X and b_X be the amplitudes received with orthogonal polarisations. The following are defined:

— The *cross-polarisation isolation: XPI* = a_C/b_X or b_C/a_X, hence:

$$\text{XPI (dB)} = 20\ \log(a_C/b_X) \text{ or } 20\ \log(b_C/a_X)\text{(dB)}$$

— The *cross-polarisation discrimination* (when a single polarisation is transmitted): XPD = a_C/a_X, hence:

$$\text{XPD (dB)} = 20\ \log(a_C/a_X)\text{(dB)}$$

In practice, XPI and XPD are comparable and are often included in the term *isolation*.

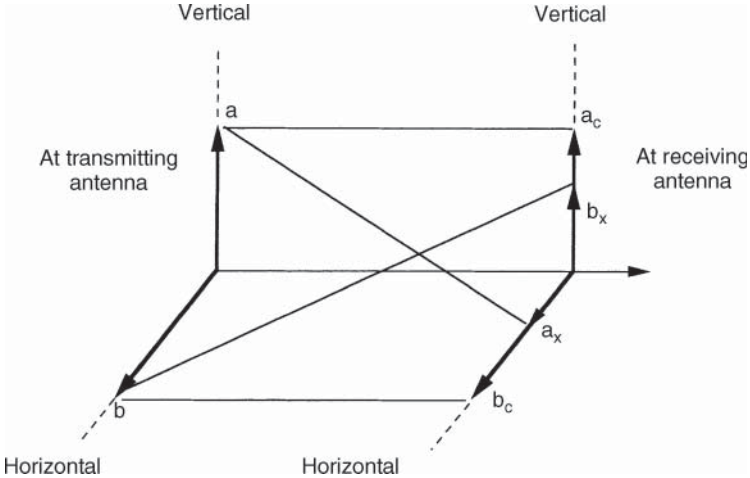


Figure 5.6 Amplitude of the transmitted and received electric field for the case of two orthogonal linear polarisations.

For a quasi-circular polarisation characterised by its value of axial ratio AR, the cross-polarisation discrimination is given by:

$$\text{XPD} = 20\ \log[(\text{AR} + 1)/(\text{AR} - 1)](\text{dB})$$

Conversely, the axial ratio AR can be expressed as a function of XPD by:

$$\text{AR} = (10^{\text{XPD}/20} + 1)/(10^{\text{XPD}/20} - 1)$$

The values and relative values of the components vary as a function of direction with respect to the antenna boresight. The antenna is thus characterised for a given polarisation by a radiation pattern for nominal polarisation (copolar) and a radiation pattern for orthogonal polarisation (cross-polar). Cross-polarisation discrimination is generally maximum on the antenna axis and degrades for directions other than that of maximum gain.

5.3 RADIATED POWER

5.3.1 Effective isotropic radiated power (EIRP)

The power radiated per unit solid angle by an isotropic antenna fed from a radio-frequency source of power P_T is given by:

$$P_T/4\pi \text{ (W/streadian)}$$

In a direction where the value of transmission gain is G_T, any antenna radiates a power per unit solid angle equal to:

$$G_T P_T/4\pi \text{ (W/streadian)}$$

The product $P_T G_T$ is called the *effective isotropic radiated power* (EIRP). It is expressed in W.

5.3.2 Power flux density

A surface of area A situated at a distance R from the transmitting antenna subtends a solid angle A/R^2 at the transmitting antenna (see Figure 5.7). It receives a power equal to:

$$P_R = (P_T G_T / 4\pi)(A/R^2) = \Phi A (\text{W}) \tag{5.10}$$

The magnitude $\Phi = P_T G_T / 4\pi R^2$ is called the *power flux density*. It is expressed in W/m^2.

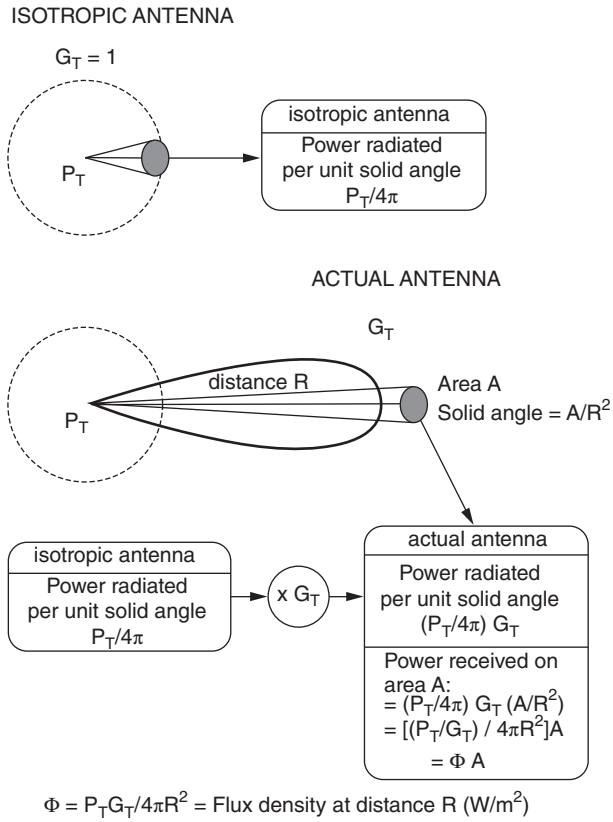


Figure 5.7 Power flux density.

5.4 RECEIVED SIGNAL POWER

5.4.1 Power captured by the receiving antenna and free space loss

As shown in Figure 5.8, a receiving antenna of effective aperture area A_{Reff} located at a distance R from the transmitting antenna receives power equal to:

$$P_R = \Phi A_{R\,\text{eff}} = (P_T G_T / 4\pi R^2) A_{R\,\text{eff}} (\text{W}) \tag{5.11}$$

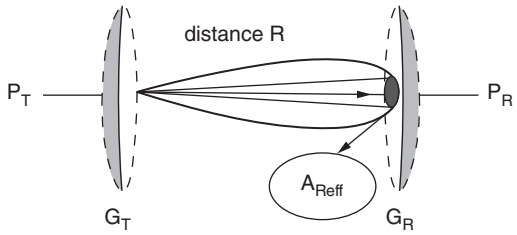


Figure 5.8 The power received by a receiving antenna.

The effective area of an antenna is expressed as a function of its receiving gain G_R according to Eq. (5.2):

$$A_{R\,\text{eff}} = G_R/(4\pi/\lambda^2)(\text{m}^2) \tag{5.12}$$

Hence an expression for the received power:

$$\begin{aligned} P_R &= (P_T G_T/4\pi R^2)(\lambda^2/4\pi)G_R \\ &= (P_T G_T)(\lambda/4\pi R)^2 G_R \\ &= (P_T G_T)(1/L_{FS})G_R\ (\text{W}) \end{aligned} \tag{5.13}$$

where $L_{FS} = (4\pi R/\lambda)^2$ is called the *free space loss* and represents the ratio of the received and transmitted powers in a link between two isotropic antennas. Figure 5.9 gives the value of L_{FS} (R_0) as a function of frequency for a geostationary satellite and a station situated at the sub-satellite point at a distance $R = R_0 = 35\,786$ km equal to the altitude of the satellite. Notice that L_{FS} is on the order of 200 dB. For any station whose position is represented by its relative latitude and longitude l and L with respect to the geostationary satellite (since the satellite is situated in the equatorial plane, l is the geographical latitude of the station), the value of $L_{FS}(R_0)$ provided by

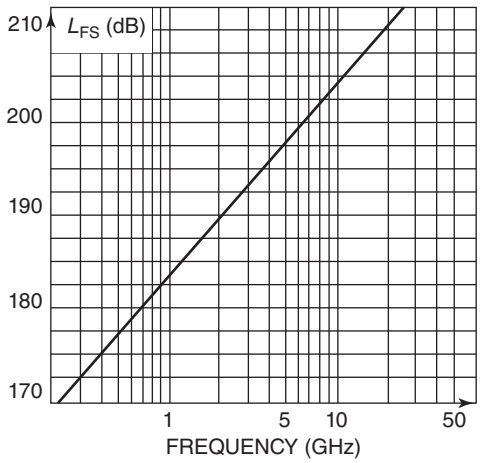


Figure 5.9 Free space loss attenuation at geostationary sub-satellite point: $L_{PS}(R_0)$.

Figure 5.9 must be corrected by the term $(R/R_0)^2$, hence:

$$L_{FS} = (4\pi R/\lambda)^2 = (4\pi R_0/\lambda)^2(R/R_0)^2 = L_{FS}\,(R_0)(R/R_0)^2$$

where $(R/R_0)^2 = 1 + 0.42(1-\cos l \cos L)$ (see Chapter 2, Eqs. (2.62) and (2.63)). The value of $(R/R_0)^2$ is between 1 and 1.356 (0–1.3 dB).

Example 5.1 Uplink received power

Consider the transmitting antenna of an earth station equipped with an antenna of diameter $D = 4$ m. This antenna is fed with a power P_T of 100 W, which is 20 dBW, at a frequency $f_U = 14$ GHz. It radiates this power towards a geostationary satellite situated at a distance of 40 000 km from the station on the axis of the antenna. The beam of the satellite receiving antenna has a width $\theta_{3\,dB} = 2°$.

It is assumed that the earth station is at the centre of the region covered by the satellite antenna and consequently benefits from the maximum gain of this antenna. The efficiency of the satellite antenna is assumed to be $\eta = 0.55$ and that of the earth station to be $\eta = 0.6$.

— The power flux density at the satellite situated at earth station antenna boresight is calculated as:

$$\Phi_{max} = P_T G_{Tmax}/4\pi R^2 (W/m^2)$$

The gain of the earth station antenna, from Eq. (5.3), is:

$$G_{Tmax} = \eta(\pi D/\lambda_U)^2 = \eta(\pi D f_U/c)^2$$
$$= 0.6(\pi \times 4 \times 14 \times 10^9/3 \times 10^8)^2 = 206\,340 = 53.1 \text{ dBi}$$

The EIRP of the earth station (on the axis) is given by:

$$(EIRP_{max})_{ES} = P_T\, G_{Tmax} = 53.1 \text{ dBi} + 20 \text{ dBW} = 73.1 \text{ dBW}$$

The power flux density is given by:

$$\Phi_{max} = P_T G_{Tmax}/4\pi R^2 = 73.1 \text{ dBW} - 10\ \log(4\pi(4 \times 10^7)^2)$$
$$= 73.1 - 163 = -89.9 \text{ dBW/m}^2$$

— The power received (in dBW) by the satellite antenna is obtained using Eq. (5.13):

$$P_R = \text{EIRP} - \text{attenuation of free space} + \text{gain of receiving antenna}$$

The attenuation of free space $L_{FS} = (4\pi R/\lambda_U)^2 = (4\pi R f_U/c)^2 = 207.4$ dB.
The gain of the satellite receiving antenna $G_R = G_{Rmax}$ is obtained using Eq. (5.3):

$$G_{Rmax} = \eta(\pi D/\lambda_U)^2$$

The value of D/λ_U is obtained using Eq. (5.7), hence $\theta_{3\,dB} = 70(\lambda_U/D)$, from which

$$D/\lambda_U = 70/\theta_{3\,dB} \text{ and } G_{Rmax} = \eta(70\pi/\theta_{3\,dB})^2 = 6650 = 38.2 \text{ dBi}.$$

Notice that the antenna gain does not depend on frequency when the beamwidth, and hence the area covered by the satellite antenna, is imposed. In total:

$$P_R = 73.1 - 207.4 + 38.2 = -96.1 \text{ dBW that is 0.25 nW or 250 pW}$$

Example 5.2 Downlink received power

Consider the transmitting antenna of a geostationary satellite fed with a power P_T of 10 W, that is, 10 dBW at a frequency $f_D = 12$ GHz, and radiating this power in a beam of width $\theta_{3\,dB}$ equal to 2°. An earth station equipped with a 4 m diameter antenna is located on the axis of the antenna at a distance of 40 000 km from the satellite. The efficiency of the satellite antenna is assumed to be $\eta = 0.55$ and that of the earth station to be $\eta = 0.6$.

— The power flux density at the earth station situated at the satellite antenna boresight is calculated as:

$$\Phi_{max} = P_T G_{Tmax}/4\pi R(\mathrm{W/m^2})$$

The gain of the satellite antenna is the same in transmission as in reception since the beamwidths are made the same (notice that this requires two separate antennas on the satellite since the diameters cannot be the same and are in the ratio $f_U/f_D = 14/12 = 1.17$). Hence:

$$(\mathrm{EIRP}_{max})_{SL} = P_T G_{Tmax} = 38.2\ \mathrm{dBi} + 10\ \mathrm{dBW} = 48.2\ \mathrm{dBW}$$

The power flux density is:

$$\Phi_{max} = P_T G_{Tmax}/4\pi R^2 = 48.2\ \mathrm{dBW} - 10\ \log(4\pi(4 \times 10^7)^2) = 48.2 - 163$$
$$= -114.8\ \mathrm{dBW/m^2}$$

— The power (in dBW) received by the antenna of the earth station is obtained using Eq. (5.13):

$$P_R = \mathrm{EIRP} - \text{attenuation of free space} + \text{gain of the receiving antenna}$$

The attenuation of free space is $L_{FS} = (4\pi R/\lambda_D)^2 = 206.1$ dB.
The gain $G_R = G_{Rmax}$ of the ground station receiving antenna is obtained using Eq. (5.3), hence:

$$G_{Rmax} = \eta(\pi D/\lambda_D)^2 = 0.6(\pi \times 4/0.025)^2 = 151\,597 = 51.8\ \mathrm{dB}$$

In total:

$$P_R = 48.2 - 206.1 + 51.8 = -106.1\ \mathrm{dBW},\ \text{that is } 25\ \mathrm{pW}$$

5.4.2 Additional losses

In practice, it is necessary to take into account additional losses due to various causes:

— Attenuation of waves as they propagate through the atmosphere
— Losses in the transmitting and receiving equipment
— Depointing losses
— Polarisation mismatch losses

5.4.2.1 Attenuation in the atmosphere

The attenuation of waves in the atmosphere, denoted by L_A, is due to the presence of gaseous components in the troposphere, water (rain, clouds, snow, and ice), and the ionosphere. A quantitative presentation of these effects is given in Section 5.7. The overall effect on the power of the received carrier can be taken into account by replacing L_{FS} in Eq. (5.13) by the path loss, L, where:

$$L = L_{FS}L_A \tag{5.14}$$

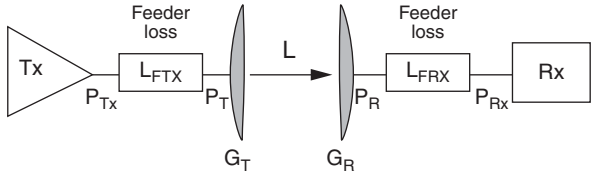


Figure 5.10 Losses in the terminal equipment.

Figure 5.10 clarifies these losses:

— *The feeder loss L_{FTX} between the transmitter and the antenna*: To feed the antenna with a power P_T, it is necessary to provide a power P_{TX} at the output of the transmission amplifier such that:

$$P_{TX} = P_T L_{FTX}(\text{W}) \tag{5.15}$$

Expressed as a function of the rated power of the transmission amplifier, the EIRP can be written:

$$\text{EIRP} = P_T G_T = (P_{TX}G_T)/L_{FTX}(\text{W}) \tag{5.16}$$

— *The feeder loss L_{FRX} between the antenna and the receiver*: The signal power P_{RX} at the input of the receiver is equal to:

$$P_{RX} = P_R/L_{FRX}(\text{W}) \tag{5.17}$$

5.4.2.2 Depointing losses

Figure 5.11 shows the geometry of the link for the case of imperfect alignment of the transmitting and receiving antennas. The result is a fallout of antenna gain with respect to the maximum gain on transmission and on reception, called *depointing loss*. These depointing losses are a function of the misalignment of angles of transmission (θ_T) and reception (θ_R) and are evaluated using Eq. (5.5). Their values are given by:

$$\begin{aligned} L_T &= 12(\theta_T/\theta_{3\text{ dB}})^2(\text{dB}) \\ L_R &= 12(\theta_R/\theta_{3\text{ dB}})^2(\text{dB}) \end{aligned} \tag{5.18}$$

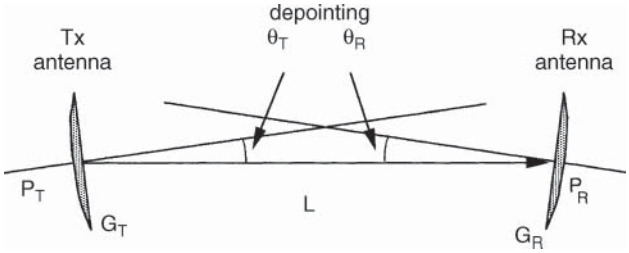


Figure 5.11 Geometry of the link.

5.4.2.3 Losses due to polarisation mismatch

It is also necessary to consider the polarisation mismatch loss L_{POL} observed when the receiving antenna is not oriented with the polarisation of the received wave. In a link with circular polarisation, the transmitted wave is circularly polarised only on the axis of the antenna and becomes elliptical off this axis. Propagation through the atmosphere can also change circular into elliptical polarisation (see Section 5.7). In a linearly polarised link, the wave can be subjected to a rotation of its plane of polarisation as it propagates through the atmosphere. Finally, with linear polarisation, the receiving antenna may not have its plane of polarisation aligned with that of the incident wave. If ψ is the angle between the two planes, the polarisation mismatch loss L_{POL} (in dB) is equal to $-20 \log \cos \psi$. In the case where a circularly polarised antenna receives a linearly polarised wave, or a linearly polarised antenna receives a circularly polarised wave, L_{POL} has a value of 3 dB. Considering all sources of loss, the signal power at the receiver input is given by:

$$P_{RX} = (P_{TX}G_{Tmax}/L_T L_{FTX})(1/L_{FS}L_A)(G_{Rmax}/L_R L_{FRX} L_{POL})(\text{W}) \tag{5.19}$$

5.4.3 Conclusion

Equations (5.13) and (5.19), which express the received power at the input to the receiver, are of the same form; they are the product of three factors:

— EIRP, which characterises the transmitting equipment:

$$\text{EIRP} = (P_{TX}G_{Tmax}/L_T L_{FTX})(\text{W})$$

This expression takes into account the losses L_{TFX} between the transmission amplifier and the antenna and the reduction in antenna gain L_T due to misalignment of the transmitting antenna.

— 1/L, which characterises the transmission medium;

$$1/L = 1/L_{FS}L_A$$

The path loss L takes into account the attenuation of free space L_{FS} and the attenuation in the atmosphere L_A.

— The gain of the receiver, which characterises the receiving equipment:

$$G = G_{Rmax}/L_R L_{FRX} L_{POL}$$

This expression takes into account the losses L_{FRX} between the antenna and the receiver, the loss of antenna gain L_R due to misalignment of the receiving antenna and the polarisation mismatch losses L_{POL}.

5.5 NOISE POWER SPECTRAL DENSITY AT THE RECEIVER INPUT

5.5.1 The origins of noise

Noise consists of all unwanted contributions whose power adds to the wanted carrier power. It reduces the ability of the receiver to reproduce correctly the information content of the received wanted carrier.

The origins of noise are as follows:

— The noise emitted by natural sources of radiation located within the antenna reception area
— The noise generated by components in the receiving equipment

Carriers from transmitters other than those that it is wished to receive are also classed as noise. This noise is described as *interference*.

5.5.2 Noise characterisation

Harmful noise power is that which occurs in the bandwidth B of the wanted modulated carrier. A popular noise model is that of white noise, for which the power spectral density N_0 (W/Hz) is constant in the frequency band involved (Figure 5.12). The equivalent noise power N(W) captured by a receiver with equivalent noise bandwidth B_N, usually matched to B ($B = B_N$), is given by:

$$N = N_0 B_N(\mathrm{W}) \tag{5.20}$$

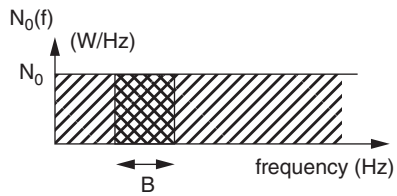


Figure 5.12 Spectral density of white noise.

Real noise sources do not always have a constant power spectral density, but the model is convenient for representation of actual noise observed over a limited bandwidth.

5.5.2.1 Noise temperature of a noise source

The noise temperature of a two-port noise source delivering an available noise power spectral density N_0 is given by:

$$T = N_0/k(\mathrm{K}) \tag{5.21}$$

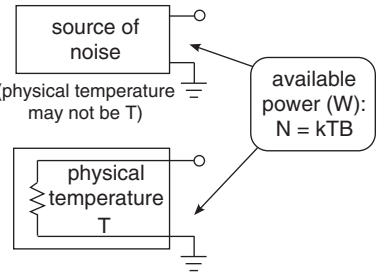


Figure 5.13 Definition of the noise temperature of a noise source.

where k is Boltzmann's constant $= 1.379 \times 10^{-23} = -228.6\,\text{dBWHz}^{-1}\,\text{K}$, and T represents the thermodynamic temperature of a resistance that delivers the same available noise power as the source under consideration (Figure 5.13). Available noise power is the power delivered by the source to a device that is impedance matched to the source.

5.5.2.2 Effective input noise temperature

The effective input noise temperature T_e of a four-port element is the thermodynamic temperature of a resistance which, placed at the input of the element assumed to be noise-free, establishes the same available noise power at the output of the element as the actual element without the noise source at the input (Figure 5.14). T_e is thus a measure of the noise generated by the internal components of the four-port element.

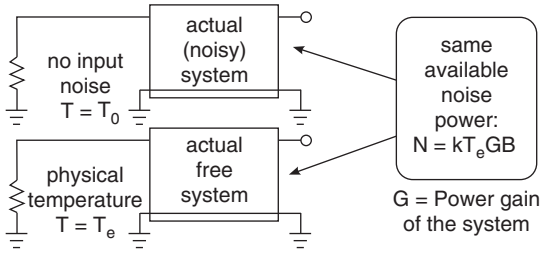


Figure 5.14 Effective input noise temperature of a four-port element.

The noise figure of this four-port element is the ratio of the total available noise power at the output of the element to the component of this power engendered by a source at the input of the element with a noise temperature equal to the reference temperature $T_0 = 290\,\text{K}$.

Assume that the element has a power gain G and a bandwidth B and is driven by a source of noise temperature T_0; the total power at the output is $Gk(T_e + T_0)B$. The component of this power originating from the source is GkT_0B. The noise figure is thus:

$$F = [Gk(T_e + T_0)B]/[GkT_0B] = (T_e + T_0)/T_0 = 1 + T/T_0 \tag{5.22}$$

The noise figure is usually quoted in decibels (dB), according to:

$$F(\text{dB}) = 10\ \log F$$

Figure 5.15 displays the relationship between noise temperature and noise figure (dB).

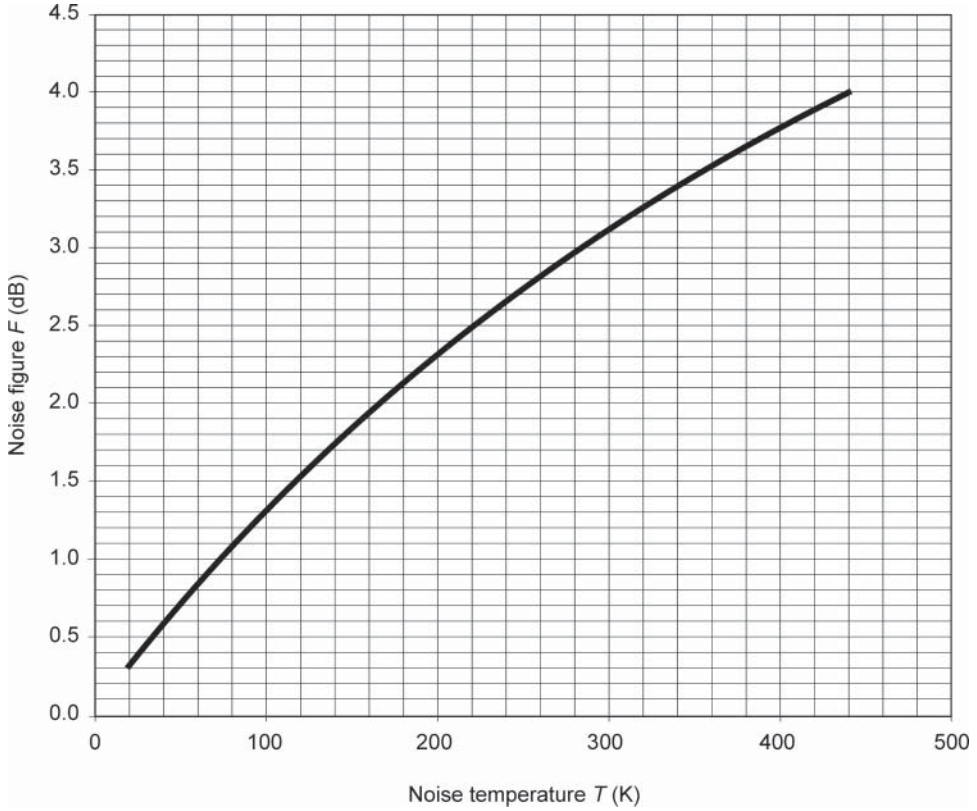


Figure 5.15 Noise figure versus noise temperature: $F(\text{dB}) = 10\log(1 + T/T_0)$ with $T_0 = 290\,\text{K}$.

5.5.2.3 *Effective input noise temperature of an attenuator*

An *attenuator* is a four-port element containing only passive components (which can be classed as resistances) all at temperature T_{ATT}, which is generally the ambient temperature. If L_{ATT} is the attenuation caused by the attenuator, the effective input noise temperature of the attenuator is:

$$T_{\text{eATT}} = (L_{\text{ATT}} - 1)T_{\text{ATT}}(\text{K}) \tag{5.23}$$

If $T_{\text{ATT}} = T_0$, the noise figure of the attenuator from a comparison of Eqs. (5.22) and (5.23), is:

$$F_{\text{ATT}} = L_{\text{ATT}}$$

5.5.2.4 *Effective input noise temperature of cascaded elements*

Consider a chain of N four-port elements in cascade, each element j having a power gain $G_j (j = 1,2, \ldots, \text{N})$ and an effective input noise temperature T_{ej}.

The overall effective input noise temperature is:

$$T_{\text{e}} = T_{\text{e1}} + T_{\text{e2}}/G_1 + T_{\text{e3}}/G_1G_2 + \ldots + T_{\text{iN}}/G_1G_2, \ldots, G_{\text{N}-1}(\text{K}) \tag{5.24}$$

The noise figure is obtained from Eq. (5.22):

$$F = F_1 + (F_2 - 1)/G_1 + (F_3 - 1)/G_1G_2 + \cdots + (F_N - 1)/G_1G_2, \ldots, G_{N-1} \tag{5.25}$$

5.5.2.5 *Effective input noise temperature of a receiver*

Figure 5.16 shows the arrangement of the receiver. By using Eq. (5.24), the effective input noise temperature T_{eRX} of the receiver can be expressed as:

$$T_{eRX} = T_{LAN} + T_{MX}/G_{LNA} + T_{IF}/G_{LNA}G_{MX}(\text{K}) \tag{5.26}$$

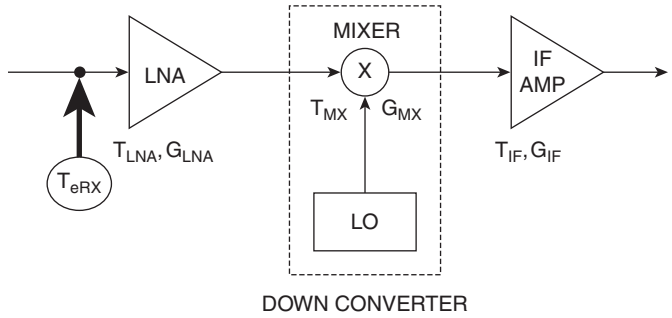


Figure 5.16 The organisation of a receiver.

Example 5.3 Low noise amplifier (LNA): $T_{LNA} = 150\,\text{K}$, $G_{LNA} = 50\,\text{dB}$
Mixer: $T_{MX} = 850\,\text{K}$, $G_{MX} = -10\,\text{dB}$ ($L_{MX} - 10\,\text{dB}$)
IF amplifier: $T_{IF} = 400\,\text{K}$, $G_{IF} = 30\,\text{dB}$
Hence:

$$\begin{aligned} T_{eRX} &= 150 + 850/10^5 + 400/10^5 10^{-1} \\ &= 150\text{K} \end{aligned}$$

Notice the benefit of the high gain of the LNA, which limits the noise temperature T_{eRX} of the receiver to that of the LNA T_{LNA}.

5.5.3 Noise temperature of an antenna

An antenna picks up noise from radiating bodies within the radiation pattern of the antenna. The noise output from the antenna is a function of the direction in which it is pointing, its radiation pattern, and the state of the surrounding environment. The antenna is assumed to be a noise source characterised by a noise temperature called the *noise temperature of the antenna* T_A (K).

Let $T_b(\theta, \varphi)$ be the brightness temperature of a radiating body located in a direction (θ, φ), where the gain of the antenna has a value $G(\theta, \varphi)$. The noise temperature of the antenna is obtained by integrating the contributions of all the radiating bodies within the radiation pattern of the antenna. The noise temperature of the antenna is thus:

$$T_A = (1/4\pi) \iint T_b(\theta, \varphi) G(\theta, \varphi) \sin\theta \, \text{d}\theta \text{d}\varphi(\text{K}) \tag{5.27}$$

There are two cases to be considered:

— A satellite antenna (the uplink)
— An earth station antenna (the downlink)

5.5.3.1 Noise temperature of a satellite antenna (uplink)

The noise captured by the antenna is noise from the earth and from outer space. The beamwidth of a satellite antenna is equal to or less than the angle of view of the earth from the satellite, which is 17.5° for a geostationary satellite. Under these conditions, the major contribution is that from the earth. For a beamwidth $\theta_{3\,\mathrm{dB}}$ of 17.5°, the antenna noise temperature depends on the frequency and the orbital position of the satellite (see Figure 5.17). For a smaller width (a spot beam), the temperature depends on the frequency and the area covered; the continents radiate more noise than the oceans, as shown in Figure 5.18. For a preliminary design, the value 290 K can be taken as a conservative value.

5.5.3.2 Noise temperature of an earth station antenna (the downlink)

The noise captured by the antenna consists of noise from the sky and noise due to radiation from the earth. Figure 5.19 shows the situation.

5.5.3.2.1 'Clear sky' conditions

At frequencies greater than 2 GHz, the greatest contribution is that of the non-ionised region of the atmosphere, which, being an absorbent medium, is a noise source. In the absence of meteorological formations (conditions described as *clear sky*), the antenna noise temperature contains contributions due to the sky and the surrounding ground (Figure 5.19a).

The sky noise contribution is determined from Eq. (5.27), where $T_b(\theta, \varphi)$ is the brightness temperature of the sky in the direction (θ, φ). In practice, only that part of the sky in the direction of the antenna boresight contributes to the integral as the gain has a high value only in that direction. As a consequence, the noise contribution of the clear sky T_{SKY} can be assimilated with the brightness temperature for the angle of elevation of the antenna. Figure 5.20 shows the clear sky brightness temperature as a function of frequency and elevation angle.

Radiation from the ground in the vicinity of the earth station is captured by the side lobes of the antenna radiation pattern and partly by the main lobe when the elevation angle is small. The contribution of each lobe is determined by $T_i = G_i(\Omega_i/4\pi)T_G$, where G_i is the mean gain of the lobe of solid angle Ω_i and T_G is the brightness temperature of the ground. The sum of these contributions yields the value T_{GROUND}. The following can be taken as a first approximation [CCIR-90b]:

$-T_G = 290$ K for lateral lobes whose elevation angle E is less than $-10°$

$-T_G = 150$ K for $-10° < E < 0°$

$-T_G = 50$ K for $0° < E < 10°$

$-T_G = 10$ K for $10° < E < 90°$

The antenna noise temperature is thus given by:

$$T_A = T_{SKY} + T_{GROUND}(\mathrm{K}) \tag{5.28}$$

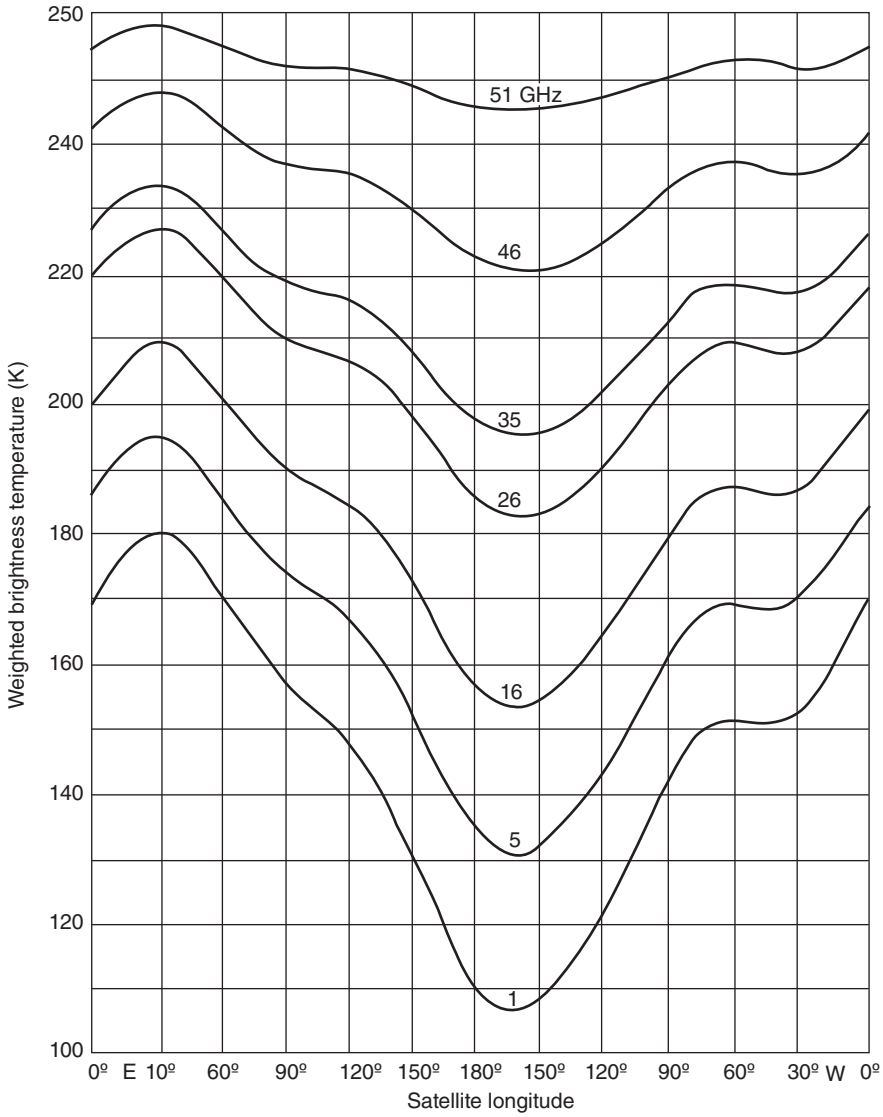


Figure 5.17 Satellite antenna noise temperature for global coverage as a function of frequency and orbital position. Source: reproduced with the permission of the American Geophysical Union.

To this noise may be added that of individual sources located in the vicinity of the antenna boresight. For a radio source of apparent angular diameter α and noise temperature T_n at the frequency considered and measured at ground level after attenuation by the atmosphere, the additional noise temperature ΔT_A for an antenna of beamwidth $\theta_{3\,dB}$ is given by:

$$\Delta T_A = T_n(\alpha/\theta_{3\,dB})^2 \quad \text{if } \theta_{3\,dB} > \alpha \text{ (K)}$$
$$\Delta T_A = T_n \qquad\qquad \text{if } \theta_{3\,dB} < \alpha \text{ (K)} \tag{5.29}$$

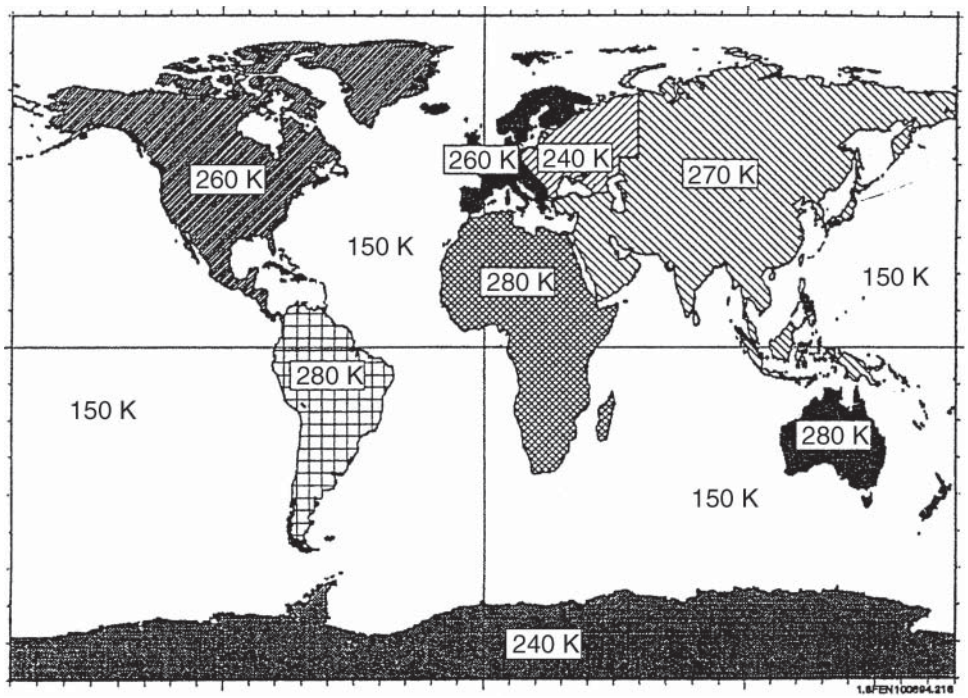


Figure 5.18 The European Space Agency (ESA)/European Telecommunications Satellite Organisation (EUTELSAT) model of the earth's brightness temperature at Ku band. Source: From [FEN-95]. Reproduced with permission.

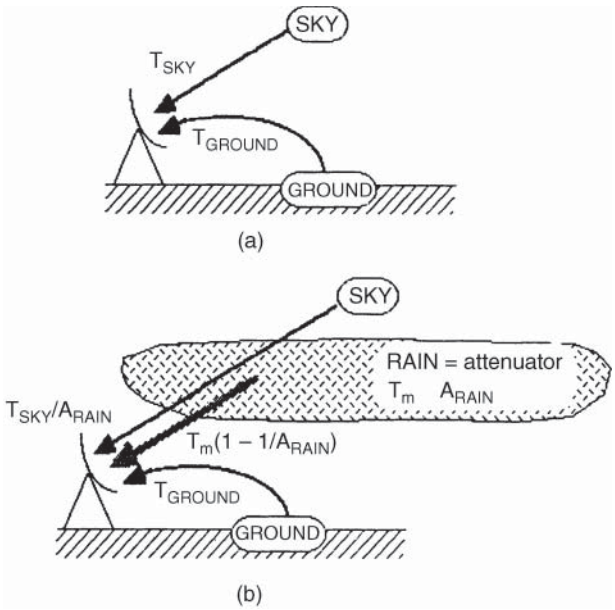


Figure 5.19 Contributions to the noise temperature of an earth station: (a) clear sky conditions; (b) conditions of rain.

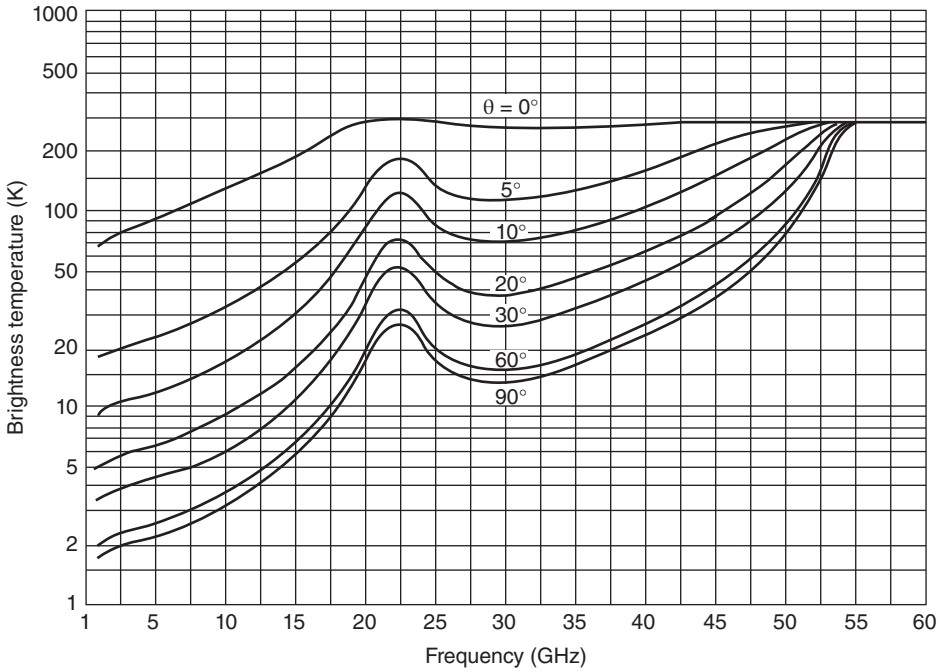


Figure 5.20 Brightness temperature of a clear sky as a function of frequency and elevation angle E for mean atmospheric humidity (7.5 g cm^{-3} humidity at ground level) and standard temperature and pressure conditions at ground level. Source: From CCIR Rep 720–2. Reproduced with the permission of the ITU [ITUR-16b].

For earth stations pointing towards a geostationary satellite, only the sun and the moon need to be considered. The sun and the moon have an apparent angular diameter of 0.5°. There is an increase of noise temperature when these heavenly bodies are aligned with the earth station pointing towards the satellite. This particular geometrical configuration can be predicted. To be more specific, at 12 GHz, a 13 m antenna undergoes a noise temperature increase due to the sun, at a time of quiet sun, by an amount ΔT_A equal to 12 000 K. The conditions of occurrence and the value of ΔT_A as a function of the antenna diameter and frequency are discussed in detail in Chapters 2 and 8. For the moon, the increase is at most 250 K at 4 GHz [CCIR-90b].

Figure 5.21 shows the variation of antenna noise temperature T_A as a function of elevation angle E for various types of antenna at different frequencies in a clear sky [CCIR-90a; ITU-02]. It can be seen that the antenna noise temperature decreases as the elevation angle increases.

5.5.3.2.2 Conditions of rain

The antenna noise temperature increases during the presence of meteorological formations, such as clouds and rain (Figure 5.19b), which constitute an absorbent, and consequently emissive, medium. Using Eq. (5.23), the antenna noise temperature becomes:

$$T_A = T_{SKY}/A_{RAIN} + T_m(1 - 1/A_{RAIN}) + T_{GROUND}(\text{K}) \tag{5.30}$$

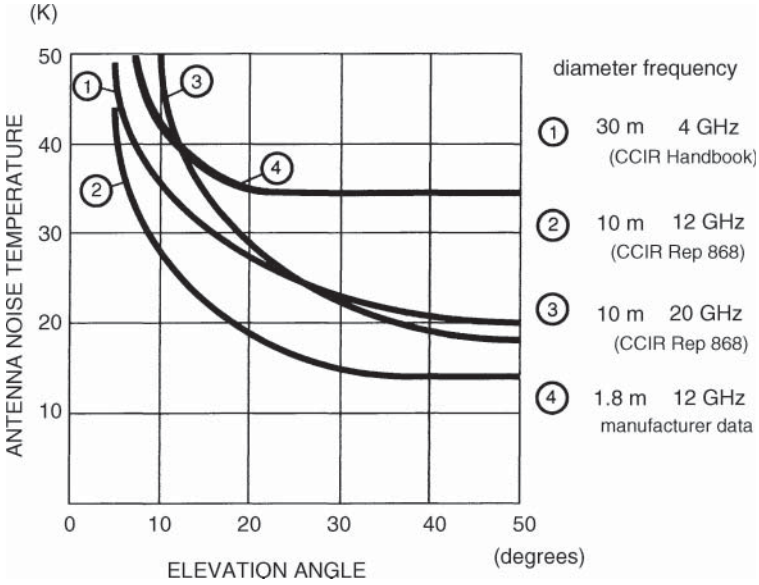


Figure 5.21 Typical values of antenna noise temperature T_A as a function of elevation angle E. Curve 1: diameter = 30 m, frequency = 4 GHz. Source: From [ITU-85]. Reproduced with the permission of the ITU. Curve 2: diameter = 10 m, frequency = 12 GHz. Curve 3: diameter = 10 m, frequency = 20 GHz. Curve 4: diameter = 1.8 m, frequency = 12 GHz. Source: reproduced with the permission of Alcatel Telspace.

where A_{RAIN} is the attenuation and T_m is the mean thermodynamic temperature of the formations in question. For T_m, a value of 275 K can be assumed ([THO-83; CCIR-82a].

5.5.3.2.3 Conclusion

In conclusion, the antenna noise temperature T_A, is a function of:

— Frequency
— Elevation angle
— Atmospheric conditions (clear sky or rain)

Consequently, the figure of merit of an earth station must be specified for particular conditions of frequency, elevation angle, and atmospheric conditions.

5.5.4 System noise temperature

Consider the receiving equipment shown in Figure 5.22. This consists of an antenna connected to a receiver. The connection (feeder) is a lossy one and is at a thermodynamic temperature T_F (which is close to T_0 = 290 K). It introduces an attenuation L_{FRX}, which corresponds to a gain $G_{FRX} = 1/L_{FRX}$ and is less than 1 ($L_{FRX} \geq 1$). The effective input noise temperature T_e of the receiver is T_{eRX}.

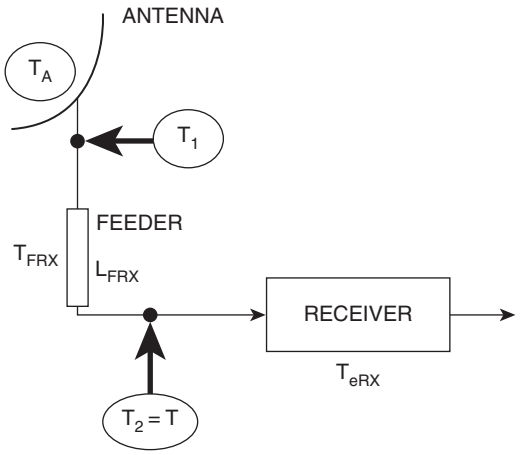


Figure 5.22 A receiving system: T is the system noise temperature at the receiver input.

The noise temperature may be determined at two points as follows:

— At the antenna output, before the feeder losses, temperature T_1
— At the receiver input, after the losses, temperature T_2

The noise temperature T_1 at the antenna output is the sum of the noise temperature of the antenna T_A and the noise temperature of the subsystem consisting of the feeder and the receiver in cascade. The noise temperature of the feeder is given by Eq. (5.23). From Eq. (5.24), the noise temperature of the subsystem is $(L_{FRX} - 1)T_F + T_{eRX}/G_{FRX}$. Adding the contribution of the antenna, considered as a noise source, this becomes:

$$T_1 = T_A + (L_{FRX} - 1)T_F + T_{eRX}/G_{FRX}(\text{K}) \tag{5.31}$$

Now consider the receiver input. This noise must be attenuated by a factor L_{FRX}. Replacing G_{FRX} with $1/L_{FRX}$, one obtains the noise temperature T_2 at the receiver input:

$$T_2 = T_1/L_{FRX} = T_A/L_{FRX} + T_F(1 - 1/L_{FRX}) + T_{eRX}(\text{K}) \tag{5.32}$$

This noise temperature T_2, which takes into account the noise generated by the antenna and the feeder together with the receiver noise, is called the *system noise temperature T* at the receiver input. Notice that measurement of noise at the considered point would reflect only the noise contribution upstream of this point. Actually, the system noise temperature takes into account all sources of noise within the receiving equipment.

Example 5.4 Consider the receiving system of Figure 5.22 with the following values:

— Antenna noise temperature: $T_A = 50\,\text{K}$
— Thermodynamic temperature of the feeder: $T_F = 290\,\text{K}$
— Effective input noise temperature of the receiver: $T_{eRX} = 50\,\text{K}$

The system noise temperature at the receiver input will be calculated for two cases: (i) no feeder loss between the antenna and the receiver and (ii) feeder loss $L_{FRX} = 1\,\text{dB}$. Using Eq. (5.31), $T = T_A/L_{FRX} + T_F(1–1/L_{FRX}) + T_{eRx}$:

For case (i), $T = 50 + 50 = 100\,\text{K}$
For case (ii), $T = 50/10^{0.1} + 290(1-1/10^{0.1}) + 50 = 39.7 + 59.6 + 50 = 149.3\,\text{K}$ or around 150 K

Notice the influence of the feeder loss; it reduces the antenna noise but makes its own contribution to the noise, and this finally causes an increase in the system noise temperature. The contribution of an attenuation to the noise can quickly be estimated using the following rule: every attenuation of 0.1 dB upstream of the receiver makes a contribution to the system noise temperature at the receiver input of $290(1-1/10^{0.01}) = 6.6\,\text{K}$ or around 7 K. To realise a receiving system with a low noise temperature, it is imperative to avoid losses upstream of the receiver.

5.5.5 Conclusion

At the receiver input, all sources of noise in the link contribute to the system noise temperature T. Those sources include the noise captured by the antenna and generated by the feeder, which can actually be measured at the receiver input, plus the noise generated downstream in the receiver, which is modelled as a fictitious source of noise at the receiver input, treating the receiver as noiseless.

The noise superimposed on the received carrier power has a power spectral density given by:

$$N_0 = kT(\text{W/Hz}) \tag{5.33}$$

where k is the Boltzmann constant ($k = 1.379 \times 10^{-23}\,\text{J/K} = -228.6\,\text{dBJ K}^{-1}$).

5.6 INDIVIDUAL LINK PERFORMANCE

Link performance is evaluated as the ratio of the received carrier power, C, to the noise power spectral density, N_0, and is quoted as the C/N_0 ratio, expressed in hertz. One can evaluate the link performance using other ratios besides C/N_0; for instance:

— C/T represents the carrier power over the system noise temperature; expressed in units of watts per Kelvin (W/K), it is given by C/T = (C/N0)k, where k is the Boltzmann constant.
— C/N represents the carrier power over the noise power; dimensionless, it is given by $C/N = (C/N_0)(1/B_\text{N})$, where B_N is the receiver noise bandwidth.

5.6.1 Carrier power to noise power spectral density ratio at receiver input

The power received at the receiver input, as given by Eq. (5.19), is that of the carrier. Hence

$$C = P_\text{RX}$$

The noise power spectral density at the same point is $N_0 = kT$, where T is given by Eq. (5.32). Hence:

$$\begin{aligned} C/N_0 = {} & [(P_\text{TX}G_\text{Tmax}/L_\text{T}L_\text{FTX})(1/L_\text{FS}L_\text{A})(G_\text{Rmax}/L_\text{R}L_\text{FRX}L_\text{POL})] \\ & /[T_\text{A}/L_\text{FRX} + T_\text{F}(1 - 1/L_\text{FRX}) + T_\text{eRX}](1/k) \quad (\text{Hz}) \end{aligned} \tag{5.34}$$

This expression can be interpreted as follows:

$$C/N_0 = (\text{transmitter EIRP})(1/\text{path loss}) \times (\text{composite receiving gain/noise temperature}) \times (1/k) \quad (\text{Hz}) \tag{5.35}$$

C/N_0 can also be expressed as a function of the power flux density Φ:

$$C/N_0 = \Phi(\lambda^2/4\pi)(\text{composite receiving gain/noise temperature})\ (1/k)(\text{Hz}) \tag{5.36}$$

where $\Phi = (\text{transmitter EIRP})/(4\pi R^2)$ (W/m^2)

Finally, it can be verified that evaluation of C/N_0 is independent of the point chosen in the receiving chain as long as the carrier power and the noise power spectral density are calculated at the same point.

Eq. (5.35) for C/N_0 introduces three factors:

— EIRP, which characterises the transmitting equipment.
— 1/L, which characterises the transmission medium.
— The composite receiving gain/noise temperature, which characterises the receiving equipment. It is called the *figure of merit*, or *G/T*, of the receiving equipment.

By examining Eq. (5.34), it can be seen that the figure of merit G/T of the receiving equipment is a function of the antenna noise temperature T_A and the effective input noise temperature T_{eRX} of the receiver.

In conclusion, Eq. (5.34) boils down to:

$$C/N_0 = (\text{EIRP})(1/L)(G/T)(1/k)\ (\text{Hz}) \tag{5.37}$$

5.6.2 Clear sky uplink performance

Figure 5.23 shows the geometry of the uplink. It is assumed that the transmitting earth station is on the edge of the 3 dB coverage of the satellite receiving antenna.

The data are as follows:

— Frequency: $f_U = 14\,\text{GHz}$

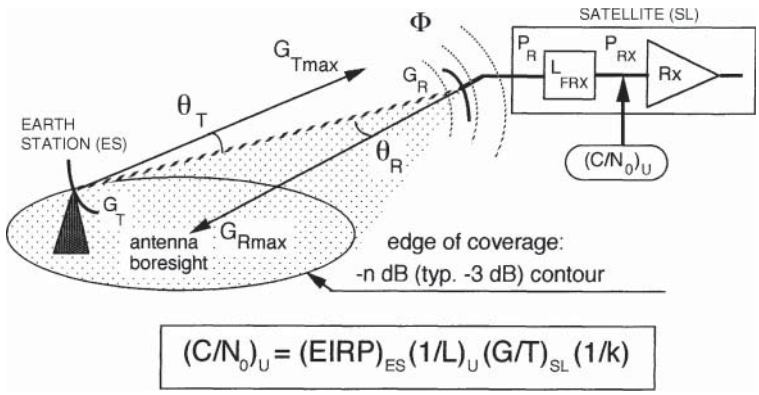


Figure 5.23 The geometry of an uplink.

— For the earth station (ES):
  Transmitting amplifier power: $P_{TX} = 100\,W$
  Loss between amplifier and antenna: $L_{FTX} = 0.5\,dB$
  Antenna diameter: $D = 4\,m$
  Antenna efficiency: $\eta = 0.6$
  Maximum pointing error: $\theta_T = 0.1°$
— Earth station-satellite distance: R = 40 000 km
— Atmospheric attenuation: $L_A = 0.3\,dB$ (typical value for attenuation by atmospheric gases at this frequency for an elevation angle of 10°)
— For the satellite (SL):
  Receiving beam half power angular width: $\theta_{3\,dB} = 2°$
  Antenna efficiency: $\eta = 0.55$
  Receiver noise figure: $F = 3\,dB$
  Loss between antenna and receiver: $L_{FRX} = 1\,dB$
  Thermodynamic temperature of the connection: $T_F = 290\,K$
  Antenna noise temperature: $T_A = 290\,K$

To calculate the EIRP of the earth station:

$$(EIRP)_{ES} = (P_{TX}G_{Tmax}/L_TL_{FTX}) \quad (W) \tag{5.38}$$

with:

$$P_{TX} = 100\,W = 20\,dBW$$
$$G_{Tmax} = \eta(\pi D/\lambda_U)^2 = \eta(\pi Df_U/c)^2 = 0.6\ [\pi \times 4 \times (14 \times 10^9)/(3 \times 10^8)]^2 = 206\,340$$
$$= 53.1\,dBi$$
$$L_T(dB) = 12(\theta_T/\theta_{3\,dB})^2 = 12(\theta_T Df_U/70c)^2 = 0.9\,dB$$
$$L_{FTX} = 0.5\,dB$$

Hence:

$$(EIRP)_{ES} = 20\,dBW + 53.1\,dB - 0.9\,dB - 0.5\,dB = 71.7\,dBW$$

To calculate the attenuation on the upward path (U):

$$L_U = L_{FS}L_A \tag{5.39}$$

with:

$$L_{FS} = (4\pi R/\lambda_U)^2 = (4\pi Rf_U/c)^2 = 5.5 \times 10^{20} = 207.4\,dB$$
$$L_A = 0.3\,dB$$

Hence:

$$L_U = 207.4\,dB + 0.3\,dB = 207.7\,dB$$

To calculate the figure of merit G/T of the satellite (SL):

$$(G/T)_{SL} = (G_{Rmax}/L_RL_{FRX}L_{POL})/[T_A/L_{FRX} + T_F(1 - 1/L_{FRX}) + T_{eRX}] \quad (K^{-1}) \tag{5.40}$$

with:

$$G_{R\,max} = \eta(\pi D/\lambda_U)^2 = \eta(\pi.70/\theta_{3\,dB})^2 = 0.55(\pi.70/2)^2 = 6650 = 38.2\,\text{dBi}$$
$$L_R = 12(\theta_R/\theta_{3\,dB})^2$$

As the earth station is on the edge of the 3 dB coverage area, $\theta_R = \theta_{3\,dB}/2$ and $L_R = 3$ dB.

$$\begin{aligned}\text{Assume } L_{POL} &= 0\,\text{dB}\\ L_{FRX} &= 1\,\text{dB}\\ \text{Given } T_A &= 290\,\text{K}\\ T_F &= 290\,\text{K}\\ T_{eRX} &= (F-1)T_0 = (10^{0.3}-1)290 = 290\,\text{K}\end{aligned}$$

Hence:

$$\begin{aligned}(G/T)_{SL} &= 38.2 - 3 - 1 - 10\,\log[290/10^{0.1} + 290(1 - 1/10^{0.1}) + 290]\\ &= 6.6\,\text{dBK}^{-1}\end{aligned}$$

Notice that when the thermodynamic temperature of the feeder between the antenna and the satellite receiver is close to the antenna noise temperature, which is the case in practice, the uplink system noise temperature at the receiver input is $T_U \approx T_F + T_{eRX} \approx 290 + T_{eRX}$. It is, therefore, needlessly costly to install a receiver with a very low noise figure on board a satellite.

To calculate the ratio C/N_0 for the uplink:

$$(C/N_0)_U = (\text{EIRP})_{ES}\,(1/L_U)(G/T)_{SL}\,(1/k) \quad \text{(Hz)} \tag{5.41}$$

Hence:

$$(C/N_0)_U = 71.7\,\text{dBW} - 207.7\,\text{dB} + 6.6\,\text{dBK}^{-1} + 228.6\,\text{dBW/HzK} = 99.2\,\text{dBHz}$$

Figure 5.24 summarises the variations in power level throughout the path.

5.6.3 Clear sky downlink performance

Figure 5.25 shows the geometry of the downlink. It is assumed that the receiving earth station is located on the edge of the 3 dB coverage area of the satellite receiving antenna. The data are as follows:

— Frequency: $f_D = 12$ GHz
— For the satellite (SL):
  Transmitting amplifier power: $P_{TX} = 10$ W
  Loss between amplifier and antenna: $L_{FTX} = 1$ dB
  Transmitting beam half power angular width: $\theta_{3\,dB} = 2°$
  Antenna efficiency: $\eta = 0.55$
— Earth station-satellite distance: R = 40 000 km
— Atmospheric attenuation: $L_A = 0.3$ dB (typical attenuation by atmospheric gases at this frequency for an elevation angle of 10°)

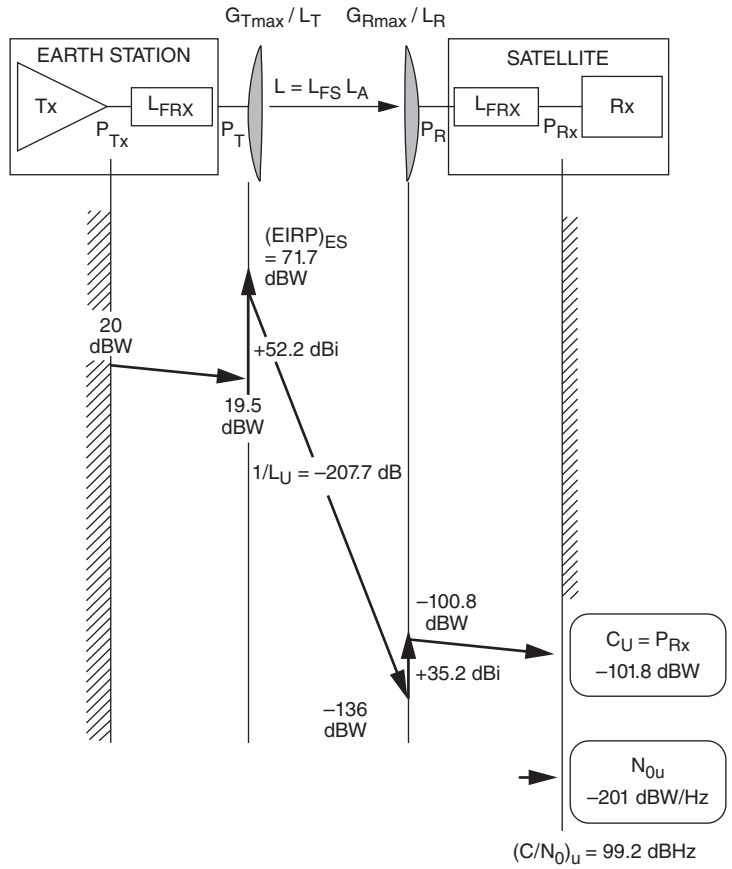


Figure 5.24 Variations in power for the clear sky uplink.

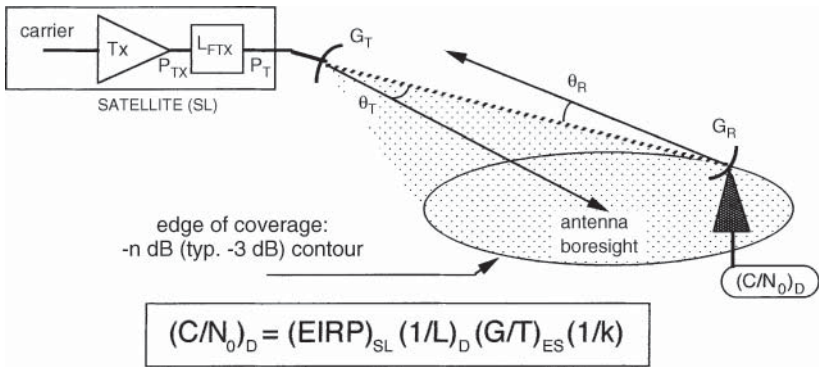


Figure 5.25 The geometry of a downlink.

— For the earth station (ES):
Receiver noise figure: $F = 1$ dB
Loss between antenna and receiver: $L_{FRX} = 0.5$ dB
Thermodynamic temperature of the feeder: $T_F = 290$ K
Antenna diameter: $D = 4$ m
Antenna efficiency: $\eta = 0.6$
Maximum pointing error: $\theta_R = 0.1°$
Ground noise temperature: $T_{GROUND} = 45$ K

To calculate the EIRP of the satellite:

$$(\text{EIRP})_{SL} = P_{TX} G_{Tmax} / L_T L_{FTX} \quad \text{(W)} \tag{5.42}$$

with:

$$P_{TX} = 10 \text{ W} = 10 \text{ dBW}$$
$$G_{Tmax} = \eta(\pi D/\lambda_D)^2 = \eta(\pi 70/\theta_{3\ dB})^2 = 0.55(\pi 70/2)^2 = 6650 = 38.2 \text{ dBi}$$
$$L_T(\text{dB}) = 3 \text{ dB(earth station on edge of coverage)}$$
$$L_{FTX} = 1 \text{ dB}$$

Hence:

$$(\text{EIRP})_{SL} = 10 \text{ dBW} + 38.2 \text{ dBi} - 3 \text{ dB} - 1 \text{ dB} = 44.2 \text{ dBW}$$

To calculate the attenuation on the downlink (D):

$$L_D = L_{FS} L_A \tag{5.43}$$

with:

$$L_{FS} = (4\pi R/\lambda_D)^2 = (4\pi R f_D/c)^2 = 4.04 \times 10^{20} = 206.1 \text{ dB}$$
$$L_A = 0.3 \text{ dB}$$

Hence:

$$L_D = 206.1 \text{ dB} + 0.3 \text{ dB} = 206.4 \text{ dB}$$

To calculate the figure of merit G/T of the earth station in the satellite direction:

$$(G/T)_{ES} = (G_{Rmax}/L_R L_{FRX} L_{POL})/T_D(K^{-1})$$

T_D is the downlink system noise temperature at the receiver input given by:

$$T_D = T_A/L_{FRX} + T_F(1 - 1/L_{FRX}) + T_{eRX}$$

and:

$$G_{Rmax} = \eta(\pi D/\lambda_D)^2 = \eta(\pi D f_D/c)^2 = 0.6(\pi \times 4 \times 12 \times 10^9/3 \times 10^8)^2$$
$$= 151\,597 = 51.8 \text{ dBi}$$
$$L_R(\text{dB}) = 12(\theta_R/\theta_{3\ dB})^2 = 12(\theta_R D f_D/70c)^2 = 0.6 \text{ dB}$$
$$L_{FRX} = 0.5 \text{ dB}$$
$$L_{POL} = 0 \text{ dB}$$

$T_A = T_{SKY} + T_{GROUND}$ with $T_{SKY} = 20\,K$ (see Figure 5.20 for $f = 12\,GHz$ and $E = 10°$) and $T_{GROUND} = 45\,K$, from which $T_A = 65\,K$:

$$T_F = 290\,K$$

$$T_{eRX} = (F-1)T_0 = (10^{0.1} - 1)290 = 75\,K$$

Hence:

$$T_D = 65/10^{0.05} + 290(1 - 1/10^{0.05}) + 75 = 164.5\,K$$

then:

$$(G/T)_{ES} = 51.8 - 0.6 - 0.5 - 10\ \log[65/10^{0.05} + 290(1 - 1/10^{0.05}) + 75]$$

$$= 28.5\,dBK^{-1}$$

To calculate the ratio C/N_0 for the downlink:

$$(C/N_0)_D = (EIRP)_{SL}\,(1/L_D)(G/T)_{ES}\,(1/k) \quad (Hz) \tag{5.44}$$

Hence:

$$(C/N_0)_D = 44.2\,dBW - 206.4\,dB + 28.5\,dBK^{-1} + 228.6\,dBW/HzK$$

$$= 94.9\,dBHz$$

Figure 5.26 summarises the variations of power level throughout the path.

5.7 INFLUENCE OF THE ATMOSPHERE

On both the up- and downlinks, the carrier passes through the atmosphere. Recall that the range of frequencies concerned is from 1 to 30 GHz. From the point of view of wave propagation at these frequencies, only two regions of the atmosphere have an influence – the troposphere and the ionosphere. The troposphere extends practically from the ground to an altitude of 15 km. The ionosphere is situated between around 70 and 1000 km. The regions where their influence is maximum are in the vicinity of the ground for the troposphere and at an altitude on the order of 400 km for the ionosphere.

The influence of the atmosphere has been mentioned previously in order to introduce the losses L_A due to atmospheric attenuation into Eq. (5.14) and in connection with antenna noise temperature. However, other phenomena can occur. Their nature and significance is now explained.

The predominant effects are those caused by absorption and depolarisation due to tropospheric precipitation (rain and snow). Dry snow has little effect. Although wet snowfalls can cause greater attenuation than the equivalent rainfall rate, this situation is rare and has little effect on attenuation statistics. Effects are particularly significant for frequencies greater than 10 GHz. The occurrence of rain is defined by the percentage of time during which a given rainfall rate is exceeded. Low rainfall rates with negligible effects correspond to high percentages of time (typically 20%); these are described as *clear sky* conditions. High rainfall rate, with significant effects, correspond to small percentages of time (typically 0.01%); these are described as *rain* conditions. These effects can degrade the quality of the link below an acceptable threshold. The availability of a link is thus directly related to the rainfall rate time statistics. In view of their importance, the effects of precipitation are presented first. The effects of other phenomena are examined later.

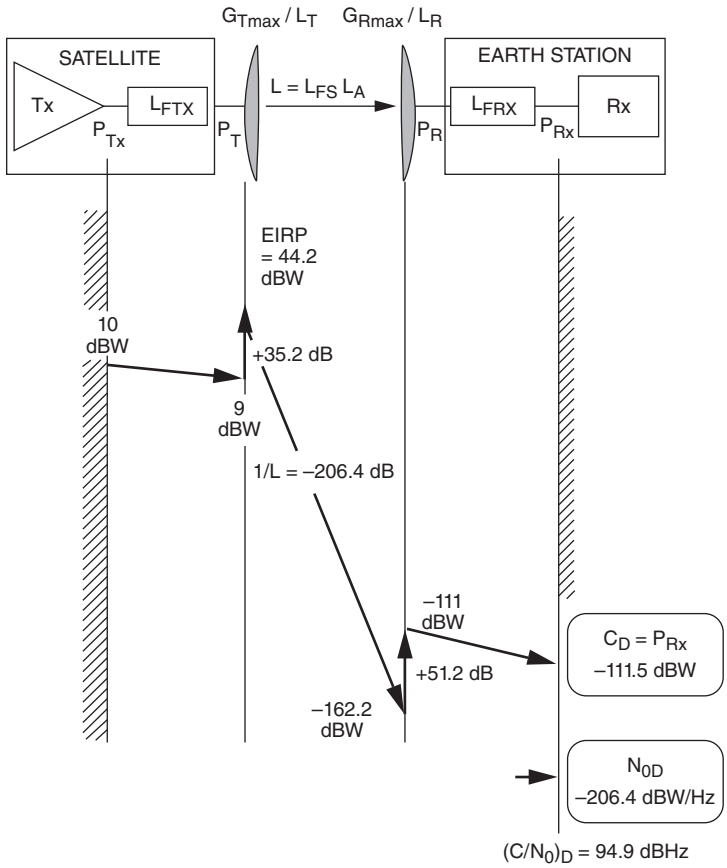


Figure 5.26 Variations in power for the clear sky downlink.

5.7.1 Impairments caused by rain

The intensity of precipitation is measured by the rainfall rate R_p expressed in mm/h. The temporal precipitation statistic is given by the cumulative probability distribution that indicates the annual percentage p (%) during which a given value of rainfall rate R_p (mm/h) is exceeded. In the absence of precipitation data for the location of the earth station involved in the link, data from Figure 5.27 [ITUR-17b] can be used. For instance, in Europe (Figure 5.27b), a rainfall rate of $R_{0.01}$ ($p = 0.01\%$ corresponds to $53\,\text{m}\,\text{y}^{-1}$) is around $30\,\text{mm}\,\text{h}^{-1}$ with the exception of some Mediterranean regions where the occurrence of storms (heavy precipitation for a short time interval) leads to a value of $R_{0.01} = 50\,\text{mm}\,\text{h}^{-1}$. In equatorial regions, $R_{0.01} = 120\,\text{mm}\,\text{h}^{-1}$ (Central America or Southeast Asia, for example). Rain causes attenuation and depolarisation.

5.7.1.1 Attenuation

The value of attenuation due to rain A_{RAIN} is given by the product of the specific attenuation γ_R(dB/km) and the effective path length of the wave in the rain L_e(km). That is:

$$A_{\text{RAIN}} = \gamma_R L_e \text{ (dB)} \tag{5.45}$$

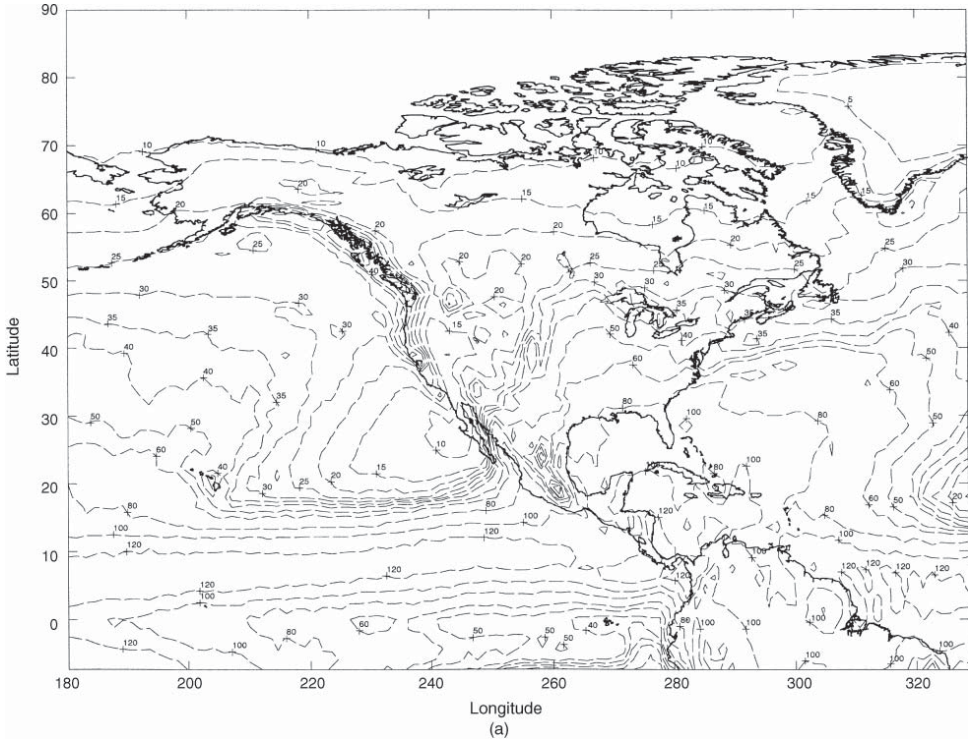


Figure 5.27 Rain intensity $R_{0.01}$ exceeded for more than 0.01% of the average year. Source: from [ITUR-17c]. Reproduced with the permission of the ITU.

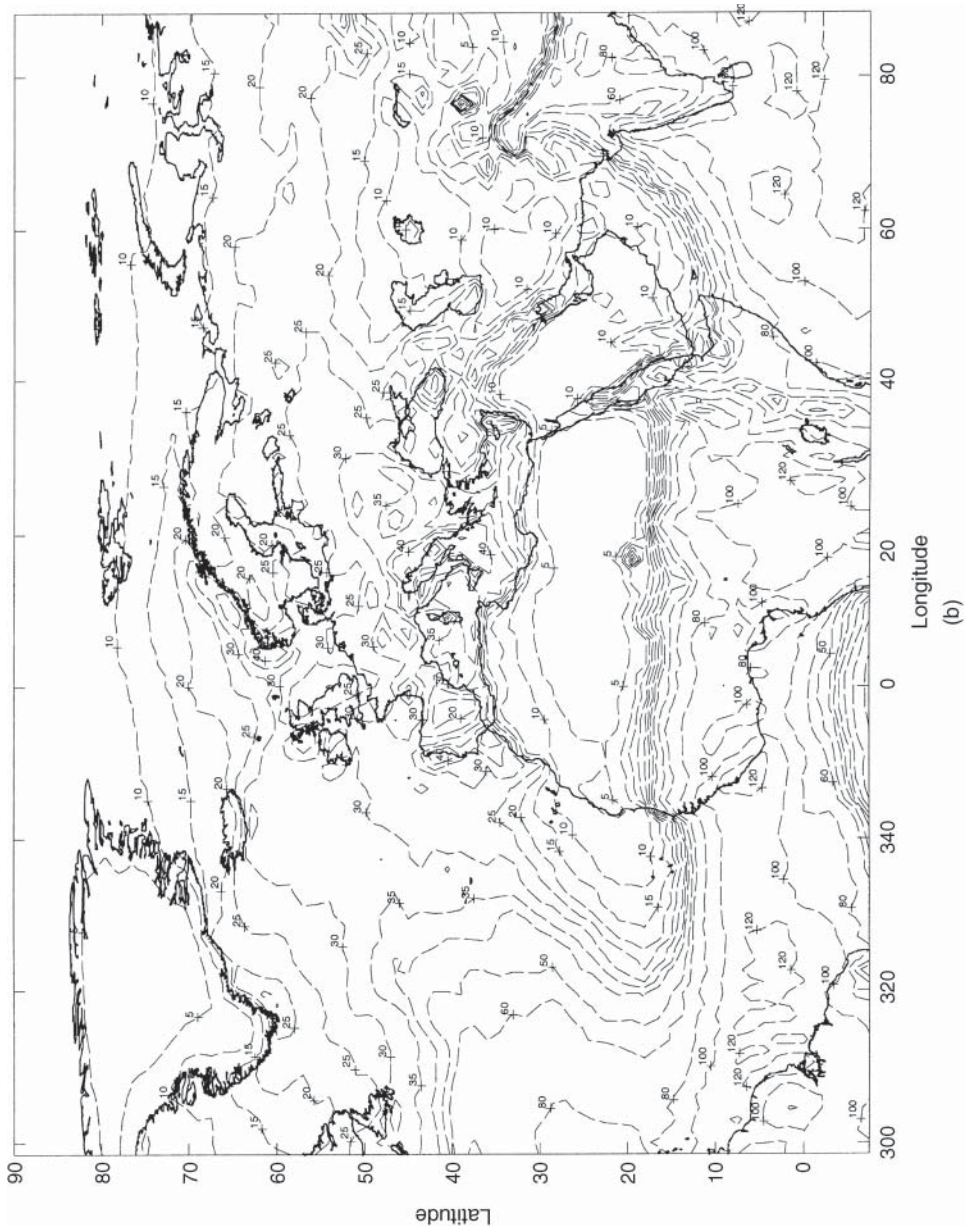


Figure 5.27 (*Continued*)

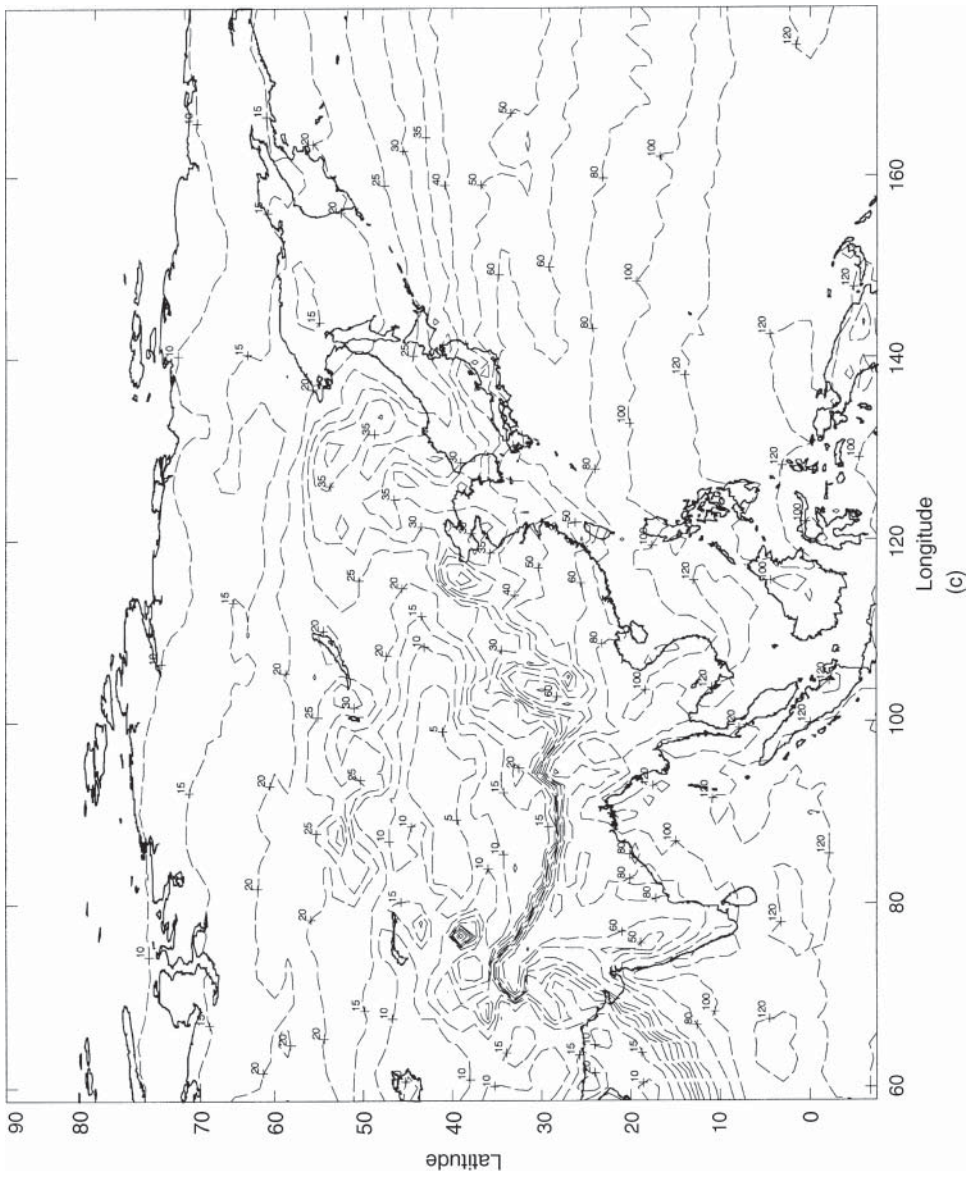


Figure 5.27 (*Continued*)

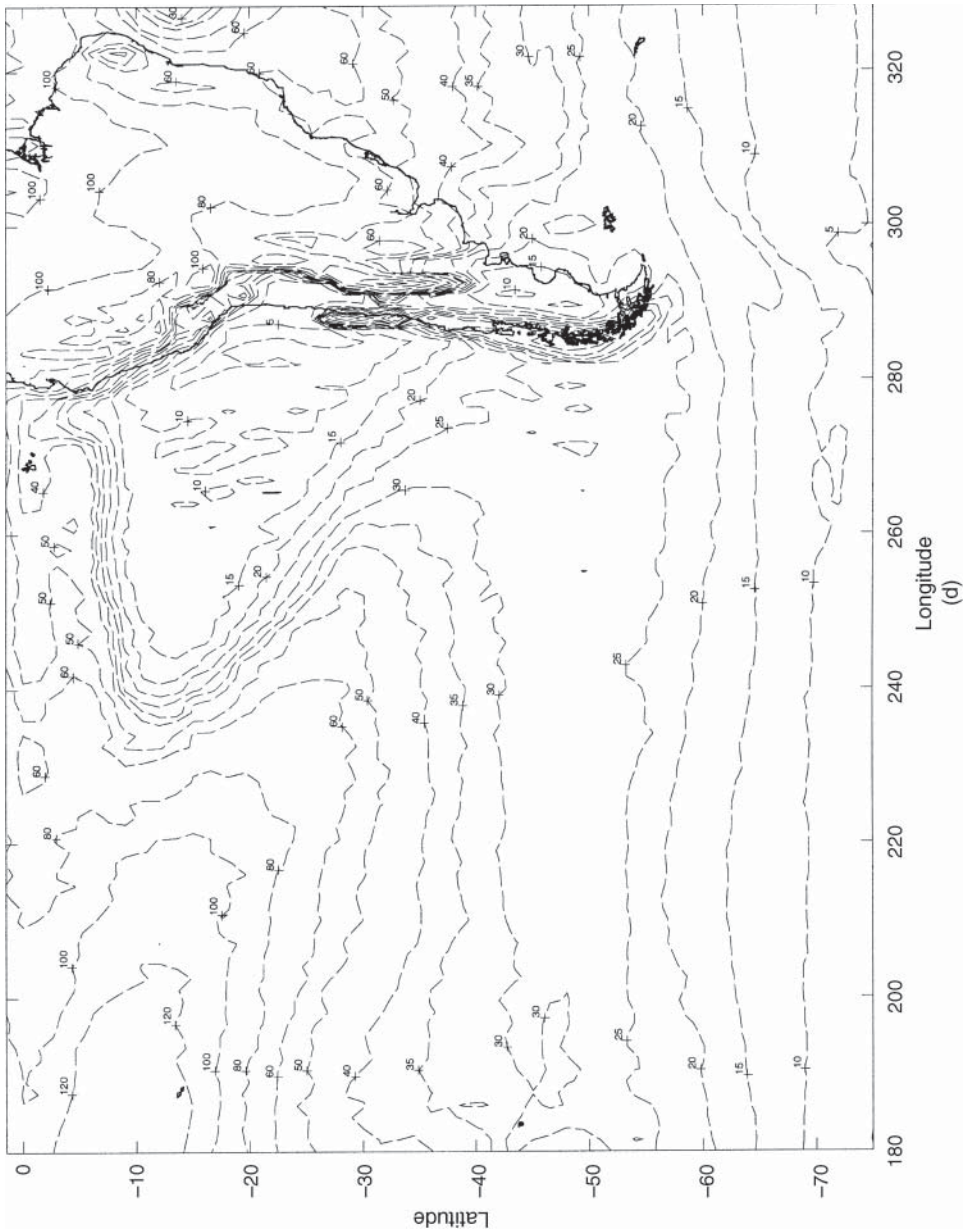


Figure 5.27 (*Continued*)

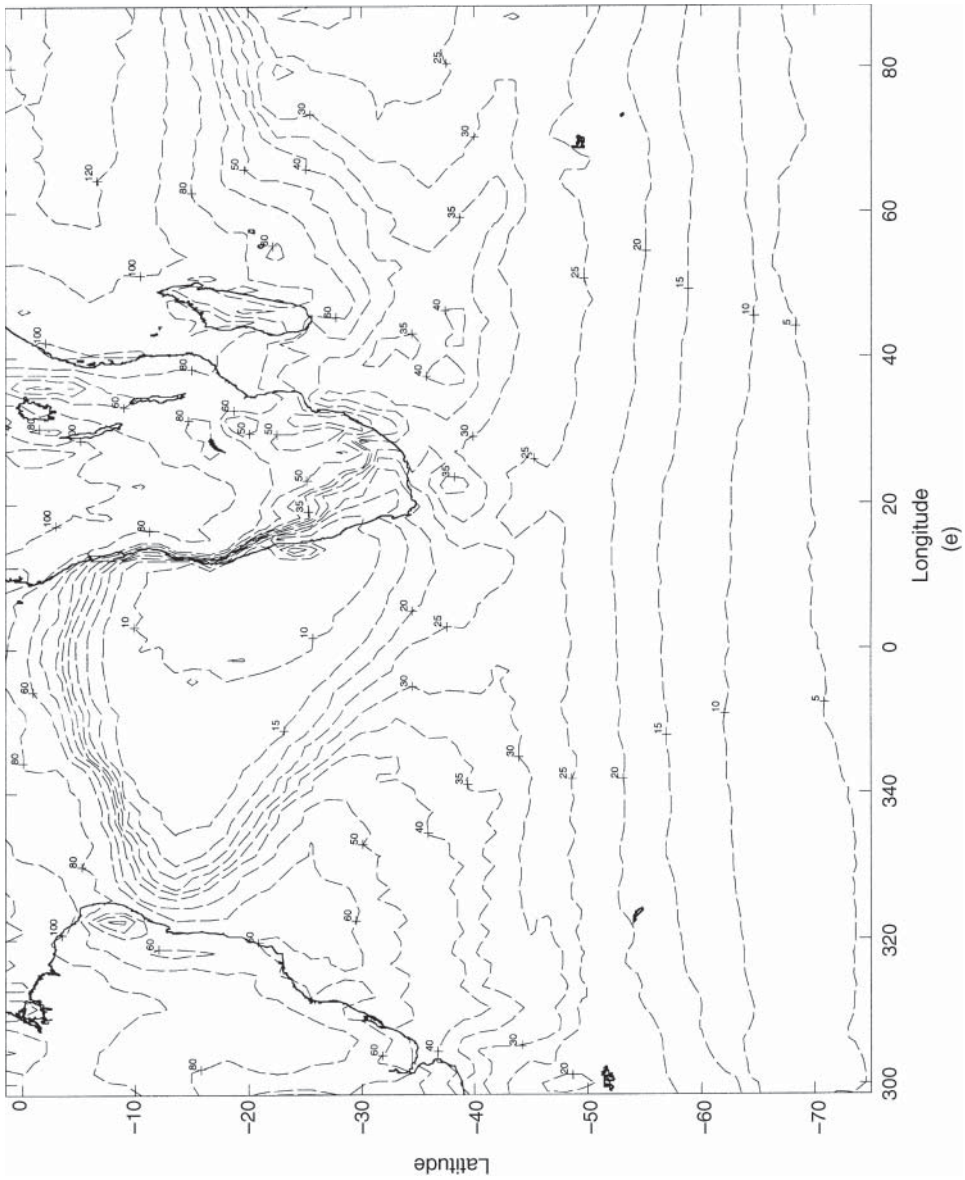


Figure 5.27 (*Continued*)

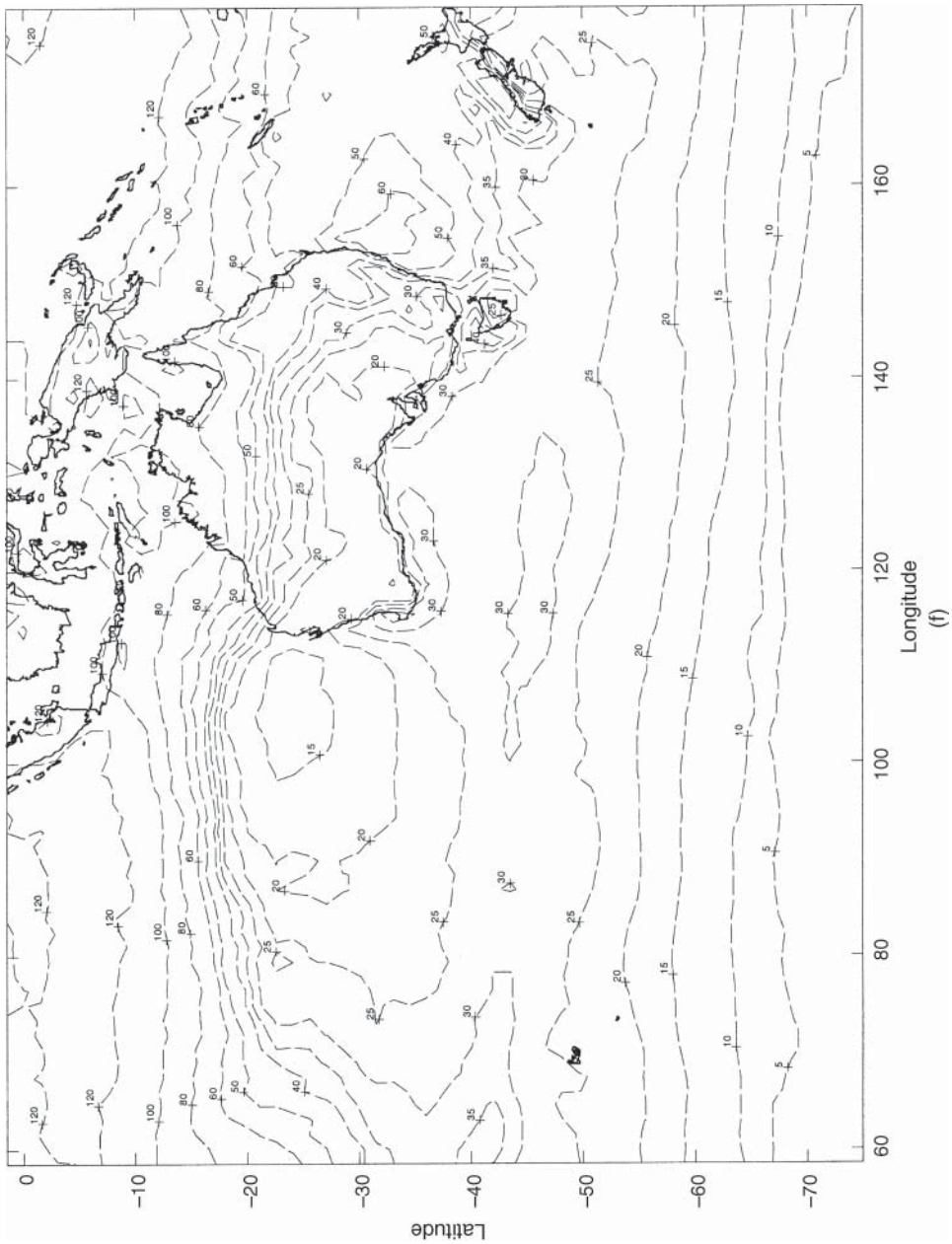


Figure 5.27 (*Continued*)

The value of γ_R depends on the frequency and intensity R_p (mm/h) of the rain. The result is a value of attenuation, which is exceeded during the percentage of time p. Determination of A_{RAIN} proceeds in several steps:

(1) Determine the rainfall rate $R_{0.01}$ exceeded for 0.01% of the time of an average year, where the earth station is located on the maps in Figure 5.27.
(2) Calculate the effective height of the rain h_R as given in [ITUR-13]:

$$h_R(\text{km}) = h_0 + 0.36 \text{ km}$$

where h_0 is the mean 0 °C isotherm height above mean sea level, given in Figure 5.28.
(3) Compute the slant-path length, L_s, below the rain height:

$$L_s = \frac{h_R - h_s}{\sin E} \quad (\text{km})$$

where h_s (km) is the earth station height above mean sea level and E is the satellite elevation angle. This is valid for $E \geq 5°$.
(4) Calculate the horizontal projection, L_G, of the slant-path length:

$$L_G = L_s \cos E \quad (\text{km})$$

(5) Obtain the specific attenuation, γ_R, as a function of $R_{0.01}$ and frequency from Table 5.1. These values are derived from the frequency-dependent coefficients given in [ITUR-05]:

$$\gamma_R = k(R_{0.01})^{\alpha} \quad (\text{dB/km})$$

where

$$k = [k_H + k_V + (k_H - k_V)\cos^2 \text{E} \cos 2\tau]/2$$

$$\alpha = [k_H\alpha_H + k_V\alpha_V + (k_H\alpha_H - k_V\alpha_V)\cos^2 \text{E} \cos 2\tau]/2k$$

and E is the elevation angle, and τ is the polarisation tilt angle relative to the horizontal ($\tau = 45°$ for circular polarisation). For a rapid but approximate estimate of γ_R, one can use the nomogram of Figure 5.29; for circular polarisation, take the mean value of the attenuation obtained for each linear polarisation.
(6) Calculate the horizontal reduction factor, $r_{0.01}$, for 0.01% of the time:
(enter L_G in km, γ_R in dB/km, f in GHz)

$$r_{0.01} = [1 + 0.78\sqrt{L_G\gamma_R/f} - 0.38(1 - \text{e}^{-2L_G})]^{-1}$$

(7) Calculate the vertical adjustment factor, $v_{0.01}$ for 0.01% of the time:

$$\zeta = \tan^{-1}\left(\frac{h_R - h_s}{L_G r_{0.01}}\right) \quad (\text{degrees})$$

$$L_R(\text{km}) = \begin{cases} L_G r_{0.01}/\cos E & \text{for } \zeta > E \\ (h_R - h_s)/\sin E & \text{otherwise} \end{cases}$$

$$\chi = \begin{cases} 36 - |\text{latitude}| & \text{if } |\text{latitude}| < 36° \\ 0 & \text{otherwise} \end{cases}$$

$$v_{0.01} = \left[1 + \sqrt{\sin E}\left(31(1 - \text{e}^{-(E/(1+\chi))})\frac{\sqrt{L_R\gamma_R}}{f^2} - 0.45\right)\right]^{-1}$$

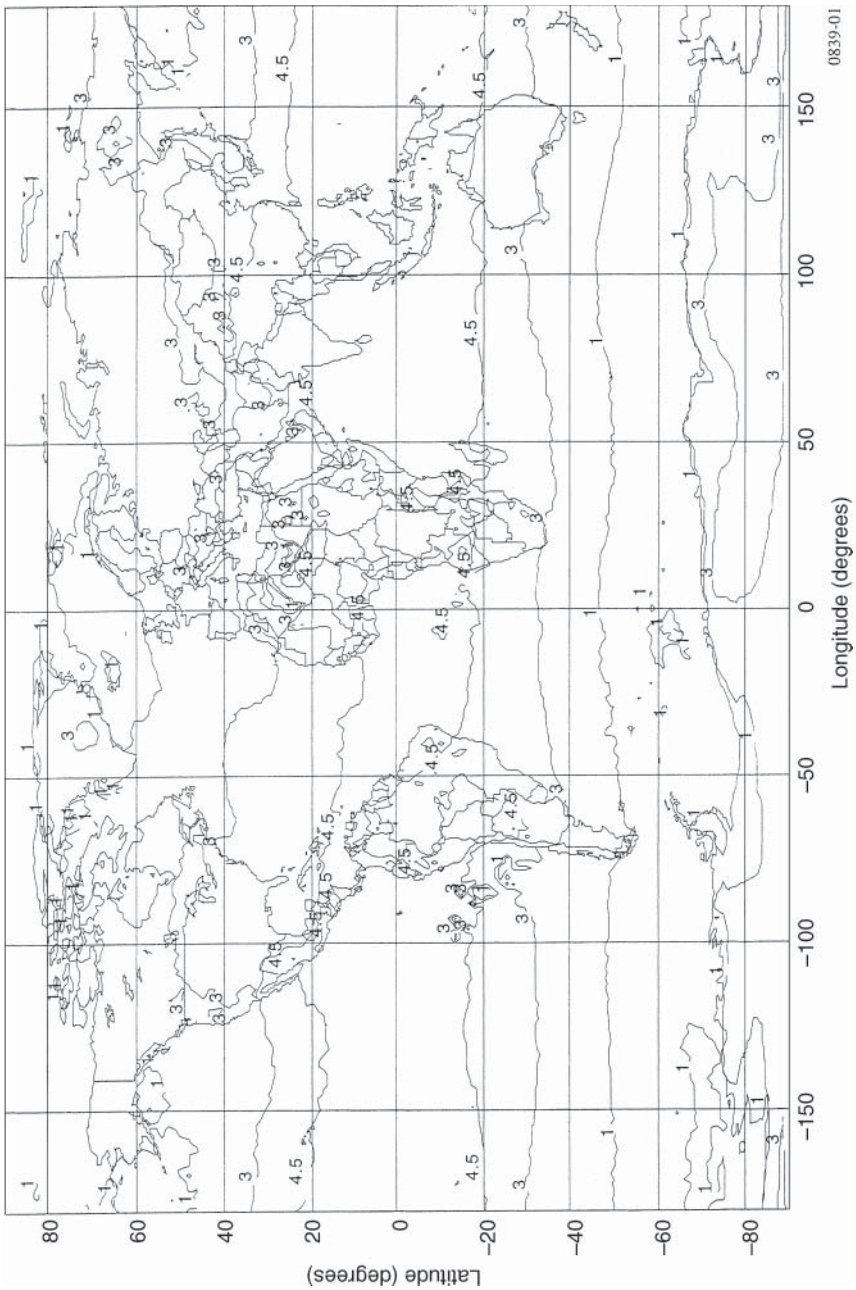


Figure 5.28 Yearly average 0 °C isotherm height above mean sea level (km).

(8) The effective path length is:

$$L_E = L_R \, v_{0.01} \quad \text{(km)}$$

(9) The predicted attenuation exceeded for 0.01% of an average year is obtained from:

$$A_{0.01} = \gamma_R \, L_E \text{ (dB)}$$

(10) The estimated attenuation to be exceeded for other percentages of an average year, in the range 0.001–5%, is determined from the attenuation to be exceeded for 0.01% for an average year:

$$\beta = \begin{cases} 0 & \text{if } p \geq 1\% \text{ or } |\text{latitude}| \geq 36° \\ -0.005(|\text{latitude}| - 36) & \text{if } p < 1\%, |\text{latitude}| < 36°, E \geq 25° \\ -0.005(|\text{latitude}| - 36) + 1.8 - 4.25 \sin E & \text{otherwise} \end{cases}$$

$$A_p = A_{0.01} \left(\frac{p}{0.01} \right)^{-(0.655 + 0.033 \ln p - 0.045 \ln A_{0.01} - \beta (1-p) \sin E)} \text{ (dB)}$$

It is sometimes required to estimate the attenuation exceeded during a percentage p_w of any month (that is the worst month). The corresponding annual percentage is given (for $1.9 \times 10^{-4} < p_w < 7.8$) by

$$p = 0.3(p_w)^{1.15} \quad (\%) \tag{5.46}$$

Performance objectives are often stipulated for $p_w = 0.3\%$ (Section 3.2). This corresponds to an annual percentage $p = 0.075\%$.

Typical values of attenuation due to rain exceeded for 0.01% of an average year can be deduced from the previous procedure for regions where the rainfall rate $R_{0.01}$ exceeded for 0.01% of an average year is in the range from 30 to 50 mm h^{-1}. This gives typically around 0.1 dB at 4 GHz, 5–10 dB at 12 GHz, 10–20 dB at 20 GHz, and 25–40 dB at 30 GHz.

5.7.1.2 Depolarisation

Cross-polarisation was identified in Section 5.2.3 as energy transfer from one polarisation to an orthogonal polarisation. Section 5.2.3 considered imperfect isolation between waves in two orthogonal polarisation states transmitted or received by a given antenna. Now consider cross-polarisation caused by wave depolarisation due to rain and ice clouds.

Rain introduces depolarisation as a result of differential attenuation and differential phase shift between two orthogonal characteristic polarisations. These effects originate in the non-spherical shape of raindrops. A commonly adopted model for a falling raindrop is an oblate spheroid with its major axis inclined to the horizontal and where deformation depends upon the radius of a sphere of equal volume. Assume the angles of inclination vary randomly in space and time.

Statistics for the cross-polarisation discrimination XPD_{rain} due to rain can be derived from rain attenuation statistics, i.e. the value of attenuation (called *copolar attenuation*), A_{RAIN} (p), which is exceeded during an annual percentage p for the considered polarisation.

Ice clouds, where high-altitude ice crystals are in a region close to the 0 °C isotherm, are also a cause of cross-polarisation. However, in contrast to rain, this effect is not accompanied by attenuation.

The cross-polarisation discrimination XPD (p) not exceeded for p% of the time is given by:

$$XPD(p) = XPD_{rain} - C_{ice}(\text{dB}) \tag{5.47}$$

Table 5.1 Values and interpolation formulas for frequency-dependent coefficients k_H, k_V, α_H, and α_V (log is base 10, i.e. log 10 = 1).

Frequency	Coefficients
$f = 1$ GHz	$k_H = 0.000\,038\,7$
	$k_V = 0.000\,035\,2$
	$\alpha_H = 0.912$
	$\alpha_V = 0.880$
1 GHz $\le f \le$ 2 GHz	$k_H = 3.870 \times 10^{-5} f_{GHz}^{1.9925}$
	$k_V = 3.520 \times 10^{-5} f_{GHz}^{1.9710}$
	$\alpha_H = 0.1694 \log f_{GHz} + 0.9120$
	$\alpha_V = 0.1428 \log f_{GHz} + 0.8800$
$f = 2$ GHz	$k_H = 0.000\,154$
	$k_V = 0.000\,138$
	$\alpha_H = 0.963$
	$\alpha_V = 0.923$
2 GHz $\le f \le$ 4 GHz	$k_H = 3.649 \times 10^{-5} f_{GHz}^{2.0775}$
	$k_V = 3.222 \times 10^{-5} f_{GHz}^{2.0985}$
	$\alpha_H = 0.5249 \log f_{GHz} + 0.8050$
	$\alpha_V = 0.5049 \log f_{GHz} + 0.7710$
$f = 4$ GHz	$k_H = 0.000\,650$
	$k_V = 0.000\,591$
	$\alpha_H = 1.121$
	$\alpha_V = 1.075$
4 GHz $\le f \le$ 6 GHz	$k_H = 2.199 \times 10^{-5} f_{GHz}^{2.4426}$
	$k_V = 2.187 \times 10^{-5} f_{GHz}^{2.3780}$
	$\alpha_H = 1.0619 \log f_{GHz} + 0.4816$
	$\alpha_V = 1.0790 \log f_{GHz} + 0.4254$
$f = 6$ GHz	$k_H = 0.001\,75$
	$k_V = 0.001\,55$
	$\alpha_H = 1.308$
	$\alpha_V = 1.265$
6 GHz $\le f \le$ 7 GHz	$k_H = 3.202 \times 10^{-6} f_{GHz}^{3.5181}$
	$k_V = 3.041 \times 10^{-6} f_{GHz}^{3.4791}$
	$\alpha_H = 0.3585 \log f_{GHz} + 1.0290$
	$\alpha_V = 0.7021 \log f_{GHz} + 0.7187$
$f = 7$ GHz	$k_H = 0.003\,01$
	$k_V = 0.002\,65$
	$\alpha_H = 1.332$
	$\alpha_V = 1.312$
7 GHz $\le f \le$ 8 GHz	$k_H = 7.542 \times 10^{-6} f_{GHz}^{3.0778}$
	$k_V = 7.890 \times 10^{-6} f_{GHz}^{2.9892}$
	$\alpha_H = -0.0862 \log f + 1.4049$
	$\alpha_V = -0.0345 \log f + 1.3411$
$f = 8$ GHz	$k_H = 0.004\,54$
	$k_V = 0.003\,95$
	$\alpha_H = 1.327$
	$\alpha_V = 1.310$

Table 5.1 (*continued*)

Frequency	Coefficients
$8\,\text{GHz} \le f \le 10\,\text{GHz}$	$k_H = 2.636 \times 10^{-6} f_{GHz}^{3.5834}$
	$k_V = 2.102 \times 10^{-6} f_{GHz}^{3.6253}$
	$\alpha_H = -0.5263 \log f_{GHz} + 1.8023$
	$\alpha_V = -0.4747 \log f_{GHz} + 1.7387$
$f = 10\,\text{GHz}$	$k_H = 0.0101$
	$k_V = 0.008\,87$
	$\alpha_H = 1.276$
	$\alpha_V = 1.264$
$10\,\text{GHz} \le f \le 12\,\text{GHz}$	$k_H = 3.949 \times 10^{-6} f_{GHz}^{3.4078}$
	$k_V = 2.785 \times 10^{-6} f_{GHz}^{3.5032}$
	$\alpha_H = -0.7451 \log f_{GHz} + 2.0211$
	$\alpha_V = -0.8083 \log f_{GHz} + 2.0723$
$f = 12\,\text{GHz}$	$k_H = 0.0188$
	$k_V = 0.0168$
	$\alpha_H = 1.217$
	$\alpha_V = 1.200$
$12\,\text{GHz} \le f \le 15\,\text{GHz}$	$k_H = 1.094 \times 10^{-5} f_{GHz}^{2.9977}$
	$k_V = 7.718 \times 10^{-6} f_{GHz}^{3.0929}$
	$\alpha_H = -0.6501 \log f_{GHz} + 1.9186$
	$\alpha_V = -0.7430 \log f_{GHz} + 2.0018$
$f = 15\,\text{GHz}$	$k_H = 0.0367$
	$k_V = 0.0335$
	$\alpha_H = 1.154$
	$\alpha_V = 1.128$
$15\,\text{GHz} \le f \le 20\,\text{GHz}$	$k_H = 4.339 \times 10^{-5} f_{GHz}^{2.4890}$
	$k_V = 3.674 \times 10^{-5} f_{GHz}^{2.5167}$
	$\alpha_H = -0.4402 \log f_{GHz} + 1.6717$
	$\alpha_V = -0.5042 \log f_{GHz} + 1.7210$
$f = 20\,\text{GHz}$	$k_H = 0.0751$
	$k_V = 0.0691$
	$\alpha_H = 1.099$
	$\alpha_V = 1.065$
$20\,\text{GHz} \le f \le 25\,\text{GHz}$	$k_H = 8.951 \times 10^{-5} f_{GHz}^{2.2473}$
	$k_V = 3.674 \times 10^{-5} f_{GHz}^{2.2041}$
	$\alpha_H = -0.3921 \log f_{GHz} + 1.6092$
	$\alpha_V = -0.3612 \log f_{GHz} + 1.5349$
$f = 25\,\text{GHz}$	$k_H = 0.124$
	$k_V = 0.1113$
	$\alpha_H = 1.061$
	$\alpha_V = 1.030$
$25\,\text{GHz} \le f \le 30\,\text{GHz}$	$k_H = 8.779 \times 10^{-5} f_{GHz}^{2.2533}$
	$k_V = 1.143 \times 10^{-4} f_{GHz}^{2.1424}$
	$\alpha_H = -0.5052 \log f_{GHz} + 1.7672$
	$\alpha_V = -0.3789 \log f_{GHz} + 1.5596$

Table 5.1 (*continued*)

Frequency	Coefficients
$f = 30$ GHz	$k_H = 0.187$
	$k_V = 0.167$
	$\alpha_H = 1.021$
	$\alpha_V = 1.000$
30 GHz $\le f \le$ 35 GHz	$k_H = 1.009 \times 10^{-4} f_{GHz}^{2.2124}$
	$k_V = 1.075 \times 10^{-4} f_{GHz}^{2.1605}$
	$\alpha_H = -0.6274 \log f_{GHz} + 1.9477$
	$\alpha_V = -0.5527 \log f_{GHz} + 1.8164$
$f = 35$ GHz	$k_H = 0.263$
	$k_V = 0.233$
	$\alpha_H = 0.979$
	$\alpha_V = 0.963$
35 GHz $\le f \le$ 40 GHz	$k_H = 1.304 \times 10^{-4} f_{GHz}^{2.1402}$
	$k_H = 1.163 \times 10^{-4} f_{GHz}^{2.1383}$
	$\alpha_H = -0.6898 \log f_{GHz} + 2.0440$
	$\alpha_H = -0.5863 \log f_{GHz} + 1.8683$
40 GHz	$k_H = 0.350$
	$k_V = 0.310$
	$\alpha_H = 0.939$
	$\alpha_V = 0.929$

where XPD_{rain} is the cross-polarisation discrimination due to rain and C_{ice} is the contribution of ice clouds, respectively given by:

$$XPD_{rain} = C_f - C_A + C_\tau + C_\theta + C_\sigma (dB)$$

$$C_{ice} = XPD_{rain}(0.3 + 0.1 \log p)/2 (dB)$$

where:

$$C_f = 30 \log f$$

$$C_A = V(f) \log A_{RAIN}(p)$$

$$V(f) = 12.8 f^{0.19} \quad \text{for} \quad 8 \le f \le 20 \quad (GHz)$$

$$V(f) = 22.6 \quad \text{for} \quad 20 \le f \le 35 \quad (GHz)$$

$$C_\tau = -10 \log[1 - 0.484(1 + \cos 4\tau)]$$

where f is the frequency (GHz) and τ is the tilt angle of the linearly polarised electric field vector with respect to the horizontal (for circular polarisation, use $\tau = 45°$).

$$C_\theta = -40 \log(\cos E) \quad \text{for} \quad E \le 60°$$

where E is the elevation angle.

$$C_\sigma = 0.0052\sigma^2$$

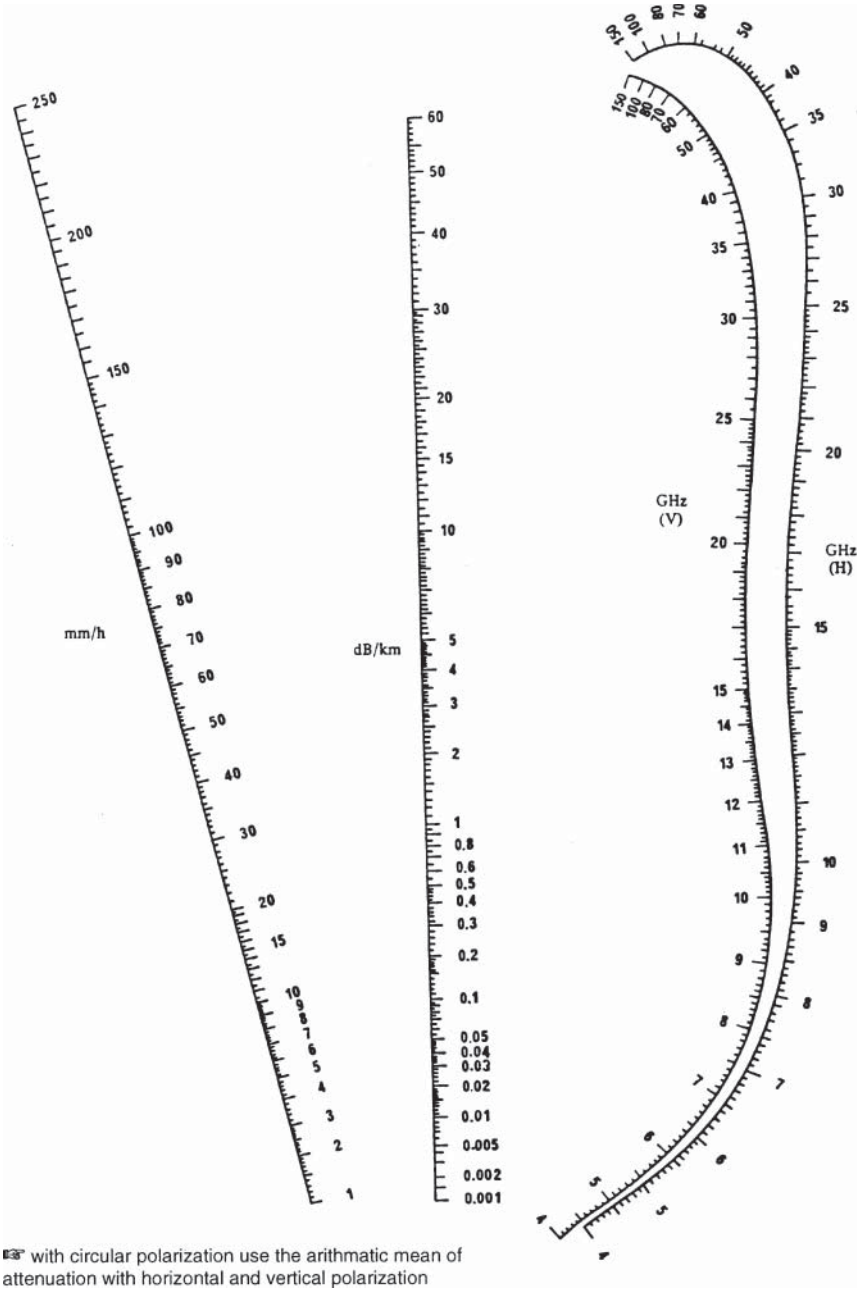


Figure 5.29 Nomogram for determination of the specific attenuation γ_R as a function of the frequency (GHz) and rain density R (mm/h). Source: CCIR Rep. 721. Reproduced with the permission of the ITU.

where σ is the standard deviation of the raindrop inclination angle distribution, expressed in degrees; σ takes the values 0°, 5°, 10°, and 15° for p = 1%, 0.1%, 0.01%, and 0.001% of the time, respectively.

Equation (5.47) is in agreement with long-term measurements for 8 GHz $\leq f \leq$ 35 GHz and elevation angle $E \leq 60°$. For lower frequencies, down to 4 GHz, one can calculate $XPD_1(p)$ at frequency f_1(8 GHz $\leq f_1 \leq$ 30 GHz) according to Eq. (5.47) and derive XPD_2(p) at frequency f_2 (4 GHz $\leq f_2 \leq$ 8 GHz) from the following semi-empirical formula:

$$XPD_2(p) = \frac{XPD_1(p) - 20\ \log[f_2[1 - 0.484(1 + \cos 4\tau_2)]^{0.5}}{f_1[1 - 0.484(1 + \cos\ 4\tau_1)]^{0.5}]}$$

where τ_1 and τ_2 are the respective polarisation tilt angles at frequencies f_1 and f_2.

5.7.2 Other impairments

5.7.2.1 Attenuation by atmospheric gases

Attenuation due to gas in the atmosphere depends on the frequency, elevation angle, altitude of the station, and water vapour concentration [ITUR-16a]. Figure 5.30 displays the attenuation for a standard atmosphere. The attenuation is negligible at frequencies less than 10 GHz and does not exceed 3 dB at 22.24 GHz (the frequency corresponding to a water vapour absorption band) for mean atmospheric humidity and elevation angles greater than 10°.

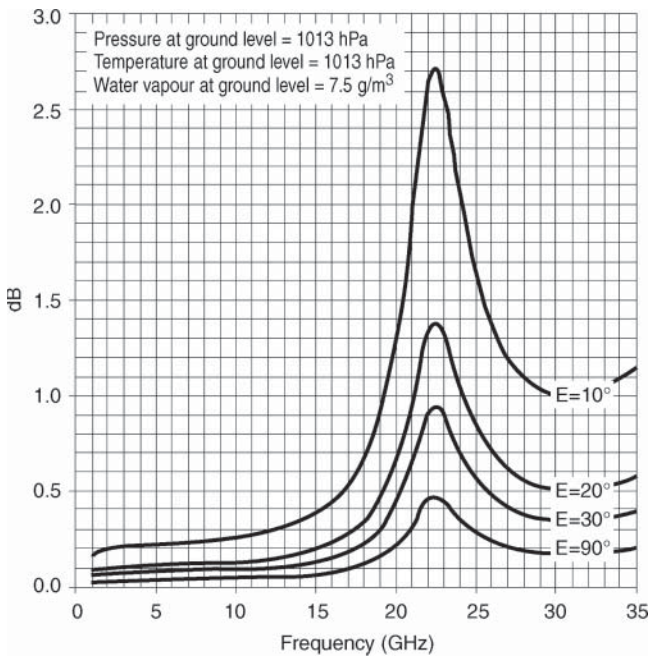


Figure 5.30 Attenuation due to atmospheric gases as a function of frequency and elevation angle E for a standard atmosphere with water vapour content at ground level of 7.5 g m^{-3}.

5.7.2.2 *Attenuation due to rain, fog, or ice clouds*

Attenuation due to rain clouds or fog can be calculated according to [ITUR-17a]. The specific attenuation γ_C is calculated as:

$$\gamma_C = KM \text{ (dB/km)} \tag{5.48}$$

where an approximate value is $K = 1.2 \times 10^{-3} f^{1.9}$ (dB/km)/(g/m^3), f is expressed in GHz from 1 to 30 GHz, K is in (dB/km)/(g/m^3), and M is water concentration of the cloud or fog (g/m^3).

Attenuation due to rain clouds and fog is usually small compared with that due to rain. This attenuation, however, is observed for a greater percentage of the time. For an elevation angle $E = 20°$, the attenuation exceeded for 1% of the year due to rain clouds is on the order of 0.2 dB at 12 GHz, 0.5 dB at 20 GHz, and 1.1 dB at 30 GHz in North America and Europe, and 0.8 dB at 12 GHz, 2.1 dB at 20 GHz, and 4.5 dB at 30 GHz in Southeast Asia. For thick fog ($M = 0.5$ g m^{-3}), the attenuation is on the order of 0.4 dB km^{-1} at 30 GHz. Attenuation due to ice clouds is smaller still.

5.7.2.3 *Attenuation by sandstorms*

The specific attenuation (dB/km) is inversely proportional to the visibility and depends strongly on the humidity of the particles. At 14 GHz, it is on the order of 0.03 dB km^{-1} for dry particles and 0.65 dB km^{-1} for particles of 20% humidity. If the path length is 3 km, the attenuation can reach 1–2 dB.

5.7.2.4 *Scintillation*

Scintillation is a variation of the amplitude of received carriers caused by variations of the refractive index of the troposphere and the ionosphere. The peak-to-peak amplitude of these variations, at Ku band and medium latitudes, can exceed 1 dB for 0.01% of the time. The troposphere and the ionosphere have different refractive indices. The refractive index of the troposphere decreases with altitude, is a function of meteorological conditions and is independent of frequency. That of the ionosphere depends on frequency and the electronic content of the ionosphere. Both are subject to rapid local fluctuations. The effect of refraction is to cause curvature of the trajectory of the wave and variation of wave velocity, and hence propagation time. The most troublesome scintillation is ionospheric scintillation; it is greater when the frequency is low and the earth station is close to the equator.

5.7.2.5 *The Faraday rotation*

The ionosphere introduces a rotation of the plane of polarisation of a linearly polarised wave. The angle of rotation is inversely proportional to the square of the frequency. It is a function of the electronic content of the ionosphere and consequently varies with time, the season, and the solar cycle. The order of magnitude is several degrees at 4 GHz. The result, for a small percentage of the time, is an attenuation L_{POL}(dB) = −20 log(cos $\Delta\psi$) of the copolar carrier (see Section 5.4.2.3) where $\Delta\psi$ is the polarisation mismatch angle due to Faraday rotation and the appearance of a cross-polarised component that reduces the value of cross-polarisation discrimination XPD. The value of XPD is given by XPD (dB) = −20 log(tan $\Delta\psi$). For the case of a rotation $\Delta\psi = 9°$ at a frequency of 4 GHz, this gives $L_{POL} = 0.1$ dB and XPD = 16 dB. As seen from the earth station, the planes of polarisation rotate in the same direction on the uplink and the downlink. It is therefore

not possible to compensate for Faraday rotation by rotating the feed system of the antenna, if the antenna is used for both transmission and reception.

5.7.2.6 Multipath contributions

When the earth station antenna is small and hence has a beam with a large angular width, the received carrier can be the result of a direct path, and contributions of significant amplitude can be received after reflection on the ground or surrounding obstacles (multipath). In the case of destructive combination (phase opposition), a large attenuation is observed. This effect is weak when the earth station is equipped with an antenna that is sufficiently directional to eliminate multipath contributions. Multipath effects become predominant for mobile communications with nondirectional terminal antennas.

5.7.3 Link impairments – relative importance

At low frequencies (less than 10 GHz), the attenuation L_A is generally small and the principal cause of degradation of the link is cross-polarisation. This is caused by the ionosphere and by high-altitude ice crystals in the troposphere. At higher frequencies, the phenomena of attenuation and cross-polarisation are both observed. These are caused essentially by atmospheric gases, rainfall, and other hydrometeors.

Statistically, these phenomena become greater when a short percentage of time is considered. The availability of the link increases when these effects can be compensated for. Compensation techniques are available and are discussed in Section 5.8.

5.7.4 Link performance under rain conditions

5.7.4.1 Uplink performance

In the presence of rain, propagation attenuation is greater due to the attenuation A_{RAIN} caused by rain in the atmosphere. This is in addition to the attenuation due to gases in the atmosphere (0.3 dB). A typical value of attenuation due to rain for an earth station situated in a temperate climate (for example, in Europe) can be considered to be $A_{\text{RAIN}} = 10$ dB. Such an attenuation would not be exceeded, at a frequency of 14 GHz, for more than 0.01% of an average year. This gives $L_A = 0.3\text{ dB} + 10\text{ dB} = 10.3\text{ dB}$.

Hence:

$$L_{\text{U}} = 207.4\text{ dB} + 10.3\text{ dB} = 217.7\text{ dB}$$

Referring to the example of Section 5.6.2, the uplink performance under rain conditions becomes:

$$(C/N_0)_{\text{U}} = 71.7\text{ dBW} - 217.7\text{ dB} + 6.6\text{ dBK}^{-1} + 228.6\text{ dBW/Hz K} = 89.2\text{ dBHz}$$

The ratio $(C/N_0)_{\text{U}}$ for the uplink would be greater than the value calculated in this way for 99.99% of an average year.

5.7.4.2 Downlink performance

Referring now to the example in Section 5.6.3, $A_{\text{RAIN}} = 7$ dB is taken as the typical value of attenuation due to rain for an earth station situated in a temperate climate (for example, in Europe) that will not be exceeded, at a frequency of 12 GHz, for more than 0.01% of an average year; this gives $L_{\text{A}} = 0.3\,\text{dB} + 7\,\text{dB} = 7.3\,\text{dB}$. Hence, $L_{\text{D}} = 206.1 + 7.3\,\text{dB} = 213.4\,\text{dB}$. The antenna noise temperature is given by:

$$T_A = T_{\text{SKY}}/A_{\text{RAIN}} + T_{\text{m}}(1 - 1/A_{\text{RAIN}}) + T_{\text{GROUND}} \quad \text{(K)} \tag{5.49}$$

Taking

$$T_{\text{m}} = 275\,\text{K}$$
$$T_A = 20/10^{0.7} + 275(1 - 1/10^{0.7}) + 45 = 269\,\text{K}$$
$$T_D = 269/10^{0.05} + 290(1 - 1/10^{0.05}) + 75 = 346\,\text{K}$$

Hence

$$(G/T)_{\text{ES}} = 51.8 - 0.6 - 0.5 - 10\,\log[269/10^{0.05} + 290(1 - 1/10^{0.05}) + 75]$$
$$= 25.3\,\text{dBK}^{-1}$$

To calculate the ratio C/N_0 for the downlink:

$$(C/N_0)_{\text{D}} = (\text{EIRP})_{\text{SL}}(1/L_{\text{D}})(G/T)_{\text{ES}}(1/k) \quad \text{(Hz)}$$

Hence:

$$(C/N_0)_{\text{D}} = 44.2\,\text{dBW} - 213.4\,\text{dB} + 25.3\,\text{dBK}^{-1} + 228.6\,\text{dBW/HzK} = 84.7\,\text{dBHz}$$

The ratio $(C/N_0)_{\text{D}}$ for the downlink would be greater than the value calculated in this way for 99.99% of an average year.

5.7.5 Conclusion

The quality of the link between a transmitter and a receiver can be characterised by the ratio of the carrier power to the noise power spectral density C/N_0. This is a function of the transmitter EIRP, the receiver figure of merit G/T, and the properties of the transmission medium. In a satellite link between two earth stations, two links must be considered – the uplink, characterised by the ratio $(C/N_0)_{\text{U}}$; and the downlink, characterised by the ratio $(C/N_0)_{\text{D}}$. The propagation conditions in the atmosphere affect the uplink and downlink differently; rain reduces the value of the ratio $(C/N_0)_{\text{U}}$ by decreasing the value of received power C_{U}, while it reduces the value of $(C/N_0)_{\text{D}}$ by reducing the value of received power C_{D} and increasing the downlink system noise temperature. Denoting the resulting degradation by $\Delta(C/N_0)$ gives:

$$\Delta(C/N_0)_{\text{U}} = \Delta C_{\text{U}} = (A_{\text{RAIN}})_{\text{U}} \quad \text{(dB)} \tag{5.50}$$

$$\Delta(C/N_0)_{\text{D}} = \Delta C_{\text{D}} - \Delta(G/T) = (A_{\text{RAIN}})_{\text{D}} + \Delta T \quad \text{(dB)} \tag{5.51}$$

5.8 MITIGATION OF ATMOSPHERIC IMPAIRMENTS

5.8.1 Depolarisation mitigation

The method of compensation relies on modification of the polarisation characteristics of the earth station (see Chapter 8). Compensation is achieved as follows:

— For the uplink, by correcting the polarisation of the transmitting antenna by anticipation so that the wave arrives matched to the satellite antenna.
— For the downlink, by matching the antenna polarisation to that of the received wave.

Compensation can be automatic; the signals transmitted by the satellite must be made available (as beacons) so that the effects of the propagating medium can be detected and the required control signal deduced.

5.8.2 Attenuation mitigation

The mission specifies a value of the ratio C/N_0 greater than or equal to $(C/N_0)_{\text{required}}$ during a given percentage of the time, equal to $(100 - p)\%$. For example, 99.99% of the time implies $p = 0.01\%$. As seen in Section 5.7.5, the attenuation A_{RAIN} due to rain causes a reduction of the ratio C/N_0 given by:

$$(C/N_0)_{\text{rain}} = (C/N_0)_{\text{clear sky}} - A_{\text{RAIN}}(\text{dB}) \ \ (\text{dBHz}) \tag{5.52}$$

for an uplink and:

$$(C/N_0)_{\text{rain}} = (C/N_0)_{\text{clear sky}} - A_{\text{RAIN}}(\text{dB}) - \Delta(G/T) \ \ (\text{dBHz}) \tag{5.53}$$

for a downlink.

$\Delta(G/T) = (G/T)_{\text{clear sky}} - (G/T)_{\text{rain}}$ represents the reduction (in dB) of the figure of merit of the earth station due to the increase of noise temperature.

For a successful mission, one must have $(C/N_0)_{\text{rain}} = (C/N_0)_{\text{required}}$; this can be achieved by including a margin M(p) in the clear sky link budget with M(p) defined by:

$$\begin{aligned} M(p) &= (C/N_0)_{\text{clear sky}} - (C/N_0)_{\text{required}} \\ &= (C/N_0)_{\text{clear sky}} - (C/N_0)_{\text{rain}} \ \ (\text{dB}) \end{aligned} \tag{5.54}$$

The value of A_{RAIN} to be used is a function of the time percentage p. It increases as p decreases.

Making provision for a margin M(p) in the clear sky link requirement implies an increase of the EIRP that requires a higher transmitting power. For high attenuations, which are encountered for a small percentage of the time and at the highest frequencies (see Section 5.7.1.1), the extra power necessary can exceed the capabilities of the transmitting equipment and other solutions must then be considered: site diversity and adaptivity.

5.8.3 Site diversity

High attenuations are due to regions of rain of small geographical extent. Two earth stations at two distinct locations can establish links with the satellite that, at a given time t, suffer attenuations of $A_1(t)$ and $A_2(t)$, respectively; $A_1(t)$ is different from $A_2(t)$ as long as the geographical

separation is sufficient. The signals are thus routed to the link less affected by attenuation. On this link, the attenuation is $A_D(t) = \min[A_1(t), A_2(t)]$. The mean attenuation for a single location is defined as $A_M(t) = [A_1(t) + A_2(t)]/2$; all values are in dB.

Two concepts are useful to quantify the improvement provided by location diversity [ITUR-17c]:

— The diversity gain
— The diversity improvement factor

5.8.3.1 *Diversity gain GD(p)*

This is the difference (in dB) between the mean attenuation at a single location $A_M(t)$ exceeded for a time percentage p, and the attenuation with diversity $A_D(p)$ exceeded for the same time percentage p. Hence, for a downlink, for example, the required margin $M(p)$ at a given location is obtained from:

$$M(p) = A_{RAIN} + \Delta(G/T) \quad \text{(dB)} \tag{5.55}$$

With site diversity, the required margin becomes:

$$M(p) = A_{RAIN} + \Delta(G/T) - G_D(p) \quad \text{(dB)} \tag{5.56}$$

5.8.3.2 *Diversity improvement factor FD*

This is the ratio between the percentage of time p_1 during which the mean attenuation at a single site exceeds the value A dB and the percentage of time p_2 during which the attenuation with diversity exceeds the same value A dB.

Figure 5.31 shows the relationship between p_2 and p_1 as a function of the distance between the two locations. These curves can be modelled by the following relations:

$$p_2 = (p_1)^2(1 + \beta^2)/(p_1 + 100\beta^2) \tag{5.57}$$

with $\beta^2 = 10^{-4}d^{1.33}$, when the distance $d > 5$ km.

Site diversity also provides protection against scintillation and cross-polarisation.

5.8.4 Adaptivity

Adaptivity involves the variation of certain parameters of the link for the duration of the attenuation in such a way as to maintain the required value for the ratio C/N_0. Several approaches can be envisaged as follows [CAS-98]:

— Assignment of an additional resource, which is normally kept in reserve, to the link affected by attenuation. This additional resource can be:
  — An increase in transmission time (such as an unoccupied frame time slot in the case of TDMA multiple access; see Chapter 6) with or without the use of error-correcting codes.
  — Use of a frequency band at a lower frequency that is less affected by the attenuation.
  — Use of higher EIRP on the uplink.

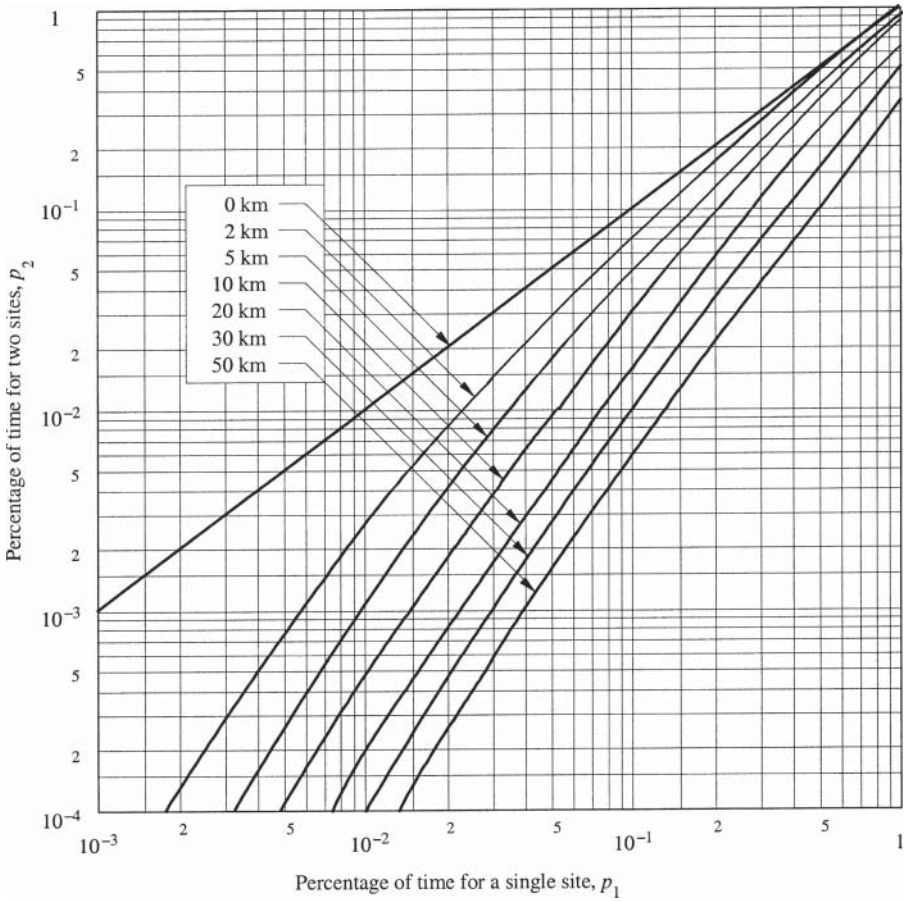


Figure 5.31 Relationship between percentage of time with and without diversity for the same attenuation (earth satellite paths).

— Reduction of capacity. In the case of digital transmission, using forward error-correction coding within the imposed bandwidth reduces the required value of C/N_0, at the cost of a reduced information bit rate R_b (see Section 4.2.7.2). The reduction in the required value of C/N_0 provides a margin equal to the C/N_0 reduction. This can be used for an overall link via a transparent satellite (see Section 5.9) or for the uplink or downlink of a regenerative satellite (see Section 5.10).

5.8.5 Cost-availability trade-off

A low unavailability (0.01% of the time, for example) corresponds to a high availability (99.99%, for the example considered). If only the effects of the propagating medium are considered as the cause of unavailability, the accepted value of unavailability represents the percentage of time p during which a given attenuation can be exceeded. When p is small (that is, the required availability is high), the value of this attenuation is high. Since the methods used to compensate

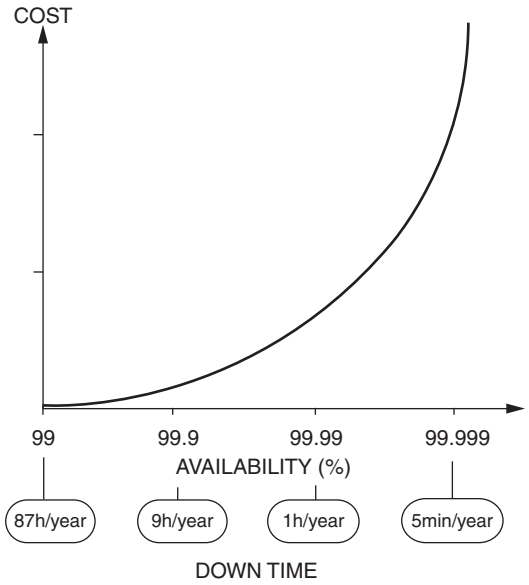


Figure 5.32 Link cost as a function of availability.

for attenuation become more costly as the attenuation increases, the specified availability has a marked effect on system cost. A typical trend is shown in Figure 5.32.

5.9 OVERALL LINK PERFORMANCE WITH TRANSPARENT SATELLITE

Section 5.6 presents the individual link performance in terms of C/N_0. This section discusses the expression for the overall station-to-station link performance: that is, the link involving one uplink and one downlink via a transparent satellite (no on-board demodulation and remodulation). Up to now, noise on the uplink and on the downlink has been considered to be thermal noise only. In practice, one has to account for interference noise originating from other carriers in the considered frequency bands and intermodulation noise resulting from multicarrier operation of nonlinear amplifiers. The overall link performance is discussed (Section 5.9.2) without intermodulation or interference, and then expressions are introduced considering interference and finally intermodulation.

The following notation is used:

— $(C/N_0)_U$ is the uplink carrier power-to-noise power spectral density ratio (Hz) at the satellite receiver input, considering no other noise contribution than the uplink system thermal noise temperature T_U.
— $(C/N_0)_D$ is the downlink carrier power-to-noise power spectral density ratio (Hz) at the earth station receiver input, considering no other noise contribution than the downlink system thermal noise temperature T_D.
— $(C/N_0)_I$ is the carrier power-to-interference noise power spectral density ratio (Hz) at the input of the considered receiver.

— $(C/N_0)_{IM}$ is the carrier power-to-intermodulation noise power spectral density ratio (Hz) at the output of the considered nonlinear amplifier.
— $(C/N_0)_T$ is the overall carrier power-to-noise power spectral density ratio (Hz) at the earth station receiver input.

5.9.1 Characteristics of the satellite channel

Figure 5.33 shows a *transparent* payload, where carriers are power amplified and frequency downconverted. Due to technology power limitations, the overall bandwidth is split into several sub-bands, the carriers in each sub-band being amplified by a dedicated power amplifier. The amplifying chain associated with each sub-band is called a *satellite channel*, or transponder. The satellite channel amplifies one or several carriers. Here is some more notation:

— C_U is the considered carrier power at the satellite receiver input; at saturation, it is denoted (CU)sat.
— P_{in} is the power at the input to the satellite channel amplifier (i = input, n = number of carriers in the channel).
— P_{on} is the power at the output of the satellite channel amplifier (o = output, n = number of carriers in the channel).
— $n = 1$ corresponds to a single-carrier operation of the satellite channel.
— $(P_{i1})_{sat}$ is the power at the input to the satellite channel amplifier at saturation in single-carrier operation.

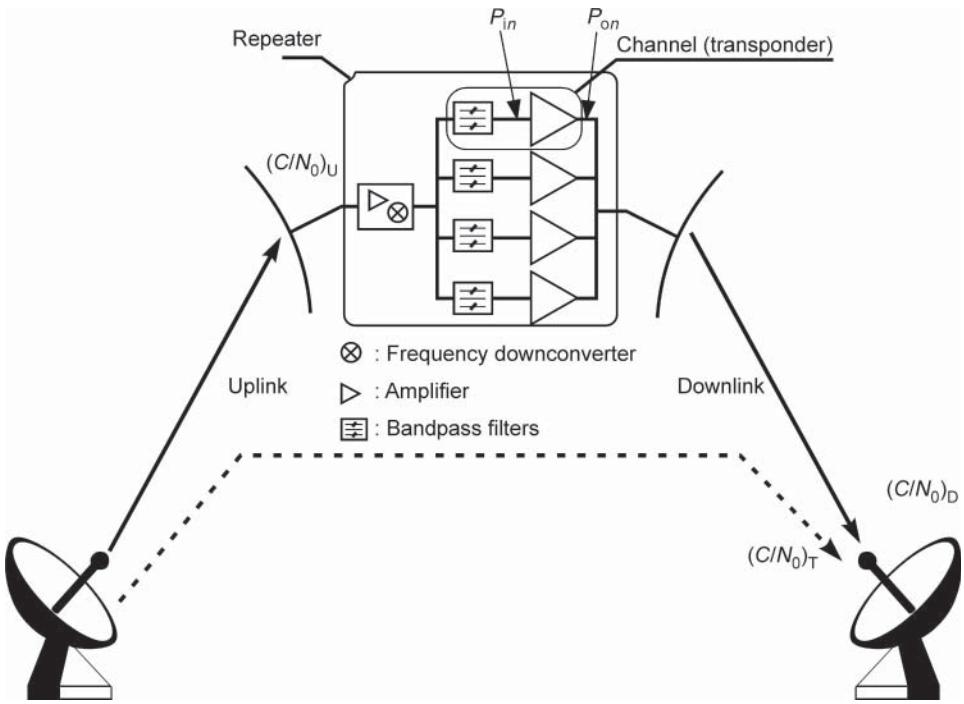


Figure 5.33 Overall station-to-station link for a transparent satellite.

— $(P_{o1})_{sat}$ is the power at the output of the satellite channel amplifier at saturation in single-carrier operation.

Saturation refers to the operation of the amplifier at maximum output power in single-carrier operation (Section 9.2.1.2). The satellite operator provides characteristic values of a satellite channel in terms of flux density at saturation, Φ_{sat}, and EIRP at saturation, $EIRP_{sat}$.

5.9.1.1 Satellite power flux density at saturation

Power flux density was introduced in Section 5.3.2. This flux is provided by the transmit earth station and is considered at the satellite receive antenna (Example 5.1). Its nominal value to drive the satellite channel amplifier at saturation is given by:

$$\Phi_{sat,\ nom} = \frac{(P_{il})_{sat}}{G_{FE}} \frac{L_{FRX}}{G_{Rmax}} \frac{4\pi}{\lambda_U^2} \quad (W/m^2) \tag{5.58}$$

where G_{FE} is the front-end gain from the input to the satellite receiver to the input to the satellite channel amplifier; L_{FRX} is the loss from the output of the satellite receive antenna to the input of the satellite receiver; and G_{Rmax} is the satellite receive antenna maximum gain (at boresight). The formula assumes that the transmit earth station is located at the centre of the satellite receive coverage (satellite antenna boresight).

In practice, the flux to be provided from a given earth station to drive the satellite channel amplifier to saturation depends on the location of the transmit earth station within the satellite coverage and the polarisation mismatch of the satellite receiving antenna with respect to the uplink carrier polarisation. Assuming that the receive satellite antenna gain in the direction of the transmit earth station experiences a gain fallout L_R with respect to the maximum gain, and a polarisation mismatch loss L_{POL}, the actual flux density to be provided by the transmit earth station is larger than or equal to:

$$\Phi_{sat} = \Phi_{sat,nom} L_R L_{POL} = \frac{(P_{i1})_{sat}}{G_{FE}} \frac{L_{FRX}}{G_{Rmax}} \frac{4\pi}{\lambda_U^2} L_R L_{POL} \ \left(W/m^2\right)$$

5.9.1.2 Satellite EIRP at saturation

EIRP was introduced in Section 5.3.1. The satellite EIRP at saturation and boresight $EIRP_{sat,max}$ relates to the satellite channel amplifier output power at saturation, $(P_{o1})_{sat}$, as follows:

$$EIRP_{sat,max} = \frac{(P_{01})_{sat}}{L_{FTX}} G_{Tmax} \ (W) \tag{5.59}$$

where L_{FTx} is the loss from the output of the power amplifier to the transmit antenna, and G_{Tmax} is the satellite transmit antenna maximum gain (at boresight).

In practice, the satellite $EIRP_{sat}$ that conditions the available carrier power at a given earth station receiver input is reduced by the transmit satellite antenna gain fallout L_T (the gain fallout is defined in the direction of the receiving earth station, with respect to the maximum gain) when the earth station is not located at the centre of transmit coverage (satellite antenna boresight):

$$EIRP_{sat} = \frac{EIRP_{sat,max}}{L_T} = \frac{(P_{01})_{sat}}{L_{FTX}} \frac{G_{Tmax}}{L_T} = \frac{(P_{01})_{sat}}{L_{FTX}} G_T \ (W) \tag{5.60}$$

5.9.1.3 Satellite repeater gain

The satellite repeater gain, G_{SR}, is the power gain from the satellite receiver input to the satellite channel amplifier output. At saturation, it is called $G_{SR\,sat}$.

$$G_{SR} = G_{FE}G_{CA} \tag{5.61}$$

where G_{Fe} is the front-end gain (from satellite receiver input to satellite channel amplifier input) and G_{CA} is the satellite channel amplifier gain.

5.9.1.4 Input and output back-off

In practice, the satellite channel power amplifier is not always operated at saturation, and it is convenient to determine the operating point Q of the satellite channel amplifier determined by the input power $(P_{in})_Q$ and the output power $(P_{on})_Q$. It is convenient to normalise these quantities with respect to $(P_{i1})_{sat}$ and $(P_{o1})_{sat}$, respectively. This defines the input back-off (IBO) and the output back-off (OBO):

$$\text{IBO} = (P_{in})_Q/(P_{i1})_{sat} \tag{5.62}$$

$$\text{OBO} = (P_{on})_Q/(P_{o1})_{sat} \tag{5.63}$$

From now on, the operating power value is denoted without the Q subscript.

5.9.1.5 Carrier power at satellite receiver input

The carrier power required at the satellite receiver input to drive the satellite channel amplifier to operate at the considered operating point Q is given by:

$$C_U = \frac{(P_{in})_Q}{G_{FE}} = \text{IBO}\frac{(P_{i1})_{sat}}{G_{FE}} \text{ (W)} \tag{5.64}$$

The carrier power can also be expressed as a function of the satellite channel amplifier output power:

$$C_U = \text{IBO}\frac{P_{on}}{G_{FE}G_{CA}} = \text{IBO}\frac{(P_{o1})_{sat}}{G_{FE}(G_{CA})_{sat}} \text{ (W)} \tag{5.65}$$

where $(G_{CA})_{sat}$ is the satellite amplifier gain *at saturation*. Finally, C_U can be expressed as:

$$C_U = \text{IBO}(C_U)_{sat} \text{ (W)} \tag{5.66}$$

where

$$(C_U)_{sat} = \frac{(P_{i1})_{sat}}{G_{FE}} = \frac{(P_{o1})_{sat}}{G_{FE}(G_{CA})_{sat}}$$

is the carrier power required at the satellite receiver input to drive the satellite channel amplifier at saturation. $(C_U)_{sat}$ can also be expressed as a function of Φ_{sat}:

$$(C_U)_{sat} = \Phi_{sat}\frac{G_{Rmax}}{L_{FRX}}\frac{\lambda_U^2}{4\pi}\frac{1}{L_R L_{POL}} \text{ (W)} \tag{5.67}$$

or

$$(C_U)_{sat} = \Phi_{sat,nom} \frac{G_{Rmax}}{L_{FRX}} \frac{\lambda_U^2}{4\pi}$$

Note that the IBO can also be expressed as the ratio of the power flux density Φ required to operate the satellite channel amplifier at the considered operating point to the satellite power flux density at saturation:

$$\text{IBO} = \frac{C_U}{(C_U)_{sat}} = \frac{\Phi}{\Phi_{sat}}$$

5.9.2 Expression for $(C/N_0)_T$

5.9.2.1 Expression for $(C/N_0)_T$ without interference from other systems or intermodulation

The power of the carrier received at the input of the earth station receiver is C_D. The noise at the input of the earth station receiver corresponds to the sum of the following:

— The downlink system noise considered in isolation ($T_D = T_2$, given by Eq. (5.32)) that defines the ratio C/N_0 for the downlink $(C/N_0)_D$ and can be calculated as in the example of Section 5.6.3 with $(N_0)_D = kT_D$.
— The uplink noise retransmitted by the satellite.

Hence:

$$(N_0)_T = (N_0)_D + G(N_0)_U \text{ (W/Hz)} \tag{5.68}$$

where $G = G_{SR}G_TG_R/L_{FTX}L_DL_{FRX}$ is the total power gain between the satellite receiver input and the earth station receiver input. G takes into account the satellite repeater gain G_{SR} from the input to the satellite receiver to the output of the satellite channel amplifier; the gain G_T/L_{FTX} of the satellite transmit antenna including the gain fallout and the loss L_{FTX} from the output of the power amplifier to the transmit antenna; the downlink path loss L_D; and the receiving station composite gain G_T/L_{FRX}. This gives

$$\begin{aligned}(C/N_0)_T^{-1} &= (N_0)_T/C_D \\ &= [(N_0)_D + G(N_0)_U]/C_D = (N_0)_D/C_D + (N_0)_U/G^{-1}C_D(\text{Hz}^{-1})\end{aligned} \tag{5.69}$$

In this expression, the term $G^{-1} = C_D$ represents the carrier power at the satellite receiver input. Hence $(N_0)_U/G^{-1}C_D = (C/N_0)_U^{-1}$. Finally:

$$(C/N_0)_T^{-1} = (C/N_0)_U^{-1} + (C/N_0)_D^{-1} \quad (\text{Hz}^{-1}) \tag{5.70}$$

In this expression:

$$\begin{aligned}(C/N_0)_U &= (P_{i1})/(N_0)_U = \text{IBO}(P_{i1})_{sat}/(N_0)_U \\ &= \text{IBO}(P_{o1})_{sat}/G_{SR\ sat}(N_0)_U \\ &= \text{IBO}(C/N_0)_{U\ sat} \quad (\text{Hz}) \\ (C/N_0)_D &= \text{OBO}(\text{EIRP}_{sat})_{SL}(1/L_D)(G/T)_{ES}(1/k) \\ &= \text{OBO}(C/N_0)_{D\ sat} \quad (\text{Hz})\end{aligned}$$

$(C/N_0)_{\text{U sat}}$ and $(C/N_0)_{\text{D sat}}$ are the values of C/N_0 for the uplink and downlink when the satellite channel operates at saturation. L_{D} represents the attenuation on the downlink and is given by Eq. (5.14) and $(G/T)_{\text{ES}}$, the figure of merit of the earth station in the satellite direction.

5.9.2.2 *Expression for $(C/N_0)_T$ taking interference into account*

Interference is the unwanted power contribution of other carriers in the frequency band occupied by the wanted carrier. A given link may suffer interference from other satellite links or from terrestrial systems operating in the same frequency bands as the considered link. Indeed, most of the frequency bands allocated to space radiocommunications are also allocated on a shared basis to terrestrial radiocommunications. To facilitate this sharing, a number of provisions have been introduced into the Radio Regulations, Articles *S* 21 and *S* 9, and a coordination procedure has been instituted between earth and terrestrial stations (Radio Regulations, Article 11).

Four types of interference between systems can be distinguished:

— A satellite interfering with a terrestrial station
— A terrestrial station interfering with a satellite
— An earth station interfering with a terrestrial station
— A terrestrial station interfering with an earth station

Figure 5.34 illustrates the geometry associated with these forms of interference.

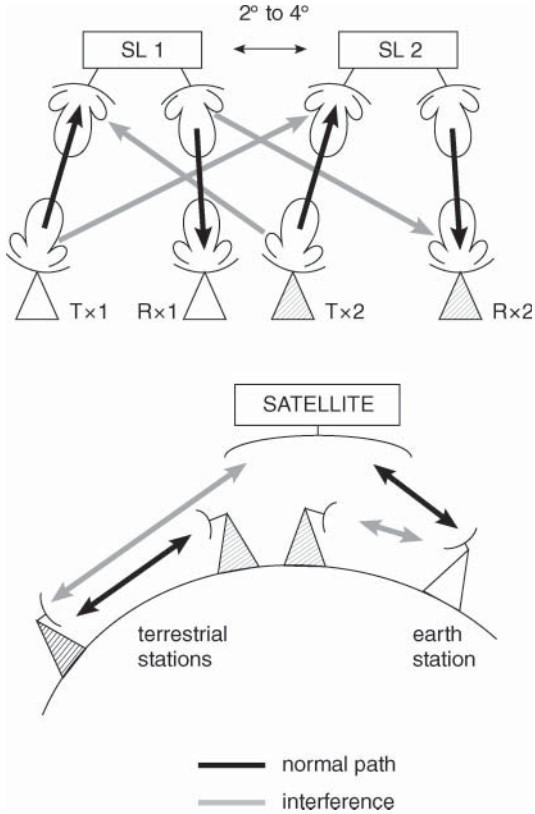


Figure 5.34 The geometry of interference between systems.

The carriers emitted by other systems are superimposed on the wanted carrier of the station-to-station link at two levels:

— At the input of the satellite repeater on the uplink
— At the input of the earth station receiver on the downlink

The effect of interference is similar to an increase of the thermal noise on the link affected by interference. It is allowed for in the equations in the form of an increase of the spectral density:

$$N_0 = (N_0)_{\text{without interference}} + (N_0)_I \quad \text{(W/Hz)} \tag{5.71}$$

where $(N_0)_I$ represents the increase of the noise power spectral density due to interference. A ratio $(C/N_0)_I$ that expresses the signal power in relation to the spectral density of the interference can be associated with $(N_0)_I$; these are $(C/N_0)_{I,U}$ for the uplink and $(C/N_0)_{I,D}$ for the downlink. This leads to modification of Eq. (5.70), replacing $(C/N_0)_U$ and $(C/N_0)_D$ with the following expressions:

$$\begin{aligned}(C/N_0)_U^{-1} &= [(C/N_0)_U^{-1}]_{\text{without interference}} + (C/N_0)_{I,U}^{-1} \quad (\text{Hz}^{-1}) \\ (C/N_0)_D^{-1} &= [(C/N_0)_D^{-1}]_{\text{without interference}} + (C/N_0)_{I,D}^{-1} \quad (\text{Hz}^{-1})\end{aligned} \tag{5.72}$$

The total expression becomes:

$$(C/N_0)_T^{-1} = (C/N_0)_U^{-1} + (C/N_0)_D^{-1} + (C/N_0)_I^{-1} \quad (\text{Hz}^{-1}) \tag{5.73}$$

where $(C/N_0)_U$ and $(C/N_0)_D$ are the values appearing in Eq. (5.70) and:

$$(C/N_0)_I^{-1} = (C/N_0)_{I,U}^{-1} + (C/N_0)_{I,D}^{-1} \quad (\text{Hz}^{-1}) \tag{5.74}$$

5.9.2.3 Expression for $(C/N_0)_T$ taking intermodulation and interference into account

When several carriers are amplified in a nonlinear amplifier, the output is not only the amplified carriers but also the intermodulation products, which appear as power at frequencies that are linear combinations of the input carrier frequencies (Section 6.5.4). Some of these intermodulation products fall in the bandwidth of the considered carrier and act as noise with spectral density $(N_0)_{IM}$. The ratio of the carrier power to the intermodulation noise spectral density is $(C/N_0)_{IM}$.

Intermodulation noise is added to the other sources of noise analysed in this chapter. Eq. (5.74) for the carrier power-to-noise power spectral density ratio for the overall station-to-station link $(C/N_0)_T$ is modified as follows:

$$(C/N_0)_T^{-1} = (C/N_0)_U^{-1} + (C/N_0)_D^{-1} + (C/N_0)_I^{-1} + (C/N_0)_{IM}^{-1} \quad (\text{Hz}^{-1}) \tag{5.75}$$

with:

$$(C/N_0)_{IM}^{-1} = (C/N_0)_{IM,U}^{-1} + (C/N_0)_{IM,D}^{-1}$$

where $(C/N_0)_{IM,U}^{-1}$ and $(C/N_0)_{IM,D}$ correspond to the generation of intermodulation noise in the transmitting earth station and the satellite repeater channel, respectively.

In this case, the expressions for the ratios $(C/N_0)_U$, $(C/N_0)_D$, and $(C/N_0)_{IM}$ are to be used with values of input and IBO and OBO for operation of the amplifier in multicarrier mode with carriers of *equal power*. The output power of the amplifier is shared among the carriers, the thermal noise and the intermodulation noise to which the interference noise for the channel is added.

Denoting by P_{in} and P_{on}, respectively, the input and output and power of one carrier among the n amplified ones, IBO and OBO are defined as follows:

— IBO *per carrier*:
  — IBO_1 = single carrier input power/single carrier input power at saturation = $P_{i1}/(P_{i1})_{sat}$ or, in dB:
  — IBO_1 (dB) = 10 log $\{P_{i1}/(P_{i1})_{sat}\}$
— OBO *per carrier:*
  — OBO_1 = single carrier output power/single carrier output power at saturation = $P_{o1}/(P_{o1})_{sat}$ or, in dB:
  — IBO_1 (dB) = 10 log$\{P_{o1}/(P_{o1})_{sat}\}$
— Total IBO:
  — IBO_t = sum of all input carrier power/single carrier input power at saturation = $\Sigma P_{in}/(P_{i1})_{sat}$ or, in dB:
  — IBO_t (dB) = 10 log$\{\Sigma P_{in}/(P_{i1})_{sat}\}$
— Total OBO:
  — OBO_t = sum of all output carrier power/single carrier output power at saturation = $\Sigma P_{on}/(P_{o1})_{sat}$
  — or, in dB:
  — OBO_t (dB) = 10 log$\{\Sigma P_{on}/(P_{o1})_{sat}\}$

With n equally powered carriers:

— $IBO_1 = IBO_t/n$ or, in dB, IBO_1 (dB) = IBO_t (dB) – 10 log n
— $OBO_1 = OBO_t/n$ or, in dB, OBO_1 (dB) = OBO_t (dB) – 10 log n

If the carriers at the amplifier input are of *unequal power*, the power at the amplifier output is shared unequally between carriers and noise. Therefore the amplifier does not have equal power gain for all carriers and a *capture effect* can arise: carriers of high power acquire more power than carriers of low power. For carriers of high power, the value of the ratio is greater than that given by Eq. (5.75). For carriers of low power, it is smaller. Generation of intermodulation products is also observed between noise on the uplink and the carriers; this effect can be taken into account in the form of an increase in the noise temperature at the channel input.

5.9.2.4 Influence of back-off

Figure 5.35 shows the variation of each of the terms in Eq. (5.75) as a function of IBO assuming the equivalent interference noise to be negligible. Because of the opposite direction of variation of the term $(C/N_0)_{IM}$ compared to that of the ratios $(C/N_0)_U$ and $(C/N_0)_D$, the value of $(C/N_0)_T$ passes through a maximum for a nonzero value of back-off. Two effects are, therefore, observed which are consequences of using the same repeater channel to amplify several carriers:

— The total power at the output of the channel is less than that which would exist in the absence of back-off.
— The useful power per carrier is reduced by allocation of part of the total power to intermodulation products.

5.9.3 Overall link performance for a transparent satellite without interference or intermodulation

It is required to establish a satellite link between two earth stations (Figure 5.33), assumed to be located at the centre of the satellite antenna's coverage. The data are as follows:

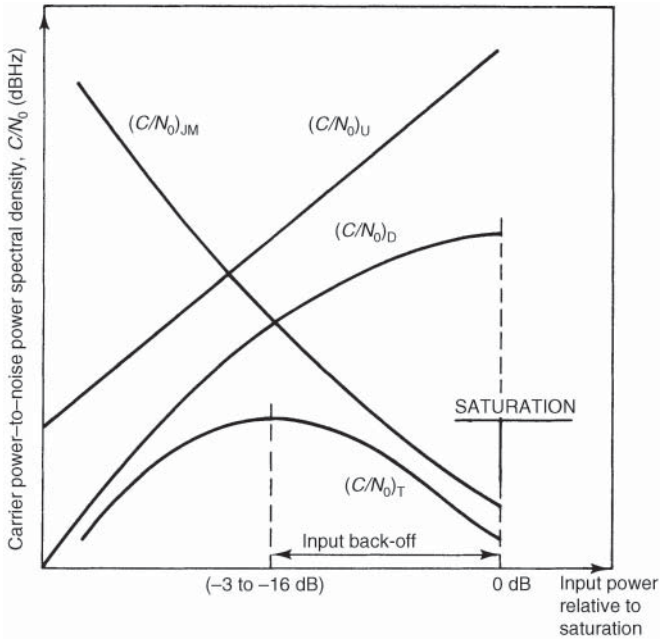


Figure 5.35 Variation of $(C/N_0)_U$, $(C/N_0)_D$, $(C/N_0)_{IM}$, and $(C/N_0)_T$ as a function of input back-off (IBO).

— Uplink frequency: $f_U = 14\,\text{GHz}$
— Downlink frequency: $f_D = 12\,\text{GHz}$
— Downlink path loss: $L_D = 206\,\text{dB}$
— For the satellite (SL):
  — Power flux density required to saturate the satellite channel amplifier:

$$(\Phi_{sat,nom})_{SL} = -90\ \text{dBW/m}^2$$

— Satellite receiving antenna gain at boresight: $G_{Rmax} = 30\,\text{dBi}$
  — Satellite figure of merit at boresight: $(G/T)_{SL} = 3.4\,\text{dBK}^{-1}$
  — Satellite channel amplifier characteristic (single carrier operation) modelled by:

$$\text{OBO(dB)} = \text{IBO(dB)} + 6 - 6\ \exp[\text{IBO(dB)}/6]$$

  — Satellite EIRP at saturation in the direction of the considered receiving earth station (i.e. at boresight of the satellite transmitting antenna)

$$(\text{EIRP}_{sat})_{SL} = 50\,\text{dBW}$$

  — Satellite transmitting antenna gain at boresight: $G_{Tmax} = 40\,\text{dBi}$

The following losses are considered:

— Satellite reception and transmission feeder losses: $L_{FRX} = L_{FTX} = 0\,\text{dB}$
— Satellite antenna polarisation mismatch loss $L_{pol} = 0\,\text{dB}$
— Satellite antenna depointing losses: $L_R = L_T = 0\,\text{dB}$ (earth stations at boresight)

— For the earth station (ES): Figure of merit of earth station in satellite direction $(G/T)_{ES} = 25\,\text{dBK}^{-1}$

It is assumed that there is no interference.

5.9.3.1 Satellite repeater gain at saturation GSR sat

$G_{SR\,sat} = (P_{01})_{sat}/(C_U)_{sat}$ where $(C_U)_{sat}$ is the carrier power required at the satellite receiver input to drive the satellite channel amplifier at saturation. From Eq. (5.60):

$$(P_{o1})_{sat} = (\text{EIRP}_{sat})_{SL} L_T L_{FTX}/G_{Tmax} \ \text{(W)}$$

Hence:

$$(P_{o1})_{sat} = 50\,\text{dBW} - 40\,\text{dBi} = 10\,\text{dBW} = 10\,\text{W}$$

From Eq. (5.67):

$$(C_U)_{sat} = (\Phi_{sat})_{SL} G_{Rmax}/L_{FRX} L_R L_{POL}(4\pi/\lambda_U^2) \ \text{(W)}$$

hence:

$$(C_U)_{sat} = -90\,\text{dBW/m}^2 + 30\,\text{dBi} - 44.4\,\text{dBm}^2 = -104.4\,\text{dBW} = 36\,\text{pW}$$

$$G_{SR\,sat} = (P_{o1})_{sat}/(C_U)_{sat} = 10\,\text{dBW} - (-104.4\,\text{dBW}) = 114.4\,\text{dB}$$

5.9.3.2 Calculation of C/N_0 for the up- and downlinks and the overall link when the repeater operates at saturation

$$(C/N_0)_{U\,sat} = (C_U)_{sat}/kT_U = (C_U)_{sat}(G/T)_{SL}/(kG_{Rmax}/L_R L_{FRX} L_{POL})$$

$$(C/N_0)_{U\,sat} = -104.4 + 3.4 - (-228.6) - 30 = 97.6\,\text{dBHz}$$

$$(C/N_0)_{D\,sat} = (\text{EIRP}_{sat})_{SL}(1/L_D)(G/T)_{ES}(1/k) \ \text{(Hz)}$$

$$(C/N_0)_{D\,sat} = 50 - 206 + 25 - (-228.6) = 97.6\,\text{dBHz}$$

$$(C/N_0)_{T\,sat}^{-1} = (C/N_0)_{U\,sat}^{-1} + (C/N_0)_{D\,sat}^{-1} \ (\text{Hz}^{-1})$$

$$(C/N_0)_{T\,sat} = 94.6\,\text{dBHz}$$

5.9.3.3 Calculation of the input and output back-off to achieve $(C/N_0)_T = 80$ dBHz and the corresponding values of $(C/N_0)_U$ and $(C/N_0)_D$

One must have:

$$(C/N_0)_U^{-1} + (C/N_0)_D^{-1} = 10^{-8}\text{Hz}^{-1}$$

Hence:

$$\text{IBO}^{-1}(C/N_0)_{U\,sat}^{-1} + \text{OBO}^{-1}(C/N_0)_{D\,sat}^{-1} = 10^{-8}\text{Hz}^{-1}$$

This gives:

$$10^{-\text{IBO(dB)}/10} + 10^{-\text{OBO(dB)}/10} = 10^{1.76}$$

with:

$$\mathrm{OBO(dB)} = \mathrm{IBO(dB)} + 6 - 6\exp(\mathrm{IBO(dB)}/6)$$

Numerical solution gives:

$$\mathrm{IBO} = -16.4\ \mathrm{dB}$$

$$\mathrm{OBO} = -10.8\ \mathrm{dB}$$

Hence:

$$(C/N_0)_{\mathrm{U}} = \mathrm{IBO}(C/N_0)_{\mathrm{U\ sat}} = -16.4\ \mathrm{dB} + 97.6\ \mathrm{dBHz} = 81.2\ \mathrm{dBHz}$$

$$(C/N_0)_{\mathrm{D}} = \mathrm{OBO}(C/N_0)_{\mathrm{D\ sat}} = -10.8\ \mathrm{dB} + 97.6\ \mathrm{dBHz} = 86.8\ \mathrm{dBHz}$$

5.9.3.4 Value of $(C/N_0)_T$ under rain conditions causing an attenuation of 6 dB on the uplink

The attenuation of 6 dB on the uplink reduces the IBO by 6 dB. The new value of IBO becomes:

$$\mathrm{IBO(dB)} = -16.4\ \mathrm{dB} - 6\ \mathrm{dB} = -22.4\ \mathrm{dB}$$

The new value of OBO corresponding to this is:

$$\mathrm{OBO(dB)} = \mathrm{IBO(dB)} + 6 - 6\exp(\mathrm{IBO(dB)}/6) = -16.5\ \mathrm{dB}$$

Hence:

$$(C/N_0)_{\mathrm{U}} = \mathrm{IBO}(C/N_0)_{\mathrm{U\ sat}} = -22.4\ \mathrm{dB} + 97.6\ \mathrm{dBHz} = 75.2\ \mathrm{dBHz}$$

$$(C/N_0)_{\mathrm{D}} = \mathrm{OBO}(C/N_0)_{\mathrm{D\ sat}} - 16.5\ \mathrm{dB} + 97.6\ \mathrm{dBHz} = 81.1\ \mathrm{dBHz}$$

and, from Eq. (5.70):

$$(C/N_0)_{\mathrm{T}} = 74.2\ \mathrm{dBHz}$$

To regain the required value $(C/N_0)_{\mathrm{T}} = 80\ \mathrm{dBHz}$, it is necessary to increase the $(\mathrm{EIRP})_{\mathrm{ES}}$ of the transmitting earth station by 6 dB.

5.9.3.5 Value of $(C/N_0)_T$ under rain conditions causing an attenuation of 6 dB on the downlink with a reduction of 2 dB in the figure of merit of the earth station due to the increase of antenna noise temperature

The value of $(C/N_0)_{\mathrm{D}}$ reduces by 8 dB, hence: $(C/N_0)_{\mathrm{D}} = 86.8\ \mathrm{dBHz}–8\ \mathrm{dB} = 78.8\ \mathrm{dBHz}$. From which: $(C/N_0)_{\mathrm{T}} = 76.8\ \mathrm{dBHz}$.

To regain the required value $(C/N_0)_{\mathrm{T}} = 80\ \mathrm{dBHz}$, it is necessary to increase the $(\mathrm{EIRP})_{\mathrm{ES}}$ of the transmitting earth station in such a way that the value of IBO satisfies the equation:

$$\mathrm{IBO}^{-1}(C/N_0)^{-1}_{\mathrm{U\ sat}} + \mathrm{OBO}^{-1}(C/N_0)^{-1}_{\mathrm{D\ sat}} = 10^{-8}\mathrm{Hz}^{-1}$$

in which:

$$(C/N_0)_{\mathrm{U\ sat}} = 97.6\ \mathrm{dBHz}$$

$$(C/N_0)_{\mathrm{D\ sat}} = 97.6\ \mathrm{dBHz} - 8\ \mathrm{dB} = 89.6\ \mathrm{dBHz}$$

This gives:

$$\text{IBO} = -13\ \text{dB}$$

$$\text{OBO} = -7.7\ \text{dB}$$

It is necessary to increase the $(\text{EIRP})_{\text{ES}}$ of the earth station transmission by $-13\ \text{dB} - (-16.4\ \text{dB}) = 3.4\ \text{dB}$.

Hence:

$$(C/N_0)_{\text{U}} = \text{IBO}(C/N_0)_{\text{U sat}} = -13\ \text{dB} + 97.6\ \text{dBHz} = 84.6\ \text{dBHz}$$

$$(C/N_0)_{\text{D}} = \text{OBO}(C/N_0)_{\text{D sat}} = -7.7\ \text{dB} + 89.6\ \text{dBHz} = 81.9\ \text{dBHz}$$

5.10 OVERALL LINK PERFORMANCE WITH REGENERATIVE SATELLITE

Figure 5.36 shows the difference between a regenerative satellite repeater and a transparent one. With the regenerative repeater, baseband signals, which have modulated the uplink carrier, are available at the output of the demodulator, and these signals are used (possibly after processing not shown in the figure) to modulate the downlink carrier. Hence the change of frequency from uplink to downlink that is obtained by mixing with the local radio-frequency oscillator in a transparent satellite is obtained in this case by modulation of a new carrier.

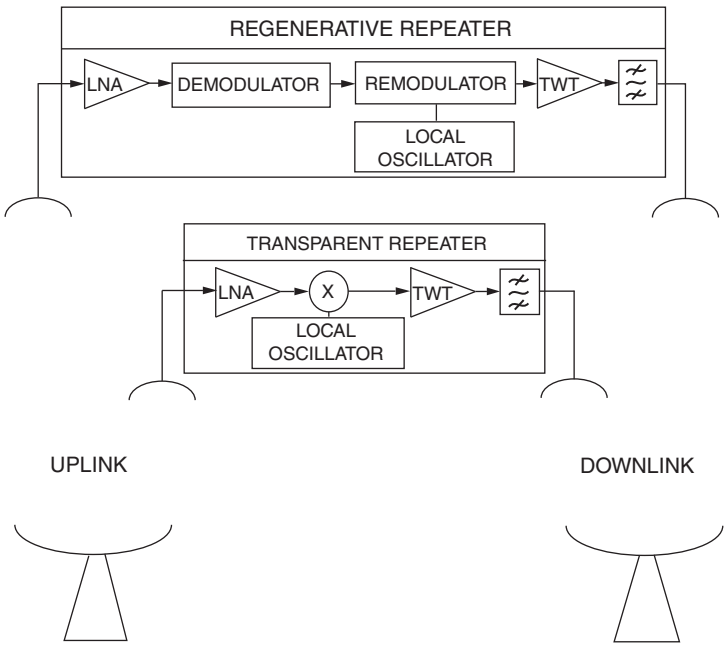


Figure 5.36 Organisation of a regenerative repeater and a transparent repeater.

5.10.1 Linear satellite channel without interference

It is assumed that the probability of error at the output of the demodulator is that given by theory (Table 4.4); that is, there is no degradation due to filtering or nonlinearities.

5.10.1.1 Link with transparent repeater

The performance of the link (Figure 5.37) is specified in terms of the bit error probability (BEP) at the output of the earth station demodulator. BEP is a function of the ratio $(E/N_0)_T$ given by Eq. (4.9) in Section 4.2.6.2 and recalled here:

$$(E/N_0)_T = (C/N_0)_T/R_c \tag{5.76}$$

where R_c is the carrier data rate and $(C/N_0)_T$ is the ratio of carrier power to noise spectral density of the station-to-station link given by Eq. (5.70) and recalled here:

$$(C/N_0)_T^{-1} = (C/N_0)_U^{-1} + (C/N_0)_D^{-1} \tag{5.77}$$

Defining $(E/N_0)_U = (C/N_0)_U/R_c$ and $(E/N_0)_D = (C/N_0)_D/R_c$ and using Eqs. (5.76) and (5.77) gives:

$$(E/N_0)_T^{-1} = (E/N_0)_U^{-1} + (E/N_0)_D^{-1} \tag{5.78}$$

5.10.1.2 Link with regenerative repeater

The performance of the link (Figure 5.38), specified in terms of the BEP, is expressed as the probability of having an error on the uplink (measured by BEP_U) and no error on the downlink ($1 - BEP_D$) or no error on the uplink ($1 - BEP_U$) and an error on the downlink (BEP_D), hence:

$$BEP = BEP_U(1 - BEP_D) + (1 - BEP_D)BEP_D \tag{5.79}$$

As BEP_U and BEP_D are small compared with 1, this becomes:

$$BEP = BEP_U + BEP_D \tag{5.80}$$

BEP_U is a function of $(E/N_0)_U$, and BEP_D is a function of $(E/N_0)_D$.

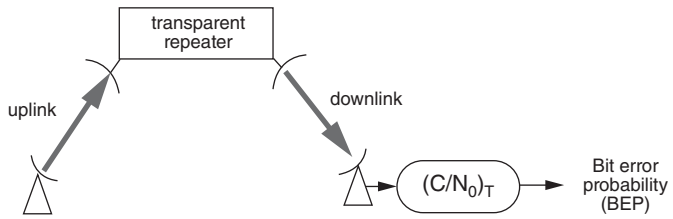


Figure 5.37 Link by transparent repeater.

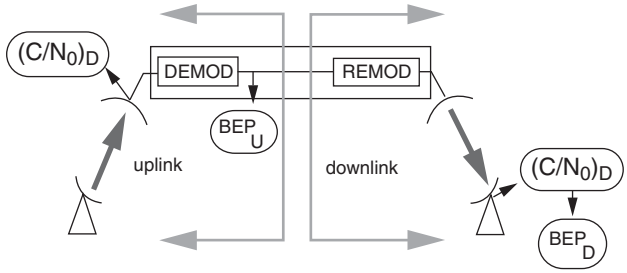


Figure 5.38 Link by regenerative repeater.

5.10.1.3 Comparison at constant bit error probability (BEP)

The value of the BEP is given as follows:

— For a transparent repeater, the value of $(E/N_0)_T$ is determined by the BEP specified for the link. The required performance is obtained for a set of values of $(E/N_0)_U$ and $(E/N_0)_D$ combined using Eq. (5.78). This is shown by curve A in Figure 5.39 for an error probability of 10^{-4} and quadrature phase shift keying (QPSK) modulation with coherent demodulation.
— For a regenerative repeater, by combining BEP_U and BEP_D from Eq. (5.80) with the constraint BEP = constant = 10^{-4} and deducing the corresponding pairs of values of $(E/N_0)_U$ and $(E/N_0)_D$, curves B and *C* in Figure 5.39 are obtained. These curves correspond to QPSK modulation with coherent demodulation (curve B) and differential demodulation (curve C) on the uplink. On the downlink, demodulation is coherent in both cases. The ratio $\alpha = (E/N_0)_U/(E/N_0)_D$ is used as a parameter.

Comparing curve A and curve B, it can be seen that the regenerative repeater provides a reduction of 3 dB in the value required for E/N_0 for the uplink and the downlink when the links are identical ($\alpha = 0$ dB). This is explained by the fact that the regenerative repeater does not transmit the amplified uplink noise along with the signal on the downlink, unlike a transparent repeater.

However, for very different values of E/N_0, this advantage disappears. For example, for a value greater than 12 dB, the two curves join; in this case, the uplink noise is negligible and the performance of the overall link reduces in both cases to that of the downlink.

Curve C indicates that by dimensioning the uplink for a value of $(E/N_0)_U$ greater than that of $(E/N_0)_D$ by about 4 dB, on-board differential demodulation can be used without degrading the global performance; this is simpler to realise than coherent demodulation.

5.10.2 Nonlinear satellite channel without interference

This corresponds more closely to a real system, since a real channel is nonlinear and band limited; the combination of nonlinearities and filtering introduces a performance degradation of the demodulator that increases as the chain of nonlinearities and filters increases. This is the case for a link with a transparent repeater (two nonlinearities and filters – on transmission at the earth station and at the transponder). With a regenerative repeater, separation of the up- and downlinks means there is now only one nonlinearity and filter per link. Figure 5.40 shows the results obtained by means of computer simulation [WAC-81] for the case where $(E/N_0)_U$ is at least 12 dB greater than $(E/N_0)_D$. This figure and other results show that, contrary to the conclusions from

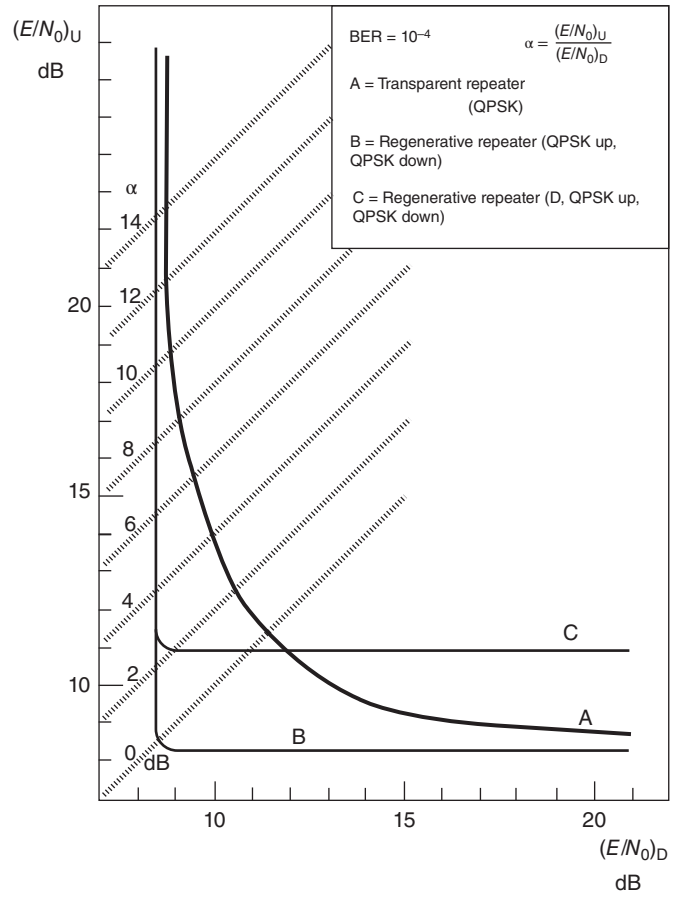


Figure 5.39 Comparison of station-to-station links by transparent repeater and regenerative repeater for the same bit error probability (BEP = 10^{-4}) (linear channel): (a) transparent repeater; (b) regenerative repeater with QPSK modulation and coherent demodulation on the uplink and the downlink; (c) regenerative repeater with QPSK modulation and differential demodulation on the uplink, and coherent demodulation on the downlink.

a linear analysis, the regenerative repeater can provide between 2 and 5 dB reduction in E/N_0 with respect to a transparent repeater even when the ratio $\alpha = (E/N_0)_U/(E/N_0)_D$ is large.

5.10.3 Nonlinear satellite channel with interference

5.10.3.1 Link with transparent repeater

The value of $(C/N_0)_T$ depends on the value of $(C/N_0)_{T\text{ without interference}}$ in the absence of interference and $(C/N_0)_I$ due to interference on the up- and downlinks (see Section 5.9.2). More precisely:

$$(C/N_0)_T^{-1} = (C/N_0)_{T\text{ without interference}}^{-1} + (C/N_0)_I^{-1} \tag{5.81}$$

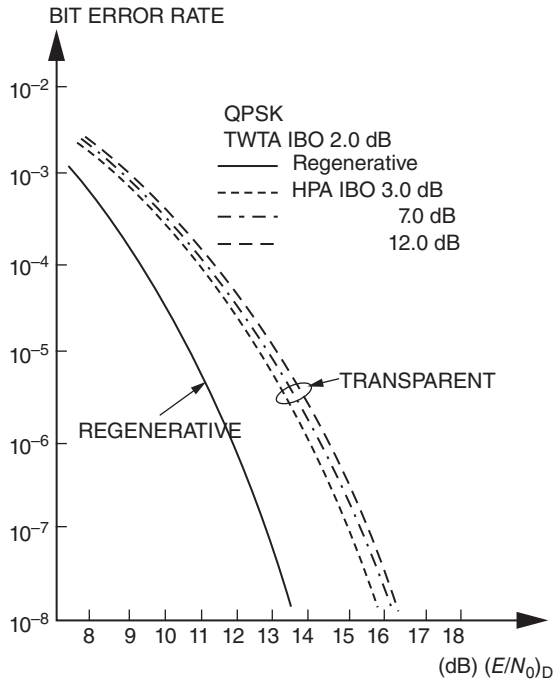


Figure 5.40 BEP as a function of $(E/N_0)_D$ for a link with a transparent repeater and for a link with a regenerative repeater in the absence of interference, for the case where $\alpha = (E/N_0)_U/(E/N_0)_D$ is high (larger than 12 dB). Travelling-wave tube amplifier (TWTA) IBO: input back-off of the on-board travelling wave tube; high-power amplifier (HPA) IBO: input back-off of the earth station transmitting amplifier. Source: [WAC-81] ©1981 IEEE. Reproduced with permission.

where:

$$(C/N_0)^{-1}_{\text{T without interference}} = (C/N_0)^{-1}_{\text{U}} + (C/N_0)^{-1}_{\text{D}}$$
$$(C/N_0)^{-1}_{\text{I}} = (C/N_0)^{-1}_{\text{I,U}} = +(C/N_0)^{-1}_{\text{I,D}}$$

By putting $E/N_0 = C/N_0/R_c$, the relations between the values of E/N_0 can be deduced from these equations:

$$(E/N_0)^{-1}_{\text{T}} = (E/N_0)^{-1}_{\text{T without interference}} + (E/N_0)^{-1}_{\text{I}} \tag{5.82}$$

From the curve in Figure 5.40 (which implies that α is large), BEP = 10^{-4} with QPSK modulation requires that $(E/N_0)_T = 11$ dB. The upper curve in Figure 5.41 shows the relation between $(E/N_0)_I$ and $(E/N_0)_{T\text{ without interference}}$ obtained from Eq. (5.82) for this case.

5.10.3.2 Link with regenerative repeater

Considering that $\alpha = (E/N_0)_U/(E/N_0)_D$ is high, the BEP of the station-to-station link is defined by the downlink BEP. A BEP of 10^{-4} requires, from Figure 5.40, $(E/N_0)_D = (E/N_0)_T = 9$ dB. The lower curve in Figure 5.41 shows the relation between $(E/N_0)_I$ and $(E/N_0)_{\text{T without interference}}$ for this case.

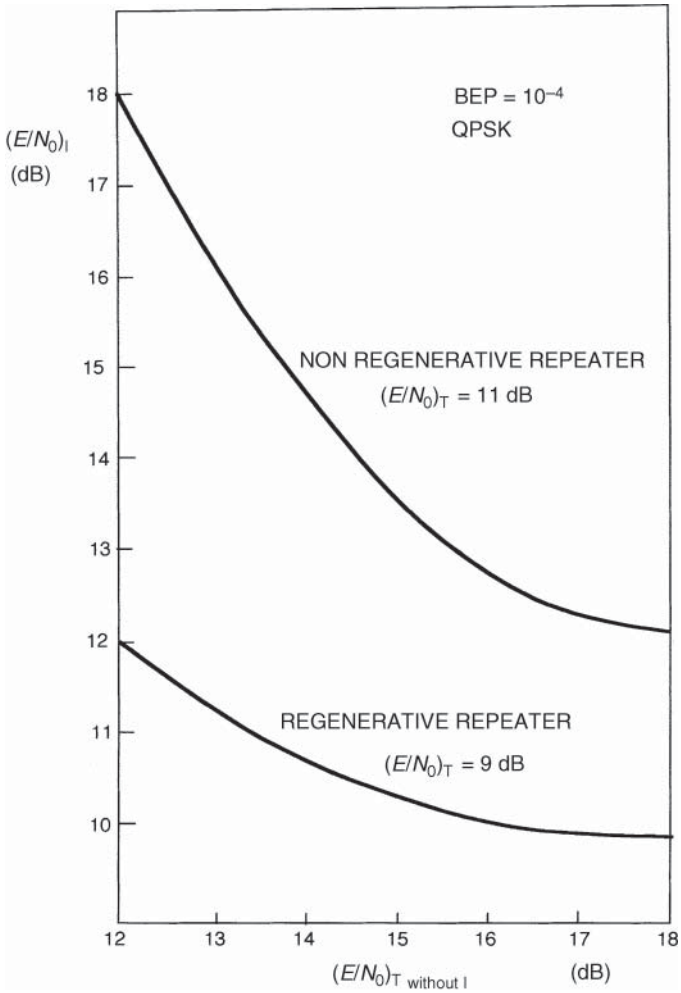


Figure 5.41 Permissible interference level; comparison of links with regenerative and transparent repeaters.

Comparison of the two curves in Figure 5.41 shows that, for a given link quality (BEP = 10^{-4}), the ratio $(E/N_0)_I$ is less for a link with a regenerative repeater. This implies that the required link performance is obtained in spite of a higher level of interference. This is a useful advantage in the context of multibeam satellites, which possibly face higher levels of interference compared to single-beam satellites (Section 5.11.2).

5.11 LINK PERFORMANCE WITH MULTIBEAM ANTENNA COVERAGE VS. MONOBEAM COVERAGE

From the previous sections, it can be noticed that the overall radio-frequency link quality depends on the gain of the satellite antenna. From Eq. (5.7), $G_{max} = 29\,000/(\theta_{3\,dB})^2$, it can be seen

that the satellite antenna gain is constrained by its beamwidth, whatever the frequency at which the link is operated. So the antenna gain is imposed by the angular width of the antenna beam covering the zone to be served (see the distinction between *coverage zone* and *service zone* in Sections 9.7 and 9.8). If the service zone is covered using a single antenna beam, this is referred to as *single-beam coverage*.

Single-beam antenna coverage displays one of these characteristics:

— The satellite may provide coverage of the whole region of the earth that is visible from the satellite (global coverage) and thus permit long-distance links to be established, for example from one continent to another. In this case, the gain of the satellite antenna is limited by its beamwidth as imposed by the coverage. For a geostationary satellite, global coverage implies a 3 dB beamwidth of 17.5° and consequently an antenna gain of no more than G_{max} (dBi) = 10 log (29000) – 20 log (17.5) = 20 dBi.
— The satellite may provide coverage of only part of the earth (a region or country) by means of a narrow beam (a zone or spot beam), with 3 dB beamwidth on the order of 1° to a few degrees. One thus benefits from a higher antenna gain due to the reduced antenna beamwidth, but the satellite cannot service earth stations situated outside this reduced coverage, and some of the earth stations that could be serviced with a global coverage are left out. These earth stations can be reached only by terrestrial links or by other satellites linked to the considered one by ISLs.

With single-beam antenna coverage, it is therefore necessary to choose between either extended coverage providing service with reduced quality to geographically dispersed earth stations, or reduced coverage providing service with improved quality to geographically concentrated earth stations.

Multibeam antenna coverage allows these two alternatives to be reconciled. Satellite extended coverage may be achieved by means of the juxtaposition of several narrow beam coverages, each beam providing an antenna gain that increases as the antenna beamwidth decreases (reduced coverage per beam). The link performance improves as the number of beams increases; the limit is determined by the antenna technology, whose complexity increases with the number of beams, and the mass. The complexity originates in the more elaborate satellite antenna technology (multibeam antennas; see Chapter 9) and the requirement to provide on-board interconnection of the coverage areas, so as to ensure within the satellite payload routing of the various carriers that are uplinked in different beams to any wanted destination beam (see Chapter 7).

5.11.1 Advantages of multibeam coverage

In Figure 5.42a, a satellite provides global coverage with a single beam of beamwidth $\theta_{3\,dB} = 17.5°$; in Figure 5.42b, the satellite supports spot beams with beamwidth $\theta_{3\,dB} = 1.75°$ with a consequently reduced coverage. In both cases, all earth stations in the satellite network are within the satellite coverage.

5.11.1.1 Impact on earth segment

The expression for $(C/N_0)_U$ for the uplink is given by (see Section 5.6.2):

$$(C/N_0)_U = (\text{EIRP})_{\text{station}}(1/L_U)(G/T)_{\text{satellite}}(1/k)\ \text{(Hz)} \tag{5.83}$$

Assuming that the noise temperature at the satellite receiver input is $T_{\text{satellite}} = 800\,\text{K} = 29\,\text{dBK}$ and is independent of the beam coverage (this is not rigorously true but satisfies a first

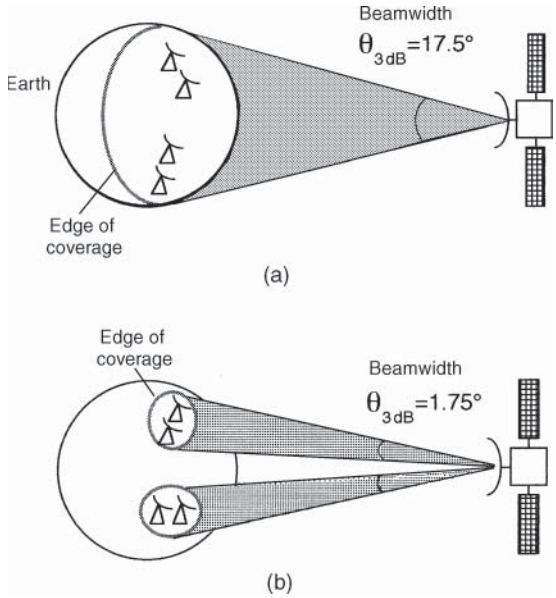


Figure 5.42 (a) Global coverage and (b) coverage by several narrow beams.

approximation), let $L_U = 200$ dB and neglect the implementation losses. Equation (5.83) becomes (all terms in dB):

$$\begin{aligned}(C/N_0)_U &= (\text{EIRP})_{\text{station}} - 200 + (G_R)_{\text{satellite}} - 29 + 228.6 \\ &= (\text{EIRP})_{\text{station}} + (G_R)_{\text{satellite}} - 0.4(\text{dBHz})\end{aligned} \tag{5.84}$$

where $(G_R)_{\text{satellite}}$ is the gain of the satellite receiving antenna in the direction of the transmitting earth station. This relation is represented in Figure 5.43 for the two cases considered:

— Global coverage ($\theta_{3\,\text{dB}} = 17.5°$), which implies $(G_R)_{\text{satellite}} = 29\,000/(\theta_{3\,\text{dB}})2 \approx 20$ dBi
— Spot-beam coverage ($\theta_{3\,\text{dB}} = 1.75°$), which implies $(G_R)_{\text{satellite}} = 29\,000/(\theta_{3\,\text{dB}})2 \approx 40$ dBi

The expression for $(C/N_0)_D$ for the downlink is given by:

$$(C/N_0)_D = (\text{EIRP})_{\text{satellite}}(1/L_D)\ (G/T)_{\text{station}}(1/k) \quad (\text{Hz}) \tag{5.85}$$

Assume that the power of the carrier transmitted by the satellite is $P_T = 10$ W $= 10$ dBW. Let $L_U = 200$ dB and neglect the implementation losses. Equation (5.85) becomes (all terms in dB):

$$\begin{aligned}(C/N_0)_D &= 10 - 200 + (G_T)_{\text{satellite}} + (G/T)_{\text{satation}} + 228.6 \\ &= (G_T)_{\text{satellite}} + (G/T)_{\text{station}} + 38.6\ (\text{dBHz})\end{aligned} \tag{5.86}$$

This relation is represented in Figure 5.44 for the two cases considered:

— Global coverage ($\theta_{3\,\text{dB}} = 17.5°$), which implies $(G_T)_{\text{satellite}} = 29\,000/(\theta_{3\,\text{dB}})^2 \approx 20$ dBi
— Spot-beam coverage ($\theta_{3\,\text{dB}} = 1.75°$), which implies $(G_T)_{\text{satellite}} = 29\,000/(\theta_{3\,\text{dB}})^2 \approx 40$ dBi

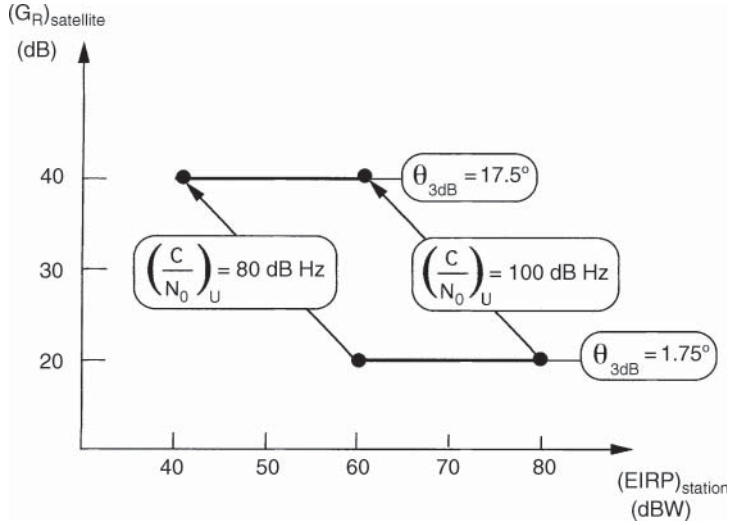


Figure 5.43 Comparison of the EIRP values required for an earth station in the case of global coverage ($\theta_{3\,dB} = 17.5°$) and in the case of a spot beam ($\theta_{3\,dB} = 1.75°$).

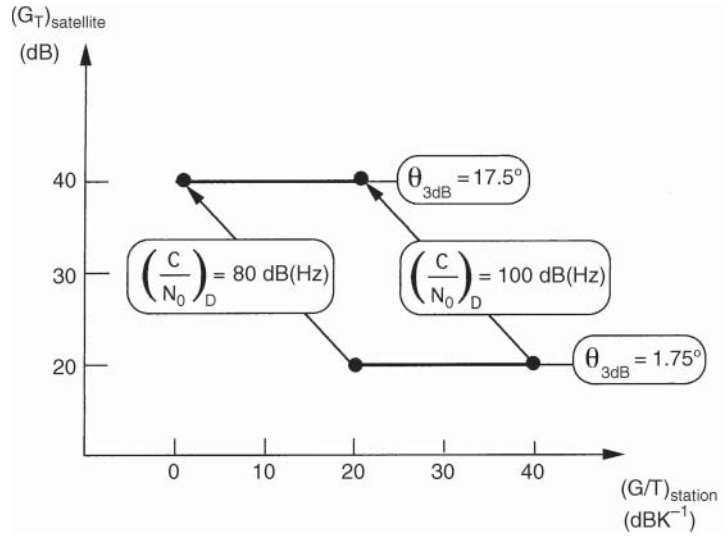


Figure 5.44 Comparison of the required values of factor of merit G/T for an earth station in the case of global coverage ($\theta_{3\,dB} = 17.5°$) and in the case of a spot beam ($\theta_{3\,dB} = 1.75°$).

In Figures 5.43 and 5.44, the oblique arrows indicate the reduction in $(EIRP)_{station}$ and $(G/T)_{station}$ when changing from a satellite with global coverage to a multibeam satellite with coverage by several spot beams. In this case, the multibeam satellite permits an economy of size, and hence cost, of the earth segment. For instance, a 20 dB reduction of $(EIRP)_{station}$ and $(G/T)_{station}$ may result in a 10-fold reduction of the antenna size (perhaps from 30 to 3 m) with a cost reduction for the earth station (perhaps from a few million Euros to a few 10 000 Euros).

If an identical earth segment is retained (a vertical displacement towards the top), an increase of C/N_0 is achieved that can be transferred to an increase of capacity, if sufficient bandwidth is available, at constant signal quality (in terms of bit error rate).

5.11.1.2 Frequency reuse

Frequency reuse consists of using the same frequency band several times in such a way as to increase the total capacity of the network without increasing the allocated bandwidth. An example has been seen in Section 5.2.3 of frequency reuse by orthogonal polarisation. In the case of a multibeam satellite, the isolation resulting from antenna directivity can be exploited to reuse the same frequency band in separate beam coverages. Figure 5.45 illustrates the principle of frequency reuse by orthogonal polarisation (Figure 5.45a) and the principle of reuse by angular beam separation (Figure 5.45b). A beam is associated with a given polarisation and a given coverage. In both cases, the bandwidth allocated to the system is B. The system uses this bandwidth B centred on the frequency f_U for the uplink and on the frequency f_D for the downlink. In the case of reuse by orthogonal polarisation, the bandwidth B is used twice only. In the case of reuse by angular separation, the bandwidth B can be reused for as many beams as the permissible interference level allows. Both types of frequency reuse can be combined.

The frequency reuse factor is defined as the number of times that the bandwidth B is used. In theory, a multibeam satellite with M single-polarisation beams, each being allocated the bandwidth B, which combines reuse by angular separation and reuse by orthogonal polarisation, may have a frequency reuse factor equal to $2M$. This signifies that it can claim the capacity that would be offered by a single-beam satellite with single polarisation using a bandwidth of $M \times B$. In practice, the frequency reuse factor depends on the configuration of the service area, which determines the coverage before it is provided by the satellite. If the service area consists of several widely separated regions (for example, urban areas separated by extensive rural areas), it is possible to reuse the same band in all beams. The frequency reuse factor can then attain the theoretical value of M. Figure 5.46 shows an example of multibeam coverage.

As the beam coverages are contiguous, the same frequency band cannot be used from one beam coverage to the next. In this example, the bandwidth allocated is divided into three equal

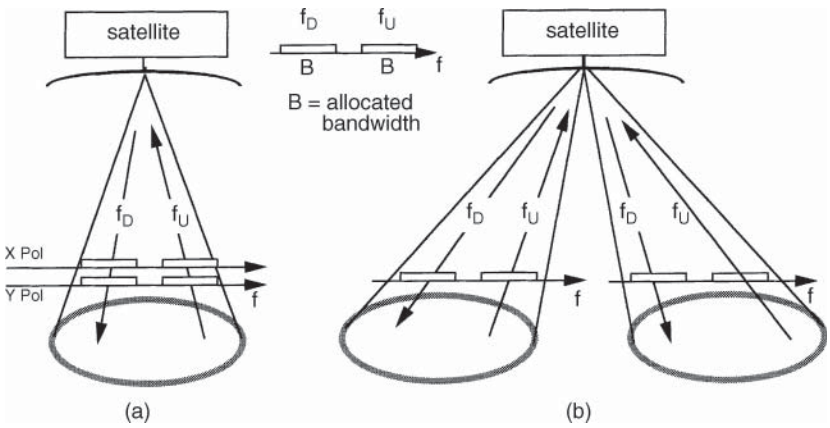


Figure 5.45 Frequency reuse with two beams ($M = 2$) by (a) orthogonal polarisation and (b) angular separation of the beams in a multibeam satellite system.

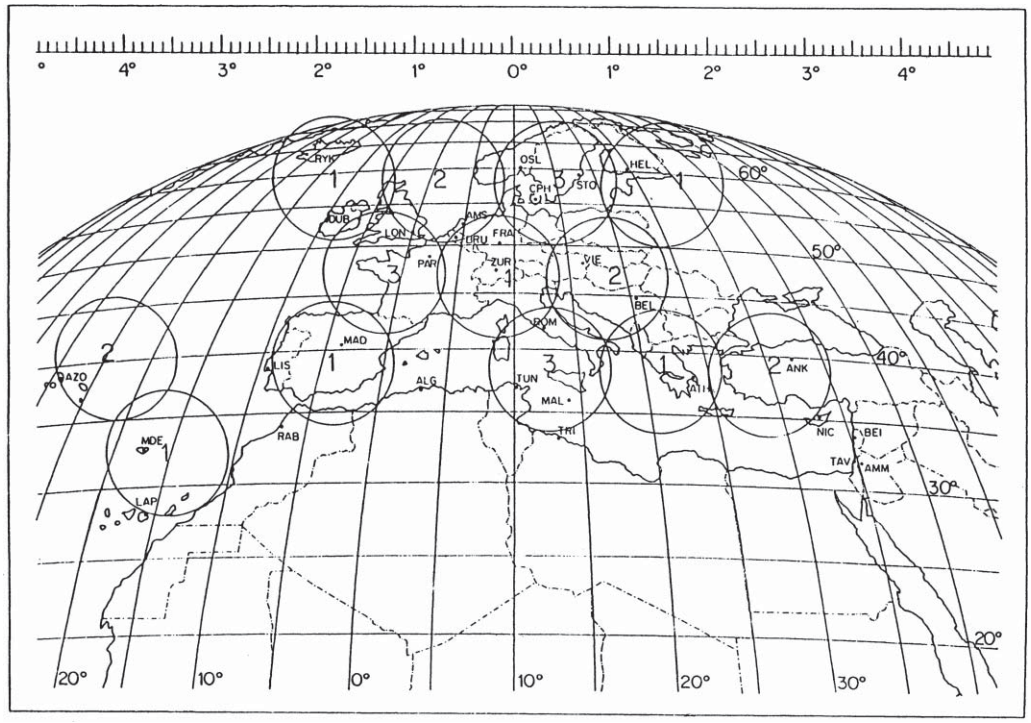


Figure 5.46 European coverage by a multibeam satellite system [LOP-82]. Source: Reproduced with the permission of the European Space Agency.

separate sub-bands, and each is used in beam coverages (1, 2, and 3) with sufficient angular separation from each other. The equivalent bandwidth, in the absence of reuse by orthogonal polarisation, has a value given by $6 \times (B/3) + 4 \times (B/3) + 3 \times (B/3) = 4.3\ B$ for $M = 13$ beams. The frequency reuse factor is then 4.3 instead of 13. With reuse by orthogonal polarisation within each beam coverage, the number of beams would be $M = 26$ and the frequency reuse factor would be 8.6.

5.11.2 Disadvantages of multibeam coverage

5.11.2.1 Interference between beams

Figure 5.47 illustrates interference generation within a multibeam satellite system, sometimes called *self-interference*. The allocated bandwidth B is divided into two sub-bands B_1 and B_2. The figure shows three beams. Beams 1 and 2 use the same band B_1. Beam 3 uses band B_2.

On the uplink (Figure 5.47a), the carrier at frequency f_{U1} of bandwidth B_1 transmitted by the beam 2 earth station is received by the antenna defining beam 1 in its side lobe with a low but nonzero gain. The spectrum of this carrier superimposes itself on that of the carrier of the same frequency emitted by the beam 1 earth station that is received in the main lobe with the maximum antenna gain. The carrier of beam 2 therefore appears as interference noise in the spectrum of the carrier of beam 1. This noise is called *co-channel interference* (CCI). Furthermore, part of the power of the carrier at frequency f_{U2} emitted by the earth station of beam 3 is introduced as a result of imperfect filtering of the input multiplexer (IMUX) filters defining the satellite channels (see Chapter 9) in the channel occupied by carrier f_{U1}. In this case, it consists of *adjacent channel interference* (ACI) analogous to that encountered in connection with frequency division multiple access in Section 6.5.3.

On the downlink (Figure 5.47b), the beam 1 earth station receives the carrier at frequency f_{D1} emitted with maximum gain in the antenna lobe defining beam 1. Downlink interference originates from the following contributions of power spectral density superimposed on the spectrum of this carrier:

— The spectra of the uplink adjacent channel and CCI noise retransmitted by the satellite.
— The spectrum of the carrier at the same frequency f_{D1} emitted with maximum gain in beam 2 and with a small but nonzero gain in the direction of the beam 1 station. This represents additional CCI.

The effect of self-interference appears as an increase in thermal noise under the same conditions as interference noise between systems analysed in Section 5.9.2. It must be included in the term $(C/N_0)_I$ that appears in Eq. (5.75). Taking into account the multiplicity of sources of interference, which become more numerous as the number of beams increases, relatively low values of $(C/N_0)_I$ may be achieved, and the contribution of this term impairs the performance in terms of $(C/N_0)_T$ of the total link. As modern satellite systems tend to reuse frequency as much as possible to increase capacity, self-interference noise in a multibeam satellite link may contribute up to 50% of the total noise.

5.11.2.2 Interconnection between coverage areas

A satellite payload using multibeam coverage must be in a position to interconnect all network earth stations and consequently must provide interconnection of coverage areas. The complexity

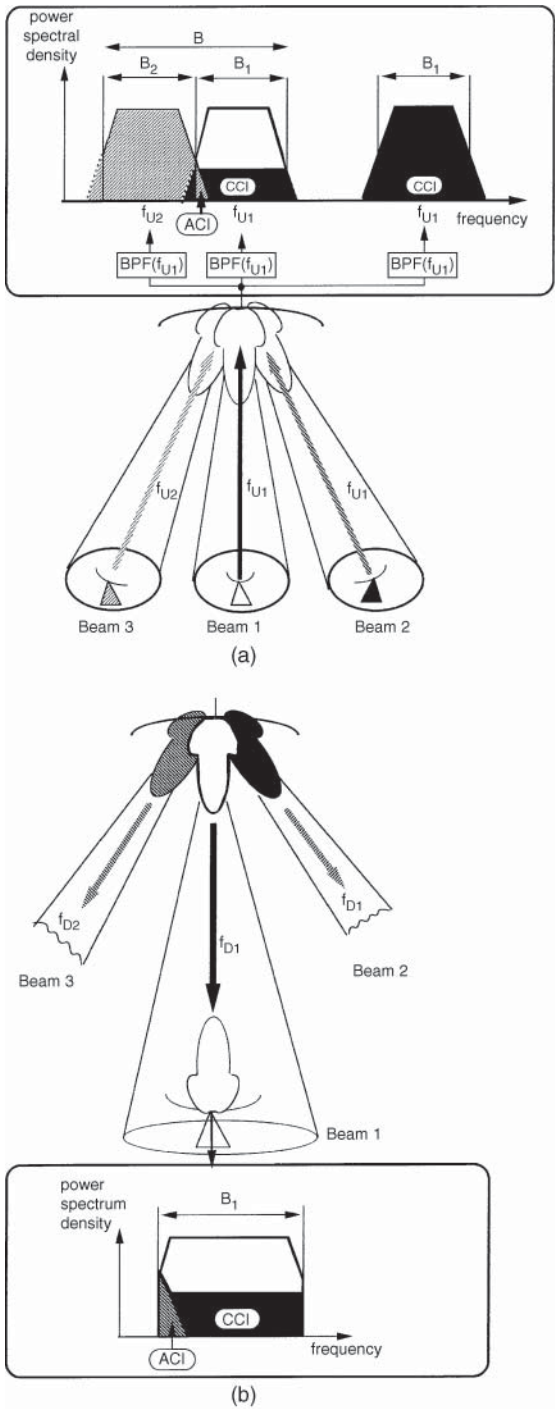


Figure 5.47 Self-interference between beams in a multibeam satellite system: (a) uplink; (b) downlink.

of the payload is added to that of the multibeam satellite antenna subsystem that is already much more complex than that of a single beam satellite.

Different techniques, depending on the on-board processing capability (no processing, transparent processing, regenerative processing, etc.) and on the network layer, are considered for interconnection of coverage:

— Interconnection by transponder hopping (no on-board processing)
— Interconnection by on-board switching (transparent and regenerative processing)
— Interconnection by beam scanning

These solutions are discussed in Chapter 7.

5.11.3 Conclusion

Multibeam satellite systems make it possible to reduce the size of earth stations and hence the cost of the earth segment. Frequency reuse from one beam to another permits an increase in capacity without increasing the bandwidth allocated to the system. However, interference between adjacent channels, which occurs between beams using the same frequencies, limits the potential capacity increase, particularly as interference is greater with earth stations equipped with small antennas.

5.12 INTERSATELLITE LINK PERFORMANCE

ISLs are links between satellites. Three types of ISLs can be considered:

— GEO to LEO links between geostationary (GEO) satellites and LEO satellites, also called interorbital links (IOLs)
— GEO to GEO links between geostationary satellites
— LEO to LEO links between low earth orbit satellites

Of course, one could consider ISLs between satellites in any type of orbit, but the listed configurations are those most considered in practice. The reader is referred to Section 7.5 for a discussion of the practical applications. Only the transmission aspects are presented here.

5.12.1 Frequency bands

Table 5.2 indicates the frequency bands allocated to ISLs by the Radio Regulations. These frequencies correspond to strong absorption by the atmosphere and have been chosen to provide protection against interference between ISLs and terrestrial systems. However, these bands are shared with other space services, and the limitation on interference level is likely to impose constraints on the choice of the defining parameters of ISLs [CCIR-90c; ITUR-93; CCIR-82b; CCIR-82c; ITUR-95; [ITUR-99]; [ITUR-02]; [ITUR-12]]. Table 5.2 also indicates the wavelengths envisaged for optical links. These result from the transmission characteristics of the components.

5.12.2 Radio-frequency links

The budget equations presented in Sections 5.1–5.6 can apply. Propagation losses reduce to free space losses since there is no passage through the atmosphere. Antenna pointing error can be

Table 5.2 Frequency bands for intersatellite links.

Intersatellite service	Frequency bands
Radio frequency	22.55–23.55 GHz
	24.45–24.75 GHz
	32–33 GHz
	54.25–58.2 GHz
Optical	0.8–0.9 μm (AlGaAs laser diode)
	1.06 μm (Nd:YAG laser diode)
	0.532 μm (Nd:YAG laser diode)
	10.6 μm (CO_2 laser)

Table 5.3 Typical values for terminal equipment of a radio-frequency intersatellite link.

Frequency (GHz)	Receiver noise factor (dB)	Transmitter power (W)
23–32	3–4.5	150
60	4.5	75
120	9	30

maintained at around a tenth of the beamwidth, and this leads to a pointing error loss on the order of 0 : 5 dB. The antenna temperature in the case of a GEO–GEO link, in the absence of solar conjunction, is on the order of 10 K. Table 5.3 indicates typical values for the terminal equipment. For practical applications, antenna dimensions are on the order of 1–2 m. Considering a frequency of 60 GHz and transmission and reception losses of 1 dB leads to:

— A receiver figure of merit G/T on the order of 25–29 dBK–1
— A transmitter EIRP on the order of 72–78 dBW

Because of the relatively wide beamwidth of the antenna (0.2° at 60 GHz for a 2 m antenna), establishing the link is not a problem. Each satellite orientates its receiving antenna in the direction of the transmitting satellite with a precision on the order of 0.1° to acquire a beacon signal that is subsequently used for tracking.

The development of high-capacity, radio-frequency ISLs between geostationary satellite systems implies reuse of frequencies from one beam to another. In view of the small angular separation of the satellites, it is preferable to use narrow-beam antennas with reduced side lobes in order to avoid interference between systems. Consequently, and in view of the limited antenna size imposed by the launcher and the technical complexity of the deployable antennas, the use of high frequencies is indicated. The use of optical links may be usefully considered in this context.

5.12.3 Optical links

In comparison with radio links, optical links have specific characteristics that are briefly described here. For a more complete presentation, refer to [KAT-87; GAG-91, Chapter 10; IJSC-88; WIT-94; BEG-00].

5.12.3.1 Establishing a link

Two aspects should be indicated:

— The small diameter of the telescope is typically on the order of 0.3 m. In this way, one is freed from congestion problems and aperture blocking of other antennas in the payload.
— The narrowness of the optical beam is typically 5 μrad. Notice that this width is several orders of magnitude less than that of a radio beam, and this is an advantage for protection against interference between systems. But it is also a disadvantage since the beamwidth is much less than the precision of satellite attitude control (typically 0.1° or 1.75 μrad). Consequently, an advanced pointing device is necessary; this is probably the most difficult technical problem.

There are three basic phases to optical communications:

— *Acquisition:* The beam must be as wide as possible in order to reduce the acquisition time. But this requires a high-power laser transmitter. A laser of lower mean power can be used that emits pulses of high peak power with a low duty cycle. The beam scans the region of space where the receiver is expected to be located. When the receiver receives the signal, it enters a tracking phase and transmits in the direction of the received signal. On receiving the return signal from the receiver, the transmitter also enters the tracking phase. The typical duration of this phase is 10 seconds.
— *Tracking*: The beams are reduced to their nominal width. Laser transmission becomes continuous. In this phase, which extends throughout the following, the pointing-error-control device must allow for movements of the platform and relative movements of the two satellites. In addition, since the relative velocity of the two satellites is not zero, a lead-ahead angle exists between the receiver line of sight and the transmitter line of sight. As demonstrated next, the lead-ahead angle is larger than the beamwidth and must be accurately determined.
— *Communications*: Information is exchanged between the two ends.

5.12.3.2 Lead-ahead angle

Consider two satellites, S1 and S2, respectively moving with velocity vectors $\mathbf{V}_{S1}$ and $\mathbf{V}_{S2}$, whose components orthogonal to the line joining S1 and S2 at time t are, respectively, the two vectors represented in Figure 5.48 by $\mathbf{V}_{T1}$ and $\mathbf{V}_{T2}$.

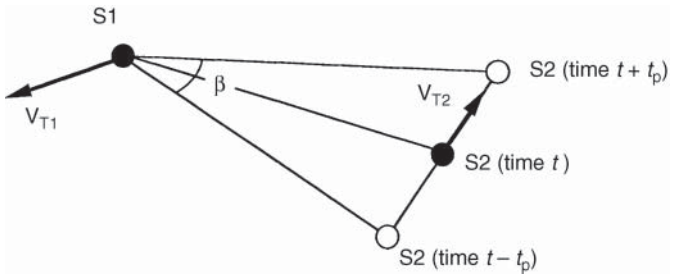


Figure 5.48 Lead-ahead angle for link between two satellites S1 and S2 with velocity vector components $\mathbf{V}_{T1}$ and $\mathbf{V}_{T2}$ in a plane perpendicular to the line joining S1 and S2 at time t; t_p is the propagation time of a photon from S1 to S2.

The propagation time of a photon from S1 to S2 is $t_p = d/c$, where d is the distance between the two satellites at time t and c is the speed of light ($c = 3 \times 108\ \text{m s}^{-1}$).

The lead-ahead angle β is given by:

$$\beta = 2 \mid \mathbf{V}_{T1} - \mathbf{V}_{T2} \mid /c \ \text{(rad)} \tag{5.87}$$

where $\mid \mathbf{V}_{T1} - \mathbf{V}_{T2} \mid$ is the modulus of the difference vector $\mathbf{V}_{T1} - \mathbf{V}_{T2}$.

Two situations are now considered: ISLs between two geostationary satellites; and interorbital links between a geostationary satellite and a low earth orbiting satellite.

5.12.3.2.1 GEO satellites separated by angle α

As both satellites are on the same circular orbit (Figure 5.49), the velocity vectors $\mathbf{V}_{S1}$ and $\mathbf{V}_{S2}$, which are tangential to the orbit, have equal modulus:

$$\mid \mathbf{V}_{S1} \mid = \mid \mathbf{V}_{S2} \mid = \omega(R_0 + R_E) = 3075\ \text{m/s}$$

where:

ω is the angular velocity of a geostationary satellite $= 7.293 \times 10^{-5}\ \text{rad s}^{-1}$
R_0 is the altitude of a geostationary satellite = 35 786 km
R_E is the earth radius = 6378 km

The component vectors $\mathbf{V}_{T1}$ and $\mathbf{V}_{T2}$, perpendicular to the line joining S1 and S2 at time t, both lie in the plane of the orbit and are opposite. They are at an angle $(\pi/2 - \alpha/2)$ with respect to vectors $\mathbf{V}_{S1}$ and $\mathbf{V}_{S2}$. Therefore:

$$\mid \mathbf{V}_{T1} - \mathbf{V}_{T2} \mid = 2\omega(R_0 + R_E)\cos(\pi/2 - \alpha/2) = 2\omega(R_0 + R_E)\sin(\alpha/2) \quad \text{(m/s)} \tag{5.88}$$

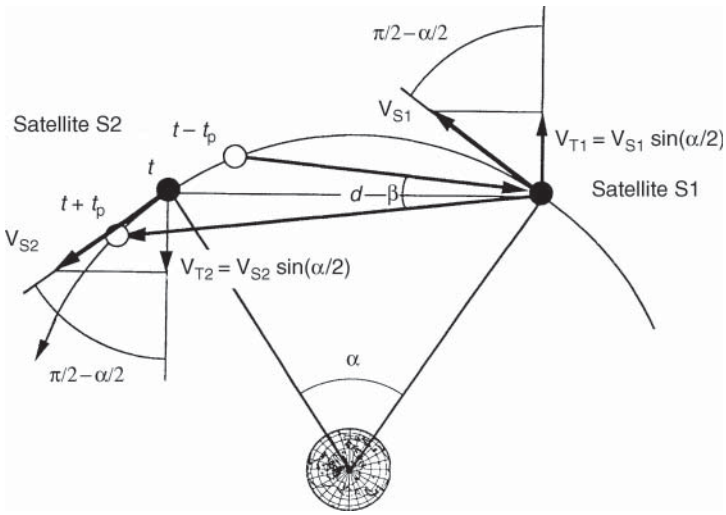


Figure 5.49 Lead-ahead angle for intersatellite links between two geostationary satellites.

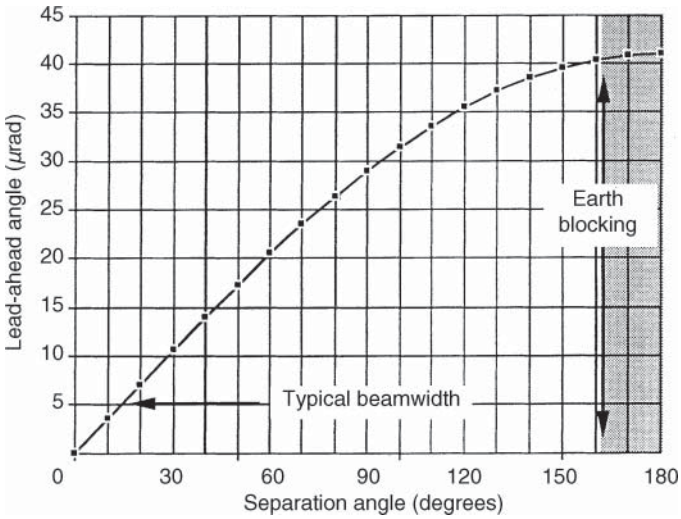


Figure 5.50 Lead-ahead angle as a function of the separation angle between two geostationary satellites.

From Eq. (5.87):

$$\beta = 2 \mid \mathbf{V}_{T1} - \mathbf{V}_{T2} \mid /c = 4\omega(R_0 + R_E)\sin(\alpha/2)/c \ \text{(rad)} \tag{5.89}$$

Figure 5.50 displays the lead-ahead angle β as a function of the separation angle α between the two geostationary satellites. Note that, for a separation angle larger than 15°, the lead-ahead angle is larger than the beamwidth (typically 5 μrad): for instance, $\beta = 10.6$ μrad for $\alpha = 30°$; $\beta = 20.5$ μrad for $\alpha = 60°$, and $\beta = 35.5$ μrad for $\alpha = 120°$.

5.12.3.2.2 A GEO satellite and a LEO satellite with circular orbit

The relative velocity of the two satellites (Figure 5.51) varies with time, and so does the value of the lead-ahead angle. Its maximum value is obtained when the LEO satellite crosses the equatorial plane. Denoting as i the LEO satellite orbit inclination, then:

$$\mid \mathbf{V}_{T1} - \mathbf{V}_{T2} \mid = \{|\mathbf{V}_{S1}|^2 + |\mathbf{V}_{S2}|^2 - 2 \mid \mathbf{V}_{S1} \mid\mid \mathbf{V}_{S2} \mid \cos i\}^{1/2} \tag{5.90}$$

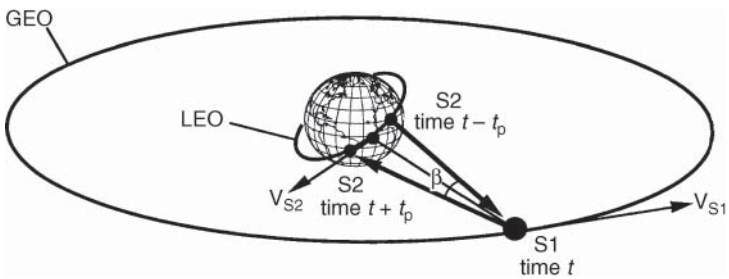


Figure 5.51 Lead-ahead angle at a GEO satellite for interorbital links between it and a LEO satellite.

where:

$$| \mathbf{V}_{S1} | = \omega_{GEO}(R_0 + R_E) = 3075 \text{ m/s}$$
$$| \mathbf{V}_{S2} | = \omega_{LEO}(h + R_E)$$

h is the LEO satellite altitude and $\omega_{LEO} + \mu^{1/2}(h + R_E)^{-3/2}$ is the LEO satellite angular rate ($\mu = 3.986 \times 10^{14}$ m^3 s–^{2}).

From Eq. (5.87), the lead-ahead angle is given by.

$$\begin{aligned}\beta &= 2 \mid V_{T1} - V_{T2} \mid /c \\ &= (2/c)\{|V_{S1}|^2 + |V_{S2}|^2 - 2 \mid V_{S1} \mid\mid V_{S2} \mid \cos i\}^{1/2} \quad \text{(rad)}\end{aligned} \tag{5.91}$$

The lead-ahead angle is the same for the two satellites. Considering $i = 98.5°$ and $h = 800$ km, then $\beta = 57$ μrad. Note this value is even larger than for ISLs between two geostationary satellites.

5.12.3.3 Transmission

Laser sources operate in single- and multi-frequency modes. In single-frequency mode, spectral width varies between 10 and 10 MHz. In multi-frequency mode, it is from 1.5 to 10 nm. The power emitted depends on the type of laser. Table 5.4 gives orders of magnitude.

Modulation can be internal or external. Internal modulation implies direct modification of the operation of the laser. External modulation is a modification of the light beam after its emission by the laser. The intensity, frequency, phase, and polarisation can be modulated. Phase and polarisation modulation are external. Intensity and frequency modulation can be internal or external. Polarisation modulation requires the presence of two detectors in the receiver, one for each polarisation. Because of this, it is preferable to reserve polarisation for multiplexing of two channels.

The intensity distribution of a laser beam, as a function of angle with respect to the maximum intensity, follows a Gaussian law. The on-axis gain is given by:

$$G_{Tmax} = 32/(\theta_T)^2 \tag{5.92}$$

where θ_T is the total beamwidth at $1/e^2$ where $e = 2.718$. The choice of θ_T depends on the pointing accuracy. With imprecise pointing, a large θ_T is better but gain is lost. If θ_T is reduced, there is benefit in gain but the pointing error loss increases. It can be shown that, if the pointing error is essentially an alignment error, the (maximum gain × pointing error loss) product is maximum when $\theta_T = 2.8 \times$ (pointing error) [[KAT-87], p. 51]. In general, for a pointing error of any kind, the beamwidth may be adapted to the pointing error.

Table 5.4 Typical values of transmitted power for lasers.

Type of laser	Wavelength (μ)	Transmitted power
Solid state (laser diode)		
AlGaAs	0.8–0.9	About 100 mW
InPAaGa	13–1.5	About 100 mW
Nd:YAG	1.06	0.5–1 W
Nd:YAG	0.532	100 mW
Gas laser		
CO_2	10.6	Several tens of watts

In addition to losses due to pointing error, transmission losses and degradation of the wavefront in the emitting optics occur.

5.12.3.4 Transmission loss

Transmission loss reduces to the free space loss:

$$L = (4\pi R/\lambda)^2 \tag{5.93}$$

where λ is the wavelength and R is the distance between transmitter and receiver.

5.12.3.5 Reception

The receiving gain of the antenna is given by:

$$G_R = (\pi D_R/\lambda)^2$$

where D_R is the effective diameter of the receiver antenna.

The receiver can be of a *direct detection* (Figure 5.52) or a *coherent detection* receiver (Figure 5.53). With direct detection, the incident photons are converted into electrons by a photodetector. The subsequent baseband electric current at the photodetected output is amplified and then detected by a matched filter. With coherent detection, the optical signal field associated with the incident photons is mixed with the signal from a local laser. The resulting optical field is converted into a bandpass electric current by a photodetector and is subsequently amplified by an intermediate frequency amplifier. The demodulator detects the useful signal either by envelope detection or by coherent demodulation.

The receiving losses include optical transmission losses and, for coherent detection, losses associated with the degradation of the wavefront. (The quality of the wavefront is an important characteristic for optimum mixing of the received signal field and the local oscillator field at the photodetector front end.) Filtering, to reject out-of-band photons, also introduces losses,

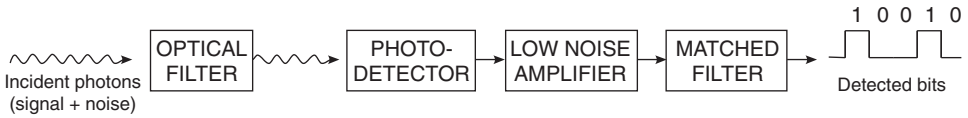


Figure 5.52 Optical direct detection receiver.

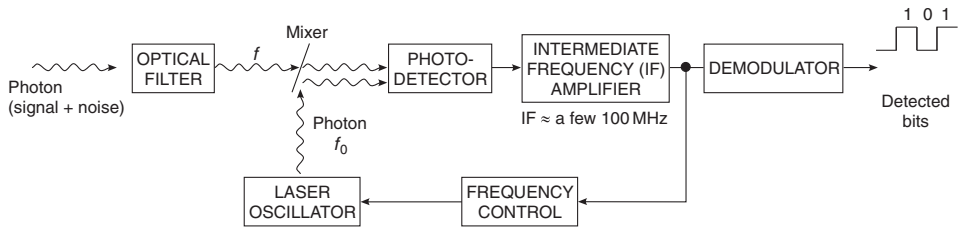


Figure 5.53 Optical coherent detection receiver.

since the transmission coefficient reduces with bandwidth. A typical filter width is from 0.1 to 100 nm.

The signal-to-noise power ratio at the detector output depends on the type of detection. For direct detection (Figure 5.52):

$$S/N = I_{\mathrm{S\,dd}}^2 / i_{\mathrm{dd}}^2$$

The quantity $I_{\mathrm{S\,dd}}$ represents the signal current intensity:

$$I_{\mathrm{S\,dd}} = (P_{\mathrm{S}}/hf)\eta_p eG \quad \text{(A)} \tag{5.94}$$

where:

P_{S} = useful optical signal power (W)
h = Planck's constant = 6.6×10^{-34} J Hz^{-1}
f = laser frequency (Hz)
η_{p} = quantum efficiency of the photodetector, typically 0.8 for an avalanche photodetector (APD)
e = electron charge = 1.6×10^{-19} C
G = photodetector gain, on the order of 50 – 300 for APD and 10^4 to 10^6 for a vacuum tube photomultiplier

Also (P_{S}/hf) represents the number of photons received per second, and $K = \eta_{\mathrm{p}} e/hf$ represents the sensitivity of the photodetector (A/W). Hence:

$$I_{\mathrm{S\,dd}} = KGP_S \text{ (A)} \tag{5.95}$$

The quantity i_{dd} represents the rms noise current intensity:

$$i_{\mathrm{dd}}^2 = i_{\mathrm{nS}}^2 + i_{\mathrm{nB}}^2 + i_{\mathrm{nD}}^2 + i_{\mathrm{nT}}^2 \ \text{(A}^2\text{)} \tag{5.96}$$

$i_{\mathrm{nS}}^2 = 2eKP_S G^2 f(G) B_{\mathrm{N}}$ is the signal shot noise.
$i_{\mathrm{nB}}^2 = 2eKP_{\mathrm{n}} G^2 f(G) B_{\mathrm{N}}$ is the background shot noise.
$i_{\mathrm{nD}}^2 = 2ei_0 B_{\mathrm{N}}$ is the dark current shot noise.
$i_{\mathrm{nT}}^2 = N_0 B_{\mathrm{N}}$ is the thermal noise of the electronic amplifying circuits.

In these formulae, P_{n} is the received background optical noise power (W), $f(G)$ is a multiplying factor taking into account the noise generated in the photodetector by secondary electrons (typically $f(G) = a + bG$ where $a \approx 2$ and $b \approx 0:01$), i_0 is the dark current intensity (A), N_0 is the electronic amplifier thermal noise spectral density (A^2/Hz), and B_{N} is its noise bandwidth (Hz).

If all sources of noise besides the signal shot noise could be eliminated ($i_0 = 0$, $P_{\mathrm{n}} = 0$, $N_0 = 0$, $f(G) = 1$), one would achieve the *quantum-limited* S/N value, given by:

$$(S/N)_{\mathrm{q1}} = \eta_{\mathrm{p}} P_{\mathrm{S}} / 2hfB_{\mathrm{N}} \tag{5.97}$$

For coherent detection (Figure 5.53):

$$S/N = I_{\mathrm{S\,cd}}^2 / (i_{\mathrm{dd}}^2 + i_{\mathrm{LO}}^2)$$

The quantity $I_{\mathrm{S\,cd}}$ represents the signal current intensity:

$$I_{\mathrm{S\,cd}} = KG\eta_{\mathrm{m}} L_{\mathrm{p}} (2P_s P_{\mathrm{LO}})^{1/2} \text{ (A)} \tag{5.98}$$

where:

η_m = mixing efficiency
L_P = loss due to polarisation mismatch

The local oscillator power i_{dd} represents the rms noise current intensity for direct detection, as determined from Eq. (5.96); i_{LO} represents a supplementary source of noise, i.e. the local oscillator rms noise current intensity:

$$i_{LO}^2 = 2eKP_{LO}G^2 f(G)B_{IF}$$

where B_{IF} is the noise bandwidth of the intermediate frequency amplifier (Hz).

The quantity P_{LO} can be increased to the point where i_{LO} is the predominant source of noise:

$$S/N \approx I_{S\,cd}^2 / i_{LO}^2 = \eta_m L_P \eta_p Ps / f(G) B_{IF} hf$$

Coherent detection confers a higher value of S/N than direct detection. In theory, one could achieve the quantum-limited S/N value, $(S/N)_{ql}$, given by Eq. (5.97), considering $\eta_m = 1$; $L_P = 1$; $f(G) = 1$; and $B_{IF} = 2B_N$. However, in the case of alignment error between the local oscillator and the beam signal, mixing efficiency is degraded. This type of detection cannot therefore be used both for acquisition and tracking. Unless high data rates are involved, there is no advantage in weight or power from using coherent detection techniques for communications along with a separate direct detection receiver for acquisition and tracking, compared to the situation where a direct detection receiver is used for both. For high data rates (typically greater than 1 Gbit s^{-1}), the power required for direct detection is excessive, and one may consider resorting to coherent detection for the communications function.

5.12.4 Conclusion

The choice between radio and optical links depends on the mass and power consumed. In general terms, it can be said that the advantage is with radio links for low throughputs (less than 1 Mbit s^{-1}). For high-capacity links (several tens of Mbit/s), optical links command attention.

For a link involving one uplink, one or more optical ISLs, and one downlink, overall station-to-station link performance should be established in the same way as overall link performance for regenerative satellites. Indeed, the implementation of intersatellite links, be it at radio frequency or with optical technology, is mainly of interest when on-board demodulation is available, so as to provide flexible on-board switching.

REFERENCES

[BEG-00] Begey, D.L. (2000). Laser cross links systems and technology. *IEEE Communications Magazine* **38** (8): 126–132.

[CAS-98] Castanet, L., Lemorton, J., and Bousquet, M. (1998). Fade mitigation techniques for new satcom services at Ku band and above: a review. In: *4th Ka Band Utilisation Conference, Venice*, 119–128.

[CCIR-82a] CCIR. (1982). Propagation data required for space telecommunication systems. Report 564.

[CCIR-82b] CCIR. (1982). Frequency sharing between the inter-satellite service when used by the fixed-satellite service and other space services. Report 874.

[CCIR-82c] CCIR. (1982). Sharing between the Inter-satellite service and broadcasting satellite service in the vicinity of 23 GHz. Report 951; also referred to as ITU-R Report BO.951-0.

[CCIR-90a] CCIR. (1990). Earth-station antennas for the fixed-satellite service. Report 390.

[CCIR-90b] CCIR. (1990). Contributions to the noise temperature of an earth-station receiving antenna. Report 868.

[CCIR-90c] CCIR. (1990). Factors affecting the system design and the selection of frequencies for inter-satellite links of the fixed-satellite service. Report 451.

[FEN-95] Fenech, H.T., Kasstan, B., Lindley, A. et al. (1995). G/T predictions of communications satellites based on new earth brightness model. *International Journal of Satellite Communications* **13** (5): 367–376.

[GAG-91] Gagliardi, R.M. (1991). *Satellite Communications*, 2e. Van Nostrand Reinhold.

[ITU-02] ITU (2002). *ITU Handbook on Satellite Communications*, 3e. Wiley.

[ITUR-93] ITU-R. (1993). Reference earth-station radiation pattern for use in coordination and interference assessment in the frequency range from 2 to about 30 GHz. Recommendation S.465.

[ITUR-95] ITU-R. (1995). Sharing between the inter-satellite service involving geostationary satellites in the fixed-satellite service and the radionavigation service at 33 GHz. Recommendation S.1151.

[ITUR-99] ITU-R. (1999). Sharing between spaceborne passive sensors of the Earth exploration-satellite service and inter-satellite links of geostationary-satellite networks in the range 54.25 to 59.3 GHz. Recommendation S.1339.

[ITUR-02] ITU-R. (2002). Sharing of inter-satellite link bands around 23, 32.5 and 64.5 GHz between non-geostationary/geostationary inter-satellite links and geostationary/geostationary inter-satellite links. Recommendation S.1591.

[ITUR-05] ITU-R. (2005).Specific attenuation model for rain for use in prediction methods. Recommendation P.838-3.

[ITUR-12] ITU-R. (2012). Protection criteria and interference assessment methods for non-GSO inter-satellite links in the 23.183-23.377 GHz band with respect to the space research service. Recommendation S.1899.

[ITUR-13] ITU-R. (2013). Rain height model for prediction methods. Recommendation P.839-4.

[ITUR-16a] ITU-R. (2016). Attenuation by atmospheric gases. Recommendation P.676-11.

[ITUR-16b] ITU-R (2016) Radio noise, P series - radiowave propagation. Recommendation P.372-13.

[ITUR-17a] ITU-R. (2017). Attenuation due to clouds and fog. Recommendation P.840-7.

[ITUR-17b] ITU-R. (2017). Characteristics of precipitation for propagation modelling. Recommendation P.837-7.

[ITUR-17c] ITU-R. (2017). Propagation data and prediction methods required for the design of Earth-space telecommunication systems. Recommendation P.618-13.

[KAT-87] Kaitzman, M. (ed.) (1987). *Laser Satellite Communications*. Prentice-Hall.

[LOP-82] Lopriore, M., Saitto, A., and Smith, G.K. (1982). A unifying concept for future fixed satellite service payloads for Europe. *ESA Journal* **6** (4): 371–396.

[IJSC-88] Peters, R.A. (ed.) (1988). *International Journal of Satellite Communications* **6** (2): 77–240. Special Issue on Intersatellite links.

[THO-83] Thorn, R.W., Thirlwell, J., and Emerson, D.J. (1983). Slant path radiometer measurements in the range 11–30 GHz at Martlesham heath, England. In: *3rd International Conference of Antennas and Propagation, ICAP 83*, 156–161. IEEE.

[WAC-81] Washira, M., Arunachalam, V., Feher, K., and Lo, G. (1981). Performance of power and bandwidth efficient modulation techniques in regenerative and conventional satellite systems. In: *International Conference on Communications, Denver*, 37.2.1–37.2.5. IEEE.

[WIT-94] Wittig, M. (1994). Optical space communications. *Space Communications* **12** (2).

6 MULTIPLE ACCESS

This chapter deals with techniques that allow several carriers from several earth stations to access the satellite. Between the transmitting and receiving antennas, the communication payload of the satellite incorporates a repeater that consists of one or more channels, called *transponders*, operating in parallel on different sub-bands of the total bandwidth used (see Chapter 9). Information transfer between several earth stations implies the establishment of several simultaneous station-to-station communications channels (connections) through a given satellite repeater channel.

Only networks based on a single-beam antenna payload are considered in this chapter. The more complex case of multibeam antenna payload networks is treated in Chapter 7. In the context of a single-beam payload, the carriers transmitted by earth stations access the same satellite receiving antenna beam, and these same earth stations can receive all the carriers retransmitted by the satellite transmit antenna.

After introducing the concepts of layered transmission and traffic parameters, this chapter successively analyses routing information and multiple access by presenting the following three fundamental techniques:

— Frequency division multiple access (FDMA)
— Time division multiple access (TDMA)
— Code division multiple access (CDMA)

Towards the end of this chapter, fixed and on-demand assignment and random assignment are explained for applications on different traffic requirement; and finally a Section 6.10 concludes the chapter.

6.1 LAYERED DATA TRANSMISSION

The exchange of information between source and destination in a communications network involves a lot of interacting functions ranging from generation of the electric signal representing the information to the smart display of the information to the end user. In order to master interactions and facilitate design, it is useful to identify and group together tasks of a similar nature and clarify the interactions between the various groups in a well-structured architecture. The functions of systems are therefore divided into *layers*, with sets of rules (called *protocols*) taking care of the exchange of information between layers.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

The layering principle allows for the definition of the concept of a reference model. In the 1980s, the International Organisation for Standardisation (ISO) derived the seven-layer reference model called the Open Systems Interconnection (OSI) reference model based on clear and simple principles (see Chapter 7, Figure 7.1). The full description of the OSI model is discussed in Section 7.1.

The features of Layer 1 deal with what has been presented in Chapters 3, 4, and 5. Layer 1 is the *Physical layer*, which specifies electrical interfaces (bits and waveforms) and the physical transmission medium (propagation channel). In a satellite network, Layer 1 consists of the modulation and channel coding techniques that enable the bit stream to be transmitted and the radio links acting as the physical transmission media.

The features of Layer 2 are relevant to this chapter and Chapter 7. Layer 2 is the *Data Link layer* delivering data streams appearing free of undetected transmission errors to the layer above (the Network layer). A special sublayer called *Medium Access Control* (MAC) deals with the sharing of the physical resource between communicating terminals.

6.2 TRAFFIC PARAMETERS

6.2.1 Traffic intensity

In 1946, the International Telecommunication Union (ITU) (formally Comité Consultatif International Télégraphique et Téléphonique CCITT) decided to use the *erlang* as a traffic-intensity unit to honour Angner Krarup Erlang (1 January 1817–3 February 1929) – a Danish mathematician, statistician, and engineer – for his contributions to the field of telecommunications with the invention of traffic engineering and queueing theory.

The traffic intensity A is defined as

$$A = R_{\text{call}} T_{\text{call}} \quad \text{(erlang)} \tag{6.1}$$

where

R_{call} = mean number of calls per unit time (s^{-1})
T_{call} = mean duration of a communication (s)

6.2.2 Call blocking probability

It is assumed that the number of users generating calls is much greater than the number of communication channels C provided between end terminals, and blocked calls are not stored. Under these conditions, the Erlang B formula indicates the probability that n channels are occupied ($n \leq C$):

$$E_n(A) = (A^n/n!) \Big/ \sum_{k=0}^{k=C} (A^k/k!) \tag{6.2}$$

The probability of blocking is given by ($n = C$):

$$B(C, A) = E_C(A) \tag{6.3}$$

The *B* means *blocking* in the Erlang *B* formula: a new arriving call will be blocked (cannot be served) and will be cleared from the system. If the call is re-tried, it is treated as a new call without any effect on the system.

Figure 6.1a displays the required number of communication channels versus traffic intensity, given the probability of call blocking. An approximation for the required number of channels given the traffic intensity A is:

$$C = A + \alpha A^{1/2} \tag{6.4}$$

where α is the exponent in the $10^{-\alpha}$ blocking probability objective.

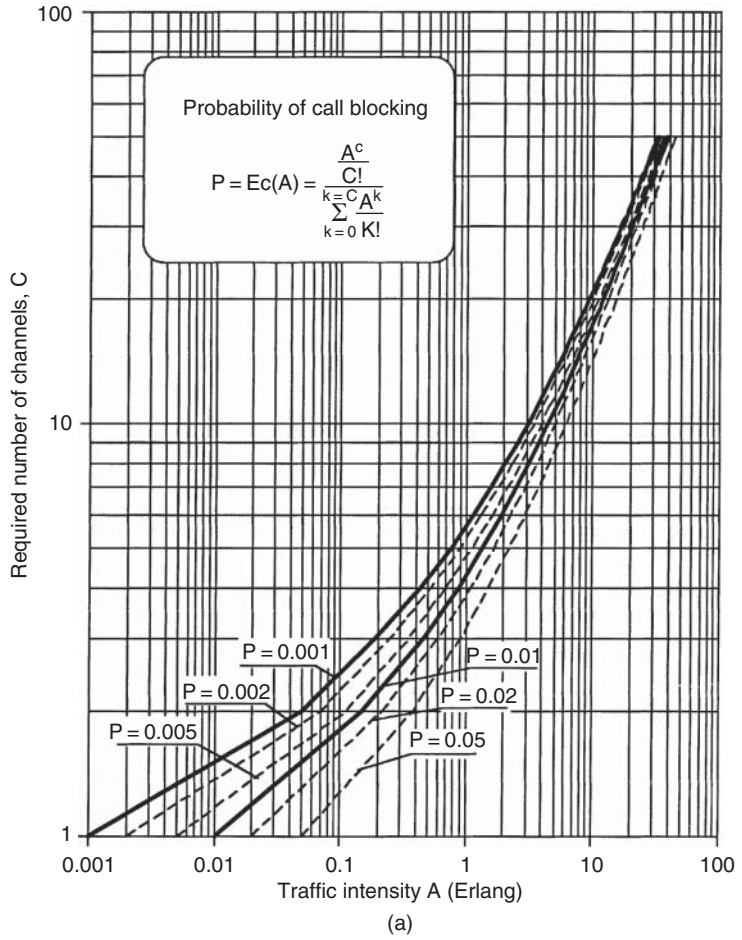


Figure 6.1 (a) Required number of communication channels to ensure connection setup with a given call-blocking probability as a function of the traffic intensity. (b) Probability of calls being delayed for a system with offered traffic A and number of channels n.

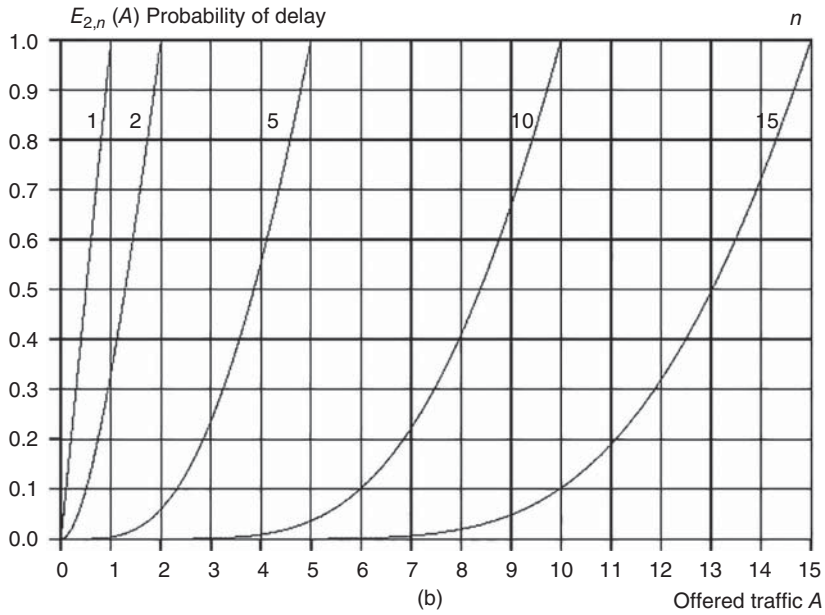


Figure 6.1 (*Continued*)

6.2.3 Burstiness

Once the connection is established, transfer of information may occur. *Burstiness* relates to the intermittent transfer of data. Information is conveyed in the form of data bursts that are generated at random intervals. Typically the situation arises when a human-operated personal computer is activated by its operator after some thinking time, or from the specific protocols that are used for data transfer with information being segmented by the transmitting terminal and segments being acknowledged in the form of short messages by the receiving terminal prior to further transmission by the transmitting terminal [MAR-04]. A common example is the traffic generated by the Transmission Control Protocol/Internet Protocol (TCP/IP). Burstiness is defined as the peak bit rate of the active information source divided by the average bit rate:

$$\mathrm{BU} = R/\lambda L \tag{6.5}$$

where R is the peak bit rate (bit/s), λ is the message generation rate (s^{-1}), and L is the length of the message (bits). Continuous information corresponds to low burstiness (stream traffic with BU of order 1–5), whereas highly intermittent traffic is characterised by high burstiness (BU = 10^3–10^5).

6.2.4 Call delay probability

Erlang also developed the Erlang C formula (C means *congestion*): if a new call arrives when all channels are occupied, the call can wait and hope that some calls already in the system will finish soon. The Erlang C formula (also called Erlang's second formula) is able to calculate the

Table 6.1 Examples of numerical evaluation of Erlang's C formula

$E_{2,n}(A)$	A = 1	2	3	4	5	6	7	8
n = 1	1							
2	0.837209	1						
3	0.683544	0.81203	1					
4	0.541353	0.639053	0.779783	1				
5	0.413265	0.484305	0.584838	0.738041	1			
6	0.30191	0.350903	0.418875	0.519508	0.683785	1		
7	0.209438	0.241221	0.284377	0.346339	0.442825	0.61383	1	
8	0.136903	0.156157	0.181713	0.217271	0.270131	0.356981	0.52614	1
9	0.083751	0.094584	0.108636	0.127591	0.154558	0.195981	0.267736	0.422384
10	0.047707	0.053352	0.060512	0.069893	0.082715	0.101299	0.130654	0.183963
11	0.025232	0.027957	0.031342	0.03566	0.041358	0.049222	0.06078	0.07943
12	0.012385	0.013607	0.015096	0.016951	0.019326	0.022474	0.026848	0.033337
13	0.00565	0.006161	0.006773	0.00752	0.008452	0.009647	0.011237	0.013454

probability that a call will have to wait and how long it has to wait given the traffic load and capacity of the system:

$$E_{2,n}(A) = \frac{\dfrac{A^n}{n!}\dfrac{n}{n-A}}{\displaystyle\sum_{i=0}^{n-1}\frac{A^i}{i!} + \frac{A^n}{n!}\frac{n}{n-A}} \quad where \;\; A < n$$

This gives the probability of delay when calls arrive for a system with load of A and number of channels n (see Figure 6.1b). Table 6.1 shows examples of numerical evaluation of the Erlang C formula.

The relationship between Erlang's C formula and B formula are as follows:

$$E_{2,n}(A) = \frac{E_{1,n}(A)}{1 - A\{1 - E_{1,n}(A)\}/n}, \qquad \text{A < n}$$

or

$$\frac{1}{E_{2,n}(A)} = \frac{1}{E_{1,n}(A)} - \frac{1}{E_{1,n-1}(A)}$$

If the call is delayed, it has to wait in a queue. Let L is the queue length. The probability of queue length $L > 0$ can be calculated as:

$$p\{L > 0\} = \frac{A}{n}E_{2,n}(A)$$

The mean queue length can be calculated as:

$$L_n = \frac{A\,E_{2,n}(A)}{n - A}$$

The mean queue length, given the queue L > 0, can be calculated as

$$L_{nq} = \frac{n}{n - A}$$

Table 6.2 Required capacities in a network containing three stations

	To station		
From station	A	B	C
A	—	C_{AB}	C_{AC}
B	C_{BA}	—	C_{BC}
C	C_{CA}	C_{BC}	—

6.3 TRAFFIC ROUTING

Given a demand for traffic in a network of N earth stations, it is necessary to establish an adequate information transfer capacity between each pair of stations. This capacity is calculated as a function of demand expressed by traffic intensity and acceptable blocking probability (a typical value is 0.5–1%) using the curves in Figure 6.1.

Let C_{XY} be the capacity, expressed as a number of communication channels for transfer of traffic from station X to station Y. The set of capacities required for exchanges between the N stations is described by a matrix of dimension N with a zero leading diagonal ($C_{XX} = 0$). Table 6.2 shows the required capacities in a network containing three stations (X = A, B, C; Y = A, B, C).

Information transfer occurs in accordance with the techniques described in Chapter 4 and implies modulation of the radio-frequency carrier relayed by the satellite channel. At network level, two techniques are considered for traffic routing (Figure 6.2):

— One carrier per station-to-station link
— One carrier per transmitting station

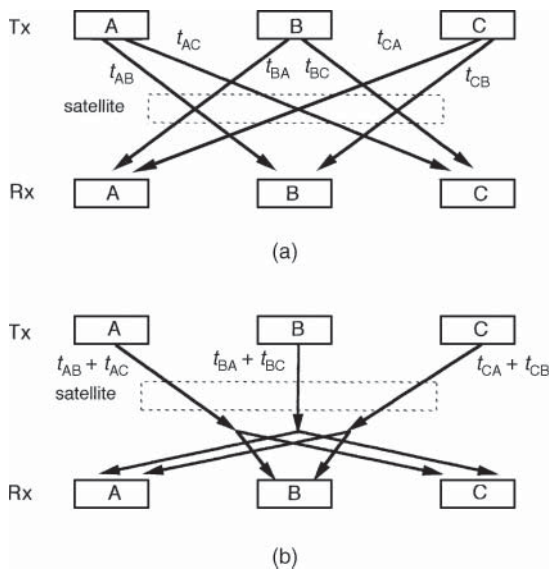


Figure 6.2 Traffic routing techniques: (a) one carrier per station-to-station link; (b) one carrier per transmitting station.

6.3.1 One carrier per station-to-station link

As shown in Figure 6.2a, one carrier carries the traffic t_{XY} from station X to station Y. The number of carriers is equal to the number of nonzero coefficients in the matrix, which is $N(N-1)$. The matrix coefficients define the required capacity for each carrier.

6.3.2 One carrier per transmitting station

As shown in Figure 6.2b, the broadcasting property of the satellite is used; this enables every station to receive all the carriers transmitted to the satellite (for a single-beam satellite). Under these conditions, it can be seen that the task of carrying all the traffic from station X to all other stations can be assigned to a single carrier. The number of carriers is equal to the number of stations N. The capacity of each carrier is given by the sum of the coefficients of the row of the matrix that corresponds to the transmitting station.

6.3.3 Comparison

It can be observed that the approach in Figure 6.2a leads to a greater number of carriers than the approach in Figure 6.2b, and each carrier has a smaller capacity. However, the receiving station receives only traffic that is intended for it in Figure 6.2a; in Figure 6.2b, the receiving station Y must extract 'X to Y' traffic from the total traffic conveyed by the carrier received from station X.

The choice between these two approaches is an economic one. It depends on considerations such as the number of satellite channels, the bandwidth of the satellite channel, and the multiple-access technique used. In general, the fact that a large number of carriers is relayed by the satellite has a greater penalty than having to transmit carriers of higher capacity. Therefore, the 'one carrier per transmitting station' approach is the most often used.

6.4 ACCESS TECHNIQUES

The problem of multiple access arises when several carriers are handled simultaneously by a satellite repeater that is a nodal point of the network. The satellite repeater consists of several adjacent channels (called *transponders*), whose bandwidth is a fraction of the total repeater bandwidth (Chapter 9). Any specific carrier falls within a given repeater channel. In terms of multiple access, there are two aspects to be considered:

— Multiple access to a particular repeater channel (i.e. a transponder)
— Multiple access to a satellite repeater

6.4.1 Access to a particular satellite channel (or transponder)

Each satellite repeater channel (transponder) amplifies every carrier whose spectrum falls within its passband at a time when the channel is in an operational state. The resource offered by each channel can thus be represented in the form of a rectangle in the time–frequency plane. This rectangle represents the bandwidth of the channel and its duration of operation (Figure 6.3). In the absence of special precautions, carriers would occupy this rectangle simultaneously and mutually interfere. To avoid this interference, it is necessary for receivers (an earth station receiver for

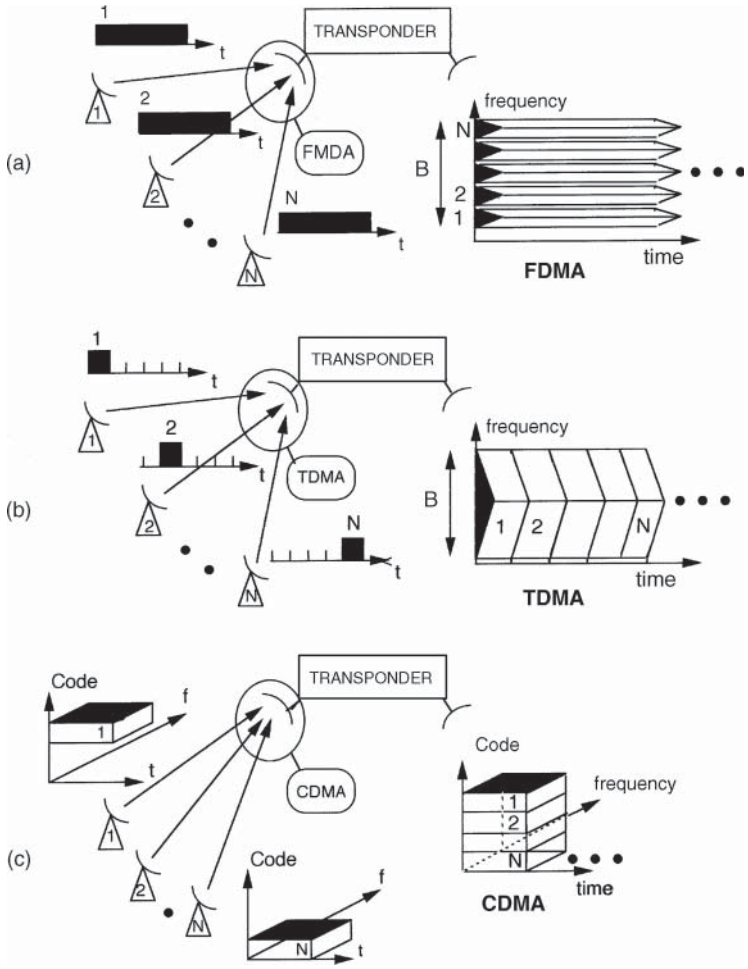


Figure 6.3 The principles of multiple access: (a) frequency division multiple access (FDMA); (b) time division multiple access (TDMA); (c) code division multiple access (CDMA). (B = channel (transponder) bandwidth.)

a transparent satellite and an on-board satellite receiver for a regenerative satellite) to be able to discriminate between the received carriers. This discrimination can be achieved:

— As a function of the carrier energies in *frequency* domain. If the spectra of the carriers each occupy a different sub-band, the receiver can discriminate between carriers by filtering. This is the principle of *frequency division multiple access* (FDMA, Figure 6.3a).
— As a function of the carrier energies in *frequency* domain. Several carriers received sequentially by the receiver can be discriminated by temporal gating even if they occupy the same frequency band. This is the principle of *time division multiple access* (TDMA, Figure 6.3b).
— By the addition of a *signature* that is known to the receiver and is specific to each carrier. This ensures identification of the carrier even when all the carriers occupy the same frequency

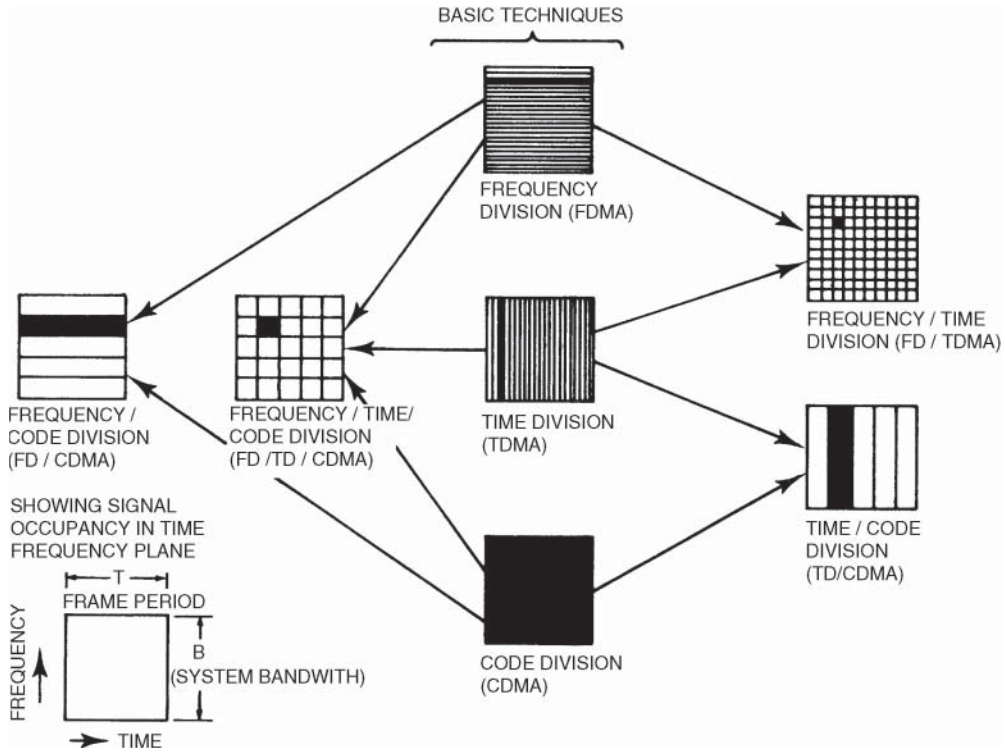


Figure 6.4 Combination of the three fundamental types of multiple access into hybrid access types.

band simultaneously. The signature is most often realised by means of pseudorandom codes (pseudonoise [PN] codes); hence the name *code division multiple access* (CDMA, Figure 6.3c). The use of such codes has the effect of broadening the carrier spectrum in comparison with that which it would have if modulated only by the useful information. This is why CDMA is also sometimes called *spread spectrum multiple access* (SSMA).

Several types of multiple access as defined earlier can be combined; Figure 6.4 illustrates the range of combinations.

6.4.2 Multiple access to the satellite repeater

Multiple access to a particular repeater channel (transponder) implies prior multiple access to the satellite repeater. Access to a satellite repeater is achieved as a function of the frequency and polarisation of the carrier. For every carrier with a given polarisation and frequency, there is obligatory FDMA access to the repeater together with FDMA, TDMA, or CDMA access to each channel. The corresponding combinations in Figure 6.4 can thus be considered as representative of multiple access to a satellite repeater. In all cases, the spectral occupation of a carrier must not exceed the channel bandwidth.

6.4.3 Performance evaluation – efficiency

The *efficiency*, η, of a multiple-access scheme is conveniently evaluated by the ratio of the capacity available from the transponder in the considered multiple-access mode to the capacity that would be available if the transponder were accessed by a single carrier occupying the full bandwidth of the transponder operated at saturation power:

$$\eta = \text{multiple access capacity/single access capacity at transponder saturation}$$

The capacity of a carrier is equal to the information bit rate R_b conveyed by the carrier. This is sometimes called the carrier *throughput*. The *efficiency* of a multiple-access scheme then is the ratio of the sum of the throughputs of all accessing carriers to the maximum throughput of a single carrier in the transponder. Efficiency then appears as a *normalised throughput*.

6.5 FREQUENCY DIVISION MULTIPLE ACCESS (FDMA)

The bandwidth of a repeater channel is divided into sub-bands; each sub-band is assigned to one of the carriers transmitted by the earth stations. With this type of access, the earth stations transmit continuously, and the channel conveys several carriers simultaneously at different frequencies. It is necessary to provide guard intervals between each band occupied by a carrier to avoid interference as a result of imperfections of oscillators and filters. The receiver selects the required carrier in accordance with the appropriate frequency. The intermediate frequency (IF) amplifier provides the filtering.

Depending on the multiplexing and modulation techniques used, several transmission schemes can be considered. In each case, the satellite channel carries several carriers simultaneously, and the nonlinear transfer characteristic of the satellite channel is the cause of a major problem – that of *intermodulation* between the carriers. This can be avoided using regenerative satellites (Section 7.4.3.3).

6.5.1 TDM/PSK/FDMA

The baseband signals at the earth station are digital. They are combined to form a time division multiplexing (TDM) signal. The binary stream representing this multiplex signal modulates a carrier by phase shift keying (PSK) that accesses the satellite repeater channel at a particular frequency at the same time as other carriers on other frequencies from other stations. To minimise intermodulation products, and consequently the number of carriers (see Section 6.3.2), traffic routing is preferably performed according to the 'one carrier per transmitting station' principle. The TDM signal contains the total traffic from the transmitting earth station to all other stations. Figure 6.5 shows an example of a network of three stations.

6.5.2 SCPC/FDMA

The baseband signals at the earth station each modulate a carrier individually. This is called *single connection per carrier* (SCPC). Each carrier accesses the satellite repeater channel on its particular frequency at the same time as other carriers on other frequencies from the same or other stations. Information routing is thus performed according to the 'one carrier per link' principle.

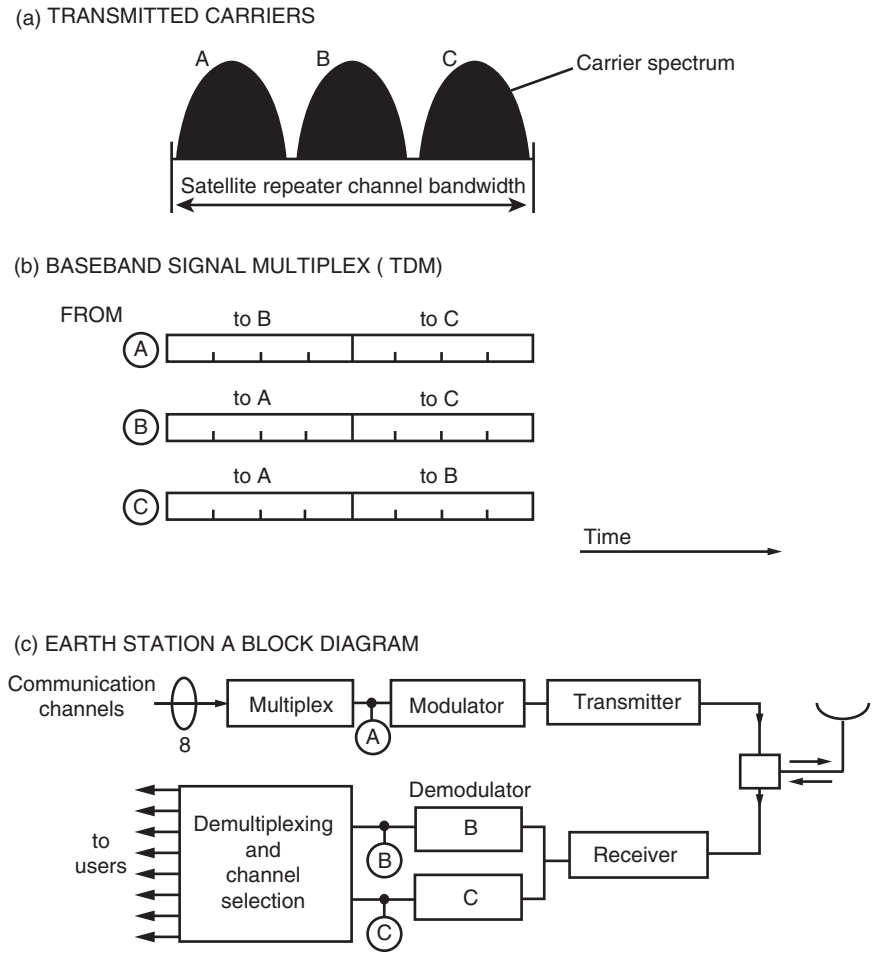


Figure 6.5 A three-station FDMA system with 'one carrier per transmitting station' routing.

6.5.3 Adjacent channel interference

As shown in Figure 6.6, the channel bandwidth is occupied by several carriers at different frequencies. The channel transmits these to all the earth stations situated in the coverage area of the satellite antenna. The carriers must be filtered by the receiver at each earth station, and this filtering is easier to realise when the carrier spectra are separated from each other by a wide frequency guard band. However, the use of wide guard bands leads to inefficient use of the channel bandwidth and a higher operating cost, per carrier, of the space segment. There is, therefore, a technical and economic compromise to be made. Whatever the compromise chosen, part of the power of a carrier adjacent to a given carrier will be captured by the receiver tuned to the frequency of the carrier considered. This causes noise due to interference, called *adjacent channel interference* (ACI). This interference is additional to the interference between systems analysed in Section 5.9.2 and can be included in the term $(C/N_0)_I$ that appears in expression (5.73) for $(C/N_0)_T$.

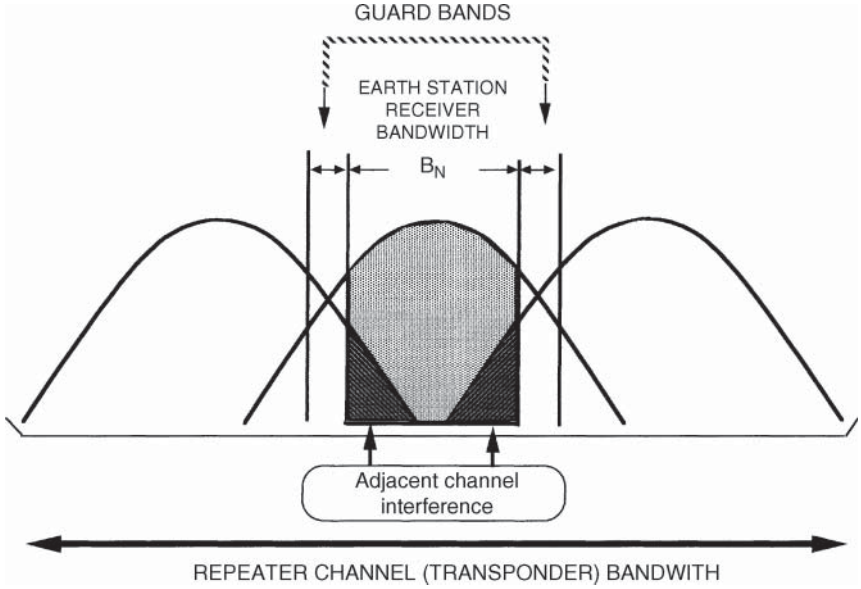


Figure 6.6 The spectrum of FDMA carriers and adjacent channel interference.

6.5.4 Intermodulation

6.5.4.1 Definition of intermodulation products

It was seen in Section 5.9.1 that a satellite repeater channel has a nonlinear transfer characteristic. By the nature of FDMA, this repeater channel simultaneously amplifies several carriers at different frequencies. The earth station itself has a nonlinear power amplifier, and this amplifier can be fed by several carriers at different frequencies. In general, when N sinusoidal signals at frequencies $f_1, f_2, \ldots, f_N$ pass through a nonlinear amplifier, the output contains not only the N signals at the original frequencies but also undesirable signals called *intermodulation products*. These appear at frequencies f_{IM} that are linear combinations of the input frequencies thus:

$$f_{IM} = m_1 f_1 + m_2 f_2 + .. \ \ldots \ + m_N f_N \tag{6.6}$$

where $m_1, m_2, \ \ldots \ , m_N$ are positive or negative integers.

The quantity X is called the *order* of an intermodulation product such that:

$$X = |m_1| + |m_2| + .. \ \ldots \ + |m_N| \tag{6.7}$$

When the centre frequency of the passband amplifier is large compared with its bandwidth, which is the case for a satellite repeater channel (compare the centre frequency of several GHz to the bandwidth of a few tens of MHz), only the odd order intermodulation products, where $\Sigma mi = 1$, fall within the amplifier bandwidth. Moreover, the amplitude of the intermodulation products decreases with the order of the product. Hence, in practice, only products of order 3, and to a lesser extent 5, are significant. Figure 6.7 shows the generation of intermodulation products from two unmodulated carriers at frequencies f_1 and f_2. It can be seen that, in the case of

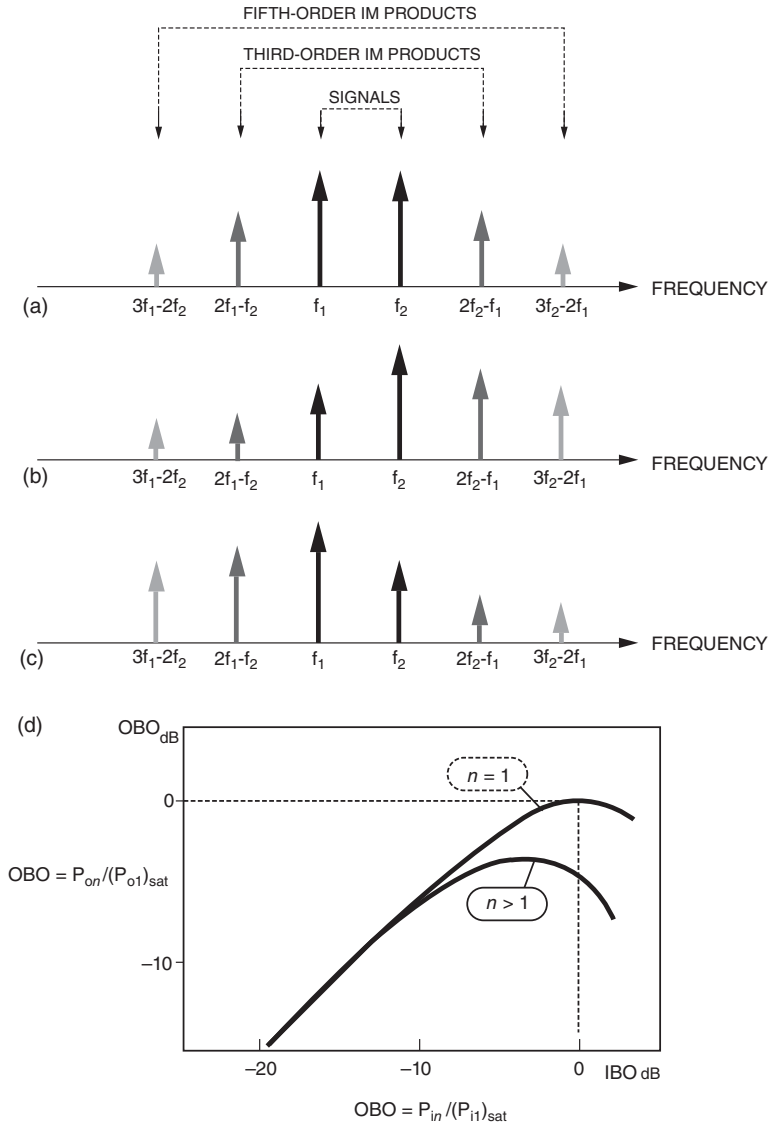


Figure 6.7 Intermodulation products for the case of two sinusoidal signals (unmodulated carriers): (a) equal amplitudes; (b) and (c) unequal amplitudes; (d) transfer characteristic of a nonlinear amplifier in single ($n = 1$) and multicarrier ($n > 1$) operation (IBO = input back-off, OBO = output back-off).

unmodulated carriers of unequal amplitude, the intermodulation products are greater at high frequencies if the carrier of greater amplitude is that which has the higher frequency and at low frequencies if the carrier of greater amplitude is that which has the lower frequency. This indicates the advantage of locating the most powerful carriers at the extremities of the channel bandwidth, as the corresponding most powerful intermodulation products then fall out of channel bandwidth and do not propagate on the downlink.

6.5.4.2 *Transfer characteristic of a nonlinear amplifier in multicarrier operation*

Figure 6.7d shows the power-transfer characteristic of a satellite repeater channel in single-carrier operation. In general, the form of this characteristic is valid for every nonlinear amplifier. It is now necessary to extend this model to the case of multicarrier operation. For this, the following notation is used:

(P_{i1}) = carrier power at the amplifier input (i = input) in single-carrier operation ($n = 1$ carrier)
(P_{in}) = power of one carrier (among n) at the amplifier input in multicarrier operation (n carriers)
(P_{o1}) = carrier power at the amplifier output (o = output) in single-carrier operation ($n = 1$ carrier)
(P_{on}) = power of one carrier (among n) at the amplifier output in multicarrier operation (n carriers)
($P_{IMX,n}$) = power of intermodulation product of order X at the amplifier output in multicarrier operation (n carriers)

The definition of input back-off (IBO) and output back-off (OBO), given in Section 5.9.1 for the case of single-carrier operation, is generalised to the case of multicarrier operation as follows:

$$\text{Input back-off of one carrier among } n \text{ : } \mathrm{IBO}_1 = (\mathrm{P}_{in})/(\mathrm{P}_{i1})_{sat}$$

$$\text{Output back-off of one carrier among } n \text{ : } \mathrm{OBO}_1 = (\mathrm{P}_{on})/(\mathrm{P}_{o1})_{sat}$$

In the previous expressions, the subscript *sat* indicates the value of the quantity considered at saturation. Figure 6.7d shows a typical variation of OBO as a function of IBO for single-carrier and multicarrier operation of a transponder.

6.5.4.3 *Intermodulation noise*

When the carriers are modulated, the intermodulation products are no longer spectral linear since their power is dispersed over a spectrum that extends over a band of frequencies [GAG-91]. If the number of carriers is sufficiently high, superposition of the spectra of the intermodulation products leads to a spectral density that is nearly constant over the whole of the amplifier bandwidth, and this justifies considering intermodulation products as white noise (noise with constant power spectral density denoted $(N_0)_{IM}$).

6.5.4.4 *The carrier-to-intermodulation noise spectral density power ratio $(C/N_0)_{IM}$*

The power spectral density of intermodulation noise, $(N_0)_{IM}$, depends on the transfer characteristic of the amplifier and the number and type of carriers amplified (Section 9.2). A carrier power-to-intermodulation noise power spectral density ratio $(C/N_0)_{IM}$ can be associated with each carrier at the amplifier output. This ratio can be deduced from an amplifier characteristic of the type given in Figure 6.7d by estimating, for example, $(N_0)_{IM}$ as $(P_{IMX,n})/B$, where B is the bandwidth of the modulated carrier. Hence:

$$(C/N_0)_{IM} = (P_{on}/P_{IMX,n})B$$

Figure 6.8 shows a typical variation of $(C/N_0)_{IM}$ as a function of IBO and the number of carriers. It can be seen that the ratio $(C/N_0)_{IM}$ becomes smaller near saturation (the nonlinear characteristic is more severe) and as the number of carriers increases (a combined effect of a smaller share per carrier of the total power available from the repeater channel, and an increase of the total power of the intermodulation products).

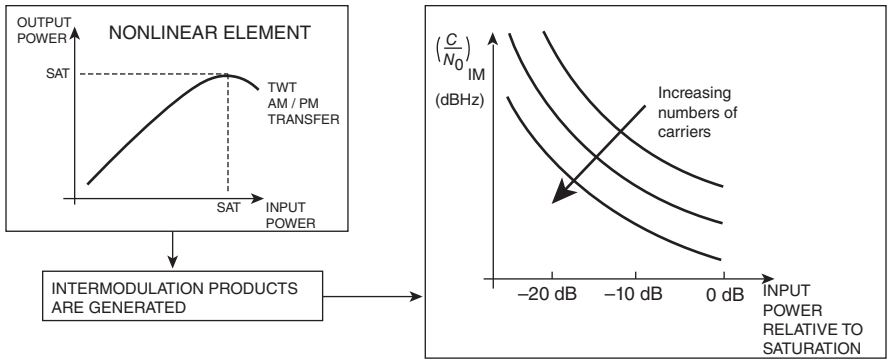


Figure 6.8 Variation of $(C/N_0)_{IM}$ as a function of back-off and number of carriers.

6.5.5 FDMA efficiency

It can be seen from Figure 5.35 that the value of $(C/N_0)_T$ is always less than the value obtained in single-carrier operation at saturation. On the other hand, the maximum value of $(C/N_0)_T$ becomes less as the back-off is increased, and this is the case when the number of carriers increases. Figure 6.9 shows the relative variation of the total capacity of a satellite repeater channel of 36 MHz bandwidth. The transmission scheme is of the TDM/PSK/FDMA type using quadrature phase shift keying (QPSK) modulation. The carriers have an equal share of the bandwidth and the power of the satellite transponder. As the number of carriers increases, the power available to each carrier reduces, and this implies use of forward error correction (FEC) schemes to maintain the target bit error rate (BER) at the demodulator output of each carrier. The throughput of each carrier decreases, and so does the total throughput, which is the sum of the throughputs of the individual carriers. Figure 6.9 thus represents the efficiency of an FDMA system as a function of the number of accesses; the transponder bandwidth is 36 MHz, and this corresponds to a single access capacity using QPSK of 54 Mbps. With 10 individual FDMA carriers, the efficiency is $\eta = 40\%$ and the total throughput is 0.4×54 Mbps = 21.6 Mbps.

6.5.6 Conclusion

FDMA is characterised by continuous access to the satellite in a given frequency band. This technique has the advantage of simplicity. However, it has some disadvantages:

— Lack of flexibility in case of reconfiguration. To accommodate capacity variations, it is necessary to change the frequency plan, and this implies modification of transmitting frequencies, receiving frequencies, and filter bandwidths of the earth stations.
— Loss of capacity when the number of accesses increases due to the generation of intermodulation products, and the need to operate at a reduced satellite transmitting power (back-off).
— The need to control the transmitting power of earth stations in such a way that the carrier powers at the satellite input are the same in order to avoid the capture effect. This control must be performed in real time and must adapt to attenuation caused by rain on the uplinks.

In telecommunications, the capture effect or frequency modulation (FM) capture is a phenomenon associated with FM reception in which only the stronger of two signals as, or near, the same frequency or channel will be demodulated.

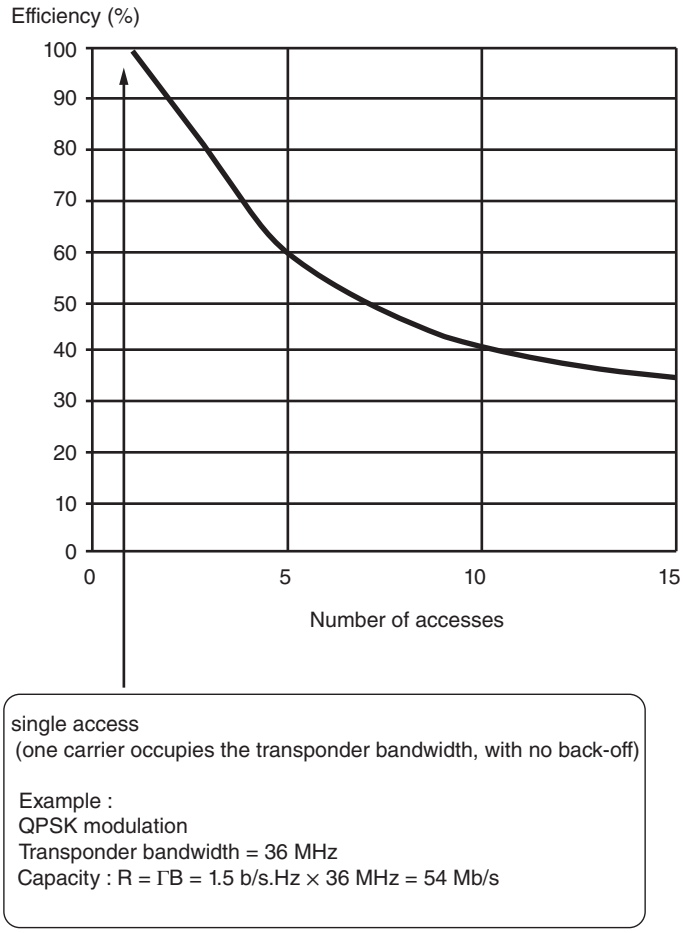


Figure 6.9 Efficiency of an FDMA transmission; the curve indicates the relative variation of the total throughput of a transponder with a bandwidth of 36 MHz as a function of the number of accesses, which is the number of TDM/QPSK/FDMA carriers. The value indicated as 100% represents the total capacity (54 Mbps) of the multiplex, which modulates the carrier for the case of single access to the repeater channel, operated at saturation.

FDMA is the oldest access technique, and it remains widely used despite the disadvantages. It tends to perpetuate itself due to investments made in the past and its known operational advantages, which include the absence of synchronisation between earth stations.

6.6 TIME DIVISION MULTIPLE ACCESS (TDMA)

Figure 6.10 shows the operation of TDMA. The earth stations transmit one after another *bursts* of carrier with duration T_B. All bursts of carrier have the same frequency and occupy the full repeater channel bandwidth. Hence the satellite repeater channel carries one carrier at a time. Bursts are inserted within a periodic time structure of duration T_F, called a *frame*.

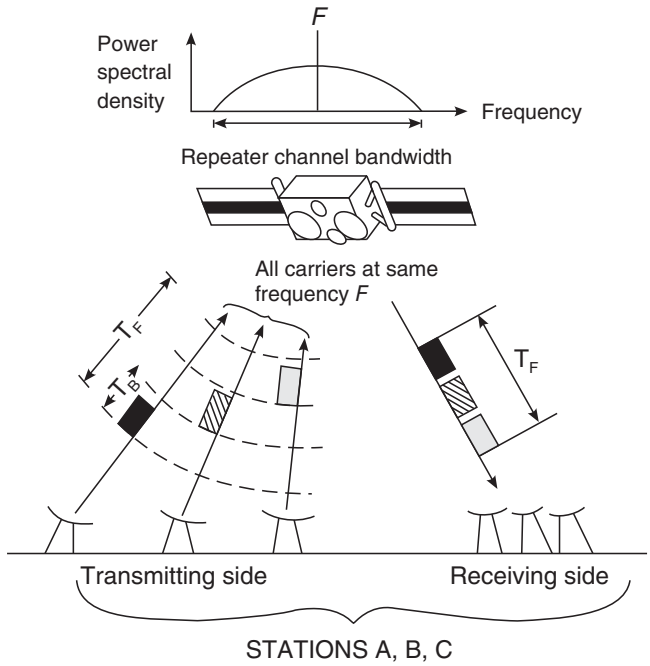


Figure 6.10 Operation of time division multiple access (TDMA).

6.6.1 Burst generation

The burst corresponds to the transfer of traffic from the station considered. With the technique of one carrier per station-to-station link, introduced in Section 6.3.1, the station transmits $N-1$ bursts per frame, where N is the number of stations on the network and the number of bursts P in the frame is given by $P = N(N-1)$. With the technique of one carrier per transmitting station, introduced in Section 6.3.2, the station transmits a single burst per frame with traffic from the station to all others in the network, and the number of bursts P in the frame is equal to N. Each burst thus contains several sub-bursts of traffic from station to station. Due to the decrease of throughput of the channel as the number of bursts increases (see Section 6.6.5) the 'one carrier per transmitting station' technique is generally retained.

Figure 6.11 illustrates burst generation. The earth station receives traffic in the form of a continuous binary stream of rate R_b from the network or user interface. This information must be stored in a buffer memory while waiting for the burst transmission time. When this time arises, the contents of the memory are transmitted in a time interval equal to T_B. The bit rate R that modulates the carrier is thus given by:

$$R = R_b(T_F/T_B) \quad \text{(bit/s)} \tag{6.8}$$

The value of R is high when the burst duration is short, and consequently the transmission duty cycle (T_B/T_F) of the station is low. Hence, for example, if $R_b = 2$ Mbps and $(T_F/T_B) = 10$, modulation occurs at 20 Mbps. Notice that R represents the total capacity of the network; that is the sum of the station capacities in bit/s. If all stations have the same capacity, the duty cycle (T_F/T_B) represents the number of stations on the network.

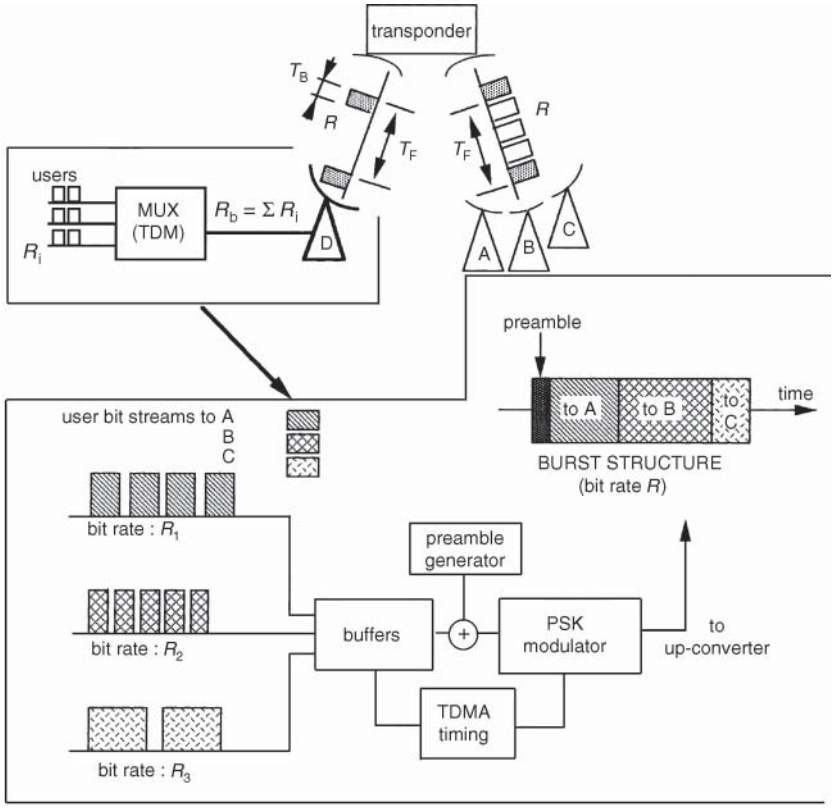


Figure 6.11 Burst generation with the 'one carrier per transmitting station' technique: R_i = user rate (bit/s), R_b = information rate of the multiplex (bit/s) = ΣR_i, R = rate in each burst (bit/s), T_B = burst duration (s), T_F = frame duration (s).

The structure of a burst can be seen in Figure 6.11 and is further detailed in Figure 6.12. It consists of a header, or preamble, and a traffic field. The header permits:

— The demodulator of the receiving earth station, in the case of coherent demodulation, to synchronise its local oscillator to the received carrier. The first part of the header is a bit sequence that provides a constant carrier phase for rapid carrier recovery.
— The detector of the receiving earth station to synchronise its bit-decision clock to the symbol rate. The second part of the header is a bit sequence providing alternating opposite phases.
— The earth station to identify the start of a burst by detecting, by means of a correlator, a group of bits called a *unique word* (UW). The third part of the header is the unique word that enables the receiver to resolve carrier phase ambiguity in the case of coherent demodulation.
— The transfer of service messages between stations (telephone and data) and signalling.

Knowing the start of the burst and the bit rate, and having (if required) resolved the phase ambiguity, the receiver can then identify all the bits occurring after the unique word.

The traffic field is located at the end of the header, and this corresponds to the transmission of useful information. In the case of the 'one carrier per transmitting station' technique, the traffic

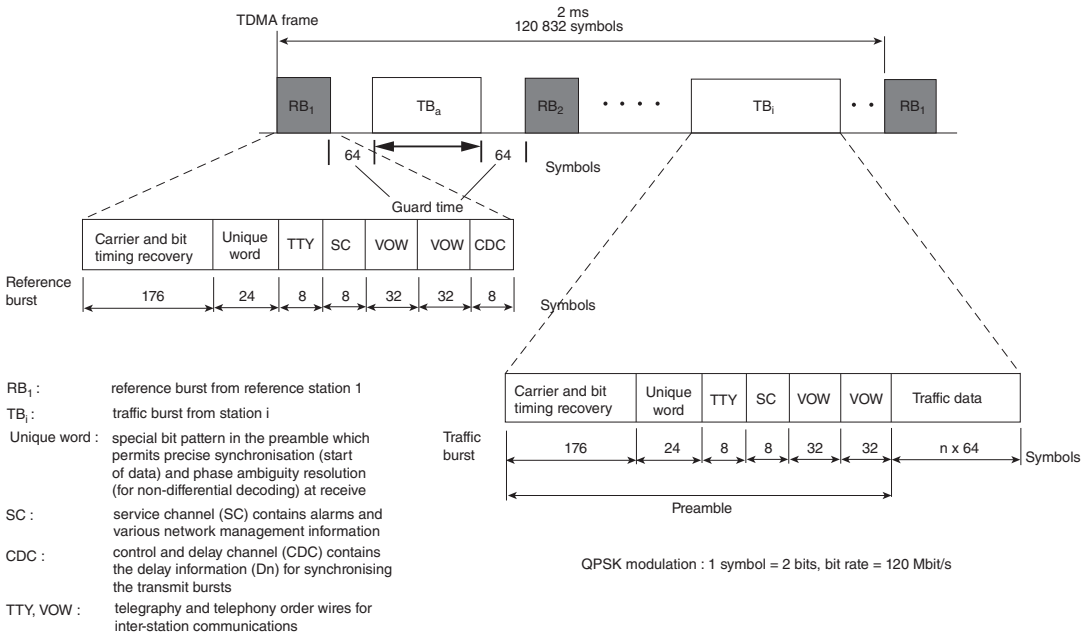


Figure 6.12 Frame structure (Intelsat/Eutelsat standard). Source: reprinted from CCIR-Rep 88 with the permission of the ITU.

field is structured in sub-bursts that correspond to the information transmitted by the station to each of the other stations.

6.6.2 Frame structure

Figure 6.12 shows the frame as an example used in the Intelsat and Eutelsat networks and provides detailed insight into the burst structure. The length of the frame is 2 ms. The frame is formed at satellite level. It consists of all the bursts transmitted by the earth stations placed one after the other, if transmission synchronisation of the stations is correct. To take into account synchronisation imperfections, a period without transmission, called a *guard time*, is provided between each burst. In Figure 6.12, the guard time occupies 64 symbols or 128 bits, and this corresponds to a time interval of 1 microsecond. Notice the presence of two types of burst:

— Those of *traffic stations*, with a header of 280 symbols, or 560 bits, and a traffic field structured in multiples of 64 symbols in accordance with the capacity of each station.
— Those of *reference stations*, with a header of 288 symbols, or 576 bits, and without a traffic field. The reference station is the station that defines the frame clock by transmitting its reference burst; all the network traffic stations must synchronise themselves to the reference station by locating their burst with a constant delay with respect to the reference station burst, called the *reference burst*. Because of its fundamental role in correct operation of the network, the reference station is replicated. This is why there are two reference bursts per frame with identical contents; each one is transmitted by each of the two mutually synchronised reference stations.

6.6.3 Burst reception

On the downlink, each station receives all bursts in the frame. Figure 6.13 illustrates the processing at the receiving station.

The receiving station identifies the start of each burst of the frame by detection of the unique word; it then extracts the traffic that is intended for it and is contained in a sub-burst of the traffic field of each burst. This traffic is received discontinuously at bit rate R. To restore the original bit rate R_b in the form of a continuous binary stream, the bits after demodulation are stored in a buffer memory for one frame period and are read out at a rate R_b during the following frame.

It is fundamental for identification of the burst contents that the receiving station must be able to detect the unique word at the start of each burst. The unique word detector establishes correlation between each bit sequence at the output of the receiver bit detector that is of the same length as the unique word and a replica of the unique word stored in the correlator memory. Only received sequences that produce a correlation peak greater than a threshold are retained as unique words. The performance of the unique word detector is measured by two quantities [FEH-83]:

— The probability of non-detection, which is the probability of not detecting the presence of a unique word at the start of burst reception.
— The probability of a false alarm, which is the probability of falsely identifying the unique word in any binary sequence, for example in the traffic field.

The probability of non-detection decreases when:

— The BER of the link decreases

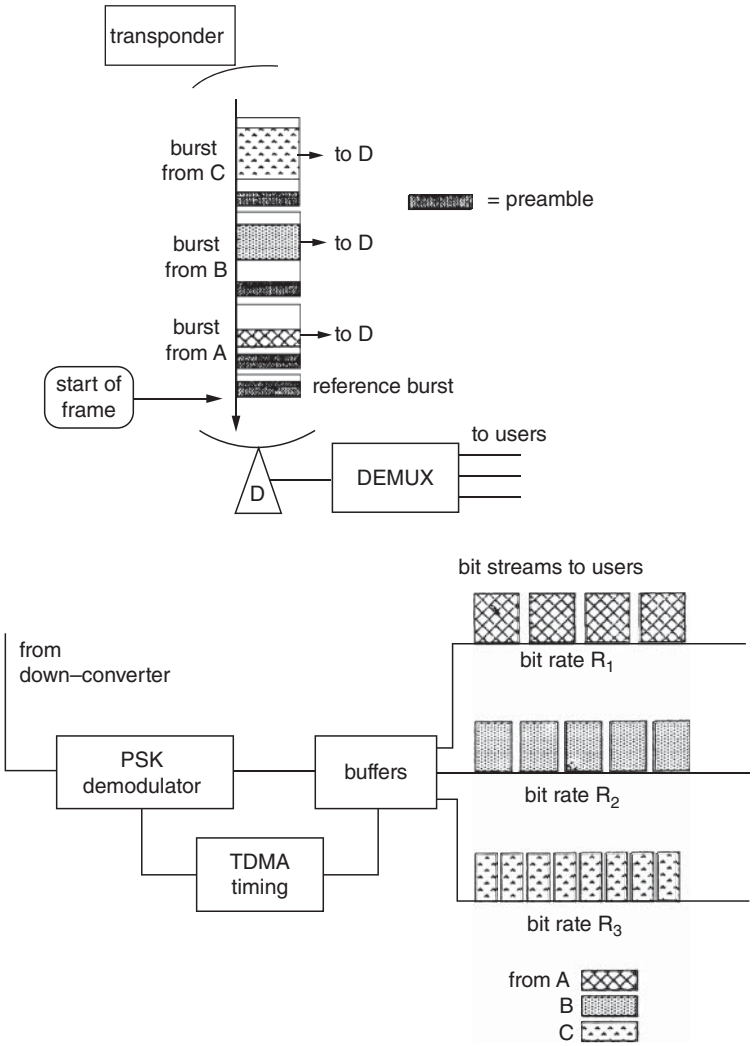


Figure 6.13 Burst reception.

— The length of the unique word decreases
— The correlation threshold decreases

The probability of a false alarm is independent of the BER on the link and decreases when:

— The length of the unique word increases
— The correlation threshold increases

A compromise must, therefore, be found; in practice, the probability of a false alarm is reduced without increasing the probability of non-detection by taking advantage of a priori knowledge

of the frame structure in order to perform correlation only in the time intervals when the unique word is expected.

6.6.4 Synchronisation

Synchronisation between the different stations in the network is necessary to avoid burst overlap from others in the frame. Such an overlap would generate a level of interference that would prevent the earth station receiver from detecting bits properly. Before discussing the synchronisation techniques, it is important to establish the order of magnitude of the disturbances associated with the imperfections of the geostationary satellite orbit.

6.6.4.1 Residual movements of a geostationary satellite

The orbit control of the satellite defines a station-keeping *box* whose typical dimensions are 0.1° in longitude and latitude. Furthermore, the eccentricity of the orbit is limited to a maximum value on the order of 4×10^{-4}. The satellite thus moves, as indicated in Figure 6.14, in a box on the order of 75 km × 75 km × 35 km. This introduces an altitude variation of around 35 km with a periodicity of 24 hours, which has two effects:

— *A variation in round-trip propagation time of around 250 μs*: This quantifies the magnitude of potential daily displacement of a burst in the frame in the absence of corrective action. This value is to be compared with the frame duration (from 2 to 20 ms).
— *A Doppler effect*: If the maximum displacement velocity of the satellite is considered to be 10 km h^{-1}, the Doppler effect causes displacement of the position of a burst in the frame from one station at a rate of around 20 ns s^{-1}. With a guard time between two bursts of 1 μs, and assuming the particular case of displacement in opposite directions of two consecutive bursts in the frame, the time for the drift to absorb the guard time between the two bursts is on the order of $(1/2)(1 \times 10^{-6}/20 \times 10^{-9})$s = 25 seconds. This determines the timescale for undertaking corrective action. Notice that this time is greater than the round-trip propagation time

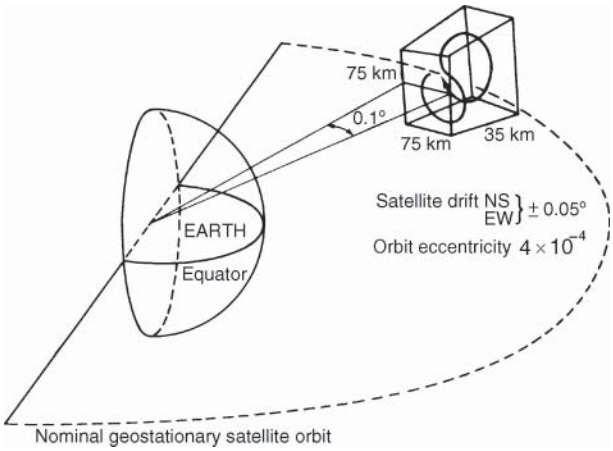


Figure 6.14 Evolution of the volume occupied by a geostationary satellite in the course of an orbital period (24 hours).

of the bursts and indicates that control of the burst position can be based on observation of position error.

6.6.4.2 Relationship between the start of a frame on transmission and reception

Any station $n(n = 1, 2, \ldots, N)$ must transmit its burst in such a way that it arrives at the satellite with a delay d_n with respect to the reference burst. As shown in Figure 6.15, the value of the delay d_n is specific for each station. The set of values of d_n determines the arrangement of bursts in the frame, called the *burst time plan*. Positioning is correct when station n transmits with a delay d_n with respect to the time of the *start of transmit frame* ($SOTF_n$). This time $SOTF_n$ is the instant at which the station should transmit in order to position its burst in the frame time slot occupied by the reference burst. The problem of synchronising station n is thus that of determining $SOTF_n$. Once this instant is known, it is merely necessary for station n to transmit with a delay d_n with respect to $SOTF_n$.

With a single-beam satellite, station n receives all of the frame on the downlink. Detection of the unique word of the reference burst determines the start time of the received frame, *start of*

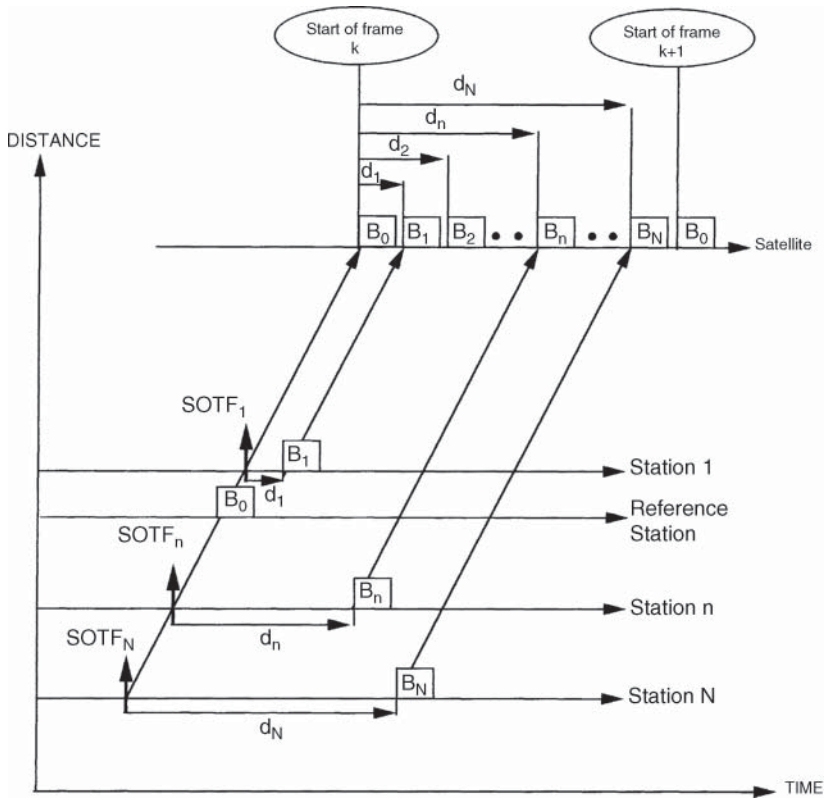


Figure 6.15 Burst-time plan within the frame: each station n locates its burst at satellite level with a delay d_n $(n = 1, 2, \ldots N)$ with respect to the reference burst B_0 that defines the start of the frame. The vertical arrow at station n indicates the start of the transmit frame ($SOTF_n$) for this station.

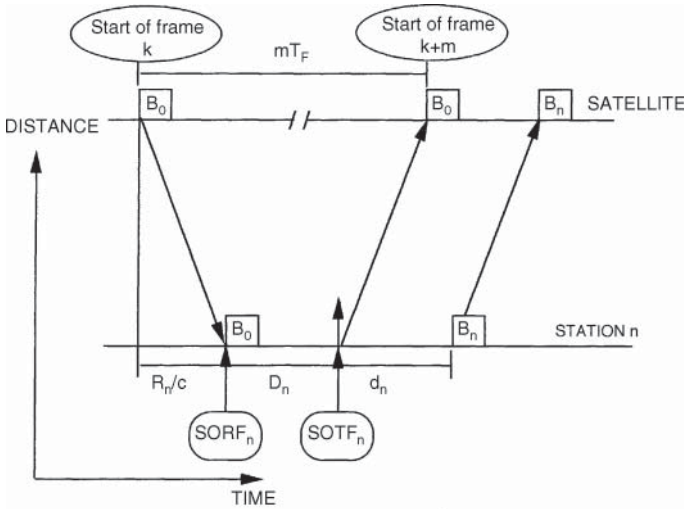


Figure 6.16 The relationship between the frame start times on transmission $SOTF_n$ and on reception $SORF_n$ for a station n.

receive frame ($SORF_n$). Figure 6.16 shows the time relationship between the $SOTF_n$ and the $SORF_n$; $SORF_n$ is equal to the start time of the frame (k) at the satellite plus the propagation time on the downlink R_n/c, where R_n is the distance of the satellite from ground station n and c is the velocity of light. The start time of frame ($k+m$), where m is an integer, is equal to $SOTF_n$ plus the propagation time on the uplink R_n/c. The time separating the start of frame (k) and frame ($k+m$) at the satellite is, by definition, mT_F. Hence the relationship:

$$SOTF_n - SORF_n = D_n = mT_F - 2R_n/c \quad \text{(s)} \tag{6.9}$$

For this quantity to be positive, it is necessary to choose m such that mT_F is greater than the value of $2R_n/c$ for station n that is furthest from the satellite. For example, if the value of m is taken as 14 and the frame duration $T_F = 20$ ms, the maximum roundtrip propagation time for the most distant station should not exceed 280 ms.

In summary, station n identifies $SORF_n$ by detecting the unique word of the reference burst and transmits at an instant $D_n + d_n$ later. Depending on the method of determining the value of D_n, two synchronisation techniques can be distinguished:

— Closed-loop synchronisation
  Open-loop synchronisation

6.6.4.3 Closed-loop synchronisation

Figure 6.17 illustrates this method. Station n observes the position of its burst in the frame relative to the reference burst by measuring the time between detection of the unique word of the reference burst and detection of the unique word of its own burst. Let $d_{on}(j)$ be the value observed on reception of the frame for which the value $D_n(j)$ had been used to determine the transmission time. The difference $e_n(j) = d_{on}(j) - d_n$ is the *burst position error*. The station then increases the value

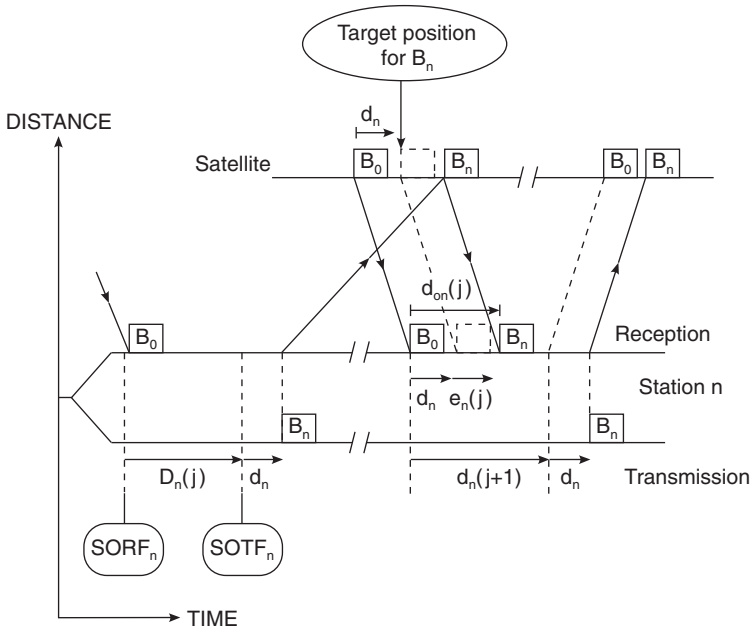


Figure 6.17 Closed-loop synchronisation: station *n* observes the position of its burst and consequently corrects the transmission time.

of D_n according to the following algorithm:

$$D_n(j+1) = D_n(j) - e_n(j) \quad \text{(s)} \tag{6.10}$$

and uses the new value of D_n to determine the transmission time. Notice that the minimum time necessary to make a correction is equal to the round-trip propagation time for the station furthest from the satellite, which is on the order of 280 ms.

6.6.4.4 Open-loop synchronisation

This is used particularly for networks with assignment on demand where the burst position of traffic stations is controlled by the reference station (see Section 6.8); this method relies on knowledge of the satellite position and calculation of the distance R_n between the satellite and each ground station. The satellite position can be provided by the orbit control station (space segment). If decoupling of responsibility between the space segment and the ground segment is required, two auxiliary stations must be provided in addition to the reference station. Figure 6.18 illustrates the technique; the two auxiliary ranging stations B and C and the reference station A measure the propagation time of their bursts. The two auxiliary ranging stations communicate these values to the reference station, which determines the satellite position by triangulation and calculates the distance of the satellite from each network station. The reference station transmits sequentially to all traffic stations *n* the value of D_n calculated from Eq. (6.9) using the control and delay channel (CDC) as signalling channel of the reference burst (see Figure 6.12). Notice that the time before correction is equal to the time required to measure the propagation time (one round

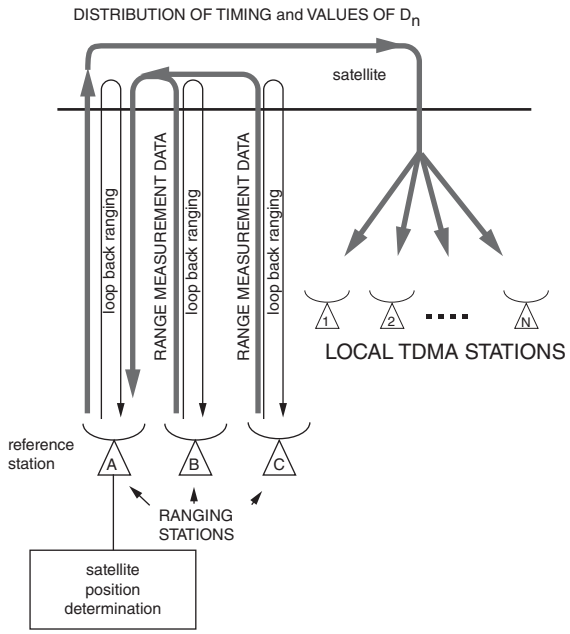


Figure 6.18 Open-loop synchronisation.

trip) plus the time required for transmission of this information by the two auxiliary stations to the reference station (one round trip) plus the calculation time and the time to distribute the values of D_n. This time can amount to several seconds and consequently implies longer guard times than in the case of closed-loop synchronisation.

6.6.4.5 Acquisition of synchronisation

Acquisition of synchronisation by a station is achieved each time the station wishes to enter the network. Operation can be in a closed or open loop.

In closed-loop synchronisation, the station transmits a low-power burst, generally modulated by a pseudorandom sequence, observes its position, corrects it to give its nominal position, and then operates on full power to transmit the useful information. Modulation by a pseudorandom sequence facilitates acquisition by virtue of the autocorrelation properties of the sequence, which permit measurement of the position error, and energy dispersion that limits interference by the entering station onto bursts transmitted from stations that are active.

In open-loop synchronisation, the entering station receives the value of D_n from the reference station and transmits at an instant $D_n + d_n$ after receiving the reference burst.

6.6.5 TDMA efficiency

In single-carrier operation, the maximum throughput is $R = B\Gamma$, where B is the channel bandwidth (Hz) and Γ is the spectral efficiency of the modulation (bit/s Hz). With multiple access, the actual total throughput is $R - (1 - \Sigma t_i / T_F)$, where Σt_i represents the sum of the times not devoted

to transmission of traffic (guard times plus burst headers). The TDMA efficiency is thus:

$$\eta = 1 - \sum t_i / T_F \tag{6.11}$$

The efficiency is greater when the frame duration T_F is high and when Σt_i is small.

The efficiency depends on the number P of bursts in the frame. Let p be the number of bits in the header and g the equivalent duration in bits of the guard time. Assuming that the frame contains two reference bursts, this gives:

$$\eta = 1 - (P + 2)(p + g)/RT_F \tag{6.12}$$

where R is the bit rate of the frame (bit/s).

The efficiency as a function of the number of accesses, which is the number of stations N on the network, depends on the traffic routing technique adopted. It is known (Section 6.6.1) that:

— In the case of a 'one carrier per link' technique, $P = N(N - 1)$.
— In the 'one carrier per transmitting station' routing technique, $P = N$.

Since efficiency is low when P is high, the advantage of adopting the latter technique is obvious.

6.6.5.1 Frame duration considerations

A long frame duration requires a higher storage capacity in the transmitting and receiving earth station buffer memories. Furthermore, the frame duration conditions information delivery delay from one terrestrial network-station interface to another station–terrestrial network interface; indeed, this delay is equal to the round-trip propagation time increased by the transmission and reception storage time. Since the storage time is at most equal to the duration of one frame, this gives:

$$\text{Interface-to-interface delivery delay} = \text{Round-trip propagation time} + 2T_F \quad \text{(s)} \tag{6.13}$$

For telephone transmission, it is considered that the propagation time *between subscribers* must not exceed 400 ms. Considering that the round-trip propagation time of radio waves does not exceed 278 ms and allowing 30 ms for the sum of the propagation times in the terrestrial tails, then T_F should be kept lower than:

$$T_F \leq \frac{1}{2}(400 - 278 - 30) = 46 \text{ ms} \tag{6.14}$$

In practice, frame durations typically range from 750 µs to 20 ms.

6.6.5.2 Guard time and header considerations

For a given frame duration, the efficiency increases as Σt_i is decreased. This implies:

— *A reduction of the guard times*: This approach is limited by the precision of the synchronisation method. A closed-loop method is preferable to an open-loop one in this respect.
— *A reduction of the headers*: It is important to provide circuits in the receivers for rapid carrier and bit-timing recovery. One can also consider designing the link for differential demodulation

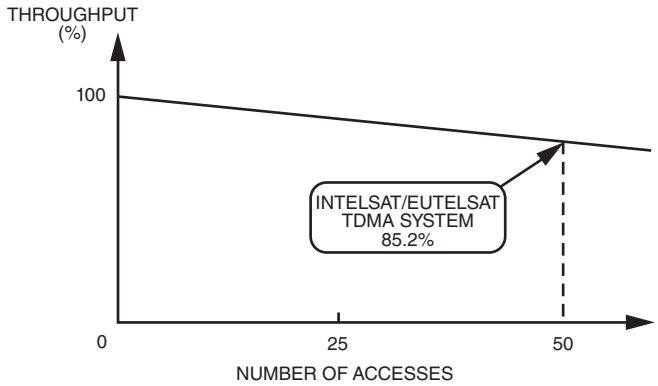


Figure 6.19 The efficiency of the Intelsat/Eutelsat TDMA system; the 100% value indicated for a single access corresponds to the capacity of the single carrier that passes through the transponder and is transmitted continuously.

instead of coherent demodulation, at the expense of an increased bit-error probability. Finally, one can attempt to reduce the duration of the unique word, but this involves an increase in the probability of a false alarm in detection of the unique word (Section 6.6.3).

Example 6.1 The variation of throughput as a function of the number of bursts of traffic P, equal to the number N of traffic stations or the number of accesses, can be examined by inserting the values of the Intelsat/Eutelsat standard indicated in Figure 6.12 into Eq. (6.12). Taking $p = 560$, $g = 128$, $R = 120.832$ Mbps, and $T_F = 2$ ms gives:

$$\eta = 1 - 2.85 \times 10^{-3} (P + 2) \tag{6.15}$$

This expression is represented by Figure 6.19. Notice the relatively slow decrease (in comparison to that of Figure 6.9 for FDMA) of efficiency as a function of the number of accesses. For instance, for 50 accesses, the efficiency is still 85%.

6.6.6 Conclusion

TDMA is characterised by access to the channel during a time slot. This has certain advantages:

— At each instant, the satellite repeater channel amplifies only a single carrier that occupies all of the repeater channel bandwidth; there are no intermodulation products, and the carrier benefits from the saturation power of the channel.
— TDMA efficiency remains high for a large number of accesses.
— There is no need to control the transmitting power of the stations.
— All stations transmit and receive on the same frequency whatever the origin or destination of the burst; this simplifies tuning.

TDMA, however, has certain disadvantages:

— The need for synchronisation implies complex procedures and the provision of two reference stations. Fortunately, these procedures can be automated and computer driven.

— The need to increase power and bandwidth as a result of high burst bit rate, compared to continuous access, as with FDMA, for instance.

Consider a station-to-station link. The quality objective is specified in terms of error probability. The imposed value determines the required value of the ratio E/N_0. The ratio C/N_0 for the overall link is determined by the relation established in Chapter 3 and recalled here:

$$C/N_0 = (E/N_0)R \tag{6.16}$$

It can be seen that C/N_0 is proportional to R, for which the expression is given by Eq. (6.8). For a capacity R_b, a station must be dimensioned in power and bandwidth to transmit a bit rate R that is high when the duty cycle T_B/T_F is low, as shown by Eq. (6.8). (In FDMA, the station transmits at bit rate R_b, and consequently the required C/N_0 is smaller.) This disadvantage of TDMA is partly compensated for by the higher power provided by the satellite repeater channel on the downlink compared with the FDMA case, where back-off is necessary.

Overall, TDMA provides better utilisation of the space segment due to the higher efficiency in the case of a large number of accesses. Furthermore, TDMA can be computer driven, with the burst time plan calculated by software and changed automatically according to the reported traffic load in the earth station. So demand assignment (discussed in Section 6.8) is easily implemented.

6.7 CODE DIVISION MULTIPLE ACCESS (CDMA)

With CDMA, network stations transmit continuously and together on the same frequency band of the satellite repeater channel. There is, therefore, interference between the transmissions of different stations, and this interference is resolved by the receiver, which identifies the *signature* of each transmitter; the signature consists of a binary sequence, called a *code,* which is combined with the useful information at each transmitter. The set of codes used must have the following properties:

— Each code must be easily distinguishable from a replica of itself shifted in time.
— Each code must be easily distinguishable regardless of other codes used on the network.

Transmission of the code combined with the useful information requires the availability of a greater radio-frequency bandwidth than that required to transmit the information alone using the techniques described in Chapter 4. This is the reason one refers to *spread spectrum* transmission. Two techniques are used in CDMA:

— Direct sequence (DS)
— Frequency hopping (FH)

6.7.1 Direct sequence (DS-CDMA)

6.7.1.1 The principle

Figure 6.20 illustrates the principle: the binary message to be transmitted, $m(t)$, of bit rate $R_b = 1/T_b$, is coded in non-return to zero (NRZ) so that $m(t) = \pm 1$ and is multiplied by a binary sequence $p(t)$, itself coded in NRZ so that $p(t) = \pm 1$. The bit rate of the sequence $p(t)$ is $R_c = 1/T_c$

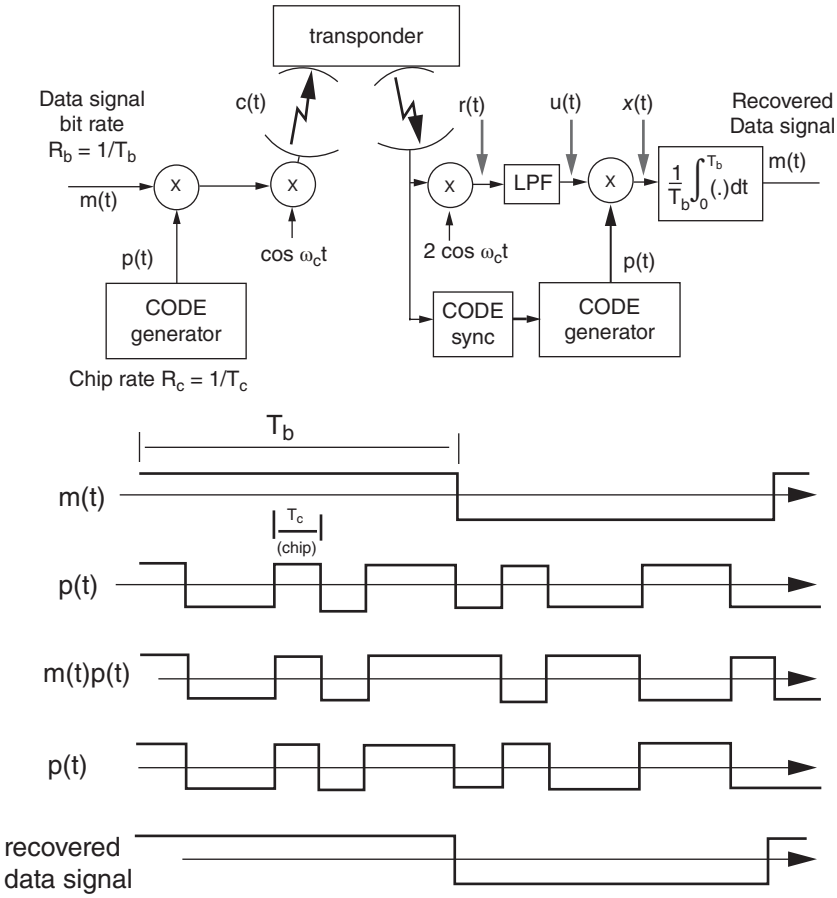


Figure 6.20 Direct sequence CDMA (DS-CDMA).

and greater (by 10^2–10^6) than the bit rate R_b. The binary element of the sequence $p(t)$ is called a *chip* in order to distinguish it from the binary element (bit) of the message, and therefore R_c is called the *chip rate*. The composite signal $m(t)p(t)$ then modulates a carrier by PSK (binary phase shift keying [BPSK], for example – see Section 4.2.1) whose frequency is common to all network stations. The transmitted carrier $c(t)$ can be expressed by:

$$c(t) = m(t)p(t)\ \cos\ \omega_c t \quad \text{(V)} \tag{6.17}$$

At the receiver, the signal is coherently demodulated by multiplying the received signal by a replica of the carrier. Neglecting thermal noise, the signal $r(t)$ at the input of the detector low-pass filter (LPF) is given by:

$$\begin{aligned} r(t) &= m(t)p(t)\cos\ \omega_c t(2\cos\ \omega_c t) \\ &= m(t)p(t) + m(t)p(t)\cos 2\,\omega_c t \qquad \text{(V)} \end{aligned} \tag{6.18}$$

The LPF eliminates the high-frequency components at $2\omega_c$ and retains only the low-frequency component $u(t) = m(t)p(t)$. This component is then multiplied by the local code $p(t)$ in phase with

the received code. In the product, $p(t)^2 = 1$. At the output of the multiplier this gives:

$$x(t) = m(t)p(t)p(t) = m(t)p(t)^2 = m(t) \qquad \text{(V)} \tag{6.19}$$

This signal is then integrated over one bit period to filter the noise. The transmitted message $m(t)$ is recovered at the integrator output.

6.7.1.2 *Spectral occupation*

The spectrum of the carrier $c(t)$, of power C and frequency F_c, is given by:

$$C(f) = \left(\frac{C}{R_c}\right) \left\{\frac{\sin[\pi(f - F_c)/R_c]}{\pi(f - F_c)/R_c}\right\}^2 \qquad \text{(W/Hz)} \tag{6.20}$$

In Figure 6.21, for comparison purposes, this spectrum is superimposed on that which the carrier would have if modulated by the message $m(t)$ alone. It can be seen that, with CDMA, $c(t)$ has a spectrum that is broadened by the spreading ratio R_c/R_b. This is the result of combining the message with the chip sequence. It will now be shown that this combination permits multiple access.

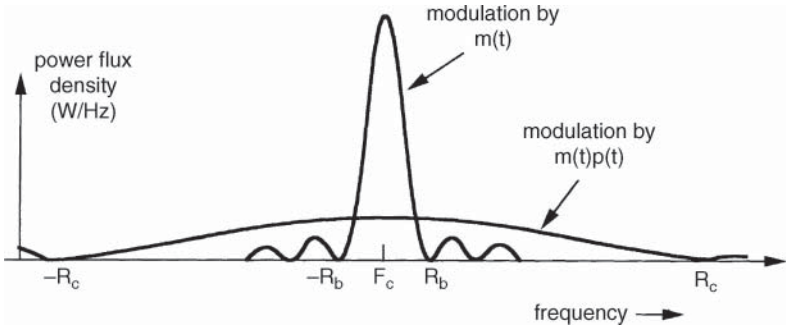


Figure 6.21 The spectrum of the carrier in DS-CDMA together with the spectrum that the carrier would have if modulated by the message $m(t)$.

6.7.1.3 *Realisation of multiple access*

The earth station receives from the channel the wanted carrier $c(t)$ superimposed on the carriers $c_i(t)(i = 1, 2, \ldots, N-1)$ of the $N-1$ other users transmitted on the same frequency; hence:

$$r(t) = c(t) + \sum c_i(t) \qquad \text{(V)} \tag{6.21}$$

with:

$$c(t) = m(t)p(t)\cos\ \omega_c t$$

$$\sum c_i(t) = \sum m_i(t)p_i(t)\cos\ \omega_c t$$

The multiplier output signal is given by:

$$x(t) = m(t)p(t)^2 + \sum m_i(t)p_i(t)p(t) = m(t) + \sum m_i(t)p_i(t)p(t) \qquad \text{(V)} \tag{6.22}$$

The message is now superimposed on noise due to interference. If care has been taken to choose codes with a low cross-correlation function, this noise will be small. Multiplication of $\Sigma m_i(t)p_i(t)$ by $p(t)$ at the receiver implies spreading the spectrum of each of the messages $m_i(t)$ that have already been spread. The noise spectral density $\Sigma m_i(t)p_i(t)p(t)$ is consequently low. The interference noise power in the bandwidth of the useful message $m(t)$ is thus low.

In the preceding discussion, it has been assumed that multiplication by the chip sequence is performed on the binary message at baseband. It should be noted that Eq. (6.17) is also obtained by multiplying the carrier by the chip sequence after the carrier has been modulated by the binary message. In the same way, the operations of demodulation and despreading can be reversed at the receiver. If spread-spectrum transmission is used to allow multiple access, it is preferable on reception to proceed firstly to despreading and then to demodulation. Otherwise, as described earlier for reasons of simplicity of explanation, coherent demodulation necessitates recovery of the reference carrier in a spectrum (obtained by nonlinear processing of spread and modulated carriers) containing the other reference carriers with higher power levels. By proceeding firstly to despreading, the spectra of the unwanted carriers are spread, and recovery of the required reference carrier is performed under more favourable signal-to-noise conditions. On transmission, technological simplicity tends to a preference for spreading before modulation.

6.7.1.4 Protection against interference

The signals transmitted by systems sharing the same frequency band as that used by the network can be narrow-band carriers (medium-capacity TDM/PSK/FDMA carriers, for example). Let $f(t)\cos \omega_c t$ be such a carrier. The signal at the multiplier output is:

$$x(t) = m(t) + f(t)p(t) \qquad \text{(V)} \tag{6.23}$$

The interference noise is spread by the receiver. The interference power in the bandwidth of the useful message $m(t)$ is small.

This property is useful:

— For military applications, when one wishes to avoid interference from an enemy transmitting a high power in a narrow band. (Spread-spectrum transmission also provides the possibility of transmitting with discretion in view of the low spectral density of the carrier.)
— For civil applications, when one wishes to receive signals with small antennas in congested bands (for example at 4 GHz); due to the wide aperture of the antenna beam, the station receives carriers from adjacent satellites with a relatively high power. Spreading the spectrum of the carriers by the receiver limits the interference power from adjacent satellites in this case.

6.7.1.5 Protection against multipath

A link suffers from *multipath* when the radio wave follows paths of different lengths and arrives at the receiver in the form of a useful signal accompanied by replicas delayed in time. This arises, for example, in mobile satellite links where the downlink wave is captured at the same time as its reflections from surrounding objects. The reflected signals thus appear as interference. If the time delay between the direct wave and the reflected waves is greater than the duration T_c of a chip, there is no longer correlation between the received code and the local code for the reflected waves, and the spectrum of the reflected signals is spread. This provides protection against multipath.

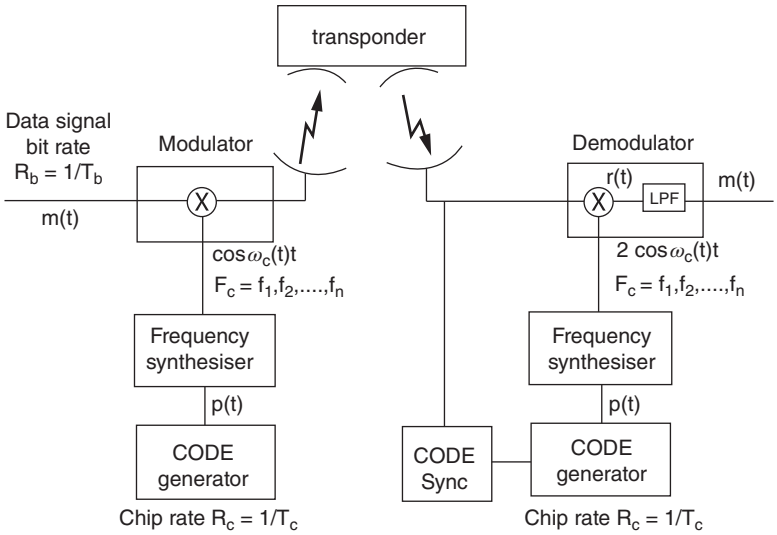


Figure 6.22 Frequency hopping (FH-CDMA).

6.7.2 Frequency hopping CDMA (FH-CDMA)

6.7.2.1 The principle

Figure 6.22 illustrates the principle. The binary message $m(t)$ to be transmitted at bit rate $R_b = 1/T_b$ is coded in NRZ. It modulates a carrier whose frequency $F_c(t) = \omega_c(t)/2\pi$ is generated by a frequency synthesiser controlled by a binary sequence (code) generator. This generator delivers chips with a bit rate R_c. The principle is illustrated by means of modulation by BPSK, although other types of modulation can be adopted, particularly frequency shift keying (FSK). The transmitted carrier is thus of the form:

$$c(t) = m(t)\cos\ \omega_c(t)t \tag{6.24}$$

The carrier frequency is determined by a set of $\log_2 N$ chips, where N is the number of possible carrier frequencies. It changes each time the code has generated $\log_2 N$ consecutive chips. The carrier frequency thus changes in hops. The hop rate is $R_H = R_c/\log_2 N$.

At the receiver, the carrier is multiplied by an unmodulated carrier generated under the same conditions as at the transmitter. If the local code is in phase with the received code, the multiplier output signal is:

$$r(t) = m(t)\cos\ \omega_c(t)t \times 2\cos\ \omega_c(t)t = m(t) + m(t)\cos\ 2\omega_c(t)t \tag{6.25}$$

The second term is eliminated by the LPF of the demodulator.

6.7.2.2 Spectral occupation

Three types of system can be considered:

— *One hop per bit*: $R_H = R_b$ that there is one frequency hop per information bit.

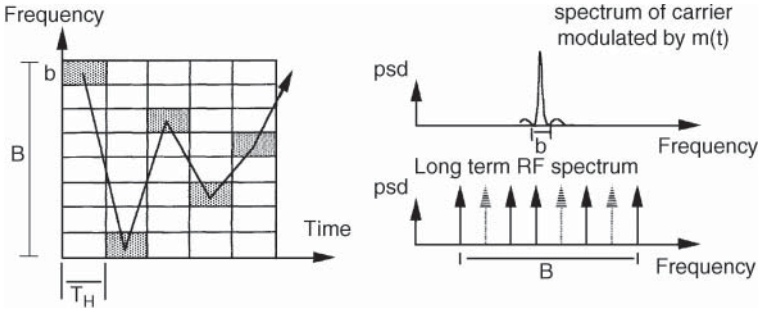


Figure 6.23 Spectral distribution in FH-CDMA for $R_H \ll R_b$; (psd: power spectral density).

— *Fast frequency hopping (FFH)*: $R_H \gg R_b$ that there are several frequency hops per bit.
— *Slow frequency hopping (SFH)*: $R_H \ll R_b$ that a frequency hop covers several bits.

Figure 6.23 shows an example of transmission with $R_H \ll R_b$. The short-term carrier spectrum (the spectrum for a period $T_H = 1/R_H$) has the characteristics of a BPSK carrier modulated by a binary stream of bit rate R_b and consequently occupies a bandwidth b approximately equal to R_b. The long-term spectrum consists of the superposition of the N carriers of the short-term spectrum. Hence it has a wider spectrum B. The spreading factor is B/b. The progress of the transmission of a carrier can be represented on the frequency-time grid in Figure 6.23 where each square represents one frequency state of the carrier at a given time.

6.7.2.3 *Realisation of multiple access*

The various network carriers follow different trajectories on the grid in Figure 6.23. At the receiver, only the carrier whose trajectory coincides with that of the carrier regenerated by the local synthesiser is demodulated. Hence the multiplier output signal, during an interval T_H when the synthesiser frequency is constant and equal to $\omega_c/2\pi$, is:

$$r(t) = \left[m(t) \cos \omega_c t + \sum m_i(t) \cos \omega_{ci}(t) \right] \times 2 \cos \omega_c t \tag{6.26}$$

At the output of the LPF, one finds $m(t)$ accompanied by noise caused by the possible presence of carriers such that $\omega_{ci} = \omega_c$. The probability of such an event is small when the number of frequency bands on the grid is high, and hence the spectrum-spreading factor B/b is large. The spectral density of the long-term interference noise spectrum can thus be made small.

6.7.2.4 *Protection against interference*

In a similar manner to the case of direct sequence CDMA, interference caused by fixed frequency carriers is subject to spectrum spreading at the receiver, which limits the noise power in the bandwidth of the useful message $m(t)$.

6.7.3 Code generation

Figure 6.24 shows an example of a technique for the generation of a pseudorandom code sequence. The generator consists of a set of r flip-flops forming a shift register with a set of

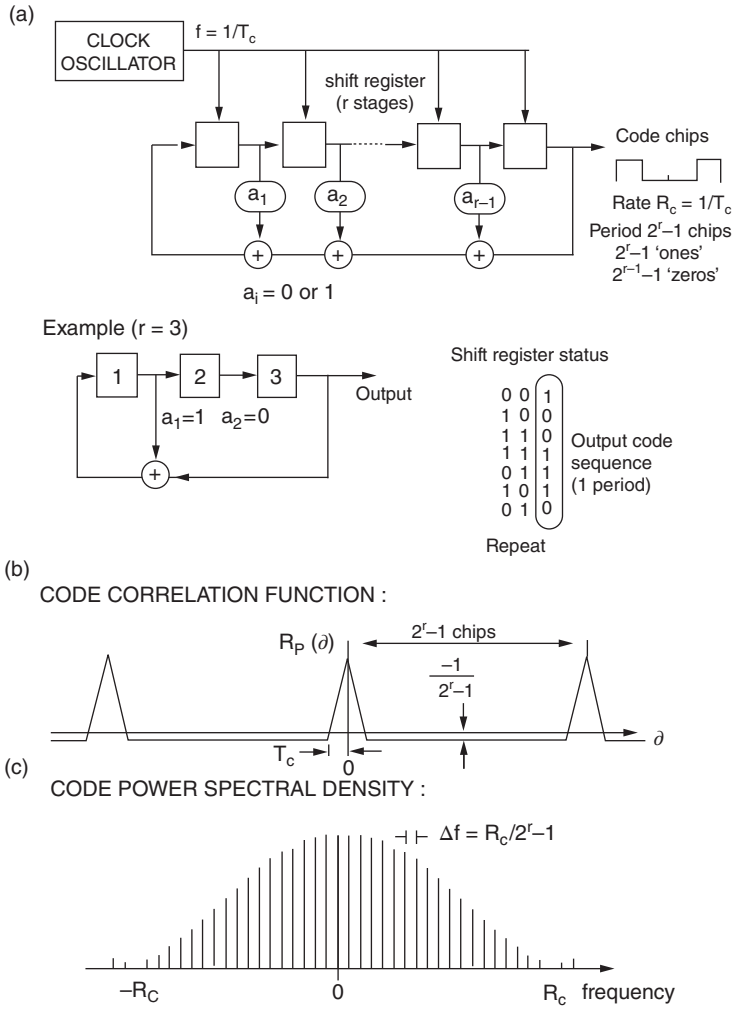


Figure 6.24 Pseudorandom sequence: (a) generation; (b) autocorrelation function; (c) power spectral density.

feedback paths provided with exclusive OR operations. The state of the flip-flops changes at the clock rate R_c. The stream of chips at the output is periodic with a period of $2^r - 1$ chips, and each period contains $2^{r-1} - 1$ chips equal to 0 and 2^{r-1} chips equal to 1. Figure 6.24 also shows the form of the autocorrelation function of the sequence together with its frequency spectrum.

6.7.4 Synchronisation

Synchronisation of the receiver pseudorandom sequence generator and the pseudorandom sequence that spreads the spectrum of the received carrier is mandatory for achieving multiple access. Only if synchronised can the receiver detect the useful message $m(t)$. Synchronisation

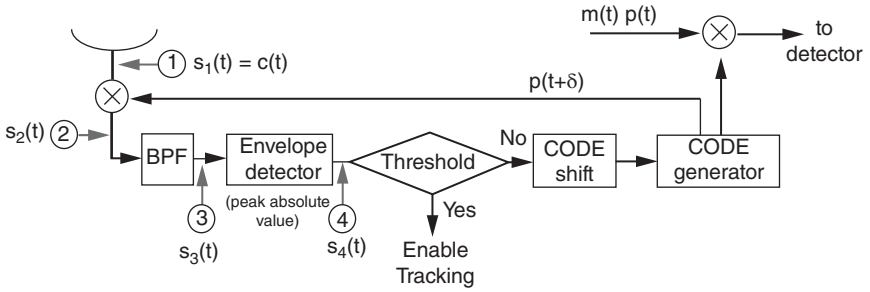


Figure 6.25 The principle of code acquisition in a DS-CDMA system.

consists of two phases: sequence acquisition and tracking. The principle of acquisition is illustrated for the case of direct sequence transmission (DS-CDMA).

6.7.4.1 Acquisition

Figure 6.25 shows the principle of a possible acquisition technique for DS-CDMA; the received carrier $c(t) = s_1(t)$ is multiplied by the locally generated sequence $p(t + \partial)$. This is not in phase with the received sequence $p(t)$, and the shift is denoted by ∂. The multiplier output $s_2(t)$ is fed to a band-pass filter (BPF) filter that is centred on the carrier frequency ω_c and has a bandwidth wide with respect to the spectrum of $m(t)$ but narrow with respect to the spectrum of $p(t)$. The filter thus has the effect of averaging the product $p(t)p(t + \partial)$, and the filter output signal can be expressed by:

$$s_3(t) = m(t)\overline{p(t)\, p(t + \partial)} \cos\, \omega_c t \tag{6.27}$$

An envelope detector follows; it detects the peak value of the filter output signal. As the amplitude of the carrier modulated by $m(t)$ is constant, the signal at the envelope-detector output provides the absolute value of the autocorrelation function of $p(t)$, hence:

$$s_4(t) = \overline{|\, p(t)p(t + \partial)\, |} = |R_P(\partial)| \tag{6.28}$$

It is known (Figure 6.24c) that this function has a pronounced maximum for $\partial = 0$. The amplitude of the output voltage of the envelope detector is measured for a given value of ∂; then, if this voltage is less than a fixed threshold, ∂ is incremented by an amount equal to the duration of a chip T_c. The operation is repeated until the amplitude of the envelope-detector output exceeds the fixed threshold indicating that the correlation peak for $\partial = 0$ has been achieved. One then proceeds to the tracking mode.

It is good practice to accumulate the results of several measurements for a given ∂ by placing an integrator, with a time interval equal to several periods of the pseudorandom sequence, between the envelope detector and the threshold detector.

6.7.4.2 Tracking

Figure 6.26 shows the principle of the tracking technique; the acquisition loop is duplicated with an Advance branch and a Delay branch. The signal produced by the pseudorandom sequence generator in the Advance branch is $p(t + T_c/2)$; that produced in the Delay branch is $p(t - T_c/2)$. The two signals at the envelope detector outputs are subtracted to produce an error

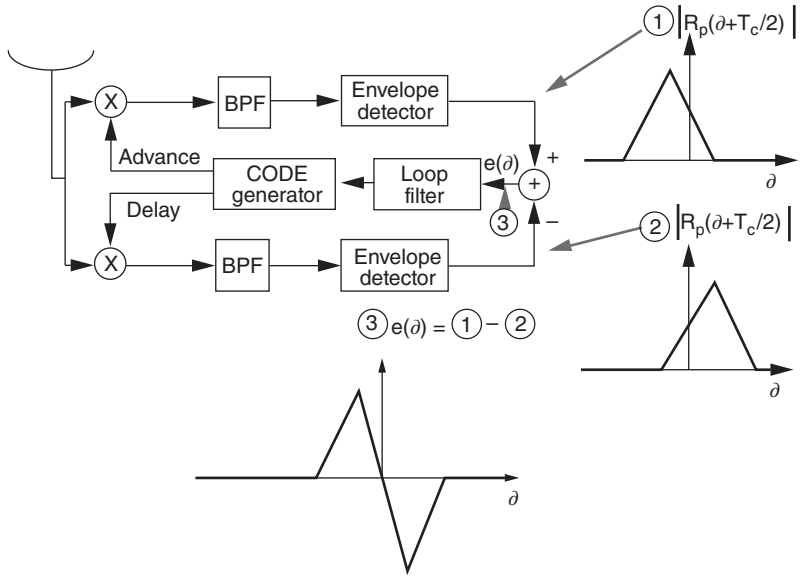


Figure 6.26 The principle of code tracking in a DS-CDMA system.

signal $e(\partial) = |R_P(\partial + T_c/2)| - |R_P(\partial - T_c/2)|$, which, after filtering, controls the advance or delay of the sequence generator. The sign of $e(\partial)$ indicates the direction of the correction to be performed, and variation of $e(\partial)$ as a function of ∂ has the form characteristic of an error signal in a control loop.

Some realisations, for acquisition and for tracking, replace the envelope detector with an energy detector (quadratic detector) [SIM-85], Vol. 3, Chapter 2. This does not modify the principle, but it does modify the form of the error signal characteristic. Other possibilities include numerical calculation of the convolution between the received signal and the locally generated code [GUO-90].

6.7.5 CDMA efficiency

The efficiency of CDMA can be considered the ratio of the total capacity provided by a repeater channel in the case of single access, which is a single carrier modulated without spectrum spreading, and that of a repeater channel transmitting several CDMA carriers simultaneously. The total capacity of the repeater channel is then the product of the capacity of one carrier and the number of carriers, which is the number of accesses.

6.7.5.1 Maximum number of accesses

Consider the case of direct sequence modulation (DS-CDMA). Assume for simplicity that the N received carriers are all of equal power C. The useful carrier power at the receiver input is thus C. As the information rate carried by this carrier is R_b, the energy per information bit is $E_b = C/R_b$. Neglecting thermal noise in the noise power at the receiver input and retaining only the contribution of interference noise, the noise power spectral density N_0 at the receiver input is approximately $N_0 = (N-1)C/B_N$, where B_N is the equivalent noise bandwidth of the receiver.

This gives:

$$E_b/N_0 = B_N/R_b(N-1) \tag{6.29}$$

The spectral efficiency $\Gamma = R_c/B_N$ of the digital modulation used can be introduced into this expression. This then gives:

$$E_b/N_0 = R_b/R_b(N-1)\Gamma \tag{6.30}$$

As the quality of the link is stipulated by a given error rate, the value of E_b/N_0 is imposed. From this the maximum number of accesses N_{max} is deduced from (6.30) and is given by:

$$N_{max} = 1 + (R_c/R_b)/\Gamma(E_b/N_0) \tag{6.31}$$

CDMA can also support FEC transmission (Section 4.3). The coded bit rate is now R_b/ρ. The utilised bandwidth remains unchanged as the chip rate R_c is kept the same. So the maximum number of accesses is:

$$N_{max} = 1 + \rho(R_c/R_b)/\Gamma(E_b/N_0)_2 \tag{6.32}$$

where ρ is the code rate and $(E_b/N_0)_2$ is the related required value of E_b/N_0.

In Eqs. (6.31) and (6.32), the first term, unity, can be neglected for practical calculations. The maximum number of accesses for coded transmission $(N_{max})_2$, as given by (6.32), is:

$$(N_{max})_2 = (N_{max})_1 \rho G_{cod} \tag{6.33}$$

where $(N_{max})_1$ is the number of accesses without coding, ρ is the code rate, and $G_{cod} = (E_b/N_0)_1/(E_b/N_0)_2$ is the decoding gain with typical values given in Table 4.7; ρG_{cod} is larger than 1, so the number of accesses has increased.

Obviously, from Eq. (6.29), the capacity of CDMA is interference limited by $N_0 = (N-1)C/B_N$. An interference-reduction technique, in connection with transmission of telephony signals, is voice activation, whereby a carrier is transmitted only during time intervals when users are speaking. Introducing the voice activity factor on a telephone channel, τ, the interference term in expression (6.30) becomes $(N-1)\tau$, and Eq. (6.32) becomes:

$$\begin{aligned} N_{max} &= 1 + \rho(R_c/R_b)/\Gamma\tau(E_b/N_0)_2 \\ &\approx \rho(R_c/R_b)/\Gamma\tau(E_b/N_0)_2 \end{aligned} \tag{6.34}$$

With $\tau = 0.4$, the number of accesses is increased by 2.5.

6.7.5.2 CDMA efficiency

The maximum total throughput is equal to $N_{max}R_b$. The throughput of a single-carrier modulated without spectrum spreading and occupying a bandwidth B_N would be the chip rate R_c. The throughput η of CDMA is thus given by the ratio:

$$\eta = N_{max}R_b/R_c \tag{6.35}$$

Example 6.2 Consider a CDMA network occupying the whole of a 36 MHz satellite repeater channel. The receiving bandwidth is $B_N = 36$ MHz. It is assumed that each carrier has a capacity equal to 64 kbps. With BPSK modulation of theoretical spectral efficiency $\Gamma = 1$ bps Hz, the chip rate is $R_c = B_N/\Gamma = 36$ Mbps and the spreading ratio R_c/R_b is $36\times10^6/64\times10^3 = 563$. $\eta = N_{max}/563$.

Table 6.3 shows the maximum number of accesses, the maximum total throughput of the network, and the resulting efficiency for a required bit error probability. The CDMA efficiency, on

Table 6.3 Performance of a CDMA access network using a 36 MHz repeater channel and binary phase shift keying (BPSK); each carrier has a capacity of 64 kbps

Required bit error probability	E_b/N_0 (dB)	Maximum number of accesses N_{max}	Maximum total throughput (Mbps)	Efficiency (%)
10^{-4}	8.4	82	5.3	15
10^{-5}	9.6	62	4.0	11
10^{-6}	10.5	51	3.3	9

the order of 10%, is low, compared for example with that of TDMA (Section 6.6.5). The values in the table are optimistic: thermal noise is neglected, user codes are assumed to be orthogonal, and no account is taken of degradation due to the demodulator.

6.7.6 Conclusion

CDMA operates on the principle of spread-spectrum transmission, recalled in Figure 6.27. The code sequence that serves to spread the spectrum constitutes the signature of the transmitter.

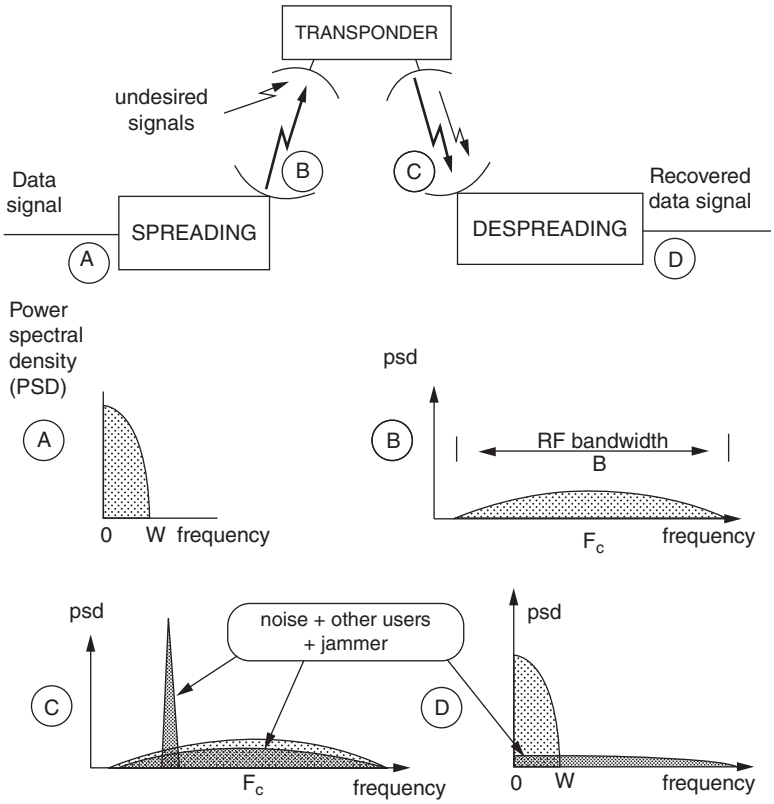


Figure 6.27 Spread-spectrum transmission in a code division multiple access system.

The receiver recovers the useful information by reducing the spectrum of the carrier transmitted in its original bandwidth. This operation simultaneously spreads the spectrum of other users in such a way that these appear as noise of low spectral density.

CDMA has the following advantages:

— It is simple to operate since it does not require any transmission synchronisation between stations. The only synchronisation is that of the receiver to the sequence of the received carrier.
— It offers useful protection properties against interference from other systems and interference due to multiple paths; this makes it attractive for networks of small stations with large antenna beamwidth and for satellite communication with mobiles.
— With multibeam satellites, it offers the potential of 100% frequency reuse between beams (Chapter 5).

The main disadvantage is poor efficiency, on the order of 10%, as a large bandwidth of the space segment is used for a low total network capacity with respect to the throughput of a single unspread carrier. This comment applies only in a single-beam network. As indicated earlier, the possibility of reusing frequency between adjacent beams greatly improves the overall efficiency. Another limitation consists of the limited number of codes (and therefore the number of simultaneous users) offering the required performance in term of inter-correlation properties.

6.8 FIXED AND ON-DEMAND ASSIGNMENT

6.8.1 The principle

Traffic routing implies access by each carrier transmitted by the earth stations to a radio-frequency channel. For each of the three fundamental modes (FDMA, TDMA, and CDMA) described in the previous sections, each carrier is assigned a portion of the resource offered by the satellite, i.e. a satellite channel (a frequency band, a time slot, or a fraction of the total power keyed to a code) or a part of it. This assignment can be defined once and for all (a fixed assignment) or in accordance with requirements (on-demand assignment).

With *fixed assignment,* the capacity allocated each earth station is fixed independently of the traffic demand from the terrestrial network to which it is connected. An earth station can receive a traffic request from the network to which it is connected greater than the capacity that is allocated to it. It must then refuse some calls; this is a blocking situation, in spite of the fact that other stations may have excess capacity available. Because of this, the resource constituted by the satellite network is poorly exploited.

With *on-demand assignment,* the satellite network resource can be assigned in a variable manner to the various stations in accordance with demand. There will, therefore, be the possibility of transferring capacity from stations with low demand to stations with excess demand.

With on-demand assignment CDMA or FDMA, a given capacity is allocated on request to a given transmitting station by assigning to that station for the duration of the connection a given code within a set of orthogonal codes or a given frequency band. On-demand assignment is straightforward in the 'one carrier per station-to-station link' routing technique (Section 6.3) combined with the 'single channel (connection) per carrier' (SCPC) scheme (Section 6.5.2). When considering a 'one carrier per transmitting station' on-demand assignment technique, then a variable proportion of connections to users connected to different earth stations has to be imple-

mented in the earth station multiplexer. If the 'one carrier per station-to-station link' technique is considered but with multiple connections per carrier, then the earth station must be equipped with several transmitters (up to as many as other stations in the network), with variable-capacity multiplexers. This means the equipment can be very expensive or there will be a lack of flexibility.

On-demand assignment TDMA offers the greatest flexibility; on-demand assignment is achieved by adjusting the length and position of bursts, requiring a coordinated burst-time plan change. This only slightly increases the earth station hardware complexity as the earth stations already have synchronisation equipment. It becomes practical and economic to implement with the development of electrical and software control technologies. Capacity increments can be as small as one communication channel, and the assignment can be performed on a call-by-call basis.

6.8.2 Comparison between fixed and on-demand assignment

Consider a satellite network containing 20 stations. Each station must transmit traffic to the 19 other stations via a satellite repeater channel whose capacity S is 1520 communication channels. A blocking probability of 0.01 is required. The traffic intensity per communication channel will be calculated for the case of fixed assignment and for that of on-demand assignment.

6.8.2.1 Fixed assignment

The total capacity of the repeater channel is shared among 20 stations. Each station thus has $1520/20 = 76$ communication channels available. These channels are shared among the 19 destinations. There are, therefore, $76/19 = 4$ channels per destination. The maximum traffic intensity A must be determined such that the blocking probability $B(C = 4, A) = E_{C=4}(A)$ remains less than 0.01. It is found, by using Eq. (6.2) or Figure 6.1, that $A = 0.87$ erlang, which is 0.217 erlang per communication channel.

6.8.2.2 On-demand assignment

The total capacity S of the repeater channel can be assigned to any station, whatever the destination. One must realise the condition $B(S = 1520, \underline{A}) = E_{S=1520}(A) < 0.01$ that leads to $A = 1491$ for the 1520 communication channels and hence an intensity $A/1520 = 0.98$ erlang per communication channel.

6.8.3 Centralised or distributed management of on-demand assignment

Management is centralised when it is performed in a single station; this implies that ordinary traffic stations send demand messages to the central station, which determines the assignment of resources and transmits this assignment to the whole network. Management is distributed when stations transmit their demands on a common signalling channel. These demands are taken into account by each station and the state of the resources is updated at each station.

Table 6.4 presents a comparison of the advantages and disadvantages of centralised management with respect to those of distributed management.

Table 6.4 Comparison of centralised and distributed control of demand assignment

Principles	Advantages	Disadvantages
Centralised control		
Traffic stations send requests to the central station.	Stations do not have to perform the assignment.	Correct operation of the network depends on that of the control station.
	Low equipment cost.	Reduced reliability.
Determination of resource assignment by the central station and distribution of this assignment to the whole network.	Reduced signalling – the whole network does not need to be informed of requests.	The need for a redundant control station.
		The setup time of a connection is penalised by the double hop.
Distributed control		
Requests are transmitted on a common signalling channel.	No control station.	More complex equipment at each station.
	Better network reliability.	
Each station updates the state of the resources.	Reduced connection setup time.	Higher cost.
		More signalling.

6.8.4 Conclusion

Fixed assignment is recommended for networks involved in routing large volumes of traffic between a small number of stations of high capacity. On-demand assignment provides better utilisation of the satellite network in the case of a large number of stations of low capacity per access with large variations in demand. Each station can thus benefit occasionally from a greater capacity than it would have in the case of a fixed assignment. Management of the assignment implies a connection setup time on the order of a second. When the connection is required for several minutes, this setup time is of no consequence. The choice of a demand assignment technique must, therefore, take the following aspects into account:

— Specifications on the user side: traffic density, number of destinations, and blocking probability.
— The gain resulting from the operation of demand assignment. This involves comparing the increase in revenue resulting from higher traffic throughput for a given blocking probability with the increased expense involved in the installation of equipment to manage demand assignment.
— The choice between centralised and distributed management.

The setup time of the connection can be a decisive factor for some types of traffic, such as data exchanged between data-processing systems. The traffic generated in communication between computers, or between computers and computer terminals, is characterised by a large variation in the duration of messages and the inter-arrival time between messages. Furthermore, the user often imposes a clause concerning transmission time that can be short compared with the time interval between messages. Under these conditions, the setup time of a connection can exceed the utilisation time, and this corresponds to inefficient use of the network. On the other hand, the setup time of the connection can lead to an unacceptable transmission time. Under these circumstances, it is preferable to resort to *random access* as described in Section 6.9.

6.9 RANDOM ACCESS

This type of access is well suited to networks containing a large number of earth stations where each earth station is required to transmit short, randomly generated messages with long silent times between messages. The principle of random access is to permit transmission of messages almost without restriction in the form of limited-duration packets, to which correspond bursts of modulated carriers, which occupy all or part of the bandwidth of the repeater channel. It is, therefore, multiple access with *time division* and *random transmission*. The possibility of collisions between carrier bursts at the satellite is accepted. In the case of collision, the earth station receiver is confronted with interference noise that can compromise demodulation with target BER and correct message identification. Retransmission of all or part of the burst is necessary.

The performance of random access is measured in terms of the *normalised throughput* and the *mean transmission delay*. Normalised throughput has been defined previously as *efficiency* and is restated here as the ratio of the volume of traffic *delivered at the destination* to the maximum volume that could be transmitted in the available bandwidth. The subtlety here is that the traffic delivered is not necessarily equal to the actual traffic load, and actually is less, as some of the messages must be transmitted several times. Hence, the transmission time (delay) is a random variable depending on the number of transmissions of a given message. Its mean value indicates the mean time between the generation of the message and its correct reception by the destination station.

Random access has been the object of numerous studies since 1970. Its practical application has become important in the context of private networks using small stations (very-small-aperture terminals [VSATs]) that have been widely developed to provide satellite communication between computers and distant terminals. A description of protocols based on random and on-demand access used in VSAT networks is available in [MAR-04].

6.9.1 Asynchronous protocols

6.9.1.1 The ALOHA protocol

Figure 6.28 illustrates the principle of multiple random access in accordance with the ALOHA protocol [ABR-77, HAY-81, ABR-73]. The packets (identified as X and Y in the figure) are transmitted by each earth station without any restriction on the time of transmission. It is, therefore, an asynchronous protocol. In the absence of collisions (Figure 6.28a), the destination earth station (denoted Z) properly identifies the packet content and transmits an acknowledgement for correct reception in the form of a short acknowledgement packet (ACK).

Figure 6.28b illustrates the case of a collision. The destination station receiver is not able to identify the message and does not send an acknowledgement. If an acknowledgement is not received within a fixed time interval after transmission, the transmitting station retransmits the message; the time interval is set to a value slightly greater than twice the round-trip propagation time of the carrier wave. This new transmission occurs after a random time that is determined independently at each station in order to avoid a further collision.

Consider a satellite channel and M earth stations with bursty traffic having average total packet-generation rate λ, i.e. λ/M per earth station (s^{-1}). The duration of any emitted packet is fixed and equal to $\tau(s)$. The probability of a new packet being generated (generated traffic) is given by:

$$S = \lambda\tau \quad \text{(per packet)}$$

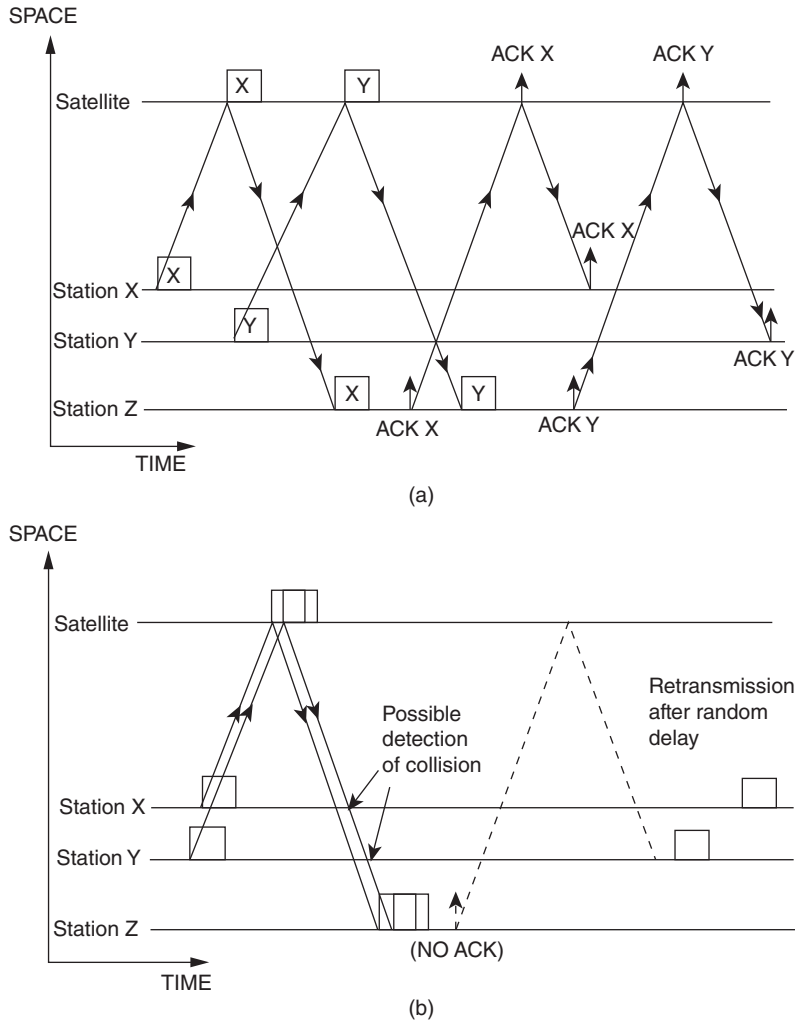


Figure 6.28 Distance–time diagram illustrating the principle of the ALOHA random multiple access protocol: (a) without collision; (b) with collision.

Due to collision, some of these packets are retransmitted, and the satellite channel conveys both the new packets and the retransmitted ones.

The probability of a packet arriving at the satellite channel input (channel load) is G (packets per packet length); G is larger than S due to retransmitted packets. Assuming the packet emission follows a Poisson process, the probability that k packets arrive at the satellite channel during any interval of t packets duration is:

$$\text{Prob}[k, t] = \frac{(Gt)^k \exp(-Gt)}{k!}$$

Consider a packet arriving at the satellite channel; this occurs with probability G. A collision occurs if one or more arriving packets overlap partially or totally with the considered packet,

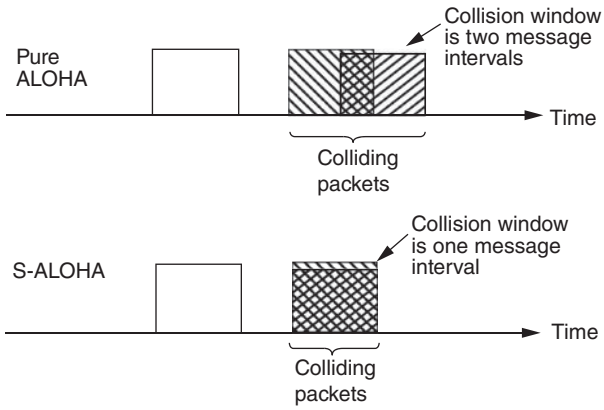


Figure 6.29 Collision diagrams with the ALOHA and slotted ALOHA (S-ALOHA) protocols.

i.e. they fall within a (collision) window $t = 2$, as shown in Figure 6.29 (pure ALOHA). The probability of *no collision* is given by:

$$\text{Prob[no collision]} = \text{Prob}[k = 0, t = 2] = \exp(-2G)$$

The probability of *successful delivery* is given by:

$$\begin{aligned} S &= \text{Prob[success]} = \text{Prob [a packet arrivies] Prob[no collision]} \\ &= G \times \text{Prob}[k = 0, t = 2] \\ &= G\exp(-2G) \end{aligned} \tag{6.36}$$

S and G are expressed as a number of packets per time slot equal to the common packet duration. S corresponds to the *normalised throughput* and is a measure of the *efficiency* of the access scheme (notice that λ is the packet throughput and is $1/\tau$ the maximum packet rate; the ratio is $S = \lambda\tau$). This curve is shown in Figure 6.30. Figure 6.31 shows the variation of mean transmission time as a function of S. It can be seen that the ALOHA protocol does not exceed a normalised throughput of 18%, and the mean transmission time increases as the traffic increases due to the increasing number of collisions and packet retransmissions.

6.9.1.2 The selective reject (SREJ) ALOHA protocol

With asynchronous transmission, collision between packets is most often partial. With the ALOHA protocol, the coherence of the packet is destroyed by even a partial collision. This leads to retransmission of the contents of the whole packet although only a part has suffered a collision. The SREJ-ALOHA protocol [RAY-87] has been designed to avoid a complete retransmission. The transmitted packet is divided into subpackets, each having its own header and protocol bits. When a collision occurs, only the subpackets involved are retransmitted. The efficiency of the protocol is greater than that of the ALOHA protocol. The practical limit, on the order of 30%, is caused by the addition of headers to the subpackets. The SREJ-ALOHA protocol is well suited to applications in which the messages have variable lengths.

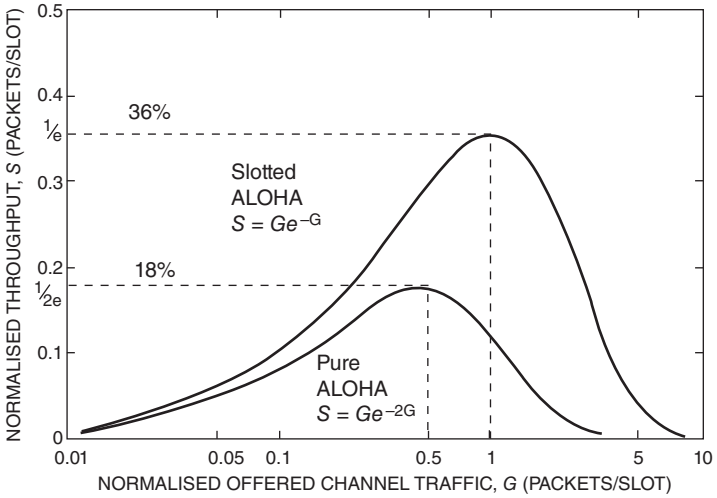


Figure 6.30 Transmission efficiency.

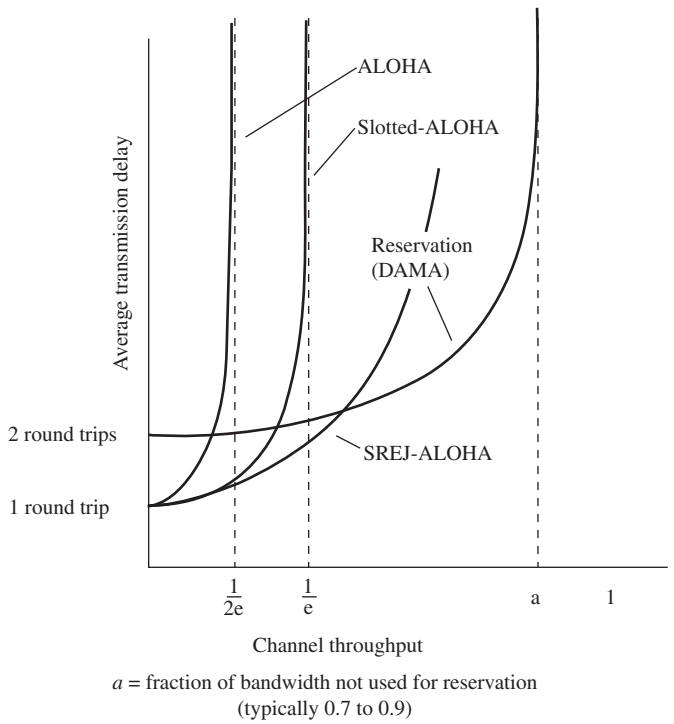


Figure 6.31 Mean transmission time.

6.9.1.3 The time-of-arrival collision resolution algorithm protocol

The time-of-arrival collision resolution algorithm protocol [CAP-79] provides an improvement to the ALOHA protocol by avoiding the possibility that a packet that has already been subjected to a collision encounters another packet during its retransmission. To achieve this, stations avoid transmitting new packets in the time slots provided for retransmission of packets that have suffered a first collision. This protocol implies a procedure for identifying packets that have suffered a collision and the setting up of temporary coordination of transmissions. The efficiency is from 40–50%. However, such a protocol tends to be complex to implement.

6.9.2 Protocols with synchronisation

6.9.2.1 The slotted ALOHA (S-ALOHA) protocol

Transmissions from stations are synchronised in such a way that packets are located at the satellite in time slots defined by the network clock and equal to the common packet duration. Hence there cannot be partial collisions; every collision arises from complete superposition of packets. The timescale of the collision is thus reduced to the duration of a packet, whereas with the ALOHA protocol the timescale is equal to the duration of two packets, as shown in Figure 6.29. With S-ALOHA, the probability of no collision is given by:

$$\text{Prob[no collision]} = \text{Prob}[k = 0, t = 2] = \exp(-G)$$

The probability of successful delivery is given by:

$$\begin{aligned} S &= \text{Prob[success]} = \text{Prob [a packet arrivies]}\ \ \text{Prob[no collision]} \\ &= G \times \text{Prob}[k = 0, t = 1] \\ &= G\exp(-G) \end{aligned} \tag{6.37}$$

The curve is represented in Figure 6.30 and displays the increased efficiency due to synchronisation.

6.9.2.2 The announced retransmission random access (ARRA) protocol

This protocol increases the efficiency of S-ALOHA by introducing a frame structure that permits numbering of time slots. Each packet incorporates additional information indicating the slot number reserved for retransmission in case of collision. This protocol enables collisions between new messages and retransmissions to be avoided. The efficiency is on the order of 50–60%.

6.9.3 Protocols with assignment on demand

These protocols are intended to increase efficiency even more by an advance capacity reservation procedure (demand assignment multiple access [DAMA]). A station reserves a particular time slot within a frame for its own use. Reservation can be implicit or explicit [RET-80]:

— Implicit reservation is reservation by occupation; that is, every slot occupied once by the packet from a given station remains assigned to this station in the frames that follow. This

protocol is called R-ALOHA [CRO-73, ROB-73]. The disadvantage is that a station is in a position to capture all the time slots of a frame for itself. The advantage is the absence of setup time for a reservation.

— Explicit reservation involves a station sending a request to occupy certain time slots to a control centre. Two examples of this protocol are R-TDMA and contention-based priority-oriented demand assignment (C-PODA) [JAC-78]. The disadvantage of these protocols is the establishment time, which can be prohibitive in some interactive applications. Figure 6.31 shows the variation of transmission time for protocols of the DAMA type with explicit reservation as a function of the efficiency (normalised throughput).

The efficiency can be as high as 70–90% depending on the fraction of capacity used for the signalling information associated with the reservation procedure.

6.10 CONCLUSION

There is a large variety of solutions to the problem of multiple access to a satellite by a group of network stations. The choice of access type depends above all on economic considerations; these are the global costs in terms of investment and operating costs and the benefits in terms of revenues.

General indications can be given according to the type of traffic. For traffic characterised by long messages implying continuous or quasi-continuous transmission of a carrier (for example, telephone traffic, television transmission, and videoconferencing), FDMA, TDMA, and CDMA access techniques are the most appropriate. If the volume of traffic per carrier is large and the number of accesses is small (trunking), FDMA has the advantage of operational simplicity. When the traffic per carrier is small and the number of accesses is large, FDMA loses much in efficiency of usage of the space segment and TDMA and CDMA are the best candidates. However, TDMA

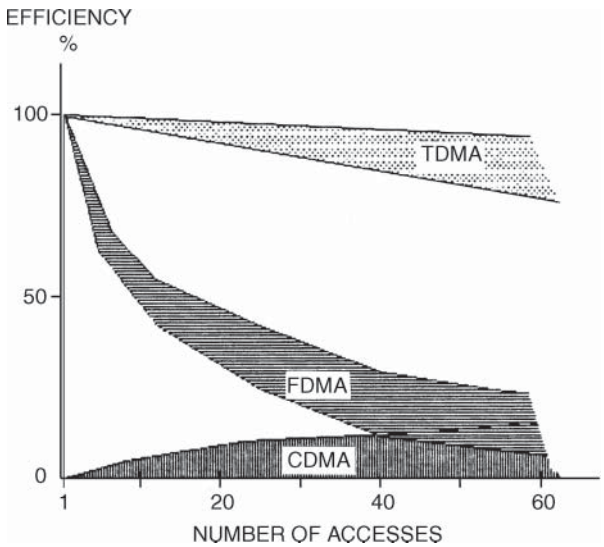


Figure 6.32 Comparison of efficiency for different multiple access schemes. Efficiency of 100% corresponds to the throughput achieved with *one access* only (one carrier within a single repeater channel, operated at saturation).

requires relatively costly earth station equipment. For small stations exposed to inter-system interference, CDMA may be preferred despite its low efficiency.

Selection of FDMA or TDMA multiple access also implies a choice between fixed and on-demand assignment. Economic considerations will prevail; the increase in revenue resulting from higher traffic is compared with the increased expense involved in the installation of equipment to control on-demand assignment. Figure 6.32 summarises the results stated previously relating to the efficiency of the various multiple-access schemes.

For traffic characterised by short messages and random generation with long silent times between messages, random access is the most appropriate. Figure 6.31 illustrates the choice between a short transmission delay with low efficiency (pure ALOHA) or a higher efficiency with a longer transmission delay (S-ALOHA or DAMA).

REFERENCES

[ABR-73] N. Abramson (1973) Packet switching with satellites, NCC AFIPS Conference Proceedings, 42, pp. 695–702.

[ABR-77] Abramson, N. (1977). The throughput of packet broadcasting channel. *IEEE Transactions on Communications* **25** (1): 117–128.

[CAP-79] Capetenakis, J.I. (1979). Tree algorithms for packet broadcast channels. *IEEE Transactions on Information Theory* **25** (5): 505–513.

[CRO-73] Crowther, W., Rettberg, R., and Walden, D. (1973). A system for broadcast communications: reservation ALOHA. In: *Proceedings of the 6th International System Science Conference*, 371–374. Hawaii.

[FEH-83] Feher, K. (1983). *Digital Communications*. Prentice Hall.

[GAG-91] Gagliardi, R.M. (1991). *Satellite Communications*, 2e. Van Nostrand Reinhold.

[GUO-90] Guo, X.Y., Maral, G., Marguinaud, A., and Sauvagnac, R. (1990). A fast algorithm for the pseudonoise sequence acquisition in direct sequence spread spectrum systems. In: *Proceedings of the Second International Workshop on Digital Signal Processing Techniques Applied to Space Communications, Turin (Italy)*. ESA-WPP-019. ACM.

[HAY-81] Hayes, J.F. (1981). Local distribution in computer communications. *IEEE Communications Magazine* **19** (2): 6–14.

[JAC-78] Jacobs, I.M., Binder, R., and Hoversten, E.V. (1978). General purpose packet satellite networks. *Proceedings of the IEEE* **66** (11): 1448–1467.

[MAR-04] Maral, G. (2004). *VSAT Networks*, 2e. Wiley.

[RAY-87] Raychaudhuri, D. (1987). Stability, throughput and delay of asynchronous selective reject ALOHA. *IEEE Transactions on Communications* **35** (7): 767–772.

[RET-80] Retnadas, G. (1980). Satellite multiple access protocols. *IEEE Communications Magazine* **18** (5): 16–22.

[ROB-73] Roberts, L.G. (1973). Dynamic allocation of satellite capacity through packet reservation. In: *National Computer Conference*, 711–716. ACM.

[SIM-85] Simon, M.K., Omura, J., Scholtz, R.A., and Levitt, B.K. (1985). *Spread Spectrum Communications*. Computer Science Press.

7 SATELLITE NETWORKS

This chapter starts by briefly introducing the basics of network layering models and protocols used to deliver different types of service to end users. Those fundamentals are the starting point for any satellite network design, as it is of high importance that satellite networks align with the protocols and interfaces that terrestrial networks implement [SUN-14].

The chapter then presents the fundamental concepts of satellite network architectures. The basic characteristics of a satellite network are introduced, in terms of topology, connectivity, and types of link before a description of typical satellite network architectures, including ones that use intersatellite links (ISLs). Issues regarding routing and switching with transparent and regenerative processing are dealt with. Examples given relate to: meshed networks for point-to-point communications, star networks for point-to-multipoint (data distribution) or multipoint-to-point (data collection) communications, broadcasting and multicasting to fixed and mobile terminals, and hybrid networks that mix star and meshed topology. Then Internet protocol (IP) networks based on digital video broadcasting via satellite (DVB-S) and DVBRCS are briefly introduced, and the Internet transmission control protocol (TCP) and its enhancements for satellite networks are presented. Finally, IPv6 over satellite networks is discussed.

7.1 NETWORK REFERENCE MODELS AND PROTOCOLS

7.1.1 Layering principle

As introduced in Chapter 6, the exchange of information between source and destination in a communications network involves a lot of interacting functions. In order to master interactions and facilitate design, it is useful to identify and group tasks of similar nature, and clarify the interactions between the various groups in a well-structured architecture. The system functions are therefore divided into *layers*, with sets of rules (called *protocols*) taking care of exchange of information between layers.

A *reference model* provides all the roles devoted to each group of functions of a certain kind corresponding to a layer and defines the required interfaces with the neighbouring layers. All parties can operate and communicate with each other if they follow the roles defined in the reference model.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

A *layer* is designed to offer certain services to the layers above, shielding those layers from the details of how the services are actually implemented. Each layer has an interface with primitive operations (data types consisting of values, ways to access data, and operations, ways to process data) that are used to access the offered services. Entities are the active element in each layer, such as user terminals (UTs), switches, and routers. Peer entities are the entities in the layer capable of communication with the same protocols.

A *protocol* is the rules and conventions used in conversation by agreement between communicating parties. Basic protocol functions include segmentation and reassembly, encapsulation, connection control, ordered delivery, flow control, error control, routing, and multiplexing. Protocols are needed to enable parties to understand each other and make sense of received information.

A *protocol stack* is a list of protocols (one protocol per layer). A network *protocol architecture* is a set of layers and protocols.

International standards are important to achieve global acceptance. Protocols described in the standards are often in the context of reference models, as there are many different standards developed. The layering principle is an important concept for network protocols and reference models. In the 1980s, the International Organisation for Standardisation (ISO) derived the seven-layer reference model shown in Figure 7.1.

7.1.2 Open Systems Interconnection (OSI) reference model

The Open Systems Interconnection (OSI) reference model is the first complete reference model developed as an international standard. The clear and simple principles that were applied to arrive at the seven layers can be summarised as follows:

— A layer defines a level of abstraction that should be different from any other layer.
— Each layer performs a well-defined function.

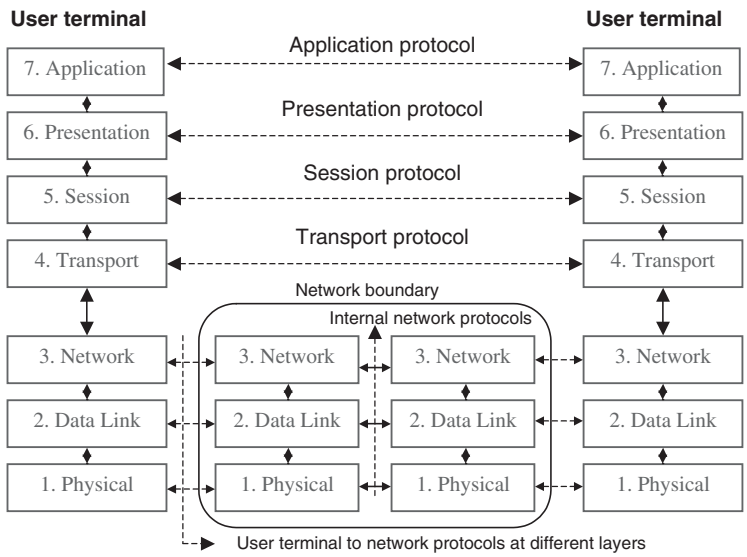


Figure 7.1 OSI/ISO seven-layer reference model.

— The function of each layer should be chosen to lead to internationally standardised protocols.
— The layer boundaries should be chosen to minimise information flow across the interface.

The following are brief descriptions of each layer of the seven-layer OSI reference model:

Layer 1: The *Physical* layer (PL) specifies electrical interfaces and the physical transmission medium. In a satellite network, layer 1 consists of modulation and channel coding techniques that enable the bit stream to be transmitted in specific formats and allocated frequency bands. The radio links act as the physical transmission media.

Layer 2: The *Data Link* layer provides a line that appears free of undetected transmission errors to the Network layer. A special sublayer called *Medium Access Control* (MAC), deals with the sharing of the physical resource between communicating terminals. This is the subject of the multiple access techniques (FDMA, TDMA, CDMA, etc.) discussed in Chapter 6 and demand assignment multiple access (DAMA). Broadcasting networks have additional issues in the Data Link layer, e.g. how to control access to the shared medium.

Layer 3: The *Network* layer routes packets from source to destination. The functions include network addressing, congestion control, accounting, disassembling and reassembling, and coping with heterogeneous network protocols and technologies. In a broadcasting network, the routing problem is simple, as the source address is the same for all packets and the packets always follow the same path to their destinations. The routing protocol is often thin or even non-existent.

Layer 4: The *Transport* layer provides a reliable (error-free) data-delivery service for processes utilising the transmitted data at higher layers. It is the highest layer of the services associated with the provider of communication services, guaranteeing ordered delivery, error control, flow control, and congestion control.

The higher layers are related to user data services:

Layer 5: The *Session layer* provides the means for cooperating presentation entities to organise and synchronise their dialogue and to manage the data exchange.

Layer 6: The *Presentation layer* is concerned with data transformation, data formatting, and data syntax.

Layer 7: The *Application layer* is the highest layer of the ISO architecture. It provides services to application processes.

When designing satellite networks, focus is primarily put on layers 1–4; but a good understanding of the upper layer processes and performances is also necessary, for the satellite network not to degrade the end-to-end quality of service (QoS) of the communication. As for the satellite payload itself, it can be involved in layer 1 or layer 2 processing. Extension to layer 3 processing is also possible, thanks to the evolution of on-board processor (OBP) technology.

One major trend in any telecommunications network is to move towards all-IP network technologies. Satellite networks are following the same trend.

7.1.3 IP reference model

Originally, the IPs were not developed by any international standardisation organisation. They were developed by a US Department of Defense (DoD) research project to connect a number of different networks designed by different vendors into a network of networks (the 'Internet' – internetworking different types of networks). It was initially successful because it delivered a few basic services such as file transfer, electronic mail, and telnet for remote logon across a

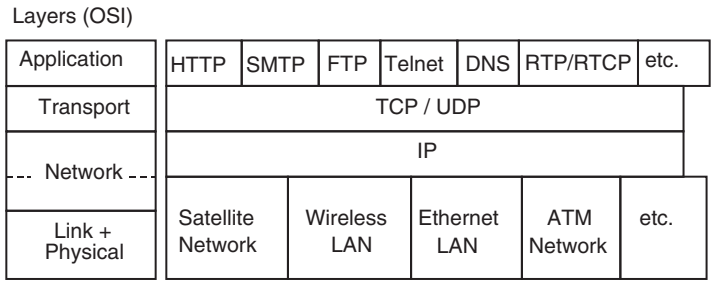


Figure 7.2 The Internet protocol reference model.

very large number of different networks including local area networks (LANs), metropolitan area networks (MANs), and wide area networks (WANs).

The main part of the IP reference model is the suite of TCPs and IPs known as the TCP/IP protocols. The IPs allow the construction of very large networks with little central management.

As with all other communications protocols, TCP/IP is composed of different layers. Figure 7.2 shows the IP reference model and examples of protocols and applications at the different layers: IP, TCP, user datagram protocol (UDP), hypertext transfer protocol (HTTP), simple mail transfer protocol (SMTP), file transfer protocol (FTP), protocol to interface terminal and applications (Telnet), real-time transfer protocol (RTP), real-time transfer control protocol (RTCP), etc.

7.1.3.1 The Network layer: IP

The IP Network layer is based on a datagram approach, providing only *best-effort service*, i.e. without any guarantee of QoS. IP is responsible for moving packets of data from router to router according to a four-byte destination IP address (in the IPv4 mode) until the packets reach their destination. Management and assignment of IP addresses is the responsibility of the Internet authorities.

Hosts in the same subnet can communicate with each other directly but have to communicate via router or gateway if they are on different subnets. The routers working together and managed by the same authority run the same routing protocols.

Typically, the routers are organised into an autonomous system (AS). An AS is a connected group of one or more IP prefixes run by one or more network operators with a single clearly defined routing policy.

Normally, with ASs, all routers run the same routing protocol. As networks have evolved, multiple routing protocols can co-exist. A routing policy is a set of rules that determine how traffic is managed within an AS, to which a single, or multiple operator(s) must adhere to.

The routing protocols within the AS are called *internal routing protocols* (such as routing information protocol [RIP]and open shortest path first protocol [OSPF]) and those outside the AS are called *external routing protocols* (such as board gateway protocol [BGP]).

7.1.3.2 Transport layer: TCP and UDP

TCP and UDP are Transport layer protocols of the IP reference model. They originate at the endpoints of bidirectional communication flows, allowing for end-user terminal services and applications to send and receive data across the Internet.

TCP is responsible for verifying the correct delivery of data between clients and servers that are hosts connected to the Internet. Data can be lost in the intermediate network. TCP adds support to detect errors or lost data and retransmit them until the data is correctly and completely received. Therefore TCP provides a reliable service, though the network underneath may be unreliable; i.e. operation of IPs does not require reliable transmission of packets, but reliable transmission can reduce the number of retransmissions and thus improve performance.

UDP provides a best-effort service, as it does not attempt to recover any error or packet loss. Therefore, it is a protocol providing unreliable transport of user data. But this can be very useful for real-time applications, as retransmission of any packet may cause additional delay and hence could cause more problems than losing packets.

7.1.3.3 Application layer

The Application layer protocols are designed as functions of user terminals or servers. The classic Internet Application layer protocols include HTTP for the Web, FTP for file transfer, SMTP for email, Telnet for remote login, and DNS for domain name services. Others include RTP and RTCP for real-time services and still others for dynamic and active web services. All of these applications are provided by the Internet with TCP/IP protocols.

7.2 REFERENCE ARCHITECTURE FOR SATELLITE NETWORKS

Satellite networks are used to provide two major types of services: TV services (associated with broadcast services) and telecommunication services (associated with two-way communication services, symmetric telephony, or asymmetric – Internet access).

One or more satellite networks can be deployed under the coverage of a single satellite and operated by a satellite network operator. It relies on a ground segment and utilises some satellite on-board resources (through the satellite channels that are used). The ground segment is composed of a *user segment* and a *control and management segment*, which are also considered as parts of the space segment.

In the user segment, one finds *satellite terminals* (STs) connected to the end-user customer premises equipment (CPE), directly or through a LAN and *hub* or *gateway stations*, sometimes called *network access terminals* (NATs), connected to terrestrial networks:

— Satellite terminals are earth stations connected to CPE, sending carriers to or receiving carriers from a satellite. They constitute the satellite access points of a network; when the satellite network is a DVB-RCS network – digital video broadcast with return channel via satellite (designed according to the DVB-RCS standard) – satellite terminals are also called return channel satellite terminals (RCSTs).
— CPE is also called a *user terminals* (U) and includes equipment such as telephone sets, television sets, and personal computers and smartphones. UTs are independent of network technology and can be used for terrestrial as well as satellite networks.
— The *gateway earth station* (GW) provides internetworking functions between the satellite network and the Internet or a terrestrial network.

The control and management segment consists of:

— A mission and network management centre (MNMC) in charge of non-real-time, high-level management functions for all the satellite networks that are deployed in the coverage of a satellite.

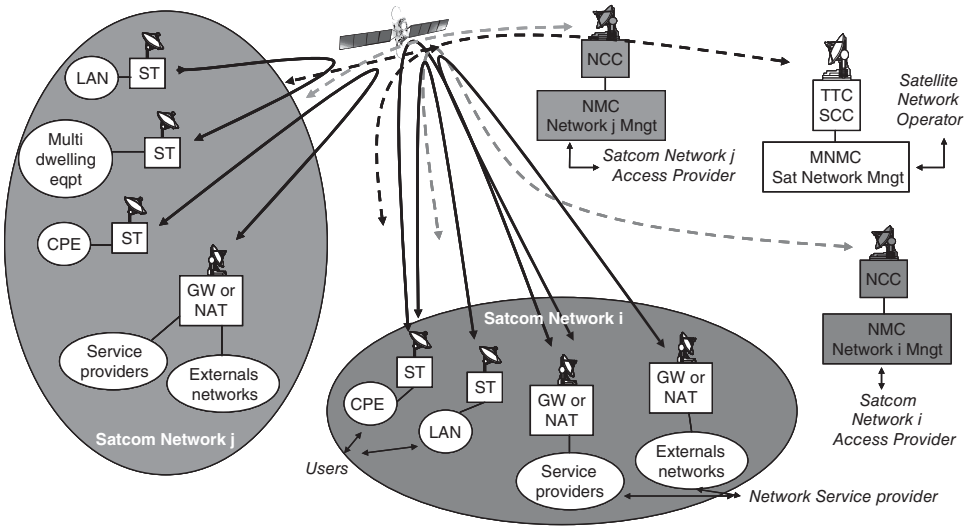


Figure 7.3 Satellite network components.

— Network management centres (NMCs), also called interactive network management centres (INMCs), for non-real-time management functions related to a single satellite network.
— Network control centres (NCCs) for real-time control of the connections and associated resources allocated to terminals that constitute one satellite network.

A satellite network (also called a *satcom network*) comprises a set of satellite terminals, one or more gateways, and one NCC that is operated by one operator and uses a subset of the satellite resources (or capacity). Figure 7.3 illustrates two satellite networks deployed under a satellite's coverage and the links that can be established between their constituents.

7.3 BASIC CHARACTERISTICS OF SATELLITE NETWORKS

Satellite networks are characterised by their topology (meshed, star, or multi-star), the types of link they support, and the connectivity they offer between earth stations.

7.3.1 Satellite network topology

7.3.1.1 Meshed network topology

In a meshed network, every node is able to communicate with every other node (Figure 7.4a). A *meshed satellite network* consists of a set of earth stations that can communicate with one another by means of satellite links consisting of radio-frequency carriers. Figure 7.4b shows an example of a meshed satellite network with three earth stations. As discussed in Chapter 6, this entails that several carriers can simultaneously access a given satellite (satellite access is frequency division multiple access [FDMA]), and subsets of these carriers simultaneously access a given transponder (transponder access can be FDMA, time division multiple access [TDMA], code division multiple access, [CDMA], or a combination).

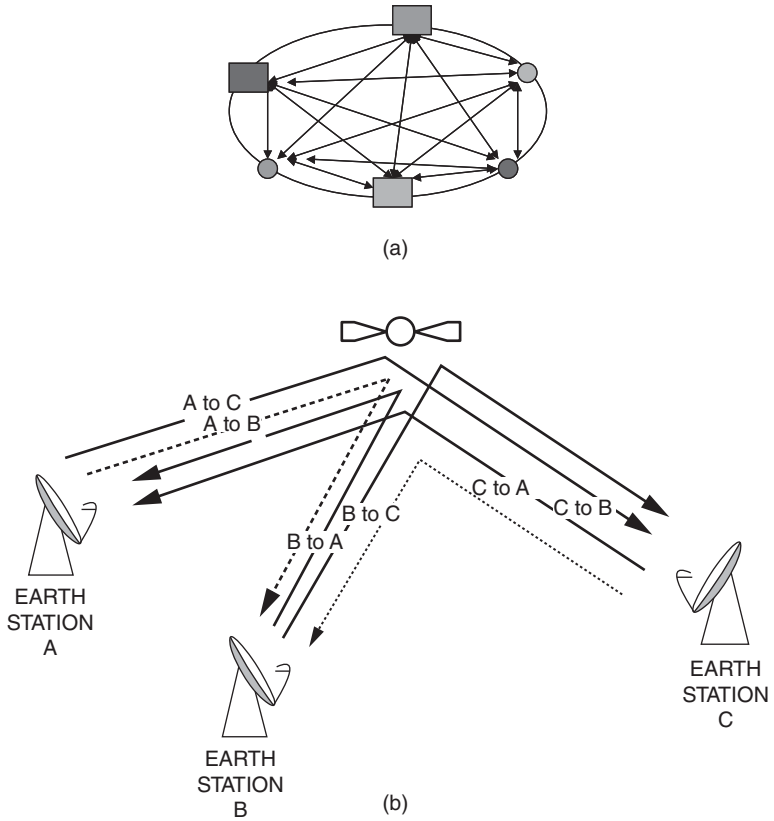


Figure 7.4 Meshed network topology: (a) in the abstract; (b) an example with three earth stations (arrows represent information flow as conveyed by the carriers relayed by the satellite).

A meshed satellite network can rely on a transparent or regenerative satellite. In the case of a transparent satellite, the radio-frequency link quality between any two earth stations in the network must be high enough to provide the end users with service achieving the target bit error rate (BER). This implies sufficient effective isotropic radiated power (EIRP) and *G/T* for each earth station, given the satellite transponder operating point. In the case of a regenerative satellite, the on-board demodulation of the signal puts fewer constraints on the EIRP and *G/T* of the earth stations.

7.3.1.2 Star network topology

In a star network, each node can communicate only with a single central node, often called the *hub* (Figure 7.5a). In a multi-star topology, several central nodes (hubs) are identified. The other nodes can communicate only with those central nodes (Figure 7.5b). A *star satellite network* consists of earth stations that can communicate only with a central earth station called the *hub*. Figure 7.5c provides an example of a star satellite network.

Generally, the hub is a large earth station (antenna size from a few meters to more than 10 m) with higher EIRP and *G/T* than the other earth stations in the network. A star network topology

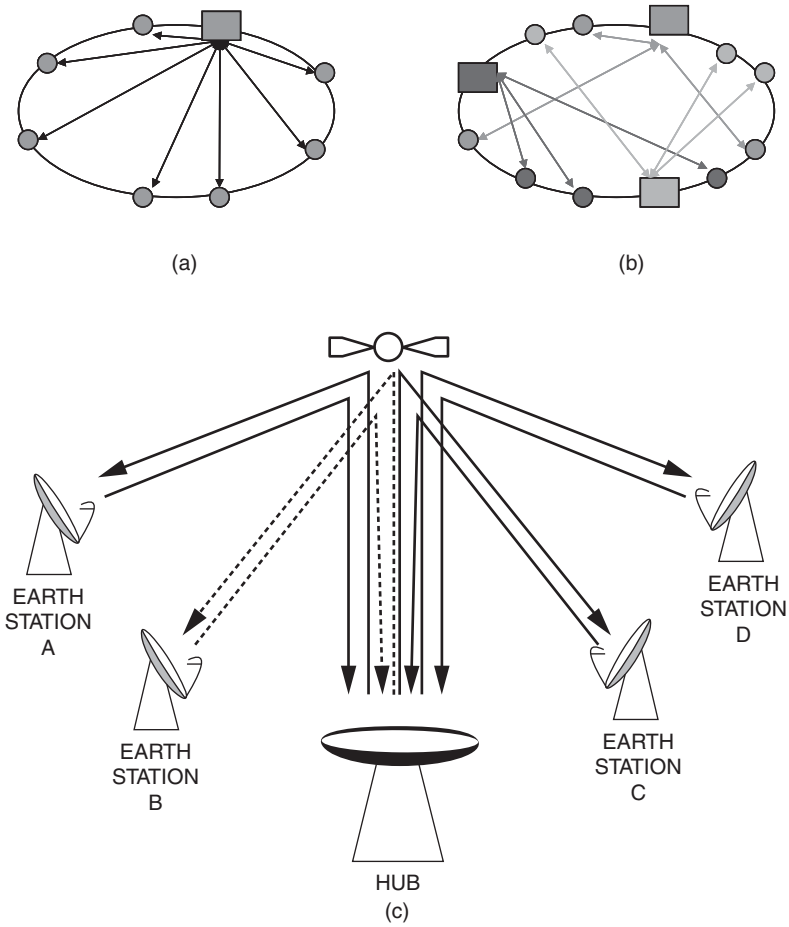


Figure 7.5 Star network topology: (a) a single-hub star; (b) a multi-star in the abstract; (c) an example with four earth stations and one hub (arrows represent information flow as conveyed by the carriers relayed by the satellite).

places fewer constraints on the EIRP and *G/T* of the earth stations than a meshed network topology relying on a transparent satellite, due to the fact that the earth stations communicate with a large earth station (the hub). This architecture is popular among networks populated with small earth stations (antenna size of about 1 m) called *very-small-aperture terminals* (VSATs) [MAR-04]. The link from any earth station to the hub is called an *inbound link* or *return link*. The link from the hub to the other earth stations is called the *outbound link* or *forward link*.

7.3.2 Types of link

Two types of link can be established through a satellite network: *unidirectional links*, where one or several stations only transmit and other earth stations only receive; and *bidirectional links*, where earth stations both transmit and receive. Unidirectional links are usually associated with a star topology in satellite broadcast-oriented networks. Bidirectional links can be associated with a star or meshed topology and are required to transport any two-way telecommunication services.

7.3.3 Connectivity

Connectivity characterises the way nodes of a network are connected to each other. Figure 7.6 presents the types of connectivity that can be encountered in telecommunications networks.

When a communication link is established through a satellite network, two levels of connectivity need to be distinguished: the connectivity required at the service level and the connectivity required on board the satellite. The connectivity at the service level defines the type of connection that is necessary between CPEs or network equipment, and between satellite terminals or gateways, to provide the service required by end users. This connectivity is principally processed on the ground and relies on identifiers associated with sessions and layer 3 and layer 2 connections.

Figure 7.7 illustrates two examples of communication services. The Internet access service is characterised by a star or multi-star topology with multipoint-to-point connectivity: customer traffic goes necessarily through a point of presence (POP), and the CPE is connected to the nearest POP of the user's ISP. A virtual private network (VPN) service is characterised by a meshed topology with point-to-point connectivity (multipoint to multipoint connectivity can also be requested for VPN multicast services). This service allows for interconnecting different LANs of a company to form a single LAN.

Satellite on-board connectivity defines how the satellite network resources are switched on board in order to meet the service-level connectivity requirements. It therefore depends on how the satellite resources (beams, channels, carriers, etc.) are organised on both satellite up- and downlinks and, primarily, on the type of coverage that the satellite system provides. In the case of global coverage, any user within the coverage can, in principle, be connected to any

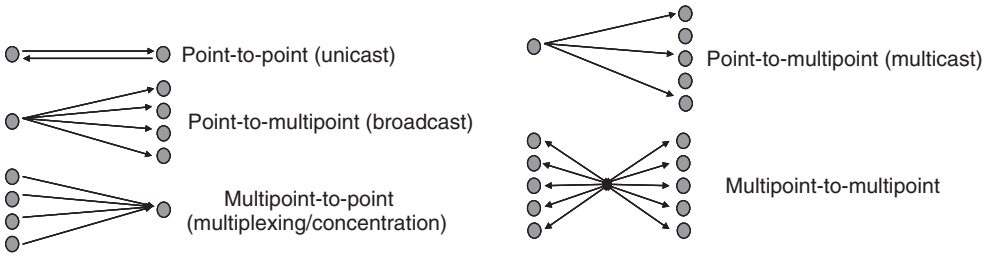


Figure 7.6 Types of network connectivity.

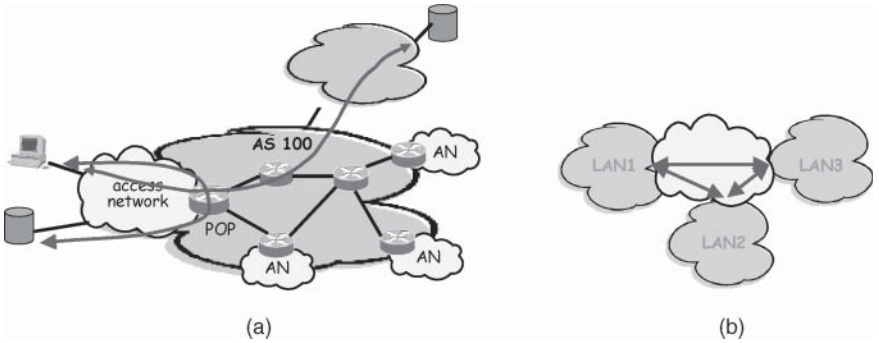


Figure 7.7 Communication services: (a) Internet access; (b) virtual private network (VPN).

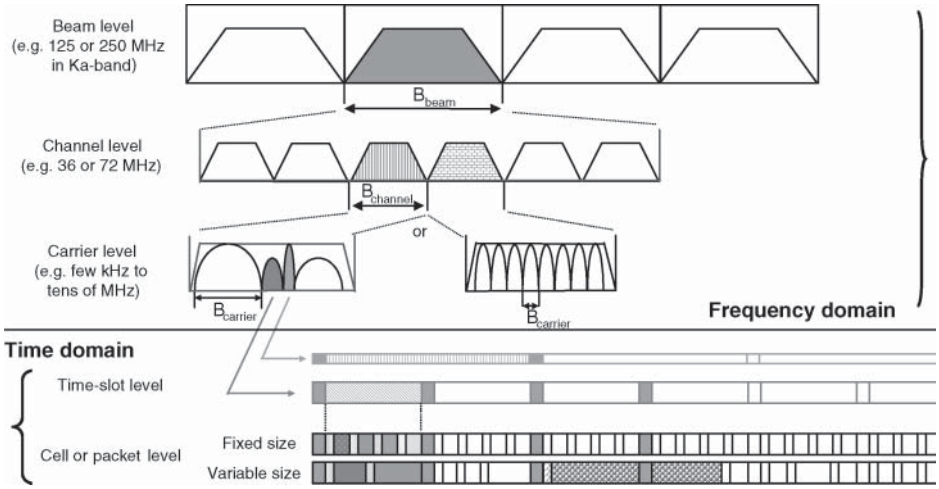


Figure 7.8 Types of resource that can be interconnected on board the satellite.

other user. In multibeam coverage, interconnection of any user within different beams of the coverage requires on-board interconnection of beams and of the resources allocated to those beams.

Figure 7.8 illustrates the different types of resources that may need to be interconnected on board the satellite. They correspond to different levels of granularity and imply different types of processing (see Section 7.4). Satellite on-board connectivity can be provided at the following levels:

— *Spot-beam:* The whole frequency resource allocated to a beam is switched on board. This can correspond to a channel or several channels (typically 125 or 250 MHz in Ka band).
— *Channel:* This is equivalent to the frequency resource that is classically transmitted through a transponder (typically 36 or 72 MHz).
— *Carrier:* This can be an FDMA carrier transmitted by a satellite terminal or earth station, or a multi-frequency time division multiple access (MF-TDMA) carrier shared by several satellite terminals (typically from a few kHz up to tens of MHz depending on the earth station radio capability).
— *Time slot:* This corresponds to time division multiplex (TDM) or TDMA time slots.
— *Burst, packet, or cell:* This corresponds to any type of layer 2 packet, up to IP datagrams. A burst is a group of packets or cells travelling together on the same link; a packet has a variable size and contains a block of data; and a cell has a fixed size.

7.4 SATELLITE ON-BOARD CONNECTIVITY

As discussed in Chapter 5, multibeam satellite systems make it possible to reduce the size of earth stations and hence the cost of the earth segment. Frequency reuse from one beam to another permits an increase in capacity without increasing the bandwidth allocated to the system. A satellite payload using multibeam coverage must be in a position to interconnect all network earth stations and consequently must provide interconnection of coverage areas.

Depending on the on-board processing capability and the network layer, different techniques are considered for interconnection of coverage:

— Transponder hopping (used when there is no on-board processing)
— On-board switching (used when there is transparent and regenerative processing)
— Beam scanning

7.4.1 On-board connectivity with transponder hopping

The band allocated to the system is divided into as many sub-bands as there are beams. A set of filters on board the satellite separates the carriers in accordance with the sub-band occupied. The output of each filter is connected by a transponder to the antenna of the destination beam. It is necessary to use a number of filters and transponders at least equal to the square of the number of beams. Figure 7.9 illustrates this concept for an example with two beams. According to the type of coverage, the earth stations must be able to transmit or receive on several frequencies and polarisations in order to hop from one transponder to another. Table 7.1 indicates the type of frequency agility required to ensure interconnection between beams according to the type of coverage. The capacity offered to traffic can be varied between beams, within the total capacity defined by the system bandwidth, by modification of the sub-band assignments, and hence by

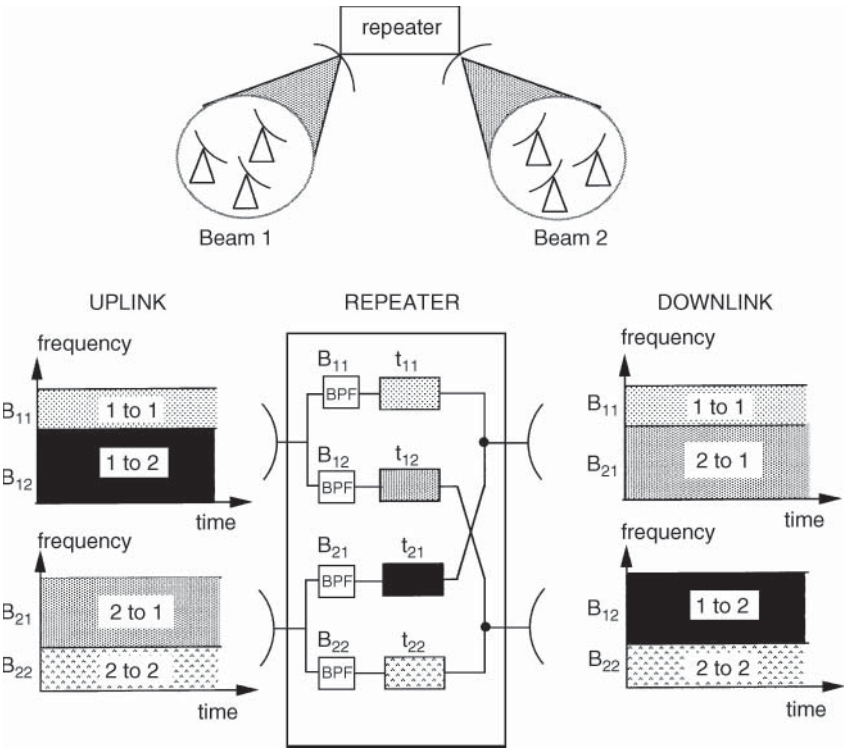


Figure 7.9 Interconnection by transponder hopping (two beams).

Table 7.1 Frequency agility required to ensure beam interconnection in accordance with the type of coverage.

Type of coverage		
Uplink	Downlink	Frequency and polarisation agility
Global	Global	On transmission or reception
Spot	Global	On reception
Global	Spot	On transmission
Spot	Spot	On transmission and reception

modification of the connections between input filters and transponders. This operation is realised by telecommand, from time to time, in accordance with long-term fluctuations in traffic.

7.4.2 On-board connectivity with transparent processing

Beam switching by transponder hopping is a solution when the number of beams is low. Because the number of transponders increases at least as the square of the number of beams, with a large number of beams the satellite payload becomes too complex and too heavy. It is therefore necessary to consider on-board switching at a lower granularity, and shift from beam switching to channel switching. Two types of technology can provide this kind of connectivity: analogue technology using an intermediate frequency-switching matrix, one example of which is known as satellite switched/TDMA (SS/TDMA), and digital technology using baseband (BB) processing equipment, in particular digital transparent processors (DTP).

7.4.2.1 Analogue transparent switching

The principle of analogue transparent switching is illustrated in Figure 7.10. The payload includes a programmable switching matrix having a number of inputs and outputs equal to the number of beams. This matrix connects each uplink beam to each downlink beam by way of a receiver and a transmitter. The number of repeaters is thus equal to the number of beams. The distribution control unit (DCU) associated with the switch matrix establishes the sequence of connection states between each input and output during a period of time in such a way that the carriers arriving at the satellite in each beam are routed to the destination beams.

When the period of time separating two connection states is a frame, since interconnection between two beams is cyclic, stations must store traffic from users and transmit it in the form of bursts when the required interconnection between beams is realised. This technique can thus be used in practice only with digital transmission and access of the TDMA type. This is why it is called *satellite-switched* time division multiple access (SS-TDMA). The granularity of the connectivity provided by the satellite-switched technique is a time slot of a high-capacity frequency carrier.

7.4.2.1.1 Frame organisation

Figure 7.11a shows the organisation of a frame for a three-beam satellite. The frame contains a synchronising field and a traffic field. Bursts from traffic stations are routed to their destinations in the traffic field. The traffic field contains a succession of switch states. During a given

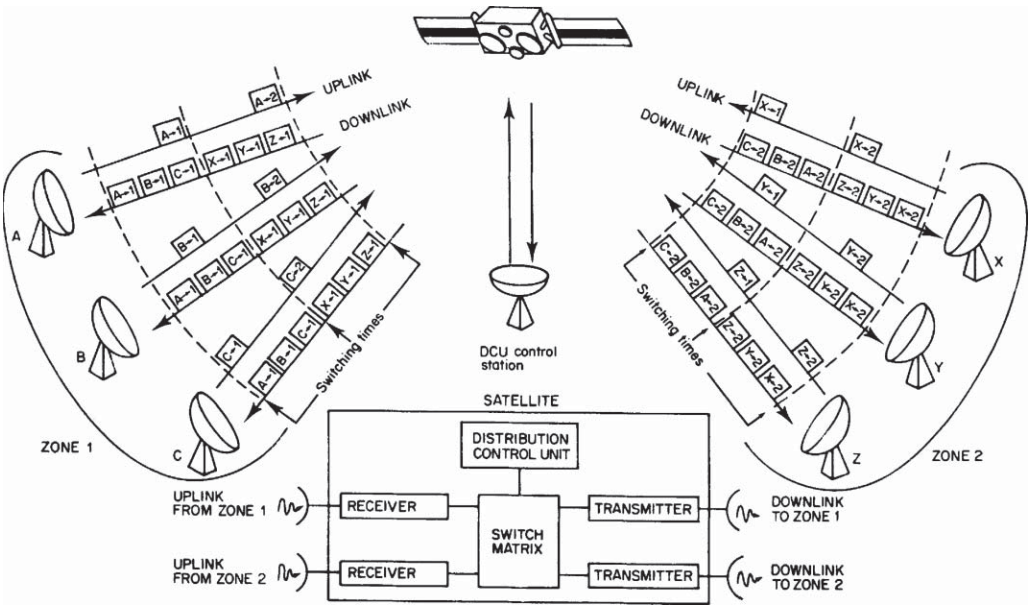


Figure 7.10 Analogue transparent on-board switching (SS-TDMA).

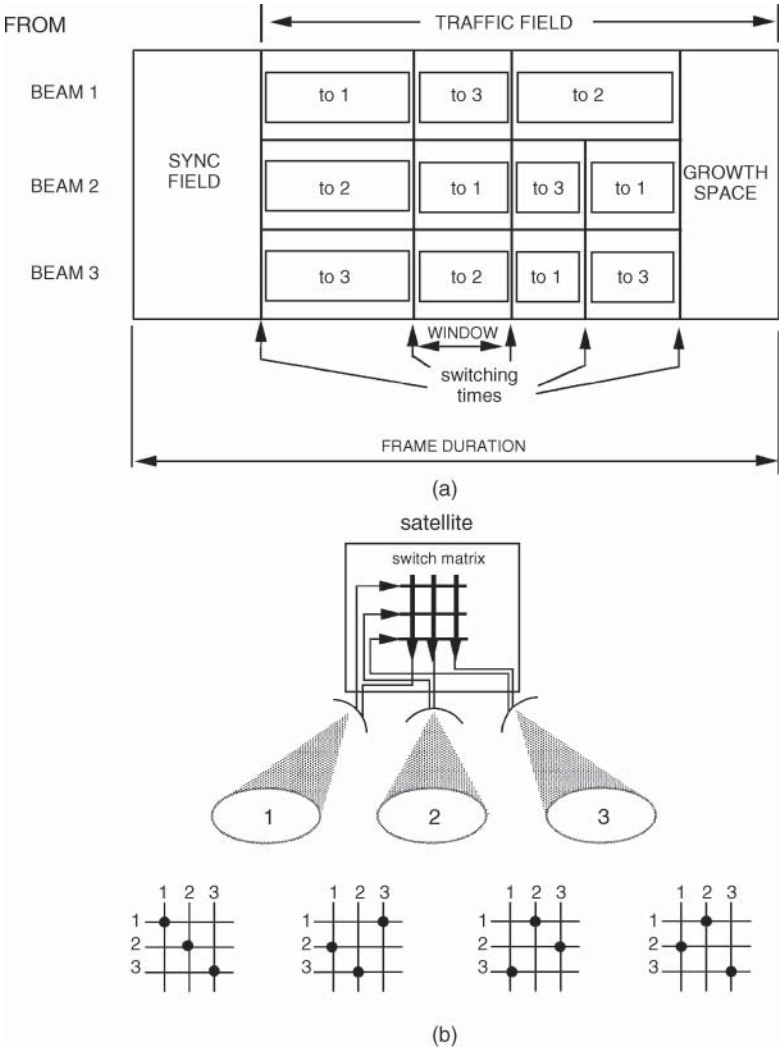


Figure 7.11 Three-beam SS-TDMA satellite: (a) frame organisation; (b) switch state sequence during the active part of the traffic field.

switch state, the switching matrix retains the same connection state. The traffic field also contains a growth space in case traffic demand is less than the capacity. The duration of a connection between an up beam and a down beam is called a *window*. A window can extend over the duration of several switch states. Figure 7.11b shows the switch state sequence implemented by the switch matrix in order to route traffic according to the frame organisation of Figure 7.11a.

7.4.2.1.2 Window organisation

Figure 7.12 shows how bursts are positioned in a window time interval. The figure shows bursts that are transmitted by stations A, B, and C in the window corresponding to a connection from

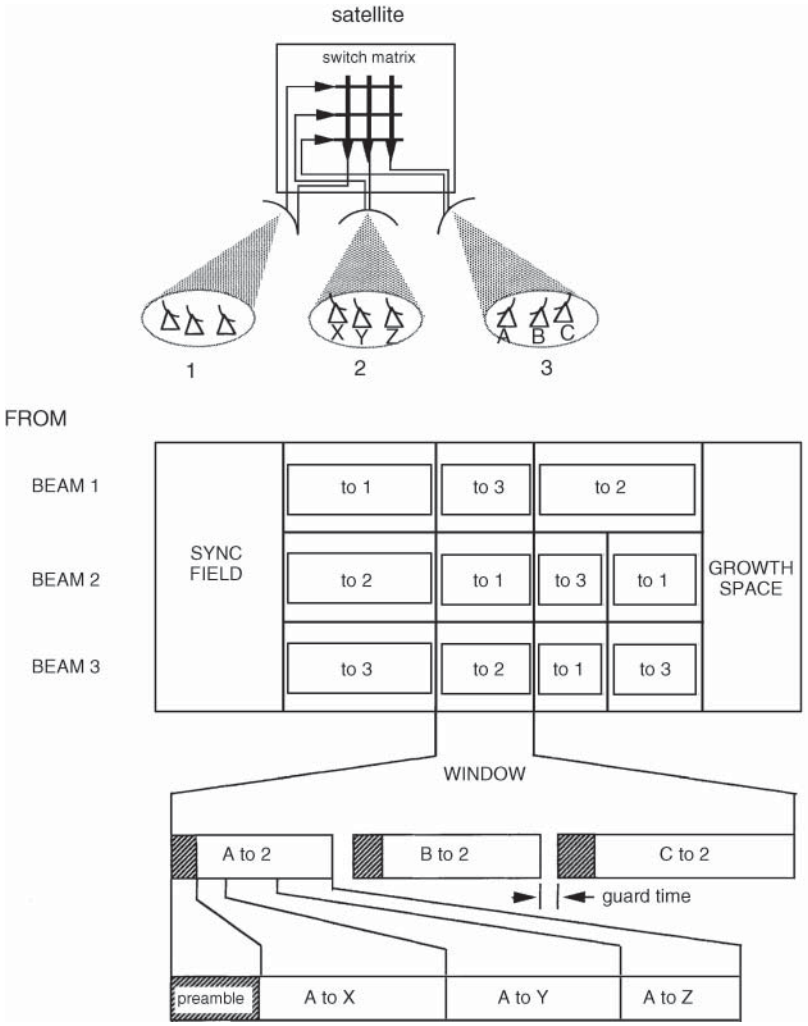


Figure 7.12 Window organisation.

beam 3 to beam 2. Each burst transmitted by a station during the window time considered consists of several sub-bursts that contain station-to-station information.

7.4.2.1.3 Assignment of packets in the frame – burst time plan (BTP)

The assignment of bursts in the frame, called the *burst time plan* (BTP), should maximise the use of the satellite transponders. Transponders are exploited best when the windows are occupied entirely by traffic bursts. This is possible only when traffic distribution between beams is balanced. In practice, this is not always the case. A traffic matrix can be established that describes the traffic demand from one beam to another. For example, for a three-beam satellite (1, 2, and 3), the matrix is as shown in Table 7.2, where t_{XY} represents the traffic demand from beam X to beam Y.

Table 7.2 Traffic demand matrix for a three-beam satellite.

	To beam	1	2	3	
From beam	1	t_{11}	t_{12}	t_{13}	S_1
	2	t_{21}	t_{22}	t_{23}	S_2
	3	t_{31}	t_{32}	t_{33}	S_3
		R_1	R_2	R_3	

In Table 7.2, the sum of each row $S_i (i = 1, 2, 3)$ represents the traffic uplinked by all stations in beam i. The sum of each column $R_j (j = 1, 2, 3)$ represents the beam traffic downlinked in beam j. In the case of balanced traffic distribution between beams, the sums S_i and R_j are equal. Otherwise, one of these sums is greater than all the others, and the corresponding line of the matrix (row or column) is called the *critical line*.

It can be shown that the minimum time to route the traffic bursts transmitted by all stations is the time required to transmit the traffic of the critical line of the traffic matrix at the rate considered [INU-81]. Numerous algorithms permit filling of frames with bursts in such a way that the active traffic field duration is at a minimum. A classification of these algorithms is given in [MAR-87].

An SS-TDMA network can operate with fixed or demand assignment. With demand assignment, variations of capacity allocated to stations are obtained by variations of burst length as with TDMA (Section 6.6). Variation of station burst length is accompanied by variation of the position of bursts of other stations, and consequently a change in the assignment of other bursts (BTP change). For changes that involve a sequence of switch states, the new sequence of switch states must be loaded into the DCU memory by means of a dedicated link (which can be the telecommand link). The assignment changes occur at the start of a *superframe* consisting of an integer number of TDMA frames in order to guarantee synchronisation of the change among all stations and the satellite.

7.4.2.1.4 Frame efficiency

The definition of efficiency is given in Section 6.6.5. The expression for it is:

$$\eta = 1 - \Sigma t_i / T_F \tag{7.1}$$

where Σt_i represents the sum of the times not devoted to information transmission (dead times) and T_F is the frame duration. There are five components of Σt_i:

— The synchronisation field.
— The packet headers and guard times, including the intervals reserved for switching of the on-board matrix (a station must transmit several times within a frame if its packets are destined for several beams; as each packet has a header, there are more dead times).
— Unfilled frame windows due to unbalanced traffic distribution among the beams (the critical line determines the minimum duration of the switching modes; some frame windows will not be filled, thereby leaving transponders inactive; see Figure 7.13).
— Less than optimum burst assignment at a given instant. Dead times may appear in windows until a BTP change occurs.
— Growth space, if traffic demand in the beam corresponding to the critical line is less than the capacity of a transponder.

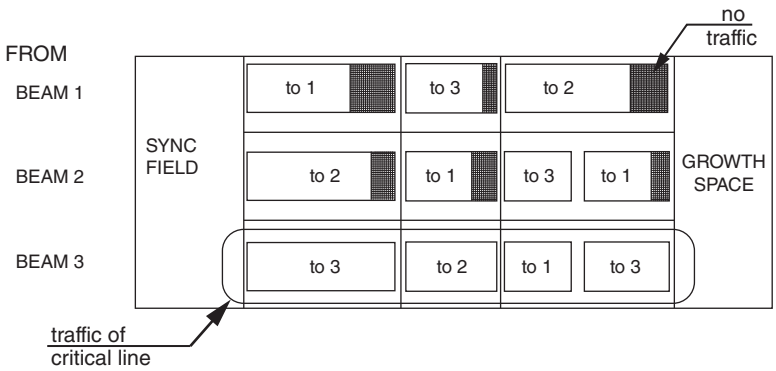


Figure 7.13 SS-TDMA frame packing in the case of non-uniform traffic distribution. The critical line of the matrix corresponds to S_3 – that is, the traffic uplinked on beam 3 determines the minimum duration of the sequence of switching modes.

Overall, it is difficult to provide values since throughput depends on traffic distribution. Simulations have indicated values of 75–80% [TIR-83]. For every assumption, the efficiency is less than with a single-beam satellite.

7.4.2.2 Digital transparent switching

When connectivity is required at granularity smaller than a channel, analogue technology may not be efficient, because it leads to an increased payload complexity. Digital technology, relying on digital filtering and switching, can be introduced. Figure 7.14 illustrates the principle of a DTP that enables the switching of uplink carriers from one spot beam to another spot beam and the transposition of frequency.

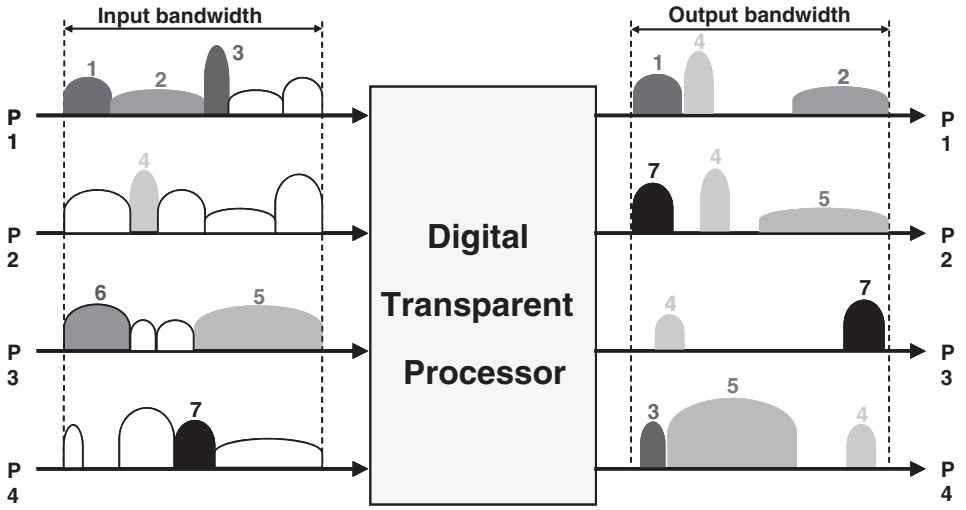


Figure 7.14 Principle of a digital transparent processor.

7.4.3 On-board connectivity with regenerative processing

The availability on board the satellite of binary digits obtained after demodulation and decoding offers several opportunities, and in particular allows the introduction of some layer 2 switching on board the satellite. The link budget issues with regenerative payloads are discussed in Chapter 5. The payload implementation is discussed in Chapter 9.

7.4.3.1 Baseband switching

The availability of bits on board the satellite at the output of the uplink carrier demodulators permits switching between receiving and transmitting antennas to be no longer at radio frequency (RF) but at BB. Implementations of the BB switching device are presented in Section 9.4.5. The constraint of immediate routing of received information to the destination downlink disappears. This permits earth stations to transmit all their information in the same burst and hence to transmit only a single burst per frame. The number of bursts per frame is reduced, and the efficiency of the frame increases.

Figure 7.15 illustrates the variety of resources and associated granularity that can be switched on board the satellite when it embarks on regenerative processing.

7.4.3.2 Rate conversion

A change of rate between the uplink and the downlink is not possible with a transparent satellite. Stations can, therefore, be interconnected only by carriers of the same capacity, and this can be restricting. For example, interconnection of large stations carrying high-rate trunk traffic and small stations (VSATs) with low-rate traffic implies a terrestrial connection and a double hop, as shown in Figure 7.16a. In contrast, by virtue of on-board processing, the traffic between networks with different data rates can be switched at BB and combined before transmission on the various downlinks in accordance with their destination and independent of the capacity of the carrier (Figure 7.16b) [NUS-86]. To avoid bulk and excessive power consumption, only high-rate carriers that contain traffic destined for low-rate stations are routed to a high-rate demodulator. The other high-rate carriers are switched at RF and are, therefore, not demodulated.

7.4.3.3 FDMA/TDM systems

Regeneration, compared with a transparent satellite system, permits a reduction in the cost of earth stations by reducing the EIRP and the *G/T* of the stations, implementing FDMA on the uplink and TDM on the downlink. Such a scheme allows continuous transmission by the earth stations on the uplink, and transmission at saturation by the satellite amplifier on the downlink without inter-modulation noise generation. This advantage is exposed in connection with Option 4 of satellite broadcasting (Section 7.6.4).

7.4.3.3.1 Earth station EIRP reduction

The uplink is often over dimensioned so that the performance in terms of C/N_0 of the total link is determined by that of the downlink, which is limited by the power available on board the satellite. With a transparent repeater, it is necessary to provide a ratio $\alpha = (E/N_0)_U/(E/N_0)_D = (C/N_0)_U/(C/N_0)_D$ of the order of 10 dB. With a regenerative repeater, the curves in Figure 5.40 show that the error probability on the uplink becomes negligible in

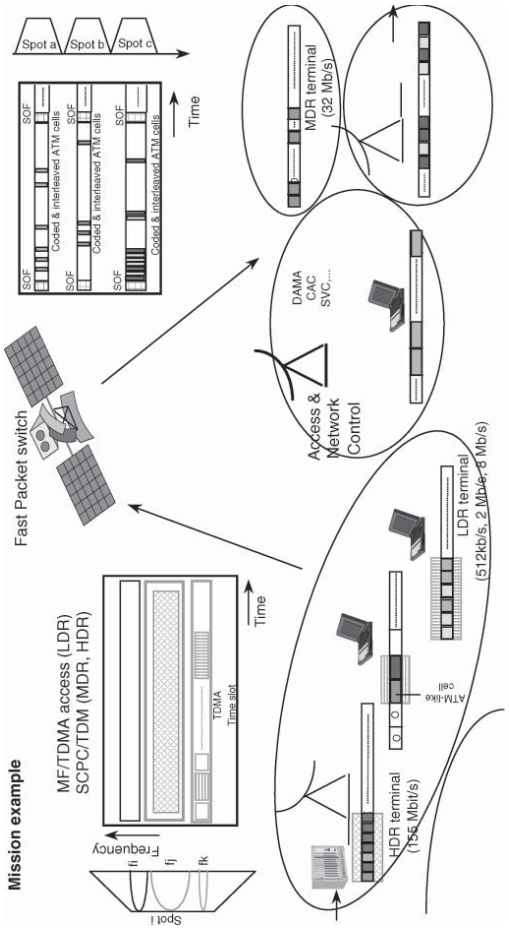


Figure 7.15 Resources switched with digital on-board processing (HDR/MDR/LDR – high/medium/low data rate).

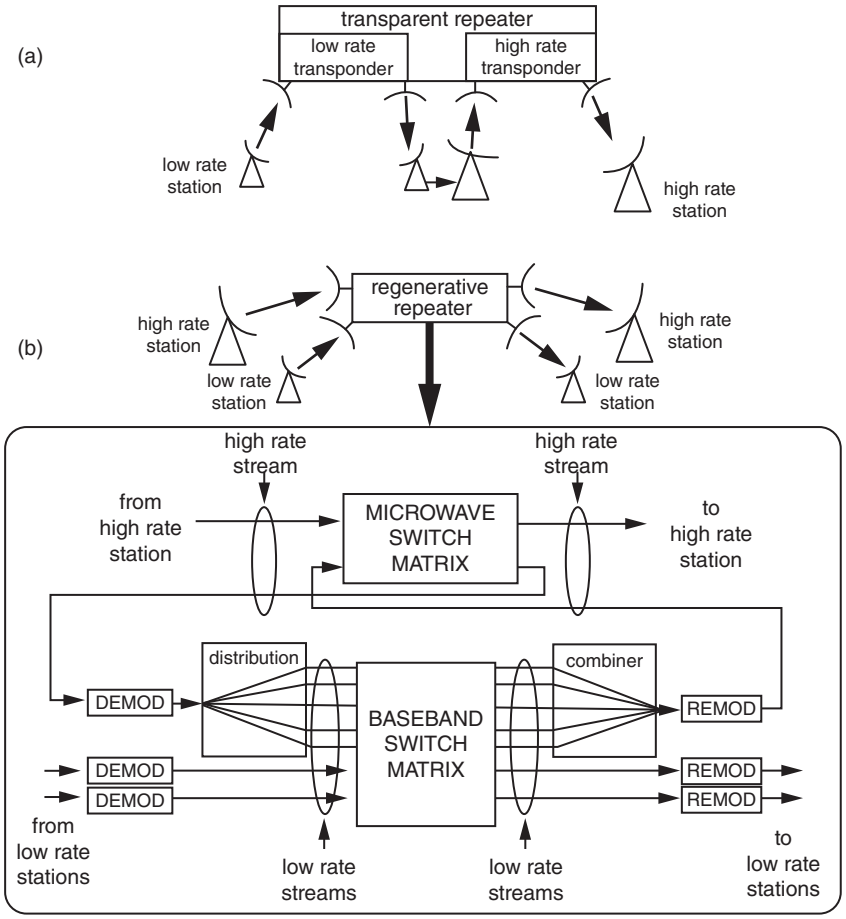


Figure 7.16 Interconnection of two networks with carriers of different capacities: (a) transparent satellite; (b) regenerative satellite.

comparison with that of the downlink when the value of the ratio a is greater than around 2 dB. This reduction of the ratio a translates into a reduction in the EIRP of the earth station and, consequently, its cost.

Another factor also permits reduction of the EIRP. As a consequence of the bit storage in the BB switching matrix, the earth stations can transmit continuously on different frequencies (FDMA multiple access), as shown in Figure 7.17. In comparison with TDMA, the rate transmitted by each station is less. The situation is:

— With TDMA:

$$(C/N_0)_U = (E/N_0)R_{TDMA}$$

— With FDMA:

$$\begin{aligned}(C/N_0)_U &= (E/N_0)R_{FDMA}\\ R_{FDMA} &= R_{TDMA}(T_B/T_F)\end{aligned} \tag{7.2}$$

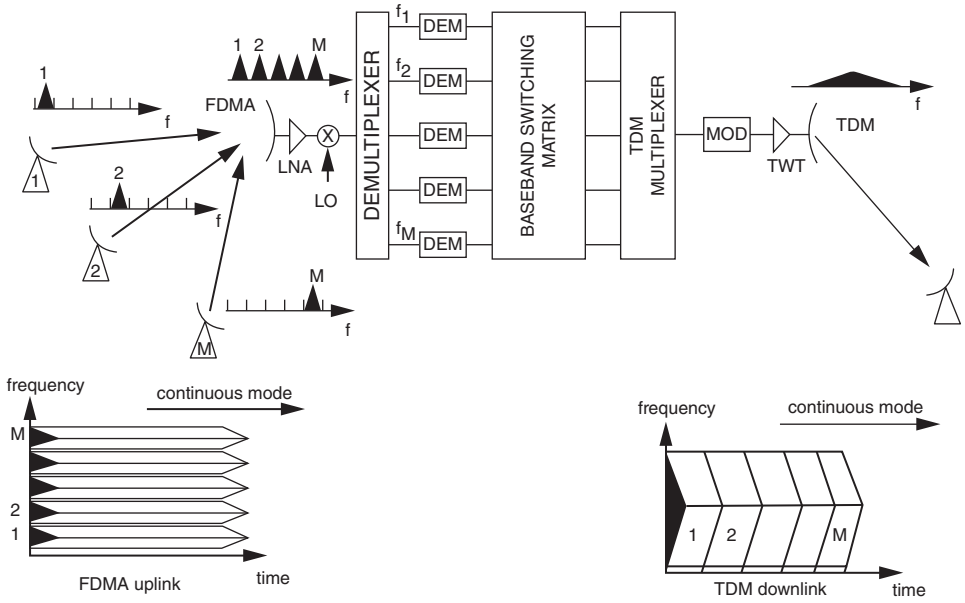


Figure 7.17 Uplink FDMA and TDM on a single downlink carrier using a regenerative repeater.

where T_B is the duration of the burst transmitted in TDMA and T_F is the frame duration. As T_B/T_F is less than 1, it can be seen that the value of C/N_0 required with FDMA is less than with TDMA.

7.4.3.3.2 Earth station G/T reduction

As shown in Figure 7.17, the satellite amplifier (TWT, if a travelling wave tube is used) handles a single carrier modulated by the multiplex of the bits destined for stations on the beam considered. In comparison with a transparent FDMA system, this amplifier operates at saturation without generating intermodulation noise on the downlink. The downlink benefits from the maximum EIRP of the satellite and the absence of intermodulation noise, so the figure of merit G/T of the earth stations can be reduced. In comparison with a transparent TDMA system, the received bursts are generated from a single carrier that originates from the satellite oscillator instead of being transmitted by different stations equipped with independent oscillators. It is, therefore, no longer necessary to provide earth station receivers with fast carrier and bit-rate acquisition circuits.

7.4.3.4 Conclusion

On-board processing offers several advantages, demonstrated here and in Chapter 5. Regenerative on-board processing payloads tolerate a higher level of interference than transparent payloads and make the earth stations simpler. However, the advantages gained have to be balanced with the additional on-board complexity and its impact on payload reliability. Moreover,

the implementation of fast computing circuitry on board the satellite places heavy demands on the power consumption of the satellite payload, and hence the satellite mass.

Finally, on-board processing payloads imply some form of a priori selection of transmission formats that make the payload specific to predefined types of service that may not turn out to be popular once the satellite is in operation. This also poses the problem of coping with unexpected changes in traffic demand (both volume and nature) and new operational procedures. This issue has been dissuasive for many years to satellite operators, who tend to prefer minimising risk rather than acquiring substantial but uncertain advantages. This constraint could be overcome using programmable hardware, such as software-defined payload.

7.4.4 On-board connectivity with beam scanning (BFN – beam-forming network)

Each coverage area is illuminated cyclically by an antenna beam whose orientation is controlled by a BFN that is part of the antenna subsystem on board the satellite. The area stations transmit or receive their bursts when the area is illuminated by a beam. Interconnection by beam scanning can be considered both with transparent and regenerative payloads.

7.4.4.1 Scanning beams with transparent payload

In the absence of on-board storage, at least two beams are necessary at a given instant – one to establish the uplink and one to establish the downlink (Figure 7.18). The illumination duration is proportional to the volume of traffic to be carried between the two areas.

7.4.4.2 Scanning beams with regenerative payload

Dynamic real-time forming of antenna beams permits consideration of single-beam satellites with a beam that sequentially scans the various regions of the service zone (Figure 7.19). The set of dwell areas that are covered sequentially by the beam form the coverage area of the system. When

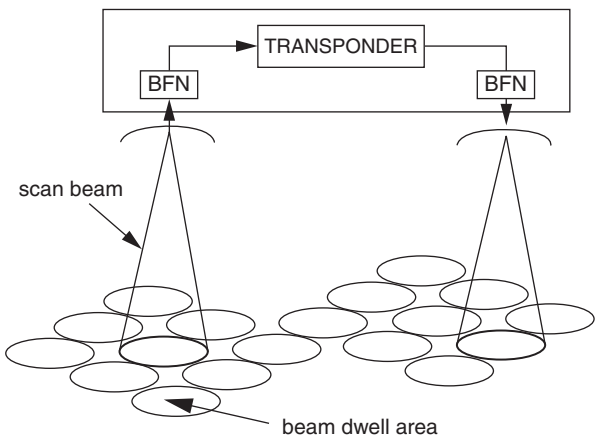


Figure 7.18 Interconnection by scanning beams.

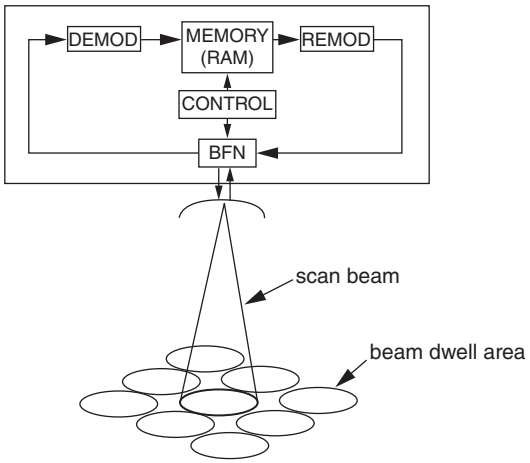


Figure 7.19 Single-beam, regenerative, scanning satellite network.

the beam is in a given dwell area, the information destined for stations in the area is extracted from the on-board memory and transmitted in multiplexed form. Simultaneously, these stations transmit information destined for stations in other dwell areas. This is stored in the on-board memory for later transmission at the time when the beam passes over the destination area.

An inherent advantage of this type of system is the disappearance of fixed simultaneous beams and hence of co-channel interference (CCI).

7.5 CONNECTIVITY THROUGH INTERSATELLITE LINKS (ISLs)

ISLs can be considered particular beams of multibeam satellites; the beams in this case are directed not towards the earth but towards other satellites. For bidirectional communication between satellites, two beams are necessary – one for transmission and one for reception. Network connectivity implies the possibility of interconnecting beams dedicated to intersatellite links and other links at the payload level.

Three classes of intersatellite link can be distinguished:

— Links between geostationary earth orbit (GEO) and low earth orbit (LEO) satellites (GEO–LEO links), also called inter-orbital links (IOL)
— Links between geostationary satellites (GEO–GEO)
— Links between low orbit satellites (LEO–LEO)

7.5.1 Links between geostationary and low earth orbit satellites (GEO–LEO)

This type of link serves to establish a permanent relay via a geostationary satellite between one or more earth stations and a group of satellites proceeding in a LEO at an altitude of the order of 500–1000 km. For economic and political reasons, one does not wish to install a network of stations that is so large that at every instant the passing LEO satellites are visible from at least one station. One or more geostationary satellites are therefore used; they are permanently and simultaneously visible both from stations and LEO satellites and serve to relay communications. This technique also permits overcoming possible limitations of the terrestrial network.

This concept is presently operated in the NASA tracking network by means of the tracking and data relay satellites (TDRSs) which, in particular, provide communication with the International Space Station. A European programme has successfully launched a data relay payload (Artemis satellite) to provide communications between the ground and LEO spacecraft.

7.5.2 Links between geostationary satellites (GEO–GEO)

7.5.2.1 Increasing the capacity of a system

Consider a multibeam satellite network. Figure 7.20 illustrates the case of a three-beam satellite (Figure 7.20a). It is assumed that the traffic demand increases and exceeds the capacity of the satellite. One solution is to launch a replacement satellite of greater capacity, and this implies risks, development costs, and the availability of a suitable launcher; alternatively, a second satellite identical to the first could be launched with the traffic shared between the two satellites. To avoid interference, the two satellites must be in sufficiently distant orbital positions but not too distant so as to provide sufficiently large common coverage. To ensure interconnectivity among all stations, it is necessary to equip all stations with two antennas, each pointing towards a different satellite (Figure 7.20b). With satellites provided with intersatellite transponders, one can implement the following scenarios:

— Equip the stations of region 1, assumed to be generating the excess traffic, with a second antenna, and retain the same configuration for the stations of regions 2 and 3 (Figure 7.20c). The intersatellite link carries the excess traffic of region 1.
— Distribute the stations, each with a single antenna, into two groups, each associated with one satellite (Figure 7.20d). The intersatellite link carries the traffic between the two groups.

The choice is economic and depends on the specific situation.

7.5.2.2 Extending the coverage of a system

An intersatellite link permits earth stations of two networks to be interconnected and hence the geographical coverage of the two satellites to be combined (Figure 7.21a). The other solutions are:

— Install an interconnecting earth station equipped with two antennas in the common part of the two coverages, if it exists (Figure 7.21b).
— Make a connection, by means of the terrestrial network, from the stations of one network to a station of the other network situated on the common border of the two coverages (Figure 7.21c).

7.5.2.3 Increasing the minimum elevation angle of earth stations

Long-distance links by a single satellite require earth stations with a small elevation angle, sometimes less than 10°. This causes a degradation of G/T for the receiving station (see Section 5.5.3) and increases the risk of interference with terrestrial microwave relays. If the link passes through two geostationary satellites connected by an intersatellite link, the elevation angle increases.

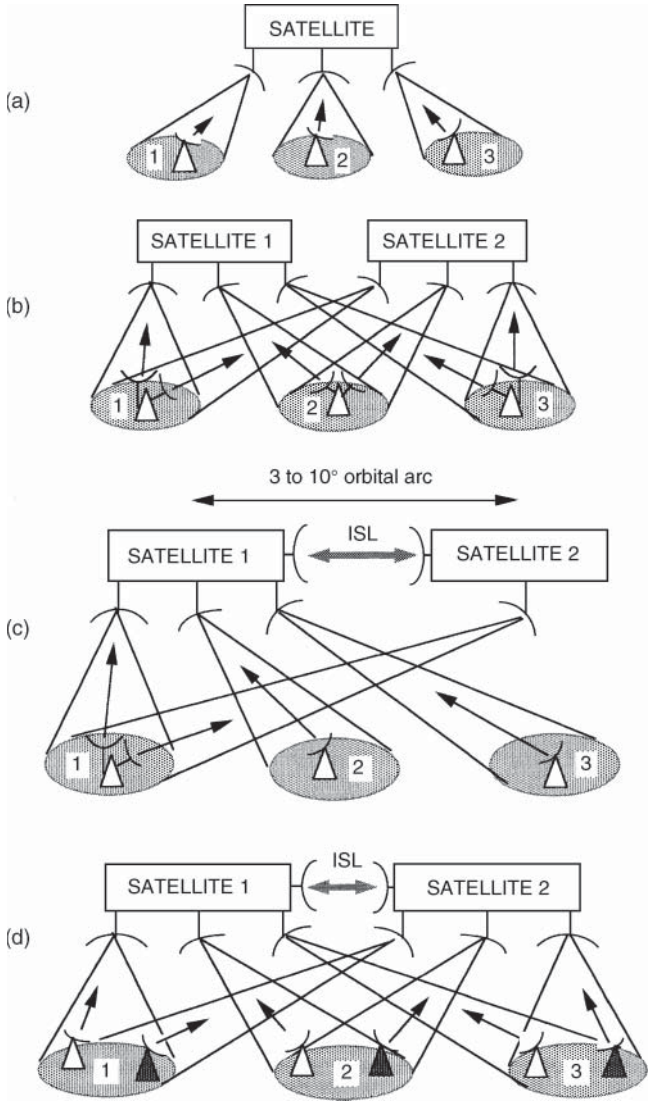


Figure 7.20 Using an intersatellite link to increase the capacity of a system without heavy investment in the earth segment: (a) network with a single satellite; (b) a second satellite is launched to increase the capacity of the space segment – the stations must be equipped with two antennas; (c) with an intersatellite link, only the stations of the most heavily loaded region must be equipped with two antennas; (d) the stations are distributed between the two satellites; the intersatellite link carries the traffic between the two groups of stations.

Hence, a link by a single satellite with an elevation angle of 5° becomes, with two satellites separated by 30°, a link with an elevation angle of 20° for equatorial stations (Figure 7.22) and 15° for stations at a latitude of 45°. This would be the case between London and Tokyo, for example, with two satellites above the Indian Ocean.

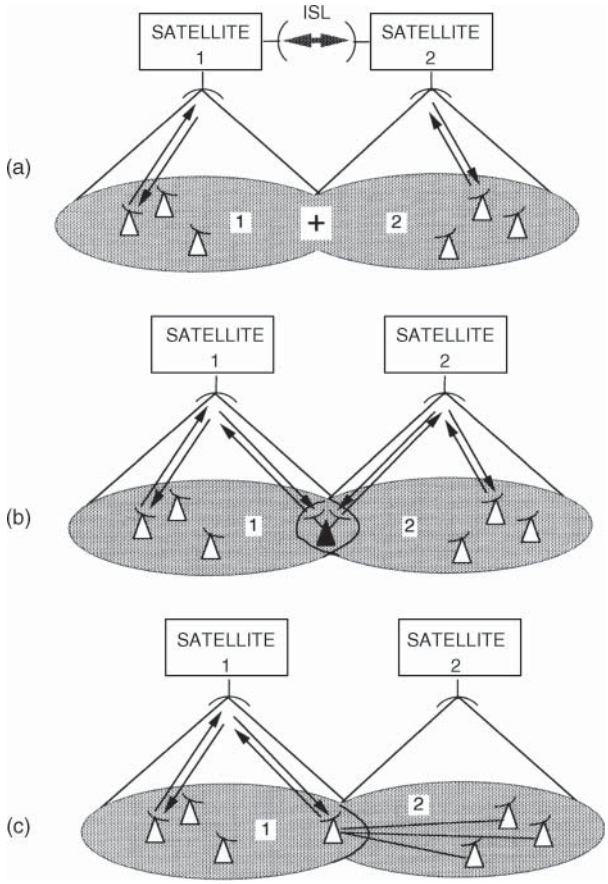


Figure 7.21 Extension of system coverage: (a) interconnecting the stations of each coverage by an intersatellite link; (b) interconnecting, without an intersatellite link, by a station common to the two networks; (c) interconnecting, without an intersatellite link, by a terrestrial network.

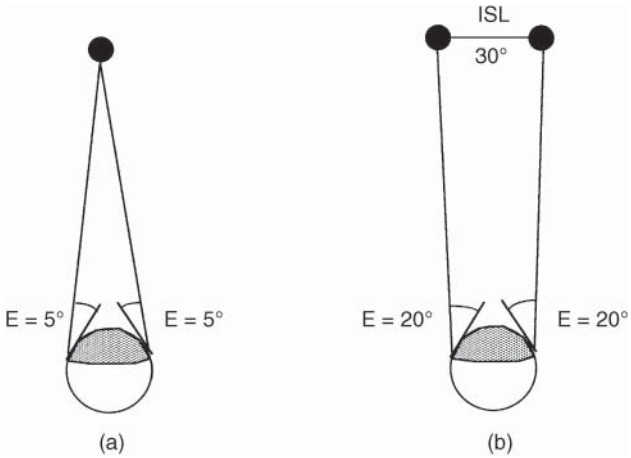


Figure 7.22 Increasing the minimum elevation angle of earth stations.

7.5.2.4 Reducing the constraints on orbital position

The orbital position of a satellite is often the result of a conflict, resolved by means of a procedure called *co-ordination*, between the desire of the satellite operator to ensure coverage of the service area under the best conditions and the need to avoid interference with established systems. The problem becomes acute above continents and particularly for the orbital arc above the American continent. Intersatellite links, when they permit traffic to be shared among several satellites in different orbital positions, provide the operator with some latitude in the positioning of satellites. Figure 7.23 shows an example of a solution by positioning two satellites at the extremities of the congested arc while guaranteeing coverage of the whole of the United States to the operator [MOR-89].

7.5.2.5 Satellite clusters

The principle is to locate several separate satellites in the same orbital position with a separation of tens of kilometres and interconnection by intersatellite links. This concept was proposed many years ago [VIS-79; WAD-80; WAL-82]. The satellites are thus all in the main lobe of an earth station antenna and appear equivalent to a single large-capacity satellite that would be too large to be launched by an existing launcher. The cluster is formed by successive launching

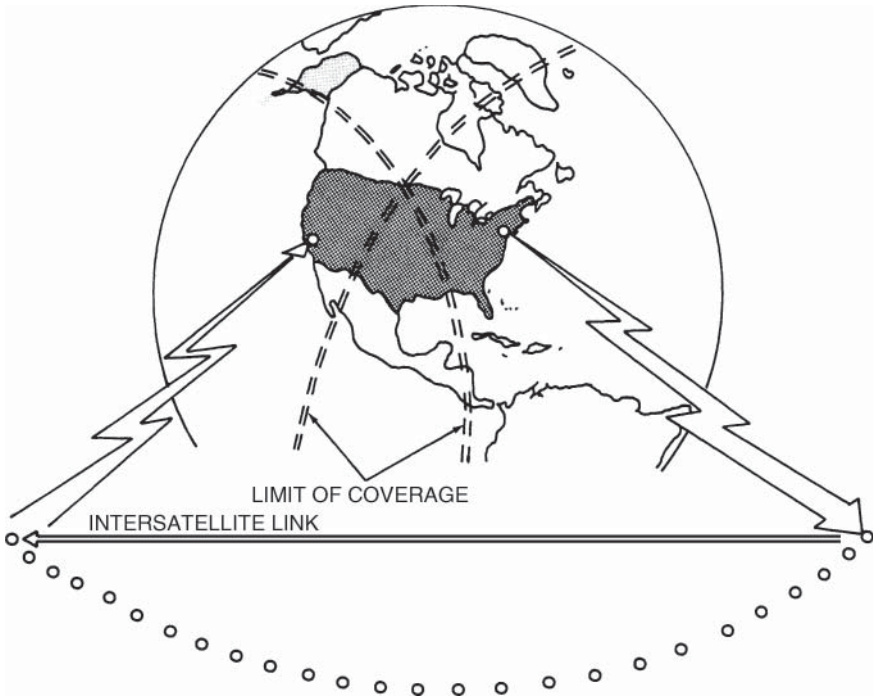


Figure 7.23 Complete coverage of the United States in spite of saturation of the orbital arc [MOR-89]. Source: reproduced with permission from Morgan, N.L. and Gordon, G.D. (1989). *Communications Satellite Handbook*, © John Wiley & Sons.

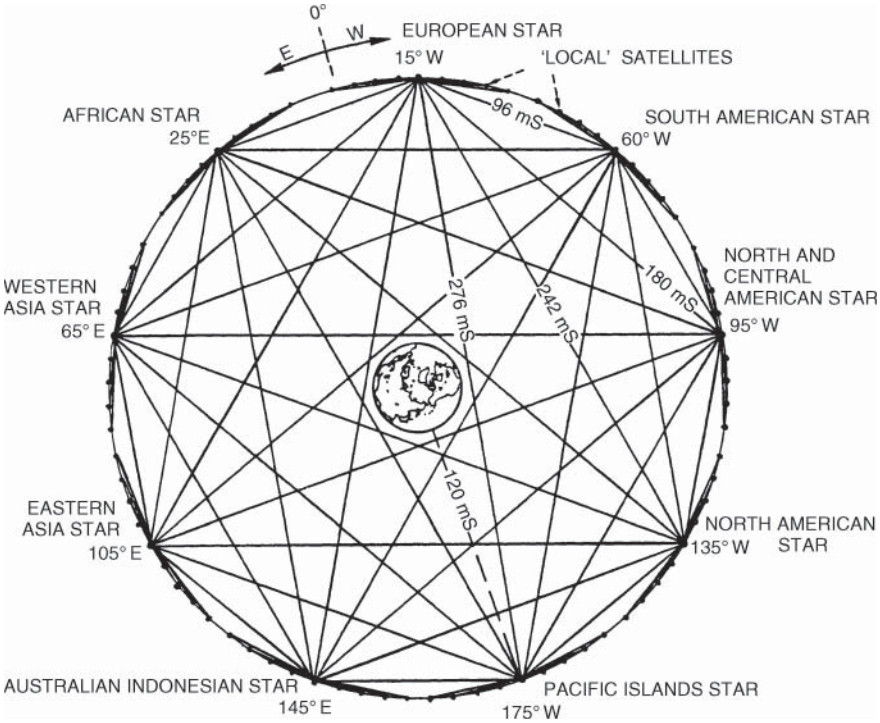


Figure 7.24 A global network [GOL-82]. Source: reproduced with the permission of the American Institute of Aeronautics and Astronautics.

of the satellites into the same place close to each other. As all the satellites are subjected to the same perturbations, orbital control is simplified, although station-keeping manoeuvres should be carefully phased. In the case of breakdown of a satellite, it can be replaced in the cluster. Finally, the configuration of the cluster can be modified in accordance with traffic demands. This concept of co-located satellites is used extensively for satellite broadcasting in order to make use of the wide radio-frequency spectrum available at Ku band. The bandwidth (about 2 GHz) is distributed between several tens of satellite channels implemented on several co-located satellites that are all seen in the beam of the user antenna, allowing delivery of hundreds of TV programmes. However, none of these clusters use intersatellite links. The concept may regain interest in the near future, as an alternative to the use of very large and powerful GEO satellites, considering clusters of small GEOs with standardised functions, possibly reconfigurable, that could be produced in large numbers at reduced cost.

7.5.2.6 A global network

Figure 7.24 shows the design of a global network based on nine geostationary Star satellites, which establish a basis for worldwide communication, and a set of local satellites connected to them by regional intersatellite links [GOL-82].

7.5.3 Links between low earth orbit satellites (LEO–LEO)

Satellites orbiting in LEO present the advantage of significantly minimising transmission delay, which is of high interest for some services (typically voice). However, a single satellite is visible from earth during a very short period of time, thus limiting the duration of communication. This disadvantage can be reduced in a network containing a large number of satellites that are connected by intersatellite links and equipped with switching devices between beams. An example of a network of this type is proposed in [BRA-84; BIN-87]. The Iridium system is another example of a constellation of 66 satellites deployed during 5 May 1997–20 June 2002 [LEO-91]. Iridium Next – the next generation of the Iridium constellation – was launched during 14 January 2014–30 December 2018.

7.5.4 Conclusion

Intersatellite links make the following configurations possible:

— The use of a geostationary satellite as a relay for permanent links between low orbit satellites and a network of a small number of earth stations.
— An increase in system capacity by combining the capacities of several geostationary satellites.
— The planning of systems with a higher degree of flexibility.
— Consideration of systems providing a permanent link and worldwide coverage using low orbit satellites as an alternative to systems using geostationary satellites.

Optical technology is more advantageous in terms of mass and power consumption for high-capacity links.

7.6 SATELLITE BROADCAST NETWORKS

A satellite *broadcast network* consists of a transmitting hub station and a number of receive-only earth stations, and uses the resource of one or several channels (transponders) of a communications satellite. It relies on a star topology and point-to-multipoint connectivity. Links are unidirectional, from the hub towards the earth stations. The hub is generally a rather large earth station while the receive-only earth stations can be very small (typical antenna size of the order of 0.5 m). The cheap cost of such small earth stations (less than 100 Euros) makes them affordable to end users who wish to receive at-home television or audio programmes directly from a satellite; these end users are usually owners of their earth station. This kind of network offers what is called a *direct to home* (DTH) service. The hub is located either at the broadcaster's facilities (and is then operated by the broadcaster itself) or at some other location (when it is most often operated by the satellite operator). As per the Radio Regulations terminology, the uplink is called a *feeder link*, and the hub is often called the *feeder earth station*. The outbound link (feeder link from the hub) is received by all end-user earth stations. In such networks, there is no inbound or return link.

An evolution of this network architecture and its associated services consists of introducing interactivity thanks to a low-data-rate return link transmitted from the earth stations towards the hub. This allows the offering of interactive TV (iTV) or video-on-demand services. Those networks can actually be considered akin to the star networks discussed in Section 7.7.

When focusing on broadcast services only, several options are open to the broadcaster.

7.6.1 Single uplink (one programme) per satellite channel

Figure 7.25 illustrates this option (Option 1). The carrier uplinked from its feeder station by each individual broadcaster occupies a full channel leased from the satellite operator. Different broadcasters share the capacity of a given satellite, each accessing a different channel. The carrier is broadcast at full EIRP by the satellite. This option was popular in the early days of satellite broadcasting when analogue FM transmission was used, requiring maximum EIRP from the satellite and a large bandwidth to allow the use of small receive earth stations at the end user home. However, it is rather expensive as the broadcaster must pay for the full lease of a transponder to transmit only one television programme. Now that compressed digital TV requires a carrier whose bandwidth can be less than a transponder's typical bandwidth (a DVB-S carrier can be as narrow as a few MHz, while a transponder bandwidth can be as large as 72 MHz), two options are considered: either several carriers (programmes) per channel, or multiplexing of several (digital) TV programmes on the same carrier.

7.6.2 Several programmes per satellite channel

With compressed digital television and standards such as DVB-S, the television carrier has a reduced bandwidth and can be transmitted within a fraction of the bandwidth of a satellite channel. Several broadcasters can share the lease of a given transponder, using an FDMA access scheme. This is illustrated in Figure 7.26 (Option 2). However, as usual with FDMA, the satellite transponder must be operated with some back-off, i.e. at reduced EIRP with respect to the maximum EIRP at saturation, which may penalise the quality of the service, unless the end customer uses a larger (and more costly) earth station, or modern satellites with very large EIRP are considered.

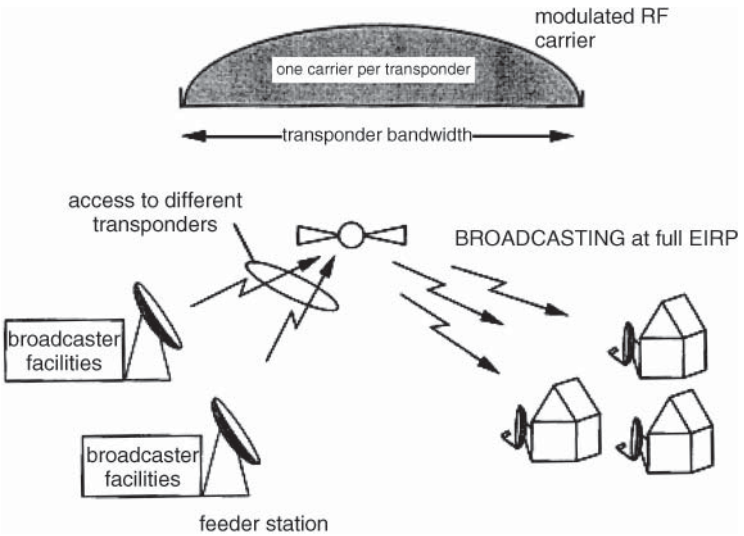


Figure 7.25 Single uplink (programme) to satellite transponder – option 1.

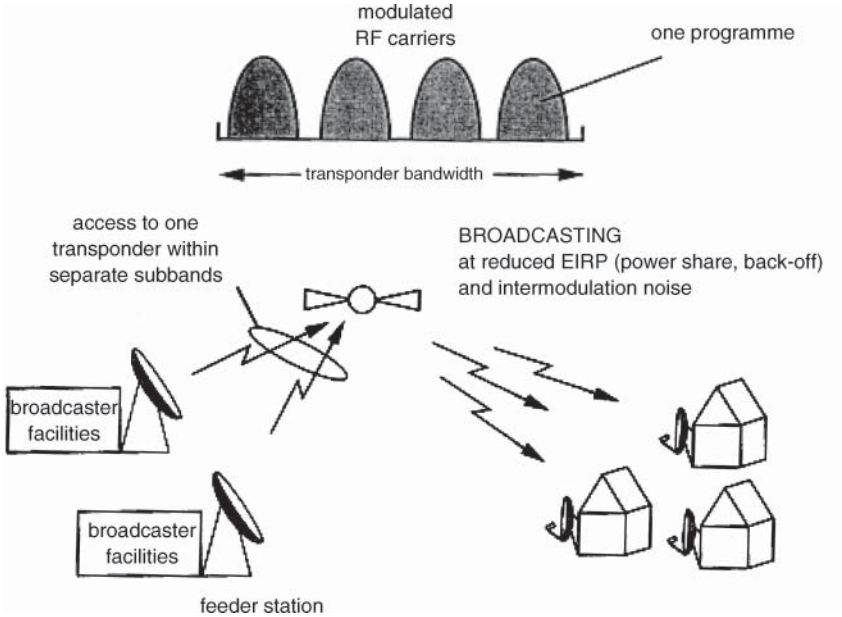


Figure 7.26 Several programmes per satellite transponder (FDMA access to the satellite transponder) – option 2.

7.6.3 Single uplink with time division multiplexing (TDM) of programmes

To benefit from the maximum satellite EIRP, it is preferable to avoid FDMA access by the feeder links and use the full channel bandwidth with a single carrier (as with Option 1) conveying a TDM of programmes from several broadcasters. This is illustrated in Figure 7.27 (Option 3). Notice that multiplexing takes place prior to modulation, and this has to be performed at the feeder station. The feeder station is usually located at the satellite operator premises, and the broadcasters must forward their programmes by terrestrial (or satellite) links to the feeder earth station.

7.6.4 Multiple uplinks with time division multiplexing (TDM) of programmes on downlink

With a regenerative satellite and on-board processing, it is possible to multiplex the programmes on board the satellite. This is illustrated in Figure 7.28 (Option 4). The broadcasters transmit their individual programmes on a single carrier, accessing the satellite transponder in an FDMA mode as with Option 2, keeping full control of the operation of the (small) feeder earth station, which can be located anywhere in the receive coverage of the satellite, possibly in different countries. On board the satellite, the individual carriers are demodulated. Then they are time multiplexed, and the resulting digital TDM is used to modulate a downlink carrier. This carrier, whose bandwidth can occupy a full transponder, is broadcast at full EIRP of the satellite transponder. This offers the advantages of Option 3.

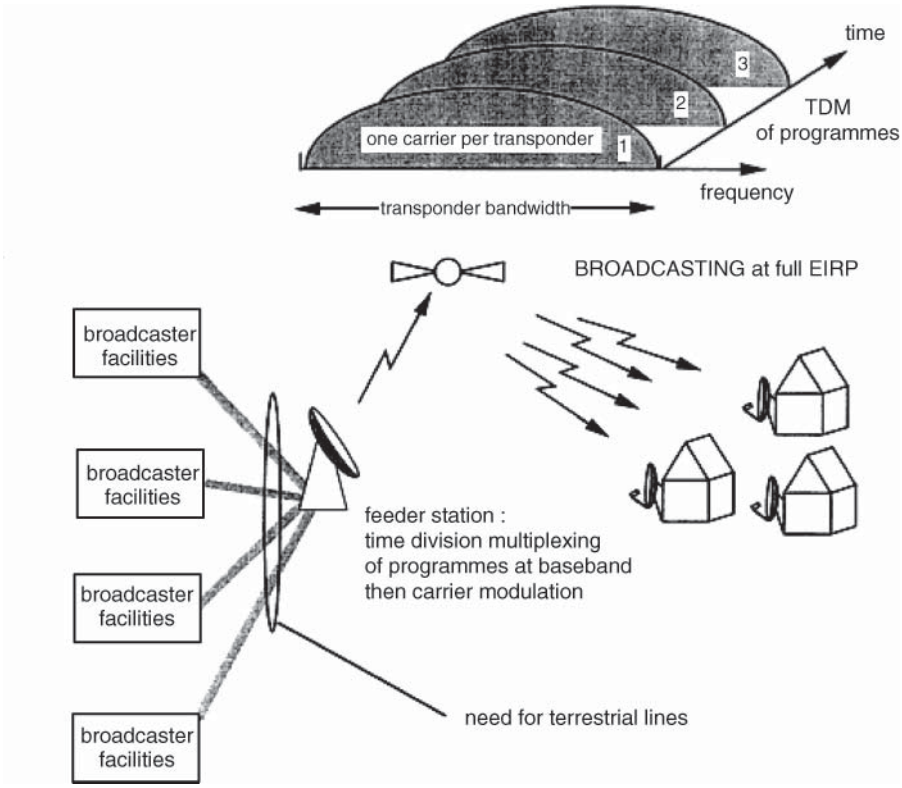


Figure 7.27 Single uplink with time division multiplex (TDM) of the programmes – option 3.

7.7 BROADBAND SATELLITE NETWORKS

A *broadband satellite network* consists of one or several gateways (or hubs) and a number of satellite terminals with receive and transmit capability, and uses the resource of one or several channels (transponders) of a communication satellite. It can rely on a variety of network topologies (star, multi-star, mesh, or hybrid star/mesh) and provide a variety of types of connectivity. Links are bidirectional. The characteristics of the satellite terminals and the gateways or hubs can vary a lot according to the market that is addressed. The consumer market calls for cheap and highly integrated satellite terminals, with the gateways being rather large earth stations. The professional market may address higher-end satellite terminals that have the capability to aggregate traffic generated by a LAN.

Broadband satellite networks are designed to offer most of the services provided by terrestrial Internet networks. Internet service provision by satellite is mainly addressed through the widely accepted *digital video broadcasting* (DVB) standards family, and in particular its satellite versions (DVB-S the original version [ETSI-06], DVB-S2 the second generation [ETSI-14d], and DVB-S2X the extension to DVB-S2 [ETSI-15b]), where the data formatting, initially designed for transporting video and audio streams, has been extended to carry IP datagrams. The satellite-specific DVB-return channel satellite standard (DVB-RCS is the original standard, and DVB-RCS2 is the new generation of standard that has been published) provides the specification for the return

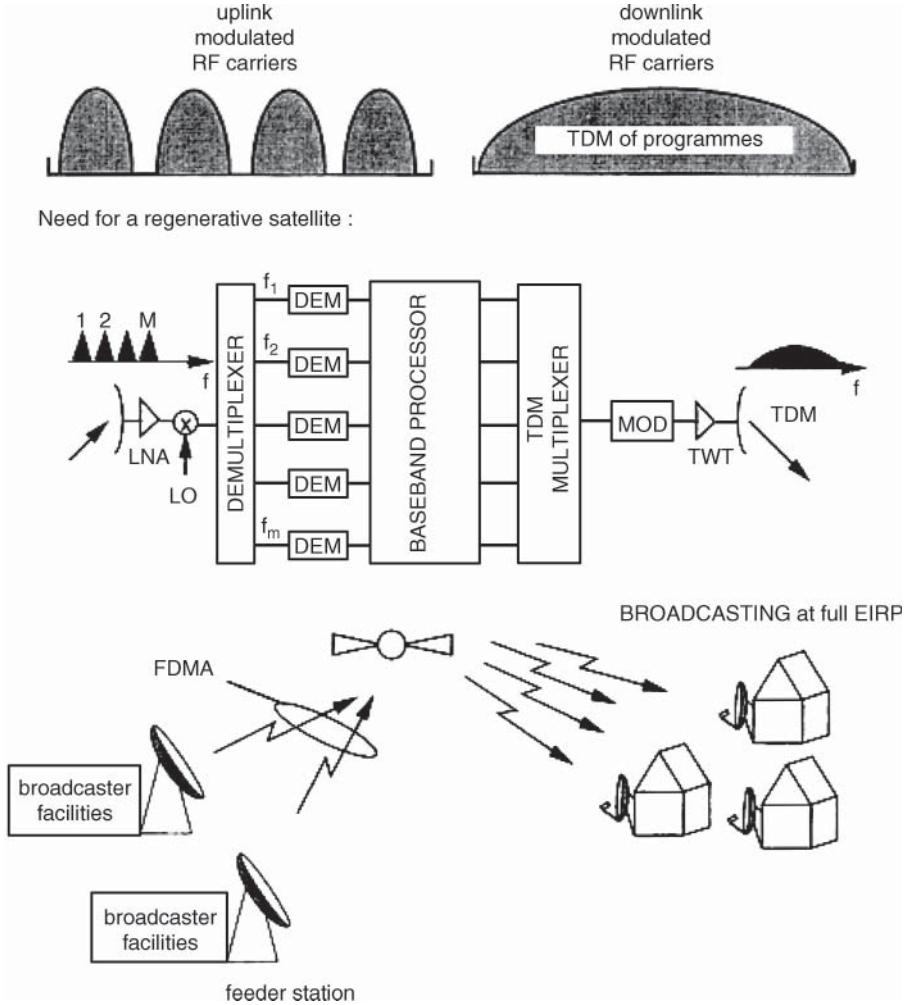


Figure 7.28 Regenerative satellite with on-board processing allows multiple FDMA uplinks and time division multiplex (TDM) of programmes on the downlink – option 4.

traffic flows from DVB-RCST to gateways [ETSI-14e; ETSI-14f; ETSI-14g; ETSI-14h; ETSI-14i]. Several standards from the European Telecommunication Standardisation Institute (ETSI) and the Internet Expert Task Force (IETF) deal with the details of implementation of IP networking protocols and network architecture. These standards and technical specifications have speeded up satellite systems for both broadcasting and networking services. The standards of interest are listed in the References section at the end of the chapter.

7.7.1 Overview of DVB-RCS/RCS2 and DVB-S/S2/S2X networks

This section provides an example of a network, following the generic description introduced in Section 7.2. For convenience of clear explanations, we use DVB-S and DVB-RCS to represent

all generations of the standards in our discussions here unless some special features need to be explained and clarified.

7.7.1.1 Network characteristics

A DVB-S and DVB-RCS network displays the following characteristics:

— The uplink of the RCST uses MF-TDMA according to the DVB-RCS standard, moving picture experts group (MPEG) profile.
— The downlink from satellite to RCST is fully compatible with the DVB-S standards.
— The satellite system supports symmetric predictive traffic, as well as bursty traffic generated by a large number of users, owing to dynamic allocation.
— The satellite system supports interworking with terrestrial networks such as PSTN and ISDN as well as private IP networks belonging to service providers (SPs).
— The satellite system supports integrated IP-based data services and native MPEG video broadcasting.
— A satellite star connectivity network supports single-hop connectivity between satellite network users and terrestrial network users through the satellite gateway (GW). A satellite mesh connectivity network supports single-hop connectivity between satellite network users; it requires an OBP that allows routing of MPEG packets from uplink to downlink beams in a flexible way, possibly with data replication on board to support multicast services.

7.7.1.2 Management station (MS)

The management station (MS) consists of the NMC and the NCC. The NMC manages all network elements and the network and service provisioning. The NCC manages the control of the interactive network, e.g. it serves satellite access requests from the users of the system.

7.7.1.3 Satellite gateway (GW)

The GW provides interworking functions between the satellite and terrestrial networks, such as the telephony networks and broadband multimedia Internet/intranet-oriented ground networks. The GW could be part of the hub's functionality in the star topology of a transparent satellite access network. The GW can provide service guarantees to subscribers based on different QoS criteria and different subscription levels of services.

The GW is composed of different and configurable elements depending on the needs of connectivity. For example, a GW may consist of interactive receive decoders (IRDs), an IP router, a multi-conference unit (MCU), and a voice and video gateway or gatekeeper.

7.7.1.4 Return channel satellite terminal (RCST)

The user earth station (in the terminology of Chapter 1) is called a RCST in this chapter to be compliant with the terminology of the DVB-RCS standard. It consists of two main units: the indoor unit (IDU) and the outdoor unit (ODU).

Uplink transmission functions consist of BB to RF up-conversion (the BB signal modulates the carrier at intermediate frequency [IF], and then the carrier is up-converted to RF band), and RF amplification so that the signal is amplified before transmission via the transmit antenna.

The achieved EIRP must meet the link budget requirements. On downlink reception, the RCST functions consist of RF low-noise amplification; down-conversion to IF; IF DVB-S demodulation; and decoding at BB.

The ODU consists of the RF transmitter, the RF receivers, and the antenna. The IDU contains the DVB-S and DVB-RCS modem and the interface to the local network. The RCST may be connected to an Internet subnet in a LAN.

The RCST allows users to communicate with each other either in a single satellite hop (mesh connectivity) or with a double satellite hop through the GW (star connectivity); it also allows users to communicate with terrestrial network users through the GW in a single satellite hop, e.g. telecommunication or IP-based Internet services.

7.7.1.5 On-board processor (OBP)

If a payload with on-board processing is considered, the OBP is the core of a satellite mesh system (see Figure 7.4). It combines both DVB-RCS and DVB-S satellite transmission standards to allow full cross-connectivity between the uplink and downlink beams to allow UTs to send and receive signals from the other terminals via satellite.

The uplink DVB-RCS carriers are down-converted from RF to a low IF; the baseband processor (BBP) de-multiplexes, demodulates, and decodes carriers at IF in order to generate a single multiplex of MPEG-2 packets compliant to the DVB-S/S2 standard using routing information coming from uplink packets; after channel encoding (FEC) with selectable code rate, the modulation with QPSK (DVB-S) or higher modulations (DVB-S2/S2X) is performed at IF to generate a carrier that is then frequency up-converted into the downlink frequency.

7.7.1.6 Network interfaces

The following interfaces are used for the components of the satellite network to communicate with each other (refer to Figure 7.44, later in the chapter):

— *T interface:* A user interface (UI) between RCST IDU and UTs (hosts) or LANs.
— *N interface:* The interface between NCC and RCST for control and signalling to support the user plane (U-plane) services (synchronisation, DVB tables, and connection control signalling).
— *M interface:* The interface between NMC and RCST for management purposes (simple network management protocol [SNMP] and management information base [MIB] interactions).
— *U interface:* The interface between the satellite payload and the RCST physical interface (the air interface).
— *P interface:* The logical interface between two RCSTs for transaction of peer layer signalling traffic and user data traffic.
— *Interface:* The interface between NCC and OBP interface for OBP control and management (if applicable).

7.7.2 Protocol stack architecture for broadband satellite networks

Satellite broadband networking is based on the assumption that all IETF IPs should be supported in the gateway stations. Indeed, satellite networks are considered one radio technology in the same way as wireless terrestrial technologies. Therefore, the focus of satellite networking has been put on the lower layers of the protocol stack. The satellite protocol stack architecture

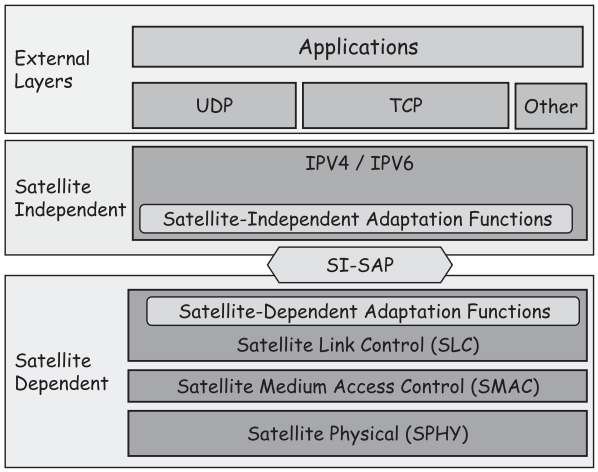


Figure 7.29 Satellite-dependent and satellite-independent protocol layers with BSM protocol architecture reference model. Source: reproduced with the kind permission of the ETSI.

is described by using the concepts of the layering principle and the general protocol stack architecture for IP interworking: the two major layers are the satellite Physical layer and the Link layer.

The Link layer consists of *Medium Access Control* (MAC) and *Logic Link Control* sublayers. The reason to split the Link layer into sublayers is for adaptation of the *satellite-dependent functions* and *satellite-independent functions*.

The protocol architecture reference model for broadband satellite multimedia (BSM) is illustrated in Figure 7.29 which shows the satellite-dependent and satellite-independent adaptation functions [ETSI-07].

The RCST may have peer-to-peer communications with other RCSTs for mesh communications or with other GWs for star communications. In terms of the OSI/ISO seven-layer reference model, the satellite-independent service access protocol (SI-SAP) [ETSI-15c; ETSI-15d] is positioned between the Link and Network layers. The satellite network protocols consist of the *satellite link control* (SLC), *satellite medium access control* (SMAC), and *physical* (PHY) layers.

The SLC sublayer exchanges IP datagrams with the Network layer; the SMAC sublayer has transport functions for transmission bursts of MPEG packets and for reception of MPEG packets contained in the TDM as well as for transportation of generic stream encapsulation (GSE) packets; the RCST Physical layer is responsible for transmitting the data over the physical medium with synchronisation and bit-error correction functions.

7.7.3 Physical layer and MAC layer

The Physical layer is the lowest layer of the protocol stack. It receives the data frames from the satellite MAC layer and transmits the data organised in packets over the physical medium.

7.7.3.1 Transport stream MPEG packets

MPEG packet transport streams constitute the format used to transport data on the forward link using DVB-S standards in the satellite networks. There are different packet formats can be

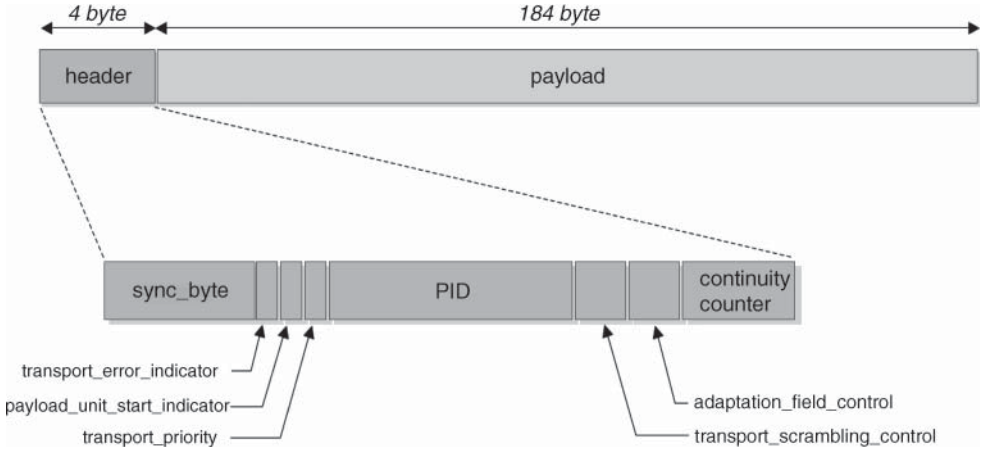


Figure 7.30 Transport stream MPEG packets.

selected DVB-S, GSE, and asynchronous transfer mode (ATM). The concept of transport streams originates in the MPEG standard where a predefined data structure made of packets of given length has been defined to carry variable-length video and audio elementary stream packets (ESPs) [ISO/IEC-96]. As shown in Figure 7.30, each MPEG-2 transport packet includes a 1472-bit (184 bytes) payload data field and a 32-bit (4 bytes) header.

The header starts with a known synchronisation byte. A set of flag bits is used to indicate how to process the payload. A 13-bit packet identifier (PID) is used to uniquely identify each received stream. The PID allows the receiver to differentiate each stream from the others. The MPEG packet format and the structure of the header are explained as the following:

— 8 bits: sync byte (sync the decoder –47hex-start of TP)
— 1 bit: transport error indicator
— 1 bit: payload unit start indicator (PUSI)
— 1 bit: transport priority
— 13 bits: packet identifier (PID)
— 2 bits: transport scrambling control
— 2 bits: adaptation field control
— 4 bits: continuity counter

7.7.3.2 Return link Physical layer

The return link Physical layer has the following three main functions in order to provide mechanisms and parameters to allow the SMAC to correctly transmit the data flow in the physical means:

— The transmission function performs signal BB processing and radio transmission between the RCSTs and the satellite. This includes randomisation for energy dispersal and adequate binary transitions using a pseudorandom binary sequence (PRBS), channel coding (FEC) with Reed–Solomon/convolutional or turbo-coding (TC), and QPSK modulation based on a root raised cosine filtering with a 0.35 roll-off.

— The synchronisation function meets the very tight requirements of the MF-TDMA access in time and frequency of RCSTs. The network clock reference (NCR) gives a 27-MHz reference for frequency synchronisation. All signals are synchronous.
— The power-control function compensates for variations in the radio channel and minimises interference between signals received at satellite level, in order to optimise the system capacity and availability.

The RCST reports physical parameter values to the NCC to allow traffic and signalling flow monitoring (synchronisation and power control functions). In reply, the SLC layer provides real-time configuration parameters, logon parameters, and traffic packets to be transmitted on the air interface.

7.7.3.3 Return link multiple-access technique

The uplink satellite access scheme is MF-TDMA, as illustrated in Figure 7.31. When the MPEG transport stream format is selected, the MF-TDMA uplink is based on the transport of bursts of up to 24 MPEG packets and customised packets for the logon and synchronisation processes.

The signal is structured into segmentations of superframes, frames, and time slots for the purpose of allocating physical resources. These segmentations are transmitted over a pool of carriers, which constitutes the MF-TDMA channel. The NCC allocates a series of bursts to each active RCST. Each burst is defined by a frequency, a bandwidth, a start time, and a duration following the MF-TDMA schema.

7.7.3.3.1 Time slots and bursts

Time slots are the smallest capacity that can be allocated to a given terminal. Each time slot comprises a single burst plus guard time at each edge of the burst. The guard time is required to cope with system timing errors and RCST power switch on–off transience. Three types of burst are considered: traffic (TRF) bursts, common signal channel (CSC) bursts, and synchronisation (SYNC) bursts. These bursts contain a fixed-length preamble for timing recovery and for phase-ambiguity suppression, set to 255 symbols.

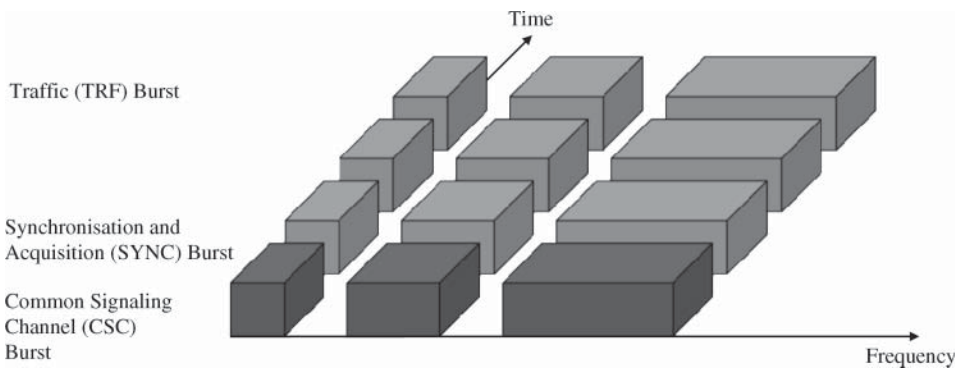


Figure 7.31 Multi-frequency time division multiple access (MF-TDMA).

— *Traffic (TRF) time-slot format:* TRF bursts are used for carrying data traffic as well as signalling and other control messages. A TRF burst is made up of a preamble and an encoded data packet. The last 48 symbols of the preamble constitute the unique word (UW) used for burst detection. The encoded packet can be GSE, MPEG, or ATM cells. The number of MPEG-2 packets (up to 24) in one burst depends on the code and number of bursts selected for the uplink frame.
— *Synchronisation (SYNC) time-slot format:* SYNC bursts are required to accurately position RCST burst transmission during fine synchronisation and synchronisation maintenance procedures, as well as to send capacity requests. SYNC bursts are composed of a 16-byte satellite access control (SAC) field, of which 64 bits are used for requests of capacity, 16 bits for the MAC field, 24 bits for a group/logon ID, 16 CRC bits for error detection, and 8 stuffing bits to make up the SAC into 16 bytes (inserted before CSC). The SAC is randomised and coded and a preamble is added for burst detection.
— *Common signal channel (CSC) time-slot format:* CSC bursts are used by an RCST to identify itself during the logon process. It has a total length of 16 bytes, of which 24 bits are used for a field to describe the RCST capabilities; 48 bits for the RCST MAC address; 40 bits for a reserved field; and 16 bits for a cyclic redundancy check (CRC) field. These fields are randomised and coded and a preamble is added for burst detection.

7.7.3.3.2 Frames

A frame consists of time slots, defined as a portion of time and frequency on the uplink from RCST to the satellite. The uplink frame duration is fixed regardless of the carrier data rate and turbo-coding rate. As a default value, the frame duration is 69 632 ms, which corresponds to 1 880 064 programme clock reference (PCR) count intervals. A frame spans a pool of carriers. Depending on the carrier rate, the frame could be structured in sub-frames of 2^i for C_i carrier rate, where $i = 1, \ldots, 5$.

Consider the frame for a C1 carrier rate. A frame is configured in terms of the number of MPEG-2 packets per burst and the number of bursts per frame. Other possible carrier types are composed of sub-frames that have the same structure as the C1 carrier rate. Table 7.3 describes the frame composition based on the turbo-coding (TC) rate and the number of TRF bursts contained in the frame. Each TRF burst contains an integer number of MPEG packets and is allocated to a single RCST.

A configuration of one TRF burst per frame or per sub-frame is oriented for video services. Table 7.4 gives the main parameters of different possible carriers for the case of 18 TRF packets per frame.

Table 7.3 Frame composition.

TRF Bursts per frame (for C1) or per subframe (for other data rates)	TC = 4/5		TC = 3/4	
	MPEG packets per burst	MPEG packets per frame	MPEG packets per burst	MPEG packets per frame
Configuration 1: 6 TRF	4	24	3	18
Configuration 2: 18 TRF	1	18	1	18
Configuration 3: 1 TRF	24	24	24	24

Note: The possible turbo codes are detailed in the DVB-RCS standard (see [ETSI-09b], clause 8.5.5.4), but only two (4/5 and 3/4) have been used here for this example of frame composition.
Source: reproduced with the kind permission of the ETSI.

Table 7.4 Carrier main parameters (18 TRF packets per frame).

Carrier type	C1	C2	C3	C4
Max information bit rate per carrier (kbps)	388.79	777.57	1555.15	3110.29
Transmission QPSK symbol rate (ksps)	350.99	701.98	1403.95	2807.90
Number of carriers per MF-TDMA channel	64	32	16	8
MF-TDMA channel information bit rate (Mbps)	24.88	24.88	24.88	24.88

Source: reproduced with the kind permission of the ETSI.

7.7.3.3.3 Superframes

A superframe is composed of a number of consecutive frames, defined as a portion of time and frequency. The default number of frames in a superframe is 2, but it may have from 1 to 31 frames.

A superframe identifies a set of resources that can be allocated to a group of RCSTs. A superframe identifier identifies the uplink resources accessed by a given set of RCSTs, so that different sets of RCSTs can be managed separately. In a typical implementation, each superframe defines a set of carriers. Figure 7.32 describes the superframe organisation in time and frequency.

For each superframe, the allocation of time slots is communicated to the RCST via the time burst table plan (TBTP) as described in DVB-RCS standard. An RCST is allowed to transmit bursts only in time slots that were allocated to it (dedicated access) or on random-access time slots (contention access). Some time slots (such as SYNC bursts) may be assigned to the RCSTs on the basis of a period much longer than one superframe. The period for these time slots is system dependent, but is typically in the order of 1 second.

7.7.3.3.4 Carrier type and frame composition

An MF-TDMA channel is defined as a certain bandwidth divided into a number of sub-bands. Each sub-band may consist of a particular carrier type or a combination of several carrier types.

Three types of carrier are considered:

— *Logon carrier:* A CSC burst and a number of TRF bursts or CSC bursts and SYNC bursts
— *Synchronisation carrier:* SYNC bursts and TRF bursts or SYNC bursts alone
— *Traffic carrier:* TRF bursts

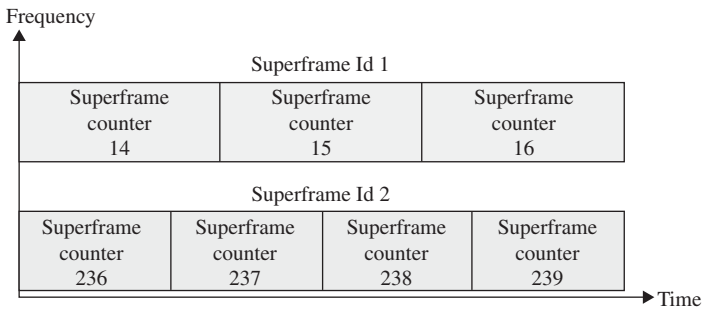


Figure 7.32 Superframe organisation. Source: reproduced with the kind permission of the ETSI.

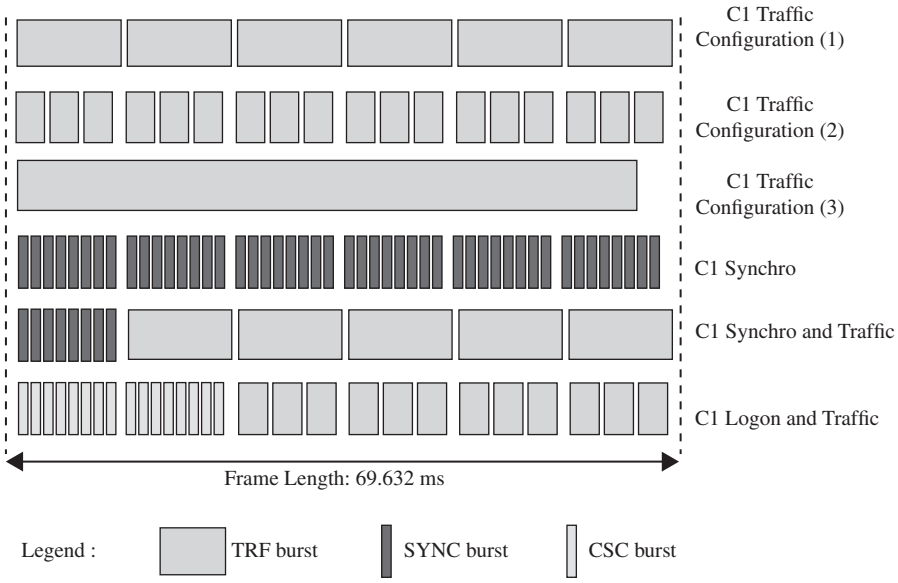


Figure 7.33 Examples of C1 carrier configurations. Source: reproduced with the kind permission of the ETSI.

Logon and synchronisation carriers are C1 carriers. Different traffic carriers may be defined. The number of carriers per sub-band depends on the bandwidth occupied by each class of carrier. Figure 7.33 shows examples of C1 carrier configurations.

7.7.3.3.5 Uplink MF-TDMA channel frequency plan

The frame duration is partitioned into time slots according to carrier type (CSC, SYNC, or TRF), carrier rate (C1, C2, C3, and C4), and carrier configuration (number of MPEG-2 packets per burst).

A typical configuration of the MF-TDMA uses a bandwidth of 36 MHz that is divided into four sub-bands of 9 MHz; the sub-bands are configured with 16 C1 carriers, 8 C2 carriers, 4 C3 carriers, or 2 C4 carriers. Each of the sub-bands may be configured independently. At least one sub-band is configured with C1 carriers for logon and synchronisation needs.

7.7.3.4 Forward link Physical layer

The forward link format conforms to the DVB-S and DVB-S2 standards. A list of the latest versions of the standards is provided in the References section at the end of the chapter for further detailed reading.

With DVB-S, the waveform is constituted, the randomised TDM MPEG-2 packets are channel encoded (RS and convolutional coding with interleaving, see Chapter 4), and, after pulse shaping with roll-off of 0.35, the carrier is modulated with QPSK. Any of the possible convolutional rates (1/2, 2/3, 3/4, 5/6, and 7/8) defined in the DVB-S standard could be used.

The DVB-S2 standard offers a higher capacity on the forward link thanks to the use of more efficient (higher coding gain, see Chapter 4) channel coding with a Bose–Chaudhuri–Hocquenghem

Table 7.5 DVB-S2 waveform configurations (N: normative, O: optional).

System configurations		Broadcast services	Interactive services
QPSK	1/4, 1/3, 2/5	O	N
	1/2, 3/5, 2/3, 3/4, 4/5, 5/6, 8/9, 9/10	N	N
8PSK	3/5, 2/3, 3/4, 5/6, 8/9, 9/10	N	N
16APSK	2/3, 3/4, 4/5, 5/6, 8/9, 9/10	O	N
32APSK	3/4, 4/5, 5/6, 8/9, 9/10	O	N

Source: reproduced with the kind permission of the ETSI.

(BCH) code concatenated with low-density parity check (LDPC) codes, and higher spectral efficiency modulation formats [MOR-04]. Four types of modulation are available: QPSK (2 bit $(s\,Hz^{-1})^{-1}$), 8PSK (3 bit $(s\,Hz^{-1})^{-1}$), 16APSK (4 bit $(s\,Hz^{-1})^{-1}$), and 32APSK (5 bit $(s\,Hz^{-1})^{-1}$).

Multiple combinations of modulation and code rates can be selected depending on the link budget conditions. This could be done in a static manner to compensate for link budget variations depending on the satellite terminal position in the satellite coverage. Table 7.5 shows the system configurations.

This allows also the implementation of an adaptive coding and modulation (ACM) strategy to mitigate RF link impairments. In the ACM method, a transmitter changes coding and modulation methods adaptively and transmits modulation coding (MODCOD) information with data according to the data reception performance of a receiver; the receiver changes the decoding/demodulation methods of the received signal according to the MODCOD information.

With DVB-S2, the input data packets are organised in user packets with an added 8-bit CRC codeword (stream adaptation). Multiple input streams can be sliced and multiplexed (merger/slicer) in packets with a 10-byte header (BBHeader data field). The packets are randomised, and the BB frame is encoded. In the FEC encoder, outer encoding of BCH codes and internal encoding of LDPC codes are performed, so that each parity bit is added to the BB frame to constitute the FEC frame. The FEC frame is made of blocks of 64 800 bits for normal frames or 16 200 bits for short frames depending on the block length of the LDPC code. The Physical layer (PL) framing unit divides the FEC frame block into slots of 90 symbols for actual transmission through modulation. Information on the start points of each frame (SOF), signalling information about MODCOD informing a transmission method, and a pilot signal are inserted to constitute the PL frame. Figure 7.34 illustrates the application of mode adaptation to transport stream MPEG packets.

The FEC frame formed by the BCH and LDPC encoding has a fixed length (64 800 bits or 16 200 bits) regardless of the coding rate and modulation method, and therefore the length of the PL frame is variable depending on the modulation method and the coding rate. The BB frame length is aligned to the block length of the BCH code (Figure 7.35). The data field is composed data field length (DFL) bits, where $K_{BCH}-80 \geq DFL \geq 0$. $K_{BCH}-80$ corresponds to the difference in bits between the BCH uncoded block length, which depends on the FECFRAME length (normal or short), the coding rate, and the length (10 bytes) of the BBHEADER length.

For broadcasting applications, the data field is filled to the maximum capacity ($K_{BCH}-80$ bits); for unicast applications, the data field may include an integer number of user packets. This allows for correct recovery of the user information when ACM is utilised. As a consequence, padding is required to complete the constant length (K_{BCH}) BBFRAME. This also happens whenever available data is not sufficient to fill the BBFRAME. The stream-adaptation subsystem is responsible for providing padding when $DFL < K_{BCH}-80$, and scrambling the information at the

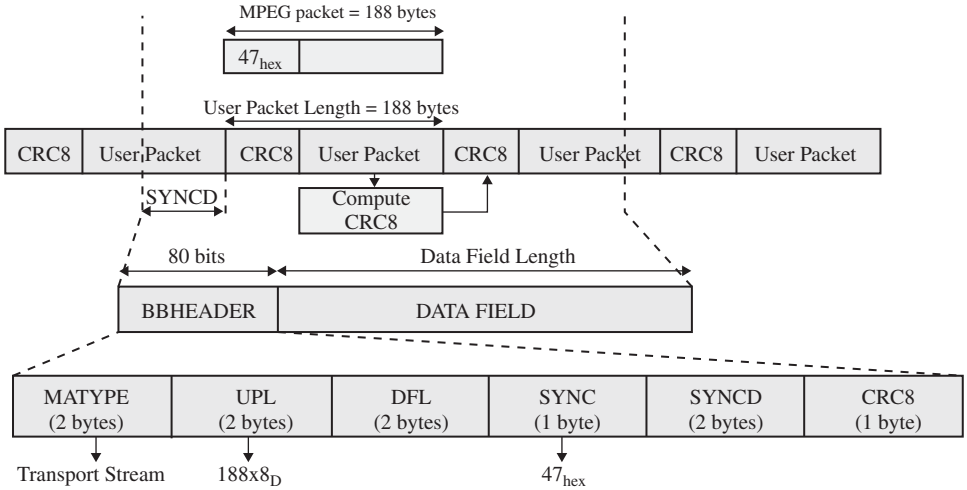


Figure 7.34 DVB-S2 mode adaptation to transport stream MPEG packets. Source: reproduced with the kind permission of the ETSI.

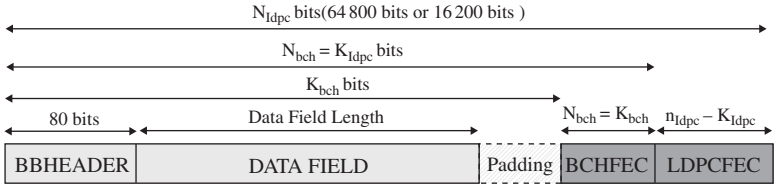


Figure 7.35 The FEC frame. Source: reproduced with the kind permission of the ETSI.

encoder input. The BB header (Figure 7.34) is composed of the MATYPE field (input stream characteristics, roll-off), the UPL field (user packet length), DFL, SYNC (copy of user packet sync byte), SYNCD (distance in bits from beginning of data field to first user packet), and an 8-bit CRC.

7.7.4 Satellite MAC layer

The SMAC layer is above the Physical layer and below the IP Network layer. It provides transmission services to the IP Network layer. The SMAC is responsible for transmission and reception of packets to and from the Physical layer.

In the user plane, the SMAC layer interfaces the PHY layer in order to send traffic bursts and also to receive all the MPEG-2 packets contained in the TDM, filtering the packets according to their packet identifier before passing them to the upper layer.

The control plane functions of the SMAC layer include the logon and synchronisation of the RCST. The SMAC layer sends logon content (specific CSC bursts, 48-bit MAC address, and 24-bit terminal capacity) in an S-ALOHA mode and capacity requests (specific SYNC bursts) to the Physical layer.

7.7.4.1 Transport mechanism protocol stack

The RCST transport mechanism protocol stack for traffic messages is based on the DVB-RCS standard for the uplink (ATM or MPEG packets) and on the MPEG-2 standard in the downlink (MPEG-2 transport stream). The asynchronous transfer mode adaptation layer (AAL) is used to send IP datagrams over ATM cells. Multi-protocol encapsulation (MPE), unidirectional lightweight encapsulation (ULE), or GSE packing is used to send IP datagrams over MPEG-2 packets. Figure 7.36 shows the protocol stack for native IP terminals.

Return signalling messages are composed of control (CTRL) and management (MNGM) messages mainly dedicated to the connection control protocol message, which are data unit labelling method (DULM) encapsulated, and specific logon (CSC) and synchronisation (SYNC) bursts.

Capacity requests are transmitted via the SAC field associated with each SYNC burst. Different types of capacity request are available:

— *Continuous rate assignment (CRA):* Fixed capacity is allocated to RCST (period = 1 superframe).
— *Rate-based dynamic capacity (RBDC)*: RCST asks for the data rate; the request is processed by NCC and the request lifetime is set to two superframes (default value).
— *Volume-based dynamic capacity (VBDC)*: RCST asks for the volume; capacity requests are cumulative.
— *Absolute volume-based dynamic capacity (AVBDC)*: As for VBDC, but each request replaces the previous one.
— *Free capacity assignment (FCA)*: Capacity is allocated to RCST without signalling

7.7.4.2 MPEG-2, DVB-S, and DVB-RCS tables

The transmission of signalling information is based on the mechanisms and procedures described by the DVB standards and on the use of MPEG-2, DVB-S, and DVB-RCS tables and messages. Forward signalling tables are encapsulated in programme-specific information (PSI) or service information (SI) based on the DVB-S standard. RCS terminal information messages (TIM) use digital storage medium – command and control (DSM-CC) encapsulation as defined in DVB-RCS standard.

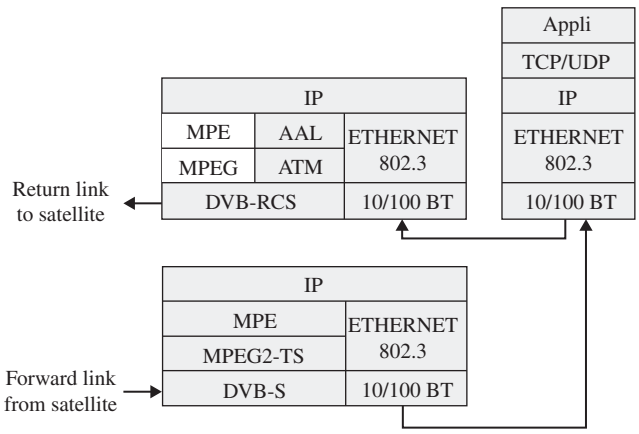


Figure 7.36 Protocol stack for native IP RCS terminals. Source: reproduced with the kind permission of the ETSI.

[ISO/IEC-96] specifies the service information (SI), which is referred to as PSI. The PSI data provides information to enable automatic configuration of a receiver for demultiplexing and decoding various streams of programme within the multiplex. The PSI data is structured into four types of table [ETSI-16; ETSI-09a]:

— *Program association table (PAT)*: For each service in the multiplex, the PAT indicates the location and the PID values of the transport stream of the corresponding program map table (PMT).
— *Conditional access table (CAT)*: This table provides information on the conditional access (CA) used in the multiplex.
— *Program map table (PMT)*: This table identifies and indicates the location of the streams that make up each service and the location of the program clock reference fields for a service.
— *Network information table (NIT):* This table describes the transport streams participating in a DVB network (identified by its network identifier) and carrier information to find the RCS service identifier and its related transport stream (TS) identifier and satellite link parameters.

In addition to the PSI, data is needed to provide identification of services and events to users. This data is structured in tables specified in the DVB standard with details of the syntax and semantics, such as the following:

— The bouquet association table (BAT) provides information regarding bouquets. As well as giving the name of the bouquet, it provides a list of services for each bouquet.
— The service description table (SDT) contains data describing the services in the system, e.g. names of services, the service provider, etc.
— The event information table (EIT) contains data concerning events or programmes such as the event name, start time, duration, etc.
— The time and date table (TDT) gives information relating to the present time and date. This information is given in a separate table because it is frequently updated.

The format and semantics of DVB-RCS table descriptions follow the DVB-RCS standard specification [ETSI-09b]:

— *RCS map table (RMT)*: This table describes transport stream tuning parameters to access forward link signalling (FLS) services. The RMT may contain one or more linkage descriptors each pointing to one FLS service. Each FLS service carries a set of the following signalling tables: SCT, FCT, TCT, SPT, CMT, TBTP, and TIMs for a defined RCST population.
— *Superframe composition table (SCT):* This signalling table describes the subdivision of the network into superframes and frames. The table contains, for each superframe, a superframe identification, a centre frequency, an absolute start time expressed as a NCR value, and a superframe count.
— *Frame composition table (FCT):* This signalling table describes the partitioning of the frames into time slots.
— *Time-slot composition table (TCT):* This signalling table defines the transmission parameters for each time-slot type identified by the time-slot identifier. It provides information about the time-slot properties such as symbol rate, code rate, preamble, payload content (TRF – traffic, CSC – common signal channel, ACQ – acquisition, SYNC – synchronisation), and others.
— *Satellite position table (SPT):* This signalling table contains the satellite ephemeris data required to update the burst position at regular intervals.
— *Correction message table (CMT):* This signalling table is sent by the NCC to groups of RCSTs. It advises the logged-on RCSTs what corrections (to burst frequency, timing, and amplitude) is made to their transmitted bursts.

— *Terminal burst table plan (TBTP):* This message is sent by the NCC to a group of terminals. It contains the assignment of a contiguous block of time slots. Each traffic assignment is described by the number of the start time slot in the block and a repetition factor giving the number of consecutive time-slot allocations.
— *Multicast map table (MMT):* This table provides the RCST with the PID to decode to receive a certain IP multicast session.
— *Terminal information message (TIM):* This message is sent by the NCC either to an individual RCST addressed by its MAC address (unicast message) or as a broadcast to all RCSTs using a reserved broadcast MAC address and contains static or quasi-static information about the forward link such as configuration.

In the satellite system, all the tables and messages are formatted and distributed by the NCC. However, depending on the source of information (NCC, NMC, or service provider), the types of tables (PSI, SI, or DVB-RCS), and the way of updating content, different classifications are specified in the standard.

7.7.4.3 IP datagram encapsulation

Different segmentation and reassembly solutions have been developed for packing IP datagrams over MPEG-2 transport stream packets or other layer 1 containers such as the BB frames with DVB-S2 [IETF-05]. The basic mechanism is MPE. A reduced overhead is obtained with ULE or the newly developed GSE solutions.

7.7.4.3.1 Multi-protocol encapsulation (MPE)

MPE provides a mechanism for transporting data network protocols on top of the MPEG-2 transport streams (MPEG-TS) in DVB networks [ETSI-15a]. It has been optimised for carriage of the IP, but it can be used for transportation of any other network protocol by using LLC/SNAP encapsulation (Figure 7.37). It covers unicast, multicast and broadcast. The encapsulation allows secure transmission of data by supporting encryption of the packets.

An MPE section is packetised video, audio, or data. The section can address up to 64 kb, but can be any length if the beginning and end of the section can be identifiable. A section length is not an exact multiple of MPEG packet payloads, and it is possible that the last MPEG packet of an MPE section may be almost empty. In order to improve satellite utilisation bandwidth, it is possible to start a new section in the middle of an MPEG packet, right after the end of the previous section. This may be done thanks to the PUSI flag of the MPEG2-TS; IETF-14b header and a one-byte pointer occupying the first byte of the MPEG2-TS payload.

7.7.4.3.2 Unidirectional lightweight encapsulation (ULE)

ULE is a mechanism for transporting data network protocols on top of the MPEG-2 transport streams (MPEG-TS) with reduced overhead [IETF-14b]. Protocol data units (PDUs), such as Ethernet frames, IP datagrams, and other network-layer packets, used for transmission over an MPEG-2 transport multiplex, are passed to an encapsulator. This formats each PDU into a subnetwork data unit (SNDU) by adding an encapsulation header and an integrity-check trailer. The SNDU is fragmented into a series of one or more MPEG-TS packets that are sent over a single TS logical channel.

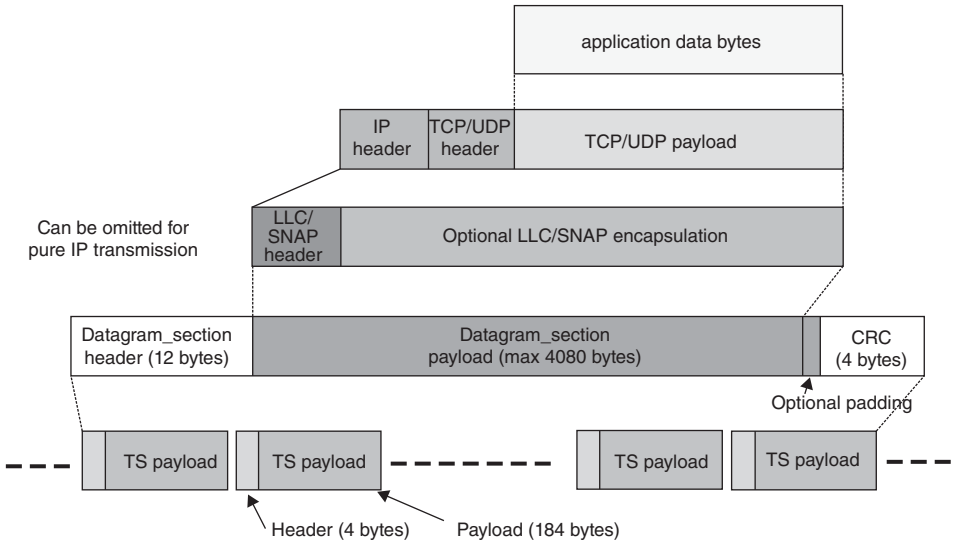


Figure 7.37 Multi-protocol encapsulation (MPE) section within MPEG2-TS. Source: reproduced with the kind permission of the ETSI.

Figure 7.38 The ULE data structure. Source: reproduced with the kind permission of the ETSI.

The ULE data structure is illustrated in Figure 7.38. The encapsulation header is composed of:

— A 1-bit destination address absent field (D)
— A 15-bit length field indicating the length of section from the next payload type field to the CRC trailer
— A 16-bit payload type field (0x0800: IPv4 payload; 0x86DD: IPv6 payload)
— A 6-byte SNDU field providing the MAC destination address

The trailer consists of a 32-bit CRC. The D and length field combine in the 0xFFFF value to indicate that all data has been transmitted (END indicator).

7.7.4.3.3 Generic stream encapsulation (GSE)

MPE is the DVB standard for the encapsulation of data and other content on MPEGTS packets. The DVB-S2 standards have features of backward compatibility modes for carrying MPEG-TS, as well as generic modes for carrying arbitrary packets of variable length. These are referred to as *generic streams*. They are intended to transport a sequence of data bits or data packets, possibly organised in frames, but with no specific timing or rate constraints. GSE was introduced to improve the efficiency to carry IP data (and other network and link-layer packets) over BB frames of DVB-S2 with reduced overhead (3%) [ETSI-11]. The protocol data units (PDUs) delivered by the Network layer are encapsulated with a GSE header to constitute the GSE packet of variable size. The GSE packets are concatenated in the BB frame data field. If the BB frame data field is

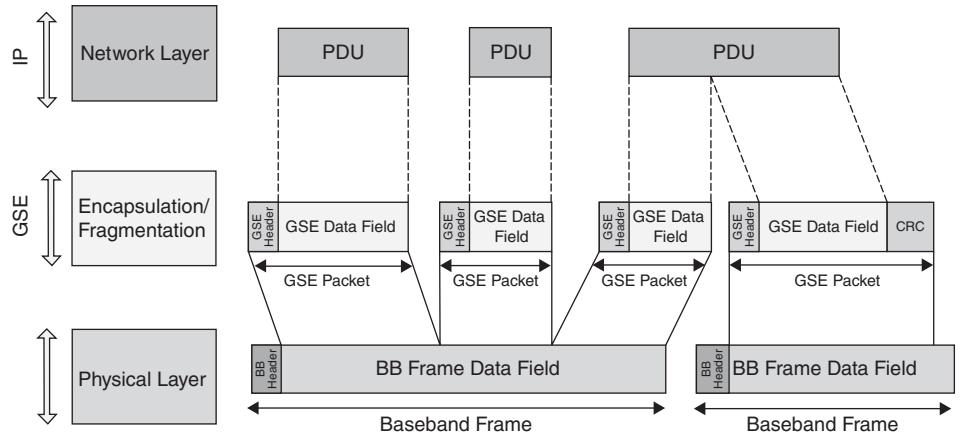


Figure 7.39 Generic stream encapsulation. Source: reproduced with the kind permission of the ETSI.

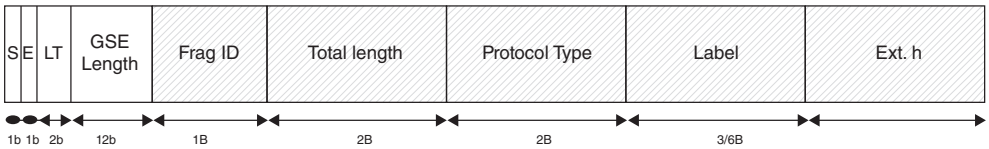


Figure 7.40 GSE header.

not fully occupied, the next PDU is segmented into two GSE packets, one fitting the remaining space of the BB frame data field, the other being transmitted in the next BB frame. Figure 7.39 illustrates the encapsulation and segmentation concept [ETSI-14a; ETSI-14b; ETSI-14c].

The GSE header structure (Figure 7.40) has some similarities with the ULE one. The destination address is set in the label field if required (packet filtering). A 2-bit label type indicator specifies the type of label (0, 3, 6 byte, or label reuse). The fragmentation identifier is used for reassembly when GSE packets are spread over different BB frames.

The following provide details about the semantics of GSE packet header:

— *Start_Indicator (S)*: A value of 1 indicates that this GSE packet contains the start of the encapsulated PDU. A value of 0 indicates that the start of the PDU is not present in this GSE packet. For padding, the Start_Indicator is set to 0.
— *End_Indicator (EI)*: A value of 1 indicates that this GSE packet contains the end of the encapsulated PDU. A value of 0 indicates that the end of the PDU is not present in this GSE packet. For padding, the End_Indicator is set to 0.
— *Label_Type_Indicator (LTI)*: This is a 2-bit field. For Start and Complete GSE packets, it indicates the type of label field in use. For Intermediate and End GSE packets, it is set to 11; For Padding GSE packets, it is set to 00; 01 Indicates that a 3-byte label is present and is used for filtering. 10 is for broadcast that no label field is present. All receivers shall process this GSE packet. This combination is used also in non-broadcast systems when no filtering is applied at layer 2, but IP header processing is utilised.
— *GSE_Length*: This 12-bit field indicates the number of bytes following in this GSE packet, counted from the byte following this GSE_Length field. The GSE_Length field allows for a length of up to 4096 bytes for a GSE packet. The GSE_Length field points to the start of the

following GSE packet. If the GSE packet is the last in the BB frame, it points to the end of the BB frame data field or the start of the padding field. For End packets, it also covers the CRC_32 field.

— *Frag_ID*: This is present when a PDU fragment is included in the GSE packet; while it is not present if Start_Indicator and End_Indicator are both set to 1. All GSE packets containing PDU fragments belonging to the same PDU contain the same Frag_ID. The selected Frag_ID is not reused on the link until the last fragment of the PDU has been transmitted.

— *Total_Length*: This field is present in the GSE header carrying the first fragment of a fragmented PDU. The 16-bit field carries the value of the total length, defined as the length, in bytes, of the protocol type, label (6-byte label or 3-byte label), extension headers, and the full PDU. The receiver performs a total length check after reassembly. It may also use the total length information for preallocation of buffer space. Although the length of a single GSE packet is limited to almost 4096 bytes, larger PDUs are supported through fragmentation, up to a total length of 65 536 bytes.

— *Protocol_Type*: This 16-bit field indicates the type of payload carried in the PDU, or the presence of a Next-Header. The set of values that may be assigned to this field is divided into two ranges, similar to the allocation of Ethernet. The two ranges are:

  ○ *Type 1, Next-Header type field*: The first range of the Type space corresponds to the range of values 0–1535 decimal. These values may be used to identify link-specific protocols and/or to indicate the presence of extension headers that carry additional optional protocol fields (e.g. a bridging encapsulation). The range is subdivided into values less than 256 and greater than 256, depending on the type of extension. The use of these values is coordinated by an IANA registry.

  ○ *Type 2, EtherType-compatible type field*: The second range of the Type space corresponds to the values between 0x600 (1536 decimal) and 0xFFFF. This set of type assignments follows DIX/IEEE assignments (but excludes use of this field as a frame-length indicator). All assignments in this space use the values defined for EtherType. The following two Type values are used as examples (taken from the IEEE EtherTypes registry): 0x0800 is for IPv4 payload and 0x86DD for IPv6 payload.

— *6_byte_Label (or 3_byte_Label)*: This 48-bit (or 24-bit) field contains the 6-byte (or 3-byte) label used for addressing.

— *data_byte*: These bytes contain a concatenation of any extension header bytes, and the PDU data. The optional extension header bytes are used to carry one or more extension header(s). The extension header format is defined by the ULE specification [IETF-14b].

— *CRC_32*: This field is only present in a GSE packet that carries the last PDU fragment.

7.7.5 Satellite Link Control layer

The SLC layer is constituted as a set of control functions and mechanisms that mainly ensures the access of IP packet flows to the Physical layer and controls their transfer between distant points. As shown in Figure 7.41, the RCST SLC layer consists of the following functions:

- *Session control function:* The session refers to a communication session. It is a semi-permanent interactive information exchange between communicating devices that is established at a certain time and torn down at a later time. The session control function includes functions of forward link acquisition, logon/logoff procedure, synchronisation procedure, etc.
- *Resource control function:* It is in charge of capacity request generation, buffer scheduling and traffic emission control, allocation message processing (TBTP), and signalling emission control.

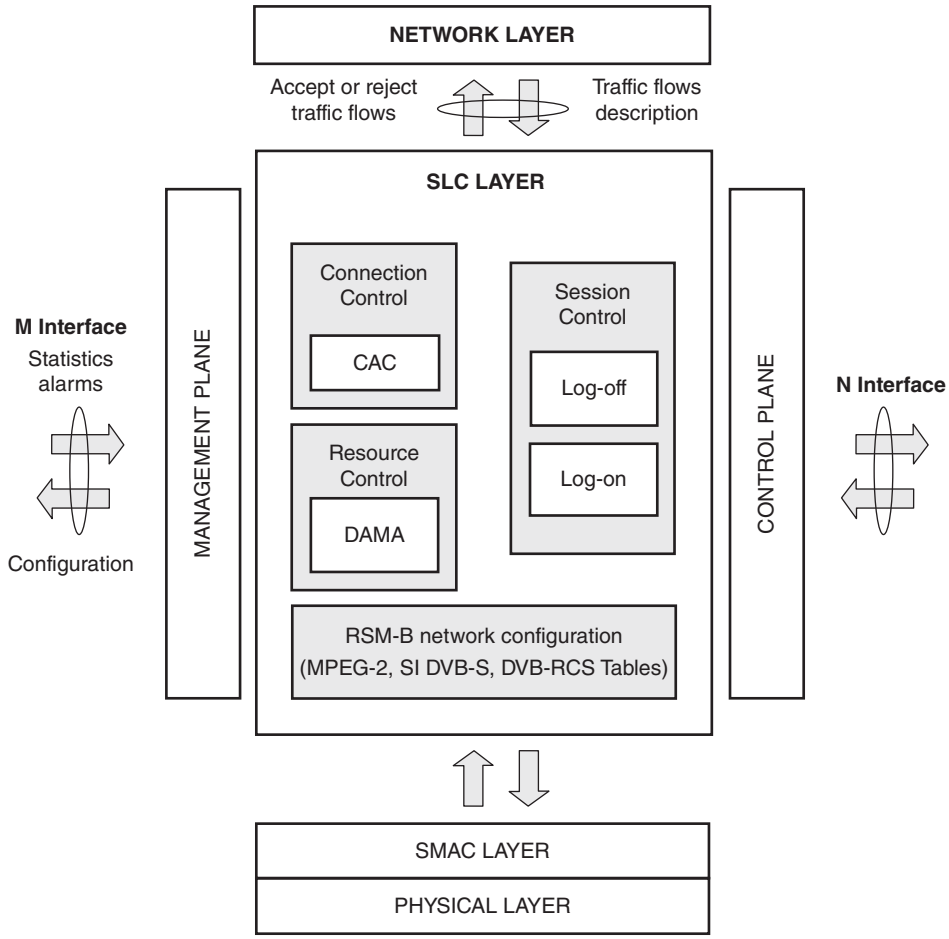


Figure 7.41 RCST satellite link control layer functions. Source: reproduced with the kind permission of the ETSI.

- *Connection control function*. It is in charge of the establishment, release, and modification of connections between two or several RCSTs and between one RCST and the NCC.

The RCST Physical layer interfaces with the management plane (M-plane) and the control plane (C-plane). The RCST reports alarms and statistics to the NMC through the M-plane and is managed and configured by the NMC (based on SNMP and MIB interactions); the RCST reports physical parameter values to the NCC to allow traffic and signalling flow monitoring (synchronisation and power control functions) through the C-plane. In reply, the SLC provides real-time configuration parameters, logon parameters, and traffic packets to be transmitted on the air interface.

The session control procedures are based on the DVB-RCS standard for interactive information exchange between RCST and NCC. Session control is implemented through signalling message exchanges between the RCSTs and the NCC in the session control context.

The resource control is based on the time-slot allocation procedure defined in DVB-RCS standard. The burst-time composition plan is defined thanks to the BTP tables: SCT, FCT, and TCT. Terminal burst-time assignment is given by the terminal burst time plan (TBTP) table.

7.7.5.1 Connection control

This section explains the basic concepts of the *connection control protocol* (C2P), which defines the mechanisms and messages required for the acceptance, establishment, modification, and release of connections. A connection consists of streams; a stream consists of channels; and a channel consists of IP flows.

7.7.5.1.1 Connection

A *connection* is defined as the means for propagating packets (traffic or signalling) with the same priority from one satellite network reference point to one (unicast) or more (multicast or broadcast) distant network reference points.

These satellite network reference points correspond to RCSTs or GWs. Between two RCSTs/GWs there can be as many connections as different priority levels are defined in the system. In a satellite system based on DVB-S/RCS, there are four levels of priority. Therefore, for a satellite mesh system, a maximum of four connections may be established between two RCSTs/GWs. Each connection is identified thanks to a connection reference identifier that allows each RCST/GW to locally identify all the active connections present.

The C2P information element (IE) fields allow for the association of various attributes to the connection according to end-user service needs.

7.7.5.1.2 IP flow

An *IP flow* consists of a number of IP packets having the same source and destination addresses. A connection may carry one or several unitary IP flows. Each RCST is capable of identifying IP flows thanks to a multifield classification.

For example, an IP flow may be identified in terms of IP source and destination addresses, differentiated service code point (DSCP) value, protocol type, and source and destination port numbers. The multifield filtering criteria is configured thanks to a type of flow table on each RCST.

7.7.5.1.3 Channel

A *channel* is the logical access link between an RCST and all its destination RCSTs sharing the same beam. A channel is associated with a physical route and a specific MF-TDMA uplink resource through the TBTP. It is possible to map either a single or a number of connections to one channel depending on QoS and routing considerations. The whole capacity allocated per channel is shared between all the connections established on this channel. Each channel is identified by a channel identifier.

7.7.5.1.4 Stream

A *stream* refers to the MPEG-2 TS flow of packets in DVB. Therefore each connection is identified in terms of MPEG-2 TS stream identifiers. These stream identifiers are also called program

identifiers (PIDs) following MPEG-2 TS nomenclature. In the case of a bidirectional connection, two stream identifiers would uniquely identify the transmission and the reception of traffic. In the case of a unidirectional connection, only one stream identifier is required to identify the transmission or the reception of traffic.

7.7.5.1.5 Connection type

There are two types of connection: *signalling connections* for control and *management and traffic connections* for user data.

Signalling connections are used for communication between the MS (NCC and NMC) and the RCSTs. Each signalling connection may convey only control and management information. Signalling connections are implicitly opened at terminal logon without the need for C2P messages. Therefore, no real connection reference identifier is assigned to them. All the information required for a signalling connection is contained in the logon messages received by the RCST.

Signalling connections are used to send C2P control messages to the NCC and management SNMP messages to the NMC. Each connection has different PID values for transmission and reception, thanks to the logon messages. Different internal queue buffers are assigned to each signalling connection in the RCST. Both connections share the time slots allocated on the reserved signalling channel identifier 0. These connections correspond to the control and management plane of the RCST.

7.7.6 Quality of service

The QoS is configured based on application requirements, traffic classes in the RCST, IP flow classification at the IP layer, and satellite Link layer connection parameters and their adaptations [ETSI-15f; ETSI-15g].

The IP Network layer QoS is defined as a probability of meeting the QoS required by each IP flow from source host to destination host. The Network layer QoS depends on the traffic profile of the satellite system and the traffic profile of the RCST. The Network layer QoS parameters can be defined only when these traffic profiles are agreed upon.

The Network layer QoS is based on IP flows and has to be mapped onto the SLC layer for the most suitable transmission parameters depending on the application involved.

7.7.6.1 QoS requirements from the Application layer

Different applications have different QoS requirements. Real-time applications such as video and voice transmissions are very time sensitive but less sensitive to data loss. They require faster turnaround than non-real-time applications such as file transfer, email, and Web. Non-real-time applications are less time sensitive but very sensitive to data loss.

7.7.6.2 Traffic classes in the RCST

QoS at the RCST level is based on adapting the satellite connection parameters to what the application requires. This requires identifying each application type and managing each of the application flows. The RCST is capable of classifying IP traffic into several IP flows. Each IP flow is identified using a multi-field filter definition. The queuing and scheduling between IP flows depend on the QoS strategy defined within the RCST.

The following traffic priorities are defined in the DVB-RCS terminals:

— *Best effort* or *low priority* (*LP*): Used by applications that do not have specific delay and jitter constraints. This traffic receives the lowest priority in the transmission scheduler of the terminal. A token bucket or weighted fair queue (WFQ) algorithm is used in order to avoid this queue getting blocked by the real-time, non-jitter-sensitive traffic queue.
— *Real-time, non-jitter-sensitive,* or *high priority* (*HP*): Used by applications sensitive to delay but not to jitter. This traffic receives the highest priority in the transmission scheduler of the terminal.
— *Real-time, jitter-sensitive,* or *high priority with jitter constraints (HPj):* Used by applications sensitive to both delay and jitter. This traffic gets specific transmission resources ensuring that the jitter produced by the TDMA transmission scheme of DVB-RCS is minimised. These transmission resources are isolated from other types of traffic.
— *Streaming* or *streaming priority (StrP):* Typically used for video traffic or volume-based applications.

7.7.6.3 Flow classification at the IP layer

The flow-classification mechanism allows RCSTs and GWs provision of different behaviours for different types of application. The IP data-flow identification is performed by the RCST and the GWs upon the following fields in the IP packet [ETSI-15h]:

— Source address
— Destination addresses
— TCP/UDP source and destination port numbers
— DSCP value
— Protocol type carried in the payload

The DSCP is defined within the differentiated service (DiffServ) QoS architecture (see [ETSI-15h; IETF-02; IETF-18]). Routers in the DiffServ domain deal with each IP packet according to the DSCP value (expedited forwarding, assured forwarding, best effort).

RCSTs are configured with a set of masks on these fields. The packets received on the RCST are classified according to these masks into different types of flow.

7.7.6.4 Link layer connection QoS adaptation

Each time a new IP packet flow enters the system, the RCST or the GW must determine if a suitable connection exists for carrying this flow and create it if it does not exist. In this case, the flow type is used to determine the connection parameters to be requested to the NCC, in particular the priority and the bandwidth parameters, such as sustainable data rate (SDR) mapped on the CRA, and peak data rate (PDR) mapped on RBDC.

The association between flow types and connection parameters is configured in the RCST MIB. Up to five flow types plus a default can be defined. The functional description of each entry in the MIB is as follows:

— IP header mask:
  - Source address and mask
  - Destination address and mask

- Source port number range
- Destination port number range
- Protocol type and mask

— SLC-C2P parameters:

- Activity timer
- Priority
- SDR return
- PDR return
- SDR forward
- PDR forward
- Directionality

Based on the type of flow classification, the RCST automatically estimates:

— *The SLC-C2P parameters*: Type of connectivity (unicast/multicast); directionality (unidirectional/bidirectional); high priority (HP) or low priority (LP) traffic type; guaranteed data rate and PDR for that traffic type reception and transmission; activity timeout to release radio resources when no more traffic is present.
— *Buffering queuing*: Distinction between HP and LP buffers (SMAC buffers).

For all traffic, it is possible to define a guaranteed and maximum bit rate. If a guaranteed bit rate is configured for a specific traffic category, as soon as a first packet of a flow requesting this level of QoS enters the satellite network, the guaranteed capacity is reserved. Once no more packets for this kind of traffic go through the network, the capacity allocation timeouts and the radio resources are freed. If needed, the terminal requests more capacity, which is assigned up to the maximum peak rate.

This type of reservation is called *soft state reservation*: the reserved resource is released implicitly; with a traditional hard state connection resource, the resource is not released until a release signalling is issued.

Figure 7.42 illustrates a typical arrangement of unicast connections from one RCST-A attached to one subnet transmitting traffic towards the TDM 1 (it has one HP connection with guaranteed

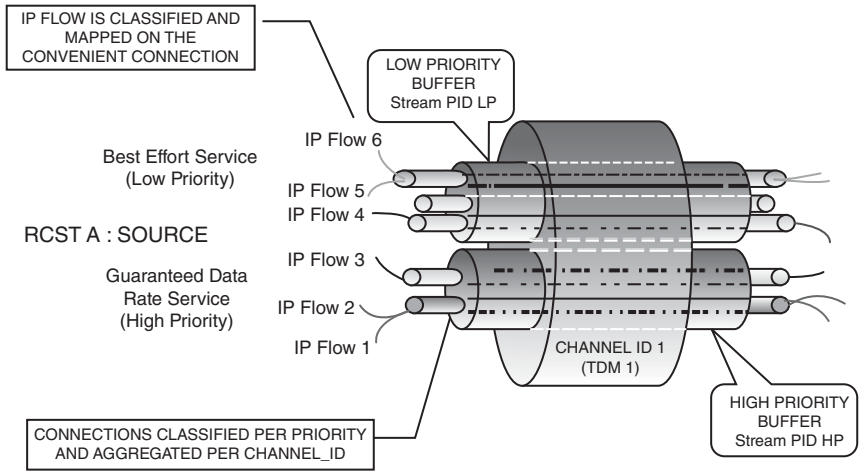


Figure 7.42 Relationships between the connection, channel identifier, and PIDs. Source: reproduced with the kind permission of the ETSI.

data rate services containing IP flows 1, 2, and 3; and one LP connection with best effort services containing IP flows 5 and 6):

- The HP connection towards RCST3 is identified by (ch_ID-1, PID A-1 HP, MAC address RCST3).
- The LP connection towards RCST2 is identified by (ch_ID-1, PID A-1 LP, MAC address RCST2).

7.7.7 Network layer

The Network layer provides end-to-end connections with the external interfaces of the network switch and routers for the traffic data to network services for many different applications, such as VoIP, IP multicast, Internet access, and LAN interconnection.

The RCST Network layer user plane (U-plane) has the following interfaces:

— IP datagrams with the SLC layer
— IP datagrams between the RCST and the user terminal (UT)

In the control plane, the Network layer implements a variety of control plane features as needed to support the provided services.

7.7.7.1 The IPv4 packet header format

An IP datagram consists of a header part and a payload part. The header format is shown in Figure 7.43. The header has a 20-byte fixed part and a variable-length optional part. It is transmitted in big-endian order: from left to right, with the length-order bit of the version field going first. On little-endian machines, software conversion is required on both transmission and reception in the UTs and routers.

The *version field* keeps track of which version of the protocol the datagram belongs to. By including the version in each datagram, each router can deal with the datagram accordingly.

Since the header length is not constant, a field in the header, *Internet head length* (*IHL*), is provided to tell how long the header is, in 32-bit words (4 bytes). The minimum value is 5, which

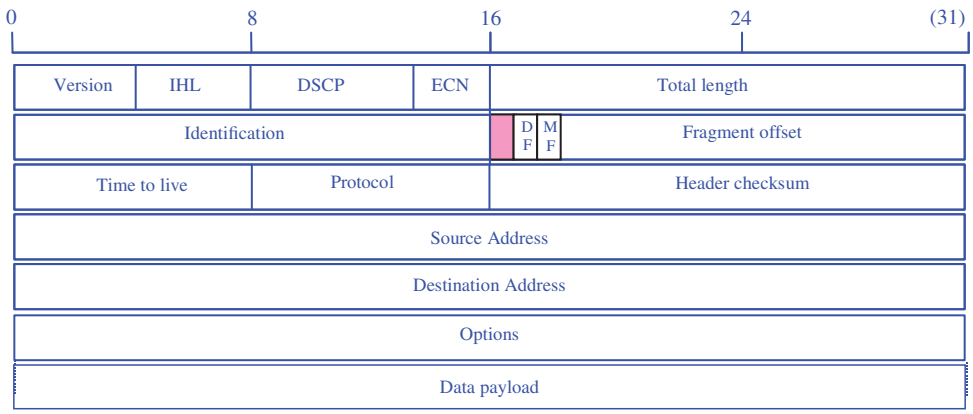


Figure 7.43 The IPv4 header.

applies when no options are present. The maximum value of this 4-bit field is 15, which limits the header to 60 bytes, and thus the options field to 40 bytes:

Differentiated services code point (DSCP): Originally defined as the type of service (ToS), this field specifies differentiated services (DiffServ).
Explicit congestion notification (ECN): This field is defined to allow end-to-end notification of network congestion without dropping packets. It is an optional feature for two endpoints to use it with the support of underlying network.

The *identification field* is needed to allow the destination host to determine which datagram a newly arrived fragment belongs to. All the fragments of a datagram contain the same identification value. Datagrams can be sent in one piece if required (if the destination is incapable of putting the pieces back together) by marking with the DF (don't fragment) bit or more pieces by marking the MF (more fragment). The fragment offset tells where in the current datagram this fragment is positioned. All fragments except the last in a datagram must be a multiple of 8 bytes, the elementary fragment unit. Since 13 bits are provided, there is a maximum of 8192 fragments per datagram, giving a maximum datagram length of 65 536 bytes.

The *time to live field* is a counter used to limit packet lifetimes. It is supposed to count time in seconds, allowing a maximum lifetime of 255 seconds. It is now used for hop count. It must be decremented on each hop from one router to the next. When it hits zero, the packet is discarded and a warning packet is sent back to the source host.

The *protocol field* indicates which Transport layer process is used (TCP, UDP, etc.) so that when the Network layer has assembled a complete datagram, it knows what to do with it. It can also be a Network layer protocol such as internet control message protocol (ICMP) or address resolution protocol (ARP).

The *header checksum* verifies the header only. Checksum is useful for detecting errors generated by wrong memory words inside a router. Note that the header checksum must be recomputed at each hop, because at least one field always changes.

The *source address* and *destination address* indicate the network number and host number that identify uniquely the host and the network it attached to.

The *Options field* was designed to allow subsequent versions of the protocol to include information not present in the original design. Currently five options are defined:

— *Security:* Specifies how secret the datagram is.
— *Strict source routing:* Gives the complete path to be followed.
— *Loose source routing:* Gives a list of routers not to be missed.
— *Record route:* Makes each router append its IP address.
— *Timestamp:* Makes each router append its address and timestamp.

7.7.7.2 IP addressing

The RCST has routing functions. It may host several subscriber subnets. The subscriber subnet interface with the RCST is called the user interface (UI). Each of these subnets may be composed of public or private IP addresses. An IP packet is delivered directly if the destination host is connected to the subnet directly; otherwise it is forwarded to a router that is selected by the routing protocol or a default router. Reaching several subnets through the RCST requires the use of at least one router between these subnets.

7.7.7.2.1 Public IP address

Every host and router on the Internet has an IP address, which encodes its network number and host number. IP addresses are 32 bits long and are used in the source address and destination address field of IP packets.

7.7.7.2.2 Private IP address

Defined in RFC-1918, three blocks of IP address have been reserved for private internets. If the user wants to reach the Internet, this can be performed either through a tunnel or through a network address translation (NAT) function as defined in RFC-1631.

7.7.7.3 IP routing and address resolution

IP routing is the Network layer function. The routing function in the satellite mesh network is organised in a decentralised router. Part of the routing functions are located in the RCSTs/GWs and the other part within the NCC, in a client–server type of architecture. The NCC is the routing server and the RCSTs/GWs are the clients.

Each time a client needs to route an IP packet, it asks the server for the information required to route this packet. The routing information sent by the server is saved in the client. Each time an IP packet is incoming to the satellite mesh network, the RCST or GW determines where to send the packet, the final target being to get the destination equipment MAC address. The RCST or the GW looks within its routing tables and, if the route on the satellite path does not exist, issues an *address resolution protocol* (ARP) request towards the NCC, through the C2P connection request message.

Like a router, the RCST performs IP routing functions. Each time an IP packet enters the RCST, the RCST determines where to send the packet, aiming to get either the destination equipment MAC address or the next hop router (NHR) MAC address.

There is a close relationship between the routing and addressing functions, and between the connection control and management. The connection interconnects distant points across the satellite network as a route across any type of network. The connection between end points is not possible without the knowledge of transit and end points (address) and paths linking these points (routing information).

All this information is centralised in the NCC to set up the connection between end points. The end point can be any user equipment hosted on a subnet located behind an RCST (UI side). The equipment is identified by a unique IP address belonging to one of the subnet masks attached to the RCST. However, since transmission across the satellite network is based upon the MPEG-2 TS packet format, the knowledge of MPE MAC addresses of end RCSTs is mandatory to establish a connection. The NCC provides the mechanisms to associate the IP address of user equipment to the MPE MAC address of the RCST. The mechanism mapping the IP address to the MAC address is called the *address resolution mechanism*.

In order to speed up the connection establishment procedure, the ARP function and the connection establishment can be carried out simultaneously, i.e. the connection establishment request message from the RCST also contains an ARP request; and the NCC response contains both ARP response (destination MAC address and subnets) and connection parameters. These are implemented in one transaction.

The RCST routing table is configured with:

— One or several prefixes covering all the private IP address ranges of all the subscribers of the satellite network.
— One or several prefixes identifying the public IP address ranges allowed in the satellite network (this may also be specific to each RCST).
— Optionally, a default router. Any packet with address not shown on the routing table is sent to the default router.

Depending on the satellite network-addressing plan, the RCST may have a default router or not. A single default router is authorised per RCST. Use of the default routing entry depends on the types of service supported by the RCST.

7.7.8 Regenerative satellite mesh network architecture

The generic concepts and characteristics of DVB-based satellite networking discussed previously are complemented in this section with the specific features of a regenerative satellite network, and with the implementation of IP multicast in a star and mesh configuration.

The satellite network is based on a regenerative satellite and the use of DVB-S2 and DVB-RCS standards. DBV-S2 allows, in addition, the implementation of ACM techniques that bring a significant increase in system capacity and efficient ways to mitigate propagation impairments (critical at Ka band).

The satellite network incorporates the following elements:

— On-board processor (OBP)
— Management station (MS)
— Regenerative satellite gateway (RSGW)
— Return channel satellite terminals (RCSTs)

The components of the satellite network communicate with each other through the interfaces discussed in Section 7.7.1.6 and the O interface between NCC and OBP for OBP control and management, as shown in Figure 7.44.

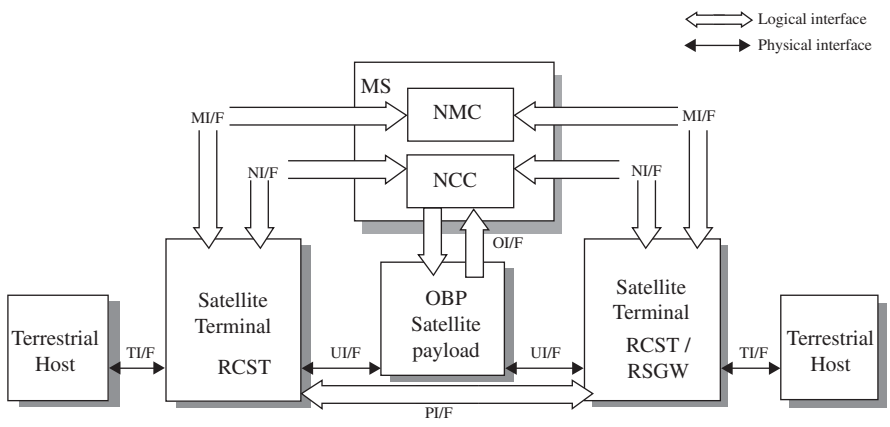


Figure 7.44 Interfaces in the regenerative satellite mesh network architecture. Source: reproduced with the kind permission of the ETSI.

The OBP is the core of the regenerative satellite mesh system. It combines both DVB-RCS and satellite transmission standards with the regenerative multibeam payload. OBP allows routing of MPEG packets from DVB-RCS uplinks to DVB-S2 downlinks in a flexible way to allow full cross-connectivity between the different uplink and downlink beams. Multicast services can be supported (see Section 7.7.8.3) thanks to data replication on board.

Figure 7.45 shows the networking concepts that can be implemented thanks to on-board processing. All the DVB-RCS UTs (RCSTs) connect via the satellite with on-board processing to the gateway (RSGW), forming a network of star topology. The multibeam cross-connectivity capabilities offered by OBP allow also that all the DVB-RCS UTs are interconnected, forming a network of mesh topology. The satellite star connectivity network supports single-hop connectivity between satellite network users and terrestrial network users through the RSGW. The satellite mesh connectivity network supports single-hop connectivity between satellite network users.

From the network point of view, as shown in Figure 7.46, a satellite network can be used to support an IP network, where RCSTs implement routing functions in the same way as IP routers and the OBP acts as a circuit switch at MPEG-2 level.

7.7.8.1 Physical layer

The satellite payload is composed of a multibeam antenna and a regenerative repeater with down-converters, demodulators, BBP, modulators, and up-converters. The down-converter converts the input RF carriers from each uplink beam to IF prior to carrier demodulation; the BBP routes the data packets from different uplink beams aiming at a given downlink beam into multiplexes of MPEG-2 packets; each packet stream is modulated at IF into QPSK (or higher-order

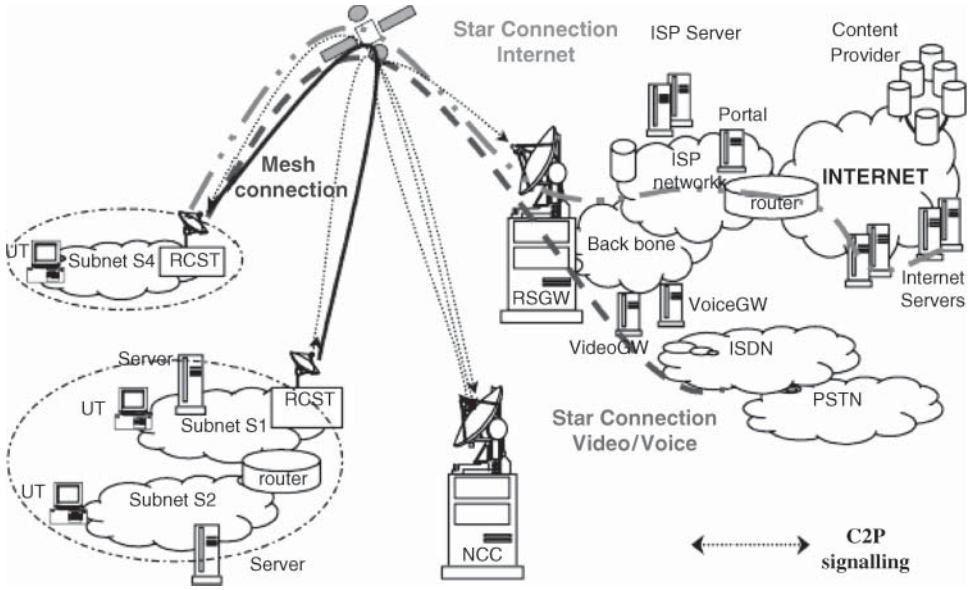


Figure 7.45 The regenerative star and mesh network for DVB-RCS user terminals (RCST). In the star network, RCSTs are linked to the gateway earth station (RSGW); in the mesh network, they are linked with any other. Source: reproduced with the kind permission of the ETSI.

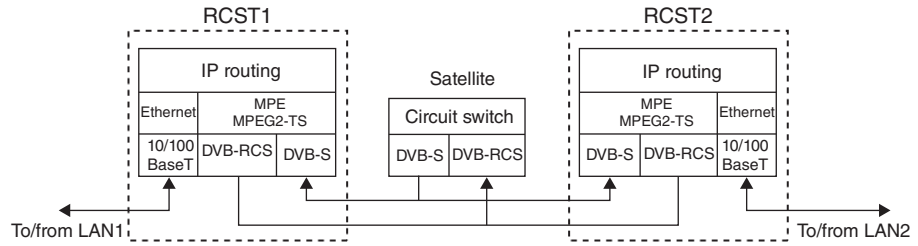


Figure 7.46 Mesh network protocol stack with satellite payload providing on-board switching. Source: reproduced with the kind permission of the ETSI.

modulations with DVB-S2) constellations, and the modulated carriers are up-converted to RF downlink beam frequencies.

7.7.8.1.1 Uplink (compliant with DVB-RCS)

The uplink waveform is compliant with the DVB-RCS standard. The uplink satellite access scheme is MF-TDMA. The uplink is based on the transport of bursts of up to 24 MPEG packets and the transport of customised packets for the logon and synchronisation processes.

The encoding may use any of the turbo codes defined in [ETSI-09b]; typically, two turbo codes (4/5 and 3/4) are used for illustration in the following. The pulse shape of the QPSK modulated signal is based on a root raised cosine filtering with a 0.35 roll-off. Table 7.6 shows the uplink transmission configuration parameters.

7.7.8.1.2 Downlink (compliant with DVB-S)

The downlink conforms to the DVB-S in this example. The on-board DVB processor performs synchronous multiplexing of fixed-length MPEG-2 packets from the different MF-TDMA uplink channels into a TDM downlink signal. Once a TDM bit stream is obtained from the multiplexer section, it is encoded (Reed–Solomon and convolutional coding). All the possible convolutional rates defined in the DVB-S standard could be used (1/2, 2/3, 3/4, 5/6, and 7/8). The transmission rate selected for the downlink is 54 Mbps, which is a QPSK symbol rate of 27 mega symbols per second.

Table 7.6 Uplink transmission configuration parameters.

	CSC bursts	SYNC bursts	TRF bursts
Payload	16 bytes	16 bytes	MPEG packet basis
Modulation	QPSK	QPSK	QPSK
Coding	CRC-16 and turbo code (see note)	CRC-16 and turbo code (see note)	CRC-32 and turbo code (see note)
Inner coding order	Natural		
Filtering	Root raised cosine filtering/roll-off 0.35		

Note: The possible turbo code values are detailed in the DVB-RCS standard (see [ETSI-09b], Clause 8.5.5.4).
Source: reproduced with the kind permission of the ETSI.

Taking into account the premise of allocating an integer number of uplink channels within the downlink frame independently of the code selected, the number of C1 (518.3 kbit/s) uplink carriers that are mapped in the downlink and the maximum number of MPEG-2 packets during a frame as a function of the convolutional rate (CVR) are shown in Table 7.7. Table 7.8 shows the maximum bit rate per TDM according to the CVR.

7.7.8.2 MAC layer tables

The regenerate satellite mesh (RSM) network is a DVB-RCS oriented network. The FLS information is therefore based on the mechanisms and the procedures described by the DVB-RCS standards, and based on the use of MPEG-2, DVB-S, and DVB-RCS tables and messages as discussed in Section 7.7.4.2.

With regenerative systems, the stream of NCR packets used by the RCSTs to regenerate their internal clocks and aid network synchronisation is located on board the satellite. Where the NCR is generated at a ground hub station, the delays between hub and satellite (and the reverse path) must be measured; an allowance for the measurement error is made in the guard time. An on-board NCR clock aids a system architecture that encompasses forward link switching [NEA-01]. The NCR is derived from an on-board reference clock. It is conveyed in the PCR insertion TS packet and distributed by the OBP. There is one NCR counter per downlink TDM. These counters must be aligned in order to ensure system synchronisation.

7.7.8.3 IP multicast

The Network layer provides end-to-end connections with the external interfaces of the network switch and routers for the traffic data to network services for many different applications, such as VoIP, IP multicast [ETSI-15e], Internet access, and LAN interconnection.

Table 7.7 Packets per frame in a downlink TDM.

CVR	Packets per frame
1/2	48 carriers (i.e. $96 \times 1/2) \times 24 = 1152$ packets
2/3	64 carriers (i.e. $96 \times 2/3) \times 24 = 1536$ packets
3/4	72 carriers (i.e. $96 \times 3/4) \times 24 = 1728$ packets
5/6	80 carriers (i.e. $96 \times 5/6) \times 24 = 1920$ packets
7/8	84 carriers (i.e. $96 \times 7/8) \times 24 = 2016$ packets

Source: reproduced with the kind permission of the ETSI.

Table 7.8 Downlink TDM bit rate.

D/L CVR	1/2	2/3	3/4	5/6	7/8
Raw data rate per TDM including RS and CVR (Mbps)	54	54	54	54	54
Reed–Solomon coding factor = 188/204	0.92	0.92	0.92	0.92	0.92
TDM data rate excluding RS (Mbps)	49.76	49.76	49.76	49.76	49.76
TDM data rate excluding RS and CVR (Mbps)	33.18	37.32	37.32	41.47	43.54

Source: reproduced with the kind permission of the ETSI.

The satellite mesh network system supports two types of IP multicast services based on two types of topology:

— *Star IP multicast:* Multicast flows are dynamically forwarded from a GW to several RCSTs. Multicast sources are in the terrestrial network and forward their multicast flows towards the GW.
— *Mesh IP multicast:* Multicast flows are statically forwarded from a source RCST to several destination RCSTs. Multicast sources are in the terrestrial network and forward their multicast flows to a source RCST.

Figure 7.47 illustrates the network topology for star IP multicast. The network entities involved in the star IP multicast network topology include user terminals (UT), RCSTs, the router on the IP subnet of an RCST, and a gateway (RSGW).

The star IP architecture is based on the Internet group management protocol (IGMP) architecture as defined in [ETSI-04b]. The forwarding of multicast flows is dynamic: the IGMP protocol is running between the RCSTs and GW for multicast group membership setup.

The GW forwards multicast flows on the uplink only if at least one RCST has requested to join the multicast group corresponding to this flow.

The GW includes an IGMP adapter. The IGMP adapter is an IGMP proxy optimised to a satellite environment following specification [ETSI-04a]. It performs specific functions to improve the signalling load induced by the IGMP protocol for the satellite networks. It has an interface towards GW-RCSTs and an interface towards a multicast edge router.

Figure 7.48 illustrates an example of mesh IP multicast network topology, which has the same types of network entity as the star topology, except the GW.

Mesh IP multicast consists in the distribution of IP multicast flows from a source RCST over all TDMs in the same satellite network to offer mesh IP multicast services.

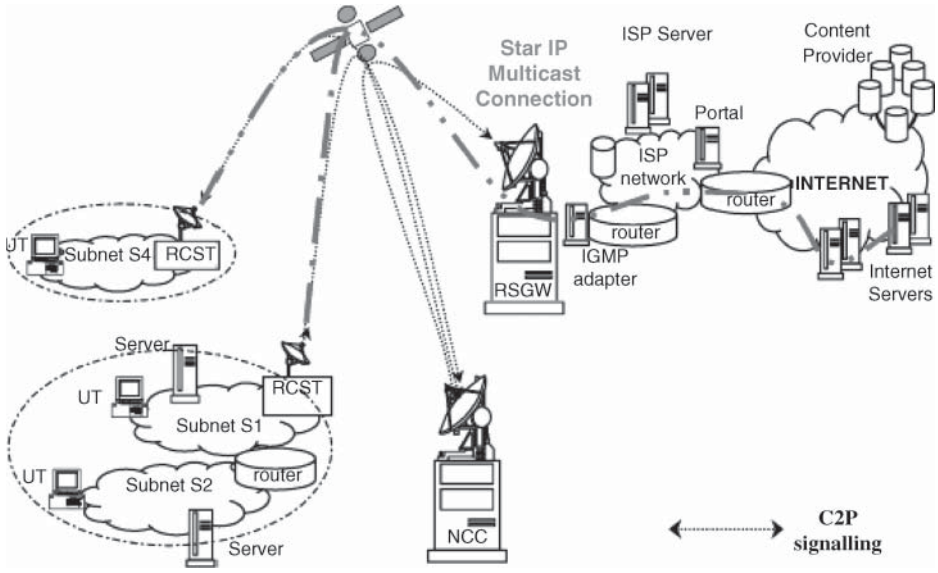


Figure 7.47 Star IP multicast network topology. Source: reproduced with the kind permission of the ETSI.

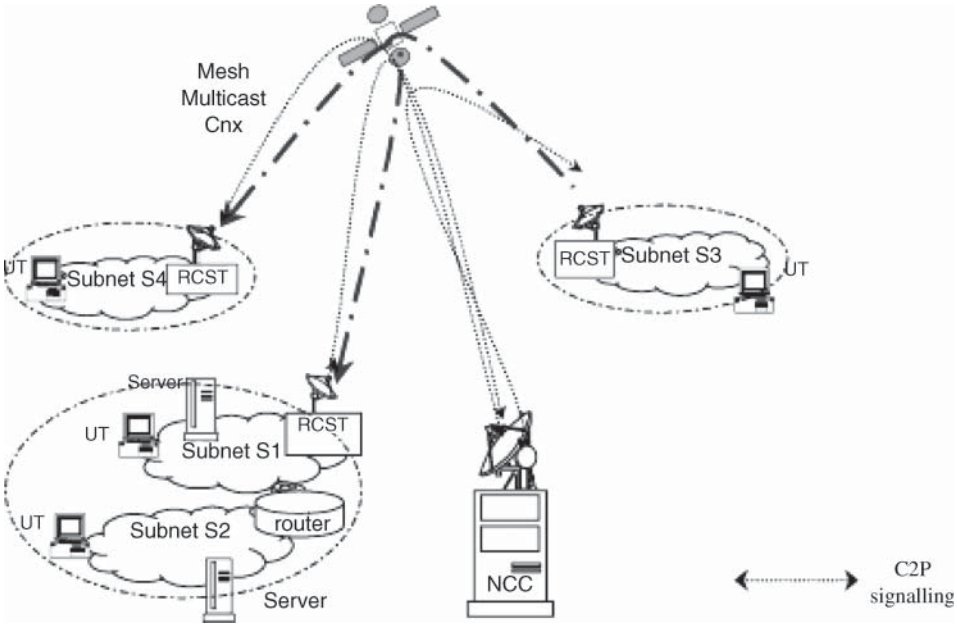


Figure 7.48 Mesh IP multicast network topology. Source: reproduced with the kind permission of the ETSI.

The UT is on the IP subnet connected to the RCST through the UI. It has an IGMP host function to subscribe/de-subscribe to a multicast group. The RCST processes subscriptions of UTs on the UI. On the air satellite interface, an RCST has no IGMP function. It forwards multicast data flow received from the air satellite interface to its UI according to its group membership table when requested. In addition, an RCST has the list of IP multicast group addresses that are authorised to be forwarded to the air satellite interface. This list is defined per RCST and is configured by management. An RCST forwards these authorised IP multicast flows from the UI to the air satellite interface over a point-to-multipoint connection.

Each satellite network managed by a service provider has a pool of IP multicast addresses assigned as recommended in [IETF-98]. Each RCST has a pool of IP multicast addresses authorised to be forwarded.

7.8 TRANSMISSION CONTROL PROTOCOL

The TCP is the protocol between the UTs. It is a connection-oriented, end-to-end protocol. It provides reliable inter-process communication between pairs of processes in host computers. The TCP assumes that it can obtain a simple, potentially unreliable datagram service from the lower-level protocols (such as IP). In principle, the TCP should be able to operate over a wide spectrum of communication systems ranging from hardwired LANs, packet-switched networks, and circuit-switched networks to wireless local area networks (WLANs), wireless mobile networks (3G/4G/5G), and satellite networks.

7.8.1 TCP segment header format

Figure 7.49 illustrates the TCP segment header, which contains the following fields:

— *Source port and destination port:* These groups of 16 bits specify the source and destination port numbers to be used as an address so that processes in the source and destination computers can communicate with each other by sending and receiving data. The IP address identify a host.
— *Sequence number:* These 32 bits identify the first data octet in this segment (except when the SYN control bit is present). If SYN is present, the sequence number is the initial sequence number (ISN) and the first data octet is ISN + 1, which is expected by the receiving host in the Internet; the port number identified the application within the host to receive the IP packet.
— *Acknowledgement number:* These 32 bits, if the ACK control bit is set, contain the value of the next sequence number the sender of the segment is expecting to receive. Once a connection is established, this is always sent. In this way the receiver acknowledges all the packets it received.
— *Data offset:* These 4 bits contain the number of 32-bit words in the header. This indicates where the data begins in the stream of the data bytes. The TCP header (even one including options) is an integer number 32 bits long.
— *Reserved:* These 6 bits are reserved for future use (they must be zero by default).
— *Control bits:* These 6 bits (from left to right) have the following meanings:
  - *URG:* Urgent pointer field significant
  - *ACK:* Acknowledgement field significant
  - *PSH:* Push function
  - *RST:* Reset the connection
  - *SYN:* Synchronise sequence numbers
  - *FIN:* Finish; no more data from sender
— *Window:* These 16 bits contain the number of data octets, beginning with the one indicated in the acknowledgement field, which the sender of this segment is willing to accept.
— *Checksum:* This field consists of 16 bits.
— *Urgent pointer:* These 16 bits communicate the current value of the urgent pointer as a positive offset from the sequence number in this segment.

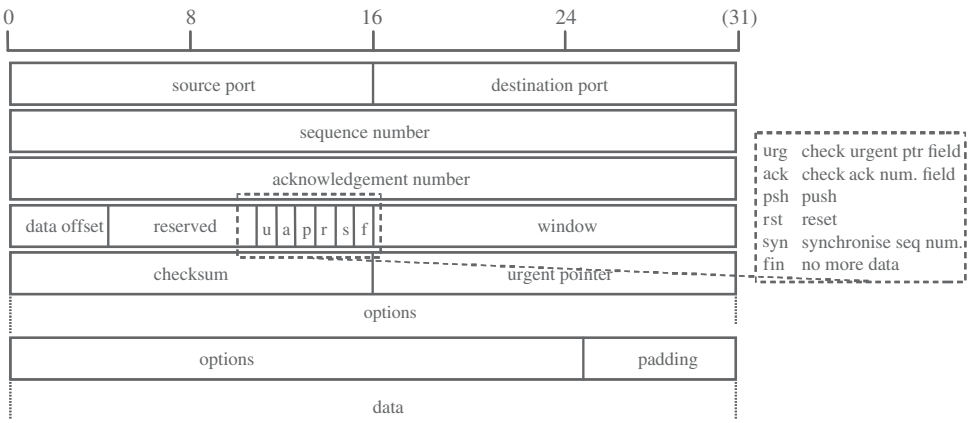


Figure 7.49 TCP segment header.

— *Options and padding:* These fields have variable length; the options field allows additional functions to be introduced to the protocol.
— *Padding:* The TCP header padding is used to ensure that the TCP header ends, and data begins, on a 32-bit boundary. The padding is composed of zeros.

To identify the separate data streams that a TCP may handle, the TCP provides a port identifier. Since port identifiers are selected independently by each TCP, they might not be unique. To provide for unique addresses within each TCP, the Internet address identifying the TCP is concatenated with a port identifier to create a unique socket throughout all subnetworks connected in the Internet. The socket is an implementation of the TCP protocol.

A TCP connection is fully specified by the pair of sockets at the ends. A local socket may participate in many connections to different foreign sockets. A connection can be used to carry data in both directions (a full-duplex connection).

The TCPs are free to associate ports with processes however they choose. However, several basic concepts are necessary in any implementation. Well-known sockets are a convenient mechanism for a priori associating socket addresses with standard service. For instance, the Telnet-server process is permanently assigned to socket 23; FTP data to 20 and FTP control to 21; TFTP to 69; SMTP to 25; POP3 to 110; and HTTP to 80.

7.8.2 Connection setup and data transmission

A connection is specified in the system call OPEN by the local and foreign socket arguments. In return, the TCP supplies a (short) local connection name by which the user refers to the connection in subsequent calls. To store this information, there is a data structure called a transmission control block (TCB).

The procedures to establish connections utilise the synchronise (SYN) control flag and involve an exchange of three messages. This exchange has been termed a *three-way handshake*. The connection becomes established when sequence numbers have been synchronised in both directions. The clearing of a connection also involves the exchange of segments, in this case carrying the FIN control flag.

The data that flows on the connection may be thought of as a stream of octets. The sending process indicates in each system call SEND that the data in that call (and any preceding calls) should be immediately pushed through to the receiving process by setting of the PUSH flag.

The sending TCP is allowed to collect data from the sending process and to send that data in segments at its own convenience, until the push function is signalled; then it must send all unsent data. When a receiving TCP sees the PUSH flag, it must not wait for more data from the sending TCP before passing the data to the receiving process. There is no necessary relationship between push functions and segment boundaries. The data in any particular segment may be the result of a single SEND call, in whole or part, or of multiple SEND calls.

7.8.3 Congestion control and flow control

One of the functions in the TCP is end-host-based congestion control for the Internet. This is a critical part of the overall stability of the Internet. In the congestion-control algorithms, TCP assumes that, at the most abstract level, the network consists of links for packet transmission and queues for buffering the packets. Queues provide output buffering on links that can be temporarily oversubscribed. They smooth instantaneous traffic bursts to fit the link bandwidth.

When demand exceeds link capacity long enough to cause queue buffer overflow, packets are lost. One practice for discarding packets consists of dropping the most recent packet (tail dropping). TCP uses sequence numbering and *acknowledgements* (ACKs) on an end-to-end basis to provide reliable, sequenced, once-only delivery. TCP ACKs are cumulative, i.e. each one implicitly acknowledges every segment received so far. If a packet is lost, the cumulative ACK ceases to advance.

Since the most common cause of packet loss is congestion in traditional wired network technologies, TCP treats packet loss as an indicator of network congestion (such an assumption is not applicable in wireless or satellite networks where packet loss is more likely caused by transmission errors). This happens automatically, and the subnetwork does not need to know anything about IP or TCP. It simply drops packets whenever it must, though some packet-dropping strategies are fairer than others.

TCP recovers from packet loss in two ways. The most important is by a retransmission timeout. If an ACK fails to arrive after a certain period of time, TCP retransmits the oldest unacknowledged packet. Taking this as a hint that the network is congested, TCP waits for the retransmission to be acknowledged before it continues, and it gradually increases the number of packets in flight as long as a *timeout* does not occur again.

A retransmission timeout can impose a significant performance penalty, as the sender is idle during the timeout interval and restarts with a congestion window of one segment following the timeout (*slow start*). To allow faster recovery from the occasional lost packet in a bulk transfer, an alternate scheme, known as *fast recovery*, can be introduced [IETF-99b]. Fast recovery relies on the fact that when a single packet is lost in a bulk transfer, the receiver continues to return ACKs to subsequent data packets, but they do not actually acknowledge any data. These are known as *duplicate acknowledgements* or *dupacks*. The sending TCP can use duplicate acknowledgements as a hint that a packet has been lost and it can retransmit it without waiting for a timeout. Duplicate acknowledgements effectively constitute a *negative acknowledgement* (NAK) for the packet whose sequence number is equal to the acknowledgement field in the incoming TCP packet. TCP currently waits until three duplicate acknowledgements are seen before assuming a loss has occurred; this helps avoid unnecessary retransmission in the face of an out-of-sequence delivery.

In addition to congestion control, TCP also deals with flow control to prevent the sender overrunning the receiver. The TCP *congestion avoidance* [IETF-99b] algorithm is the end-to-end system congestion control and flow control algorithm used by TCP. This algorithm maintains a *congestion window* between the sender and receiver, controlling the amount of data in flight at any given point in time. Reducing the congestion window reduces the overall bandwidth obtained by the connection; similarly, raising the congestion window increases the performance, up to the limit of the available bandwidth.

TCP probes for available network bandwidth by setting the congestion window at one packet and then increasing it by one packet for each ACK returned from the receiver. This is the *slow start* mechanism. When a packet loss is detected (or congestion is signalled by other mechanisms), the congestion window is set back to 1, and the slow start process is repeated until the congestion window reaches half of its setting before the loss. The congestion window continues to increase past this point, but at a much slower rate than before to avoid congestion. If no further losses occur, the congestion window eventually reaches the window size advertised by the receiver. Figure 7.50 illustrates an example of the congestion-control and congestion-avoidance algorithm.

7.8.4 Impact of satellite channel characteristics on TCP

The Internet differs from a single network because different parts may have different topologies, bandwidth, delays, and packet sizes. TCP was formally defined in [IETF-81] and updated in

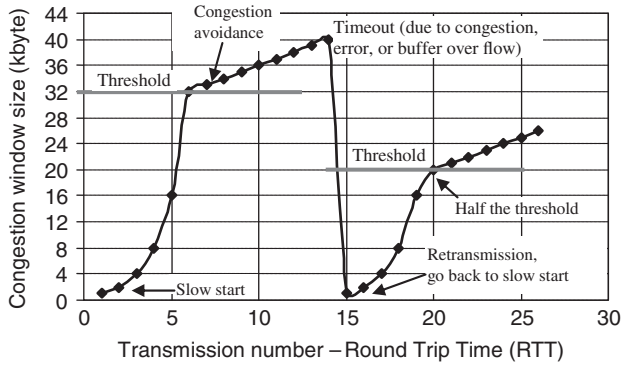


Figure 7.50 Congestion control and avoidance thresholds.

[IETF-89], and extensions are given in [IETF-14a]. TCP is a byte stream, not a message stream, and message boundaries are not preserved end to end. All TCP connections are full-duplex and point-to-point connections. As such, TCP does not support multicasting or broadcasting.

The sending and receiving TCP entities exchange data in the form of segments. A *segment* consists of a fixed 20-byte header (plus an optional part) followed by zero or more data bytes (see Figure 7.49). Two limits restrict the TCP segment size:

— The 65 535 byte IP payload ([IETF-97] describes adapting TCP and UDP to use IPv6, which supports larger datagrams).
— The network maximum transfer unit (MTU). In practice, it is a few thousand bytes and thus defines the upper boundary of the segment size. For example, Ethernet can support a MTU size up to 1500 bytes.

Satellite channels have several characteristics that differ from most terrestrial channels. These characteristics may degrade the performance of TCP, including:

— *Long feedback loop:* Due to the propagation delay with satellite transmission, it takes a significant time for a TCP sender to determine whether a packet has been successfully received at the final destination. This delay affects interactive applications such as Telnet, voice over IP (VoIP), as well as some of the TCP congestion-control algorithms.
— *Large delay * bandwidth (DB) product:* The (DB) product defines the amount of data a protocol should have in flight (data that has been transmitted, but not yet acknowledged) at any one time to fully utilise the available channel capacity. The delay used in this equation is the round trip delay (RTT, end-to-end), and the bandwidth is the capacity of the bottleneck link in the network path. Because the delay in satellite environments is significant, TCP needs to keep a large number of packets sent and waits for acknowledgement.
— *Transmission errors:* Satellite channels may exhibit a higher BER than terrestrial networks. TCP is designed to use all packet drops as signals of network congestion and reduces its window size in an attempt to alleviate the congestion. In the absence of knowledge about why a packet was dropped (congestion at the network transmission errors or corruption at the receiver), TCP must assume the drop was due to network congestion to avoid congestion collapse. Therefore, packets dropped due to corruption cause TCP to reduce the size of its sliding window, even though these packet drops do not signal congestion in the network.

— *Asymmetric use:* Due to the constrained (power-limited link) data rate on the return channel, satellite networks are often asymmetric. When surfing the Web, users always get more date than they send. This asymmetry may have an impact on TCP performance.
— *Variable round-trip times (RTTs)*: In LEO constellations, the propagation delay to and from the satellite varies over time. In particular, connections may be interrupted when handover occurs from one satellite to another satellite; hence delay time can change suddenly due to handover.
— *Intermittent connectivity:* In non-GEO satellite orbit configurations, TCP connections must be transferred from one satellite to another or from one ground station to another from time to time. This handoff may cause packet loss if not properly performed.

7.8.5 TCP performance enhancement (PEP) protocols

According to the principle of protocols, each layer of the protocol should only make use of the services provided by the protocol below it to provide services to the protocol above it. TCP is a Transport layer protocol providing end-to-end, connection-oriented services. Any function between the TCP connection or the IP below it should not disturb or interrupt the TCP data transmission or acknowledgement flows. In this section, we first list the standard mechanisms and then discuss two of the methods widely used in the satellite community: TCP spoofing and TCP cascading (also known as split TCP) to improve TCP performance. These techniques are known as *performance enhancement protocols*.

7.8.5.1 Standard mechanisms

Various techniques and mechanisms have been developed to improve the performance of TCP over satellite networks [IETF-99a; CHO-00; SUN-00]:

— Slow-start enhancement with larger initial window
— Loss-recovery enhancement
— Selective acknowledgement (SACK) enhancement
— Fast retransmission and fast recovery
— Detecting corruption loss
— Congestion-avoidance enhancement
— Multiple data connections
— Sharing TCP state among similar connections
— TCP header compression
— Acknowledgement enhancement

These can be implemented independently or in combination.

7.8.5.2 TCP spoofing

TCP spoofing aims at getting around the slow start for GEO satellite networks. A router implemented before the data is transmitted over the satellite link sends back acknowledgements for the TCP data to give the sender the illusion of a short delay path. The router then suppresses acknowledgements returning from the receiver and takes responsibility for retransmitting any segments lost downstream of the router.

There are a number of problems with this scheme:

— The router must do a considerable amount of work after it sends an acknowledgement. It must buffer the data segment because the original sender is now free to discard its copy (the segment has been acknowledged), so if the segment gets lost between the router and the receiver, the router has to take full responsibility for retransmitting it. One side effect of this behaviour is that if a queue builds up, it is likely to be a queue of TCP segments that the router is holding for possible retransmission. Unlike IP datagrams, this data cannot be deleted until the router gets the relevant acknowledgements from the receiver.
— It requires symmetric paths: the data and acknowledgements must flow along the same path through the router. However, in much of the Internet, asymmetric paths are quite common.
— It is vulnerable to unexpected failures. If a path changes or the router crashes, data may be lost. Data may even be lost after the sender has finished sending and, based on the router's acknowledgements, reported data successfully transferred.
— It does not work if the data in the IP datagram is encrypted, because the router is then unable to read the TCP header.

7.8.5.3 *Cascading TCP (split TCP)*

With cascading TCP, also known as *split TCP*, a TCP connection is divided into multiple connections, with a special TCP connection running over the satellite link. Indeed, the TCP running over the satellite link can be modified to run faster, taking into account the specific features of the satellite link.

Because each TCP connection is terminated, cascading TCP is not vulnerable to asymmetric paths. It works well also in cases where applications actively participate in TCP connection management (such as Web caching). Otherwise, cascading TCP has the same problems as TCP spoofing.

7.9 IPV6 OVER SATELLITE NETWORKS

IP is in transition from the current IP version 4 (IPv4) protocol to IP version 6 (IPv6), which the IETF has developed as a replacement. It has a potential impact on all layers of protocols for trading off processing power, buffer space, bandwidth, complexity, implementation costs, and human factors. In satellite networking, the following two scenarios need to be considered:

— *The satellite network is IPv6-enabled*. This raises issues on UTs and terrestrial IP networks. In practice, it is much easier to upgrade terrestrial UTs and network equipment. Even if all networks were IPv6-enabled, it is still a bandwidth efficiency problem due to the large overhead of IPv6.
— *The satellite network is IPv4-enabled*. This faces similar problems to the previous scenario, but satellite networks may be forced to evolve to IPv6 if all terrestrial networks and terminals are running IPv6. In terrestrial networks, where bandwidth is plentiful, one can afford to delay evolution at the cost of bandwidth. In satellite networks, such a strategy may not be practical. Hence, timing, stable IPv6 technologies, and evolution strategies all play an important role.

7.9.1 IPv6 basics

IPv6 incorporates support for a flow identifier within the packet header for fast packet switching, which the network can use to identify flows, much as VPI/VCI are used to identify streams of ATM cells. Resource reservation protocol (RSVP) helps to associate with each flow a flow specification that characterises the traffic parameters of the flow, much as the ATM traffic contract is associated with an ATM connection.

IPv6 can support integrated services with QoS thanks to such mechanisms and the definition of protocols such as RSVP. It extends the IPv4 protocol to address the problems of the current Internet. The IPv6 protocol:

— Supports more host addresses
— Reduces the size of the routing table
— Simplifies the protocol to allow routers to process packets faster
— Has better security (authentication and privacy)
— Provides different types of service including real-time data
— Aids multicasting (allows scopes)
— Allows mobility (roaming without changing address)
— Allows the protocol to evolve
— Permits coexistence of old and new protocols
— Provides a strategy for evolution

Compared to IPv4, IPv6 has made significant changes to the packet format for the purpose of achieving the objectives of the next-generation Internet with the Network layer functions. Figure 7.51 shows the IPv6 packet header format. The functions of its fields are summarised as follows:

— *Version:* The same as in IPv4 (6 for IPv6 and 4 for IPv4).
— Traffic class*:* Identifier for packets with specific real-time delivery requirements. The six most-significant bits hold the Differentiated Services (DS) field the same as IPv4. The remaining two bits are used for ECN, also the same as IPv4.

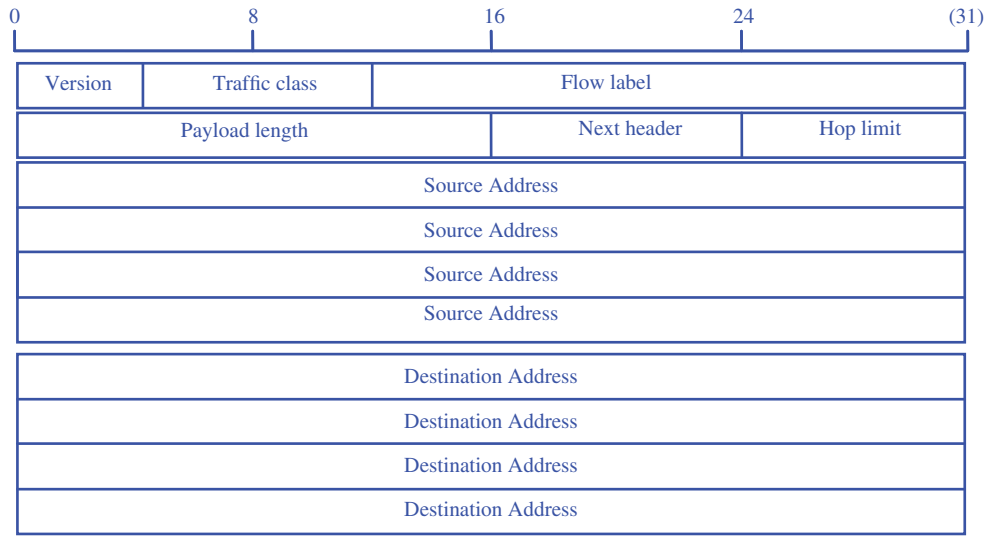


Figure 7.51 IPv6 packet header format.

— *Flow label:* Allows source and destination to set up a pseudoconnection with particular properties and requirements.
— *Payload:* The number of bytes following the 40-byte header (instead of the total length as in IPv4),
— *Next header:* The transport handler to which the packet is passed (similar to the protocol field in IPv4).
— *Hop limit:* A counter used to limit packet lifetime to prevent the packet from staying in the network forever (similar to the time-to-live field in IPv4).
— *Source and destination addresses:* The network number and host number (four times larger than in IPv4).
— Extension headers*:* Similar to the options in IPv4 (see Table 7.9).

Each extension header consists of a next header field and type, length, and value fields. In IPv6, the optional features of the IPv4 become mandatory features: security, mobility, multicast, and transitions. IPv6 tries to achieve an efficient and extensible IP datagram in that:

— The IP header contains fewer fields, which enables efficient routing performance.
— Extensibility of the header offers better options.
— The flow label gives efficient processing of the IP datagram.

7.9.2 IPv6 transitions

The transition is a very important aspect of IPv6 for a successful migration towards IPv6. Many new technologies fail because of the lack of transition scenarios and tools. IPv6 was designed with transition and strategies for transitions from the beginning. For end systems, it uses a dual-stack approach; for network integration, it uses tunnels (some sort of translation from IPv6-only networks to IPv4-only networks).

Dual stack means a node has both IPv4 and IPv6 stacks and addresses. IPv6-enabled applications request both the IPv4 and IPv6 addresses of the destination. The DNS resolver returns IPv6, IPv4, or both addresses to the application. Applications choose the address and can communicate with IPv4 nodes via IPv4 or with IPv6 nodes via IPv6.

7.9.3 IPv6 tunnelling through satellite networks

Tunnelling IPv6 in IPv4 is a technique to encapsulate IPv6 packets into IPv4 packets with the IP packet header protocol field of 41. Many topologies are possible, including router to router, host to router, and host to host. The tunnel endpoints take care of the encapsulation. This process is transparent to the intermediate nodes. Tunnelling is a vital transition mechanism. In the

Table 7.9 IPv6 extension headers.

Extension header	Description
Hop-by-hop options	Miscellaneous information for routers
Destination options	Additional information for the destination
Routing	Loose list of routers to visit
Fragmentation	Management of datagram fragments
Authentication	Verification of the sender's identity
Encrypted security payload	Information about the encrypted contents

tunnelling technique, the tunnel endpoints are explicitly configured and tunnel endpoints must be dual stack nodes.

An IPv4 address is the endpoint for the tunnel. It requires reachable IPv4 addresses. Tunnel configuration implies manual configuration of the source and destination IPv4 addresses and the source and destination IPv6 addresses. Configuration can be between two hosts, one host and one router, or two routers of different IPv6 networks.

7.9.4 6to4 translation via satellite networks

The 6to4 translation is a technique to interconnect isolated IPv6 domains over an IPv4 network with automatic establishment of the tunnel. It avoids the explicit tunnels used in the tunnelling technique by embedding the IPv4 destination address in the IPv6 address. It uses the reserved prefix '2002::/16' (meaning 6to4). It gives a full 48 bits of the address to a site based on its external IPv4 address. The IPv4 external address is embedded: 2002:<ipv4 ext. address > ::/48 with a format of '2002:<ipv4add > :<subnet > ::/64'.

To support 6to4, an egress router implementing 6to4 must have a reachable external IPv4 address. It is a dual-stack node. It is often configured using a loopback address. Individual nodes do not need to support 6to4. The prefix 2002 may be received from router advertisements. It does not need to be dual stack.

Translation from IPv6 to IPv4 gives rise to the following issues:

— The IPv4 external address space is much smaller than that of IPv6.
— If the egress router changes its IPv4 address, it has to renumber the full IPv6 internal network.
— There is only one entry point available. It is difficult to have multiple network entry points for redundancy.

Application aspects of IPv6 transitions give other problems:

— Support of IPv6 in the operating system and in applications is unrelated.
— Dual stack does not mean having both IPv4 and IPv6 applications.
— DNS does not indicate which IP version is to be used.
— It is difficult to support many versions of applications.

The application transitions of different cases can be summarised as follows:

— IPv4 applications in a dual-stack node should be ported to IPv6.
— IPv6 applications in a dual-stack node should map the IPv4 and IPv6 addresses using '::FFFF:x.y.z.w' to make IPv4 applications work on an IPv6 dual stack.
— IPv4/IPv6 applications in a dual-stack node should use a protocol-independent API.
— IPv4/IPv6 applications in an IPv4-only node should be dealt with on a case-by-case basis, depending on the application and operating system support.

7.10 CONCLUSION

This chapter has discussed many important concepts related to satellite networks. These include star and meshed networks over satellites, on-board switching, broadcasting, and multicasting over satellite networks. It has also discussed protocol-layering principles, OSI and IP reference models, IPs, satellite IP networks based on DVB-S/S2/S2X and DVB-RCS/RCS2, the Transport

layer protocol TCP, and its enhancements for satellite networks. Much research has been carried out to solve problems with IP over satellite. Future challenges reside in the efficient deployment of IPv6 over satellites, interworking with other access technologies such as WLAN, WiMAX, and so on [FAN-07].

REFERENCES

[BIN-87] Binder, R., Huffman, S.D., Guarantz, I., and Vena, P.A. (1987). Crosslink architectures for a multiple satellite system. *Proceedings of the IEEE* **75** (1): 74–82.

[BRA-84] Brayer, K. (1984). Packet switching for mobile earth stations via low orbiting satellite network. *Proceedings of the IEEE* **72** (11): 1627–1636.

[CHO-00] Chotikapong, Y. and Sun, Z. (2000). Evaluation of application performance for TCP/IP via satellite links. Presentation at the IEE Colloquium on Satellite Services and the Internet.

[ETSI-04a] ETSI. (2004). satellite earth stations and systems (SES); broadband satellite multimedia (BSM) services and architectures; IP interworking over satellite; multicast group management; IGMP adaptation. TS 102 293 V1.1.1.

[ETSI-04b] ETSI. (2004). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM) services and architectures; IP interworking via satellite; multicast functional architecture. TS 102 294 V1.1.1.

[ETSI-06] ETSI. (2006). Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation system for broadcast, interactive services, news gathering and other broadband satellite applications. EN 302 307 V1.1.2.

[ETSI-07] ETSI. (2007). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); services and architectures. TR 101 984 V1.2.1.

[ETSI-09a] ETSI. (2009). Digital video broadcasting (DVB); guidelines on implementation and usage of service information (SI). TR 101 211 V1.9.1.

[ETSI-09b] ETSI. (2009). Digital video broadcasting (DVB); interaction channel for satellite distribution systems. EN 301 790.

[ETSI-11] ETSI. (2011). Digital video broadcasting (DVB); generic stream encapsulation (GSE) implementation guidelines. TS 102 771 V1.2.1

[ETSI-14a] ETSI. (2014). Digital video broadcasting (DVB); generic stream encapsulation (GSE); part 1: protocol TS 102 606-1 V1.2.1.

[ETSI-14b] ETSI. (2014). Digital video broadcasting (DVB); generic stream encapsulation (GSE); part 2: logical link control (LLC). TS 102 606-2 V1.2.1

[ETSI-14c] ETSI. (2014). Digital video broadcasting (DVB); generic stream encapsulation (GSE); part 3: robust header compression (ROHC) for IP. TS 102 606-3 V1.1.1.

[ETSI-14d] ETSI. (2014), Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications; part 1: DVB-S2. EN 302 307-1 V1.4.1.

[ETSI-14e] ETSI. (2014). Digital video broadcasting (DVB); second generation DVB interactive satellite system (DVB-RCS2); part 1: overview and system level specification. TS 101 545-1 V1.2.1.

[ETSI-14f] ETSI. (2014). Digital video broadcasting (DVB); second generation DVB interactive satellite system (DVB-RCS2); part 2: lower layers for satellite standard. EN 301 545-2 V1.2.1.

[ETSI-14g] ETSI. (2014). Digital video broadcasting (DVB); second generation DVB interactive satellite system (DVB-RCS2); part 3: higher layers satellite specification. TS 101 545-3 V1.2.1.

[ETSI-14h] ETSI. (2014). Digital video broadcasting (DVB); second generation DVB interactive satellite system (DVB-RCS2); part 4: guidelines for implementation and use of EN 301 545-2. TR 101 545-4 V1.1.1.

[ETSI-14i] ETSI. (2014). Digital video broadcasting (DVB); second generation DVB interactive satellite system (DVB-RCS2); part 5: guidelines for the implementation and use of TS 101 545-3. TR 101 545-5 V1.1.1.

[ETSI-15a] , ETSI. (2015). Digital video broadcasting (DVB); DVB specification for data broadcasting. EN 301 192 V1.6.1.

[ETSI-15b] ETSI. (2015). Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications; part 2: DVB-S2 extensions (DVB-S2X). EN 302 307-2 V1.1.1.
[ETSI-15c] ETSI. (2015). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); guidelines for the satellite independent service access point (SI-SAP). TR 102 353 V1.2.1.
[ETSI-15d] ETSI. (2015). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); address management at the SI-SAP. TS 102 460 V1.2.1.
[ETSI-15e] ETSI. (2015). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); multicast source management. TS 102 461 V1.2.1.
[ETSI-15f] ETSI. (2015). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); QoS functional architecture. TS 102 462 V1.2.1.
[ETSI-15g] ETSI. (2015). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); interworking with IntServ QoS TS 102 463 V1.2.1.
[ETSI-15h] ETSI. (2015). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); interworking with DiffServ QoS. TS 102 464 V1.2.1.
[ETSI-16] ETSI. (2016). Digital video broadcasting (DVB); specification for service information (SI) in DVB systems. EN 300 468 V1.15.1.
[FAN-07] Fan, L., Cruickshank, H., and Sun, Z. (eds.) (2007). *IP Networking over Next-Generation Satellite Systems*. Springer.
[GOL-82] Golden, E. (1982). The wired sky. In: *AIAA 9th International Conference, San Diego*, 174–180. AIAA.
[IETF-81] IETF. (1981). Transmission control protocol. DARPA Internet program protocol specification. RFC 793.
[IETF-89] IETF. (1989). Requirements for Internet hosts – communication layers. RFC 1122.
[IETF-97] IETF. Borman, D. (1997). TCP and UDP over IPv6 jumbograms. RFC 2147.
[IETF-98] IETF. Meyer, D. (1998). Administratively scoped IP multicast. RFC 2365.
[IETF-99a] IETF. Allman, M., Glover, D., and Sanchez, L. (1999). Enhancing TCP over satellite channels using standard mechanisms, BCP 28. RFC 2488.
[IETF-99b] IETF. Allman, M., Paxson, V., and Stevens, W. (1999). TCP congestion control. RFC 2581.
[IETF-02] IETF. Grossman, D. (2002). New terminology and clarifications for diffserv. RFC 3260.
[IETF-05] IETF. Montpetit, M.-J., Fairhurst, G., Clausen, H. et al. (2005). A framework for transmission of IP datagrams over MPEG-2 networks. RFC 4259.
[IETF-14a] IETF. Borman, D., Braden, B., Jacobson, V., and Scheffenegger, R. (2014). TCP extensions for high performance. RFC 7323.
[IETF-14b] IETF. Fairhurst, G. (2014). IANA guidance for managing the unidirectional lightweight encapsulation (ULE) next-header registry. RFC 7280.
[IETF-18] IETF. Fairhurst, G. (2018). Update to IANA registration procedures for pool 3 values in the differentiated services field codepoints (DSCP) registry. RFC 8436.
[INU-81] Inukai, T. and Campanella, S.J. (1981). On board clock correction for SS/TDMA and baseband processing satellites. *COMSAT Technical Review* **11** (1): 77–102, Spring.
[ISO/IEC-96] ISO/IEC (1996) Generic coding of moving pictures and associated audio information, ISO IEC 13818.
[LEO-91] Leopold, R.J. (1991). Low earth orbit global cellular communications network. In: *ICC'91*, 1108–1111. IEEE.
[MAR-04] Maral, G. (2004). *VSAT Networks*, 2e. Wiley.
[MAR-87] Maral, G. and Bousquet, M. (1987). Performance of fully variable demand assignment SS-TDMA system. *International Journal of Satellite Communications* **5** (4): 279–290.
[MOR-04] Morello, A. and Reimers, U. (2004). DVB-S2, the second generation standard for satellite broadcasting and unicasting. *International Journal of Satellite Communications and Networking* **22** (3): 249–268.
[MOR-89] Morgan, W.L. and Gordon, G.D. (1989). *Communications Satellite Handbook*. Wiley.
[NEA-01] Neale, J. and Bégin, G. (2001). Terminal timing synchronisation in DVB-RCS systems using on-board NCR generation. *Space Communications* **17** (1–3): 257–266.

[NUS-86] Nuspl, P.P., Peters, R., and Abdel-Nabi, T. (1986). On-board processing for communications satellite systems. In: *7th International Conference on Digital Satellite Communications*, 137–148. VDE-Verlag GmbH.

[SUN-00] Sun, Z., Chotikapong, Y., and Chaisompong, C. (2000). Simulation studies of TCP/IP performance over satellite. Presentation at the 18th AIAA International Communication Satellite Systems Conference and Exhibit, Oakland.

[SUN-14] Sun, Z. (2014). *Satellite Networking: Principles and Protocols*, 2e. Wiley.

[TIR-83] Tirro, S. (1983). Satellites and switching. *Space Communications and Broadcasting* **1** (1): 97–133.

[VIS-79] Visher, P.S. (1979). Satellite clusters. *Satellite Communications* **3** (9): 22–27.

[WAD-80] Wadsworth, D.V.Z. (1980). Satellite cluster provides modular growth of communications functions. In: *International Telemetering Conference ITC80*, 209. IFT.

[WAL-82] Walker, J.G. (1982). The geometry of satellite clusters. *Journal of the British Interplanetary Society* **35**: 345–354.

8 EARTH STATIONS

This chapter is devoted to the organisation of earth stations. In particular, it treats the various aspects and subsystems that determine the performance of the satellite communications system. Consequently, the characteristics of the antenna, transmitting and receiving subsystems, and communication equipment are examined. On the other hand, equipment whose specification is not directly associated with satellite communication, such as that for interfacing with the terrestrial network, switching, multiplexing, and supplying electric power, is considered only in its functional aspect.

8.1 STATION ORGANISATION

The general organisation of an earth station is shown in Figure 8.1. It consists of an antenna subsystem, with an associated tracking system, a transmitting section, and a receiving section. It also includes equipment to interface with the terrestrial network together with various monitoring and electricity supply installations. This organisation is not, in principle, fundamentally different from that of other telecommunication stations such as those for terrestrial microwave links. Among the special features, there is the tracking system, although it may be rather simple in certain cases.

The antenna is generally common to transmission and reception for reasons of cost and bulk. Separation of transmission and reception is achieved by means of a diplexer. Antennas are often capable of transmitting and receiving on orthogonal polarisations (circular or linear) in order to permit reuse of frequencies (see Section 5.2.3).

The tracking system keeps the antenna pointing in the direction of the satellite in spite of the relative movement of the satellite and the station. Even in the case of a geostationary satellite, orbital perturbations cause apparent displacements of the satellite that are, however, limited to the *station-keeping box* (see Section 2.3.4). Furthermore, the station can be installed on a mobile vehicle, the location and direction of which vary with time.

The performance required of the tracking system varies in accordance with the characteristics of the antenna beam and the satellite orbit. For small antennas, the tracking system can be eliminated (fixed mounting), and this enables costs to be reduced.

The size and complexity of stations depend on the service to be provided and the effective isotropic radiated power (EIRP) and figure of merit (G/T) of the satellite. The simplest stations permit reception only and are equipped with a parabolic antenna, which may have a diameter of a few tens of centimetres. The largest are the first Intelsat Standard A stations with antennas of 32 m diameter.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

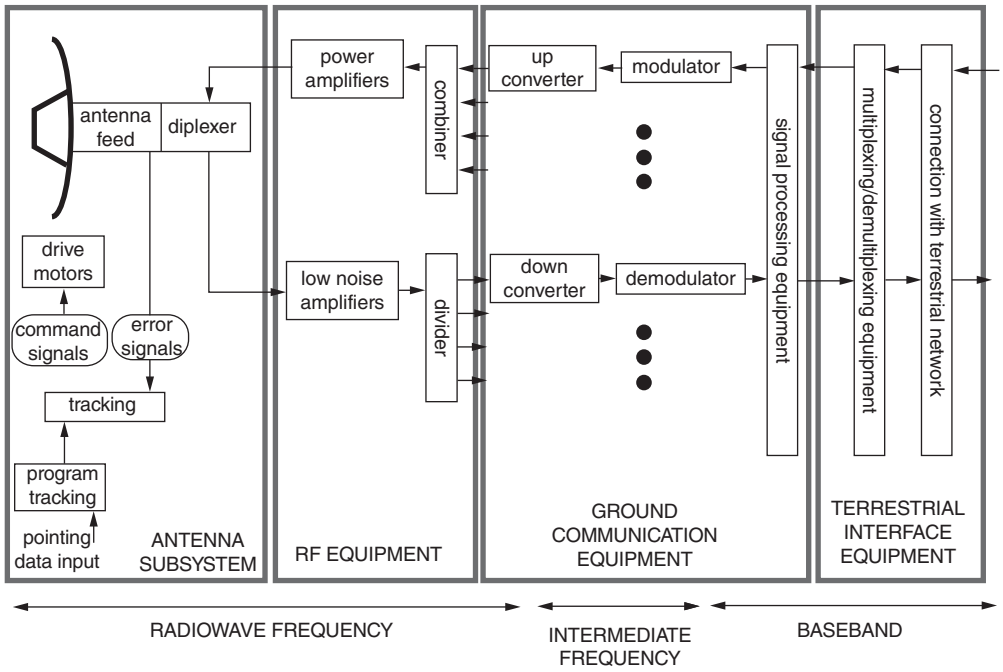


Figure 8.1 The organisation of an earth station.

8.2 RADIO-FREQUENCY CHARACTERISTICS

The characteristics that determine the radio-frequency (RF) performance of earth stations occur in the link-budget expressions for the uplink and the downlink, which were discussed in Chapter 5:

$$(C/N_0)_U = (P_T G_T)_{ES}(1/L_U)(G/T)_{SL}(1/k) \quad \text{(Hz)} \tag{8.1}$$

where $(P_T G_T)_{ES}$ is the EIRP of the earth station, and

$$(C/N_0)_D = (P_T G_T)_{SL}(1/L_D)(G/T)_{ES}(1/k) \quad \text{(Hz)} \tag{8.2}$$

where $(G/T)_{ES}$ is the figure of merit of the earth station.

The expressions are established for a particular link as characterised by the frequency of its carrier, the polarisation of the wave, the type of modulation, and the bandwidth occupied. An earth station usually transmits and receives several carriers for which the characteristics, particularly the EIRP, can be different.

It should also be noted that the transmitting and receiving gains are related to each other since the same antenna is used for transmitting and receiving.

8.2.1 Effective isotropic radiated power (EIRP)

The EIRP is the product $(P_T G_T)_{ES}$ of the available power of the carrier considered at the antenna input and the transmission gain of the antenna in the direction of the satellite at the frequency considered (see Section 5.3).

8.2.1.1 *Available power* P_T

The carrier power P_T available at the antenna input is a function of the rated power $(P_{HPA})_{ES}$ of the power amplifier, the *transmit feeder losses* $(L_{FTX})_{ES}$ between the amplifier output and the antenna input and the *power reduction* $(L_{MC})_{ES}$ required by multicarrier operation:

$$P_T = (P_{HPA})_{ES}(1/L_{FTX})_{ES}(1/L_{MC})_{ES} \quad \text{(W)} \tag{8.3}$$

In most applications, a particular earth station transmits more than one carrier to the satellite. This situation exists when access to the satellite is in frequency division multiple access (FDMA) (see Section 6.5) or with a multibeam satellite when destination selection is by transponder hopping (see Section 7.4.1). The configuration of the earth station amplifying system depends on the way in which the carriers to be transmitted are combined (see Section 8.4.2). $(P_{HPA})(1/L_{MC})$ is the available power on the considered carrier at the output of the power amplifying system, referred to for simplicity as the *transmitting amplifier power* (P_{TX}) in Section 5.4.4.2. The power reduction L_{MC} is then either the amplifier *output back-off* (OBO) (for coupling before amplification) or the *loss of the coupling device* (for coupling after amplification). The rated power P_{HPA} of the amplifier is the output power at saturation with one carrier, also referred to as $(P_{o1})_{sat}$.

8.2.1.2 *Transmission gain* G_T

The maximum gain $G_{T\,max}$ for an antenna of given diameter D is defined for a carrier frequency f_U (cf. Eq. (5.3)) by:

$$G_{Tmax} = \eta_T(\pi D f_U/c)^2 \tag{8.4a}$$

where η_T is the transmission efficiency of the antenna and c is the velocity of light. As the antenna never points perfectly at the satellite, the actual transmission gain G_T differs from the maximum gain $G_{T\,max}$ by a factor L_T, which is a function of the pointing angle error:

$$G_T = (G_{T\,max}/L_T)_{ES} \tag{8.4b}$$

The depointing loss L_T has a value (cf. Eq. (5.18)):

$$L_T = 10^{1.2(\theta_T/\theta_{3dB})^2}$$

where θ_T is the pointing error angle whose value depends on the type of tracking system, and $\theta_{3\,dB}$ is the half-power beamwidth of the antenna at the considered frequency. The influence of the type of tracking is examined in Section 8.3.7.7.

8.2.1.3 *Limitation of the EIRP*

In order to limit interference between satellite systems, the ITU specifies a limit to the value of off-axis EIRP [ITUR-06]. Consider the EIRP density (per 40 kHz) of earth stations operating in the fixed satellite service (FSS) at Ku band. At any off-boresight angle θ larger than 2.5°, in any direction within 3° of the geostationary satellite orbit, the EIRP density should not exceed the values in Table 8.1. For any direction in the region outside 3° of the geostationary satellite orbit, the limits in Table 8.1 may be exceeded by no more than 3 dB.

Table 8.1 Limits on EIRP density

Off-boresight angle	Maximum EIRP (dBW/40 kHz)
$2.5° \leq \theta \leq 7°$	$39–25 \log \theta$
$7° < \theta \leq 9.2°$	18
$9.2° < \theta \leq 48°$	$42–25 \log \theta$
$48° < \theta \leq 180°$	0

8.2.2 Figure of merit of the station

The figure of merit $(G/T)_{ES}$ is defined at the station receiver input as the ratio of the composite receiving gain G to the system noise temperature T of the earth station.

8.2.2.1 Composite receiving gain G

The *composite receiving gain* G is determined from the actual antenna gain G_R by allowing for the losses L_{FRX} suffered by the carrier in the feeder between the antenna interface and the receiver. The real receiving antenna gain G_R differs from the maximum gain $G_{R\,max}$ by a factor L_R, which is a function of the pointing error angle α_R. The maximum gain $G_{R\,max}$ for an antenna of given diameter D is defined at a downlink frequency f_D by $G_{R\,max} = \eta_R(\pi D f_D/c)^2$, where η_R is the efficiency of the receiving antenna:

$$G = (G_R/L_{FRX})_{ES} = (G_{R\,max}/L_R)_{ES}(1/L_{FRX})_{ES} \tag{8.5}$$

The values of the *pointing error angles* θ_R and θ_T are usually identical. On the other hand, the corresponding depointing losses L_R and L_T are not the same for receiving and transmitting for a given value of $\theta_R = \theta_T$. Indeed, L_R and L_T depend on the downlink and uplink frequencies f_D and f_U, which are different, and on the type of tracking. The influence of the type of tracking on the gain is examined in Section 8.3.7.7.

8.2.2.2 System noise temperature T

The concept of *system noise temperature* T is explained in Section 5.5.4. The noise temperature is given by:

$$T = (T_A/L_{FRX})_{ES} + T_F(1 - 1/L_{FRX})_{ES} + T_{eRX} \quad \text{(K)} \tag{8.6}$$

The system noise temperature T is a function of the antenna noise temperature T_A, the feeder losses L_{FRX}, the thermodynamic temperature T_F of this feeder, and the effective noise temperature T_{eRX} of the receiver. Recall that the antenna noise temperature T_A of an earth station depends on the meteorological conditions; attenuation due to rain causes an increase in the noise temperature of the sky (see Section 5.5.3.2). The antenna noise temperature also depends on the elevation angle.

The figure of merit G/T of the station is thus defined for a minimum elevation angle and clear sky conditions. Taking account of rain conditions in the link budget leads to a reduction $\Delta(G/T)$ of the earth station figure of merit.

8.2.3 Standards defined by international organisations and satellite operators

In the early years, international satellite communications services were provided by international organisations. These organisations (now privatised) have defined various standards for earth stations operating in connection with the satellites they operate. These standards specify numerous parameters, e.g. the figure of merit G/T, for different services and applications.

8.2.3.1 Intelsat standards

The characteristics of earth stations used in the Intelsat network are grouped into Intelsat earth station standards (IESS) modules (according to IESS-101, Rev 61, 10 March 2005):

- *Group 1 (Series 100, IESS N°101) – Introductory*: Introduction and approved IESS document list.
- *Group 2 (Series 200, IESS N° 207, and IESS N° 208) – Antenna and RF Equipment Characteristics*: Classes of authorised stations (antenna performance, *G/T*, side-lobe level, etc.).
- *Group 3 (Series 300, N° 307–311, and IESS N° 314–317) – Modulation and Access Characteristics*: Access, modulation and coding, and carrier EIRP.
- *Group 4 (Series 400, IESS N° 401, IESS N° 402, IESS N° 408–412, IESS N° 415, IESS N°417–420, and IESS N°422–424) – Supplementary*: Additional specifications such as the characteristics of the satellites, geographical advantage, intermodulation levels, and service circuits.
- *Group 5 (Series 500, IESS N° 501–503) – Baseband Processing*: System specifications such as digital circuit multiplication equipment (DCME) and digital TV transmission.
- *Group 6 (Series 600, IESS N° 601) – Generic Earth Station Standards*: Performance characteristics for earth stations accessing the Intelsat space segment for international and domestic services not covered by other IESS (Standard G).
- *Group 7 (Series 700, IESS N° 701, and IESS N° 702) – Intelsat Managed Telecommunications Networks*: Performance requirements for the Internet trunking service and dedicated video solutions (space segment only).

8.2.3.1.1 Standards development

The much lower performance of the first communication satellites (Intelsat I – Early Bird – in 1965) imposed large dimensions on earth stations. Standard A Intelsat stations were thus characterised by a figure of merit G/T of 40.7 dBK^{-1} (the 32 m antenna diameter is equipped with a monopulse tracking system and a receiving amplifier with a noise temperature less than 30 K using a maser followed by a parametric amplifier). These characteristics lead to high-cost stations (on the order of $10 M); however, this is only a small part of the cost, in comparison with the cost of the space segment. The requirements were later reduced (revised Standard A, see Table 8.2) to take advantage of increased satellite performance.

In the mid 1970s, a new standard of earth station appeared, Standard B, which was intended to facilitate the development of low-traffic links. With a G/T of 31.7 dBK^{-1} (9 dB less than Standard A), the Standard B station permits the use of an antenna of around 11 m diameter with a simplified tracking system (step by step). This type of station, operated in SCPC/PSK at 64 kbit s^{-1} or in FDM/CFM, makes routes on the order of tens of channels profitable, which is not the case with Standard A stations.

The availability of frequency bands at 14/11 GHz, with the Intelsat V generation (in 1981), led to the specification of a third standard earth station, Standard C, characterised by a clear sky G/T

Table 8.2 Intelsat earth station standards

Standard	Frequency (GHz) up/down	Minimum G/T (dBK^{-1})	Antenna diameter (m)	Services
A (old)	6/4	$40.7 + 10\log(f_{GHz}/4)$	30	TV or FDM/FM/FDMA or
A	6/4	$35 + 10\log(f_{GHz}/4)$	16	TDM/PSK/TDMA (IESS 201)
B	6/4	$31.7 + 10\log(f_{GHz}/4)$	11 to 14	TV or SCPC/QPSK or FDM/FM/FDMA (IESS 202)
C (old)	14/11	$39 + 10\log(f_{GHz}/11)$	14 to 18	TV or FDM/FM/FDMA or
C	14/11	$37 + 10\log(f_{GHz}/11)$	11 to 13	TDM/PSK/TDMA (IESS 203)
E-1	14/11	$25 + 10\log(f_{GHz}/11)$	2.4 to 4.5	IBS and IDR
E-2		$29 + 10\log(f_{GHz}/11)$	4.5 to 7	(IESS 205)
E-3		$34 + 10\log(f_{GHz}/11)$	7.7	
F-1	6/4	$22.7 + 10\log(f_{GHz}/4)$	3.7 to 4.5	IBS and IDR (IESS 207)
F-2		$27 + 10\log(f_{GHz}/4)$	5.5 to 7.5	
F-3		$29 + 10\log(f_{GHz}/4)$	7.3 to 9	
G	6/4 and 14/11	Unspecified	Up to 4.5 m	International and domestic services not covered by Standards A–F earth stations
H-2	6/4	$15.1 + 10\log(f_{GHz}/4)$	1.8	Intelsat DAMA, VSAT IBS
H-3		$18.3 + 10\log(f_{GHz}/4)$	2.4	or broadband VSAT
H-4		$22.1 + 10\log(f_{GHz}/4)$	3.7	
K-2	14/11	$19.8 + 10\log(f_{GHz}/11)$	1.2	VSAT IBS or broadband VSAT
K-3		$23.3 + 10\log(f_{GHz}/11)$	1.8	

of 37 dBK^{-1}. Its antenna diameter, on the order of 15 m, makes it the equivalent at 14/11 GHz of a Standard A station. It was an expensive type of station (more than $5 M), principally suited to trunk (heavy) traffic routes. Again, requirements have been softened at a later stage.

From 1983, enhancement of the characteristics of Intelsat satellites (higher EIRP and greater reuse of frequencies), extensive use of digital techniques (particularly error correcting codes), and a proliferation of new services have led to the introduction of new IESS in association with more and more varied access and modulation systems:

- Standard E stations have nominal *G/T*s of 25.0 dBK^{-1} (Standard E-1) with antenna diameter of about 3 m (Ku band); 29.0 dBK^{-1} (Standard E-2) with antenna diameter of about 5 m (Ku band); and 34.0 dBK^{-1} (Standard E-3) operating in the 14/11 or 14/12 GHz bands for Intelsat Business Services (IBS), IDR international service, and broadband VSAT (BVSAT) services.

- Standard F stations have nominal *G/T*s of 22.7 dBK^{-1} (Standard F-1) with antenna diameter of about 4 m (C band); 27.0 dBK^{-1} (Standard F-2) with antenna diameter 6.3 m (C band); and 29.0 dBK^{-1} (Standard F-3) with antenna diameter 7.3 m operating in the 6/4 GHz bands for IBS, IDR international service, CFDM/FM carriers, and broadband VSAT (BVSAT) services.
- Standard G stations have no specified *G/T* with antenna diameter ranging from less than 1 m up to 7 m, for access at C band or Ku band to the Intelsat space segment for international and domestic services not covered by earth stations of Standards A through F. Standard G earth stations also allow the use of modulation and access techniques other than those specified and approved by Intelsat and only define performance characteristics in terms of general RF boundary conditions.
- Standard H stations have nominal *G/T*s of 15.1 dBK^{-1} with antenna diameter of about 2 m (Standard H-2); 18.3 dBK^{-1} with antenna diameter of about 2.4 m (Standard H-3); and 22.1 dBK^{-1} (Standard H-4) operating in the 6/4 GHz bands for Intelsat DAMA, VSAT IBS, and broadband VSAT (BVSAT) services.
- Standard K stations have nominal *G/T*s of 19.8 dBK^{-1} (Standard K-2) and 23.3 dBK^{-1} (Standard K-3) operating in the 14/11 GHz and 14/12 GHz bands for Intelsat VSAT IBS and broadband VSAT (BVSAT) services.

Table 8.3 shows more detailed information on the associated IBS and VSAT IBS services.

Table 8.3 Intelsat IBS and VSAT IBS services

Characteristic	IBS	VSAT IBS
Network types	Open, closed	Open, closed
Satellites	VI, VII/VIIA, VIII, and IX	VI, VII/VIIA, VIII, and IX
Beams	All beams	All beams, except global
Transmit or receive earth station standards	A, B, C, E, and F	Standards E-1, F-1, H, and K
Terrestrial network connectivity	No connection to public-switched network (PSN)	Connectivity to public-switched packet data network (PSPDN)
Information rate	64 kbit s^{-1} to 2.048 Mbit s^{-1} (open network) 64 kbit s^{-1} to 8.448 Mbit s^{-1} (closed network)	64 kbit s^{-1} to 8.448 Mbit s^{-1} (open and closed network)
Forward error correction (FEC)	Rate 1/2 and 3/4 convolutional encoding/Viterbi decoding (open network) Unspecified (closed network)	Rate 1/2 and 3/4 convolutional encoding/Viterbi decoding (open network) Unspecified (closed network)
Reed-Solomon (219, 201) outer coding	Optional (open and closed network)	Mandatory (open network); optional (closed network)
Modulation	QPSK	QPSK, BPSK
Quality (open network)	Clear sky: $\leq 10^{-8}$ BER(≥95.9% of the year) Threshold: 10^{-3} BER (≥99.96% of the year)	Clear sky: $\ll 10^{-10}$ BER (≥95.9% of the year) Threshold: $\leq 10^{-10}$ BER (≥99.6% of the year)

8.2.3.1.2 Access and modulation modes

IESS modules 301 (FDM/FM), 302 (CFDM/FM), 303 (SCPC/QPSK), 305 (SCPC/CFM), and 306 (TV/FM), dealing with analogue transmission operated in FDMA, were removed on 31 October 2002. The following carrier types are considered further here:

- SCPC/QPSK (IESS 303) has four options in connection with Standard A and B earth stations:
  — Voice or voice band data at or below 4.8 kbit s^{-1}
  — Voice or voice band data above 4.8 kbit s^{-1} using (120, 112) FEC
  — Digital data at 48 or 50 kbit s^{-1} using rate 3/4 convolutional coding
  — Digital data at 56 kbit s^{-1} using rate 7/8 convolutional coding
- TDMA (IESS-307) consists of a 120 Mbit s^{-1} time division multiple access (TDMA) transmission from and to Standard A earth stations within 80 MHz hemi and zone beam transponders at 6/4 GHz. Under clear sky conditions, the TDMA system typically provides a nominal bit error ratio (BER) of better than 10^{-10}. Under degraded sky conditions, the TDMA system is expected to provide a nominal BER of better than 10^{-6} for 99.96% of the year. The main features of the system are:
  — The TDMA frame length is nominally 2 ms.
  — QPSK modulation with coherent demodulation is used.
  — Absolute encoding is employed (i.e. differential encoding is not employed).
  — Forward error correction (FEC) is applied to the traffic bursts.
  — Digital speech interpolation (DSI) is used.
  Each satellite channel (transponder) is served by two reference stations, enabling the terminals to operate with primary and backup reference bursts. Each reference station generates one reference burst per transponder to perform the following functions:
  — Provide open loop acquisition information to traffic terminals and other reference stations
  — Provide synchronisation information to traffic terminals and other reference stations
  — Provide burst-time plan change control
  — Provide common synchronisation across multiple satellite transponders, which permits transponder hopping
  — Provide voice and teletype order wires
- QPSK/IDR (IESS 308) is transmission of intermediate data rate (IDR) (from 64 kbit s^{-1} to 44.736 Mbit s^{-1}) digital carriers using convolutional encoding/Viterbi decoding and QPSK modulation in an FDMA mode in connection with Standard A, B, C, E, and F earth stations. Connections are designed to provide a BER of equal to or better than 2×10^{-8} for more than 95.90% of the year and 2×10^{-7} for more than 99.36% of the year under clear sky conditions, and 7×10^{-5} for more than 99.96% of the year and 10^{-3} for more than 99.98% of the year during degraded (rain) conditions. An optional G.826 Plus Quality is available [ITUT-02] thanks to Reed–Solomon outer coding, and it provides a BER of equal to or better than 10^{-9} for more than 95.90% of the year and 10^{-8} for more than 99.36% of the year under clear sky conditions, and 10^{-6} for more than 99.96% of the year and 10^{-5} for more than 99.98% of the year during degraded (rainy) conditions.
- IBS (IESS 309) stands for Intelsat Business Services. Two distinct offerings are defined:
  — IBS employ digital carriers, which use QPSK modulation with FDMA technique. IBS is designed for communication between Standard A, B, C, E, and F earth stations, which may function as national gateways, urban gateways, or customer-premises installations. The service is not intended to be used for public-switched telephony. IBS networks can be operated in either open network or closed network configurations. Connections are designed to provide a BER of better than 10^{-3} for more than 99.96% of the year. The clear-sky BER performance of IBS is typically less than 10^{-8} at C band and virtually error-free at Ku band.

— VSAT IBS (IBS to and from small earth stations) employ FDMA digital BPSK or QPSK modulated carriers. BPSK modulation is used, whenever necessary, to reduce signal power density, thereby maximising the number of possible VSAT IBS earth station connectivities that can be established. VSAT IBS networks can be operated in either open network or closed network configuration. VSAT IBS carriers are defined as those that terminate at or originate from Standard E-1, F-1, H, and K earth stations only. The baseline coding/modulation scheme for VSAT IBS is rate 1/2 convolutional encoding/Viterbi decoding concatenated with Reed–Solomon (219, 210) outer coding and QPSK modulation. For VSAT IBS carrier transmissions to receiving earth stations larger than Standards E-1, F-1, H, and K, users can request, as an alternative, the use of rate 3/4 convolutional encoding/Viterbi decoding concatenated with Reed–Solomon (219, 210) outer coding. Connections are designed to provide a BER of equal to or better than 10^{-10} (Rate 1/2 or 3/4 FEC) for more than 99.6% of the year. At C band, the clear-sky BER performance provided is much better than 10^{-10} (i.e. practically error-free) due to using rate 1/2 (or 3/4) convolutional encoding/Viterbi decoding concatenated with Reed–Solomon (219, 201) outer coding. At Ku band, the clear-sky BER performance using either FEC scheme is essentially error-free for all information rates by virtue of the larger link margin allocated to mitigate the effects of rain attenuation.

8.2.3.2 Eutelsat standards

The Eutelsat earth station standards (EESS) are published by Eutelsat to provide users with a common source of reference for performance characteristics required from earth stations and associated equipment for access to the Eutelsat space segment and the establishment of communication links. The EESS initially comprised six groups of documents:

— EESS 100, covering an introduction and overview of the documents.
— EESS 200 (telephony services), covering the T-2 Standard for the 120 Mbit s^{-1} TDMA service (EESS 200), the TDMA/DSI System Specification (EESS 201), the DCME Specification (EESS 202), and the intermediate rate digital carrier (IDC) earth station standards I-1, I-2, and I-3 (EESS 203). IDC digital carriers in the Eutelsat system use coherent QPSK modulation and rate 3/4 or rate 1/2 convolutional coding with Viterbi decoding.
— EESS 300 (TV services), covering all TV uplink earth stations, whether they are intended for high-quality contribution links, TV transmissions to leased transponders, or temporary TV transmissions, including satellite news gathering (SNG).
— EESS 400, covering generic EESS providing TV, telephony, or data services, and containing the minimum technical and operational requirements for accessing leased capacity (Standard L).
— EESS 500, dealing with satellite multi services (SMS):
  - The SMS system employs QPSK/FDMA/SCPC for transmitting mainly data carriers. Four standard earth station types are defined for the SMS open network service: Standards S-0, S-1, S-2, and S-3. Standards S-0 and S-1 have the same *G/T*, but Standard S-0 is required to be able to operate over a wider frequency range and to be equipped for dual-polarisation operation.
  - EESS 501 covers the standard structured utilisation or *open network* aspects. It specifies the baseband and modulation equipment for QPSK transmissions using either rate 3/4 or rate 1/2 FEC with Viterbi decoding. With rate 1/2 FEC, the user bit rates covered range from 64 kbit s^{-1} to 2 Mbit s^{-1} (8 Mbit s^{-1} for rate 3/4 FEC).
  - EESS 502 deals with Standard M, for accessing the SMS transponders in PSK/FDMA/SCPC mode by non-standard structured types of SMS carriers. These transmissions are

referred to as *closed network*. VSAT earth stations fall in this category, with antennas on the order of 0.9–1.8 m.

— EESS 600 (G) deals with the EUTELTRACS system, a two-way data communication and position reporting service for mobiles, operating in the 11–12/14 GHz bands.

From October 2007, all previously published Eutelsat standards (Standards L, M, S, and I) have been merged into a single one keeping the designation of Standard M. The purpose of this standard is to define the minimum technical and operational requirements under which 'Approval to Access the EUTELSAT Space Segment' may be granted to an applicant. The existing EESS and guidelines are grouped together and maintained as EESS documents (according to EESS 101 (G) 2008). Table 8.4 displays the list of EESS and related guidelines.

Another group of documents, entitled the *Eutelsat Systems Operations Guides* (ESOGs), provide all Eutelsat space segment users with the necessary information for the successful operation of earth stations within the Eutelsat satellite system: ESOG 100, ESOG 110, ESOG 120, ESOG 140, ESOG 160, ESOG 210, ESOG 220, ESOG 230, and ESOG 240.

8.2.3.3 Other fixed and broadcast satellite service operators

Large satellite operators such as SES, Hispasat, etc., typically define sets of earth station requirements in order to avoid interference to and from adjacent satellites and to guarantee a given quality of service (QoS). This section displays, as an example, some requirements defined by SES, which is a worldwide network of satellite operators, such as SES ASTRA, SES Americom, SES New Skies (earth station requirements SES-NEWSKIES/REG/CMG/001, 2006).

The SES New Skies satellites (NSS-5, NSS-6, NSS-7, NSS-703, NSS-806, NSS-10, NSS-11, IS-603, ASTRA 2B) operate with both left-hand circular polarisation (LHCP) and right-hand circular polarisation (RHCP) uplinks and downlinks at C band and with linear orthogonal polarisations at Ku and Ka band (some satellites also operate at C band with linear polarisation). The required values of transmit polarisation discrimination for stations greater than 2.5 m diameter are shown in Table 8.5.

8.2.3.3.1 C, Ku, and Ka band antenna transmit off-axis gain

It is mandatory that the off-axis transmit antenna copolarised gain of the earth station at an angle u measured between the main beam electrical boresight and the direction considered shall be no higher than the values in Table 8.6.

The off-axis transmit antenna cross-polarised gain of the earth station for all antennas shall not exceed the levels $19–25 \log \theta$ dBi for $2.5 \leq \theta \leq 7$.

8.2.3.3.2 Uplink EIRP

The maximum allowable uplink EIRP values that may be assumed for transmission planning and operational purposes are shown in Table 8.7. Uplink spectral density limits for transmission planning in Ka band is addressed on a case-by-case basis.

8.2.3.3.3 Out-of-band emissions

Stations should perform a wideband out-of-band emissions check. Limits apply over the frequency band specified in Table 8.8, and at off-axis angles greater than 7° (for Ku-band terminals) and 11° (for C-band terminals), providing the band is outside of the user allocations.

Table 8.4 Eutelsat earth station standards (EESS) and related guidelines (EESS 101 G)

EESS No.	Issue/Rev.	Date of issue	Earth station standard	Title	Pages revised since last version	Status
100 (G)	11/1	October 2008	–	Overview of Eutelsat Earth Station Standards	All	
203	7/0	February 2005	1-1, 1-2, 1-3	Intermediate Rate Digital Carrier (IDC) Earth Station Standard I	All	Obsolete since 24/10/2007
400	12/0	August 2006	L	Minimum Technical and Operational Requirements for Earth Stations transmitting to Leased Capacity in the Eutelsat Space Segment–Standard L.	All	Obsolete since 24/10/2007
500	9/1	April 2005	S-l, S-2, S-3	Satellite Multiservice System (SMS) Earth Station Standard S	All	Obsolete since 24/10/2007
501 (G)	3/0	March 2004	–	SMS QPSK/FDMA System Specification	All	
502	11/1	October 2008	M	Minimum Technical and Operational Requirements	Addition of Ka-sat, definition of *a*	
502-Addendum	1/1	October 2008	Mx	Nomenclature of Standard M-x	Page 1	
700 (G)	2/0	October 2007	–	Overview of the Eutelsat Satellite Fleet	All	

Table 8.5 Transmit polarisation discrimination

Band	Performance
C – circular	Voltage axial ratio of 1.09
C – linear	XPD = 30 dB
Ku	XPD = 35 dB

Table 8.6 Off-axis transmit antenna copolarised gain

Antenna Diameter	Copolar gain in direction θ	Off-axis angle θ (degrees)
Diameter > 100	29–25.logθ dBi	$1 \leq \theta \leq 20$
λ/D	−3.5 dBi	$20 \leq \theta \leq 26.3$
	32–25.logθ dBi	$26.3 \leq \theta \leq 48$
	10 dBi	$\theta > 48$
Diameter ≤ 100	29–25.logθ dBi	100. $\lambda/D \leq \theta \leq 20$
λ/D	−3.5 dBi	$20 \leq \theta \leq 26.3$
	32–25.logθ dBi	$26.3 \leq \theta \leq 48$

Table 8.7 Maximum allowable uplink EIRP in Ku band

EIRP power spectral density	Off-axis angle
33–25 logθ dBW/40 kHz	$2.5 < \theta < 7$
+12 dBW/40 kHz	$7.0 < \theta < 9.2$
36–25 logθ dBW/40 kHz	$9.2 < \theta < 48$
−6 dBW/40 kHz	$48 < \theta < 180$

These requirements apply at all conditions from no high-power amplifier (HPA) drive to full operational load and include all periodic and random components of emission including re-radiation of external interferences and harmonics, spurious signals, intermodulation, and noise.

8.2.3.4 Inmarsat standards

The international organisation for mobile maritime telecommunication services, Inmarsat began operating in 1979 to provide satellite communications for ship management and distress and safety situations. Later, the global mobile satellite network was extended to offer land mobile and aeronautical communications. In 1999, Inmarsat became a private company offering a wide range of mobile satellite communications services.

The Inmarsat network comprises four components:

- *Mobile-earth stations* (MESs) operate in L band with a partitioning depending on the type of service. The overall band used in the mobile-to-satellite direction (uplink of return link) is between 1626.5 and 1660.5 MHz, and the band used in the satellite-to-mobile direction (downlink of forward link) is between 1525 and 1559 MHz. The polarisation used is right circular for the uplink and left circular for the downlink.

Table 8.8 Out-of-band emissions limits

Frequency band	Power spectral density
3.4–10.7	55 dBpW/100 kHz
10.7–11.7	61 dBpW/100 kHz
11.7–21.2	78 dBpW/100 kHz
21.2–25.5	67 dBpW/100 kHz

- *Satellites*: Several geostationary satellites are positioned above four ocean regions (Pacific, Indian, and Atlantic East and West). The latest generation of Inmarsat-4 satellites entered into service in 2006 offering personal multimedia communications with data rates from 144 to 432 kbit s^{-1}.
- *Network coordination stations (NCSs):* The transmission and reception of signals are coordinated by four NCSs, one for each satellite coverage region.
- *Land-earth stations (LESs):* A call to or from an MES is routed via an Inmarsat satellite from or to an LES for connection to the national and international phone and data networks. The frequencies used for satellite-to-LES links (feeder links) belong to the FSS and are operated at C band.

Different services and standards have been defined for MESs:

- *Inmarsat-A* is characterised by a clear sky G/T greater than -4 dBK^{-1} for an elevation angle greater than 10°, EIRP of 37 dBW, and parabolic antenna of diameter on the order of 90 cm. For operation at sea, the antenna is mounted on a stabilised platform and provided with a tracking device. Transportable units also exist for land applications (typically 40 kg with cost in the range of $30 000). The station permits transmission of two-way, direct-dial analogue telephone channels in SCPC/FM mode, telex, Group 3 (9.6 kbit s^{-1}) facsimile, and data (9600 bit s^{-1}) (in SCPC/BPSK/TDMA). Also available is a high-speed data service (64 kbit s^{-1}).
- *Inmarsat-B* was introduced into service in 1993 to provide a digital version of the Inmarsat-A voice service. It is characterised by a clear sky G/T greater than -4 dBK^{-1}, EIRP of 33 dBW, using a parabolic antenna of diameter on the order of 90 cm for transmission of telephony at 16 kbit s^{-1} in SCPC/OQPSK, Group 3 facsimile, and duplex data at 9.6 kbit s^{-1}; high-speed data (64 kbit s^{-1}) is also available as a special service. A maritime unit is about 100 kg including tracking devices, while a transportable unit is about 20 kg.
- *Inmarsat-C* is a two-way, packet data service via lightweight (about 7 kg), low-cost terminals small enough to be hand-carried or fitted to any vessel, vehicle, or aircraft. The MES is characterised by a clear sky G/T greater than -23 dBK^{-1} and an EIRP in the range 11–16 dBW with an omnidirectional antenna. The station permits two-way, store-and-forward text messaging or data up to 32 kB in length at an information bit rate of 600 bit s^{-1}. This standard has been approved for use under the Global Maritime Distress and Safety System (GMDSS).
- *Inmarsat-M* is a portable (about 10 kg) mobile satellite phone introduced in 1993, making possible direct-dial 4.8 kbit s^{-1} voice-encoded telephone calls, Group 3 facsimile, data, and group call services from a briefcase-sized terminal. Maritime versions are fitted with radome tracking antennas of about 60 cm (about one-eighth the volume of Inmarsat A/B equivalents) and weighing about 50 kg. It is characterised by a clear sky G/T of -10 to -12 dBK^{-1} and EIRP of 19–27 dBW.

- *Inmarsat-Phone (Mini-M)* exploits the spot-beam power of the Inmarsat-3 satellite to provide voice, facsimile, and data services at 9.6 kbit s^{-1} from a small, notebook-sized terminal weighing less than 2 kg.
- *Inmarsat-E* provides a global maritime distress alerting service. An emergency position indicating radio beacon (EPIRB), measuring between 22 and 70 cm high and weighing about 1.2 kg, sends the vessel's location and an automatic message to maritime rescue coordination centres usually within two minutes of entering the water. It is one component of the GMDSS.
- *Inmarsat-D +* is a two-way data communications service from equipment the size of a personal CD player available with an integrated global positioning system (GPS) facility for tracking, tracing, short data messaging, and supervisory control and data acquisition (SCADA).
- The *Inmarsat-FLEET (F77, F55, F33)* family provides large and small vessels access to Internet, email, and voice communications, at data rates up to 128 kbit s^{-1}, along with a distress and alerting service (for F77).

The Inmarsat global area network (GAN) offers voice telephony and high-speed, wireless data transmission at 64 kbit s^{-1} via a small portable mobile satcom unit (MSU) the size of a notebook computer. GAN offers two types of service: *mobile ISDN* and *mobile packet data*. Mobile ISDN is suited for data-intensive applications such as video conferencing, image transfer, and broadcast-quality voice telephony. Using standard ISDN interfaces, it connects with corporate applications. Mobile packet data is charged according to the amount of data sent, rather than communication time. It is suited to web-based applications, such as intranet access and electronic commerce. By means of channel aggregation, an information bit rate of $4 \times 64 = 256$ kbit s^{-1} is possible.

Broadband global area network (BGAN) offers voice, facsimile, and broadband data transfers, extending to 492 kbit s^{-1} over a shared bearer and providing enterprise users with an integrated high-speed data and voice solution using briefcase, laptop, and palmtop size terminals. These services are made possible thanks to the large satellite EIRP and *G/T* obtained using the large deployable (about 9 m, 80 m^2) multibeam antennas of Inmarsat-4 satellites. Located initially on the Atlantic and Indian oceans, the first two satellites provided service over the Americas, Europe, Africa, the Middle East, and Asia. Following the successful launch of the third Inmarsat-4 (I-4) satellite (initially planned as an on-ground spare) on 18 August 2008, Inmarsat repositioned its I-4 satellites in order to optimise the network, so as to offer the Inmarsat BGAN service on a global basis. The I-4 satellites are more directly positioned over the landmasses, which means optimised service performance for BGAN users.

Inmarsat also provides aeronautical services with a range of AERO terminal types:

- *AERO-C* is the aeronautical equivalent of the Inmarsat-C station and provides a low-data-rate (600 bit s^{-1}), store-and-forward messaging and data-reporting service for aircraft. AERO-C can be used for weather and flight plan updates, maintenance and fuel requests, and business and personal communications.
- *AERO-H* supports multiple simultaneous voice telephony services, Group 3 facsimile at 4.8 kbit s^{-1} and real-time, two-way data communications up to 10.5 kbit s^{-1} for passengers, airline operation, and administrative data applications, anywhere within the global beam. The equipment comprises a steerable, high-gain antenna and suitable avionics. AERO-H+ is an evolution of AERO-C using the Inmarsat-3 spot beams to offer voice at 4.8 kbit s^{-1} and more robust performance, as well as improved operational costs.
- *AERO-I* is designed for short- and medium-haul aircraft and certified by the Civil Aviation Authority for air-traffic management and safety purposes. It offers cockpit and passenger

phone and facsimile communications, packet data from 600 bit s^{-1} to 4.8 kbit s^{-1}, and online access to ground-based information sources and services.

- *AERO-L* provides a low-gain aeronautical satellite communications service offering real-time, two-way, air-to-ground data exchange at 600 bit s^{-1}. It complies with International Civil Aviation Organisation (ICAO) requirements for safety and air traffic.
- *AERO Mini-M* is designed for small corporate aircraft and general aviation users, for voice, facsimile, and 9.6 kbit s^{-1} data. An externally mounted antenna links to a small terminal weighing about 4.5 kg.
- *SWIFT64* is a circuit-mode and packet-mode, aeronautical, high-speed data service to support the full range of 64 kbit s^{-1} ISDN-compatible communications and TCP/IP Internet connectivity. Both services have been designed to meet the needs of aircraft passengers, corporate users, and the flight deck and are based on technology developed by Inmarsat for land-based services. They are designed to take advantage of Inmarsat AERO-H/H + installations.

Table 8.9 summarises the uses and some technical characteristics of Inmarsat stations.

8.3 THE ANTENNA SUBSYSTEM

The characteristics required for an earth station antenna are as follows:

- High directivity, in the direction of the nominal satellite position (for useful signals)
- Low directivity in other directions, in particular that of nearby satellites, to limit interference with other systems
- Antenna efficiency as high as possible for both frequency bands (uplinks and downlinks) on which the antenna operates
- High isolation between orthogonal polarisations
- The lowest possible antenna noise temperature
- Continuous pointing in the direction of the satellite with the required accuracy
- Limitation, as far as possible, of the effect of local meteorological conditions (such as wind, temperature, etc.) on the overall performance

8.3.1 Radiation characteristics (main lobe)

The antennas used, usually parabolic reflectors, are of the type with a radiating aperture. The performance and properties of radiating apertures are treated in numerous works [JOH-84]. The characteristic parameters of an antenna were discussed in Section 5.2. For an earth station antenna, the important parameters that characterise the radiation of the major lobe are the gain, angular beamwidth, and polarisation isolation.

The antenna gain arises directly in the expressions for the EIRP and the G/T of the station. The antenna beamwidth determines the type of tracking system used in accordance with the particular characteristics of the satellite orbit. The value of polarisation isolation determines the ability of an antenna to operate in a system with frequency reuse by orthogonal polarisation. Assuming that the carrier powers of orthogonal polarisations are the same, the interference introduced by the antenna from one carrier to the other is equal to the polarisation isolation, which must, therefore, be greater than a specified value. By way of example, Intelsat advocates, for certain standards and applications, a value less than 1.06 for the axial ratio (AR) in the direction of a satellite with new antennas. This corresponds to a carrier power-to-interference power ratio $(C/N)_I$ greater than 30.7 dB.

Table 8.9 A summary of Inmarsat earth station standards and associated services

Designation of terminal	Operational area	Category of service	Max. radiated power (dBW)	Radiated power (relative to 4 kHz) (dBW)	Channel bandwidth (kHz)	Frequency pattern (kHz)	Type of modulation
Inmarsat-A	Land, sea	Speech, fax, and data transmission	36.0	25.0	50	25 (interleaved)	FM
Inmarsat-A high-speed data	Land, sea	Speech, fax, and data transmission	36.0	23.0	80	100	QPSK
Inmarsat-B	Land, sea	Speech, fax, and data transmission	33.0	27.3	15	20	O-QPSK
Inmarsat-B high-speed data	Land, sea	Data transmission 64 kbit s^{-1}	33.0	20.0	80	100	O-QPSK
Inmarsat-C	Land, sea, air	Data transmission 600 bit s^{-1}	10.5	10.5	0.6	5	BPSK
Inmarsat Mini-C	Land, sea	Data transmission 600 bit s^{-1}	7.0	7.0	0.6	5	BPSK
Inmarsat-D/D +	–	–	0.0	0.0	–	1	32FSK Rx/2 FSK Tx
Inmarsat-M	Land	Speech, fax, and data transmission	25.0	24.0	5	10	O-QPSK
Inmarsat-M	Sea	Speech, fax, and data transmission	27.0	26.0	5	10	O-QPSK
Inmarsat Mini-M	Land	Speech, fax, and data transmission	17.0	17.0	3.5	12.5	O-QPSK
Inmarsat Mini-M	Land, sea, air	Speech, fax, and data transmission	14.0	14.0	3.5	12.5	O-QPSK

Inmarsat GAN	Land	Speech, fax, and data transmission	14.0	14.0	3.5	5	O-QPSK
Inmarsat GAN high speed	Land	Data transmission 64 kbit s^{-1} (incl. MPDS)	25.0	15.0	40	45	16QAM
Inmarsat F77	Sea	Speech, fax, and data transmission	22.0	22.0	3.5	5	O-QPSK
Inmarsat F77	Sea	Fax transmission 9.6 k	29.0	23.3	15	20	O-QPSK
Inmarsat F77 high speed	Sea	Data transmission 64 kbit s^{-1} (incl. MPDS)	32.0	22.0	40	45	16QAM
Inmarsat F55	Sea	Voice	20.0	20.0	3.5	5	O-QPSK
Inmarsat F55	Sea	Fax transmission	22.0	16.3	15	20	O-QPSK
Inmarsat F55	Sea	–	25.0	15.0	40	45	16QAM
Inmarsat F33	Sea	Speech transmission	20.0	20.0	3.5	7.5	O-QPSK
Inmarsat F33	Sea	Fax and data transmission	20.0	14.3	15	20	O-QPSK
Inmarsat F33	Sea	Data transmission (MPDS)	21.0	11.0	40	45	16QAM
Inmarsat Swift 64	Sea	Speech, fax, and data transmission	14.0	14.0	3.5	5	O-QPSK

(*Continued*)

Table 8.9 (*continued*)

Designation of terminal	Operational area	Category of service	Max. radiated power (dBW)	Radiated power (relative to 4 kHz) (dBW)	Channel bandwidth (kHz)	Frequency pattern (kHz)	Type of modulation
Inmarsat Swift 64 high speed	Air	Data transmission 64 kbit s^{-1} (incl. MPDS)	22.5	12.5	40	45	16QAM
Inmarsat Aero H	Air	Speech and data transmission	19.5	13.3	16.8	17.5	QPSK
Inmarsat Aero I	Air	Speech and data transmission	18.0	15.8	6.7	7.5	QPSK
Inmarsat Aero L	Air	Data transmission	9.0	9.0	1.5	2.5	QPSK
Inmarsat Regional BGAN	Land	Data transmission	12.0	−1.5	90	100	n/4-QPSK
Inmarsat BGAN	Land	Speech and data transmission (72 kbit s^{-1})	11.0	−3.0	100	100	QPSK
Inmarsat BGAN	Land	Speech and data transmission (144 kbit s^{-1})	14.5	0.5	100	100	16-QAM
Inmarsat BGAN	Land	Speech and data transmission (432 kbit s^{-1})	10.8	−6.2	200	200	16-QAM

8.3.2 Side-lobe radiation

Most of the power is radiated (or acquired) in the major lobe. However, a non-negligible amount of power is dispersed by the side lobes. The side lobes of an earth station antenna determine the level of interference with other orbiting satellites.

To limit interference, a reference diagram has been proposed for frequencies in the range 2–30 GHz [ITUR-10]:

$$G(\theta) = 32 - 25\ \log\theta \ \text{ for } \ \theta_{\min} \leq \theta < 48^{\circ} \ \text{(dBi)} \tag{8.7a}$$

$$G(\theta) = -10 \qquad \text{for } 48^{\circ} \leq \theta < 180^{\circ}\text{(dBi)} \tag{8.7b}$$

where $\theta_{\min} = 1^{\circ}$ or $100\lambda/D$ degrees, whichever is greater (λ = wavelength, D = diameter).

In order to further limit interference for geostationary satellite systems, [ITUR-04] recommends that antenna manufacturers produce antennas such that the gain G of at least 90% of the side-lobe peaks does not exceed

$$G = 29 - 25\log\theta(dBi) \tag{8.8a}$$

where θ is the off-axis angle. The requirement should be met for any off-axis direction for which $1^{\circ} \leq \theta \leq 20^{\circ}$ and for any off-axis direction within 3° of the geostationary orbit (Figure 8.2). These requirements should be met for θ between 1° or 100 λ/D degrees, whichever is the greater, and 20° for any off-axis direction that is within 3° of the geostationary satellite orbit.

The use of offset mountings with two reflectors seems well suited to obtaining both good RF characteristics (gain and polarisation isolation) for the major lobe and low side-lobe levels. Reduction of side-lobe levels can also be obtained by combining radiation patterns generated by auxiliary sources.

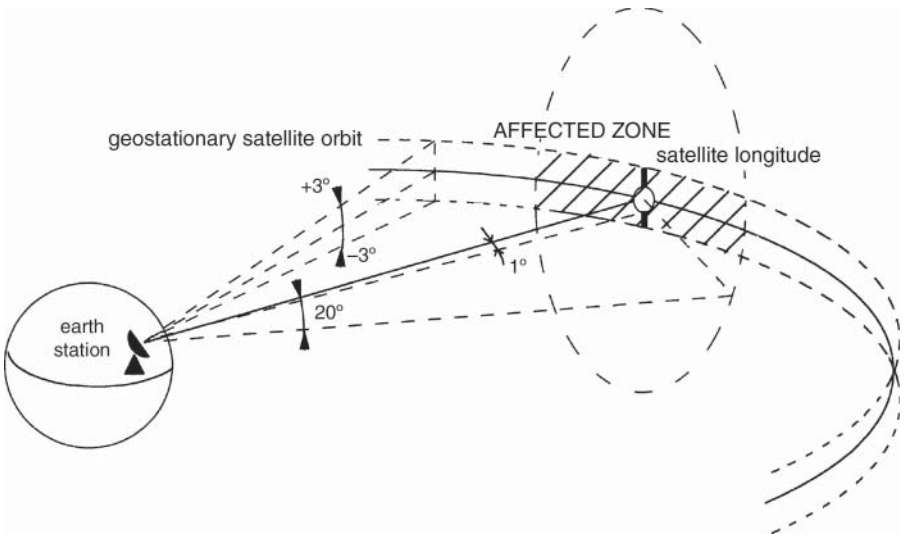


Figure 8.2 Zone around the geostationary satellite orbit to which the design objective for earth station antennas applies.

8.3.3 Antenna noise temperature

The concept of antenna noise temperature was presented in Section 5.5.3. For an earth station, the noise acquired by the antenna originates from the sky and surrounding ground radiation (Figure 5.19). It depends on the frequency, elevation angle, and atmospheric conditions (clear sky or rain). The type of antenna mounting also has an influence on the contribution of radiation from the ground, and this is discussed in Section 8.3.4. Before that, an important phenomenon characterised by an increase of antenna temperature during conjunction of the sun and the satellite is presented [VUO-83a; RAU-85; MOH-88].

Conjunction of the sun and the satellite corresponds to a situation where the earth station, the satellite, and the sun are aligned. In practice, the increase of antenna noise temperature also occurs in the vicinity of these conditions due to the nonzero width of the antenna beam, which captures the noise from the sun even when it is not exactly behind the satellite. Furthermore, the sun is not a point source. Hence, an increase of antenna noise temperature causes significant performance degradation that may lead to link unavailability when the direction of the centre of the sun is within a solid angle of width θ_i (see Section 8.3.3.4): that is, once per day for several days around the date of conjunction (geometric alignment).

The geometric considerations relating to the occurrence conditions of this phenomenon are discussed in [LUN-70; LOE-83; GAR-84; DUR-87] and in Section 2.2.5.7, where expressions are given which permit the number of days and the daily duration to be calculated as a function of θ_i.

8.3.3.1 Brightness temperature of the sun

The increase of antenna noise temperature during conjunction depends on the *brightness temperature* of the sun in the band of frequencies concerned. The brightness temperature of a point on the surface of the sun varies as a function of the wavelength, its position within the solar disc, and solar activity. Various models have been developed to estimate the mean brightness temperature of the sun as a function of the wavelength. An approximate expression for the mean brightness temperature of the sun excluding periods of solar activity is proposed in [RAU-85] for operation in C band:

$$T_{SUN} = (1.9610^5/f)[1 + (\sin 2\pi\{[\log 6(f - 0.1)]/2.3\})/2.3](\mathrm{K}) \tag{8.9}$$

where f is the frequency in GHz.

Another approximate expression [VUO-83b] leads to values in K band that are close to those given by the models of Van de Hulst and Allen [SHI-68]:

$$T_{SUN} = 120000\, f^{-0.75}(\mathrm{K}) \tag{8.10}$$

where f is the frequency in GHz.

These expressions give mean values of the brightness temperature on the solar surface, which is assumed to be quiet. In reality, variations from one point to another are large and greater at low frequencies; at 4 GHz, the temperature varies from 25 000–70 000 K. At 12 GHz, the temperature of the centre of the sun (the cold point) is around 12 000 K and the mean temperature over the whole disc is on the order of 16 000–19 000 K. These brightness temperature variations of the sun as a function of wavelength for the RF domain are illustrated by the curves of Figure 8.3.

In periods of intense solar activity, a considerable increase in the brightness temperature can be observed, particularly at low frequencies – more than 50% during 1% of the time in C band [CCIR-90; ITUT-02]. At 12 GHz, the brightness temperature may be as high as 28 000 K.

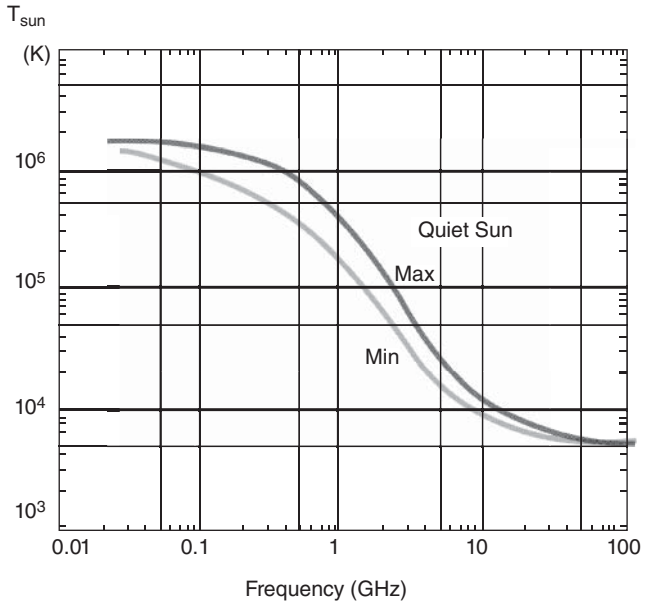


Figure 8.3 Brightness temperature across the surface of the sun as a function of frequency (quiet sun).

The apparent diameter of the sun itself also depends on the wavelength of the radiation and varies inversely with frequency, particularly below 10 GHz. In Ku band, it is slightly greater than the apparent diameter of the solar disc in the visible region: that is, around 0.5°.

8.3.3.2 Increase of the noise temperature during conjunction

The increase of antenna noise temperature is obtained by integrating the product of the brightness temperature $T_{SUN}(\theta; \varphi)$ and the antenna gain $G(\theta; \varphi)$ (with the functions defined in spherical coordinates $(\theta; \varphi)$) over the solid angle through which the sun is viewed [HO-61]:

$$\Delta T_A = (1/4\pi) \int \int_{\text{solar disc}} T_{SUN}(\theta, \varphi) G(\theta, \varphi) \sin\theta \, d\theta \, d\varphi (\text{K}) \tag{8.11}$$

An approximate estimate of the increase in noise temperature is obtained by considering that the antenna radiation pattern is concentrated within a beam of equivalent width θ_e and considering the mean brightness temperature of the sun. The increase in the antenna noise temperature is then proportional to the ratio of the solid angle through which the sun is viewed to the solid angle that corresponds to the equivalent width θ_e of the antenna beam if this width is greater than the apparent diameter of the sun. Otherwise, the increase in antenna noise temperature is equal to the brightness temperature of the sun:

$$\begin{aligned} \Delta T_A &= T_{SUN}(0.5/\theta_e)^2 \quad \text{if } \theta_e > 0.5° (\text{K}) \\ \Delta T_{A\,\text{max unpol}} &= T_{SUN} \quad \text{if } \theta_e < 0.5° (\text{K}) \end{aligned} \tag{8.12}$$

The apparent diameter of the sun is taken to be 0.5°. The equivalent width θ_e of the beam can be taken to be the half-power beamwidth $\theta_{3\,\text{dB}}$.

Electromagnetic waves originating from the sun have a random polarisation. The source of an earth station antenna operating with frequency reuse by orthogonal polarisation is equipped with a polariser that only enables waves arriving with the proper polarisation to be received. Under these conditions, the noise power acquired from the sun, and in turn the increase of noise temperature of the corresponding antenna, is reduced by half. So, when a polariser is used, the increase of antenna noise temperature has the following value:

$$\Delta T_{\text{A max}} = 0.5\, \Delta T_{\text{A max unpol}} \text{ (K)} \tag{8.13}$$

with $\Delta T_{\text{A max unpol}}$ given by Eq. (8.12).

Equations (8.12) show that, for an antenna of small diameter (at a given wavelength), the increase in noise temperature is smaller than for a large antenna. For example, in Ku band (f = 12 GHz), an antenna of 1.2 m with a polariser ($\theta_{3\,\text{dB}} = 1.5°$) has an antenna temperature increase of 900 K, while an antenna of 5 m with a polariser ($\theta_{3\,\text{dB}} = 0.35°$) suffers an increase of 8000 K.

8.3.3.3 Permissible increase of antenna noise temperature with solar conjunction

The link budget under normal operating conditions (clear sky) often retains a margin M_1 with respect to the C/N_0 (the ratio of carrier power-to-noise power spectral density), which is necessary to obtain the nominal required service quality (for a given percentage of the time). The quality objectives with rain generally provide for a permissible degradation for smaller percentages of the time. This value of degradation acts as an additional margin M_2 on the ratio C/N_0 with respect to that necessary under normal operating conditions. These margins $M_1 + M_2$ (in dB), when referred to the figure of merit G/T of the earth station, permit an increase $\Delta T_{\text{A acc}}$ (acc = acceptable) to be accepted in the antenna noise temperature of the earth station.

For the margin $M = M_1 + M_2$, the acceptable increase $\Delta T_{\text{A acc}}$ in the antenna noise temperature is determined from Eq. (8.6):

$$\Delta T_{\text{A acc}} = T(10^{0.1M} - 1)L_{\text{FRX}} \text{ (K)} \tag{8.14}$$

where M is the available margin expressed in dB, T is the clear sky system noise temperature, and L_{FRX} is the connection losses between the antenna interface and the receiver input.

Conversely, for an increase ΔT_{A} in the antenna temperature, the degradation $\Delta(C/N)$ or $\Delta(C/N_0)$ of the carrier to noise or carrier to noise spectral density ratio is equal to the relative increase $\Delta T/T$ of the system noise temperature and is given by:

$$\Delta(C/N) = \Delta(C/N_0) = \Delta T/T) = 10\, \log[TL_{\text{FRX}}/(TL_{\text{FRX}} + \Delta T_{\text{A}})] \tag{8.15}$$

where T is the clear sky system noise temperature and L_{FRX} represents the feeder losses between the antenna interface and the receiver input.

8.3.3.4 Angular diameter of the zone of solar interference

The *zone of solar interference* θ_{i} is defined as the region of the sky such that, when the centre of the sun is within it, the antenna noise temperature T_{A} exceeds the acceptable limit $T_{\text{A acc.}}$

The variation of antenna temperature when the apparent movement of the sun crosses the antenna beam is now examined (Figure 8.4) (this figure assumes that maximum solar interference is obtained at antenna boresight, i.e. the day when solar conjunction occurs):

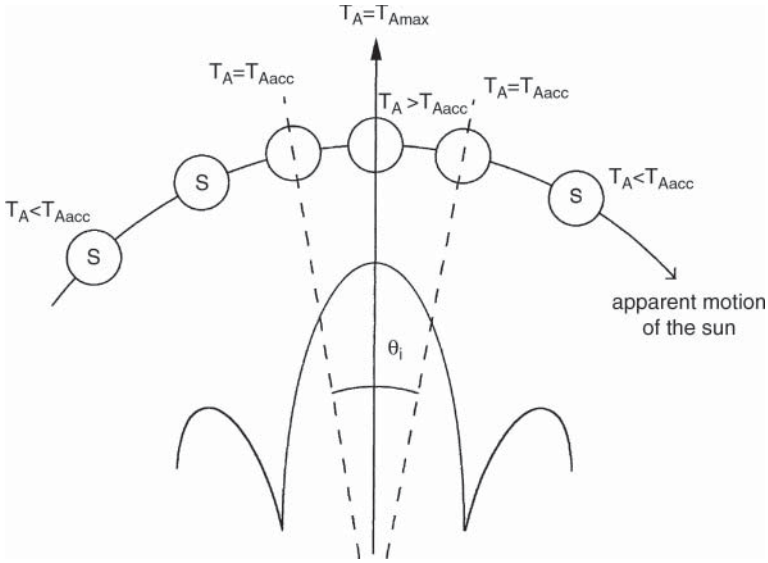


Figure 8.4 Antenna noise temperature variation with apparent motion of the sun.

- As long as the sun is far from the antenna axis, the antenna temperature is equal to its nominal clear sky value (the service quality is greater than the nominal value as a consequence of the margin M_1).
- As the sun approaches the beam, the antenna noise temperature, and hence the system noise temperature increases. At first the available margin M_1 compensates for the increase in system temperature and the nominal quality objective remains fulfilled.
- When the increase in system temperature exceeds the available margin M_1, the quality falls below the nominal objective. As long as the quality remains greater than the value corresponding to degraded operation, with the reservation that the cumulative total of the corresponding durations is less than the specified percentage of time for the degraded mode, operation of the system remains assured. The additional increase of the acceptable system temperature corresponds to the margin M_2.
- When the antenna noise temperature is such that the quality objective in degraded mode is no longer satisfied ($T_A = T_{A\,acc}$, a function of $M_1 + M_2$), interruption of the service occurs. The position of the centre of the sun with respect to the axis of the antenna beam thus defines the angular half diameter of the solar interference zone (Figure 8.4).
- During traversal of the interference zone, the antenna noise temperature T_A is greater than the acceptable limit $T_{A\,acc}$ and the service is interrupted.
- The service is re-established when the quality again becomes greater than the objective in degraded mode as the sun leaves the interference region.

The *angular diameter* θ_i of the zone of solar interference thus depends on the diameter of the antenna beam and the ratio of the increase acceptable antenna noise temperature $T_{A\,acc}$ to the maximum antenna noise temperature $T_{A\,max}$ during conjunction. Figure 8.5 gives the value of the angular diameter θ_i as a function of the ratio $T_{A\,acc}/T_{A\,max}$ and the value of the ratio D/λ for the antenna [CCIR-90].

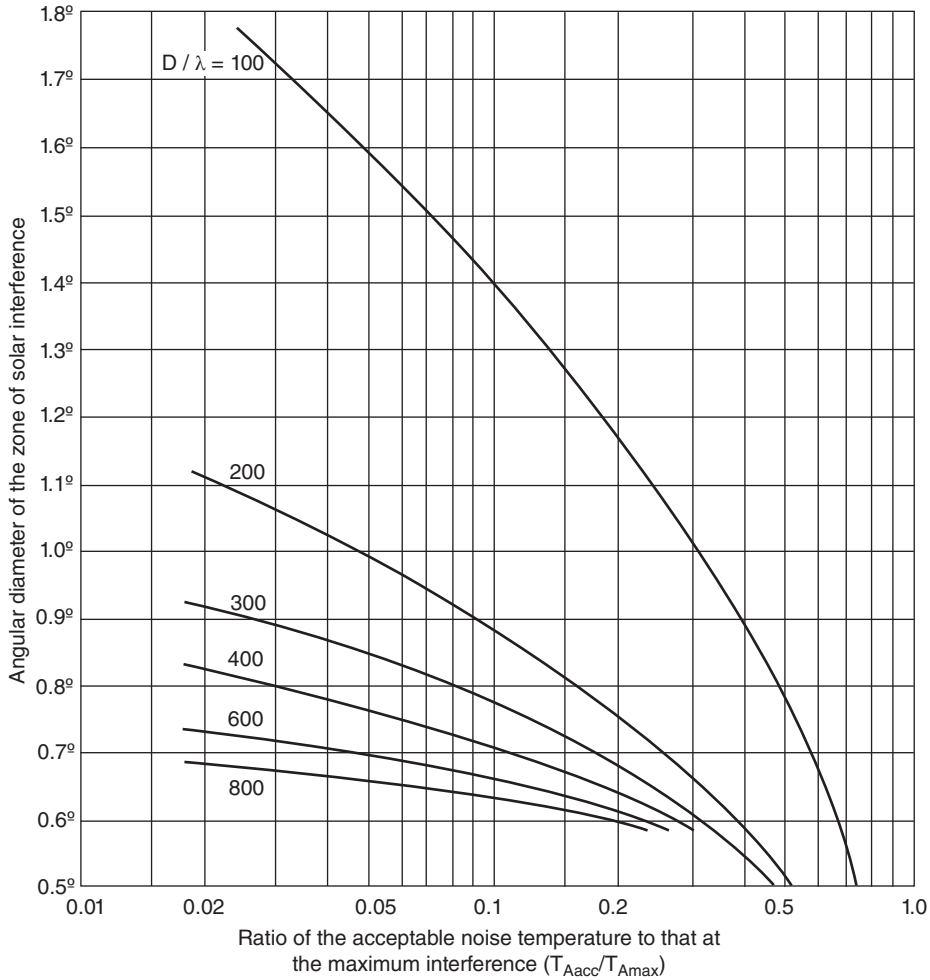


Figure 8.5 Angular diameter of the zone of interference as a function of the ratio of the acceptable noise temperature ($T_{\text{A acc}}$) to that at the maximum solar interference ($T_{\text{A max}}$). Source: reprinted from [CCIR-90] with the permission of the ITU.

Example 8.1 Consider the receiver system characterised by:

- Clear sky system noise temperature $T = 400\,\text{K}$
- Total margin $M = M_1 + M_2 = 5\,\text{dB}$
- Feeder losses $L_{\text{FRX}} = 0.5\,\text{dB}$

The acceptable increase $\Delta T_{\text{A acc}}$ in the antenna temperature calculated from Eq. (8.14) has a value of 970 K.

A station equipped with an antenna of 1.2 m diameter with a polariser operating at 12 GHz whose antenna temperature increase $\Delta T_{\text{A max}}$ is 890 K when in sun–satellite conjunction (see the example in Section 8.3.3.2) thus continues to function (by fulfilling the quality objective in degraded mode) in spite of the solar interference.

On the other hand, a station equipped with an antenna of greater diameter suffers an interruption of service. Neglecting the clear sky antenna noise temperature without solar interference, the ratio between the acceptable antenna noise temperature and the maximum antenna noise temperature during conjunction has a value of $970/8000 = 0.12$. Consider an antenna of 5 m diameter ($D/\lambda = 200$); from Figure 8.5, the angular diameter θ_i of the solar interference zone is 0.85°.

The duration of service interruption for a geostationary satellite under these conditions is calculated. The apparent movement of the sun is $360°/24 \times 60 = 0.25°/\text{min}$ (see Section 2.2.5.7); if the aiming direction of the antenna remains fixed, the maximum duration T_i of service interruption is:

$$T_i = (\theta_i/0.25)\text{min} = 4\theta_i \quad \text{(min)} \tag{8.16}$$

θ_i, the angular diameter of the zone of solar interference, is expressed in degrees. For the example considered, the duration T_i has a value of $4 \times 0.85°$ or 3.4 minutes.

When the available margin is small, it can be assumed that solar interference that leads to an interruption of service occurs as soon as the solar disc penetrates the major lobe of the antenna. The antenna beamwidth to be considered is thus on the order of $2\theta_{3\,\text{dB}}$. The angular diameter u_i of the solar interference zone can thus be expressed approximately as:

$$\theta_i = 2\theta_{3\text{dB}} + 0.5° \tag{8.17}$$

8.3.4 Types of antenna

The antennas considered belong to various classes:

- Horn antenna
- Phased array antenna
- Parabolic antenna

The horn antenna allows a high figure of merit to be achieved, but it is expensive and cumbersome even if its bulk can be reduced with a folded horn. This type of antenna was used at the start of space communication for experimental links with the Telstar satellite (Pleumeur Bodou in France). This technology is no longer in use.

Phased array antennas have an advantage when the beam is in constant movement, as is the case for stations mounted on mobiles; however, the technology remains relatively difficult and costly, which limits the use of this type of antenna.

The most common antennas are those with parabolic reflectors. The three principal mountings are:

- *Symmetrical* or *axisymmetric* mounting
- *Offset* mounting
- *Cassegrain* mounting

8.3.4.1 Antenna with symmetrical parabolic reflector

Figure 8.6 illustrates the mounting of an antenna with a parabolic reflector, which has symmetry of rotation with respect to the principal axis on which the primary feed is placed at the focus. The main weakness of this mounting is that the feed supports and the feed itself have a masking

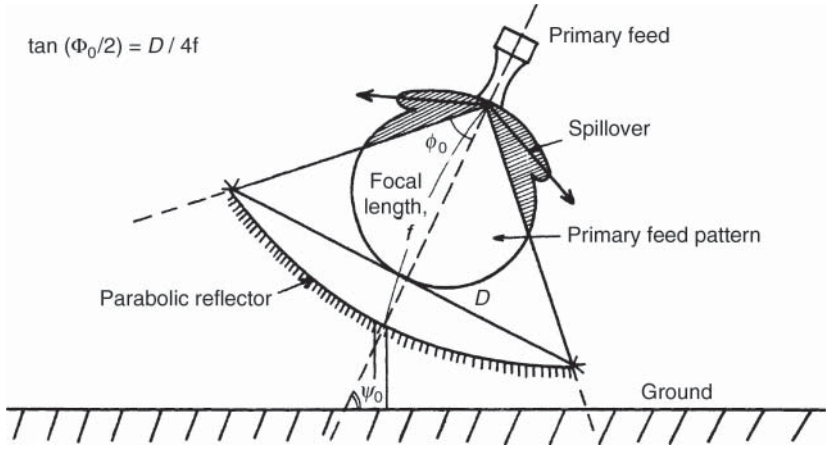


Figure 8.6 Axisymmetric parabolic reflector antenna.

effect on the radiating aperture (aperture blocking). This blocking leads to a reduction of antenna efficiency and an increase in the level of the side lobes due to diffraction by the obstacles.

Furthermore, the primary feed faces the earth, and the part of the radiation pattern of the primary feed that does not intercept the reflector (spillover) easily captures the radiation emitted by the ground, and this makes a relatively large contribution to the antenna noise temperature (several tens to around 100 K).

Spillover is attenuated if the amplitude of the primary feed radiation is reduced at the edge with respect to its value at the centre (tapering). To obtain a low noise temperature, a directional primary feed and a long focal length are necessary. The antenna is thus cumbersome and badly suited to the installation of microwave circuits immediately behind the feed; the bulk of such circuits would have a substantial masking effect.

8.3.4.2 Offset mounting

Offset reflector mounting enables microwave circuits to be located immediately behind the primary feed without masking effects. It does not involve, as the name might suggest, offsetting the feed with respect to the focus, but the use of that part of the parabola situated on one side of the vertex for the reflector profile (Figure 8.7). Presently used for antennas of small diameter (1–4 m), this mounting is little used for large antennas, for which the Cassegrain mounting is preferred. With offset mounting, the spillover is again towards the ground and the antenna temperature remains high.

8.3.4.3 Cassegrain mounting

With *Cassegrain mounting* (Figure 8.8), the phase centre of the primary feed is situated at the first focus *S* of an auxiliary hyperbolic reflector. The other focus *R* of the auxiliary reflector coincides with the focus of the main parabolic reflector. If *D* is the aperture diameter of the parabolic reflector and f_d its focal length, the solid apex angle $2\Phi_o$, under which the reflector is viewed, is given by:

$$\tan(\Phi_0/2) = D/4f_d \tag{8.18}$$

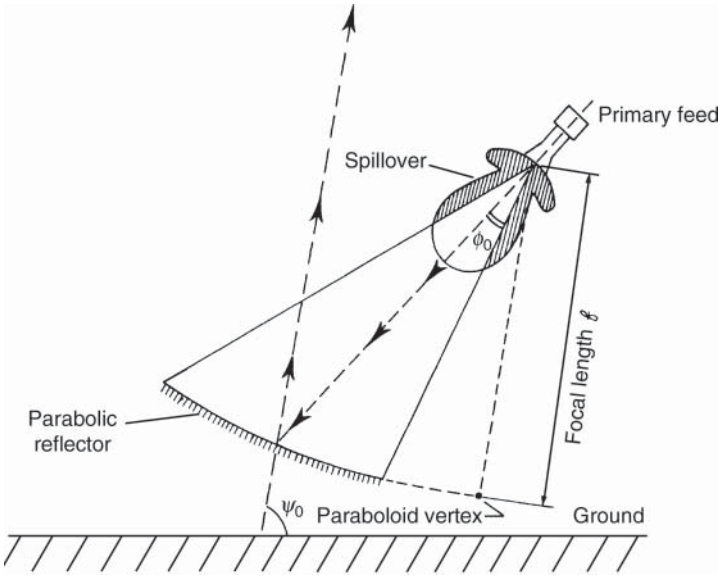


Figure 8.7 Offset-fed parabolic reflector antenna.

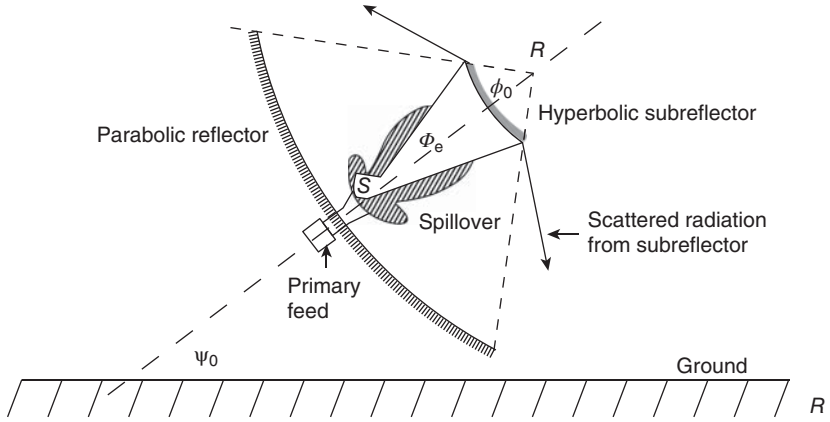


Figure 8.8 Dual-reflector Cassegrain antenna.

The performance of a Cassegrain antenna is evaluated by using the concept of an equivalent paraboloid. An equivalent parabolic reflector antenna is defined as an antenna that has a single reflector of identical diameter to that of the main reflector of the Cassegrain antenna and a focal length equal to that of the Cassegrain assembly. The equivalent paraboloid is thus of diameter D and focal length f_e and characterised by the apex angle $2\Phi_e$ of the auxiliary reflector as viewed from the focus S (Figure 8.9).

The Cassegrain antenna is thus less cumbersome although it retains the advantage of antennas with a long focal length: the antenna noise temperature is low (i) since the greatest part of the spillover is no longer directed towards the ground but towards the sky and (ii) because

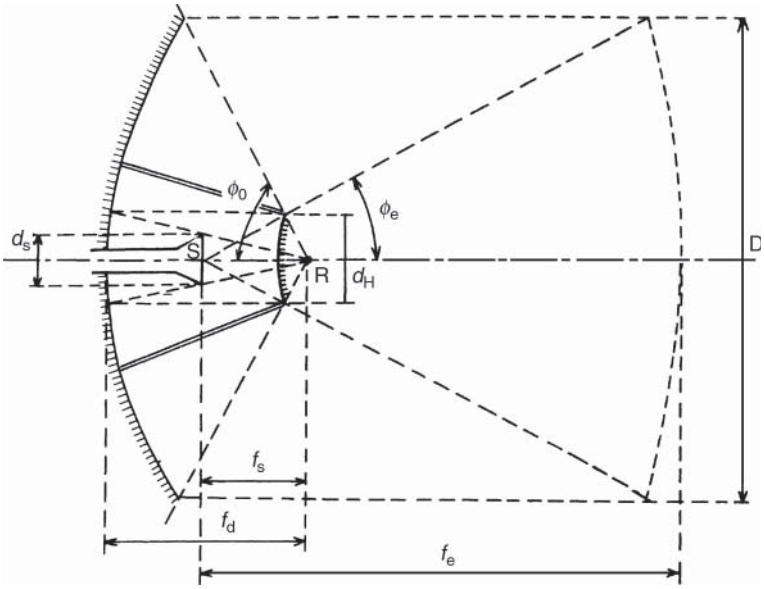


Figure 8.9 Single reflector with equivalent focal length of a dual-reflector Cassegrain antenna.

the spillover is reduced; the high value of equivalent focal length permits the use of directional primary feeds, and the low values of the actual focal lengths f_d and f_s of the parabolic and hyperbolic reflectors attenuate the spillover.

Another advantage is that microwave circuits can easily be located immediately behind the primary feed, which is located behind the main reflector. The effect of losses in the link are thus limited. However, for antennas of large diameter (e.g. 30 m), this equipment is situated at a significant height above the ground. To facilitate maintenance, it is possible to install it at ground level in a building under the antenna by using a system of microwave mirrors to guide the radio waves from the primary feed at ground level to the focus *S* of the reflector (Figure 8.10). This arrangement enables the high losses inherent in a coaxial cable or waveguide to be avoided while permitting rotation of the antenna about two orthogonal axes and allowing feed and RF equipment to remain fixed. Because of the high cost and the reduction in antenna diameter (the new Intelsat Standard A, for example) that makes installation of RF equipment at the focus of the antenna easier, this periscope mounting is less and less used.

A disadvantage of the Cassegrain mounting is the masking effect of the auxiliary reflector. The perturbation caused by the auxiliary mirror leads to a slight reduction of gain and the 3 dB beamwidth, a noticeable increase in the level of the first side lobe, and a modification of the level or a broadening of subsequent side lobes. These effects are negligible for a small d_H/D ratio (d_H is the diameter of the auxiliary reflector). For medium-sized antennas, they can be minimised by selecting dimensions such that:

$$f_s/f_d = d_s/d_H$$
$$d_s = (2f_d\lambda/\eta_s)^{1/2} \quad (m) \tag{8.19}$$

where the notation is that of Figure 8.9, and η_s is the efficiency of the primary feed.

The masking effect of the auxiliary reflector can be overcome by choosing an offset Cassegrain mounting.

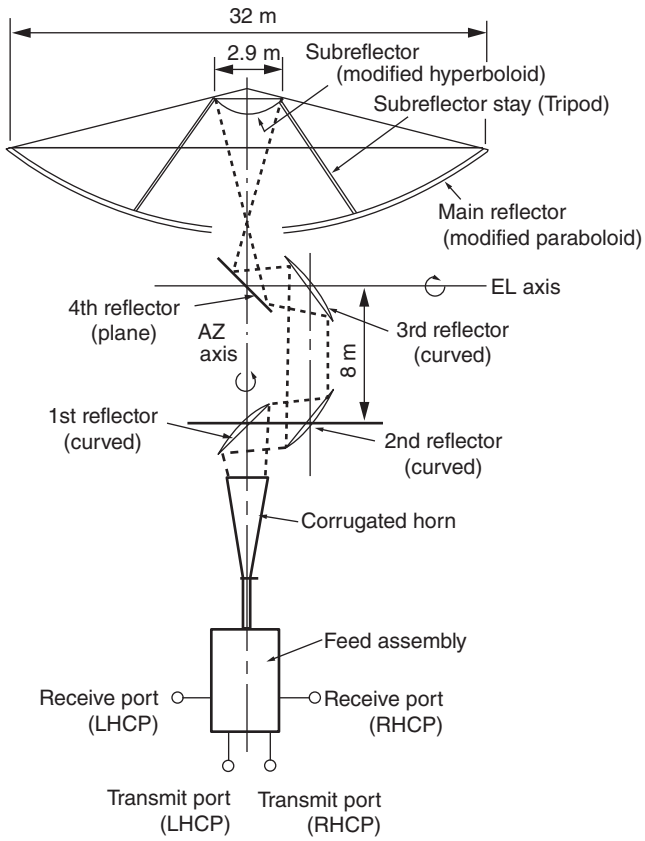


Figure 8.10 Guided-beam system.

8.3.4.4 Multibeam antennas

A small antenna suitable for the reception of several satellites grouped in one portion of the geostationary orbit has been developed by COMSAT Laboratories [KRE-80]. This multiple-beam torus antenna (MBTA) is equivalent, for each beam, to an antenna of 9.8 m aperture and has a gain on the order of 50 dB at 4–6 GHz with a noise temperature on the order of 30 K for an elevation angle of 20°.

Small multibeam antennas have been developed for direct-to-home (DTH) reception of television carriers from broadcasting geostationary satellites located at separate orbital positions. These antennas are equipped with two feeds, slightly offset from the focus, each defining a beam aiming at one of the two satellites. This allows the user to be able to receive the programmes from both satellites without having to change the antenna orientation.

8.3.5 Pointing angles of an earth station antenna

The elevation and azimuth angles that specify the direction of a satellite from a point on the earth's surface were defined in Chapter 2 as a function of the relative coordinates of the satellite

and the point concerned. This section is devoted to one application of the determination of these angles for the case of a geostationary satellite and a practical presentation in the form of a nomogram. The polarisation angle, which completes the characterisation of RF antenna pointing when plane polarised waves are used, is also defined.

8.3.5.1 Elevation and azimuth angles

The orientation of the axis of an antenna pointed towards a satellite is defined by two angles – the azimuth A and elevation angle *E*. These two angles are specified as a function of the latitude *l* and the relative longitude *L* of the station (*L* is the absolute value of the difference from the longitude of the earth station to that of the satellite).

The *azimuth angle* is the angle about a vertical axis through which the antenna must be turned, clockwise from the geographical north to bring the axis of the antenna into the vertical plane that contains the direction of the satellite. This plane passes through the centre of the earth, the station, and the satellite (Figure 8.11). The azimuth angle *A* has a value between 0° and 360°. Its value is obtained from Figure 8.12 by means of an intermediate parameter *a* determined from the family of curves and used to deduce the angle *A* using the table inserted into the figure. The curves result from the following relation, which could be used for greater accuracy:

$$a = \arctan(\tan L / \sin l) \tag{8.20}$$

The *elevation angle* E is the angle through which the antenna must be turned in the vertical plane containing the satellite to bring the boresight of the antenna from the horizontal to the direction of the satellite (Figure 8.11). The elevation angle *E* is obtained from the corresponding family of curves of Figure 8.12, which follow from the relation:

$$E = \arctan[(\cos\phi - R_E/(R_E + R_0))/(1 - \cos^2\phi)^{1/2}] \tag{8.21}$$

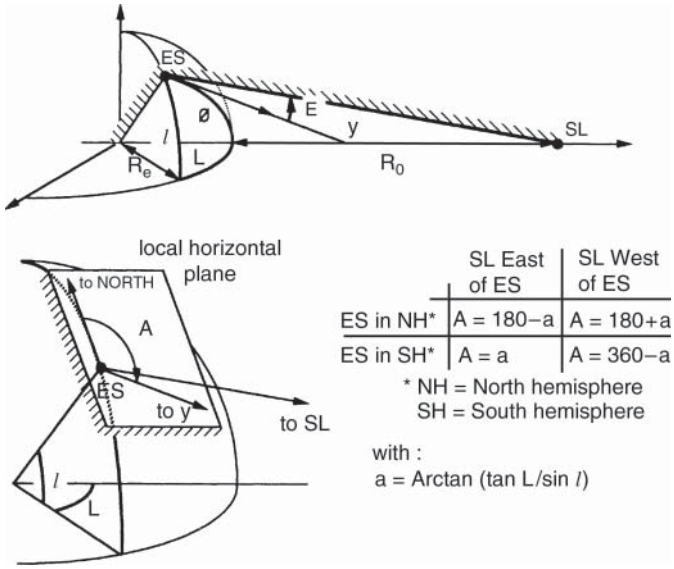
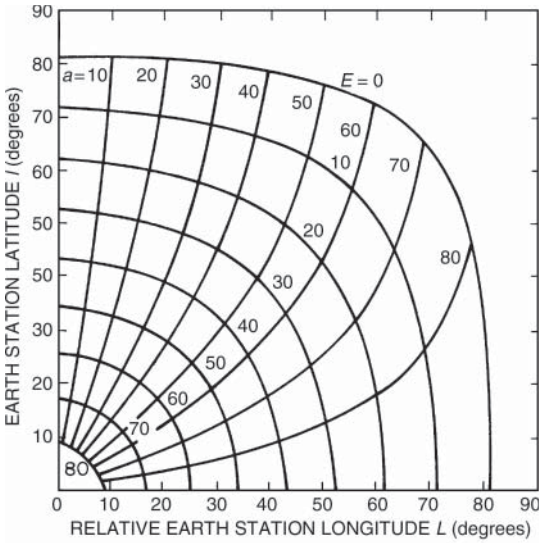


Figure 8.11 Azimuth and elevation angles.



	SL EAST OF ES	SL WEST OF ES
NORTH HEMISPHERE	$A = 180 - a$	$A = 180 + a$
SOUTH HEMISPHERE	$A = a$	$A = 360 - a$

Figure 8.12 Azimuth and elevation angles as a function of the earth station latitude and satellite relative longitude.

where:

$\cos \phi = \cos l \cos L$
R_E = radius of the earth = 6378 km,
R_0 = altitude of the satellite = 35 786 km.

8.3.5.2 Polarisation angle

Whenever the wave polarisation is linear, the earth station antenna feed must have its polarisation aligned with the polarisation plane of the received wave. This plane contains the electric field of the wave (see Section 5.2.3). The polarisation plane at the satellite contains the satellite antenna boresight and a reference direction. This reference direction is, for instance, the perpendicular to the equatorial plane for vertical (V or Y) polarisation or parallel to the plane of the equator for the horizontal (H or X) polarisation. The *polarisation angle* at the earth station is the angle ψ between the plane defined by the local vertical at the earth station and the antenna boresight, and the polarisation plane. $\psi = 0$ corresponds to the reception or emission at the earth station of a linearly polarised wave with its polarisation plane containing the local vertical. The polarisation angle at the earth station for a reference direction at the satellite in a plane perpendicular to the

equatorial plane is given by:

$$\cos\psi = \frac{\sin l\left(1 - \frac{R_E}{r}cos\phi\right)}{\sqrt{1-\cos^2\phi}\sqrt{1 - 2\frac{R_E}{r}\cos\phi + \left(\frac{R_E}{r}\right)^2\cos^2 l}} \tag{8.22a}$$

where:

r: distance from satellite to earth centre = $R_E + R_0$
R_E: earth radius = 6378 km
R_0: geostationary satellite altitude = 35 786 km
$\cos\phi = \cos l \cos L$

For a geostationary satellite, a simplified expression with an error less than 0.3° for ψ is obtained by considering the satellite at infinite distance r from the earth:

$$\cos\psi = \frac{\sin l}{\sqrt{1-\cos^2\phi}} \tag{8.22b}$$

or equivalently:

$$\tan\psi = \sin L / \tan l \tag{8.22c}$$

Values of ψ are displayed in Figure 8.13.

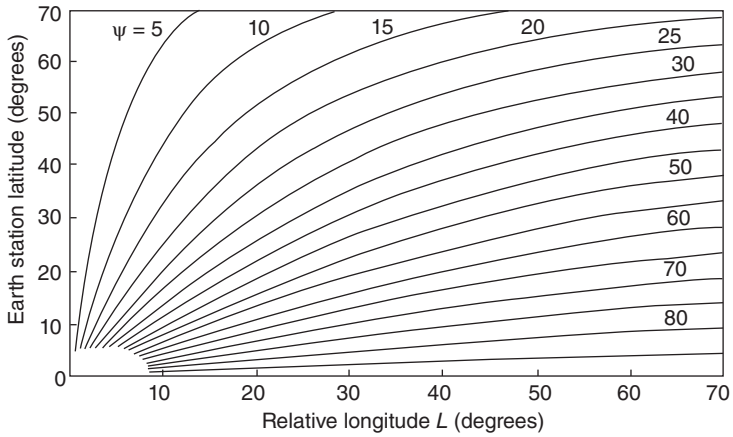


Figure 8.13 Polarisation angle ψ as a function of earth station latitude l and satellite relative longitude L, considering a reference polarisation plane for the satellite perpendicular to the equatorial plane.

8.3.6 Mountings to permit antenna pointing

For fixed stations that operate with a specific geostationary satellite, the range of angles through which the antenna is likely to be pointed is small. Its magnitude must, however, be sufficient to permit repointing to a standby satellite in case of breakdown of the first. In the more general case, it is desirable to provide equipment capable of providing beam pointing in any direction

in order to be able to establish links with different geostationary satellites or nongeostationary satellites.

Movement of the antenna results usually from movement about two axes – a primary axis that is fixed with respect to the earth and a secondary axis that rotates about the first.

8.3.6.1 Azimuth–elevation mounting

An azimuth–elevation mounting corresponds to a vertical fixed primary axis and a horizontal secondary axis (Figure 8.14) constrained to rotate about the vertical axis. Rotation of the antenna support about the vertical axis enables the azimuth angle A to be adjusted, and rotation of the antenna about the associated horizontal axis of the support then permits the elevation angle E to be adjusted. This is the mounting most commonly used for antennas of steerable earth stations.

With the previous mounting, the secondary axis may not be in the horizontal plane and hence may be at an angle other than 90° with the primary axis; this is non-orthogonal *azimuth–elevation* mounting. This mounting is useful for Cassegrain antennas since the volume within that the antenna operates for different pointing angles is reduced with respect to conventional azimuth–elevation mounting. On the other hand, coupling between the rotation about the axes is introduced, and angular displacements about the axes no longer correspond to the azimuth and elevation angles previously defined.

Azimuth–elevation mounting has the disadvantage of leading to high angular velocities when tracking a satellite passing through the vicinity of the zenith. The elevation angle then reaches 90°, which generally corresponds to a mechanical stop to prevent overtravel of the antenna about the secondary axis. To track the satellite, the antenna must thus perform a rapid rotation of 180° about the primary axis.

This constraint can be avoided by giving the pointing system an additional degree of freedom. This permits the introduction of a bias on the secondary axis support with respect to the vertical (Figure 8.15). This function may be realised, for example, by a relative rotation of two half cylinders initially with the same principal axis coincident with the primary axis and whose contact faces are not orthogonal to the axis. Once introduced, rotation of the antenna about the secondary axis for a given elevation pointing angle is thus equal to the elevation angle less the bias. Hence for pointing at the zenith ($E = 90°$), the maximum travel is not reached. This mounting is used, for example, for stations mounted on mobiles.

Another solution to avoid the problem near zenith and to give the pointing system an additional degree of freedom consists of adding a third axis of rotation orthogonal to the elevation axis (like the Y-axis in the X–Y mounting; see Section 8.3.6.2, the elevation axis being the X axis). This allows a few degrees angular shift of the boresight of the antenna with respect to the plane normal to the elevation axis. This mounting is called *three-axis mounting*. When tracking a satellite expected to pass above the station, as soon as the elevation becomes larger than a given value, the azimuth rotation is initiated and pointing at the satellite is maintained by means of an appropriate combination of elevation angle and angular shift about the third axis. This provides a long enough time to rotate azimuth by 180° without overstepping the maximum angular speed of the azimuth drive mechanism.

8.3.6.2 X–Y mounting

An *X–Y mounting* has a fixed horizontal primary axis and a dependent secondary axis that rotates about the primary axis and is orthogonal to it (Figure 8.16). This mounting does not have the disadvantage of the azimuth–elevation mounting when the satellite passes through the zenith

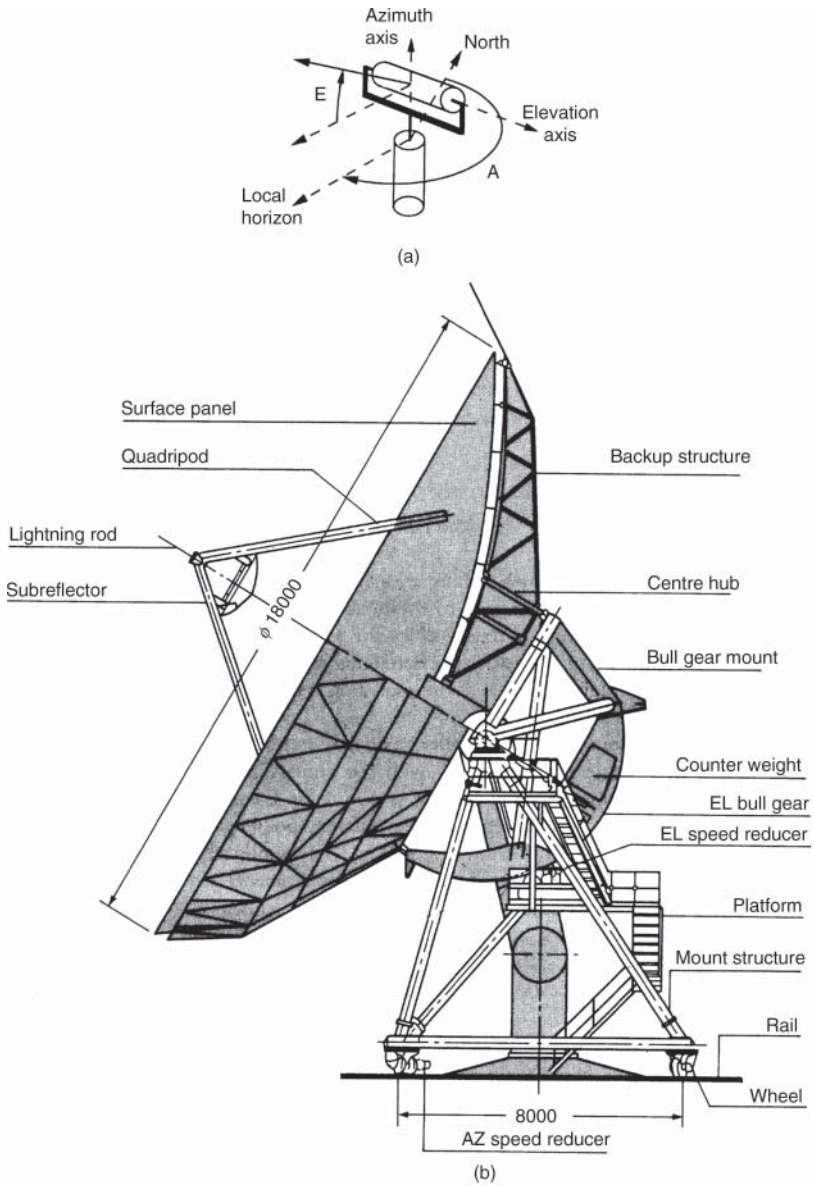


Figure 8.14 Azimuth–elevation mount. (a) Axes of rotation: antenna pointing in the direction of the satellite is obtained by rotation through an angle equal to the azimuth A about the vertical primary axis, then by rotation through an angle E (the elevation angle) about the horizontal secondary axis. (b) An example of implementation with a Standard C antenna.

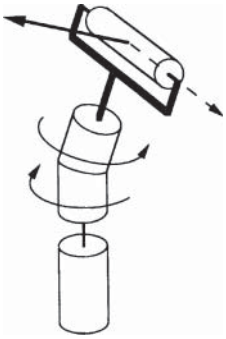


Figure 8.15 Modified azimuth–elevation mount: relative rotation of 180° of the two parts of the secondary axis support introduces an offset with respect to the vertical of the upper part of the primary axis.

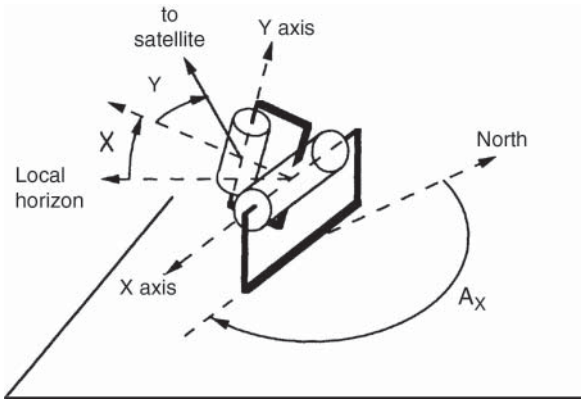


Figure 8.16 *X*–*Y* mount: Antenna pointing in the direction of the satellite is obtained by rotation through an angle *X* about the horizontal primary axis, then by rotation through an angle *Y* about the associated secondary axis of the part that rotates about the primary axis.

(a high speed of rotation about the primary axis). *X*–*Y* mounting is thus useful for satellites in low orbits rather than for geostationary satellites and stations mounted on mobiles.

For a station in the northern hemisphere, the pointing angles *X* (rotation about the primary axis from the local horizontal) and *Y* (rotation about the secondary axis from the plane perpendicular to the primary axis) are given as a function of the latitude l and the relative longitude L of the station by:

$$X = \arctan[(\tan E)/\sin A_R] \tag{8.23a}$$

$$Y = \arcsin[-\cos A_R \cos E] \tag{8.23b}$$

with A_R, the azimuth of the satellite relative to the primary axis of the mounting (*X* axis), such that $A_R = A - A_X$ and where:

- E is the elevation angle of the satellite obtained from Eq. (8.21).
- A is the azimuth of the satellite obtained from Eq. (8.20) and the table in Figure 8.12.
- A_X is the orientation of the *X* axis with respect to the north.

The angles and their projection on the horizontal plane are taken as positive in the inverse trigonometric direction.

8.3.6.3 Polar mounting

A *polar* or *equatorial* mounting corresponds to a primary axis (the *hour axis*) parallel to the axis of rotation of the earth and a secondary axis (the *declination axis*) perpendicular to the former (Figure 8.17). This mounting is used for telescopes since it permits tracking of the apparent movement of stars by rotation only about the hour axis, which thus compensates for the rotation of the earth about its line of poles.

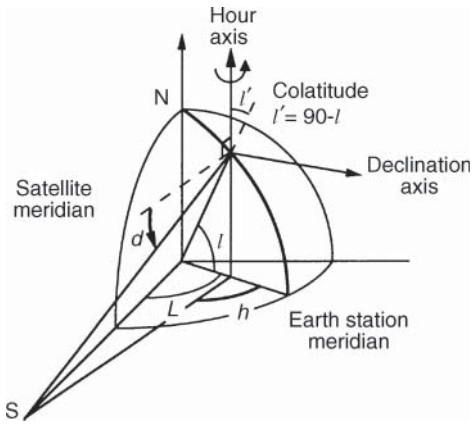


Figure 8.17 Polar mount: definition of the hour angle h (the angle with apex at the projection of the station in the equatorial plane between the station meridian plane and the satellite) and declination d (the angle between the parallel to the equatorial plane and the satellite in the plane perpendicular to the equator, which contains the station and the satellite).

This mounting is useful for links with geostationary satellites since it is possible to point the antenna at several satellites successively by rotation about the hour axis. However, the fact that the satellites are not at infinity necessitates, in principle, slight adjustments of orientation about the declination axis.

The expressions for the *hour angle h* (the rotation about the hour axis from the earth station meridian plane to the plane containing the satellite) and the *declination d* (the rotation in the plane defined by the hour axis and the satellite from the perpendicular to the hour axis to the direction of the satellite) as functions of the latitude l and the relative longitude L of the station are:

$$h = \tan^{-1}[\sin L/(\cos L - 0.15126 \cos l)] \tag{8.24a}$$

$$d = \tan^{-1}[-0.15126 \sin l \sin h/\sin L] \tag{8.24b}$$

The hour angle is positive towards the east if the relative longitude L is defined as the algebraic difference between the longitude of the satellite (degrees east) and the longitude of the station (degrees east). The value 0.151 26 corresponds to the ratio $R_E/(R_0 + R_E)$ for the nominal values of the terrestrial radius ($R_E = 6378$ km) and the nominal altitude of the geostationary satellite.

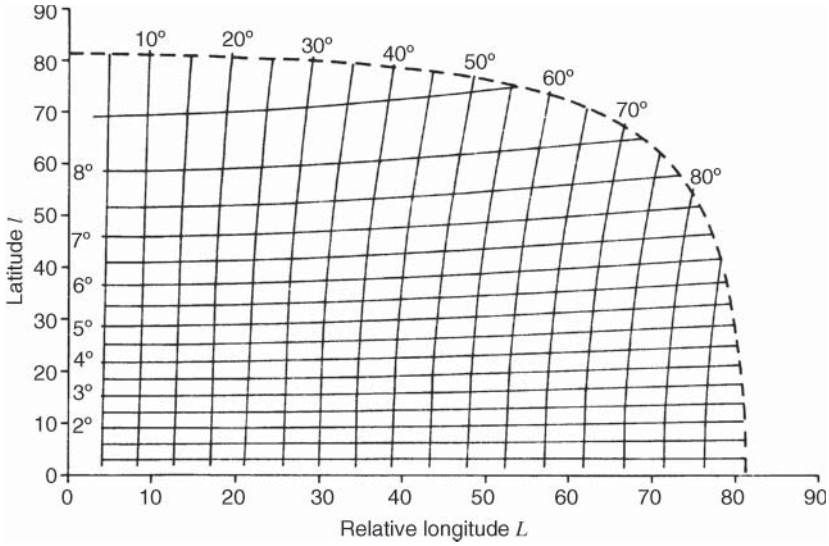


Figure 8.18 Hour angle h and declination d as a function of the latitude l of the station and its relative longitude L with respect to the satellite.

For $L = 0$, the hour angle is zero and Eq. (8.24b) is not defined. Direct determination of the declination angle leads to:

$$d_{L=0} = \tan^{-1}[-l/(6.61078) - \cos l] \tag{8.24c}$$

where the coefficient 6.610 78 corresponds to the nominal value of the ratio $(R_0 + R_E)/R_E$.

The curves in Figure 8.18 provide the values of the hour and declination angles. In practice, the orientation of the hour axis is obtained by inclining the axis with respect to the vertical in the plane of the local meridian by a value equal to 90° minus the latitude of the station (the colatitude).

For antennas of small diameter having a sufficiently large main lobe, it is possible to point at several satellites by rotation only about the hour axis. This approach is often adopted for semi-professional antennas intended for the reception of television signals transmitted by various satellites.

As the declination remains fixed, a pointing error occurs, which depends on the latitude of the station and the angular separation between the satellites. Figure 8.19 gives the declination error as a function of the latitude l of the station and the relative longitude L of the satellite with respect to the station (the nominal value of the declination is that of a satellite situated at the longitude of the station: that is, $d_{L=0}$). It should be noted that it is at medium latitudes (40°) that this error is greatest.

A reduced pointing error is obtained by fixing the declination angle at a value that results from a compromise between the values corresponding to the various satellites considered. For example, for a station of latitude 40°, a sweep of 50° of the geostationary arc on both sides of the station meridian leads to a declination error of 0.3° at the limit of the sweep if the initial declination is adjusted for a satellite in the meridian plane of the station. Through reducing the initial declination by 0.3°/2° = 0.15°, to that corresponding to correct pointing if the satellite were in the plane of the meridian: that is, $d_{L=0}(l = 40°) - 0.15°$, the maximum pointing error is equal to 0.15° when at the extreme ends of the sweep and in the plane of the meridian, and less than 0.15° when operating elsewhere.

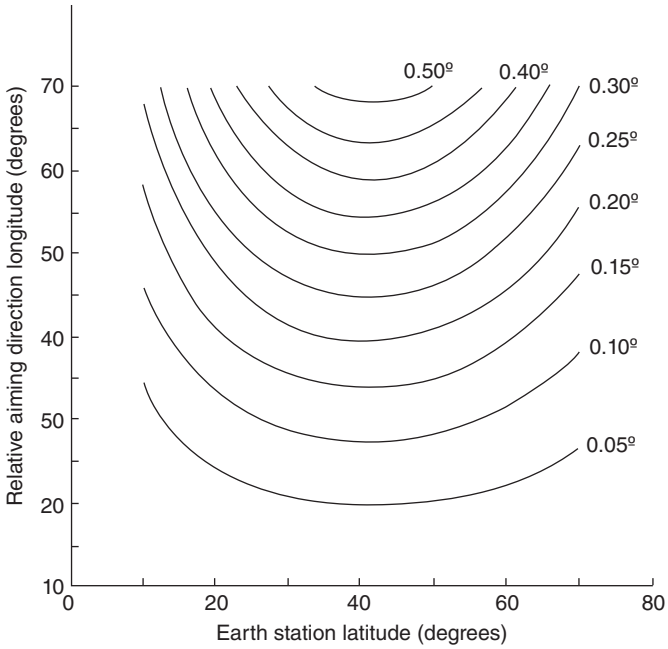


Figure 8.19 Pointing error with respect to the orbit of geostationary satellites as a function of the earth station latitude and the relative longitude of the aiming direction.

The compromise can be further improved by introducing a bias on the orientation of the hour axis. The hour axis is no longer parallel to the polar axis but is inclined towards the exterior in the plane of the meridian in such a way as to cancel the declination error when operating in the plane of the meridian while minimising it over the swept portion of the orbit (modified polar mounting). The bias to be introduced depends on the latitude of the station and the required magnitude of the sweep in longitude. Figure 8.20 gives the value of bias and declination adjustment (rotation with respect to the perpendicular to the mounting hour axis) as a function of the station latitude for an orbit sweep of ±45° on each side of the meridian. The resultant pointing error is less than 0.1°.

These values are obtained by taking account of the oblateness of the earth, which leads to the replacement of the latitude l in Eq. (8.24c) by l' (the geocentric latitude) and $R_E/(R_0 + R_E)$ by $R_E(1 - A \sin^2 l')/(R_0 + R_E)$, where A is the oblateness coefficient equal to $1/298.257$ or 3.352×10^{-3} (see Section 2.1.5.1).

8.3.6.4 Tripod mounting

Tripod mounting is well suited to geostationary satellites. The antenna is fixed to the support by means of three legs of which two are of variable length. According to the mounting used, pointing in elevation and azimuth may or may not be independent. Mounting is simple but the magnitude of pointing variation is limited (for example, 10° about a mean direction).

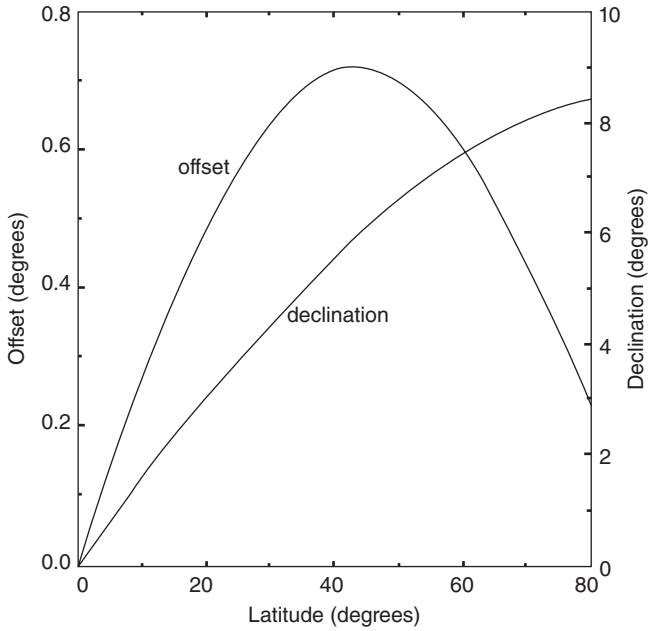


Figure 8.20 Modified polar mount: hour angle offset and declination angle as a function of the latitude of the earth station.

8.3.7 Tracking

Tracking consists of maintaining the axis of the antenna beam in the direction of the satellite in spite of movement of the satellite or station. Several types of tracking are possible and are characterised by their tracking error (pointing angle error). Choice of the type of tracking depends on the antenna beamwidth and the magnitude of apparent movement of the satellite.

8.3.7.1 The influence of antenna characteristics

The angular width of the beam directly affects selection of the type of tracking. It should be noted that, at the frequencies used, the 3 dB angular half-power beamwidth $\theta_{3\,\mathrm{dB}}$ can be small. By way of example, Figure 8.21 gives the 3 dB angular half-power beamwidth $\theta_{3\,\mathrm{dB}}$ for different frequencies as a function of antenna diameter.

The depointing loss L for a depointing angle θ with respect to the direction of maximum gain is given by (see Eqs. (5.5) and (5.18)):

$$L = \Delta G = 12(\theta/\theta_{3\mathrm{dB}})^2 \quad (\mathrm{dB}) \tag{8.25}$$

Depointing is associated with relative movement of the satellite and the direction of maximum gain of the antenna. Decisions relating to antenna installation and tracking procedure depend on the beamwidth in relation to the magnitude of apparent movement of the satellite; the determining criterion is the variation of antenna gain with depointing.

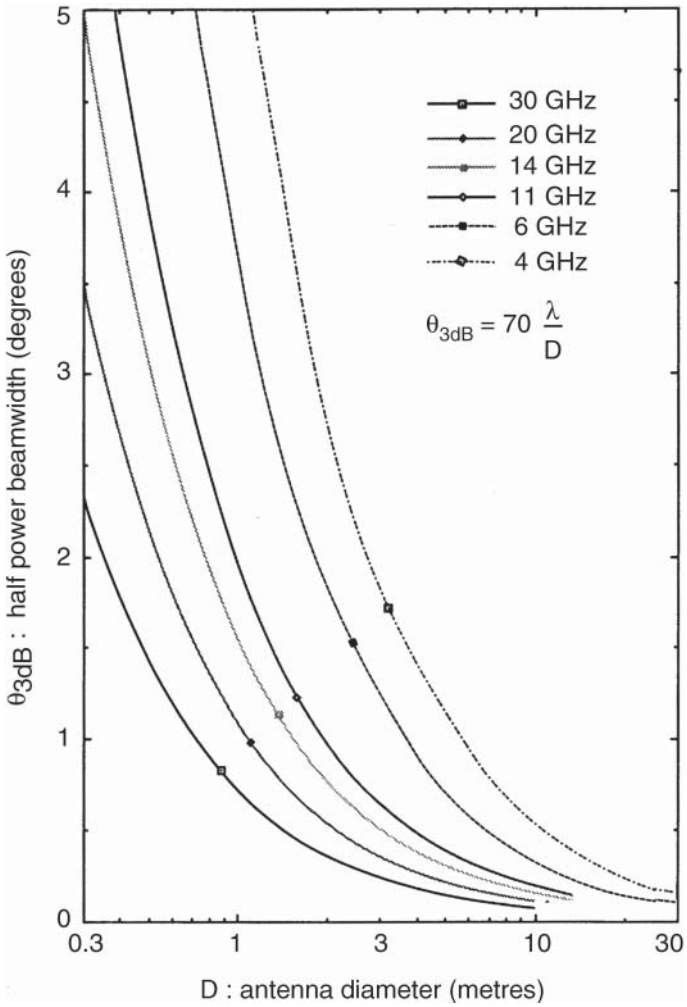


Figure 8.21 Half-power beamwidth $\vartheta_{3\,\mathrm{dB}}$ versus antenna diameter D.

Another antenna characteristic that is associated with its diameter and directly affects the performance of orientating devices is its mass. For small antennas, the mass of the parabolic reflector ranges from a few tens to several hundreds of kilograms. For large antennas, it is several tons; the moving part of the 32.5 m diameter antenna of Pleumeur Bodou IV, France, for example, weighs 185 tons. Meteorological conditions (wind speed) and the mass of the antenna itself cause deformations that vary with elevation angle.

8.3.7.2 Apparent movement of the satellite

The apparent movement of the satellite was examined in Chapter 2 as a function of the type of orbit.

For satellites on inclined elliptical orbits, movement causes a variation of elevation angle whose magnitude about the zenith varies according to the type of orbit, the number of satellites in the system, and the location of the service area.

For geostationary satellites, the apparent movement is contained within the *station-keeping box* (Section 2.3.4.3) whose dimensions impose the accuracy of station keeping (±0.1° north–south and east–west, for example). The actual movement within the window is a combination of north–south movement with a period of 24 hours due to nonzero orbit inclination (in the form of a figure of eight for a large inclination), east–west movement of the same period due to eccentricity, and a drift towards the east or west whose value and direction depend on the longitude of the satellite. The apparent speed of movement of the satellite does not exceed 2°/h; for an orbit inclination of 5°, the maximum angular velocity is 7×10^{-4}°/s (2.5°/h).

At a given time, the satellite can be anywhere within the station-keeping box. Only the satellite control station knows the position of the satellite as a function of time with some uncertainty. Distributed orbital data tables permit the times when the satellite is close to the centre of the window to be predicted (in some cases, this permits the maximum depointing of the earth station antenna to be minimised).

8.3.7.3 Fixed antenna without tracking

Tracking is not necessary when the antenna beamwidth is large in comparison with the station-keeping box of a geostationary satellite or for the case of a system of satellites on inclined elliptical orbits when the antenna beamwidth greatly exceeds the solid angle that contains the apparent movement of the active orbiting satellite.

The usable part of the beam can be defined at −0.1, −0.5, −1, or $-n$ dB in accordance with the acceptable loss of gain; the choice results from optimisation of the characteristics of the links between the satellite and the stations.

In the case of pointing towards a geostationary satellite, the maximum depointing angle that determines the dimensions of the system can be minimised, for a given window size and a given $\theta_{3\,\mathrm{dB}}$ (or λ/D ratio), by performing initial pointing when the satellite is closest to the centre of the box. Coarse orientation of the antenna is achieved from pointing angles determined with the help of Figure 8.12 or the expressions given in Section 8.3.5. Fine pointing is then obtained by searching for the maximum beacon signal level from the satellite concerned by displacing the pointing direction on each side of the axis, which is assumed to correspond to maximum gain (locating the angular values corresponding to a given decrease of level and pointing in the corresponding direction of the centre point). As a result of small variations of gain in the vicinity of the electromagnetic axis, the *initial pointing error* (θ_{IPE}) is on the order of $0.1\text{–}0.2\theta_{3\,\mathrm{dB}}$.

Considering that initial pointing is achieved when the satellite is assumed to be near the centre of the station-keeping box of half-width SKW, and designating as SPO the offset angle of the satellite with respect to the centre of the box, the *maximum value* of *depointing angle* θ_{MAX} is determined with the help of Figure 8.22:

$$\theta_{\mathrm{MAX}} = \mathrm{SKW}\sqrt{2} + \mathrm{SPO} + \theta_{\mathrm{IPE}} \tag{8.26}$$

where θ_{IPE} of the form $b\theta_{3\,\mathrm{dB}}$ is the *initial pointing error*.

SPO is either the uncertainty in the satellite position determination should one orient the antenna at the time the satellite is expected to pass at the centre of the station-keeping box, or the actual angular shift in the position of the satellite with respect to the centre of the box at the time the antenna is oriented.

θ_{MAX} is thus of the form $a + b\theta_{3\,\mathrm{db}}$ (a = constant = SKW $\sqrt{2}$ + SPO).

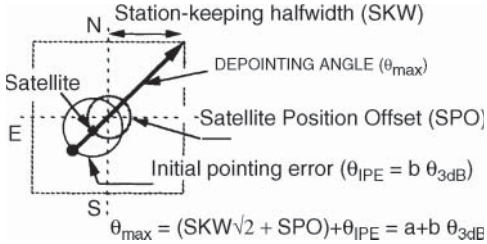


Figure 8.22 Maximum depointing angle with a fixed mount antenna.

8.3.7.4 Programmed tracking

With *programmed tracking*, antenna pointing is achieved by providing the antenna orientation control system with the corresponding values of azimuth and elevation angles at each instant. These azimuth and elevation angles are calculated in advance for successive instants, taking account of the predicted apparent movement of the satellite, and the values are stored in a memory. Pointing is then performed in open loop without determination of the pointing error between the actual direction of the satellite and the aiming direction at each instant.

The pointing error thus depends on the accuracy of knowledge of the apparent motion of the satellite from which the various pointing angles are calculated and the accuracy with which pointing in a given direction is achieved by the antenna. (Errors arise from inaccuracies in local references and the coding of the antenna orientation, feedback control errors, and so on.)

Programmed tracking is mainly used for earth station antennas of large λ/D ratio, which thus have a large enough beamwidth that high pointing accuracy is not required. If high pointing accuracy is required (a small λ/D ratio), programmed tracking is used with non-geostationary satellites to preposition the antenna in an area of the sky where the satellite will appear in such a way as to ensure acquisition by a closed-loop tracking system operating on the satellite beacon.

With geostationary satellites, programmed tracking is used for stations with mid-range λ/D values (at Ku band, typically 4 m diameter).

8.3.7.5 Computed tracking

This system is a variant of programmed tracking and is well suited to tracking geostationary satellites for antennas having an intermediate value of λ/D, *which* does not justify the use of closed-loop beacon tracking.

With *computed tracking*, a computer incorporated in the pointing system evaluates the antenna orientation control parameters. The computer uses the orbit parameters (inclination, semi-major axis, eccentricity, right ascension of the ascendant node, argument of the perigee, anomaly) and, if necessary, a model of their progression. The data in memory are, if necessary, refreshed periodically (after a few days). The system can also extrapolate the progression of the orbit parameters from daily satellite displacements that are stored in memory.

8.3.7.6 Closed-loop automatic tracking

With antennas having a small value of λ/D, and hence a small angular antenna beamwidth with respect to the apparent magnitude of satellite movement, precise tracking of the satellite

is obtained by continuously aligning the antenna direction to that of a beacon located on the satellite.

The accuracy depends on the method used to determine the direction of arrival of the beacon signal, the deviation between the direction of arrival and the actual direction of the satellite (caused by propagation aberrations), and the precision of the feedback control system.

In addition to an accuracy, which can be very high (the tracking error can be less than 0.005° with a monopulse system), an advantage of this procedure is its autonomy since tracking information does not come from the ground. Moreover, it is the only conceivable system for mobile stations whose antenna movement cannot be known a priori (if the itinerary of the mobile is known, programmed tracking could conceivably be used).

Two techniques are used for beacon tracking – tracking by sequential amplitude detection and monopulse tracking [HAW-88].

8.3.7.6.1 Sequential amplitude detection

Sequential amplitude detection tracking systems make use of variations in received signal level as a consequence of commanded displacement of the antenna pointing axis. The level variations generated in this way enable the direction of maximum gain, which corresponds to the highest received signal level, to be determined. The main source of error arises from the fact that the system is unable to distinguish a level variation due to antenna depointing from a level variation caused by a change of wave-propagation conditions. Various procedures are used: conical scanning, step-by-step tracking, and electronic tracking.

Conical scanning. The antenna beam rotates continuously about an axis, which makes a given angle (small compared with $\theta_{3\,\mathrm{dB}}/2$) with respect to the axis of maximum gain. When the direction of the satellite differs from the direction of the axis of rotation, the received level is modulated at the rate of rotation of the antenna as a function of the angular deviation between the two directions. The tracking receiver correlates this modulation with the antenna rotation to generate orientation control signals. The modulation of the received signal becomes zero when the direction of the satellite coincides with the axis of rotation. The (automatic) *tracking error* θ_{ATE} is between $0.2\ \theta_{3\,\mathrm{dB}}$ and $0.05\ \theta_{3\,\mathrm{dB}}$.

This technique, which has been used for a long time, particularly for small antennas, has been progressively abandoned in favour of step-by-step tracking, which enables a tracking accuracy of the same order of magnitude to be obtained with less mechanical complexity. Furthermore, the modulation of the uplink signal at the rate of antenna rotation can be a source of perturbations.

Step-by-step tracking. Antenna pointing is achieved by searching for the maximum received beacon signal. This proceeds by successive displacements (steps) of the antenna about each of the axes of rotation (the method is also known as *step-track* or *hill-climbing*). The direction of the subsequent displacement is determined by comparing the received signal level before and after the step. If the signal increases, the displacement is made in the same direction. If the signal decreases, the direction of displacement is reversed. The procedure is performed alternately on each of the two axes of rotation of the antenna [TOM-70].

There are several limits to the accuracy of tracking, as follows:

- The uncertainty in the direction of the maximum may be greater than the step size, which must therefore be chosen to be sufficiently small (on the order of $\theta_{3\,\mathrm{dB}}/10$) [RIC-86].
- The gain of the antenna (and hence the level of the received signal) about the direction of maximum gain varies slowly with depointing angle (the lobe has a flat top). Determination

of the direction of maximum gain is thus less precise than determination of the pronounced null of the gain characteristic of monopulse systems (see Section 8.3.7.6.2). The accuracy with which the direction of maximum gain is determined is fundamentally a function of the $\theta_{3\,\mathrm{dB}}$ beamwidth and hence λ/D.

- The system has a limited dynamic response, and one must wait for the reflector displacement before detecting each variation of the received signal.
- Finally, as with all systems where depointing information is obtained from variations of received signal level, step-by-step tracking is affected by spurious amplitude modulation of the signal level. Furthermore, it is necessary to have a sufficient C/N (carrier power-to-noise power) ratio at the input of the tracking receiver (typically 30 dB).

Taking these limitations into account, the (automatic) *tracking error* θ_{ATE} is between $0.05\,\theta_{3\,\mathrm{dB}}$ and $0.15\,\theta_{3\,\mathrm{dB}}$.

For antennas used with geostationary satellites, the system can be used either in continuous tracking mode or in *point-and-rest* mode where, after pointing the antenna in the direction of the satellite, the positions of the axes of rotation are clamped until pointing is performed again. This repointing is activated either periodically at regular intervals or on detection of a reduction of received signal level. This point-and-rest mode reduces the operations and stresses suffered by the pointing servos and the orientation motors, which permits an increased lifetime of the hardware and reduced maintenance.

Smoothed step-track. The performance of a step-by-step tracking system can be improved by reducing the sensitivity to amplitude fluctuations of the received signal. One solution consists of combining computed tracking with a step-by-step system. An estimate of the pointing direction is thus obtained from a simplified model of the apparent motion of the satellite. The step-by-step tracking system then improves the pointing with respect to the direction obtained. The direction of the satellite is then known, and this enables the model of apparent motion to be updated [EDW-83].

As the system has an estimate of the satellite direction at each instant, possible errors caused by fluctuations of signal amplitude can be detected, and this permits incorrect commands to be cancelled. Furthermore, step-by-step tracking need not be continuously activated; its role is to update the motion model periodically, and this updating enables good accuracy of computed tracking to be obtained, and augments the lifetime of the mechanical devices.

Electronic tracking. This recent technique is comparable with step-by-step tracking. The difference lies in the technique used for successive displacement of the beam in the four cardinal directions, since this is realised electronically. Depointing by a given angle is obtained by varying the impedance of four microwave devices coupled to the source waveguide; these devices are located symmetrically on each side of the waveguide in two perpendicular planes [WAT-86].

Successive deviation of the beam in four directions enables the magnitude and direction of depointing to be evaluated if the received signal does not arrive along the boresight of the antenna. The signal is actually received with different levels in accordance with the direction of deviation of the beam. An error signal is derived from a combination of successive signals, and this enables the antenna orientation to be controlled in such a way as to reduce depointing.

The system thus uses a tracking receiver with a single channel as for a step-by-step system. Furthermore, determination of pointing error is achieved without mechanical displacement of the antenna and stresses on the antenna orientation mechanism are reduced. Finally, the tracking accuracy obtained is greater as a consequence of the virtual simultaneity (on the timescale of the apparent movement of the satellite) of level measurements in each of the directions of deviation.

This results in a rapid dynamic response. The (automatic) *tracking error* θ_{ATE} can be as small as $0.01\ \theta_{3\,dB}$ [DAN-85].

8.3.7.6.2 The monopulse technique

Excitation of an antenna pattern that is specifically intended for tracking and contains a zero on the axis permits the antenna to be orientated in such a way as to cancel the received signal. The orientation command signals are generated by comparison, in a *monopulse tracking receiver*, of a reference signal and the error angle measurement signals. The reference signal is the beacon signal (the 'sum' channel Σ) extracted from the system that generates the error signals in the antenna (see Figure 8.1). The error angle measurement signals result from depointing as measured in two orthogonal planes (the 'difference' channels Δ). In this way, error signals (Δ/Σ) are available that are independent of the received signal level.

The error angle measurement signals are provided either by comparison of the waves received from four sources located around the electromagnetic axis of the antenna (multiple source monopulse) or by detection of the higher-order modes generated by depointing of the antenna in the waveguide coupled to the primary source (mode extraction).

Multiple source monopulse. Each source in a multiple-source monopulse system has a radiation pattern that is slightly shifted with respect to the principal axis of the antenna (Figure 8.23). In each of the two orthogonal planes, the difference between the two signals received from the two sources becomes zero for a wave that arrives parallel to the principal axis. This difference is, at a first approximation (in the vicinity of correct pointing), proportional to the depointing angle. Insensitivity to variations in incident signal level is ensured by normalising the difference signals with respect to the sum signal. *Tracking error* can be as small as 0.01° for an antenna having a 3 dB beamwidth of 2° (a tracking error $\theta_{ATE} = 5 \times 10^{-3}\theta_{3\,dB}$). This type of system, which was used on the first large earth stations (Pleumeur Bodou 2, for example), has been abandoned in favour of mode-extraction systems, which are easier to implement.

Mode-extraction monopulse. The mode-extraction system (Figure 8.24) uses the special propagation properties of modes of orders greater than the fundamental TE_{11} in a waveguide (transverse electric [TE] signifies a wave whose magnetic field component H along the direction of propagation is zero). For a circularly or linearly polarised wave, only the TE_{11} mode propagates if the

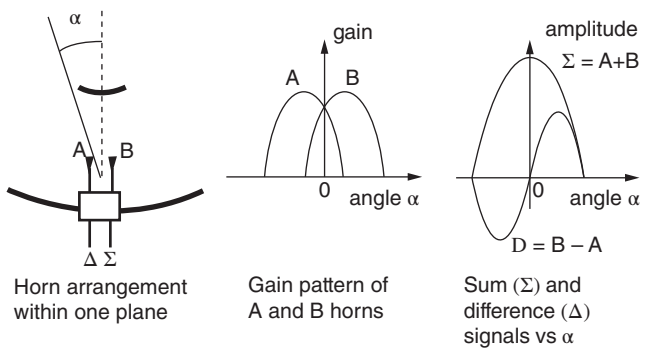


Figure 8.23 Multihorn monopulse tracking system.

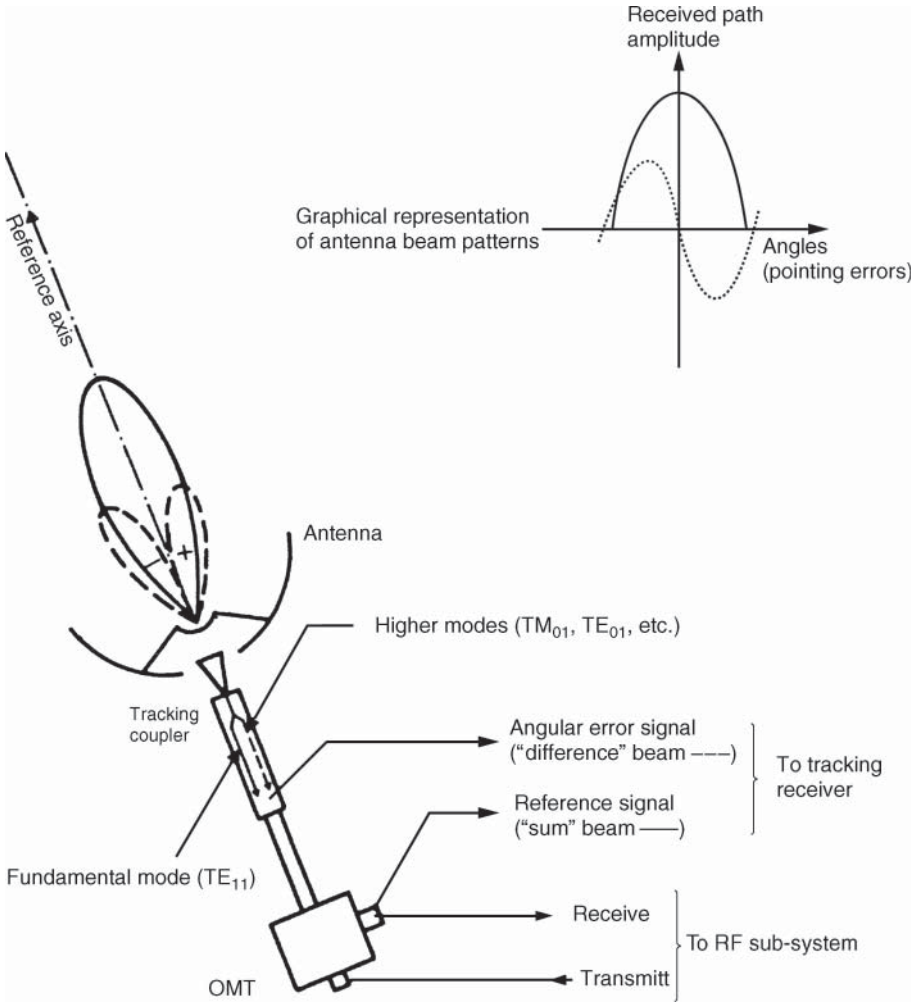


Figure 8.24 Multimode extraction monopulse tracking system. Source: reprinted from CCIR-88 with the permission of the ITU.

incident wave arrives along the axis of the guide; if the wave arrives with an angular deviation with respect to the axis of the guide, TM_{01} and TE_{21} modes are generated, and the amplitude of the TE_{11} mode decreases slightly. The TM_{01} and TE_{21} modes are odd functions of depointing and are orthogonal. With a linearly polarised wave, use of the TM_{01} and TE_{21} modes permits depointing to be determined in the two reference planes. With a circular polarised wave, depointing is determined in each plane by the phase shift between one of the TM_{01} or TE_{21} modes (only one of the two is used) and the fundamental TE_{11} mode.

Monopulse tracking systems are characterised by excellent performance in terms of tracking accuracy and speed of response. *Tracking error* θ_{ATE} is on the order of $0.02\ \theta_{3\,dB}$–$0.05\ \theta_{3\,dB}$. In contrast, these systems are costly since they require coherent tracking receivers with several channels and sources, which are difficult to manufacture. They are principally used in stations of the former Intelsat Standard A type (30 m diameter in C band). For antennas of smaller diameter (for

example, the new Standard A of 16 m diameter), step-by-step systems are economically more attractive, and the performance degradation is scarcely noticeable.

8.3.7.7 The influence of tracking type on antenna gain

Table 8.10 summarises the various types of tracking technique and related tracking errors. According to the type of tracking used, the maximum depointing angle θ_{MAX} has a particular value that may be independent of λ/D (for the case of programmed or computed tracking), the sum of a constant term, and a term that is a function of λ/D (for fixed mounting) or a function of λ/D (for most cases of closed-loop automatic beacon tracking).

Table 8.10 Tracking system performance

Tracking type	Tracking error	Gain loss
None (fixed mounting)	Initial pointing error: $\theta_{IPE} = 0.1$ $\theta_{3\,dB}$ to 0.2 $\theta_{3\,dB}$	A function of the station-keeping box
Programmed or computed	Typical: 0.01°	A function of D/λ
Conical scanning	0.05 $\theta_{3\,dB}$ to 0.2 $\theta_{3\,dB}$ (typical: 0.01°)	$DG = 0.03$–0.5 dB
Step-by-step	0.05 $\theta_{3\,dB}$ to 0.15 $\theta_{3\,dB}$ (typical: 0.01°)	DG = 0.03–0.3 dB
Electronic deviation	0.01 $\theta_{3\,dB}$ to 0.05 $\theta_{3\,dB}$ (typical: 0.005°)	DG ≤ 0.03 dB
Monopulse	0.02 $\theta_{3\,dB}$ to 0.05 $\theta_{3\,dB}$ (typical: 0.005°)	DG ≤ 0.03 dB

The gain fallout with respect to the maximum gain can be evaluated for the various cases from Eq. (8.25) as a function of the maximum value θ_{MAX} of the depointing angle:

$$\Delta G = 12(\theta_{MAX}/\theta_{3\,dB})^2 \quad \text{(dB)}$$

The minimum gain of the antenna for the maximum depointing condition θ_{MAX} as a function of λ/D and the efficiency η is thus of the form:

$$G_{MIN} = \eta(\pi D/\lambda)^2 10^{-1.2[\theta_{MAX} D/70\lambda]^2} \tag{8.27}$$

8.3.7.7.1 Fixed mounting

For a fixed mounting used with a geostationary satellite, the maximum value θ_{MAX} of the depointing angle is given by Eq. (8.26) and is of the form $a + b\ \theta_{3\,dB}$ (a is the sum of the semi-diagonal SKW $\sqrt{2}$ of the station-keeping box and the offset SPO in the satellite position with respect to the centre of the box; $b\ \theta_{3\,dB}$ is the initial pointing error θ_{IPE}).

The gain fallout is thus:

$$\Delta G = 12(b + a/\theta_{3\,dB})^2 \quad \text{(dB)}$$

This gain fallout is not the same on the up- and downlinks since the 3 dB beam-widths ($\theta_{3\,dB}$) of the antenna beams are different.

The expression for the gain G_{MIN} of the antenna for conditions of maximum depointing θ_{MAX} as a function of λ/D, efficiency η and the parameters a and b defined above can be put in the form:

$$G_{MIN} = \eta(\pi D/\lambda)^2 10^{-1.2[b+(aD/70\lambda)]^2} \tag{8.28}$$

With fixed mounting and a satellite whose apparent movement is relatively large (as for inclined elliptical orbits), the minimum gain of the antenna is given by Eq. (8.27) after determination of the maximum depointing angle between the fixed antenna aiming direction and the direction corresponding to the entry into (or departure from) service of the active satellite.

8.3.7.7.2 Programmed tracking

For programmed or computed tracking, the loss of gain depends on the beamwidth $\theta_{3\,dB}$: that is, $70\lambda/D$ and the maximum value θ_{MAX} of the depointing angle:

$$\Delta G = 12(\theta_{MAX}/\theta_{3\,dB})^2 \text{ (dB)}$$

The minimum gain of the antenna for maximum depointing conditions θ_{MAX} is given by Eq. (8.27).

8.3.7.7.3 Automatic tracking

Finally, with automatic beacon tracking, the tracking error is very often a function of the 3 dB beamwidth. The depointing angle θ_{MAX} is thus of the form $\theta_{MAX} = c\theta_{3\,dB}$ and the corresponding loss of gain is given by:

$$\Delta G = 12(c)^2 \text{ (dB)}$$

The loss of gain is constant and independent of frequency and antenna efficiency. It is therefore the same on the up- and downlinks. Hence, for a step-by-step tracking system where the tracking error is between 0.05 and 0.15 $\theta_{3\,dB}(0.05 \le c \le 0.15)$, for example, the loss of gain is between 0.03 and 0.3 dB.

The expression for the antenna gain G_{MIN} for a tracking error of the form $c\theta_{3\,dB}$ as a function of λ/D and efficiency η is thus:

$$G_{MIN} = \eta(\pi D/\lambda)^2 10^{-1.2[c]^2} \text{ (dBi)} \tag{8.29}$$

8.3.7.7.4 Conclusion

To summarise the influence of depointing error on gain, it must be emphasised that an increase in antenna diameter does not necessarily lead to an increase in antenna gain in the direction of the satellite. In fact, for a station operating with a fixed depointing angle u_{MAX} that is not directly proportional to λ/D (the case of programmed or computed tracking or fixed mounting), an increase in λ/D leads to an increase in on-axis gain. On the other hand, depending on the value of a, an increase in λ/D does not always lead to an increase in gain in a direction that makes an angle u with the direction of maximum gain (Figure 8.25). This is due to the reduction of angular beamwidth as λ/D increases.

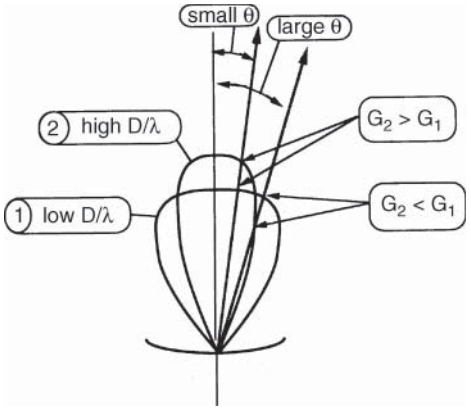


Figure 8.25 Gain variations versus D/λ as a function of the depointing angle ϑ.

As far as tracking of geostationary satellites is concerned, the typical size of the station-keeping box (±0.05°) makes use of a tracking system unnecessary (i.e. fixed mounting) up to diameters around 4 m for antennas operating in Ku band (14/11 GHz). Between 4 and 6 m, the choice between fixed mounting and a system of computed or step-by-step tracking is determined by a more detailed analysis. Above 6 m diameter, the use of a tracking system becomes mandatory. The use of step-by-step tracking systems, possibly associated with computed tracking, is tending to become universal for both medium and large stations as the result of the trade-off between cost and performance.

8.3.7.8 Antennas mounted on mobiles

With a highly directional antenna, automatic tracking can only be by closed-loop tracking of a beacon mounted on the satellite.

The difficulties of acquiring and maintaining locking of the servo loop may also require the use of an inertially stabilised platform, particularly for antennas mounted on board ships. Movement of the antenna with respect to the vessel is determined from information provided by the tracking receiver and the inertial platform.

Orientation of the beam can also be achieved with electronically controlled antennas. It may be worth controlling antenna orientation in only one axis, with the other retaining a fixed direction. This could be considered on aircraft using an array antenna that is electronically pointed in azimuth and mounted on the fuselage of an aircraft.

Finally, particularly for terrestrial mobiles, the use of fixed-zenith pointing antennas that have a sufficiently large 3 dB beamwidth (to the detriment of the gain) can be considered; this applies particularly to the case of systems using satellites in inclined elliptical orbits where the elevation angle under which the active satellite is viewed remains high (e.g. greater than 45° or 60°; see Section 2.3).

For links with geostationary satellites where the elevation angles are smaller, an omnidirectional antenna avoids the complexity and cost of a tracking system and also occupies less space.

8.4 THE RADIO-FREQUENCY SUBSYSTEM

The RF subsystem contains:

- On the receiving side, low noise amplifier (LNA) equipment and equipment for routing the received carriers to the demodulating channels.
- On the transmitting side, equipment for coupling the transmitted carriers and HPAs.

In each direction, frequency converters form the interface with the telecommunications subsystem, which operates at intermediate frequency.

8.4.1 Receiving equipment

The earth station figure of merit G/T is determined by the value of system noise temperature T that is given by Eq. (8.6):

$$T = (T_A/L_{FRX}) + T_F(1 - 1/L_{FRX}) + T_{eRX}$$

where T_A is the antenna temperature, L_{FRX} is the feeder loss between the antenna interface and the receiver input, T_F is the physical temperature of this connection, and T_{eRX} is the effective input noise temperature of the receiver.

Antenna temperature has already been mentioned in Section 8.3.3. At a given antenna temperature, the system noise temperature T is reduced by minimising the feeder loss between the antenna interface and the receiver input and limiting the effective input noise temperature of the receiver.

Noise contribution of the feeder loss is efficiently reduced by locating the first stage of the receiver as close as possible to the antenna feed. The *effective input noise temperature* T_{eRX} of the receiver is of the form (cf. Eq. (5.24)):

$$\begin{aligned} T_{eRX} = {} & T_{LNA} + (L_1 - 1)T_F/G_{LNA} + T_{MX}L_1/G_{LNA} \\ & + (L_2 - 1)T_F L_1/G_{LNA}G_{MX} + T_{IF}L_2L_1/G_{LNA}G_{MX} + \cdots \end{aligned} \tag{8.30}$$

where the effective input noise temperatures of the various stages are included with connection losses between these stages (Figure 8.26).

Equation (8.30) shows that it is necessary to use receiving equipment whose first stage has low noise and sufficiently high gain to mask the noise introduced by the following stages. In accordance with Figure 8.26 (see Section 8.4.1.2), the loss L_1 from the LNA output, located close to the antenna feed, to the frequency conversion equipment, often located at some distance from the LNA, includes either only feeder loss ($L_1 = L_{FRX}$) or feeder loss (L_{FRX}) combined with the influence of the power splitter ($L_1 = L_{FRX}L_{PS}\,n$, where L_{PS} is the power splitter insertion loss and n the power division ratio).

For small stations, frequency conversion and low noise amplification can be combined in equipment called a low noise block (LNB) converter, which is mounted behind the source (L_1 is close to zero); but the attenuation L_2 of the feeder between the converter and the following stages, usually a coaxial cable, can then make a non-negligible contribution.

The first Intelsat Standard A stations used cryogenically cooled maser amplification systems. The very high operating cost has caused these systems to be abandoned in favour of parametric amplifiers. The use of transistor amplifiers then became important first in C band and subsequently in Ku and Ka bands.

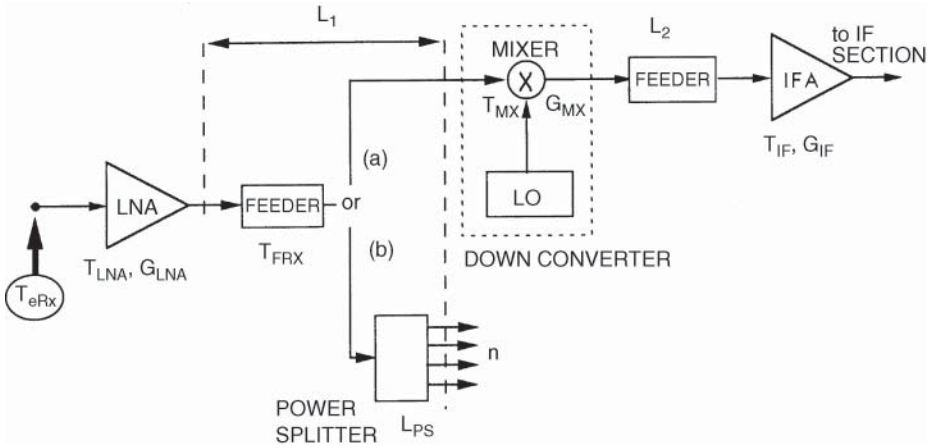


Figure 8.26 Receiver front-end block diagram: (a) fullband conversion; (b) carrier-by-carrier conversion.

8.4.1.1 Low noise amplifier

As a consequence of their junction structure, bipolar transistors cause (shot) noise other than thermal noise and can provide mediocre performance at high frequencies. On the other hand, the noise due to field effect transistors is mainly of thermal origin and can be reduced by selecting the type of semiconductor used and the geometric characteristics of the transistor. The performance in terms of noise factor is continuously improving due to the use of *gallium arsenide* (GaAs) and submicron lithography. Finally, the appearance of high electron mobility transistors (HEMTs) has enabled the noise temperature of receiving equipment to be further reduced, particularly at high frequencies (20 GHz). Table 8.11 gives typical noise temperatures as a function of the frequency band of HEMT front-end amplifiers. Peltier thermoelectric devices enable the temperature of the active element to be reduced to around −50 °C; the noise temperature of the amplifier is thus reduced in comparison with operation at ambient temperature.

Table 8.11 Noise temperature of low noise amplifiers (LNAs) using high electron mobility transistors (HEMTs)

Frequency band (GHz)	Noise temperature (K)
4	30
12	65
20	130
40	200

8.4.1.2 Frequency downconversion

Once low noise amplification has been performed, the carriers received in the frequency band used on the link are converted to a lower intermediate frequency (IF) where the operations of filtering and signal processing are simpler (see Section 8.5). The conversion may be realised either

on the whole of the frequency band (fullband conversion) used in the receiver (Figure 8.26a) or carrier by carrier (Figure 8.26b).

Fullband conversion of the frequency band is used in equipment intended for the reception of single channel carriers (SCPC). Distribution of carriers to different demodulators is performed at IF (typically 140 MHz), and selection of a particular carrier frequency (of narrow bandwidth, typically a few tens of kHz for voice and a few MHz for video) is achieved by alignment of the demodulator. Fullband conversion of the frequency band is also usual for small antennas intended for television signal reception or data transmission. In this case, the frequency converter is usually integrated with the low noise amplifier (LNA); the combination, named LNB, is mounted on the feed at the focus of the antenna. The converter output frequency is on the order of 1 GHz (900–1700 MHz), and this permits reduction of losses in the co-axial cable between the converter and the remainder of the equipment, which may be distant from the antenna.

Carrier-by-carrier conversion involves the use of frequency conversion equipment that permits the carrier concerned to be selected and converted to the IF. This IF is the same regardless of the frequency of the received RF carrier; tuning is performed at the downconverter by controlling the local oscillator frequency. This permits standardisation of IF equipment and hence cost reduction and simplified maintenance. Only the bandwidth of this equipment must be matched to that of the particular received carrier. Common values of IF are 70 and 140 MHz.

When an earth station must demodulate several carriers simultaneously, it is necessary to distribute the power at the low noise amplifier (LNA) output among the various converter channels. This is performed by a power splitter using passive devices (hybrid couplers or power dividers). The power splitter insertion loss L_{PS} adds to the feeder loss L_{FRX} from the LNA output to the power splitter input. The power split among the n channels translates into an attenuation by a factor n. Hence the total loss L_1 from the LNA output to any converter input is $L_1 = L_{FRX} L_{PS}\, n$.

8.4.2 Transmission equipment

The power per carrier P_T provided by the transmission equipment determines the value of the EIRP, which is a characteristic of the earth station for the link considered. The available carrier power P_T at the antenna input depends on the *power* P_{HPA} of the power amplifier, the *feeder loss* L_{FTX} between the output of the amplifier and the antenna interface, and the *power loss* L_{MC} entailed in multiple carrier operation, as specified by Eq. (8.3):

$$P_T = (P_{HPA})(1/L_{FTX})(1/L_{MC}) \quad \text{(W)}$$

The power amplifier characteristics vary according to the technology used, which may be travelling wave tube (TWT), klystron, or transistor. The magnitude and nature of the power loss L_{MC} entailed in multicarrier operation depends on the type of coupling, which may be performed before or after power amplification.

8.4.2.1 Power amplifiers

The power amplifier subsystem uses a tube or transistor power stage, which may be associated with a preamplifier and a linearizer. This subsystem also includes protection and control equipment and possibly a cooling system. Table 8.12 presents the main typical characteristics of these amplifiers.

Table 8.12 Power amplifier characteristics

Technology	Frequency (GHz)	Power (kW)	Efficiency (%)	Bandwidth (MHz)	Gain (dB)
Klystron	6	1–5	50	60	40
	14	0.5–3	35	90	40
	18	1.5	35	120	40
	30	0.5	30	150	40
Travelling wave tube (TWT)	6	0.1–3	40	600	50
	14	0.1–2.5	50	700	50
	18	0.5	50	1000	50
	30	0.05–0.15	50	3000	50
FET	6	5–100	30	600	30
	14	1–100	20	500	30

8.4.2.1.1 Tube amplifiers

The tube amplifiers used in earth stations are klystrons or TWTs [GIL-86]. The general organisation of these devices is similar; they consist of an electron gun, a system for focusing the electrons that enables an extended cylindrical beam to be obtained, a device that enables the kinetic energy of the electrons to be converted into electromagnetic energy, and a collector of the electrons.

In a *klystron*, the conversion device consists of a series of cavities, which are microwave resonant circuits and are traversed by the electron beam. The low-level electromagnetic wave that excites the first cavity causes modulation of the velocity of the electrons that cross it. This modulation creates an induced wave in the second cavity that in turn increases the modulation of the electron beam. The process repeats itself and is amplified in the following cavities. A RF wave at high level is thus produced at the output of the last cavity. The powers obtained range from a few hundreds of watts (around 800 W) to several kilowatts (5 kW). The bandwidth of the klystron is limited by the presence of resonant cavities in the amplification process. It is on the order of 40–80 MHz in C band (6 GHz) and 80–100 MHz in Ku band (14 GHz).

In the *travelling wave tube*, the energy transfer device is a helix that surrounds the electron beam and along which the electromagnetic wave propagates (see Figure 9.15). The helix effectively slows the wave so that the axial component of the electromagnetic wave velocity (equal to the product of the velocity of light and the ratio of the helix pitch to the length of one turn) is approximately equal to that of the electrons. Consequently, a continuous mechanism of energy transfer occurs along the helix. The electromagnetic wave gains the kinetic energy given up by the electrons. The power obtained ranges from several tens of watts (e.g. 35 W) to several kilowatts (e.g. 3 kW). The bandwidth of the TWT is large – about 600 MHz in C band at 6 GHz and about 3 GHz in Ka band at 30 GHz.

Tube amplifiers enable high powers to be produced and are therefore widely used in earth stations. The choice between TWTs and klystrons depends on the required bandwidth; for equal powers, the cost advantage is with the klystron. Available powers depend only slightly on frequency; tubes are available at 17 GHz (for feeder links) and 30 GHz.

Tube amplifiers require a suitable power supply to deliver the various voltages (up to 10 kV) required on the electrodes. These voltages must be adequately regulated (to a relative value of 10^{-3}) in order to avoid fluctuations (e.g. phase noise) of the amplified carriers. For high powers, it is necessary to provide forced air (up to around 3 kW) or circulating liquid cooling arrangements. In spite of a high power gain (40–50 dB), the power required at the input of the tube usually

requires the use of a pre-amplifier. This pre-amplifier uses a low-power TWT (from a few watts to a few tens of watts) or several transistor stages.

8.4.2.1.2 Transistor amplifiers

Semiconductor amplifiers provide powers up to about a few 100 W in C band (6 GHz) and several tens of watts in Ku band (14 GHz). These amplifiers usually use *gallium arsenide* (GaAs) field effect transistors mounted in parallel. In spite of the low powers available (which are continuously increasing with advances in technology), transistor amplifiers are increasingly used because of their low cost, linearity, and wide bandwidth.

8.4.2.1.3 Power amplifier characteristics

Nonlinearity. Power amplifiers are nonlinear. As shown in Figures 6.8 and 9.2, as the carrier power applied to the input of an amplifier is increased, there is a region of quasi-linear operation at low level after which the output power no longer increases in proportion to the input power. The maximum power obtained at the output corresponds to saturation (unless a limit on power dissipation prevents the saturation point being reached, which is the case for solid state amplifiers in particular). The *maximum output power at saturation in single carrier operation* $(P_{o1})_{sat}$ is the rated power given in the manufacturer's data sheet (P_{HPA}). With solid-state amplifiers that cannot be operated at saturation, the maximum output power is most often specified by the *1 dB compression point* (see Section 9.2.1.2). When operating with several carriers, intermodulation products appear at frequencies corresponding to linear combinations of the input carrier frequencies (see Section 6.4.3). When the carriers are modulated, the intermodulation products that fall within the useful bandwidth of the amplifier behave as noise [SHI-71]. This noise is characterised within the bandwidth of each carrier by the value of the intermodulation power spectral density $(N_0)_{IM}$.

To limit intermodulation noise when several carriers are amplified simultaneously to a value compatible with the overall link budget requirement (see Sections 5.9.2.3 and 5.9.2.4), it is advisable to operate the amplifier below the saturation region. The *output back-off* (OBO), defined as the ratio of the output power delivered on one of the n carriers (P_{on}) to the saturation power, determines the position of the operating point (Sections 5.9.1.4 and 5.9.2.4). The power delivered at the amplifier output for the carrier concerned is thus equal to:

$$(P_{on}) = P_{HPA} \times \text{OBO} \quad \text{(W)} \tag{8.31}$$

The value of back-off depends on the minimum allowed value of the earth station's contribution to the carrier power to intermodulation power spectral density ratio $(C/N_0)_{IM}$ of the link, the number of carriers, and the input–output characteristic of the amplifier.

The *total back-off* is sometimes defined as the ratio of the total power available on all N carriers to the saturation power in single carrier operation; when the carriers are of equal level, the power per carrier can be obtained by dividing the product of the amplifier saturation output power and the total OBO by n. Additional specifications (such as the *third-order intercept point*, the AM/PM *conversion* and *transfer coefficients*, and the *noise power ratio* [NPR] relating to the characteristics of amplifier nonlinearities) are given in Section 9.2.1.

Gain variations. The specified gain of a power amplifier is susceptible to variation as a function of various parameters. It is important to specify the stability of the gain: that is, the magnitude of

permitted variations as a function of the various parameters. Hence the following are specified for a particular application:

- The *stability of the gain* as a function of time (e.g. $\leq$0.4 dB/24 h) at constant input level
- The magnitude of *gain variation* as a function of frequency within the bandwidth for a given power level (e.g. $\leq$ 4 dB in 500 MHz)
- The maximum *rate of change of gain fluctuation* as a function of frequency in a specified portion of the band (e.g. $\leq$ 0.05 dB MHz^{-1})

Standing wave ratio (SWR) and propagation time. The maximum *standing wave ratio* is specified at the input and output of the amplifier and also for the load driven by it. The *group delay* within the frequency band of a power amplifier varies as a function of frequency. Meeting the specification can require the installation of propagation time equalisers in the transmission channel, usually in the stages operating at IF.

RF/DC efficiency. The (power) *RF/DC efficiency* is defined as the ratio of the output RF power (W) to the direct current (DC) power (volts × amperes) required to operate the amplifier. The efficiency depends on the operating point (back-off value) and is typically maximum at *saturation* (or at *1 dB compression point* for solid state amplifiers).

8.4.2.1.4 Comparison of power amplifier technologies

Table 8.13 summarises the advantages and drawbacks of the various technologies including TWTA, SSPA, and Klystron.

8.4.2.2 Linearisers

The use of *linearisers* is becoming more common in order to limit the effects of amplifier nonlinearity. Combined with the pre-amplifier, or located before it, most linearisers produce amplitude and phase distortion of the signal in order to compensate for the specific characteristics of the power amplifier (Figure 8.27). For a given level of intermodulation noise, the lineariser permits a reduction of back-off (in absolute value); that is, the amplifier is operated closer to saturation. The reduction of back-off provides a considerably greater available carrier power for an amplifier of given saturation power and potential cost, power consumption, and bulk reduction.

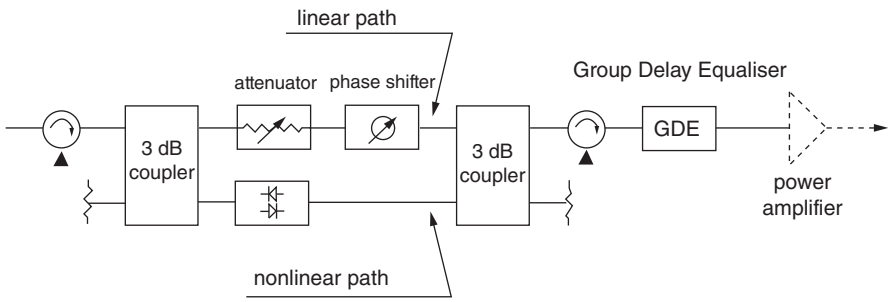


Figure 8.27 Nonlinear predistortion type lineariser.

Table 8.13 Power amplifier technologies compared

Technologies	Advantages	Disadvantages
TWTA	• Medium to large output power (35–3000 W) • Proven fielded robust performance • Good RF/DC efficiency: 30–50%; does not decrease quickly with back-off • Stable performance over temperature • Instantaneous, broadband capability • Long life • No memory effect (nonlinearity)	• Limited TWT production • No soft-fail capability in case of failure • High voltages required • Nonlinear, but linearisers exist or operation with back-off
SSPA	• High-volume production capability • Built-in soft-fail capabilities in case of single device or module failure • Inherently good linear performance for multicarrier transmission	• Limited output powers (100 s W at C and Ku bands, 10 s W at Ka band) • Highly inefficient (10–30%) • High currents • Need to be temperature compensated • Large amounts of heat at concentrated locations; heat-dissipation problems • Increased size and weight at high power levels due to added cooling requirements (air flow, heat sinks, etc.)
Klystron	• High power (several kW), headroom for back-off for optimal linear performance • Good linear performance for multicarrier • Cost effective, reliable • Good efficiency (50%)	• Narrower instantaneous bandwidth (40–90 MHz), but tunable over 500 MHz or more

8.4.2.3 Carrier pre-coupling

As stated in the introduction to this chapter, an earth station very often transmits several carriers (at different frequencies) to the satellite concerned. As the antenna interface generally has only a single input (for a given polarisation), it is necessary to multiplex these carriers, which have been modulated separately, by frequency division in order to combine them on the same physical connection.

Carrier coupling can be performed at a low power level (for example, by using hybrid couplers) *before power amplification* (Figure 8.28). Power amplification then operates in a multicarrier regime and must be operated with some OBO in order to limit intermodulation noise power. The power

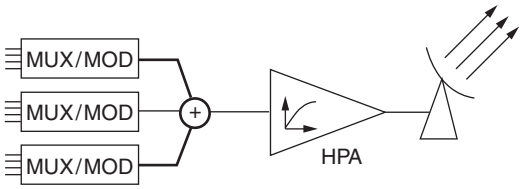


Figure 8.28 Carrier coupling prior to power amplification.

loss L_{MC} caused by multicarrier operation in this case has a value

$$(L_{MC})_{ES} = -(OBO)_{ES} \ (dB)$$

where OBO is the output back-off defined as the ratio of the available power of the carrier concerned (one of N) to the output power at saturation in single carrier operation.

The advantages of pre-coupling lie in the simplicity of coupling and the flexibility to adapt to changes in the number and bandwidth of carriers. The number of amplifiers is also minimised. On the other hand, this mode of coupling introduces a source of intermodulation noise in the earth segment that affects the overall link budget. Limitation of intermodulation noise to an acceptable value requires the use of an amplifier with sufficient back-off, and this leads to the use of a device with a saturation power much greater than the required power. Furthermore, the amplifier must have a sufficient bandwidth to amplify the different carriers (this can prohibit the use of a klystron, which would otherwise be more economic).

8.4.2.4 Carrier post-coupling

Coupling can also be performed after separate amplification of each carrier (Figure 8.29). It is then necessary to have as many amplifiers as there are carriers (plus any backup equipment). Each amplifier amplifies only one carrier; the amplifiers can therefore operate at saturation.

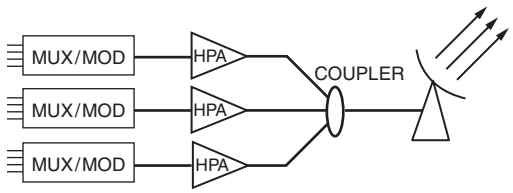


Figure 8.29 Post-amplification carrier coupling.

However, the coupling device introduces losses L_C, which can be identified with the reduction in power L_{MC} caused by multicarrier operation in Eq. (8.3):

$$(L_{MC})_{EC} = L_C \ (\text{dB})$$

Since each carrier is amplified separately, the required bandwidth is limited. Furthermore, operation to saturation enables low power, and hence lower cost, amplifiers to be used. However, the carrier coupling has to be done at high level with a minimum of loss.

Two types of device permit coupling to be realised:

- Hybrid couplers (aperiodic coupling)
- Band-pass filter multiplexers

8.4.2.4.1 Aperiodic coupling

To couple the amplifier outputs while matching the impedances, it is possible to use conventional hybrid couplers. This approach permits considerable flexibility of use since coupling is wideband, but this is accompanied by large losses. In fact, a hybrid coupler permits two signals to be combined but the power of each signal is shared between the two outputs (in equal parts for a 3 dB coupler). As a single output is used, the power on the unused output is dissipated in the matched load (half of the total power for a 3 dB coupler). This loss occurs again when coupling the sum of the first two signals with the third, and so on.

8.4.2.4.2 Coupling by multiplexer

The use of multiplexers involving band-pass filters tuned to each carrier enables the losses to be minimised to the detriment of system flexibility. Two techniques are used to combine the signals:

- Circulators route the signals after band-pass filtering and reflection on to the filter outputs in accordance with a principle similar to that used for satellite output multiplexers (OMUX) (see Section 9.2.3.2).
- Hybrid couplers (Figure 8.30) operate as follows. The signal of amplitude A corresponding to carrier 1 is divided into two components of amplitude $A/\sqrt{2}$ and phase-shifted by 90° by the first hybrid coupler. These two components pass through the band-pass filters tuned to the frequency and band of carrier 1. The two components present at the input of the second coupler after division by $\sqrt{2}$ are summed in phase at the antenna port and in phase opposition on the other port (port B) due to their phase shift of 90°. The signal of amplitude B corresponding to carrier 2 is divided into two components of amplitude $B/\sqrt{2}$ and phase shifted by 90° by the second hybrid coupler. These two components are reflected on to the outputs of the band-pass filters tuned to the frequency and band of carrier 1. The reflected components appear at the ports of the second coupler and are also summed in phase on the antenna port after division by $\sqrt{2}$.

The disadvantages of coupling by multiplexer are the loss of flexibility associated with the need to have band-pass filters that are perfectly matched to the characteristics of the carriers to be combined. On the other hand, the losses are small, on the order of a decibel.

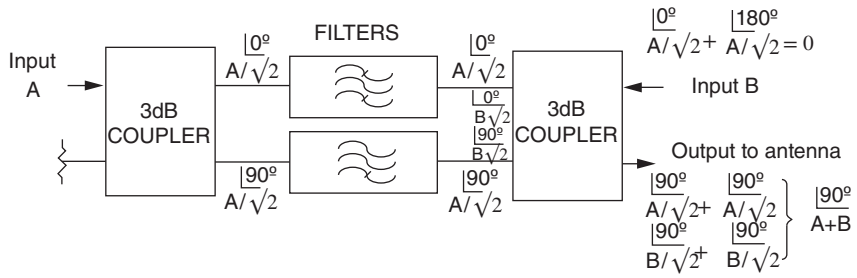


Figure 8.30 Carrier coupling with filter combiner.

8.4.2.5 Mixed coupling (pre- and post-coupling of carriers)

A combination of the two types of coupling is often used. Each amplifier amplifies a restricted number of carriers and the outputs of these amplifiers are coupled by one of the techniques (or a combination of the two) described in the previous section. This permits, for example, all carriers that are routed by a given transponder to be coupled into the same amplifier. Then, at the earth station, each amplifier can be associated with a particular satellite transponder.

8.4.3 Redundancy

To satisfy the objectives of reliability and specified availability, it is often necessary to back up the RF equipment of an earth station (see Chapter 13).

As far as the input stages are concerned, the use of a redundant receiver is normal except for small stations that usually do not incorporate redundancy. Since the operation of the station is monitored, it is rare to have more than one standby system, since maintenance of the station is guaranteed.

For the output stages, the redundancy arrangement depends on the type of coupling. With pre-coupling of the carriers, the power amplifier (single except in the case of mixed coupling) is usually replicated. With post-coupling of the carriers, it is not useful to replicate each carrier amplifier, and the use of backup equipment shared among several active units is usual.

8.5 COMMUNICATION SUBSYSTEMS

The communication subsystem on the transmission side consists of equipment for converting baseband signals to RF carriers for amplification; conversely, on the reception side, it converts the carriers at the output of the low noise amplifier to baseband signals.

The baseband signal may be either analogue or digital. In the analogue case, it can be a telephone channel in the case of a single channel per carrier (SCPC) transmission system, a multiplex of telephone channels, a television signal, or a sound programme.

In the digital case, it is usually in the form of a bitstream that corresponds to one or a multiplex of voice channels or data frames or packets.

The functions to be realised on the receiving side are as follows:

- Conversion of the carrier frequency (RF) to an IF
- Filtering and equalisation of group propagation delay
- Carrier demodulation

In the case of transmission using TDMA, it is necessary to re-establish a continuous digital stream from the packets of the received frame.

On the transmission side, if TDMA is used, it is necessary to group the bits of the baseband signal into packets that are inserted in the proper time slots provided in the frame.

Finally, as for analogue signals, the following operations are performed:

- *Modulation* of a carrier at an IF
- *Filtering* and *equalisation* of group propagation delay
- *Conversion* of the modulated carries to RF

8.5.1 Frequency translation

The function of the frequency-translation subsystem is to select a particular in-band carrier at the output of the LNA and translate the spectrum of this carrier to the chosen IF. An IF of the same conventional value for every channel permits the use of standardised equipment. The choice of IF is determined by the following considerations: on the one hand, the value must be greater than the spectral width occupied by the modulated carrier; but, on the other hand, this value must be sufficiently low to permit selective band-pass filtering of the modulated carrier. The selectivity Δf of a filter represented by its *quality factor* Q is defined by the ratio f/Q, where f is the central frequency of the filter. Assuming a quality factor of 50, if it is required to isolate a signal occupying a bandwidth of 1 MHz, the maximum operating frequency of the filter is 50 MHz. Common values of IF are 70 and 140 MHz.

Frequency translation can be of the single- or dual-conversion type. The organisation of the two types of frequency translation system for reception is described in the following sections. System architectures on the transmission side are similar.

8.5.1.1 Single frequency conversion

The frequency-translation system consists of a band-pass filter centred on the RF carrier to be received and a mixer that is also fed by the signal from a local oscillator (Figure 8.31). The input filter serves to eliminate the *image frequency*, which is a characteristic of the frequency-conversion process. If f_{LO} is the frequency of the local oscillator, two carriers of frequency $f_{LO} + f_{IF}$ and $f_{LO} - f_{IF}$ are translated to the IF f_{IF}. Only one of these two frequencies corresponds to the frequency f_c of the carrier to be received. The other frequency $f_1 = f_c + 2f_{IF}$, which is called the *image frequency*, may correspond to another carrier and must be eliminated. The separation between the required carrier and its image frequency is equal to $2f_{IF}$. For a low value of IF (e.g. 70 MHz), it is necessary to provide an RF filter with high selectivity tuned to the frequency of the carrier to be received; for example, if $f_{IF} = 70$ MHz, the selectivity must be on the order of 200 at 12 GHz.

Selection of the received carrier is achieved by changing the frequency of the local oscillator and the centre frequency of the image frequency rejection filter. Realisation of a tunable and

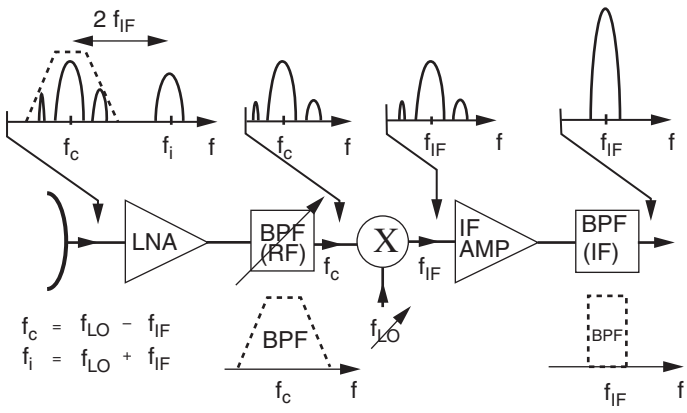


Figure 8.31 Single conversion downconverter.

easily controllable RF filter is difficult, and the double frequency changing structure is often preferred.

8.5.1.2 Dual frequency conversion

To provide frequency agility without tuning the input filter to the carrier to be received, it is necessary to keep the image frequency outside the band of frequencies within which the carrier to be received can occur (Figure 8.32).

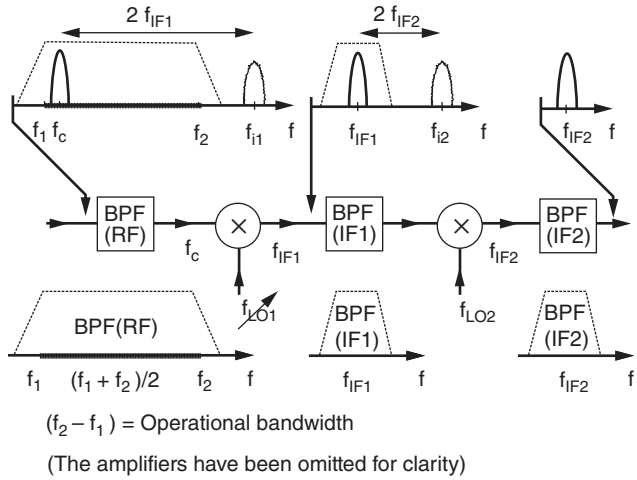


Figure 8.32 Dual conversion downconverter.

The frequency obtained after translation must be as high as the receiving bandwidth is wide. For example, for the band $f_1 = 3.625$ to $f_2 = 4.2$ GHz (width $f_2 - f_1 = 575$ MHz), downconverting the band about the IF $f_{IF1} = 1400$ MHz permits the use of an input band-pass filter with a fixed bandwidth of 575 MHz; this ensures sufficient rejection of the image frequency, which, in the worst case (when the carrier to be received is at the lower edge of the band, i.e. $f_c = f_1$), is at $f_i = f_c + 2f_{IF1} = f_1 + 2f_{IF1} = 3625 + 2 \times 1400 = 6425$ MHz: that is, 2225 MHz greater than f_2 and hence outside the passband.

The desired IF f_{IF2} (e.g. 70 MHz) is obtained by a second frequency translation, which is performed after band-pass filtering centred on the value of the first IF f_{IF1}. This filter has a sufficient bandwidth (e.g. 40 MHz) to allow any type of modulated carrier to pass. It eliminates the image frequency in the second translation since the image frequency is situated at $f_{i2} = 1540$ MHz, i.e. 140 MHz above the first IF $f_{IF1} = 1400$ MHz if the second local oscillator generates a frequency of $f_{LO2} = 1470$ MHz (if $f_{IF2} = 70$ MHz).

Selection of the received carrier frequency f_c is accomplished by setting the frequency of the first local oscillator to $f_c + 1400$ MHz. A frequency synthesiser is often used (with frequency variation in steps, for example 125 KHz). The frequency of the second oscillator remains fixed, as does the central frequency of the various band-pass filters.

Some configurations use a fixed RF source for the first oscillator. Tuning is then achieved by adjusting the frequency of the second oscillator (a synthesiser) that operates at a low frequency and is thus easier to design. On the other hand, the first IF is not of a fixed value and it is necessary for the associated band-pass filter to be tunable.

8.5.1.3 Fullband conversion

In the previous section, only one carrier was downconverted at a time, and hence there was only one carrier involved per IF channel. It is also possible to translate the whole of the received frequency band, and hence all the carriers, to the IF band at the same time. This architecture is used particularly for SCPC systems.

Figure 8.33 shows the block diagram of a subsystem for transmission and reception translation with a double change of frequency; modulated carriers in the 52–88 MHz band are translated into the 5.850–6.425 GHz band with a double frequency conversion using a second IF of 825 MHz. On the receiving side, received carriers in the 3.625–4.200 GHz band are translated to the 52–88 MHz band (with a second IF of 1400 MHz). The two translations use a single frequency synthesiser.

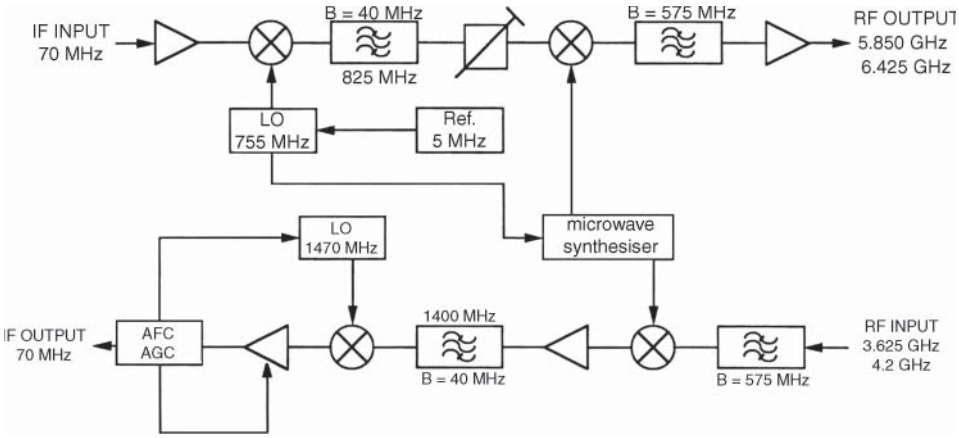


Figure 8.33 Architecture of a transmit and receive frequency converter.

8.5.1.4 Characteristics of frequency translation subsystems

Apart from the capacity for frequency agility discussed earlier, which determines the range of acceptable frequencies for input and output signals, the specified characteristics of a frequency translation subsystem are as follows:

- *Frequency stability* of the local oscillators (long-term and phase noise)
- Maximum *level of spurious* frequency components
- Long-term *gain stability* in the frequency band
- *Linearity* (the level of intermodulation products or intercept point)

8.5.2 Amplification, filtering, and equalisation

The functions of amplification, filtering, and group propagation delay equalisation are realised at IF. These operations are facilitated by retaining a *fixed IF* regardless of the RF carrier concerned.

On reception, the IF amplifier includes automatic gain control so that a constant level is provided at the input of the demodulation subsystem. On the transmission side, gain control enables the level (or the back-off) at the input of the RF amplifier to be adjusted.

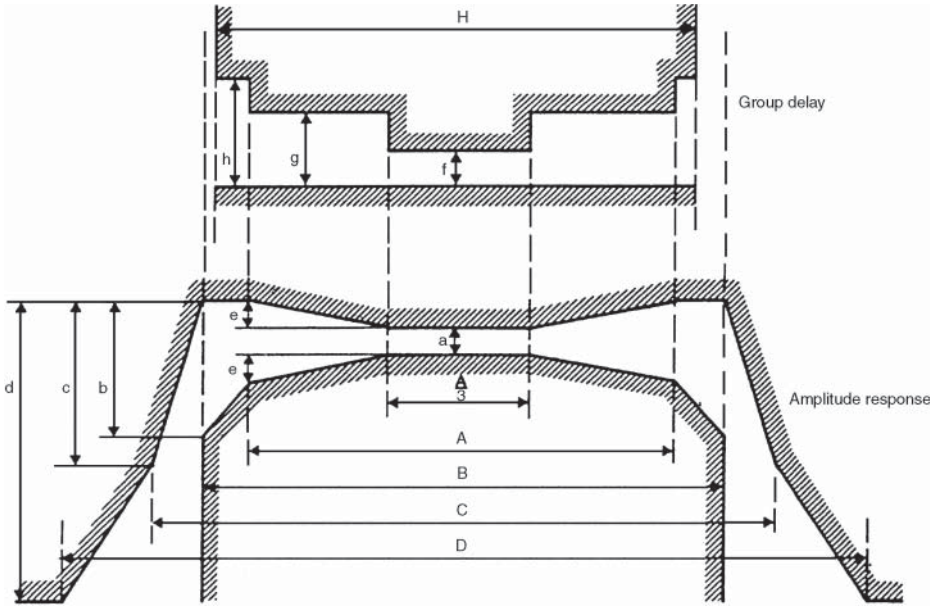


Figure 8.34 Amplitude and group delay limits for filtering at intermediate frequency (Intelsat). The values of the parameters shown in the figure are given in Tables 8.14 and 8.15 as a function of the bandwidth occupied by the modulated carrier.

Band-pass filtering at IF defines the spectrum of the modulated carrier and limits the noise bandwidth. The characteristics of this filter depend on the modulation characteristics of the carrier concerned. These filters are usually designed with transfer functions of the Butterworth or Chebyshev type using capacitors and inductors. The filter elements in the transmission and reception channels, the power amplification stages, and the satellite transponder all introduce group delay variations as a function of frequency. These variations are corrected within the useful bandwidth by means of *group propagation delay equalisers*. These equalisers are integrated into the band-pass filter or realised separately by means of inductor capacity (LC) cells of the bridged T type. By way of example, Figure 8.34 and Tables 8.14 and 8.15 give the Intelsat specifications for

Table 8.14 Characteristics of the Intelsat filtering amplitude specification

Carrier bandwidth (MHz)	A (MHz)	B (MHz)	C (MHz)	D (MHz)	a (dB)	b (dB)	c (dB)	d (dB)	e (dB)
1.25	0.9	1.13	1.15	4.0	0.7	1.15	3.0	25	0.0
2.5	1.8	2.25	2.75	8.0	0.7	1.5	2.5	25	0.0
5.0	3.6	4.50	5.25	13.0	0.5	2.0	3.0	25	0.0
10.0	7.2	9.00	10.25	19.0	0.3	2.5	5.0	25	0.1
20.0	14.4	18.00	20.50	28.0	0.3	2.5	7.5	25	0.1
36.0	28.8	36.00	45.25	60.0	0.6	2.5	10.0	25	0.3
Video	12.6	15.75	18.00	26.5	0.3	2.5	6.5	25	0.1
Video	24.0	30.00	–	–	0.5	2.5	–	–	0.3

Table 8.15 Characteristics of the Intelsat filtering group delay specification

Carrier bandwidth (MHz)	A (MHz)	H (MHz)	f (ns)	g (ns)	h (ns)
1.25	0.9	1.13	24	24	30
2.5	1.8	2.1	16	16	20
5.0	3.6	4.1	12	12	20
10.0	7.2	8.3	9	9	18
20.0	14.4	16.6	4	5	15
36.0	28.8	33.1	3	5	15
Video	12.6	14.2	6	6	15
Video	24.0	30.0	5	5	15

filtering and group delay equalisation as a function of the bandwidth occupied by the modulated carrier.

8.5.3 Modems

The operations of modulation (on the transmission side) and demodulation (on the receiving side) are realised at IF. Modem subsystems are realised in accordance with the type of baseband signal (multiplex or single channel), the type of channel coding (FEC – forward error correction), the type of modulation of the carrier, and the multiple access mode (FDMA, TDMA, MF/TDMA, CDMA).

8.5.3.1 Energy dispersion

Energy dispersion avoids the appearance of discrete frequencies in the spectrum of the modulated signal. Energy dispersion is realised by scrambling the bitstream to be transmitted before modulation (see Section 4.2.3). Scrambling is performed by modulo 2 addition of the bitstream and a pseudorandom sequence that is generated by a set of shift registers with appropriate feedback. On reception, the scrambled sequence recovered by the demodulator (containing erroneous bits) is combined with the same pseudorandom sequence that is generated locally and appropriately synchronised.

8.5.3.2 Channel coding and decoding, interleaving

Channel encoding has the objective of adding redundant bits to the information bits; the former are used at the receiver to detect and correct errors. This technique is called *forward error correction* (FEC). The detail and performance improvement of channel coding is discussed in Section 4.3. Two encoding techniques are used: block encoding and convolutional encoding.

With *block encoding*, the encoder associates r bits of redundancy with each block of n information bits; each block is coded independently of the others. The code bits are generated by linear combination of the information bits of the corresponding block. Cyclic codes, particularly the codes of Reed–Solomon (RS) and Bose–Chaudhuri–Hocquenghem (BCH), for which every code word is a multiple of a generating polynomial, are most used.

With *convolutional encoding*, $(n + r)$ bits are generated by the encoder from the $(N - 1)$ preceding packets of n bits of information; the product $N(n + r)$ defines the *constraint length* of the code. The encoder consists of shift registers and adders of the 'exclusive OR' type.

Today, based on the performances obtained with the DVB standards, concatenated coding is most often considered (Section 4.7). In that case, the FEC encoder performs outer coding and inner coding. With the DVB-S standard, the outer code is an RS block code and the inner code is a convolutional code, the code rate of which is adjusted by puncturing. Interleaving is typically implemented between the inner and outer coders in order to ensure that deinterleaving between the Viterbi decoder and the RS decoder spreads out the residual errors at the output of the Viterbi decoder and to provide quasi-error-free performance at the output of the RS decoder.

With the DVB-S2 standard, the outer code is a BCH code and the inner code is a low-density parity check (LDPC) code (Section 4.8). The selected LDPC codes use very large block lengths (64 800 bits, for applications not too critical for delays, and 16 200 bits). Code rates of 1/4, 1/3, 2/5, 1/2, 3/5, 2/3, 3/4, 4/5, 5/6, 8/9, and 9/10 are available, depending on the selected modulation and the system requirements. Recursive decoding techniques (turbo-decoding) are used. Bit interleaving on the encoded frames is implemented prior to modulation [ETSI-14; ETSI-15].

8.5.3.3 Modulation and demodulation

With digital transmission, phase modulation (BPSK, QPSK, 8PSK) or combined amplitude and phase modulation, such as 16APSK and 32APSK with DVB-S2, are used. The principle of these modulation types was presented in Section 4.2. The carrier is modulated at IF from the sine wave generated by a quartz crystal oscillator or a frequency synthesiser. Band-pass filtering limits the spectrum of the modulated carrier.

With coherent demodulation, the received modulated carrier is multiplied by an unmodulated carrier that is generated locally. Carrier recovery is achieved by passing the (modulated) received carrier through a nonlinear circuit, which produces components at frequencies in the spectrum that are multiples of that of the carrier, then by filtering one of these components and frequency division.

Frequency division introduces a phase ambiguity that must be resolved for correct detection of the signal. Filtering of the spectral components is performed either by a phase-locked loop or a passive filter. In the case of transmission using TDMA, where the equipment operates in bursts, small carrier acquisition times are to be achieved.

8.5.3.4 Burst mode operation

With TDMA or multi frequency TDMA (MF/TDMA), the modem operates in burst mode. It gets the baseband signals to be transmitted in continuous mode and supplies packets of information (stored in buffer memory) to the RF equipment that transmits these packets in bursts at the instants that correspond to their time-slot assignment in the frame. On the receiving side, the terminal accepts the modulated carrier in bursts and reconstitutes a continuous bitstream.

With MF/TDMA, the subsystem must have frequency-hopping capabilities by generating burst carriers at different frequencies (on the transmitting side) and, if relevant, by multiplexing into the demodulator bursts received at different frequencies.

On reception, the bursts received in succession are from different earth stations and thus have different phases and amplitudes. The use of a passive filter (with automatic control of the centre

frequency of the filter that corrects slow drifts of the input frequency) resolves the problem of rapid phase recovery at the start of each burst. Automatic phase control compensates for the rapid variation of frequency between consecutive bursts. Automatic gain control with a rapid response time (less than a microsecond) permits burst-to-burst amplitude variations to be compensated.

On transmission, when bursts are not being transmitted, the IF channel output level should be sufficiently low to avoid interference (a typical rejection ratio is greater than 60 dB).

The frequency-hopping technique involves selection of a particular satellite carrier frequency at which the burst concerned is transmitted according to its destination. The station is thus provided with n channels at IF corresponding to n different converters that can produce n different RF carriers. On transmission, the frequency-switching subsystem routes the bursts from the modulator to the n IF channels.

An example block diagram of a burst modem is shown in Figure 8.35. On top of the functions discussed, the modem implements a series of controlling functions:

- The receiver controller performs alignment of the data at symbol level using signals from the TDMA terminal clock, recognition of unique words and resolution of ambiguity in data demodulation, synchronisation of the buffers containing packetised data to the received burst time plan, descrambling and error correction decoding, and demultiplexing of service channels (SCs).
- The timebase generates the terminal clock. This module includes a voltage-controlled quartz crystal oscillator that synchronises itself to the reference packet clock. Other time references are obtained by division of the main clock.
- The transmission controller provides functions similar to those of the receiver controller: synchronisation of the buffers containing packetised data to the transmit burst time plan, generation of the preamble (header), data multiplexing, application of error correcting coding, and scrambling.
- The main processor provides the following functions: acquisition and synchronisation of the network, management and processing of burst time plans and, possibly, automatic test and diagnostics, management of redundancy, and assistance with maintenance.

The auxiliary processor extracts the messages from the SCs and the control delay channel (CDC) that have been demultiplexed on reception and routes them to the main processor for execution. It also participates in the acquisition and synchronisation procedures.

8.6 THE NETWORK INTERFACE SUBSYSTEM

This subsystem is the interface between baseband signals produced by, or destined for, the communication common equipment and baseband signals in the terrestrial network format. The main functions are multiplexing (and demultiplexing) of telephone channels, which may include DSI and channel multiplication (DCME), suppression (or cancellation) of echoes, and various functions particular to single channel transmission (SCPC). The new modem device also provides RJ45 interface for Ethernet connection to provide IP network services. Future satellite systems will support high-speed Internet packet services rather than traditional voice and low-speed digital services.

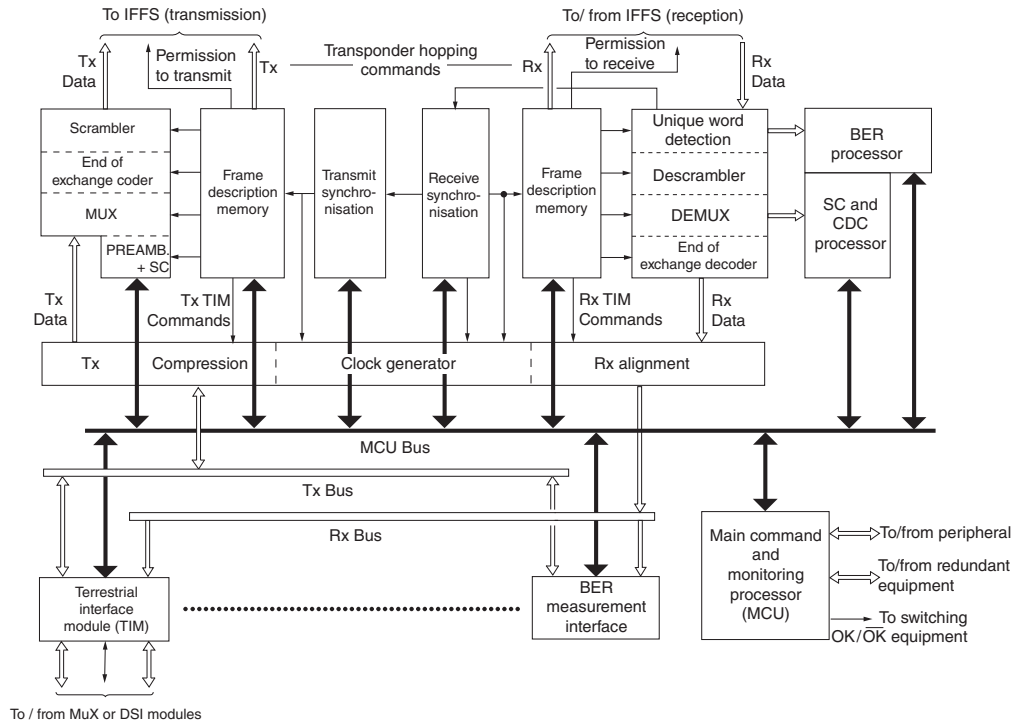


Figure 8.35 Block diagram of a common logic equipment (CLE) subsystem of a TDMA terminal.

8.6.1 Multiplexing and demultiplexing

Even if the telephone channels on the terrestrial network are already multiplexed, it is almost always necessary to rearrange the distribution of telephone channels as the multiplexing standards used on the terrestrial network and on satellite links are slightly different. Furthermore, the telephone channels arriving at the earth station on the terrestrial link do not all have the same destination. Telephone channels having the same destination are grouped into a single multiplex that modulates a carrier and is transmitted to this particular destination. Similarly, on reception, only one section of the telephone channels present in a received multiplex relate to the terrestrial network connected to the earth station concerned. These channels (or groups or supergroups of channels) are separated from the others and combined with those from other multiplexes received on the various multidestination carriers from other stations. These channels are then combined to form a terrestrial standard multiplex destined for the switching exchange connected to the earth station concerned.

Digital transmission multiplexing standards were presented in Section 3.1.1. There are two types of hierarchy, and these have a base of either 24 telephone channels (1.544 Mbit s^{-1}) or 30 telephone channels (CEPT, 2.048 Mbit/s). These hierarchies are used in terrestrial networks and on satellite links.

The multiplexing equipment in the earth stations combines bitstreams from various origins that have the same destination. Problems arise due to lack of synchronisation of the earth station clock and the bitstream clocks at diverse origins. This lack of synchronism is due to instability and drift of the oscillators and variations of propagation times on the links.

When the bitstreams are not synchronous, it is necessary to use buffer memories and possibly to add stuffing bits in order to obtain exactly the same bit rates before multiplexing. The operation is facilitated when the bitstreams are synchronous (synchronised clocks) or *plesiochronous* (clocks that are almost but not quite synchronised with accuracy on the order of 10^{-11}). In the latter case, the technique of frame slipping permits periodic readjustment of the bitstreams, and there is no need for stuffing bits (see Section 3.1.1.3.4).

8.6.2 Digital speech interpolation (DSI)

DSI exploits the silences in a telephone channel to insert bits representing the active speech of another channel into these silences (Section 3.1.1.3.3). In this way, a number m of telephone channels from the terrestrial network can be carried in a multiplex with a capacity of n digital telephone channels where $m > n$ [CAM-76; KEP-89].

Figure 8.36 illustrates this principle. Speech detectors are necessary to identify the silences in the terrestrial network channels. The DSI equipment assigns one channel of the satellite link (the *bearer channel*) to each active terrestrial telephone channel, and a connection network performs the corresponding branching operations. The assignment is arbitrary and can change from one talk spurt to another on the same terrestrial channel. The assignment information is transmitted on a signalling channel.

On reception, dedicated equipment establishes the connections between the bearer channels and the terrestrial telephone channels, in accordance with the signalling messages received, in such a way as to route the bits to their proper destination.

The performance of a DSI system is measured by the *speech interpolation gain m/n*. This gain is greatest when the number of channels to be concentrated is large. If the number of channels

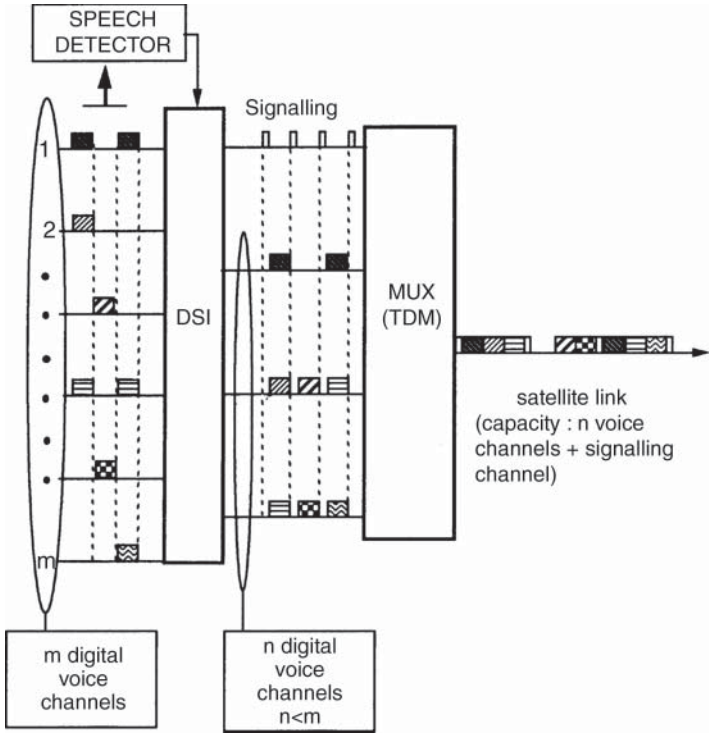


Figure 8.36 Digital speech interpolation (DSI).

exceeds 60, the gain obtained can reach 2.5. Limitation of the speech interpolation gain arises from the following:

- Degradation of quality associated with clipping when a talk spurt cannot be routed because at that instant all the bearer channels of the satellite link are occupied
- Degradation associated with temporary overloading of the signalling channel
- Terrestrial telephone channels that have an activity factor greater than that of speech (when used for data transmission, for example)

If the number of active terrestrial telephone channels at a given time exceeds the number n of available bearer channels, the bits in excess from the blocked talk spurt sample can be routed by using a *bit-stealing* technique; the least significant bit in the quantised speech sample on seven telephone channels of the terrestrial network is not transmitted. For these seven telephone channels, quantisation is performed to seven bits instead of eight; this temporarily increases the quantisation noise, but the quality degradation on the seven telephone channels is less than that which would be observed on the clipped telephone channel.

8.6.3 Digital circuit multiplication equipment (DCME)

DCME permits an improvement on DSI in the commercial exploitation of satellite telephone channels (see Intelsat Specification IESS 501, [FOR-89; YAT-89]). The system combines two

techniques for multiplying the number of telephone channels that can be transmitted on the same satellite bearer channel; these are DSI and *adaptive differential pulse code modulation* (ADPCM).

In comparison with the technique of DSI described in the previous section, a further factor of two is obtained by means of adaptive differential coding. Four bits are used to code samples of the voice signal instead of eight bits. Hence a given number of bearer channels on the satellite link can convey a greater number of terrestrial channels (on the order of five times more).

8.6.3.1 DCME organisation

A typical organisation of a circuit multiplication equipment is shown in Figure 8.37. It includes the following pieces of equipment:

- *Input data link interface (DLI)*: This interface equipment handles the terrestrial trunks at 1.544 Mbit s^{-1} or 2.048 Mbit s^{-1} and delivers 2.048 Mbit s^{-1} bitstreams; it also provides clock recovery and delivers frame synchronisation in plesiochronous mode. When converting 1.544 to 2.048 Mbit s^{-1}, the DLI introduces stuffing bits and only 24 bits out of 31 correspond to information bits.
- *Time-slot interchange (TSI):* This equipment is used when trunks at 1.544 Mbit s^{-1} are present at the input of the DLI. The TSI groups 10 streams at 2.048 Mbit s^{-1} (initially at 1.544 Mbit s^{-1}) in order to obtain 8 streams at 2.048 Mbit s^{-1} that now contain only information bits and the control bits that are generated by the equipment.
- *DSI:* This equipment consists of speech detectors associated with a noise-level monitor, possibly a delay line to anticipate speech detection, a 2100 Hz pilot tone detector, and a device to

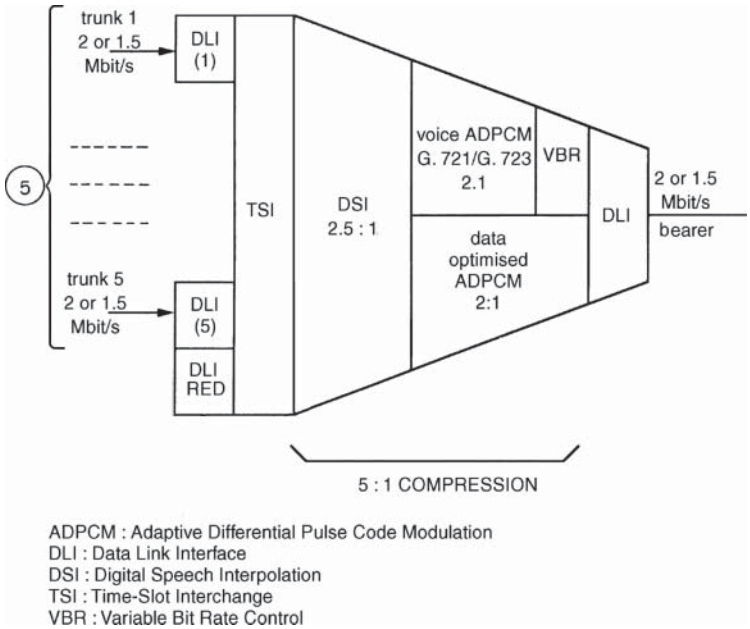


Figure 8.37 Organisation of digital circuit multiplication equipment (DCME).

distinguish between speech and data signals. The DSI equipment combines, typically, 150 terrestrial channels into 62 bearer channels. However, at a given time, the number of active voice samples delivered by the 150 terrestrial channels may exceed 62. In that case, the equipment is able to handle up to 96 simultaneous samples using bit stealing techniques.

- *Adaptive differential pulse code modulation (ADPCM):* These adaptive differential encoders use an appropriate coding algorithm that meets ITU-T Rec. G.721, and G.723. The encoders regroup the 62 bearer channels into a bitstream at 2.048 Mbit s^{-1}.
- *Variable bit rate (VBR):* Under normal conditions, PCM speech signal samples of A Law (a-law) or μ Law (u-law) are encoded in four bits. When the DSI equipment delivers more than 62 simultaneous samples, additional encoders are activated in order to create temporary bearer channels. All the encoders must then share the output bit rate, which cannot exceed the maximum value of 2.048 Mbit s^{-1}. The encoders of certain channels (other than those carrying data) then operate on three bits instead of four in accordance with a VBR procedure. These channels are selected randomly from one sample to the next. Data signals transmitted within the bandwidth of a telephone channel are processed in specially optimised encoders and coded at 32 kbit s^{-1}.
- *Output data link interface:* This equipment realises the interface between the output of the ADPCM encoders at 2.048 Mbit s^{-1} and a standard 2.048 Mbit s^{-1} (or 1.544 Mbit s^{-1}) bearer.

8.6.3.2 Circuit multiplication gain

The *circuit multiplication gain* is the ratio of the number of input terrestrial channels to the number of DCME output bearer channels. A typical value is five. In the case where the terrestrial telephone channels are distributed over a wide geographical area and, in particular, include different time zones, traffic peaks are spread over a period of time, and the probability of having a large number of channels active at the same time is low. It is thus possible to increase the number of terrestrial channels connected to the equipment, for example up to 240 channels, which are transmitted in the available 62 bearer channels. The speech interpolation gain thus becomes four by taking advantage of the different times of activity among the 240 channels; the overall circuit multiplication gain under these conditions may reach a value of eight.

A DCME is commonly used at both ends of a point-to-point satellite link. Point-to-multipoint operation is also possible.

The multiple-destination operation uses the ability of a satellite to deliver a carrier to several earth stations simultaneously. The DCME specified in IESS-501 can operate with up to four destinations. The bearer can be configured as a single pool of channels, shared by all DCME traffic, or can be segregated into two separate pools. The single-pool case, with more than one destination, constitutes the *multidestination mode* of operation, whereas the two-pool case, with one destination per pool, is called the *multiclique mode* of operation.

8.6.3.3 Multidestination mode of operation

In the multidestination mode, the DCME installed in the earth station is capable of the following:

- On the transmission side, the transmitted bearer capacity is shared among all the destinations, with indication of the destination of each sample so that the receiving equipment at each destination can identify the samples destined for it and reconstruct the corresponding telephone channels.

- On the receiving side, the telephone channels destined for the earth station are extracted from the incoming 2.048 Mbit s^{-1} bearer coming from other stations.

Note that the DCME generates one transmitted bearer and processes several received bearers. Bearers at 2.048 Mbit s^{-1} are routed either directly on IDR carriers or within bursts of a TDMA frame. Telephone channels on the terrestrial link to the switching centre are not concentrated, and the bit rate on this link is thus greater than the rate transmitted on satellite links. For example, a terrestrial link at 8.448 Mbit s^{-1} will be necessary to carry traffic exchanged on a 2.048 Mbit s^{-1} satellite link with four other locations.

8.6.3.4 Multiclique mode of operation

In multiclique mode, the samples transmitted on the bearer are arranged in several pools (*cliques*) within the bearer frame at 2.048 Mbit s^{-1}. Each clique is associated with a particular destination and contains its own assignment information from the DSI process. The Intelsat system uses up to two cliques per frame.

This approach enables the DCME to be located in the switching centre associated with the earth station. Hence the terrestrial link between the switching centre and the earth station also benefits from the circuit multiplication gain, which may, however, be smaller due to the reduced number of channels per clique.

The following operations are performed at the earth station:

- On the transmission side, the bearer, constituted of several cliques, directly modulates the multidestination carrier (IDR carrier) or is transmitted in a sub-burst of the TDMA frame.
- On the receiving side, the bearer received from a particular station contains several cliques (two in the case of Intelsat), of which only one is destined for the receiving station. Only that clique is retained and multiplexed with those from other stations to form the DCME received bearer, which is routed to the switching centre. The operation is simple to realise since the destination of cliques is known from their position in the frame. This operation is realised by dedicated equipment (the *clique sorting facility* [CSF]), in the case of IDR transmissions, or by terrestrial network interface equipment (digital non-interpolated [DNI] and direct digital interface [DDI]) in the case of TDMA operation. This permits the terrestrial link to take advantage of the circuit multiplication gain.

8.6.4 Equipment specific to SCPC transmission

With SCPC, dedicated equipment may provide activation of the carrier by speech and companding of the signal.

The use of a *speech detector* permits activation of the transmitted carriers only when speech is present on the channel concerned. This permits a reduction in the number of carriers passing through the satellite transponder at a given instant and hence a higher EIRP per carrier and less intermodulation noise.

8.6.5 Ethernet port for IP network connections

In new satellite modems, the internet protocols (IPs) are integrated as part of the indoor unit of the satellite earth stations. A modem can be configured to function as a LAN bridge or an IP

gateway through its Ethernet ports. Typically, two RJ45 connectors are provided for the Ethernet ports.

8.6.5.1 Bridge functions

When the modem is configured as an Ethernet bridge, all the terminals (such as laptops or mobile devices) are connected to the same IP subnet. Essentially, the bridge transport IP packets over satellite transparently without any processing of the IP header fields. Consequently, for simple point-to-point communications, little or no user set up is required to support IP over satellite. If the Ethernet ports are configured to be part of the bridge, then a single Ethernet connection to the modem can be used for both IP traffic and modem monitoring and control (M&C). The bridge functions maintains information on how to forward frames based on replies that are received from each terminal device in the network.

8.6.5.2 IP routing and functions

The satellite modem can also be configured as an IP gateway. Then one Ethernet port is dedicated to IP traffic and the other to M&C. One of the Ethernet ports can be configured for M&C outside the bridge functions. To communicate with the modem itself for M&C purposes, an IP address and subnet mask need to be set. The IP address can be set manually or using a Dynamic Host Control Protocol (DHCP) server on the subnetwork (subnet). Static routes are supported, allowing routing decisions to be made based on a set of explicit routing rules. Dynamic routing needs support from the standard routing protocols, such as routing information protocol (RIP) and open shortest path first (OSPF).

A default gateway IP address is provided to the modem and provides a next-hop IP address for all destinations that are not on the local subnet. This is usually the address of a router that has been set up to forward packets to the correct network. The bridge maintains information on how to forward frames based on replies that are received from each device in the network.

8.6.5.3 IP addressing

Each Ethernet port can have its own IP address. Two Ethernet ports can be bridged together, acting as a two-port switch. In bridging mode, IP addresses are not used. In routing mode, a single IP address covers both ports in the same subnet. If the M&C port is out of the bridge (i.e. the traffic port has its own IP address), then the IP traffic port and M&C port must be on different subnets.

8.6.5.4 Satellite gateways

When one TCP/IP stack is running on the base modem, there can only ever be one gateway associated with the TCP/IP stack in operation. The modem can be a M&C gateway, an IP traffic gateway with the port connected to user devices, or a satellite gateway with the port connected to satellite links.

8.6.5.5 IP traffic throughput performance

IP traffic throughput performance depends on a number of factors including one-way or two-way traffic, packet size, data rates, and the mixture of IP features. It is good practice to put a switch (or router) between the modem and local network in order to minimise the number of packets the modem has to process, as incidental network traffic (not intended for the satellite) has the potential to push the modem over its packet-processing limit. All satellite modems have TCP acceleration – performance enhancement function (PEP) functions to the maximum data rate of the modem. Header-compression functions on the IP traffic also reduce packet size.

8.6.5.6 Protocol header compressions

Internet standards for robust header compression (ROHC) include these different modes:

— Compression of IP packet header and user datagram packet (UDP) header together
— Compression of IP header, UDP header, and real-time transport protocol (RTP) headers
— Compression of Ethernet headers when transported over satellite

The 40 bytes of IP, UDP, and RTP headers are typically compressed to between 1 and 3 bytes. The headers of TCP packets can also be compressed. When the Ethernet header is compressed, the 14 bytes of the Ethernet frame (the Ethernet CRC is not sent over satellite even when compression function is not used) are typically reduced to 1 byte.

8.6.5.7 IP connectivity modes

Satellite modems can be used for the following connection modes:

— *Point-to-point mode*: One modem is transmitting to and receiving from one other modem (i.e. there is a direct satellite return path).
— *Point-to-multipoint mode*: One hub modem is transmitting to several remote modems. The remote modems may be Rx only or may transmit back to Rx-only modems at the hub that are daisy-chained together to the hub Tx modem (to allow all of the hub modems to share the hub Tx carrier).
— *Mesh network mode*: A number of remote sites each have one Tx carrier that is used to communicate with the other sites. Each site also has one Rx-modem for every site, to allow it to receive from each of the other sites.

8.7 MONITORING AND CONTROL; AUXILIARY EQUIPMENT

Monitoring of correct operation and control of the earth station are the purposes of a dedicated subsystem. Several specifications are given in this section concerning the earth station electrical power supply that is included in the auxiliary equipment.

8.7.1 Monitoring, alarms, and control (MAC) equipment

The monitoring, alarm, and control equipment of the earth station has the following purposes:

- To provide operators with the necessary information for monitoring and controlling the station (this includes measured parameters, equipment in service, switch positions, etc.) and managing traffic
- To initiate alarms in case of incorrect operation or an incident affecting the main station equipment or link performance and permit identification of the equipment involved
- To permit control of the station equipment, including bringing equipment into service, adjusting parameters, switching redundant equipment, and so on

M&C functions can be provided locally, in a centralised manner, or under the control of a computer. Locally, the functions are provided on the equipment itself by means of warning lights, indicators, and control push-buttons. With centralised control, the various functions are combined at one control centre. All the monitored parameters are available at this centre and are presented to the operator by means of various display devices (such as screens, indicators, and warning lamps). Control of a variety of equipment is possible from the console. This control and monitoring centre is situated at some distance from the equipment.

The next step is to transfer all the monitoring operations to a computer that records the parameters, selects the most important parameters for display on a standard screen, detects abnormal situations, prepares special commands, and executes them either automatically without human intervention (such as bringing replicated equipment into operation) or after approval by the operator.

With centralised or computer-aided management, it is possible to have a station without permanent staff; M&C information can be routed to a distant common network control centre by means of dedicated terrestrial lines or service channels on the satellite links.

8.7.2 Electrical power

Electrical energy is necessary for operation of the earth station equipment. This energy is obtained in most cases from the national energy distribution grid. According to the specified availability requirement, it is often necessary to take precautions against interruption of this source of energy. Three types of energy are generally available to an earth station, as follows:

- *Uninterruptible* power feeds all equipment that must operate without interruption, such as RF communications equipment, emergency lighting, etc.
- *Standby* power feeds devices that tolerate supply interruptions that can last for several minutes (such as antenna servos).
- Power *without standby* supplies noncritical circuits that can tolerate interruptions of several hours (such as air conditioning, antenna de-icing, etc.). It is generally possible, in the case of prolonged power failures, to be able to supply, on demand from the standby supply, some circuits that are normally without standby.

Energy without standby is provided by the national power grid (the *mains*). This also applies to standby energy when the mains is available. The uninterruptible supply is realised using

batteries that continuously supply the equipment concerned either directly as DC or by way of a converter if an alternating current is necessary. The national power grid provides float charging of the batteries by means of rectifiers. In the case of power failure, an electrical generator is started up automatically. This generator feeds the equipment connected to the standby energy circuits. It also replaces the mains to ensure that the batteries remain charged; the rectifier supply circuits are automatically disconnected from the mains when it fails. At the end of the failure, the generator stops, and the circuits using standby energy are switched to the mains.

8.8 CONCLUSION

From the start of the satellite communication era, earth stations have developed continuously, although the general organisation of the stations has remained unchanged. This development has been evidenced by a reduction in the size of earth stations. The diameter of antennas, initially more than 30 m, can now in some cases be as small as a few tens of centimetres. This is due to the increase of EIRP of communication satellites in association with the use of high-performance transmission techniques. This reduction is also evident in the size of the equipment used in the stations and has been made possible by the use of digital techniques and large-scale integration of components.

Use of these technologies has also enabled the processing capacity and complexity of equipment to be greatly increased. This has resulted in an increase in performance. In this way, the use of sophisticated transmission techniques, such as TDMA, spread-spectrum transmission, high-order modulation, error-correcting coding, and so on, has been made possible. Much greater ease of operation and maintenance has resulted from the use of these technologies for equipment design. For example, at the frequency-translation stage, programmable frequency synthesisers permit rapid carrier-frequency selection and high stability of the displayed frequency. Monitoring under computer control ensures continuous checking of the operation of diverse systems and rapid detection of faulty equipment, and even its replacement by replicated equipment.

Simultaneously, the appearance of new systems has permitted better exploitation of the characteristics of satellites, such as broadcasting capacities and the possibility of access to widespread users without additional cost. These systems open up the possibility of numerous telecommunication services in areas as varied as business communication, rural telecommunication, video data distribution, data broadcasting, Internet access, interactive transfers, and communication with mobiles, particularly recent development of high-speed broadband multimedia Internet services.

Many of these systems make use of small earth stations that are installed on the user's premises and provide direct telephone links (rural communication), data communications with very small aperture terminals (VSATs) on private networks, Internet access, and video reception. For communication with mobiles, mass and power constraints, as well as the use of tracking or omnidirectional antennas, should be taken into consideration.

REFERENCES

[CAM-76] Campanella, S.J. (1976). Digital speech interpolation. *COMSAT Technical Review* **6** (1): 127–157.

[CCIR-90] CCIR. (1990). Earth-station antennas for the fixed-satellite service. Report 390.

[DAN-85] Dang, R., Watson, B.K., and Davis, I. (1985). Electronic tracking systems for satellite ground stations. In: *15th European Microwave Conference*, 681–687. IEEE.

[DUR-87] Durwen, E.J. (1987). *Determination of Sun Interference Periods for Geostationary Satellite Communication Links*, 183–195. Elsevier Science.

[EDW-83] Edwards, D.J. and Terrell, P.M. (1983). The smoothed step-track antenna controller. *International Journal of Satellite Communications* **1**: 133–139.

[ETSI-14] ETSI. (2014). Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications; part 1: DVB-S2. EN 302 307-1 V1.4.1.

[ETSI-15] ETSI. (2015). Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications; part 2: DVB-S2 extensions (DVB-S2X). EN 302 307-2 V1.1.1.

[FOR-89] Forcina, G., Oei, W.S., Oishi, T., and Phiel, J. (1989). Intelsat digital circuit multiplication equipment. In: *Proceedings of the ICDSC 8th International Conference on Digital Satellite Communications, Pointe à Pître*, 795–803.

[GAR-84] Garcia, H. (1984). Geometric aspects of solar disruption in satellite communications. *IEEE Transactions on Broadcasting* **BC-30** (2, 49): 44.

[GIL-86] Gilmour, A.S. Jr. (1986). *Microwave Tubes*. Artech House.

[HAW-88] Hawkins, G.J. et al. (1988). Tracking systems for satellite communications. *IEE Proceedings* **135** (5): 393–407.

[HO-61] Ho, H.C. (1961). On the determination of the disk temperature and the flux density of a radio source using high gain antennas. *IRE Transactions on Antennas and Propagation*: 500–510.

[ITUR-02] ITU-R. (2002). Impact of interference from the sun into a geostationary-satellite orbit fixed satellite service link. S.1525-1.

[ITUR-04] ITU-R. (2004). Radiation diagrams for use as design objectives for antennas of earth stations operating with geostationary satellites. S.580.

[ITUR-06] ITU-R. (2006). Maximum permissible levels of off-axis e.i.r.p. density from earth stations in geostationary-satellite orbit networks operating in the fixed-satellite service transmitting in the 6, 13, 14 and 30 GHz frequency bands. S.524-9.

[ITUR-10] ITU-R. (2010). Reference radiation pattern for earth station antennas in the fixed-satellite service for use in coordination and interference assessment in the frequency range from 2 to 31 GHz. S.465.

[ITUT-02] ITU-T. (2002). Series G: transmission systems and media, digital systems and networks, digital networks – quality and availability targets, end-to-end error performance parameters and objectives for international, constant bit-rate digital paths and connections. G.826.

[JOH-84] Johnson, R.C. and Jasik, H. (1984). *Antenna Engineering Handbook*. McGraw-Hill.

[KEP-89] Kepley, W.R. and Kwan, A. (1989). DSI development for 16 kbit/s voice systems. In: *ICDSC 8th International Conference on Digital Satellite Communications, Pointe à Pître*, 551–559.

[KRE-80] Kreutel, R.W. and Potts, J.B. (1980). The multiple-beam Torus earth stations antennas. In: *International Conference on Communications ICC 80, Seattle*, 25.4.1–25.4.3. IEEE.

[LOE-83] Loeffler, J. (1983). Planning for solar outages. *Satellite Communications*: 38–40.

[LUN-70] Lundgren, C.W. (1970). A satellite system for avoiding serial sun-transit outages and eclipses. *Bell Technical Journal* **49** (8): 1943–1957.

[MOH-88] Mohamadi, F., Lyon, D., and Murrell, P. (1988). Effects of solar transit on Ku-band Vsat systems. *International Journal of Satellite Communications* **6**: 65–71.

[RAU-85] Rauthan, D.B. and Garg, V.K. (1985). Geostationary satellite signal degradation due to sun interference. *Journal of Aeronautical Society of India* **37** (2): 137–143.

[RIC-86] Richaria, M. (1986). Design considerations for an earth station step-track system. *Space Communications and Broadcasting* **4**: 215–228.

[SHI-71] Shimbo, O. (1971). Effects of intermodulation AM-PM conversion and additive noise in multicarrier TWT systems. *Proceedings of the IEEE* **59**: 230–238.

[SHI-68] Shimbukuro, F. and Tracey, J.M. (1968). Brightness temperature of quiet sun at centimeter and millimeter wavelengths. *The Astrophysical Journal* **6**: 777–782.

[TOM-70] Tom, N. (1970). Autotracking of communication satellite by the steptrack technique. In: *Proceedings of the IEE Conference on Earth Station Technology*, 121–126.

[VUO-83a] Vuong, X.T. and Forsey, R.J. (1983). C/N degradation due to sun transit in an operational communication satellite system. In: *Proceedings of the Satellite Communication Conference SCC-83, Ottawa*, 11.3.1–11.3.4.

[VUO-83b] Vuong, X.T. and Forsey, R.J. (1983). Prediction of sun transit outages in an operational communication satellite system. *IEEE Transaction Broadcasting* **BC-29** (4): 134–139.

[WAT-86] Watson, B.K. and Hart, M. (1986). A primary-feed for electronic tracking with circularly-polarised beacons. In: *Proceedings of the Military Microwaves Conference, Brighton*, 261–266.

[YAT-89] Yatsuzuka, Y. (1989). A design of 64 kbps DCME with variable rate coding and packet discarding. In: *Proceedings of the ICDSC 8th International Conference on Digital Satellite Communications, Pointe à Pître*, 547–551.

9 THE COMMUNICATION PAYLOAD

This chapter is devoted to a description of the satellite payload with the emphasis on design principles, characteristic parameters, and the technologies used for the equipment.

With most commercial communication satellites, the payload consists of two distinct parts with well-defined interfaces – the *repeater* and the *antennas*. This dichotomy is less clear with active antennas that closely associate the radiating elements and the amplifiers. For clarity of presentation, active antennas are considered in the parts of the chapter devoted to antennas, after amplifier technology is discussed in the sections pertaining to the repeater.

In this work, the word *repeater* designates the electronic equipment that performs a range of functions on the carriers from the receiving antenna before delivering them to the transmitting antenna. The repeater usually encompasses several channels (also called *transponders*) that are individually dedicated to sub-bands within the overall payload frequency band. The architecture differs for *single-beam transparent* repeater (Section 9.2), *regenerative* repeater (Section 9.3), and *multibeam* payload (Section 9.4). The functions and characteristic parameters of the payload are presented first. All of these are fundamental technologies for modern satellite communication systems and networks.

9.1 MISSION AND CHARACTERISTICS OF THE PAYLOAD

9.1.1 Functions of the payload

The main functions of the communications payload of a satellite are as follows:

— To capture the carriers transmitted, in a given frequency band and with a given polarisation, by the earth stations of the network. (The stations are situated within a given region [*service zone*] on the surface of the earth and are seen from the satellite within an angle that determines the angular width of the satellite antenna beam. The intersection of the satellite antenna beam with the surface of the earth defines the *receive coverage*.)
— To capture as little interference as possible. (The interference is a carrier originating from a different region or not having the specified values of frequency or polarisation.)

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

— To amplify the received carriers while limiting noise and distortion as much as possible. (The level of the received carrier is on the order of a few tens of picowatts.)
— To change the frequency of the carriers received on the uplinks to that on the downlinks (for example, from 14 to 11 GHz for Ku band and from 30 to 20 GHz for Ka band).
— To provide the power required in a given frequency band at the interface with the transmitting antenna (the power to be provided ranges from tens to hundreds of watts).
— To radiate the carriers in a given frequency band and with a given polarisation (which are characteristic of the downlink antenna beam) to a given region (service zone) on the surface of the earth. The intersection of the transmit antenna beam with the surface of the earth defines the *transmit coverage zone*.

These functions are to be realised regardless of the organisation of the payload. For a multi-beam satellite, an additional function is to route the carriers from any given uplink beam to any downlink beam. The regenerative repeater must also provide demodulation and remodulation of the carriers.

The band of frequencies allocated to the repeater can be as large as from hundreds of megahertz to several gigahertz. To facilitate power amplification, this band is usually divided into a number of sub-bands (channels or transponders) with which separate amplification chains are associated. The bandwidth of these channels is typically on the order of several tens of megahertz.

9.1.2 Characterisation of the payload

The characteristic parameters of a communication satellite payload are as follows:

— The transmitting and receiving frequency bands and polarisations for the various repeater channels
— The transmit and receive coverages
— The effective isotropic radiated power (EIRP) or the power flux density achieved in a given region (satellite transmit coverage)
— The power flux density required at the satellite receiving antenna in order to produce the performance specified at the repeater channel output (this can depend on the channel or group of channels concerned)
— The figure of merit (G/T) of the receiving system in a given region (satellite receive coverage)
— The nonlinear characteristics
— The reliability after N years for a specified number (or percentage) of channels in working order

Antenna coverage is specified in terms of the radio-frequency (RF) characteristics to be obtained at a set of reference points on the surface of the earth, which define the contour of the so-called *service zone*. The beamwidth is obtained by taking account of the various sources of antenna beam depointing and is specified by the permissible gain reduction (often taken as 3 dB) at the boundary of the coverage area (see Sections 9.7 and 9.8). The receive and transmit coverages are considered differently in general.

The EIRP or *power flux density* produced in a given region is generally specified at the edge of coverage (EOC) area under particular operating conditions for one repeater channel. It usually involves operation of the amplifier at saturation. It should be noted that there are regulatory limits to the power flux density produced on the surface of the earth by communication satellites (ITU-R Rec. SF.358).

The *minimum power flux density* at the satellite receiving antenna is defined for a given receive coverage and under particular operating conditions for the amplifier of the repeater channel (usually to obtain saturation of the amplifier).

The *figure of merit* (G/T) of the receiving system is also defined for a given receive coverage (for example, by a minimum value at the edge of coverage).

Characterisation of *nonlinearities* includes, for example, the level of third-order intermodulation products for operation with a given number of carriers of the same amplitude and particular values of output back-off (OBO) (see Section 9.2.1).

Reliability is the subject of Chapter 13.

9.1.3 The relationship between the radio-frequency characteristics

The principal parameters that characterise the payload from the point of view of the link budget are the EIRP for the downlink and the figure of merit (G/T) for the uplink. Although they characterise different links, these parameters are not independent. By considering, for simplicity, the case of a station-to-station link through the payload in the absence of interference and in single access, the carrier power-to-noise power spectral density ratio $(C/N_0)_T$ for the overall link can be written (see Eq. (5.70)):

$$(C/N_0)_T^{-1} = (C/N_0)_U^{-1} + (C/N_0)_D^{-1}(\text{Hz}^{-1})$$

In this expression, $(C/N_0)_U$ is the carrier-to-noise density ratio on the uplink, and this is proportional to the figure of merit G/T of the satellite. $(C/N_0)_D$ is the carrier-to-noise density ratio that characterises the downlink alone and is proportional to the EIRP of the channel.

For a given performance objective that determines the value of $(C/N_0)_T$, the ratios $(C/N_0)_U$ and $(C/N_0)_D$, and hence the values of the figure of merit G/T and the EIRP, are combined by an expression of the following type:

$$C = A(G/T)^{-1} + B(EIRP)^{-1} \tag{9.1}$$

where A, B, and C are constants for a given configuration. This relationship is illustrated in Figure 9.1. As for earth stations, there is, therefore, the possibility of a trade-off in the choice of parameter values. By assuming that the gains of the receiving and transmitting antennas are fixed for a given coverage, the trade-off between the power P_{TX} delivered by the output amplifier and the system noise temperature T is finally established.

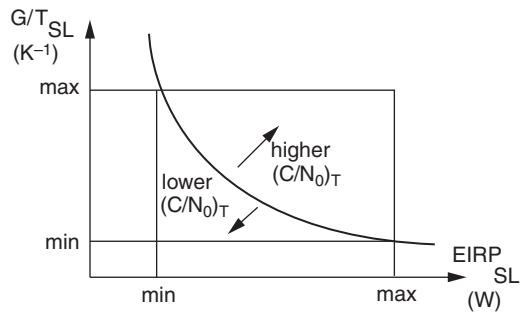


Figure 9.1 Satellite G/T versus EIRP for a given overall performance objective $(C/N)_T$.

For a specified performance objective, it is thus possible to compensate for an increase in system noise temperature by an increase in the power of the channel output amplifier (an exchange of power with noise temperature) in accordance with the constraints of power and noise figure limitation. This remains valid (with a more complex formulation) for a link with interference and intermodulation noise.

9.2 TRANSPARENT REPEATER

This part of the chapter is devoted to a presentation of the organisation and technologies of *transparent repeater* equipment associated with *single-beam* antennas on both the receiving and the transmitting sides. This configuration is often referred as a *bent-pipe* repeater in the literature. There is, therefore, only one antenna port for both directions (one input on the receiving side and one output on the transmitting side). All the earth stations are situated in the same coverage region, which makes the organisation of the network simple (no need for switching), as only the sharing of the satellite resource among stations should be considered, as discussed in Chapter 6. Since the organisation of the repeater is largely determined by the problems arising from equipment nonlinearities, characterisation of these nonlinearities is presented first.

9.2.1 Characterisation of nonlinearities

The payload equipment demonstrates nonlinear characteristics. The behaviour of the equipment thus depends on the carrier level applied at the input. This applies particularly to equipment that uses active components such as travelling wave tubes (TWTs), klystrons, and transistors. However, it also applies under certain conditions, notably at high power levels, to passive equipment such as filters and antennas; in this case, one refers to passive intermodulation (PIM) products [HOE-86; TAN-90].

The aim of the following sections is to define various parameters that are currently used. Polynomial modelling is introduced first, although it is imperfect as it deals only with amplitude, since it permits most of the principal phenomena to be easily illustrated. More elaborate models, taking into account the effects of both amplitude and phase, are then presented.

9.2.1.1 Polynomial (amplitude only) modelling of an amplifier

One of the main functions of the repeater is amplification of the carrier power level. The input–output characteristic of an ideal amplifier is merely a coefficient of proportionality. In practice, the output voltage, particularly at high levels, does not vary in proportion to the amplitude of the input signal. Various models of this phenomenon can be devised; one of the simplest is to consider the instantaneous amplitude S_o of the output carrier as a polynomial function of the instantaneous amplitude S_i of the input carrier:

$$S_o = aS_i + bS_i^3 + cS_i^5 + \cdots \quad \text{(V)} \tag{9.2}$$

where a, b, c, etc. are constants. These constants, and the order of the polynomial (only odd powers are necessary if only odd intermodulation products are considered; see Section 9.2.1.3), are selected to represent the actual characteristic of the amplifier as closely as possible.

Nonlinear phenomena also affect the phase of the output carrier, and this depends on the amplitude of the input carrier. These phenomena are not taken into account in polynomial modelling. The relative phase variation $\Delta\phi$ as a function of input power P_i can be modelled independently. For example [BER-71]:

$$\Delta\phi = a[1 - \exp(-bP_i)] + cP_i \tag{9.3}$$

where a, b, and c are constants chosen to fit the actual characteristic.

9.2.1.2 *Power transfer characteristic in single carrier operation*

If an unmodulated carrier, whose instantaneous amplitude is expressed in the form $S_i(t) = A \sin \omega_1 t$, is applied at the input of a device, expansion of the instantaneous amplitude of the output carrier S_o using Eq. (9.2) produces a sum of terms; one has angular frequency ω_1, and the others consist of harmonics with frequencies that are multiples of ω_1. In the applications considered, the bandwidth of the equipment is less than the nominal frequency. The harmonics are thus eliminated by filtering. The output power, measured across a 1 Ω impedance, for the carrier is obtained by taking half of the square of the amplitude of the resulting term of angular frequency ω_1. Hence, with polynomial modelling:

$$P_{o1} = (1/2)(aA + 3bA^3/4 + 15cA^5/24 + \cdots)^2 \quad \text{(W)} \tag{9.4}$$

where P_{o1} designates the output power (o = output) in single carrier operation (1 = single).

Introducing the input signal power $P_{i1} = A^2/2$ (the powers are defined across 1 Ω loads) gives:

$$P_{o1} = P_{i1}[a + (3b/2)P_{i1} + (15c/6)(P_{i1})^2 + \cdots]^2 \quad \text{(W)} \tag{9.5}$$

This relation constitutes the power transfer characteristic, which thus represents the output power P_{o1} at the carrier frequency as a function of the carrier input power P_{i1}.

The curve representing this characteristic has a maximum for a particular value $(P_{i1})_{sat}$ of carrier power applied at the input [BAU-85]. This maximum corresponds to the *saturation output power* (*in single carrier operation*) $(P_{o1})_{sat}$. The saturation power in single carrier operation is the value used to characterise an amplifier (e.g. TWTs or klystrons) in the manufacturer's data sheet.

The saturation output power $(P_{o1})_{sat}$ is related to the corresponding input power $(P_{i1})_{sat}$

$$(P_{o1})_{sat} = G_{sat}(P_{i1})_{sat} \quad \text{(W)} \tag{9.6}$$

where G_{sat} is the *saturation power gain of the device*.

Normalised characteristics: input and output back-off. A particular operating point (Q) of the amplifier is characterised by the pair of input and output powers $(P_{i1}, P_{o1})_Q$. It is convenient to normalise these quantities with respect to the saturation output power $(P_{o1})_{sat}$ and the input power $(P_{i1})_{sat}$ required to obtain saturation, respectively.

The normalised characteristic thus relates the magnitude $Y = P_{o1}/(P_{o1})_{sat}$ and the magnitude $X = P_{i1}/(P_{i1})_{sat}$. Figure 9.2a shows such a characteristic that relates the Y and X values in decibels for a typical TWT. For a particular operating point defined by $(P_{i1}, P_{o1})_Q$, $(X)_Q$, and $(Y)_Q$ represent the *input back-off* (IBO) and the OBO, respectively (see Section 5.9.1.4).

A simplified model of the normalised characteristic can be obtained by considering a limited expansion in Eq. (9.5). Differentiating with respect to the input power and setting this derivative to zero for $P_{i1} = (P_{i1})_{sat}$ to give the saturation power results in:

$$P_{o1} = \frac{G_{sat}\, P_{i1}}{4}\left[3 - \frac{P_{i1}}{(P_{i1})_{sat}}\right]^2 \quad \text{(W)} \tag{9.7a}$$

Normalisation in this simplified case leads to:

$$Y = (X/4)(3 - X)^2 \tag{9.7b}$$

Amplitude modulation to amplitude modulation (AM/AM) conversion coefficient. For small X, Eq. (9.7b), with values expressed in decibels, reduces to $(Y)_{dB} = (X)_{dB} + \text{constant}$. The slope of the characteristic (in dB) in Figure 9.2a is thus equal to 1: that is, for 1 dB variation of the input power, the output power also varies by 1 dB (in the linear region). The slope of the

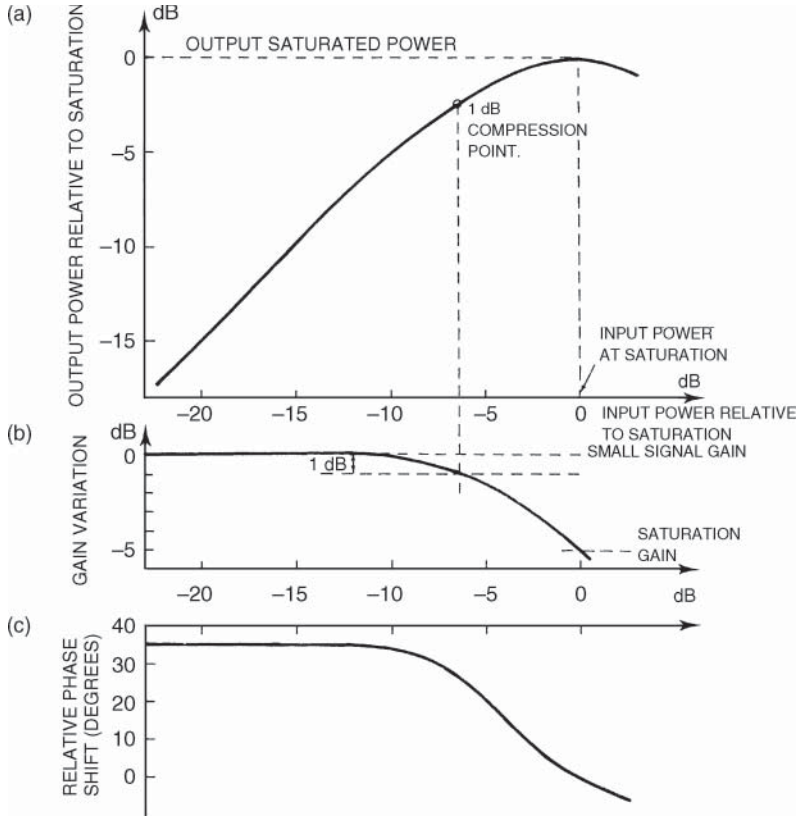


Figure 9.2 Normalised characteristics as a function of input back-off (IBO): (a) amplifier power transfer in single carrier operation; (b) power gain; (c) relative phase shift between input and output.

characteristic is called the *AM/AM conversion coefficient* and is expressed in dB per dB. This conversion coefficient, therefore, has a value of unity when the absolute value of back-off is large (for example, an OBO of less than −15 dB with a TWT). The AM/AM conversion coefficient decreases as input power increases up to saturation where it becomes zero.

Power gain. The ratio of output power P_o to input power P_i is the *power gain*. It is constant in the linear part of the characteristic, which corresponds to a low power level, where it is called the *small signal power gain* G_{ss}. The gain then decreases with the approach of saturation, as illustrated in Figure 9.2b. At saturation, it takes the value G_{sat}, *the saturation power gain of the device*.

Point of compression to 1 dB. The output power obtained when the actual characteristic deviates by 1 dB from an extension of the linear part defines the *1 dB compression point*. This point corresponds to a reduction of 1 dB in power gain.

This parameter is used to define the part of the characteristic that can be considered to be linear. In order to obtain quasi-linear operation for an amplifier module, a signal level greater than a value defined with respect to the point of compression to 1 dB (for example, 10 dB) must be prohibited. The point of compression to 1 dB is often used in the manufacturer's technical data sheet to characterise the power of a solid state amplifier (with these amplifiers, driving to saturation power may damage the device).

Amplitude modulation to phase modulation (AM/PM) conversion factor K_P. The effect of nonlinearity also appears in the phase of the signal. The device introduces phase shift between the input and the output. The relative variation of phase shift with respect to that corresponding to saturation as a function of input signal level is illustrated in Figure 9.2c. The slope of this characteristic, called the *AM/PM conversion factor* K_P, is given by:

$$K_P = \Delta\phi/\Delta P_{i1} \quad (^\circ/\text{dB}) \tag{9.8}$$

The conversion factor, expressed in degree/dB (°/dB), is maximum for a value of input power less than the saturation value by a few dB.

9.2.1.3 *Power transfer characteristic in multicarrier operation*

For multicarrier operation, the signal applied at the input of the device is now considered as the sum of sinusoidal unmodulated carriers. It is, therefore, put in the form:

$$S_i = A \sin \omega_1 t + B \sin \omega_2 t + C \sin \omega_3 t + \cdots \quad \text{(V)}$$

Expansion of the instantaneous amplitude of the output signal S_o using Eq. (9.2) results in the appearance of components at the input angular frequencies (ω_1, ω_2, etc.) and at frequencies corresponding to linear combinations of these frequencies (*intermodulation products*). These intermodulation products have been defined in Section 6.5.4. Only odd intermodulation products occur in the vicinity of the input frequencies. On the other hand, the amplitude of these intermodulation products decreases with their order. The most troublesome are third-order intermodulation products at frequencies $2f_i - f_j$ and $f_i + f_j - f_k$.

In the case of n unmodulated carriers of equal amplitude A_i, a general expression for the amplitude A_{on} of each of the n components of the output signal at the input frequencies (f_1, f_2, etc.) is given by [PRI-93], pp. 414–415]:

$$A_{on} = aA_i[1 + (3b/2a)(n - 1/2)A_1^2 + (15c/4a)(n^2 - 3n/2 + 2/3)A_i^4 + \cdots] \quad \text{(V)} \tag{9.9}$$

Under the same assumptions, the amplitude $A_{IM3,n}$ of the third-order intermodulation products at frequencies $2f_i - f_j$ is given by:

$$A_{IM3,n} = (3b/4)A_i^3\{1 + (2c/6b)A_i^2[(25/2) + 15(n - 2)] + \cdots\} \quad \text{(V)} \tag{9.10}$$

and the amplitude $A'_{IM3,n}$ of the third-order intermodulation products at frequencies $f_i + f_j - f_k$ is given by:

$$A'_{IM3,n} = (3b/2)A_i^3\{1 + (10c/2b)A_i^2[(3/2) + (n - 3)] + \cdots\} \quad \text{(V)} \tag{9.11}$$

9.2.1.4 *Characterisation of nonlinearities using two unmodulated carriers of equal amplitude*

To characterise a device, it is customary to consider two input unmodulated carriers of equal amplitude ($A = B$). The power P_{o2} of one of the output components at frequency f_1 or f_2 can then be expressed as a function of the power ($P_{i2} = A^2/2$) of one of the two input carriers. The curve representing output power variation for one of the two carriers as a function of the power of one of the two input carriers can be plotted after normalising the output and input magnitudes with respect to $(P_{o1})_{sat}$ and $(P_{i1})_{sat}$ respectively (refer to Figure 9.3).

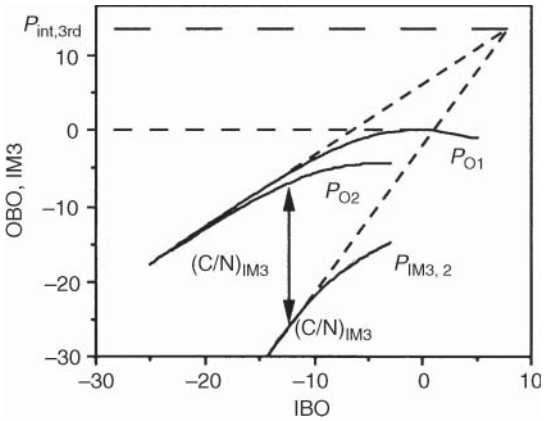


Figure 9.3 Normalised power transfer characteristic with two equal amplitude carriers.

In comparison with the curve representing single carrier operation, saturation corresponds to an output power less than $(P_{o1})_{sat}$. The maximum power that the device can deliver is effectively shared between the two carriers and the various intermodulation products. For a TWT, the difference between the maximum power in single-carrier operation and the maximum power of one of the two carriers is on the order of 4–5 dB (see Figure 9.3).

The effect can be illustrated using the polynomial model of Eq. (9.2). By considering two input carriers of equal amplitude A and frequencies f_1 and f_2, Eq. (9.2) leads to:

$$P_{o2} = P_{i1}[a + (9b/2)P_{i2} + (15c/6)(P_{i2})^2 + \cdots]^2 \quad \text{(W)} \tag{9.12}$$

where P_{o2} is the power of one of the output components at frequency f_1 or f_2 and $P_{i2} = A^2/2$ is the power of one of the input carriers.

The coefficients of this relation can be expressed as a function of the characteristic parameters for single carrier operation. By restricting the expansion in Eq. (9.12), as in the previous section, this gives:

$$(P_{o2}) = (9G_{sat}/4)(P_{i2})[1 - (P_{i2})/(P_{i1})_{sat}]^2 \quad \text{(W)} \tag{9.13a}$$

Normalising the output and input magnitudes with respect to $(P_{o1})_{sat}$ and $(P_{i1})_{sat}$, respectively, the following simplification is obtained:

$$Y' = X'(9/4)(1 - X')^2 \tag{9.13b}$$

where $Y' = P_{o2}/(P_{o1})_{sat}$ and $X' = P_{i2}/(P_{i1})_{sat}$ are the normalised magnitudes. Saturation is reached in this simplified case for $X' = 1/3$. The normalised output power is then $Y' = 1/3$.

Input and output back-off. For a particular operating point defined by $(P_{i2}, P_{o2})_{Q'}$, X' and Y' represent the *IBO* and the *OBO*, respectively. Notice that the back-offs are defined in multicarrier operation with respect to the saturation powers in single-carrier operation. In the example used for illustration, the input and output back-offs corresponding to saturation in two carrier operation are −5 dB ($X' = Y' = 1/3$).

The *total* (input or output) *back-off* may also be defined. It is the ratio of the sum of the powers (input or output) of the various carriers to the saturation power in single carrier operation. In the previous example, the total input and output back-off corresponding to saturation with two carriers would be −2 dB ($X'_T = Y'_T = 2 \times 1/3$). These concepts were introduced in Section 5.9.1.4.

Third-order intermodulation. The power $P_{IM3,2}$ of one of the third-order intermodulation products can also be plotted on the normalised diagram with respect to $(P_{o1})_{sat}$ and $(P_{i1})_{sat}$ (Figure 9.3). It can be seen that the linear part of this curve (for small values of $X' = IBO$) has a slope equal to 3 with a decibel scale. This can be verified from the simplified model. The power of the terms of frequency $2f_1 - f_2$ and $2f_2 - f_1$ in the expansion obtained from Eq. (9.2) is given by:

$$P_{IM3,2} = (P_{i2})^3[(3b/2) + (25c/2)(P_{i2}) + \cdots]^2 \quad \text{(V)} \tag{9.14}$$

where $P_{IM3,2}$ is the output power of one of the third-order intermodulation products for operation with two carriers of equal amplitude. The coefficients can be expressed as a function of the characteristic parameters of single carrier operation. By limiting the expansion of Eq. (9.14) to its first term, this gives:

$$(P_{IM3,2}) = (P_{i2})^3 G_{sat}/[(P_{i1})_{sat}]^2 \quad \text{(W)} \tag{9.15a}$$

Normalising the output and input magnitudes with respect to $(P_{o1})_{sat}$ and $(P_{i1})_{sat}$, this simplified case gives:

$$IM3 = (1/4)(X')^3 \tag{9.15b}$$

where $IM3 = P_{IM3,2}/(P_{i1})_{sat}$ and $X' = P_{i2}/(P_{i1})_{sat}$ are the normalised magnitudes. Equation (9.15b), with the values expressed in decibels, can be put in the form: $(IM3)_{dB} = 3(X')_{dB} + \text{constant}$. The slope of the characteristic (in dB) in Figure 9.3 is thus equal to three: that is, for 1 dB variation of the carrier input power, the power of one of the intermodulation products varies by 3 dB (in the linear region).

On the other hand, Eqs. (9.15b) do not show a reduction of the slope of the intermodulation product characteristic as the normalised input power increases. This originates in the limited expansion of Eq. (9.14), which is not sufficiently representative in the vicinity of saturation. By considering a larger expansion, an expression of the following form would be obtained:

$$IM3 = p(X')^3(q + rX')^2 \tag{9.16}$$

This shows a saturation effect when X' becomes large (p, q, and r are constants that are functions of the characteristic parameters of single carrier representation).

Relative level of third-order intermodulation products. The ratio $(C/N)_{IM3}$ of the output power of one of the two carriers to the power of one of the third-order intermodulation products for different values of IBO characterises the *relative level of the third-order intermodulation products*. A table of values is often provided in the technical data sheet of an amplifier to characterise the effect of the nonlinearity.

Third-order intercept point ($P_{int,3rd}$). Another parameter is currently used to characterise the effect of nonlinearity, particularly with solid-state devices. The straight lines obtained by extending the linear parts of the characteristics of the useful signal (one of the two carriers) and one of the third-order intermodulation products meet at a point called the *third-order intercept point*, $P_{int\,3rd}$.

The ordinate $(P_{int\,3rd})$ of this point (expressed as the value of the power) enables the linearity of several devices to be compared; the higher the value at the intercept, the more linear the device (for a given power). The value of the third-order intercept point is typically greater by about 10 dB than the value of the point of compression to 1 dB.

The third-order intercept point also enables the ratio $(C/N)_{IM3}$ to be determined for a given output power on one of the two carriers. By considering the linear part of Eqs. (9.13b) and (9.15b), expressed in decibels, the following is obtained:

$$(P_{o2})_{dB} = 10\log(9G_{sat}/4) + (P_{i2})_{dB} = K_1 + (P_{i2})_{dB} \quad \text{(dBW)}$$

and

$$(P_{IM3\,2})_{\mathrm{dB}} = 10\log\{G_{\mathrm{sat}}/[(P_{\mathrm{i}1})_{\mathrm{sat}}]^2\} + 3(P_{\mathrm{i}2})_{\mathrm{dB}} = K_2 + 3(P_{\mathrm{i}2})_{\mathrm{dB}} \quad (\mathrm{dBW})$$

The value of the intercept is such that:

$$P_{int\ 3\mathrm{rd}} = K_1 + (P_{\mathrm{i}2})_{\mathrm{dB}} = K_2 + 3(P_{\mathrm{i}2})_{\mathrm{dB}} \quad (\mathrm{dBW})$$

hence

$$P_{int\ 3\mathrm{rd}} = (3K_1 - K_2)/2 \quad (\mathrm{dBW})$$

Furthermore:

$$(P_{\mathrm{o}2})_{\mathrm{dB}} - P_{int\ 3\mathrm{rd}} = [(K_2 - K_1)/2] + (P_{\mathrm{i}2})_{\mathrm{dB}} \quad (\mathrm{dBW})$$

The difference (in decibels) between the output power on one of the two carriers and the power of one of the two third-order intermodulation products is then:

$$(P_{\mathrm{o}2})_{\mathrm{dB}} - (P_{\mathrm{IM3,2}})_{\mathrm{dB}} = [K_1 + (P_{\mathrm{i}2})_{\mathrm{dB}}] - [K_2 + 3(P_{\mathrm{i}2})_{\mathrm{dB}}] = K_1 - K_2 - 2(P_{\mathrm{i}2})_{\mathrm{dB}} \quad (\mathrm{dB})$$

from which:

$$(C/N)_{\mathrm{IM3}} = (P_{\mathrm{o}2})_{\mathrm{dB}} - (P_{\mathrm{IM3}})_{\mathrm{dB}} = 2[(P_{int\ 3\mathrm{rd}})_{\mathrm{dB}} - (P_{\mathrm{o}2})_{\mathrm{dB}}] \quad (\mathrm{dB}) \tag{9.17}$$

This relation is obtained by considering the linear parts of Eqs. (9.13b) and (9.15b) and is thus valid well below saturation (below the point of compression).

It should be noted that the values of $(C/N)_{\mathrm{IM3}}$ involved here are used to characterise the device (the amplifier) and correspond to operation of the amplifier in a specific mode (two unmodulated carriers of equal amplitude).

The link budget Eq. (5.75) requires the value of $(C/N_0)_{\mathrm{IM}}$. This value can be obtained as $(C/N_0)_{\mathrm{IM}} = (C/N)_{\mathrm{IM}}/B$, where B is the modulated carrier bandwidth (see Section 6.5.4.4). The value of $(C/N)_{\mathrm{IM}}$ for the considered link and nonlinear amplifier depends on the number of carriers, the distribution of carrier powers and frequencies, and the modulation scheme. $(C/N)_{\mathrm{IM}}$ can be obtained by experimental characterisation or by simulation based on modelling using, for instance, Bessel functions (see Section 9.2.1.7). This value generally differs from that of $(C/N)_{\mathrm{IM3}}$, discussed earlier.

The AM/PM conversion characteristic of the amplifier also causes generation of intermodulation products that are additional to those caused by amplitude nonlinearity. A degradation of several decibels of the $(C/N)_{\mathrm{IM}}$ ratio, as determined by considering only amplitude nonlinearity, can thus be introduced if the device is operated near saturation.

Transfer coefficient K_{T}. In multicarrier operation, nonlinear phase effects also cause transfer of the amplitude modulation of one carrier into phase modulation of other carriers. In the case of operation with two carriers, *the AM to PM transfer coefficient* K_{T} *from one carrier to another* is defined as the slope of the relative variation (with respect to the phase at saturation) of the phase of one output carrier (whose input amplitude is held constant) when the input amplitude of the other is varied.

9.2.1.5 The capture effect

Consider a nonlinear device in multicarrier operation for which the power of one of the input carriers is less than that of the other carriers and differs from them by a quantity ΔP_{i}. At the output, the difference ΔP_{o} between the power of the other carriers and that of the carrier considered is

increased. This phenomenon is called the *capture effect*. It can occur with two carriers and the simplified modelling considered earlier is used, by comparing the amplitudes of the output components when the amplitudes of the two input carriers are different. The input signal is of the form:

$$S_i = A \sin \omega_1 t + B \sin \omega_2 t \quad \text{(V)}$$

From Eq. (9.2), determination of the components $(A_{o2})_{\omega 1}$ and $(B_{o2})_{\omega 2}$ of the output signal at angular frequencies ω_1 and ω_2 gives:

$$(A_{o2})_{\omega 1} = A[a + (3b/4)A^2 + (3b/2)B^2]$$

hence

$$(P_{o2})_{\omega 1} = (P_{i2})_{\omega 1}\{1 + (3b/a)[(P_{i2})_{\omega 1}/2 + (P_{i2})_{\omega 2}]\}^2 \quad \text{(W)}$$

and

$$(B_{02})_{\omega 2} = B[a + (3b/4)B^2 + (3b/2)A^2]$$

hence

$$(P_{o2})_{\omega 2} = (P_{i2})_{\omega 2}\{1 + (3b/a)[(P_{i2})_{\omega 2}/2 + (P_{i2})_{\omega 1}\}^2 \quad \text{(W)}$$

Normalising with respect to $(P_{i1})_{sat}$ and $(P_{o1})_{sat}$:

$$(Y')_{\omega 1} = (X')_{\omega 1}\{1 - (1 - 3)[(X')_{\omega 1} + 2(X')_{\omega 2}]\}^2$$

where $(X')_{\omega 1}$ and $(Y')_{\omega 1}$ are the input and output powers at angular frequency ω_1 normalised with respect to the single carrier saturation power and:

$$(Y')_{\omega 2} = (X')_{\omega 2}\{1 - (1/3)[(X')_{\omega 2} + 2(X')_{\omega 1}]\}^2$$

where $(X')_{\omega 2}$ and $(Y')_{\omega 2}$ are the input and output powers at angular frequency v_2 normalised with respect to the single carrier saturation power.

The ratio ΔP_i of the input signal power at angular frequency ω_1 to that at angular frequency ω_2 is given by:

$$\Delta P_i = (P_{i2})_{\omega 1}/(P_{i2})_{\omega 2} = (X')_{\omega 1}/(X')_{\omega 2} = (A/B)^2$$

The ratio ΔP_o of the output powers is given by:

$$\Delta P_o = \Delta P_i\{[1 - ((X')_{\omega 1} + 2(X')_{\omega 2})/3]/[1 - ((X')_{\omega 2} + 2(X')_{\omega 1})/3]\}^2 \quad (9.18)$$

Since the normalised magnitudes $(X')_{\omega 1}$ and $(X')_{\omega 2}$ are by definition less than unity, the quantity in square brackets is always greater than unity if $(X')_{\omega 1}$ is greater than $(X')_{\omega 2}$. Since the coefficient of ΔP_i is greater than unity, the ratio of the output powers ΔP_o is thus greater than the ratio of the input powers ΔP_i.

The capture effect Δ is defined by:

$$\Delta = \Delta P_o/\Delta P_i \text{ or } (\Delta)_{dB} = (\Delta P_o)_{dB} - (\Delta P_i)_{dB} \quad (9.19)$$

Equation (9.18) shows that when the IBOs are small (a large absolute value of back-off), the capture effect disappears ($(\Delta)_{dB} = 0$ dB). The capture effect is greatest when the difference between the input signal powers is large and the absolute value of back-off is small. For example, with $(\Delta P_i)_{dB} = 10$ dB, the capture effect Δ is 0.25 dB for a total IBO of −15 and 5 dB at saturation (zero total back-off). With $(\Delta P_i)_{dB} = 2$ dB, the capture effect Δ is 1.5 dB at saturation.

9.2.1.6 Noise power ratio (NPR)

The *noise power ratio (NPR)* is a parameter used to characterise the nonlinearity of a nonlinear device such as an amplifier. It corresponds to a measure of intermodulation when an infinite number of carriers is being amplified.

Figure 9.4 illustrates the principle: the amplifier under test is fed with random white noise previously filtered by a notch filter that eliminates the noise within a narrow band (typically a band less than one-tenth of the amplifier passband) about the notch frequency f_{notch} chosen at the centre of the amplifier passband. With a perfectly linear device, no intermodulation noise should appear in the band of the notch at the output of the amplifier. Intermodulation in the amplifier produces intermodulation products that partially fill the notch. At the output of the amplifier, the noise power spectral density N_0 in a band outside the notch is compared to the intermodulation noise power spectral density $(N_0)_{IM}$ within the notch. The ratio in dB is called the NPR:

$$\text{NPR(dB)} = 10\log[N_0/(N_0)_{IM}]$$

The higher the NPR, the more linear the amplifier.

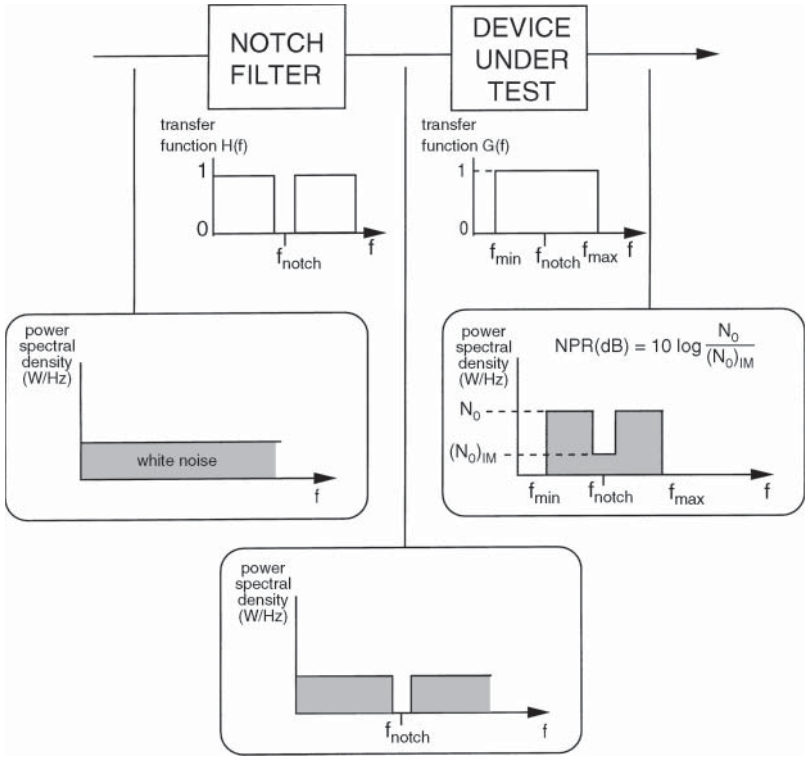


Figure 9.4 Principle of noise power ratio measurement.

9.2.1.7 Amplitude and phase models of an amplifier

The input carrier is represented as a bandpass signal of the form:

$$S_i(t) = Re\{r(t)\exp[\mathrm{j}(2\pi f_0 t + \Phi(t))]\}$$

where $r(t)$ and $\Phi(t)$ are respectively the instantaneous amplitude and phase of the input carrier.

Taking into account both amplitude and phase effects, it is convenient to characterise the nonlinear element by a complex function.

Fuenzalida's model [FUE-73]. The fundamental signal output is represented by:

$$S_o(t) = Re\{g[r(t)]\exp[\mathrm{j}(2\pi f_0 t + f[r(t)] + \Phi(t)]\} \quad \text{(V)}$$

where $g[r]$ and $f[r]$ are the output amplitude and phase functions, respectively. It is assumed they are independent of frequency (a 'memoryless model') and are modelled as a Bessel expansion:

$$g[r]\exp\{\mathrm{j}f[r]\} = \sum_{s=1}^{s=L} b_s J_1(\alpha\, sr) \quad \text{(V)}$$

where $J_1(x)$ is the first-order Bessel function (a function such that $J_{-n}(x) = (-1)^n J_n(x)$, $n = 0, 1, 2, \cdots$), b_s are complex coefficients, and α is a fitting parameter, e.g. $\alpha = 0:6$). The coefficients b_s and the number of terms L in the expansion are selected to give a best fit to the actual nonlinearity. It has been found that 10 terms are sufficient for typical nonlinear characteristics.

In the case where the input to the nonlinear device is formed by the sum of m narrowband bandpass signals, the complex amplitude of the output signal components is given by:

$$M(k_1, k_2, \cdots k_m) = \sum_{s=1}^{s=L} b_s \prod_{l=1}^{l=m} J_{kl}(\alpha s A_l(t))$$

where $A_l(t)$ represents the instantaneous amplitude of the input signal components. In the previous expression, only those components satisfying the condition $\sum_{l=1}^{l=m} k_l = 1$ are to be retained in order to be consistent with the bandpass representation of the output.

Saleh's model [SAL-81]. The nonlinearity is characterised by two-parameter algebraic formulas:

$$g[r] = \alpha_g r/(1 + \beta_g r^2)$$
$$f[r] = \alpha_f r^3/(1 + \beta_f r^2)^2$$

where the α_g, α_f, β_g, and β_f are constants depending on the nonlinearity.

9.2.2 Repeater organisation

The organisation of the repeater is determined by the mission specification and technological constraints. A *large power gain*, together with a *low effective input noise temperature* and a *high output power* over a *wide frequency band*, are to be provided. *Frequency conversion* of the carriers must also be performed.

9.2.2.1 Low noise amplifiers (LNA) and frequency conversion

Frequency conversion between the uplink and the downlink enables decoupling between the input and output of the repeater to be ensured. Re-injection of signals radiated by the output into the transponder input can thus be avoided by filtering.

Frequency conversion can be envisaged as the first operation performed on the carriers from the antenna (a *front-end mixer*). However, except in special cases, this arrangement does not enable the required *system noise temperature* specification to be satisfied because of the high noise figures of mixers. Furthermore, it is preferable to split the power gain between two sets of amplifier units operating with different input and output frequencies. This enables the danger of instability, which is inherent in a very-high-gain amplifier where all the stages operate at the same frequency, to be limited.

The repeater thus consists firstly of a low noise amplifier that provides the required value of *effective input noise temperature* at the uplink frequency. A high gain (20–40 dB) minimises the noise contribution of the mixer which follows the amplifier.

A mixer associated with a local oscillator then provides frequency conversion (Figure 9.5a). The position of the mixer in the chain is determined by the level of the signal to be converted to cope with nonlinear effects.

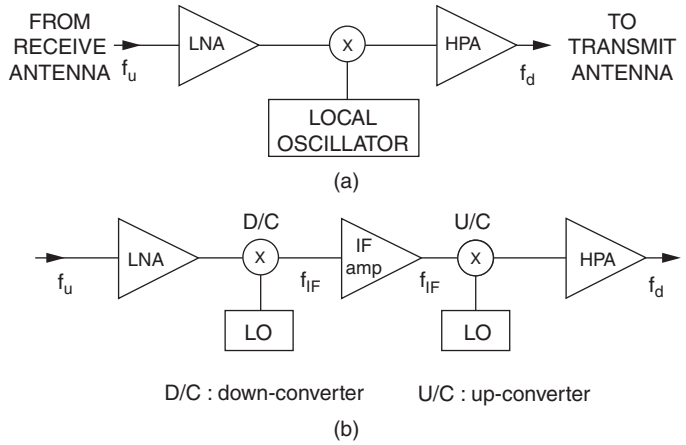


Figure 9.5 Repeater architecture: (a) single frequency conversion; (b) dual frequency conversion.

9.2.2.2 Single and double frequency conversion

After frequency conversion, and taking account of the gain of the low noise amplifier and the conversion losses of the mixer, a certain amount of gain remains to be provided in order to obtain the total required power gain. Depending on the frequency bands concerned, technological considerations can make it difficult to obtain high power gain at the downlink frequency. *Double frequency conversion*, using an intermediate frequency (IF) of lower value than the downlink frequency, is used (Figure 9.5b). The uplink signals are first down-converted to the IF (a few gigahertz) where amplification is performed. An up-converter then translates the frequency to that of the downlink.

Double frequency conversion was used for the first satellites operating in Ku band (14 = 12 GHz). It may remain appropriate for satellites operating in Ka band (30/20 GHz) and above.

Multiband satellites providing international communication services, like those of Intelsat, may have a payload in Ku band with double frequency conversion and an IF of 4 GHz. This architecture is convenient for interconnection of the Ku band payload and the payload operating in C band (6/4 GHz).

9.2.2.3 *Amplification after frequency conversion*

The signal is further amplified after conversion. The signal level increases as it progresses through the amplifying stages of the repeater. The operating point on the transfer characteristic of each stage moves progressively towards the nonlinear region (Figure 9.6a). The level of intermodulation noise is negligible for the input stages, which operate at very low level, and subsequently increases. Depending on the technology used, as characterised by its intercept or compression point, the specified maximum intermodulation noise power level may be exceeded when the signal has passed through a certain number of stages and reaches a given power level. It is then no longer feasible to provide further power amplification using the considered technology; it is necessary to choose a technology with a higher intercept point which generates less intermodulation noise at this given power level. Should one face the maximum state-of-the-art intercept point, then a technique called *channelisation* must be implemented.

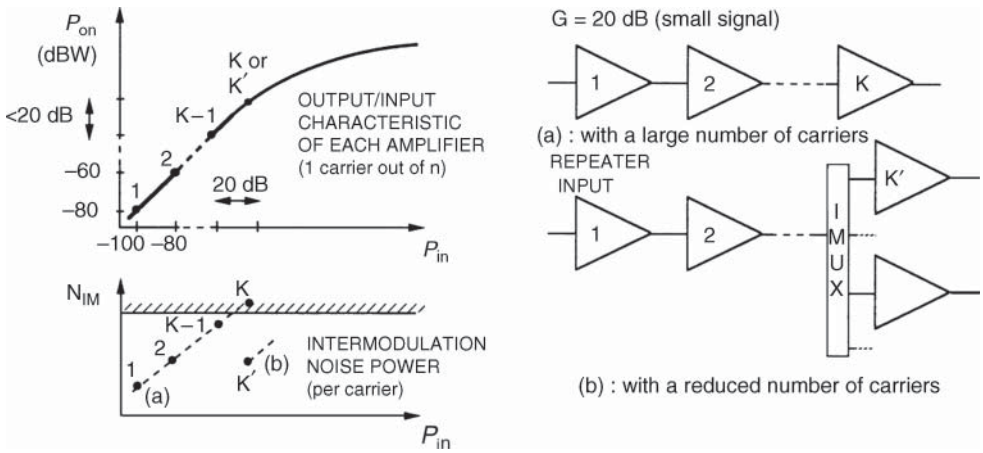


Figure 9.6 Reduction of intermodulation noise by channelisation of the frequency band.

9.2.2.4 *Channelisation of the repeater*

Intermodulation noise. The input stages of the repeater operate on the whole of the system frequency band, several hundreds of MHz. Several tens of carriers share this band and thus give rise to a large number of intermodulation products as these carriers pass through a nonlinear device. A reduction in the number of intermodulation products, and hence the level of intermodulation noise, is obtained by limiting the number of carriers passing through the same amplifier.

When the level of intermodulation noise tends to become excessive for wide band amplification, the system frequency band is divided into several sub-bands that are amplified separately (Figure 9.6b).

Channelisation. The purpose of *channelisation* of the repeater is to create *channels* (sub-bands) of reduced width. Since the number of carriers in each sub-band is less, the intermodulation noise generated by the sub-band amplification stage is much less than the intermodulation noise that would have been generated if the amplifying stage had operated on the total system bandwidth. Figure 9.6 illustrates the comparison qualitatively.

Amplification of the carriers continues within the channel until the required power level is obtained. The amplifier power is shared among the various carriers that occupy the channel. The maximum power available with off-the-shelf equipment developed for space applications is limited. Without channelisation, this maximum power must be shared among all the carriers occupying the system bandwidth. With channelisation, a limited number of carriers shares this maximum power. The available power per carrier is thus greater.

The advantages of channelisation are thus twofold:

— Power amplification with *limited intermodulation* noise due to the reduced number of carriers per amplifier
— *Increase of the total power* of the repeater by having several channels, each benefiting from the maximum power available from a single amplifier

As the band is split into parallel channels, distortion occurs when part of the energy of a carrier feeds into the channels adjacent to the nominal one within which the spectrum of the carriers should be contained. These *adjacent channel interference* (ACI) effects are minimised by a guard band between channels and by the use of filters that restrict the channel widths as closely as possible to those of ideal band-pass filters.

Channel separation is obtained by means of a set of bandpass filters called *input (de)multiplexers* (IMUX). The bandwidths of the channels range from a few tens of MHz to around a hundred MHz (e.g. 36, 40, 72, and 120 MHz). The various sub-bands are recombined in the *output multiplexer* (OMUX) after amplification in each channel. The word *transponder* (or *repeater*) is sometimes used instead of *channel* to designate the equipment that operates within a given sub-band.

Adjacent and alternate channels. The OMUX may or may not be of the adjacent channel type. With non-adjacent channels, the various channels that are combined by a given multiplexer are separated by a wide guard band, equal, for example, to the width of one channel. Realisation of the multiplexer is thus facilitated. To avoid wasting the frequency band corresponding to the guard bands and ensure using the spectrum profitably, the repeater is then divided into alternate *even* and *odd* channels by means of separate IMUXs at the beginning of the channelised section (see Figure 9.9). The channels of each group (even and odd) are then recombined by a different OMUX for each group. The OMUX outputs are connected either to two different transmitting antennas or to the two inputs of a dual mode antenna. This solution was widely used in the past when the design of quasi-ideal bandpass filters, i.e. with a very large Q factor (see Section 9.2.3.2) operating at RF (C, Ku, Ka, etc.), was not well mastered.

Currently, the *adjacent channel multiplexer* allows recombination of adjacent channels. Obtaining a good performance (i.e. narrow guard bandwidth, low insertion loss, and high isolation between channels) imposes severe constraints on the characteristics of the bandpass filters used and leads to substantial complexity in the design and optimisation of the OMUX.

9.2.2.5 Repeater channel amplification

Amplification within the repeater channel uses a preamplifier that provides the power required to drive the output stage. This preamplifier, called the *channel amplifier* (CAMP) or *driver amplifier*, is generally associated with a *variable gain* device that can be adjusted by telecommand. This

permits compensation for variations of power amplifier gain during the lifetime of the satellite. Automatic level control (ALC) may also be implemented.

The *high power amplifier* (HPA) provides the power delivered to the OMUX inputs at the output of each channel. The nonlinearity of the channel may be reduced by including a linearisers within the CAMP.

9.2.2.6 Input and output filtering

At the repeater input, a band-pass filter limits the noise bandwidth and provides a high rejection at the downlink frequencies. At the output, a band-pass filter eliminates the harmonics generated by the nonlinear elements and provides additional repeater output–input isolation. These filters must introduce the lowest possible insertion losses for useful signals. High-input filter-insertion losses cause degradation of the repeater figure of merit G/T and output filter losses cause a reduction of the EIRP.

9.2.2.7 General organisation

The organisation of a repeater with a *single frequency conversion* in accordance with the considerations discussed earlier is presented in Figure 9.7. The equipment that operates over the entire system bandwidth constitutes the *receiver*. It is followed by the IMUX that defines the beginning of the *channelised section*, the channel, and the HPAs, and the OMUX.

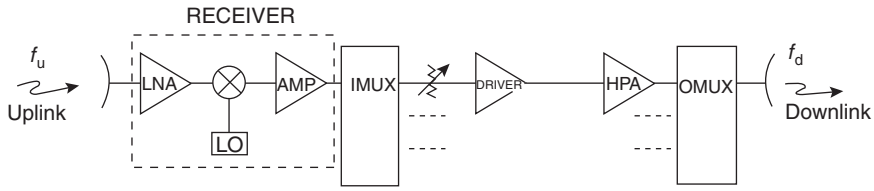


Figure 9.7 Repeater organisation with single frequency conversion.

When double frequency conversion is required, uplink frequency conversion can be performed either in the receiver on the entire system bandwidth with a single mixer, or in the channelised part that then requires as many mixers as channels. In the first case, the whole of the channelised part operates at the downlink frequency, and the mixer operates at low level over the whole band. With the second configuration, demultiplexing and part of the channel amplification is performed at IF. However, a large number of mixers, which also operate at high level, is required.

9.2.2.8 Redundancy

The organisation presented in Figure 9.7 does not include any backup equipment. In order to guarantee the required reliability at the end of life, this architecture is modified to limit *single-point failure* as far as possible; these are elements whose failure causes loss of the mission (see Chapter 13).

The input and output multiplexers do not have a backup since these are passive elements whose failure rate is very low and replication is difficult. The receiver is generally duplicated with an identical backup unit; this is *1/2 redundancy* – one active unit out of the two installed units. A switch operated by telecommand routes the signal from the antenna to the receiver in use; the receiver outputs are connected to the IMUX by means of a hybrid coupler that provides

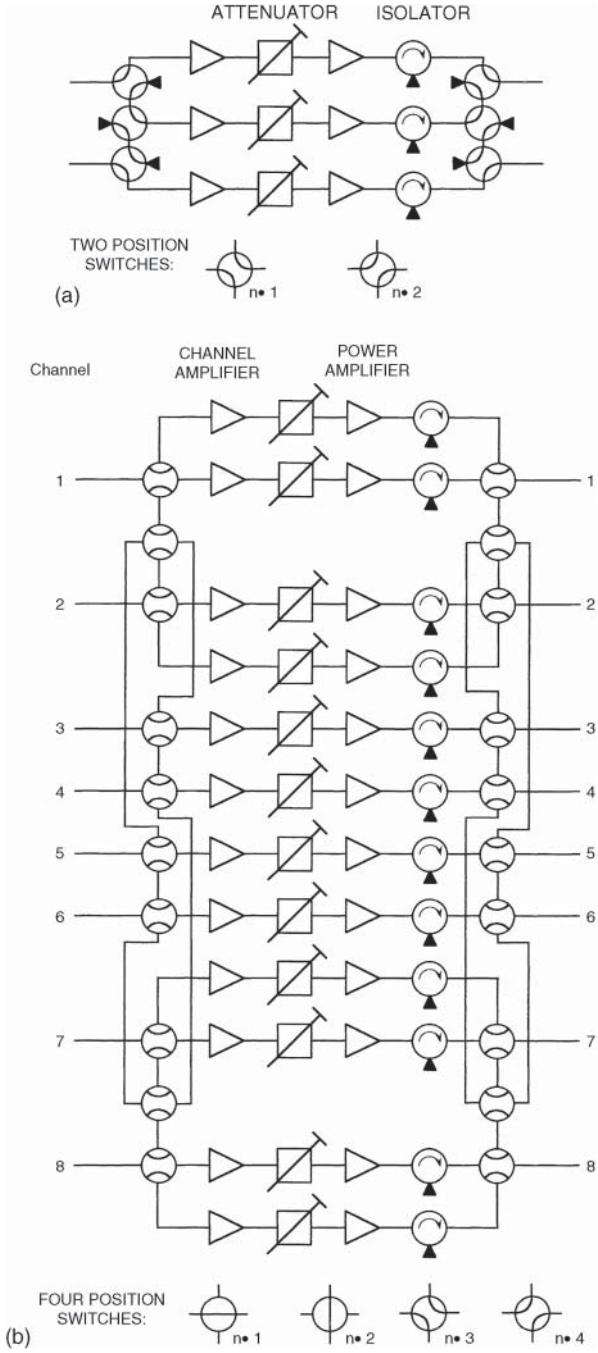


Figure 9.8 Redundancy: (a) 2/3 channel redundancy; (b) 8/12 ring redundancy.

passive routing of the signals. Various redundancy strategies (e.g. 2/4) can be used when the satellite contains several payloads.

The channel-amplifying equipment has a backup (Figure 9.8). In conventional arrangements, a given small number of IMUX output channels is shared among a larger number of amplification chains. For example, with 2/3 redundancy, a switch with two inputs and three outputs routes the signals available on two outputs of the IMUX to two out of the three installed chains. A switch with three inputs and two outputs routes the signals from the two active chains to the two OMUX inputs. In the case of failure of one of the active amplifiers, the standby unit replaces it, but a further failure involves loss of a channel. Therefore in order to increase the reliability of the channelised part, a more complex system called a *redundancy ring* is often used. With this arrangement, all the channels share a larger number of installed chains. For example, Figure 9.8b illustrates 8/12 ring redundancy, where 8 channels share 12 installed amplifying chains. Each channel at the output of the IMUX can be directed to the input of several chains by means of a set of interconnected multi-position switches. This arrangement minimises the number of switches. A large number of possible ways of substituting for failed equipment is thus available; in this way, high reliability at the end of life can be obtained.

9.2.3 Equipment characteristics

The performance of the repeater equipment determines that of the payload. The following sections review the principal units with an emphasis on performance, the influence of this performance at system level and the technologies used.

9.2.3.1 The receiver

The receiver consists of an amplifier at the uplink frequency, a frequency-conversion stage, and amplification after conversion. These elements are generally assembled in the same housing using a modular design.

The input amplifier. The amplifier at the uplink frequency is the main element that determines the figure of merit *G/T* of the transponder. This amplifier must thus have a low noise temperature and a high gain in order to limit the contribution of the noise of subsequent stages. The first satellites used tunnel diode amplifiers. Subsequently parametric amplifiers were used; their principle of operation is based on the reflection of the signal onto a negative resistance.

Nowadays low noise amplifiers (LNAs) are used, incorporating gallium arsenide (GaAs) and high electron mobility transistors (HEMTs). Typical values of noise figures obtained in the various frequency bands are given in Table 9.1.

Table 9.1 Typical low noise amplifier (LNA) characteristics

Uplink frequency band (GHz)	6	14	30	47
Noise figure (dB)	1.4	1.7	2.2	2.4

The frequency-conversion stage. The conversion stage consists of a mixer, a local oscillator, and filters. The frequency of the local oscillator is the difference between the centre frequency of the uplink band and the centre frequency of the downlink band (for a single frequency-conversion architecture and assuming a continuous frequency band). In C band, it is on the order of 2.2 GHz. In Ku band, it is, for example, 1.5, 2.58, or 3.8 GHz depending on the frequency band used on the downlink (10 : 95–11 : 2 GHz, 11 : 54–11 : 7 GHz, or 12 : 5–12 : 75 GHz) for an uplink in the 14–14.5 GHz band.

The principal characteristic parameters are:

— The *conversion loss*, i.e. the ratio of the input power (at the frequency of the uplink) to the output power (at the frequency after conversion) and the *noise figure* (typical values are 5–10 dB)
— The *stability of the frequency* generated by the local oscillator, expressed in the long term over the lifetime of the satellite (the typical value of relative frequency variation must be less than ± 1 to $\pm 5 \times 10^{-6}$) and in the short term ($<10^{-6}$) within the specified temperature range
— The *amplitude of unwanted signals*, i.e. residual input and output signals at the oscillator frequency and its harmonics (typically < -60 dBm) and spurious output signals at frequencies close to that of the main signal (typically < -70 dBc in a bandwidth of 4 kHz situated more than 10 kHz from the signal)

Traditionally, the mixer is of the double-balanced type and uses Schottky diodes. The local oscillator frequency is delivered by a frequency synthesiser based on a phase-locked loop (PLL). The frequency is produced by a voltage controlled oscillator locked to a quartz reference frequency, which may be temperature stabilised or provided with a trimming circuit to permit adjustment of the frequency by telecommand.

Amplification after frequency conversion. This stage provides gain that complements that before the channelised part. The multistage amplifier may contain a device (for instance, a PIN diode attenuator) for gain control by telecommand. High linearity is the main characteristic required for these stages. They operate over a wide bandwidth and, hence, with a large number of carrier signals whose levels may be sufficient to cause nonlinear effects. Typically, the level of third-order intermodulation products must remain less than that of the carrier by more than 40 dB (with characterisation based on two carriers of equal amplitude at the input).

The overall receiver gain is on the order of 60–75 dB. This gain must be constant with frequency over the useful bandwidth in order to avoid distortion associated with nonlinearity of the transponder output stages (amplitude-phase transfer; see Sections 9.2.1.2–9.2.1.6). Typically, the ripple should not exceed 0 : 5 dB over a range of 500 MHz. To obtain such a small ripple, matching between the various stages must be meticulous in order to minimise the *standing wave ratio* (SWR). This minimisation is facilitated by inserting an isolator (a circulator with a matched load) between each stage; this dissipates waves reflected at the interface.

9.2.3.2 Input and output multiplexers

These devices define the input and output of the channelised part. The term *multiplexer* in this case designates a passive device that is used to combine signals at different frequencies from different sources onto a single output or to route signals from a single source to different outputs according to the frequency of the signal. The multiplexers are configured as interconnected high-selectivity band-pass filters. The various architectures and performance obtained (such as channel spacing, insertion loss, and isolation between channels) depend on the interconnection technique used.

Input multiplexer (IMUX). The IMUX divides the total system bandwidth into different sub-bands. The band-pass filters used define the bandwidth of the various channels. A typical configuration involves a set of band-pass filters fed through circulators. Figure 9.9 shows an example where the channels are organised in two groups of *even* and *odd* channels. The IMUX is then divided into two parts that share the power available at the receiver output by means of a hybrid. As shown in Figure 9.9, this hybrid also permits the signals to be delivered to the IMUX to be provided from the redundant receiver thereby avoiding selection by means of a switch.

The *losses* in the multiplexer depend on the number of times the signal concerned passes through a circulator and the number of reflections at the band-pass filter inputs (the loss per

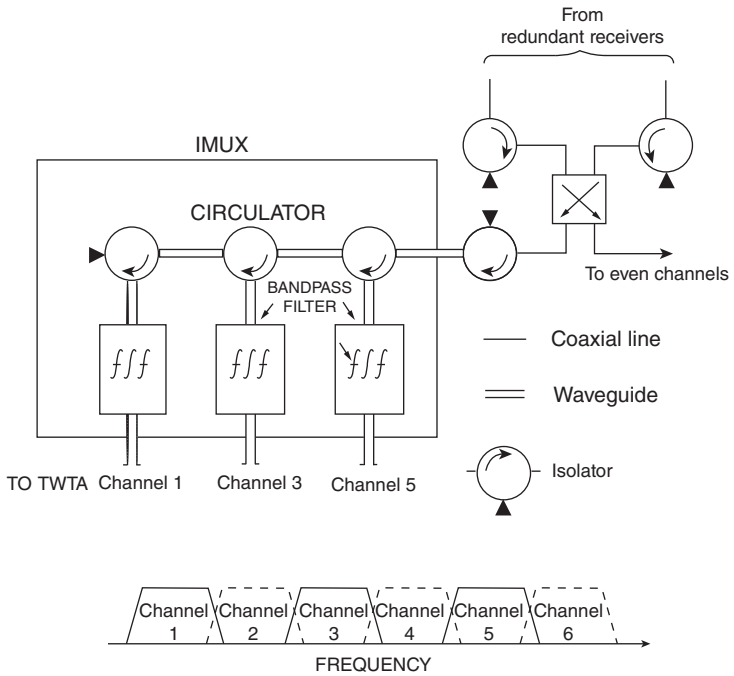


Figure 9.9 Arrangement of an input multiplexer (IMUX).

element is on the order of 0.1 dB). The losses thus differ from one channel to the other and are maximum for the channel furthest from the IMUX input. Division of the IMUX into several parts, each supporting a limited number of channels, enables the difference in losses between channels to be reduced; the losses themselves are not critical since the channel amplification compensates for them.

Output multiplexer (OMUX). The OMUX recombines the channels after power amplification. Unlike the IMUX, losses in the OMUX are critical since they lead directly to a reduction of the radiated power. Instead of using circulators, which are bulky and introduce losses, output coupling of the band-pass filters is achieved by mounting the filters on a common waveguide (a *manifold*) of which one end is short-circuited. The output of each filter, coupled to the common waveguide through an iris, must constitute a short-circuit for out-of-band signals that originate from other channels. The characteristics of each filter thus influence operation of the whole system due to interactions.

Design and optimisation of the OMUX are difficult, particularly with a narrow guard band between each channel. Previously, organisation of the channelised part into even and odd channels left a guard band between each channel with a width equal to that of one channel for each group; this involved less severe constraints on the specification of the OMUXs associated with each group of channels. Extensive research effort on modelling and software development has made possible the design of the *adjacent channel multiplexers* used on current satellites. Figure 9.10 illustrates the amplitude versus frequency response of a 12-channel Ku-band OMUX assembly.

For certain applications (for example, in the case of a backup satellite common to several systems using channels of different frequencies), multiplexers with *tunable filters* have been proposed. The frequency of each channel is then changed by telecommand by adjusting the resonant frequency of the band-pass filters using a tuning device. A mechanical flexible filtering approach could consist of two sub-filters: a pseudo-low-pass filter and a pseudo-high-pass filter.

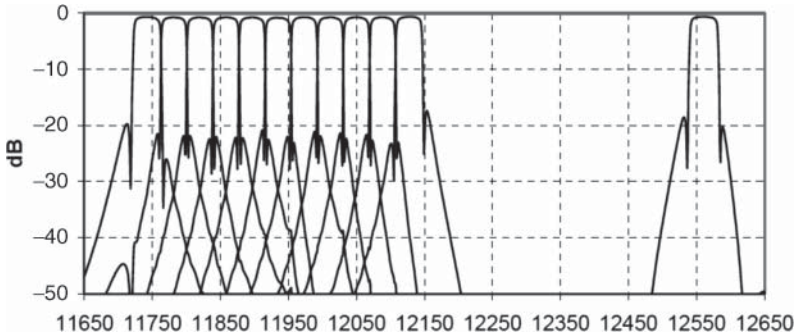


Figure 9.10 Amplitude versus frequency response of a 12-channel Ku-band OMUX assembly (Thales Alenia Space).

Each sub-filter can be tuned independently in the centre frequency by adjusting the length of the filter cavity using a movable metallic top plate that in turn changes the filter centre frequency. Because of the change of the relative centre frequency of each pseudo filter, the overall filter response can be varied in both bandwidth and centre frequency [JON-08].

Band-pass filters. The characteristics of the band-pass filters used are defined as a function of frequency by the amplitude and group delay specifications (Figure 9.11). The amplitude specifications indicate:

— The amplitude and maximum slope of the ripple of the amplitude of the transfer function within the passband
— The minimum slope of the amplitude decrease at the limit of the passband
— The minimum value of attenuation outside the passband

Gain ripples within the band are particularly critical at the IMUX, which is located before the channel power amplifiers. The ripples give rise to spurious modulation of the signal amplitude. Due to AM/PM-conversion effects in the power amplifiers, this amplitude modulation causes spurious phase modulation of the signals. This disturbs the operation of the frequency or phase demodulators of the earth station receivers and hence causes a degradation of the quality of the link.

A high slope at the extremity of the passband allows narrow guard bands between channels and thus permits maximum utilisation of the frequency bands. High out-of-band attenuation is necessary to avoid interference between channels.

The *group-delay specification* defines the maximum permissible variation of group delay within the passband. Variation of group delay causes phase shift between the spectral components of a wideband signal and hence distortion. The transfer functions are of the Chebyshev or elliptic type with several poles (four to eight). Examples of amplitude responses are given in Figure 9.12. *Group-delay equalisers* associated with the filter elements enable the required characteristics to be obtained.

Waveguide cavity filters permit the high Q factors imposed by the amplitude specifications to be obtained. Although single-mode filters were used for early realisations, bi-mode techniques, where two resonant modes are excited in the same cavity, have become dominant. This technique permits reduction by a factor of two in the number (and hence mass and volume) of cavities required for a transfer function with a given number of poles. Tri-mode cavities and even quadri-mode cavities have also been developed.

Transverse electric (TE) modes are usually used in cavities. Coupling of resonant modes between adjacent cavities is realised by irises. Coupling of modes of different kinds (TE and transverse magnetic [TM]) offers new possibilities for the realisation of multimode filters.

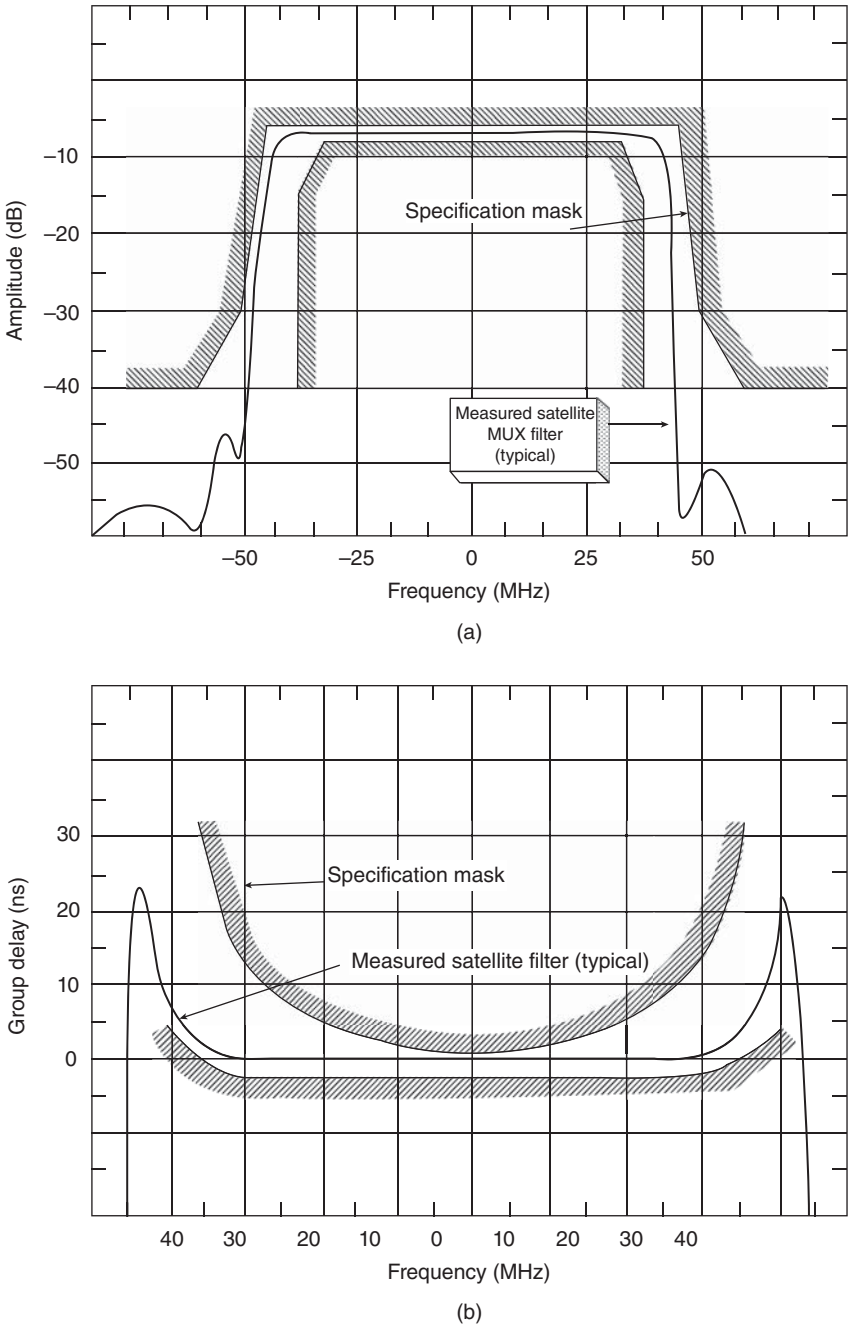


Figure 9.11 Filter mask: (a) amplitude and (b) group delay limits as a function of frequency.

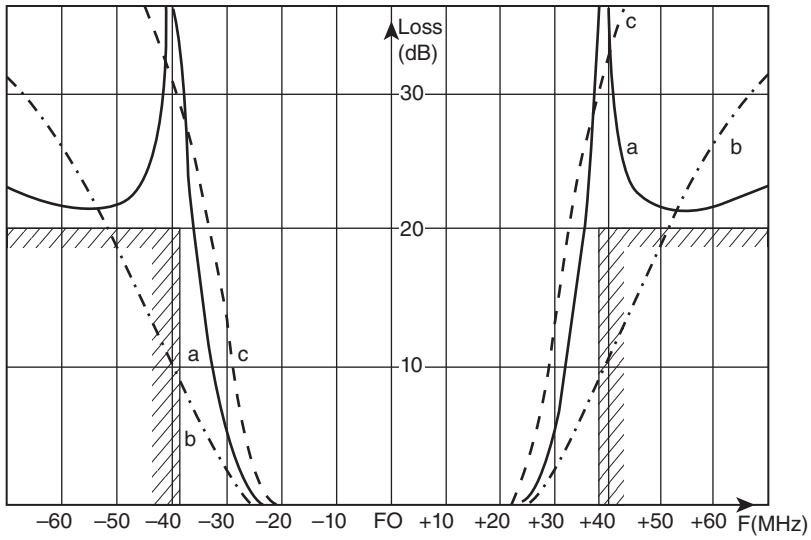


Figure 9.12 Filter responses: (a) four-pole elliptic; (b) four-pole Chebyshev; (c) six-pole Chebyshev.

In order to limit the drift of the centre frequency of the filter (typically less than 2.5×10^{-4} over the lifetime), it is important to avoid dimensional variations of the cavity as a consequence of ageing and thermal expansion. The material used must, therefore, have high mechanical stability and a low coefficient of expansion. It must also be light and a good conductor:

— Aluminium, in spite of its high coefficient of expansion (22×10^{-6} / ° C), which requires precise temperature control to avoid thermal deformation, is used since its good conductivity and low density (2.7) permit cavities to be produced.
— Realisation of resin-impregnated carbon fibre cavities seems promising due to the low coefficient of expansion (-1.6×10^{-6} / ° C), high rigidity, and low density (1.6). The complexity of the fabrication process, however, is limiting development of this technology.
— Invar, an alloy of 36% steel and 64% nickel (coefficient of expansion 1.6×10^{-6} / ° C) has a high density (8.05), but its stiffness permits the use of cavities with thin wall manufacturing. These properties cause the material to be widely used. A coating of silver within the cavity ensures good conductivity and a good surface state, which are required to obtain a high Q factor.

The size of the cavity is directly determined by the propagating wavelength within the medium that constitutes the interior. Traditionally, the interior of the cavity is empty and the cavities are large, particularly at low frequencies (C band).

Use of a material of high permittivity within the cavity (a resonator) and concentration of the field lines in a reduced volume permit fabrication of smaller cavities. Figure 9.13 shows an example of filter realisation using bi-mode cavities coupled by an iris [CAM-90].

Dielectric resonators are now commonly used for C- and Ku-band IMUXs. Typical mass per channel at C band is 240 g. This technology is also applicable to OMUXs up to 280 W with specific design to cope with power dissipation constraints.

Further reduction in the volume and mass of the multiplexer can be expected using super-conducting microwave devices enabling construction of planar filters. The related reduction both in mass and in volume is about 50%.

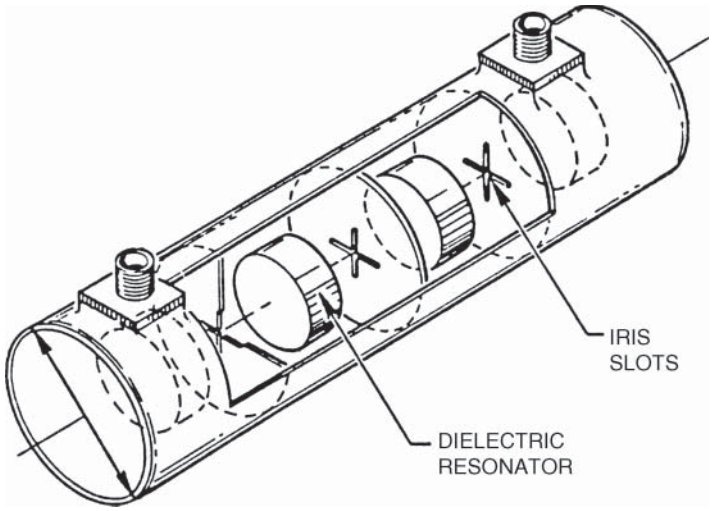


Figure 9.13 Realisation of a bi-mode cavity filter with resonator. Source: reproduced from [CAM-90] with the permission of the AIAA.

Another technique for realising multiplexers is the surface acoustic wave (SAW) technology, characterised by small size and mass. This technology can provide extremely sharp filters and inherent linear phase in the passband, but at a frequency of operation in the range of tens to hundreds of MHz. The achieved passband of each filter is narrow (several tens of kHz), but several filter responses could be combined to enlarge bandwidth. The filter band of operation requires frequency conversion from the RF bands of the satellite systems. This technology is widely used for channelisation with mobile L-band systems, where the bandwidth of satellite channels is small (tens to hundreds of kHz) because of the typical narrow width of the carrier (single connection per carrier [SCPC] with limited data rate) in combination with on-board switches that provide independent routing of channels between multiple uplink beams and downlink beams.

9.2.3.3 The channel amplifier (CAMP)

The receiver output power should be kept within a limit conditioned by the maximum acceptable level of intermodulation noise (Section 9.2.2.3), resulting from the nonlinear characteristics of the receiver. Losses in the IMUX then determine the available signal level at the input of the channel. This level is generally insufficient to drive the channel output HPA.

The CAMP or driver amplifier provides the required power gain, conventionally on the order of 20–50 dB. Good linearity is required in spite of the reduced number of carriers in the channel in order to avoid an excessive contribution to the system intermodulation noise. The use of monolithic microwave integrated circuits (MMICs) permits compact and light realisation (Section 9.6).

The amplifier is associated with an *attenuator* that enables the gain to be adjusted over a range from 0 to several dB in steps of tenths of a decibel. This attenuator, which is controllable via the payload TTC links (telemetry, tracking, and command), is typically realised with PIN diodes (P-type, Intrinsic, N-type semiconductor) whose bias is adjusted in order to vary the conductivity. This permits compensation for the subsequent HPA gain variation over the satellite lifetime or adjustment of the operating point (back-off) of this amplifier.

The amplifier may be associated with an ALC to maintain constant channel output power regardless of input power variations. It can also be associated with a *linearisers* that compensates for the nonlinear amplitude and phase characteristics of the output stage.

Several techniques can be used for linearisation. The predistortion technique, which involves passing the signal through a circuit with a transfer function opposite that of the device to be linearised, is appropriate (see Section 8.4.2.2).

Figure 9.14 illustrates the organisation of a CAMP, displaying the various functions that can be implemented.

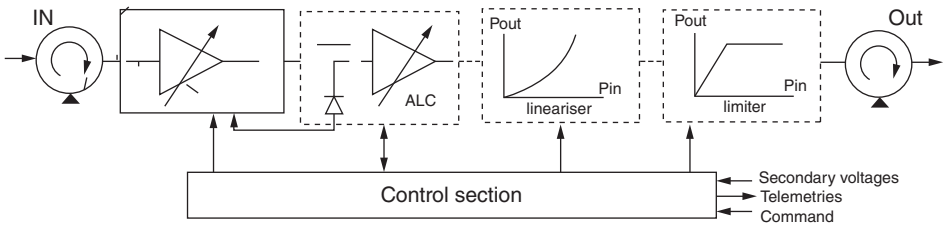


Figure 9.14 Organisation and functions of a channel amplifier (CAMP). (Thales Alenia Space).

9.2.3.4 The output high-power amplifier

The output amplifier provides the power at the output of each channel, and this determines the value of EIRP of the channel. The reference amplifier output power is defined by the single carrier saturation power.

The chosen operating point, as defined by the corresponding IBO (or OBO), results from a compromise between the available output power delivered to a given carrier and the level of intermodulation noise:

— A small (absolute) back-off (operation close to saturation) has the benefit of high power, but intermodulation noise is high since the device operates in a highly nonlinear region.
— A large (absolute) back-off limits intermodulation noise, but the available output power is reduced.

The procedure used to determine the back-off generally consists of optimising the back-off for which the carrier power-to-noise power spectral density ratio $(C/N_0)_T$ of the overall link (station to-station) is maximised (see Sections 5.9.2.3 and 5.9.2.4).

It should be noted that, in multicarrier operation, zero IBO corresponds to an operating point on the transfer characteristic beyond saturation with one carrier and hence the maximum power per carrier is obtained for a nonzero IBO (see, for example, the curve for one of the two carriers in Figure 9.3).

A particularly important parameter of the output HPA is the *efficiency*. The efficiency is defined as the ratio of RF output power to direct current (DC) electric power consumed. The difference is dissipated in the form of heat. A high value of efficiency thus leads to a reduction of electricity consumption and hence the size and mass of the satellite electrical system; the performance required of the thermal control system (as specified in terms of the heat extraction capacity) is also reduced.

Two types of power amplifier are used on satellites: travelling wave tube amplifiers (TWTAs) and transistor solid state power amplifiers (SSPAs).

9.2.3.4.1 Travelling wave tube amplifier

TWTs operate by interaction between an *electron beam* and the radio wave [AUB-92; BOS-04]. Figure 9.15 illustrates the organisation of a TWT.

The electron beam, generated by a *cathode* raised to a high temperature, is focused and accelerated by a pair of anodes. The wave propagates along a *helix*; the electron beam, whose focus is maintained by concentrically located magnets, flows within the helix. The axial velocity of the wave is artificially reduced by the helix to a value close to the velocity of the electrons. The interaction leads to a slowing of the electrons that give up their kinetic energy. The interaction between the electron beam and the electromagnetic wave to be amplified causes the electrons to slow down (on average) near the output end of the helix. The travelling wave thus gradually moves faster and faster compared to the electrons and the synchronism condition necessary for amplification is no longer fulfilled. A way to draw more power from the electron beam, and therefore to increase the electronic efficiency, is to slow down the wave on the helix progressively as it nears the output end, which strengthens the interactions between the RF waves and the electron beam. This is obtained by decreasing the helix pitch, at the expense of an increased phase distortion.

A *collector* captures the electrons at the output of the helix. Division of the collector into several stages at different potentials permits better matching to the dispersion of the residual energy of the electrons and hence an increase in the efficiency of the tube. The residual energy is to be dissipated in the form of heat. The collector conducts the heat to be dissipated either by conduction (conduction-cooled) towards the satellite radiative surfaces of the satellite (see Section 10.6) or directly into space by a self-radiating system part of the TWT (radiation-cooled). Radiation-cooling makes it possible to reduce the thermal load of the satellite and to decrease the overall platform mass for a given RF performance.

The past six decades of helix TWT development have resulted in a constant increase in the overall DC to RF conversion efficiency, up to 75% with the potential to approach 80% for commercial satellite communication applications.

Typical values of the characteristics of tubes used are:

— Power at saturation: 20–250 W
— Efficiency at saturation: 60–75%
— Gain at saturation: around 55 dB
— $(C/N)_{IM}$ at saturation: 10–12 dB (two carriers of equal amplitude)
— AM/PM-conversion coefficient K_p: around 4.5°/dB (near saturation)

An electric power supply (*electric power conditioner* [EPC]) generates the various voltages (up to 4000 V) required for operation of the tube. The efficiency is on the order of 95%, which leads to a global efficiency of 60–65%. The total mass is around 2.2 kg (tube, 0.7 kg; power supply, 1.5 kg).

The combination of the TWT and EPC into an integrated device is called a TWTA. For high-power applications, such as satellite broadcasting, the DC power supply of two TWTs by a single EPC provides a cost- and weight-effective solution. The two TWTs can be operated as single TWTAs independently or RF combined in order to provide about twice the power of each TWT.

Integration of a CAMP including a lineariser with the TWT (*linearised travelling wave tube* [LTWT]) can be achieved to reduce mass and interface complexity. A further step consists of integrating the CAMP, the lineariser, and the TWT into the same housing as the EPC. This *microwave power module* (MPM) approach offers many advantages such as savings in mass, mounting area, and harness simplification in payload integration, as well as improved electromagnetic compatibility (EMC) characteristics and reduction to a single connection to the EPC for DC and all telemetry/telecommand (TC/TM) functions of the MPM.

Other types of TWT have been investigated. They make use of a *cold cathode* instead of a heated one. Electrons are generated by an intense electric field applied to a surface fitted with sharp

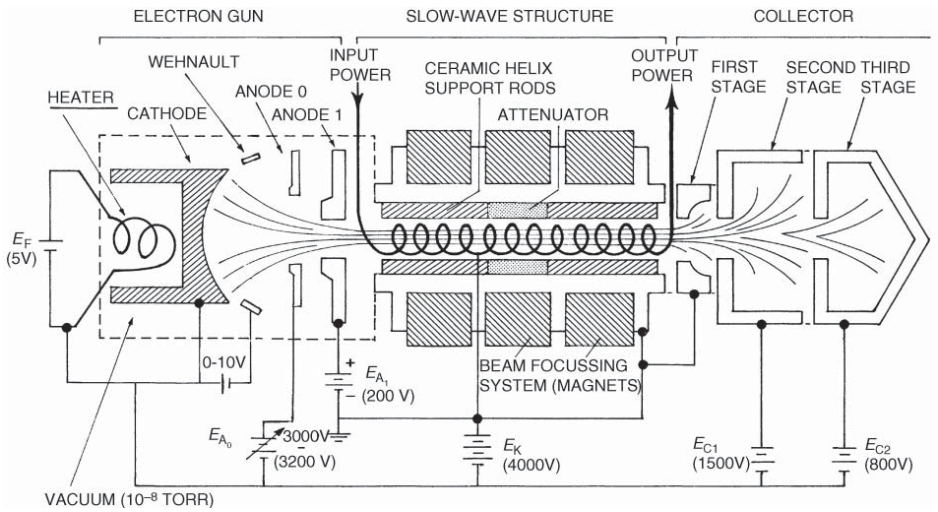


Figure 9.15 The arrangement of a travelling wave tube.

emitter tips. This allows a reduction in size of the device and an increase in efficiency as there is no longer a need to heat the cathode with a current. It is also possible to use shortened helixes using a lower voltage. These tubes are known as *mini-TWTs*.

At high frequencies (Ka band, 20 GHz), propagation conditions can cause large variations (from 5 to 25 dB) of link attenuation (see Chapter 5). A power amplifier with variable output power, in order to be able to match the channel power to the propagation conditions by telecommand, would be of interest. New generations of tubes with an *in-orbit adjustable saturated output power* are being developed. These *flexible TWTs* with optimised helix line require an EPC with adjustable anode voltage. The anode adjustment is performed by telemetry from the ground. The *flexi-tubes* allow variable saturated output power with relatively small power efficiency changes. The great advantage of the flexi-tube is that it offers significantly reduced power consumption with respect to flying higher power TWTAs operated with a given OBO. Therefore, they offer in-orbit flexibility to adapt to the needed traffic for a specific application and to new applications within its lifetime.

9.2.3.4.2 Solid-state power amplifier

SSPAs use field-effect transistors [SEY-06]. The power required is obtained by connecting transistors in parallel in the output stages (Figure 9.16). SSPAs have been used operationally in C band since the beginning of the 1980s with powers on the order of a few tens of watts. It was then anticipated that SSPAs would take over from TWTs, thanks to an appealing power-to-mass ratio and a higher linearity. However, their efficiency in linear mode operation was typically low (about 30%) and has remained so, while TWT efficiency has increased to 70%. In addition, the demand has increased for higher power (typically above 100 W per channel at Ku band). All these facts have given a competitive edge to TWTs.

Typical values for the characteristics of SSPAs are:

— Power: 20–40 W
— Efficiency: 30–45%
— Gain at saturation: 70–90 dB (depending on the number of stages)
— $(C/N)_{IM}$ at saturation: 14–18 dB (two carriers of equal amplitude)
— AM/PM-conversion coefficient K_p: around 2°/dB (near saturation)

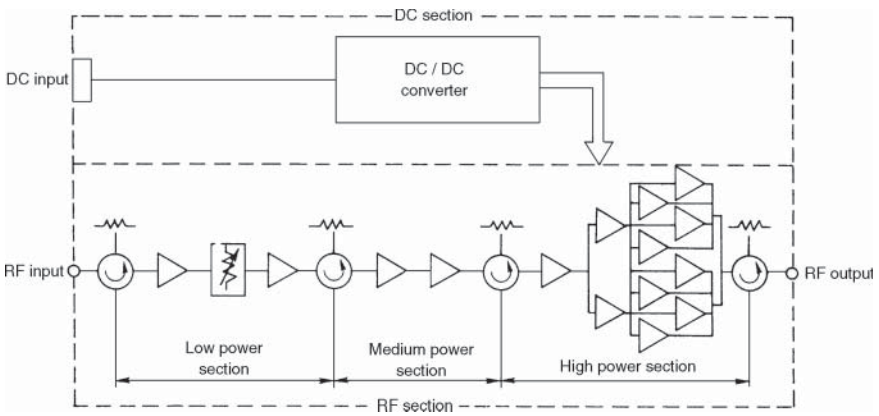


Figure 9.16 Block diagram of a solid state power amplifier.

The power supply associated with the transistor amplifier generates the required supply and bias voltages (a few tens of volts). Temperature compensation is necessary to avoid thermal drifts. The efficiency is on the order of 85–90%, and this leads to an overall efficiency of 30–45% according to the frequency band. The total mass varies from 0.8 to 1.5 kg according to the power. Table 9.2 summarises typical values of the characteristics for TWTAs and SSPAs.

Table 9.2 Summary of TWTA and SSPA characteristics

Characteristic	TWTA	SSPA
Operating band (GHz)	C, Ku, Ka	L, C
Saturated power output (W)	20–250	20–40
Gain at saturation (dB)	~55	70–90
Third-order intermodulation product relative level $(C/N)_{IM3}$ (dB)	10–12	14–18
AM/PM conversion coefficient* K_p(°/dB)	4.5	2
DC to RF efficiency including EPC†(%)	50–65	30–45
Mass including EPC (kg)	1.5–2.2	0.8–1.5
Failure in 10^9 h (FIT)	<150	<150

*Close to saturation.
†Electric power conditioner.

9.2.3.5 *Multiport power amplifiers*

With *multiport power amplifiers* (MPAs), the total available power of a set of amplifiers can be distributed in an adjustable manner between different repeater channels.

The principle of MPAs (also referred to as hybrid or *Butler matrix* amplifiers) is illustrated in Figure 9.17 [CAR-08]. It is composed of three sections: an input hybrid matrix (IHM) or input network (INET), a set of power amplifiers (PAs), and an output hybrid matrix (OHM) or output network (ONET). The input matrix provides power sharing between each of the HPAs by distributing the signal power and spectrum incident on a given input equally between all of the amplifiers with a different, predetermined phase shift at each amplifier. Each amplifier operates on all input signals, and therefore the amplifiers are assumed to operate in their linear region and to have equal gain and phase shift. The amplified signals are then fed to the output matrix that phase-shifts and combines the signals in such a manner that each of its output ports provides a single input port signal after having been amplified by all amplifiers.

The input and output matrixes are made of a number of 3 dB couplers and introduce some insertion losses. Insertion losses are critical as they have an impact on the overall MPA power

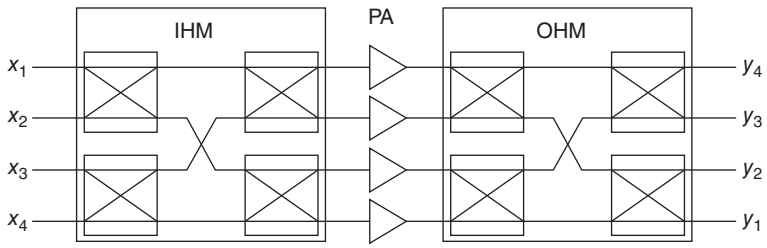


Figure 9.17 Organisation of a four-port amplifier. Source: reproduced from [CAR-08] with permission.

efficiency. Another implementation problem with MPA is that any deviation in the coupler responses and mismatches in gain and phase between power amplifiers results in cross-talk between the MPA output signals. It is therefore important to minimise the deviations and the mismatches from the ideal responses over the temperature range and over the wide frequency range. Different technologies are required for INET and ONET implementations due to different requirements: ONET implementations need to be low loss and able to handle high RF power, while INET implementations need to be lightweight and compact. The use of planar/quasi-planar technology for the realisation of matrix elements is considered for operating frequencies from L band to X Band as it offers significant savings in mass and size when compared with conventional rectangular waveguide or coaxial technology. The preferred realisation methodology for ONETs at higher frequencies (greater than 10 GHz) is rectangular waveguide technology due to inherent lower loss characteristics compared to the planar approach [JON-08].

With a single-beam payload architecture, MPAs allow the distribution of power between different repeater channels. When used in a multiple-beam payload architecture, the use of MPA offers a means to distribute the total available power between different downlink beams. For example, when the satellite mission comprises service areas with different solid angles seen from the spacecraft, the antenna gain associated with the service areas is different. The MPA concept allows compensation such that the performance on each service area could be set equal or individually to a required level. An MPA can also allow an increase in the transmitted power of a beam serving a service area subject to additional attenuation due to distance or atmospheric attenuation. In a system using MPAs, the power allocation for a channel can exceed the power available from a single power amplifier, as and when necessary.

As recombination of power from different power amplifiers is done inside the MPA structure, the use of an OMUX with specific allocated bandwidth per channel can be avoided. Using a wideband OMUX technology instead offers a means of flexibly managing the downlink frequency plan since there are no specificities related to the frequency plan. This opens opportunities that are normally not possible with a conventional payload. Indeed, in the design of a multibeam system, the capacity per beam (power and bandwidth allocated to each beam) must be apportioned in order to serve the expected traffic demand over the life of the communication system. This apportionment results in a trade-off between peak and mean traffic demands, based on the expected mean and peak traffic in each beam over time. Such trade-offs are difficult to make when the number of beams increases and the distribution of traffic in a given beam varies widely over time.

The use of multiport amplifiers offers an efficient way to manage the resources of the payload according to the operational requirements. The HPAs are not dimensioned for the worst-case power required but for the total requirement of a number of channels. This gives an operational flexibility that normally is not available where the dimensioning is performed on the maximum requirement on a per-channel basis rather than on the total power average over the number of channels employing an MPA. This enables the system design to be better optimised and used operationally in a more efficient way.

9.3 REGENERATIVE REPEATER

The advantages of *regenerative repeaters* were presented in Section 5.10 in terms of overall link performance. Regenerative repeaters provide improved link quality compared to transparent repeaters. However, they entail a more complex payload and impose some rigidity on the transmission format. The implementation of regenerative repeaters can only be justified given the other advantages of on-board processing, as presented in Section 7.4.3: adaptation of modulation

and coding to specific requirements of up- and downlinks, baseband switching, beam scanning, and frequency division multiple access (FDMA)/time division multiplex (TDM) for the best match of satellite and earth station features.

Figure 5.36 compares the basic functions of transparent and regenerative satellites: with a regenerative repeater, uplink carriers are demodulated. The recovered baseband signals then modulate an on-board generated carrier, preventing the uplink noise being retransmitted on the downlink.

The on-board processing section depends on the type of application. The switching issues are discussed in Sections 7.4.3 and 9.4.5.

A regenerative repeater contains various units of which some, such as the low noise amplifiers (LNAs), mixers, IF amplifiers, HPAs, and RF filters are similar to those of a conventional transponder and are discussed in Section 9.2.

The equipment specific to a regenerative satellite consists of the demodulating and remodulating equipment and the baseband signal processing equipment. The signals carried by a regenerative transponder are digital (at least at baseband, since the noise added to RF links gives the received carriers an analogue character). The specific equipment is thus designed to process digital signals.

Demodulation can be either coherent or differential according to the digital modulation anticipated for the uplink. In particular, for applications that use the uplink in FDMA mode, possibly with a single connection per carrier (SCPC/FDMA), it is useful to demodulate the various carriers simultaneously.

9.3.1 Coherent demodulation

Quadrature phase shift key (QPSK) modulation in association with a time division multiple access (TDMA) mode provides good performance. Conventional QPSK demodulator structures are well suited. For carrier recovery, conventional structures using PLLs suffer from acquisition times that are too long for operation in burst mode (the phenomenon of *hang-up*), and specific architectures have been considered. It is also necessary to resolve the phase ambiguity of the recovered carrier. This can be achieved by using unique word detection. Finally, the demodulator contains circuits for digital clock recovery and digital signal restoration.

Signal filtering at IF before demodulation must be optimised in accordance with the uplink characteristics. This applies particularly to the transmit filter at the earth station so as to limit degradation due to inter-symbol interference (ISI). This degradation leads to an increase of the bit error rate (BER) with respect to the theoretical value. Filters of the raised cosine type are currently used.

9.3.2 Differential demodulation

Modulation using differential coding permits differential demodulation and avoids carrier recovery (see Section 4.2.6). This type of modulation was used with the first generation of regenerative satellites. Indeed, coherent demodulation requires a carrier recovery arrangement whose complexity involves an increase of mass and power consumption and poses reliability problems. The penalty for the simplified demodulator is an increase of at least 2.3 dB in the power required on the uplink if four-phase modulation is considered (see Table 4.5). Differential demodulation works by comparing the received waveform during a symbol duration and the previous same waveform delayed by one symbol duration. It is, therefore, necessary to deploy a delay line with a stable performance, particularly with respect to temperature. Various

technologies have been developed that make use of filters on silica substrates, microstrip lines on dielectric substrates, waveguide filters, etc.

9.3.3 Multicarrier demodulation

One of the advantages of a regenerative satellite is the ability to use FDMA on the uplinks while having TDM carriers on the downlinks. This enables the transmitting power of earth stations to be reduced and the maximum benefit to be gained from the power of a CAMP that operates close to saturation (see Section 7.4.3.3). A large number of carriers is thus present at the input of the satellite transponder, and these must be demodulated.

One approach is to use a bank of band-pass filters centred on the uplink carriers and to connect each filter to a demodulator. This leads to a high mass and power consumption when a large number of carriers is involved.

When the various uplink carriers have the same data rate and are equally distributed in frequency, block demodulation of all the carriers may be considered. The approach and complexity depend strongly on the existence, or otherwise, of synchronisation of the symbol clock of the digital signals carried on the various carriers.

Several techniques are available for realising a *multicarrier demodulator* (MCD). One uses baseband processing of the signal. After changing the frequency of the carriers to the vicinity of the baseband, time samples of the composite signal are taken and analysed by a digital signal-processing algorithm. This processing can be performed on all carriers by combining multiphase networks and using the fast Fourier transform (FFT) or on each individual carrier after demultiplexing using an array of digital filters or a tree partition with successive divisions of the spectrum by two.

The processing can be performed at IF by using the *chirp* Fourier transform by means of SAW filters mounted in transmultiplexers [KOV-91]. The filter output signal is a time domain representation of the short-term spectrum of the frequency-multiplexed input signal. The use of optical techniques in association with SAW devices could permit demultiplexing demodulation of several carriers within the same circuit.

Instead of permanently assigning a carrier at each station, it is also possible to share a set of links at different frequencies with the same data rate between stations. TDMA is then used, and stations select a frequency for each burst for which a time interval in the frame is available (multi-frequency time division multiple access [MF-TDMA] or multi-carrier time division multiple access [MC-TDMA]). This is the access technique used with DVB-RCS/RCS2/RCS2x (see Section 7.7.3.3).

In this context, operation of a MCD on board the satellite is not simple. The bursts from different stations have different frequencies and clock rates, and reduction or even suppression of the preamble at the beginning of each burst does not permit frequency and clock-rate recovery from one burst to the other. It is thus necessary to ensure synchronisation of the earth station clocks, for example by adjusting these clocks to a reference on board the satellite.

The various techniques presented, however, remain appropriate to the realisation of a multi-carrier demodulator. The carrier recovery circuit in particular must be optimised when the preamble is suppressed or of limited length. The circuit can exploit the coherence of bursts between successive frames or use a nonlinear estimation method.

9.4 MULTIBEAM ANTENNA PAYLOAD

A *multibeam antenna payload* features several antenna beams that provide coverage of different service zones (Section 5.11). As received on board the satellite, the carriers appear at the outputs

of one or more receiving antennas. The carriers at the repeater output must be fed to the various transmitting antennas. Two basic configurations are possible:

— Each receiving-transmitting beam combination constitutes an independent network.
— The stations within different coverage regions belong to a unique network, and station-to-station connections must be established between any pair of stations situated in different service zones.

In the first configuration, the payload contains as many independent repeaters as there are receiving–transmitting beam combinations. These repeaters operate in different frequency bands (e.g. 6/4 and 14/12 GHz), possibly with two orthogonal polarisations for two receiving-transmitting beams in a given frequency band.

The second configuration corresponds particularly to the concept of the multibeam satellite discussed in Chapter 7. Interconnectivity between the different beams must be established. This is achieved by repeater channel (transponder) hopping (see Section 7.4.1) or with on-board processing (see Section 7.4.2 for transparent switching and Section 7.4.3 for regenerative switching).

In the rest of this section, a system containing M receiving beams (uplinks) in one or more frequency bands and N transmitting beams (downlinks) with the possibility of interconnecting any pair of beams is considered.

9.4.1 Fixed interconnection

The interconnections between beams are decided upon designing the payload and then implemented once and for all during manufacture. The receiver coverage is often common to all regions, possibly using two orthogonal polarisations (M = 1 or 2 according to whether one or two polarisations are used).

The satellite contains as many active receivers as there are uplink beams. At the receiver outputs, the IMUXs divide the frequency band into different repeater channels; the number of channels is, a priori, a multiple of the number N of transmitting regions if the traffic between regions is balanced and the channel widths are the same.

Unlike the single-beam satellite, where all channels are grouped on transmission with a unique destination region, the multibeam repeater contains as many OMUXs as transmitting beams; each of these multiplexers combines the channels allocated to the beam concerned. Selection of the destination region is achieved by choosing the uplink carrier frequency so that, after frequency conversion, it falls within the band of one of the channels allocated to the region concerned (transponder hopping; see Section 7.4.1).

9.4.2 Reconfigurable (semi-fixed) interconnection

Unlike the arrangement of the previous section, association of repeater channels with transmitting antenna inputs is not explicit. Using switches controlled by telecommand (mechanically actuated microwave switches), it is possible to reconfigure the payload by changing the branching between the channel output and the inputs of the multiplexers associated with the inputs to the transmitting antennas. This enables the capacity of a beam (that is, the width or number of repeater channels allocated to the beam) to be adapted in accordance with changing traffic demand in the service regions during the lifetime of the satellite. Of course, the number of possible configurations is limited and is predefined upon design.

These facilities are available on the Intelsat satellites, for example. They were further exploited on the Eutelsat II satellites by permitting the frequency band used for the downlink beams to be

selected. With these satellites, the uplink frequencies (14–14.5 GHz) were converted into three separate frequency bands that exploit the various segments available for downlinks in Ku band in Region 1 (10.95–11.2, 11.45–11.7, and 12.5–12.75 GHz). The two orthogonal polarisations are used on each uplink and downlink frequency band. The organisation of the repeater is represented in [MAR-09, Figure 9.16]. The channels were arranged in three groups corresponding to the downlink frequency sub-bands for each polarisation and fed to the inputs of the transmitting antennas. It is possible to modify the branching of certain channels by controlling the switches located between the multiplexers and the CAMPs. A high degree of flexibility of channel management according to the type of signal to be transmitted (such as telephony or television) is thus obtained.

Switch arrangements can be designed to provide both reconfigurable interconnection and channel equipment redundancy, with the aim of reducing the total number of switches. These arrangements allow the selection of p channels from a total of n; they are amplified by p amplifying chains among m: such an arrangement is called $p/n/m$. An example is given in Figure 9.18, where the switching matrix allows the selection of $p = 6$ active channels out of $n = 10$, and where $m = 9$ chains of amplification provide the 6/9 redundancy. This arrangement is called 6/10/9.

9.4.3 Transparent on-board time domain switching

This refers to the on-board implementation of the satellite switched time division multiple access (SS/TDMA) techniques discussed in Section 7.4.2.1. Multibeam satellites may require rapid reconfiguration (within a few hundred nanoseconds) of the interconnections between beams. They must, therefore, be equipped with a fast switching device. A *switching matrix* is used to interconnect the receivers associated with uplink beams and the downlink beam transmitters sequentially. The switches route the modulated carriers without demodulation, either at RFs (those of up- or downlinks) or at IF to facilitate the implementation of the switch. Fast switching implies the use of:

— Switches using active elements
— An on-board device to control the switching sequence such as a distribution control unit (DCU)

9.4.3.1 Solid-state switching elements

The first switching matrices developed used PIN diodes as the switching elements. This is the case for the 6×6 (actually 10×6 due to redundancy) RF switching matrix used on Intelsat VI [ASS-81]. PIN diodes were replaced later by field-effect transistors (FETs) that provide better isolation (60 dB), a shorter switching time (less than 0.1 ns instead of 10–100 ns), and gain (on the order of 15 dB with two stages in cascade); this enables the losses inherent in the architecture to be partially compensated.

9.4.3.2 Switching matrix architectures

Of the various conceivable architectures for an $N \times N$ (or $M \times N$) matrix, only two are capable of distributing the information present on one of the uplink beams onto several downlinks (broadcast mode). These are the architectures with power splitters and combiners on each input and output (*power divider–combiner architecture*) and with cascaded directional couplers (*coupler cross-bar architecture*). These architectures are presented in Figure 9.19.

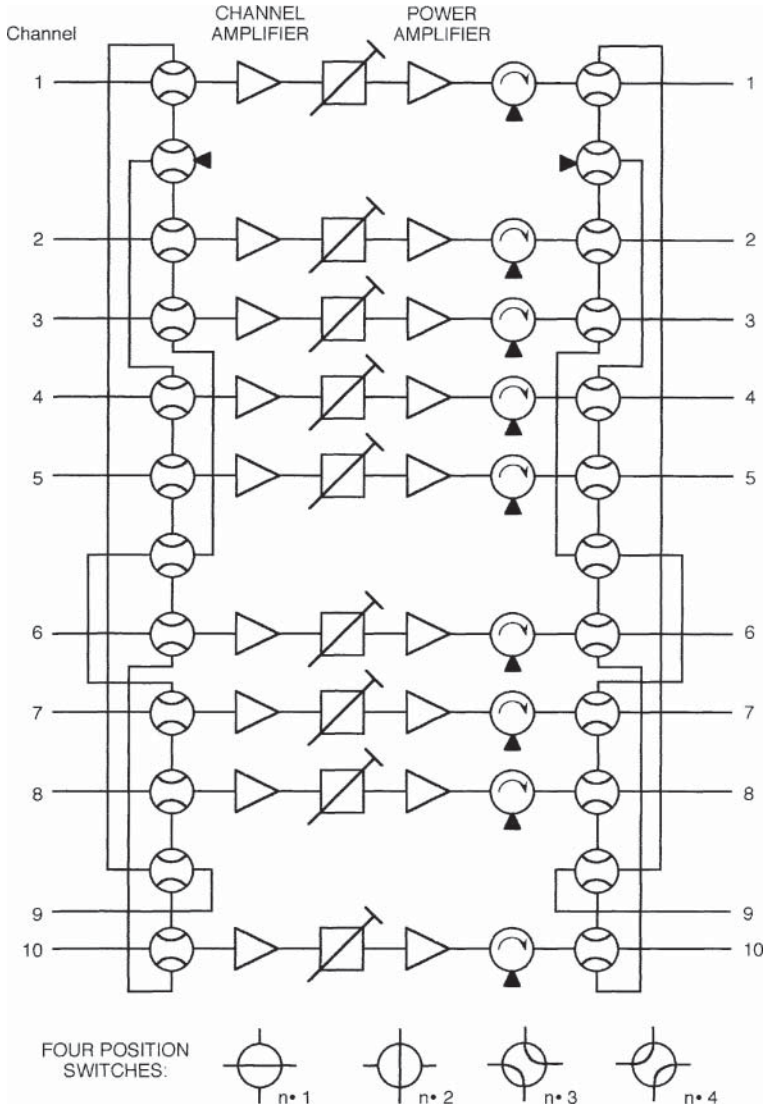


Figure 9.18 An arrangement to achieve $p/n/m$ redundancy.

The divider–combiner architecture (Figure 9.19a) uses $N(M)$ power dividers to separate the input into N channels and N power combiners with $N(M)$ inputs at the output. The various divider outputs are connected to one input of each of the combiners by way of an on–off switch. The connection between the input of the divider concerned and the output of the combiner is achieved or not according to the closed or open state of the switch. The matrix has a cubic form that makes access to the inside elements difficult and favours coupling between channels.

The matrix architecture (cross-bar) uses $N(M)$ input lines and N output lines with interconnecting elements located at the intersections (Figure 9.19b). These elements consist of two directional

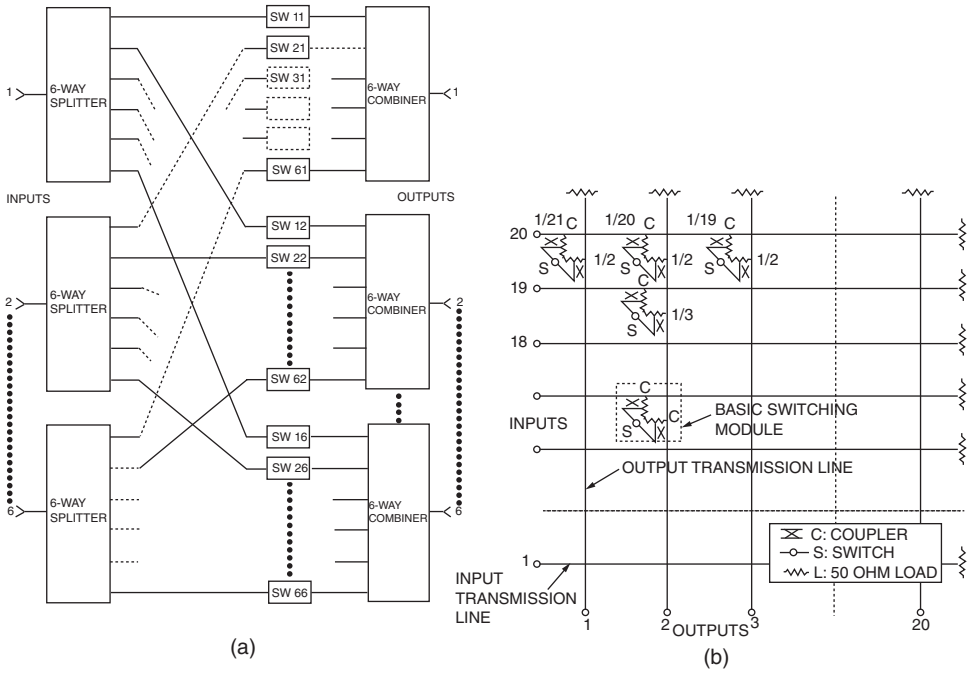


Figure 9.19 Switch matrixes: (a) with splitters and combiners; (b) cross-bar.

couplers with an on–off switch between them. The resultant matrix has the advantage of being planar and well suited to the use of microwave integrated circuit technologies.

There are two possible approaches to realising the couplers for the switching elements – identical and distributed coupling coefficients. In the first approach, all the directional couplers of one row or one column have the same coupling coefficient. Realisation of couplers is simplified, but the power varies from one crossover point to another.

The other approach enables minimum insertion loss to be obtained by operating all switches at the same power level. In order to distribute the power equally among the N crossover points on the same input line, the couplers have different coupling coefficients equal to $1/(N+1)$,$1/N$, … , 1/2 for the 1st, 2nd, … , Nth crossover points. It is thus necessary to realise couplers with different coupling coefficients of a precise value. Furthermore, performance is affected by impedance variations between switches.

Correct operation of the matrix in spite of the failure of one or more switching elements is obtained by producing matrices that contain more rows and/or columns than necessary (more than the number of uplink and downlink beams). Redundancy is thus provided, and switches located at the input (or output) are used to route signals from the inputs to the active lines or from the active columns to the outputs. A fault-detecting device can be integrated into the switching matrix.

9.4.4 On-board frequency domain transparent switching

This section discusses on-board implementation of techniques that are extensions to the concept of transponder hopping discussed in Section 7.4.1. Indeed, in the interconnection scheme

provided by transponder hopping, earth stations are required to hop by changing transmit frequency from one transponder to another depending on the destination beam. In the following discussion, switching is implemented on board in the frequency domain, thus allowing constant frequency operation for the earth stations. This is often referred to as *satellite switching frequency division multiple access* (SS/FDMA).

The switching is done either at a low (a few hundred MHz) intermediate frequency (IF) to ease the filtering process performed in the analogue domain thanks to SAWs filters, or by digital filtering using FFT algorithms.

9.4.4.1 Intermediate frequency domain switching

In the context of a non-regenerative repeater, a set of band-pass filters is used as a switch and routing is performed according to the frequency of the uplink. The difference compared with the transponder hopping techniques used with current satellites lies in the bandwidth and number of filters used; interconnections between the beams of satellites are realised on the basis of the width of one satellite channel (typically 36–72 MHz), and the number of channels is limited by considerations of bulk and mass. An SS/FDMA system could contain a large number of filters with a bandwidth matched to the capacity of the beams concerned. The use of variable bandwidth filters or selection of those whose bandwidth is matched from a group of filters permits on-demand assignment of the resource (the frequency band) as a function of the traffic to be transmitted from one beam to the other. SAW and magnetostatic surface wave (MSW) technologies permit realisation of narrow bandwidth filters of low mass and bulk. The bandwidth can be made variable by combining band-pass filters of adjacent widths.

In view of the mass budget, this technology is applicable to satellite systems whose bandwidth is on the order of a few tens of MHz, which is typical of that of mobile satellite systems.

9.4.4.2 Digital transparent processing (DTP)

For satellite systems whose bandwidth is on the order of several hundreds of MHz, processing carriers with SAW filters at IF is not suitable. Received carriers are down-converted to near-zero frequency, and sampled. Digital samples are fed to several parallel processing chains in order to decrease the data rate in each chain to a value compatible with low-power technologies (complementary metal-oxide-semiconductor [CMOS]). Real-to-complex conversion is performed in digital by a so-called *analytical head*. Time domain complex samples are converted in the frequency domain by FFT processors. A switching matrix made of cascaded/parallelised application-specific integrated circuits (ASICs) ensures the proper connectivity between input and output sub-bands. Any channel can be switched from any input sub-band to any output sub-band (see Section 7.4.2.2 and Figure 7.14). An inverse FFT restores the carrier samples in the time domain, which are then converted back to analogue. Although the repeater incorporates digital processing, it is still a transparent repeater as the carriers are not demodulated.

9.4.5 Baseband regenerative switching

The baseband switching device routes the packets from a particular uplink beam to the appropriate downlink beam (Section 7.4.3.1). Different architectures are possible for performing this function. Among these, three stage structures of the *time–space–time* (TST) type and *single-stage* T type structures are evident. Baseband switching implies, a priori, organisation of the data in the form of a frame. Clock-realignment circuits using buffer memories may be necessary.

In a TST structure, the bursts from the uplinks are stored in buffer memories for the duration of one frame. These bursts are then extracted from the memories and physically routed by a switching network to the output buffer memories associated with the downlinks [PEN-84; MOA-86]. Figure 9.20 is an example of such an architecture.

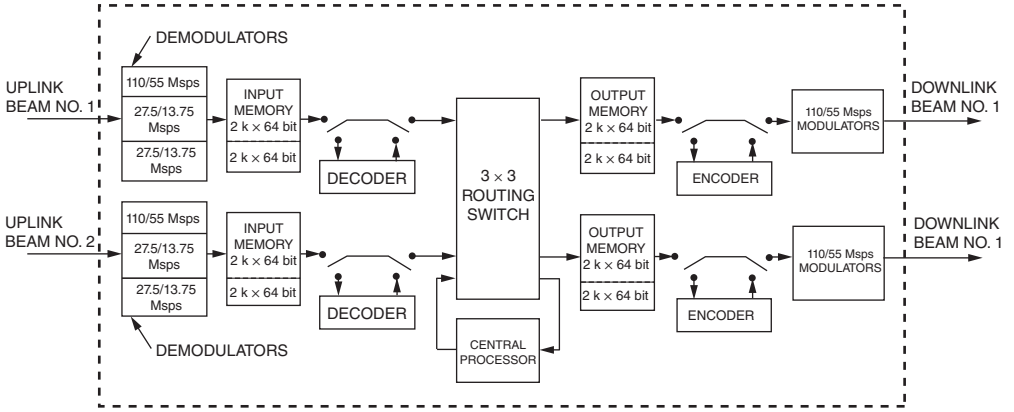


Figure 9.20 Architecture of TST type switching. Source: reproduced from [NAD-88] with the permission of the AIAA.

The TST structure suffers from complexity due to the need to find a path through three stages for all bursts to be routed and the increased number of components required to realise a non-blocking switching network (where it is always possible to find a path) in comparison with a blocking network.

These disadvantages can be avoided by means of a single-stage network where routing can be provided in a simple manner. Synchronous binary words arriving at the various inputs are first converted into parallel form and then transferred successively, via a time-multiplexed bus, to various sections of memory. These words are represented by the number of the time interval that corresponds to a combination of the number of the input on which the word was present and the number of the word in the frame. The process is repeated until all the words in the frame on each input are stored in memory. Words are written into memory at successive addresses in accordance with the destination coordinates of the word as defined by the number of the output and the number of the word in the frame. Hence the word to be transmitted on the first output in the first time interval is stored at address 0, the word to be transmitted on the second output in the first time interval is stored at address 1, and so on. Reading and transfer of words on to the bus is thus realised simply under the control of a counter. The write address for each word is provided, on arrival of the word at the corresponding input, in a synchronous manner by a control device.

A single-stage structure also suffers from certain failings; in particular, it does not permit multidestination broadcasting of bursts. Other structures (such as a modified T type stage and an S type stage with a buffer or time multiplexing) are possible depending on the particular context.

The presence of the baseband signal offers additional possibilities such as changing the digital data rate and using error-correcting coding to combat attenuation due to rain.

9.4.5.1 First-generation regenerative payloads

The first regenerative satellites were designed mostly to interconnect high-data-rate earth stations on the order of 100–200 Mbit s^{-1} through a few beams. Examples are Italsat and the NASA

Advanced Communications Technology Satellite (ACTS). These examples are discussed in more detail in [MAR-09, Section 9.4]. Both used TDMA access, with fixed beams for Italsat and a combination of fixed and scanning beams for ACTS.

The Italsat satellite contains three payloads of which one is regenerative with baseband switching that interconnects six narrow beams at 30/20 GHz. Details of the regenerative payload are given in [SAG-87; MOR-88].

The functions of the baseband switch (BBS) matrix are as follows:

— Synchronisation of the demodulated bursts with the on-board clock
— Routing of bursts to the appropriate modulator
— Generation of reference bursts to synchronise the earth station network

The ACTS satellite contains a payload at 30/20 GHz that uses two multibeam antennas, one for reception and one for transmission; each generates three fixed beams and two beams with electronic pointing and an aperture of 0.3° [NAD-88]. The received carriers are routed to the switching device with a frequency change to 3 GHz by four receiving and amplifying units.

The switching device consists of an IF switching matrix and a baseband processor; one or the other is used according to the mode of operation of the payload (Figure 9.21). In the first mode of operation, the IF switching matrix interconnects the uplinks and downlinks at 220 Mbit s^{-1} using TDMA of the three fixed beams.

The baseband processor is used with the two scanning beams in a mode of operation that uses baseband switching and temporary data storage in buffer memories (see Section 7.2.4). Each beam accepts either one link at 110 Mbit s^{-1} or two frequency division multiplexed links at 27.5 Mbit s^{-1}. The links use TDMA with a 1 ms frame. In order to compensate for the large attenuation in Ka band during rain, the data rates of the earth stations concerned are divided by four, and an error-correcting code of coding rate 1/2 is also activated (see Section 4.2.7). The demodulators then operate at half the nominal data rate (55 and 13.75 Mbit s^{-1}); the bit rate transmitted in the bursts concerned is halved. Decoders and encoders are activated only for the bursts concerned in the frame; control is provided by the network control station.

9.4.5.2 Second-generation regenerative payloads

A significant interest with regenerative satellites is to take advantage of on-board regeneration to multiplex at the satellite level different information contents originating from small earth stations (typically 1–2 m antenna), and distribute this multiplex on a single carrier to small receive-only earth stations, thanks to the benefits of FDMA/TDM. Examples discussed in [MAR-09] are the Skyplex payload on Eutelsat Hot Bird satellites, for video and data broadcasting, and Worldspace, for audio and data broadcasting. A more recent implementation is the Amheris payload on Hispasat satellites.

Concurrently, DVB-S/S2 (see Sections 4.7, 4.8, and 7.5) and DVB-RCS (DVB-S with return channel via satellite, Section 7.5.1) standards have emerged and have been widely adopted by earth station manufacturers, thus offsetting the risk for the satellite operator of implementing an on-board processing system and imposing a proprietary transmission scheme.

9.4.6 Optical switching

The use of *optical switching elements* has also been investigated. With this approach, the RF carriers modulate optical sources (lasers). The light signals are dynamically switched with optical switches and are then detected to regenerate the RF carriers.

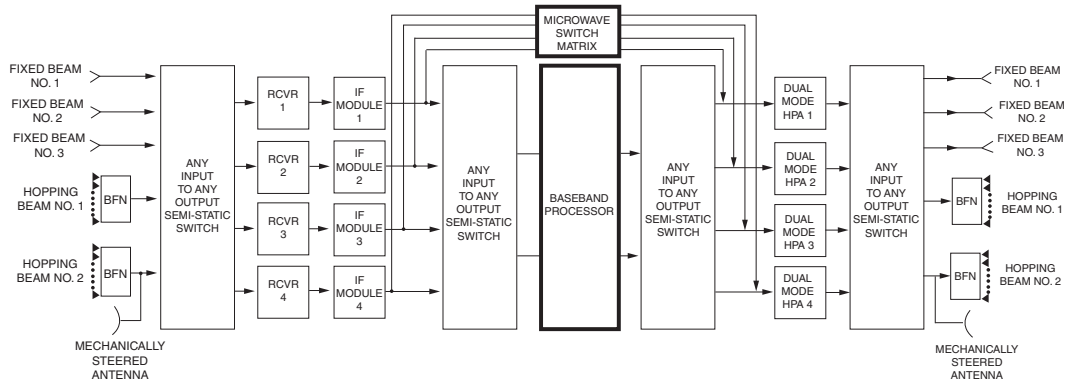


Figure 9.21 Architecture of the ACTS satellite payload. Source: reproduced from [NAD-88] with the permission of the AIAA.

Various approaches are possible for realisation of the switching elements; these include opto-electronic switching using an intermediate detector, modification of the coupling between two adjacent optical waveguides (delta–beta switch), and transmission or total internal reflection (TIR) where two optical waveguides cross by variation of the dielectric constant. A combination of the latter two techniques has permitted the development of an optical switch, which provides good performance in terms of isolation (58 dB) and switching time (several nanoseconds) and can be used to realise switching matrices of different architectures. In spite of the high insertion losses (60 dB), the advantage of an optical switching matrix lies in the associated reduction of size and mass.

9.5 INTRODUCTION TO FLEXIBLE PAYLOADS

Most current commercial spacecraft are based on dedicated designs: the frequency plan, antenna coverage contours, *G/T*, and EIRP distributions are fixed for a given communication mission and are defined for a given satellite position on geostationary satellite orbit. Typical approaches consider the use of one antenna aperture per coverage, and the number of antennas that can be accommodated limits the coverage capability of a satellite. The main drawbacks are significant non-recurring costs, relatively long development schedules, and limited flexibility capabilities.

Flexible payloads, reconfigurable in coverage, frequency plan, and routing, are an efficient answer to the following needs:

— Universal payload for in-orbit replacement of any satellite with continuity of service
— Reconfigurable payload to follow market evolution
— Standard payload for low cost and fast schedule procurement

Two types of architecture (channelised amplification and distributed amplification) can be envisaged to address the needs of satellite operators. These concepts are illustrated in Figure 9.22 [VOI-08].

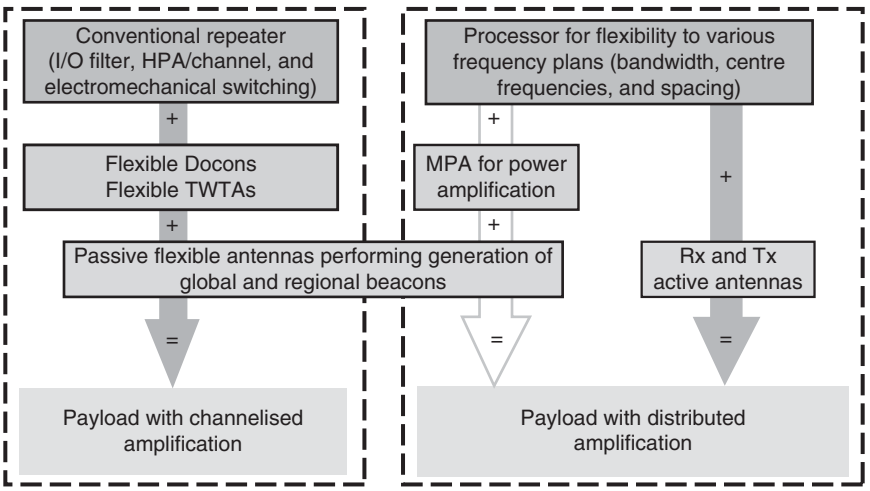


Figure 9.22 Concepts of flexible communication payloads. Source: after [VOI-08]; used by permission.

Architecture with channelised amplification. The payload remains based on a conventional repeater architecture with dedicated input/output filters and TWTA per channel. The repeater incorporates frequency downconverters (docons), which local oscillators and therefore frequency shifts can adjust, offering some flexibility in terms of adaptation of the satellite frequency plan. Flexible power amplifiers allow variable channel output power. The repeater is associated with passive flexible antennas. This type of architecture offers:

— Coverage flexibility thanks to the use of passive flexible antennas (mechanically steerable antennas or electrically reconfigurable antennas). Depending on customer needs, coverage flexibility can be limited to a highly flexible secondary coverage, the main coverage being fixed and achieved with a conventional antenna.
— Power flexibility thanks to the use of flexible TWTAs. Saturation output power is tunable within a limited range, typically larger than 3 dB, while keeping high efficiency.
— Flexible receive and transmit connectivity. The use of a flexible downconverter offers flexible selection of the receive frequency blocks. However, because of IMUX and OMUX, transmit frequencies are fixed, and therefore there is limited flexibility in terms of adaptation of the satellite frequency plan.

Architecture with distributed amplification. The amplification is performed through a bank of power amplifiers providing amplification across multiple channels. This can be achieved by either multiport amplifiers (MPAs) based on Butler matrices (a type of beam forming network [BFN]) or active transmit antennas.

This is the sole approach that avoids the use of OMUX and so is compatible with various downlink frequency plans. Channel passband filtering, gain control, and routing functions have to be implemented prior to the power-amplification section. Typical solutions consider an *on-board processor (OBP)* – either analogue or digital – operating at IF. Frequency-conversion chains before and after the processor and switching matrices permit interconnection of the processor inputs and outputs with the receive and transmit beams (see Sections 7.4.2.2 and 7.4.3.1). This more complex payload architecture offers:

— Coverage flexibility, either obtained by the use of passive antennas, flexible or not, fed by TWTAs in parallel or MPAs, or achieved by use of active antennas. With active antennas, multiple independent beams can be formed through the same antenna aperture.
— Frequency plan flexibility, where the achievable performance is strongly dependent on the OBP solutions. It can consist either of flexible selection of channel bandwidth within a limited set of values (36 or 72 MHz bandwidth, for example) or full flexibility in terms of channel bandwidth adaptation (from a few MHz up to bandwidths larger than 100 MHz). Flexible uplink and downlink beam interconnectivity is also provided within frequency plan flexibility.
— Power flexibility due to the intrinsic properties of distributed amplification.

The satellite industry works to propose innovative flexible payload solutions to develop the satellite communications business. Possible late definition of missions in the procurement process or in-orbit reconfigurable telecommunications payloads to cover various missions at different orbital positions will add value to the services and bring solutions to create and develop new applications. Flexible commercial satellites bring other advantages, such as the increased ability to provide backup satellites in the fleet by using flexible spacecraft. Designing flexible payloads will lead to a generic design and, hence, standardised satellites. The lower cost and shorter manufacturing cycles made possible by standardisation will increase the attractiveness of flexible solutions.

9.6 SOLID STATE EQUIPMENT TECHNOLOGY

In comparison with the technologies available for equipment used on the ground, it is necessary to take account of the constraints due to the specific environment in the design of on-board satellite equipment. This section first introduces the major characteristics of the environment then presents analogue and digital solid state component technology used on board satellites.

9.6.1 The environment

This environment (see Chapter 12) is characterised by:

— High cumulative radiation during the lifetime of the equipment. Depending on orbits and shielding, the components must be able to withstand cumulative doses up to 100 krad. For a geostationary satellite, taking into account a 10 mm aluminium shield, and a mission lifetime of 12 years, the total cumulative dose requirement is less, on the order of 10 krad. The radiation contains, in particular, heavy ions that may cause disruptions, *single-event upset* (SEU), and *latch-up*.
— The absence of convection, which causes problems in evacuating the heat dissipated by components in a reduced volume. The use of heat sinks (strips of copper bonded to components) enables the heat generated to be routed to the equipment case and the structure of the satellite.
— Demands for high reliability driven by the requirement for a long lifetime without the possibility of intervention for maintenance.

Single-event upset refers to the response of an integrated circuit (IC) to a single radiation event that can cause the temporary failure or change of state of the IC.

Latch-up is a failure mechanism of an IC caused by high-energy radiation whereby the circuit is unable to return to its previous state after the stimulus, i.e. impacting radiation, stops. Basically, a PNPN or an NPNP thyristor-type parasitic structure suddenly turns to an 'on' state, thereby bypassing or shorting out portions of the IC. Latch-up is a catastrophic situation that requires the shutdown of the system to clear, or a fatal condition if the chip is destroyed by overheating due to current consumption exceeding its limit. Avoiding latch-up is of particular importance in the design of the equipment on board a spacecraft.

9.6.2 Analogue microwave component technology

Conventional technology employs *hybrid circuits* using *bare* (*unencapsulated*) *chips* as active components. The microstrip circuits are produced by photolithography of the substrate (alumina, sapphire, etc.), which is covered with a layer of metallic conductor obtained by deposition under vacuum (*thin-film technology*).

MMICs are now widely used, although not all of the microwave functions on board the satellite employ MMIC technology. For instance, the first stage of the LNA of the satellite receiver is still based on discrete pseudomorphic high electron mobility transistors (PHEMT) hybrids.

An MMIC receiver incorporates about 10 MMIC chips. They perform the following functions: low-level amplification at uplink frequency, balanced mixing for frequency conversion, and amplification at intermediate and downlink frequencies. MMICs are also used for amplification and linearisation within the CAMP and the HPA.

The advantages of MMIC are size reduction compared to hybrid circuits, leading to a reduction in the mass of the equipment (the factor is typically 2.5), low cost and good uniformity for the production of large series of identical circuits, low cost for the equipment manufacturing due to the quasi-absence of tuning, reduced parts count, and interconnects, which improve reliability.

9.6.3 Digital component technology

With regenerative repeaters, the digital form of the signal permits the use of IC technology. The high processing power needed together with the constraints on mass, bulk, and power consumption require very-large-scale integration (VLSI) of the components with semi-custom implementation (ASIC). These technologies must have the following properties: radiation resistance, high noise immunity, high speed, low power consumption, and high integration density. Some of these properties are incompatible and should be traded off according to the application concerned.

Silicon (Si) is the primary semiconductor material used to produce logic circuits. Thanks to the extraordinary feature-size shrink in the 100–50 nm range (Moore's law), the use of various efficient device improvements (i.e. copper and low K materials for interconnection stack, strained Si, high K metal gate) and the adoption of silicon on insulator (SOI) wafers, the propagation delay time of logic gates has been continuously reduced, down to less than 10 ps (picosecond equal to 10^{-12} of a second). The MOS transistor, as implemented in the CMOS process, is the workhorse of logic; it offers the best gate density. Tuning the MOS transistor characteristics allow setting of the delay vs power consumption trade-off. Mixing both MOS and advanced heterojunction bipolar transistor (HBT) in the SiGe Bi-CMOS allows much higher frequency to be achieved; this is commonly used for mixed signal circuits. Compound semiconductor material (GaAs – gallium arsenide and, especially, InP – indium phosphide) circuits are thus pushed to the highest achievable frequency range (more than 100 GHz).

Radiation resistance. Thanks to the very thin oxide thickness in advanced processors (a few nanometres), CMOS resistance to cumulative doses naturally improves (up to 100 krad) and only the isolation leakage has to be circumvented by dedicated design techniques (edgeless transistors, guard rings). The sensitivity to SEU and single-event latch-up (SEL) caused by heavy ions can be reduced by, respectively, logic design techniques involving redundancies and layout techniques using guard rings. CMOS on insulating substrate (SOI) is much less sensitive to the passage of heavy ions.

Speed and power consumption. CMOS, especially when implemented as SOI provides a very low power consumption, although this consumption increases linearly with switching frequency (a CMOS gate consumes current only during a change of state). The power consumption, in the range of 10 nW/MHz/gate for a 130 nm generation, scales down by nearly 40% for each generation (90–65 nm).

Integration density. CMOS silicon technology offers aggressive integration densities such as 150–200 kgates mm^{-2} for a 130 nm generation that nearly doubles at each generation, thus allowing for the implementation of circuits including several tens of millions of gates for 130 and 90 nm, and more than 100 Mgates for 65 nm generation.

In conclusion, CMOS components completely dominate for realisation of circuits with a high integration density and high operating speed. This situation will continue thanks to the coming generation (45, 32, and 22 nm).

9.7 ANTENNA COVERAGE

A satellite communications mission specifies the coverage performance of a service zone in terms of minimum RF objectives (satellite EIRP or power flux density at the ground for the downlink coverage and satellite *G/T* or power flux density at the satellite for the uplink coverage). This performance must be achieved at a set of locations identified on the earth by their geographical

coordinates. These locations are the service zone reference points. Several concepts must be considered:

— *Service zone contour:* The contour joining the reference points as they are seen from the nominal position of the satellite
— *Geometrical contour:* A contour encompassing the service zone, as seen from the satellite, whatever the antenna pointing error
— *RF coverage:* The area within which a required value of RF performance is guaranteed

Service zone contour and geometrical aspects are discussed first. RF aspects are presented in Section 9.8.

9.7.1 Service zone contour

The service zone reference points are identified in a (x,y,z) satellite-centred coordinate system (Figure 9.23). The z axis is oriented in the satellite–earth centre direction, the y axis is perpendicular to the meridian plane of the satellite and is oriented towards the east, and the x axis is perpendicular to the xy plane and oriented so as to complete a right-handed coordinate system (it points to the north for a geostationary satellite).

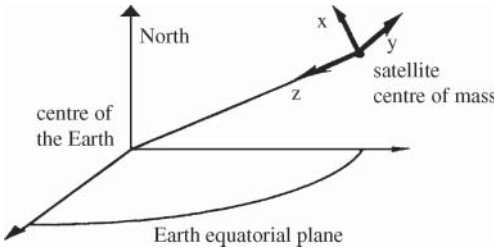


Figure 9.23 Reference coordinate system.

A point P on the surface of the earth has coordinates x_P, y_P, and z_P in the satellite-centred coordinate system. These coordinates can be calculated as a function of the altitude of the satellite and the coordinates of the sub-satellite point. For a geostationary satellite, one obtains:

$$\begin{aligned} x_P &= R_E \sin l \\ y_P &= R_E \cos l \sin L \\ z_P &= R_0 + R_E(1 - \cos l \cos L) \end{aligned} \tag{9.20}$$

where:

R_E = 6378 km, the mean earth radius
R_0 = 35 786 km, the satellite altitude
l = the latitude of point P
L = the longitude difference between the longitude of point P and that of the sub-satellite point

9.7.1.1 True view angles

The *true view* for any point P on the earth is defined as the set of two angles identifying the direction of point P as seen from the satellite. These angles are:

— The angle θ between the direction of the centre of the earth and the direction of point P. θ identifies with the nadir angle for point P.
— The angle φ between the meridian plane of the satellite (plane xz) and the plane defined by the direction of the centre of the earth and point P.

The true view angles are obtained from the coordinates x_P, y_P, and z_P of point P as follows:

$$\theta = \text{arc tan}(\sqrt{x_P^2 + y_P^2}/z_P)$$
$$\varphi = \text{arc tan}(y_P/z_P) \tag{9.21}$$

In the case of the geostationary satellite, the true view angles can be calculated directly from the relative longitude L of the satellite and point P and the latitude l of P, by replacing in Eq. (9.21) expressions for the coordinates of P as given by Eq. (9.20). Results are given in Section 9.7.4.

9.7.1.2 Representation of the service zone contour

Representation of a contour on a map poses the problem of converting from three-dimensional space to a plane.

One representation consists of using a reference plane tangential to the surface of the earth at the sub-satellite point and performing a projection of the points on the surface of the earth onto this plane (Figure 9.24a). The result bears little resemblance to the view of the earth from the satellite. In particular, with a sub-satellite point on the equator (e.g. a geostationary satellite), the poles are apparent, although in reality they are not visible.

A more realistic representation is obtained by using the same plane tangential to the surface of the earth at the sub-satellite point and hence perpendicular to the direction to the satellite–earth centre direction, but defining the contour as the intersections (oblique projections) on this plane of the lines joining the satellite and the set of the relevant points on the earth (Figure 9.24b). The map obtained in this way bears more resemblance to the view of the earth from the satellite. For instance, with a geostationary satellite, the poles are not visible.

In the latter representation, the points of the considered contour have coordinates X and Y in an xy coordinate system centred on the sub-satellite point O. The x and y axes in the tangential plane to the earth are aligned with the corresponding axes of the satellite-centred coordinate system. For a geostationary satellite, the X and Y coordinates are given by:

$$X = K \sin l$$
$$Y = K \cos l \sin L \tag{9.22}$$

where:

$$K = \frac{R_0 R_E}{R_0 + R_E(1 - \cos l \cos L)} \tag{9.23}$$

The true view angles defined in Section 9.7.1.1 can be represented on this xy plane. For example, a point P (Figure 9.25) can be represented by:

$$X = \theta \cos \varphi$$
$$Y = \theta \sin \varphi \tag{9.24}$$

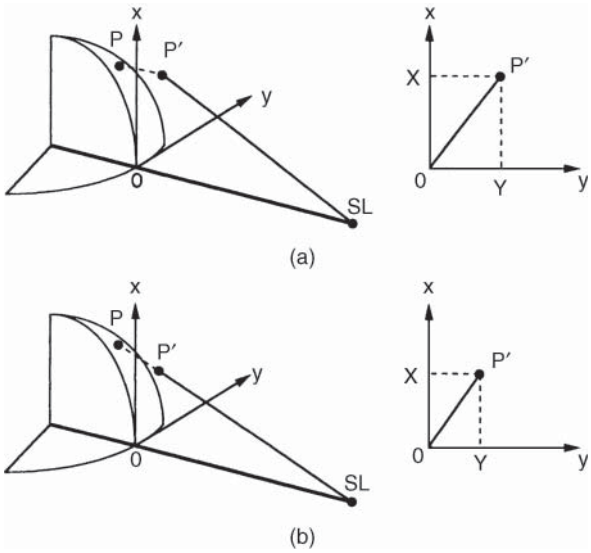


Figure 9.24 Possible representations of a point P on the earth in a plane tangential to the surface of the Earth at the sub-satellite point: (a) orthogonal projection P′ on the plane; (b) oblique projection P′ according to a true view from satellite SL.

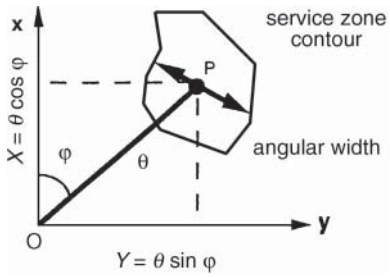


Figure 9.25 True view angles θ and φ of point P in the xy coordinate system.

The angle θ is represented by the segment OP and the angle w by the angle between Ox and OP.

The representation obtained well expresses any apparent displacement of the point in the satellite coordinate system; for example, a rotation about Oz (a variation of φ with θ constant) is represented by the arc of a circle centred on O. However, the angular distances between points are not respected. Consider two directions OP_1 and OP_2 represented respectively by the true view coordinates (θ_1,φ_1) and $(\theta_2,\varphi_2 = 0)$. The angular distance χ between P_1 and P_2 on a satellite centred sphere is obtained from spherical trigonometry:

$$\cos\chi = \cos\theta_1 \cos\theta_2 + \sin\theta_1 \sin\theta_2 \cos\varphi_1$$

hence:

$$\chi = \arccos(\cos\theta_1 \cos\theta_2 + \sin\theta_1 \sin\theta_2 \cos\varphi_1)$$

In the true view representation, the angular distance between P_1 and P_2 is represented by the length of the segment P_1P_2 and calculated from the geometry of the triangle OP_1P_2 as:

$$P_1P_2 = (\theta_1^2 + \theta_2^2 - 2\theta_1\theta_2 \cos\varphi_1)^{1/2}$$

The relative error ε in this representation is thus given by:

$$\varepsilon = (P_1P_2 - \chi)/\chi$$

For a geostationary satellite, the nadir angle of any visible point is at most equal to 8.7°; by considering the two points P_1 and P_2 at the limit of visibility, P_1 on the equator and P_2 on the meridian of the satellite, then $\theta_1 = \theta_2 = 8.7°$ and $\varphi_1 = 90°$. The angular difference as determined from the true view representation is $P_1P_2 = 12.30°$, while the actual angle on the satellite centred sphere is $\chi = 12.28°$. The error incurred is less than 0.02°, and the relative error is less than 2×10^{-3}. It is therefore permissible, at least for a geostationary satellite, to assume that angular distances are respected. Geometrical transformation (such as translation) can, therefore, be performed using this representation.

9.7.2 Geometrical contour

The geometrical contour is derived from the service zone contour, taking into account the antenna pointing error. Therefore, a zone of uncertainty is assigned to each point of the service zone contour. It is a circle centred on the considered point with radius equal to the angular pointing error. This circle should encompass all apparent displacements of the considered point, resulting from the antenna pointing error. The geometrical contour is a contour that surrounds the service zone augmented by all uncertainty zones (see Figure 9.36b, later in the chapter).

The geometrical contour should take into account the combined effects of depointing of the antenna boresight due to the satellite motion and the deformation of the service zone contour due to the relative displacement of the satellite with respect to the considered service zone.

When the antenna is provided with a pointing control mechanism, the pointing error still remains: indeed, the inaccuracy of the pointing control mechanism and the effect of the apparent movement of the satellite with respect to the service zone modify the true view angles of the points on the service zone contour. This also occurs when the coverage of a given service zone must be obtained from two different orbital positions. This is the case for a system with two geostationary satellites of identical design, or having a common standby satellite.

9.7.3 Global coverage

Global coverage is achieved when the geometrical contour encompasses the visible region of the earth corresponding to a given minimum elevation angle on the ground.

9.7.3.1 Maximum geographical contour extent

The geographical contour is limited by the great circle on the earth along which the cone having the satellite as its vertex is tangent to the earth. The elevation angle on this contour is equal to zero. For a geostationary satellite, this cone has a vertex angle 2θ equal to 17.4°.

9.7.3.2 Contour for minimum elevation angle

Earth stations located on the contour with zero elevation angle would have their antennas pointing horizontally. Compared to earth stations operating at higher elevation angles, poorer performance of the links is likely as a consequence of the increased atmospheric propagation

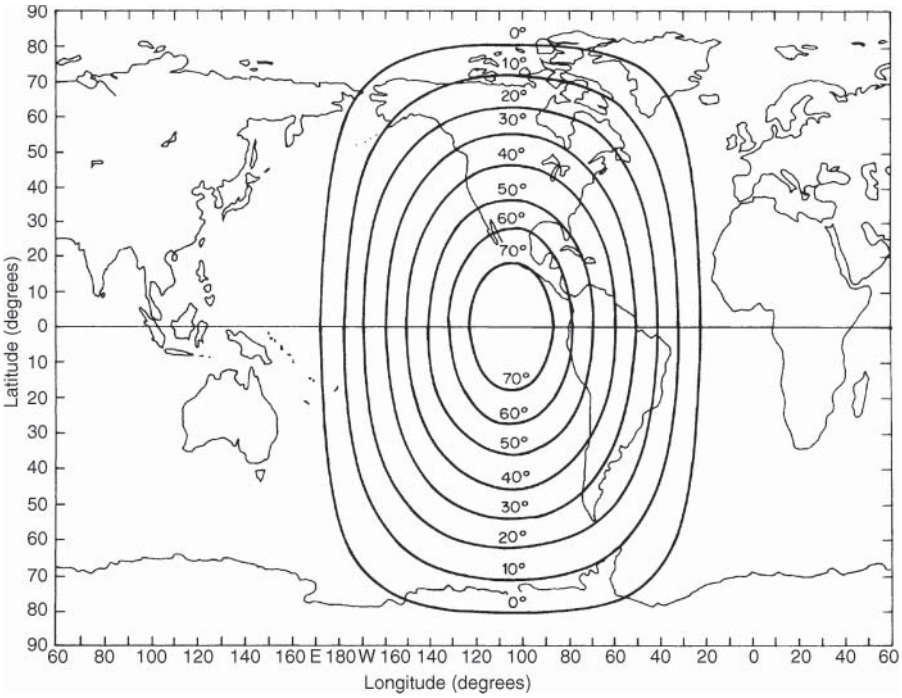


Figure 9.26 Global coverage of a geostationary satellite located at 105°W according to the minimum elevation angle value.

attenuation due to the greater length of the slant path of the wave in the atmosphere and, for the downlink, as a result of the increased earth station antenna noise temperature. So, in practice, the geometrical contour is often defined by the curve along which the direction of the satellite from the earth station makes a certain angle $E > 0°$ with the horizontal. This angle corresponds to the minimum elevation angle $E_{\min}$. For earth stations within the contour, E is larger than $E_{\min}$.

The contours for different minimum elevation angles are represented in Figure 9.26 for the case of a geostationary satellite at 105°W. A practical global coverage usually corresponds to a minimum elevation angle $E_{\min} = 10°$.

Figure 9.27 gives the elevation angle E of the earth station antenna and the nadir angle u (the angle between the satellite-to-earth-station direction and the satellite-to-earth-centre direction) as a function of the geocentric angle ϕ between the earth-centre-to-earth-station direction and the earth-centre-to-satellite direction. The angle ϕ is obtained as a function of the latitude l of the station and the relative longitude L by means of Eq. (2.63):

$$\cos \phi = \cos l \cos L$$

The angular width of the geographical region seen from a geostationary satellite for a minimum elevation angle $E_{\min}$ is fixed by the value 2θ determined for $E = E_{\min}$ in the equation referring to (2.46c):

$$2\theta = 2\text{arc} \sin[(R_E \cos E)/(R_0 + R_E)] = 2\text{arc} \sin(0.15 \cos E_{\min}) \quad \text{(degree)} \tag{9.25}$$

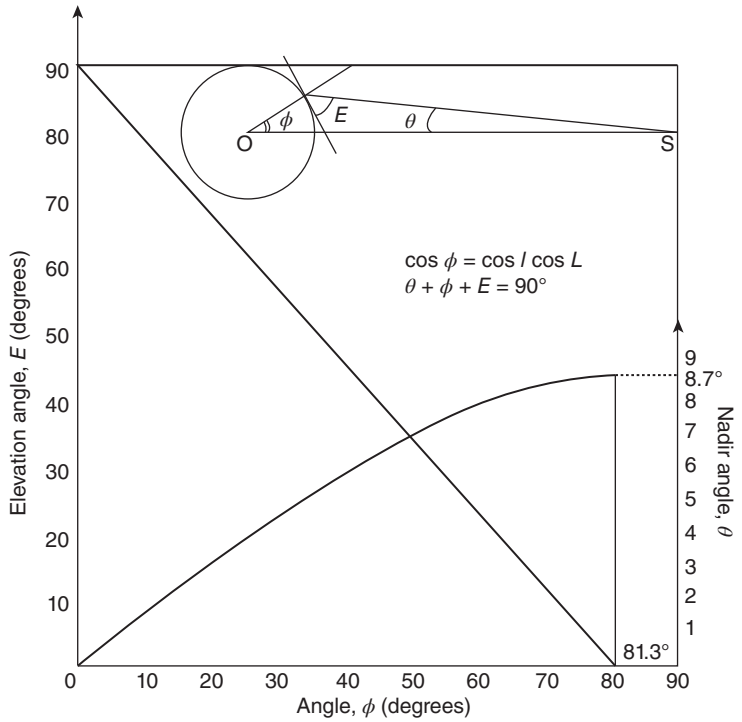


Figure 9.27 Elevation angle E and nadir angle θ as a function of geocentric angle ϕ in the plane defined by the satellite, the earth station, and the centre of the Earth.

The maximum latitude of the geographical contour is given by:

$$l_{\max} = 90^{\circ} - (\theta + E_{\min}) \quad \text{(degree)} \tag{9.26}$$

The antenna illuminating the service zone is assumed to have its boresight directed towards the centre of the earth. If the pointing error $\Delta\theta$ is taken into account, the angular width of the geometrical contour should then be equal to $2\theta + 2\Delta\theta$. If, for example, $\Delta\theta = 1°$, the angular width of the global geometrical contour for a minimum elevation angle $E_{\min} = 10°$ will be $2\theta = 2\arcsin(0.15 \sin E_{\min}) + 2\ \Delta\theta = 2\arcsin(0.15 \sin 10°) + 2° = 19°$.

9.7.4 Reduced or spot coverage

When the coverage is not global, then it relates to a particular region of the earth as viewed from the satellite. The antenna boresight does not pass through the centre of the earth but through a reference point on the surface of the earth which is defined as the *centre of coverage*.

The geometry for a geostationary satellite is illustrated in Figure 9.28. S represents the satellite and P the *centre of coverage*. In the absence of pointing error, SP represents the antenna boresight and its direction is determined by two angles in a reference frame associated with the satellite. These angles can be the two true view angles θ and φ defined in Section 9.7.1.1, which are shown in Figure 9.28 in the xy reference frame. In this frame, four other angles related to the direction of the antenna boresight SP are also displayed: α, α^*, β, β^*.

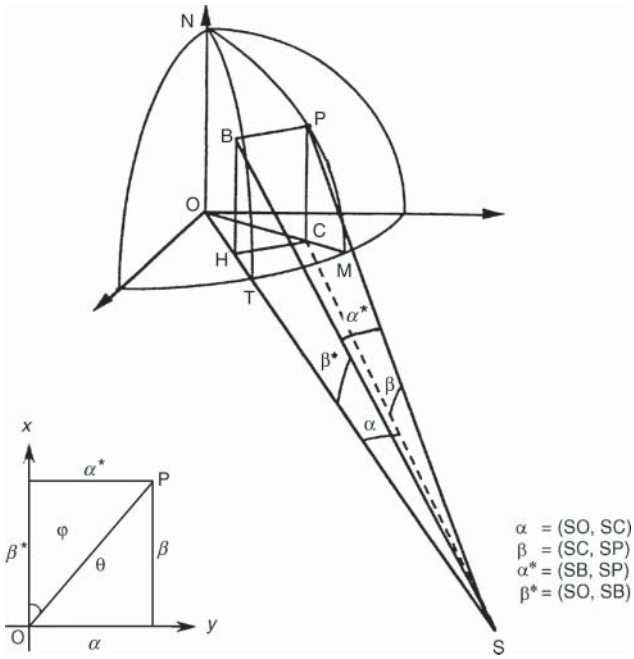


Figure 9.28 Geostationary satellite–earth geometry: definition of angles for direction of satellite antenna boresight SP from satellite S to considered location P on the earth surface.

The angle α between the direction of the centre of the earth SO and the projection SC of the satellite antenna boresight SP on the orbital plane is called the *satellite antenna azimuth angle*. It is given by:

$$\alpha = \arctan \frac{R_E \cos l \sin L}{R_0 + R_E(1 - \cos l \cos L)} \tag{9.27}$$

where $R_E = 6378$ km is the mean equatorial radius of the earth, $R_0 = 35\,786$ km is the nominal geostationary satellite altitude, and l and L are the latitude and the longitude difference, respectively, with respect to that of the satellite for point P. The angle β between the satellite antenna boresight SP and its projection SC on the orbital plane is called the *satellite antenna elevation angle*. It is given by:

$$\beta = \arctan \frac{R_E \sin l \cos \alpha}{R_0 + R_E(1 - \cos l \cos L)} \tag{9.28}$$

The angle β^* between the direction of the centre of the earth SO and the projection SB of the satellite antenna boresight SP on the meridian plane of the satellite is given by:

$$\beta^* = \arctan \frac{R_E \sin l}{R_0 + R_E(1 - \cos l \cos L)} \tag{9.29}$$

The angle α^* between the satellite antenna boresight SP and its projection SB on the meridian plane of the satellite is given by:

$$\alpha^* = \arctan \frac{R_E \cos l \sin L \cos \beta^*}{R_0 + R_E(1 - \cos l \cos L)} \tag{9.30}$$

The true view angles can now be identified. The angle θ, which is the nadir angle, is the angle between the direction of the earth centre SO and the satellite antenna boresight SP. It is given by:

$$\cos\theta = \cos\alpha\cos\beta \tag{9.31a}$$

or by:

$$\cos\theta\frac{R_0 + R_E(1 - \cos l\cos L)}{R} \tag{9.31b}$$

where R is the distance from the satellite to the earth location P. R can be calculated using:

$$R = \sqrt{R_0^2 + 2R_E(R_E + R_0)(1 - \cos l\cos L)} \tag{9.31c}$$

The angle φ is the angle between the meridian plane of the satellite and the plane defined by the direction of the centre of the earth and the satellite antenna boresight. It is given by:

$$\varphi = \arctan\frac{\sin L}{\tan l} \tag{9.32a}$$

if P is located in the northern hemisphere ($l \geq 0$), and by:

$$\varphi = \pi + \arctan\frac{\sin L}{\tan l} \tag{9.32b}$$

if P is located in the southern hemisphere ($l < 0$).

9.7.5 Evaluation of antenna pointing error

Correct pointing of the satellite antenna occurs when the direction of its boresight has true view angles θ and φ equal to the true view angles of the centre of the service zone. Considering a rigidly mounted antenna on the satellite body, depointing originates from attitude motion of the satellite about its centre of mass as a result of imperfect attitude control, and of the satellite with respect to its nominal orbital position as a result of imperfect station keeping. In addition, pointing errors result from initial satellite antenna boresight misalignment at mounting and deformations due to mechanical and thermal constraints during the satellite operation lifetime. The depointing angle $\Delta\theta$ for each source of depointing is resolved into two components: $\Delta\theta_y$ parallel to the equatorial plane, along the y axis of Figure 9.23, and $\Delta\theta_x$ in the meridian plane of the satellite, along the x axis. The overall depointing angle is then calculated as the resultant of the two components obtained as an appropriate combination of the individual components along each axis.

9.7.5.1 Depointing due to attitude motion

Attitude is determined by the angles between the satellite body mechanical axes and three reference axes: roll, pitch, and yaw (Section 10.2 and Figure 10.4). Nominal attitude is specified as the alignment of the satellite body mechanical axes and those three reference axes. Attitude motions are resolved into rotations about those three axes. A rotation about any axis translates into a depointing of the satellite antenna boresight. The resulting depointing results in the two components mentioned earlier. These components will now be expressed.

Depointing induced by rotation about the roll axis. The rotation about the roll axis is ε_R. For a geostationary satellite, a typical upper limit of ε_R is 0.05°. In general, the antenna boresight is not perpendicular to the roll axis. As a consequence, the rotation about the roll axis induces not only

a depointing component along the x axis but also a component along the y axis. Expressions for those two components are:

(a) along the x axis:

$$\Delta\theta_{R,x} = \arctan[\tan(\beta^* + \varepsilon_R)\cos\alpha] - \beta \tag{9.33a}$$

(b) along the y axis:

$$\Delta\theta_{R,y} = \arctan\left[\frac{\cos\beta^* tan\alpha^*}{\cos(\beta^* + \varepsilon_R)}\right] - \alpha^* \tag{9.33b}$$

Approximations for $\Delta\theta_{R,x}$ and $\Delta\theta_{R,y}$ can be derived:

$$\Delta\theta_{R,x} = \varepsilon_R \cos\alpha \tag{9.34a}$$

$$\Delta\theta_{R,y} = \varepsilon_R \sin\alpha^* \cos\alpha^* \left(\tan\beta^* + \frac{\varepsilon_R}{2}\right) \quad \text{(angles in radians)} \tag{9.34b}$$

For $\varepsilon_R \leq 0.05°$, the relative error introduced by the earlier approximations is less than 10^{-3}.

Depointing induced by rotation about the pitch axis. The rotation about the pitch axis is ε_P. For a geostationary satellite, a typical upper limit of ε_P is 0.02°. The induced depointing angle components are given by:

(a) along the x axis:

$$\Delta\theta_{P,x} = \arctan\left[\frac{\cos\alpha\ \tan\beta}{\cos(\alpha + \varepsilon_P)}\right] - \beta \tag{9.35a}$$

(b) along the y axis:

$$\Delta\theta_{P,y} = \arctan[\tan(\alpha + \varepsilon_P)\cos\beta^*] - \alpha^* \tag{9.35b}$$

Approximations for $\Delta\theta_{P,x}$ and $\Delta\theta_{P,y}$ can be derived (angles in radians):

$$\Delta\theta_{P,x} = \varepsilon_P \sin\beta\cos\beta\left(\tan\alpha + \frac{\varepsilon_P}{2}\right) \tag{9.36a}$$

$$\Delta\theta_{P,y} = \varepsilon_P \cos\beta^* \tag{9.36b}$$

For $\varepsilon_P \leq 0.02°$, the relative error introduced by the earlier approximations is less than 10^{-3}.

Depointing induced by rotation about the yaw axis. The rotation about the yaw axis is ε_Y. For a geostationary satellite, a typical upper limit of ε_Y is 0.3°. The induced depointing angle components are given by:

(a) along the x axis:

$$\begin{aligned} \Delta\theta_{Y,x} &= \arctan\left[\frac{\cos(\varphi + \varepsilon_Y)\tan\beta}{\cos\varphi}\right] - \beta \quad \text{if} \quad \varphi \neq 90° \\ &= -\varepsilon_Y \sin\alpha \qquad\qquad\qquad\qquad\quad \text{if} \quad \varphi \neq 90° \end{aligned} \tag{9.37a}$$

(b) along the y axis:

$$\Delta\theta_{Y,y} = \arctan\left[\frac{\sin(\varphi + \varepsilon_Y)\tan\alpha^*}{\sin\varphi}\right] - \alpha^* \qquad \text{if} \quad \varphi \neq 0^\circ$$
$$= \varepsilon_Y \sin\beta^* \qquad \text{if} \quad \varphi \neq 0^\circ \tag{9.37b}$$

Approximations for $\Delta\theta_{Y,x}$ and $\Delta\theta_{Y,y}$ can be derived (angles in radians):

$$\Delta\theta_{Y,x} = -\varepsilon_Y \cos\beta\left(\sin\alpha\cos\beta + \frac{\varepsilon_Y}{2}\sin\beta\right) \tag{9.38a}$$

$$\Delta\theta_{Y,y} = \varepsilon_Y \cos\alpha^*\left(\cos\alpha^*\beta^* - \frac{\varepsilon_Y}{2}\sin\alpha^*\right) \tag{9.38b}$$

For $\varepsilon_Y \leq 0.3^\circ$, the relative error introduced by the earlier approximations is less than 10^{-3}.

9.7.5.2 Depointing due to orbital motion

Displacement of the centre of mass of the satellite will change the direction of the considered point on the earth surface, and hence, as the antenna boresight is in a fixed direction with respect to the satellite axes, this displacement will introduce antenna depointing.

Nominal inclination and eccentricity of the geostationary orbit are equal to zero. Due to orbit perturbations, inclination and eccentricity vary with time and do not remain equal to zero. Moreover depending on the orbital location of the satellite, the satellite exhibits a long-term longitudinal drift.

Nonzero inclination mainly results in a north–south (NS) displacement with 24 hour periodicity. Nonzero eccentricity turns into east–west (EW) and radial displacements with the same periodicity. The long-term longitudinal drift is continuous, either eastward or westward. The amplitude of the overall orbital motion is specified by the station-keeping (SK) box, which is stipulated as a solid angle at the centre of the earth. The depointing resulting from these orbital motions is resolved into two x and y components, which will now be evaluated.

Depointing induced by nonzero inclination. While orbiting with nonzero inclination, the satellite crosses the plane of the equator at nodes and alternately finds itself above or below the plane of the equator (Figure 9.29). Maximal latitudinal displacement is achieved at 90° from the nodes (vertex of the orbit) and is equal to i, where i is the inclination value. On the other hand, at nodes the latitudinal displacement is 0, but the inclination turns into a depointing similar to the depointing induced by a rotation about the yaw axis with angle i. The satellite also displays a longitudinal displacement that is maximum in between and equal to $4.36 \times 10^{-3}\, i^2$ degrees, where the inclination i is expressed in degrees (Section 2.2.3). For geostationary satellites, the value of inclination i is small enough to allow neglecting this longitudinal displacement ($i < 0.1^\circ$ during most of the operational lifetime and up to about 6° at the end of the lifetime when north–south station keeping is relaxed during the so-called *inclined orbit operation*).

Consider depointing induced by the *latitudinal displacement*. The latitudinal displacement is contained within the SK box. Maximum depointing is then dependent on the north–south angular half-width NS of the window. Latitudinal displacement within the window corresponds to a rotation of the orbital plane about the line of nodes of the orbit. As a consequence of the slanting

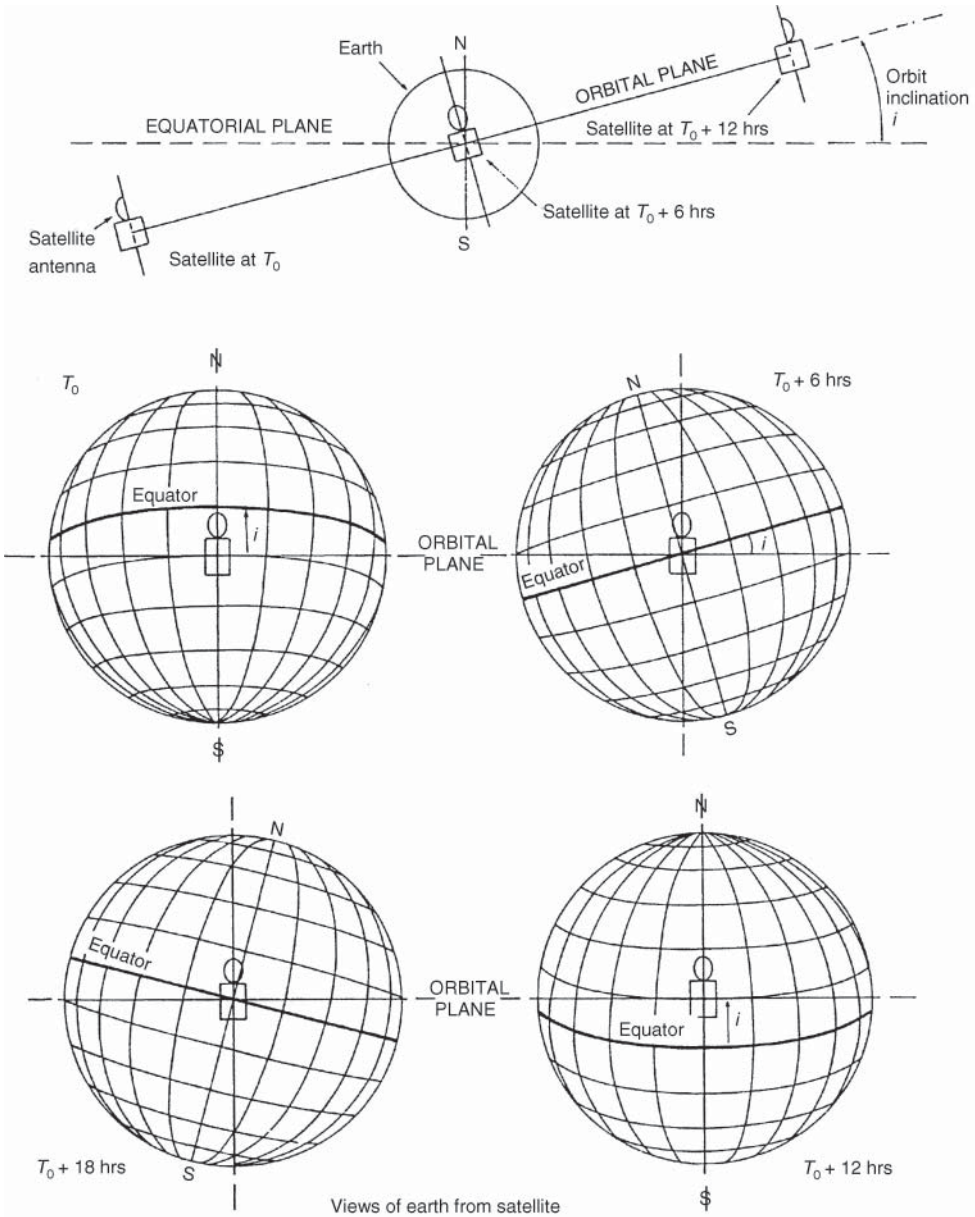


Figure 9.29 Apparent movement of the earth seen from a quasi-geostationary satellite subject to an orbit inclination $i \neq 0$.

direction of aiming with respect to the satellite meridian, this rotation induces not only a depointing component along the x axis but also a component along the y axis. Expressions for those two components are:

(a) along the x axis:

$$\begin{aligned}\Delta\theta_{\mathrm{NS},x} &= \arctan\left[\tan\beta\left(1-\frac{\cos L}{\tan l}\tan NS\right)\right]-\beta \quad \text{if} \quad l\neq 0\\ &= \arctan\left[\tan\beta-\frac{\sin\alpha}{\tan L}\tan NS\right]-\beta \qquad \text{if} \quad l=0, L\neq 0\\ &= -\frac{\mathrm{R_E}}{\mathrm{R_0}}NS \qquad\qquad\qquad\qquad \text{if} \quad l=0, L=0\end{aligned} \tag{9.39a}$$

(b) along the y axis:

$$\Delta\theta_{NS,y} = \arctan[\tan\alpha^*(1+\tan\beta^*\sin NS)]-\alpha^* \tag{9.39b}$$

Approximations for $\Delta\theta_{\mathrm{NS},x}$ and $\Delta\theta_{\mathrm{NS},y}$ can be derived:

(a) along the x axis:

$$\begin{aligned}\Delta\theta_{\mathrm{NS},x} &= -NS\ \sin\beta\cos\beta\frac{\cos L}{\tan l} \quad \text{if} \quad l\neq 0\\ &= -NS\frac{\sin\alpha}{\tan L}\ \cos^2\beta \qquad \text{if} \quad l=0, L\neq 0\\ &= -NS\frac{\mathrm{R_E}}{\mathrm{R_0}} \qquad\qquad \text{if} \quad l=0, L=0\end{aligned} \tag{9.40a}$$

(b) along the y axis:

$$\Delta\theta_{\mathrm{NS},y} = NS\sin\alpha^*\cos\alpha^*\tan\beta^* \tag{9.40b}$$

For $NS \leq 0.1°$, the relative error introduced by the earlier approximations is less than 10^{-2}.

Now consider depointing induced by the *inclination of the orbit at nodes*. The depointing is similar to that induced by a rotation about the yaw axis. Using the formulae derived above, one can obtain the expression for those two components:

(a) along the x axis:

$$\begin{aligned}\Delta\theta_{i,x} &= \arctan\left[\frac{\cos(\varphi+i)\tan\beta}{\cos\varphi}\right]-\beta \quad \text{if} \quad \varphi\neq 90°\\ &= -i\sin\alpha \qquad\qquad\qquad\qquad \text{if} \quad \varphi\neq 90°\end{aligned} \tag{9.41a}$$

(b) along the y axis:

$$\begin{aligned}\Delta\theta_{\mathrm{i,y}} &= \arctan\left[\frac{\sin(\varphi+i)\tan\alpha^*}{\sin\varphi}\right]-\alpha^* \text{if} \quad \varphi\neq 0°\\ &= i\sin\beta^* \qquad\qquad\qquad\qquad \text{if} \quad \varphi\neq 0°\end{aligned} \tag{9.41b}$$

Approximations for $\Delta\theta_{\mathrm{i,x}}$ and $\Delta\theta_{\mathrm{i,y}}$ can be derived:

(a) along the x axis:

$$\Delta\theta_{\mathrm{i,x}} = -i\cos\beta\left[\sin\alpha\cos\beta+\frac{i}{2}\sin\beta\right] \quad \text{(angles in radians)} \tag{9.42a}$$

(b) along the y axis:

$$\Delta\theta_{i,y} = i\cos\alpha^* \left[\cos\alpha^* \sin\beta^* - \frac{i}{2}\sin\alpha^*\right] \quad \text{(angles in radians)} \tag{9.42b}$$

For $i \leq 0.1°$, the relative error introduced by the earlier approximations is less than 10^{-4}.

At nodes, the nonzero inclination causes an apparent rotation of the antenna beam with respect to an earth station about the line which joins the satellite and the station concerned. The maximum value of the angle of rotation is equal to the inclination i for a station situated on the equator at the longitude of the satellite (at the sub-satellite point). In the case where the links are using waves with orthogonal polarisation, assuming that each polarised carrier would have power C at zero inclination, then for nonzero inclination i, the antenna receives on a given wanted polarisation a power equal to $C\cos^2 i$ along with an interfering orthogonally polarised component with power equal to $C\sin^2 i$. The carrier-to-interference power ratio then is $(C/N)_i = (C\cos^2 i)/(C\sin^2 i)$, whose value in dB is $-20\log(\tan i)$: that is, 19.6 dB for $i = 6°$. This may be a problem and may prevent reuse of the same band of frequencies with two orthogonal linear polarisations. Reuse of frequencies by orthogonal linear polarisation, which enables the capacity in a given band of frequencies to be doubled, requires that the satellite should be maintained on an orbit with sufficiently small inclination.

Depointing due to east–west motion. Depointing due to east–west motion in the SK box is equivalent to rotation of the reference frame about the earth polar axis. Depointing is maximum when the satellite reaches the boundary of the SK box, and therefore the depointing components express as a function of the angular half-width EW of the box:

(a) along the x axis:

$$\Delta\theta_{EW,x} = \arctan[\tan\beta(1 + \tan\alpha \sin EW)] - \beta \tag{9.43a}$$

(b) along the y axis:

$$\begin{aligned}
\Delta\theta_{EW,y} &= \arctan\left[\tan\alpha^*\left(1 - \frac{\tan EW}{\tan L}\right)\right] - \alpha^* && \text{if} \quad L \neq 0 \\
&= \arctan\left[\tan\alpha^* - \frac{\sin\beta^*}{\tan l}\tan EW\right] - \alpha^* && \text{if} \quad L = 0 \;\; l \neq 0 \\
&= -\frac{R_E}{R_0}EW && \text{if} \quad L = 0 \;\; l = 0
\end{aligned} \tag{9.43b}$$

Approximations for $\Delta\theta_{EW,x}$ and $\Delta\theta_{EW,y}$ can be derived:

(a) along the x axis:

$$\Delta\theta_{EW,x} = EW\tan\alpha\sin\beta\cos\beta \tag{9.44a}$$

(b) along the y axis:

$$\begin{aligned}
\Delta\theta_{EW,y} &= -EW\,\frac{\sin\alpha^*\cos\alpha^*}{\tan L} && \text{if} \quad L \neq 0 \\
&= -EW\frac{\sin\beta^*\cos^2\alpha^*}{\tan l} && \text{if} \quad L = 0 \text{ and } l \neq 0 \\
&= -\frac{R_E}{R_0}EW && \text{if} \quad L = 0 \text{ and } l = 0
\end{aligned} \tag{9.44b}$$

For EW $\leq 0.08°$, the relative error introduced by the approximation is less than 10^{-3}.

Depointing induced by nonzero eccentricity. Nonzero eccentricity turns into longitudinal and radial displacements with 24 hour periodicity. Each contributes to the depointing. The peak amplitude expressed in degrees of the longitudinal displacement is equal to 114*e*, and the radial displacement is *ae*, where e is the eccentricity and *a* the semi-major axis of the orbit (Eq. (2.55)). The depointing induced by the longitudinal displacement due to eccentricity is included within the depointing corresponding to a satellite position at the boundary of the SK box. This has already been dealt with in Section 9.7.5.1.

Depointing induced by radial displacement is larger when the satellite is at perigee. Expressions for the two components of depointing at perigee are:

(a) along the x axis:

$$\Delta\theta_{\mathrm{e,x}} = \arctan\left[\frac{R_0 + R_E(1-\cos l\cos L)\tan\beta}{R_0 + R_E(1-\cos l\cos L) - \mathrm{e}(R_0+R_E)}\right] - \beta \tag{9.45a}$$

(b) along the y axis:

$$\Delta\theta_{\mathrm{e,y}} = \arctan\left[\frac{R_0 + R_E(1-\cos l\cos L)\tan\alpha^*}{R_0 + R_E(1-\cos l\cos L) - \mathrm{e}(R_0+R_E)}\right] - \alpha^* \tag{9.45b}$$

Approximations for $\Delta\theta_{\mathrm{e,x}}$ and $\Delta\theta_{\mathrm{e,y}}$ can be derived:

(a) along the x axis:

$$\Delta\theta_{\mathrm{e,x}} = \mathrm{e}\left(\frac{180}{\pi}\right)\frac{R_0+R_E}{R\cos\theta}\ \sin\beta\cos\beta\ \text{(degree)} \tag{9.46a}$$

(b) along the y axis:

$$\Delta\theta_{\mathrm{e,y}} = \mathrm{e}\left(\frac{180}{\pi}\right)\frac{R_0+R_E}{R\cos\theta}\ \sin\alpha^*\cos\alpha^*\ \text{(degree)} \tag{9.46b}$$

where, see Eq. (9.31b), $R\cos\theta = R_0 + R_E\,(1-\cos l\,\cos L)$.

For $\mathrm{e} \leq 0.0008$, the relative error introduced by the earlier approximations is less than 10^{-3}.

9.7.5.3 *Evaluation of the overall depointing due to satellite motion*

Pointing errors are random and deterministic variables with varying degrees of correlation. A definition of pointing error must take into consideration the probability of occurrence of the event that has led to this value being obtained. A random variable is usually specified by its mean value and its standard deviation; the 3σ value corresponds to a probability of 99.73% that the value is not exceeded considering Gaussian variables.

When the values of the various components have been determined, the overall depointing must be evaluated by combining them. Depointing due to displacement of the satellite has been decomposed into perpendicular axes oriented along the north–south and east–west axes. Depointing due to the initial boresight misalignment may have any direction and is, a priori, independent of depointing due to displacement of the satellite (see Section 9.7.5.4).

The earlier calculations provide the individual depointing components along the x and y axes. It is required to estimate the overall contribution to depointing along each of these axes. The overall depointing angle is the resultant of these two orthogonal contributions. Two problems have to be addressed:

— What is the appropriate combination of the individual components along each x and y axis?
— How is the overall depointing angle constructed from its two x and y components?

These problems were addressed in [BEN-86], where the error components considered are identified at subsystem level. These errors are then grouped in several classes according to their temporal characteristics: constant, long-term varying, diurnally varying, and short-term varying. Here the individual components are identified at system level only, and a different approach is followed.

Combination or individual components along each axis. The overall component of the depointing along any axis is a combination of deterministic and random variables.

The *deterministic* components result from the motion of the satellite within the SK box. They are the depointing components induced by:

— North–south drift ($\Delta\theta_{\mathrm{NS}}$)
— nonzero inclination ($\Delta\theta_{\mathrm{i}}$)
— East–west drift ($\Delta\theta_{\mathrm{EW}}$)
— Nonzero eccentricity ($\Delta\theta_{\mathrm{e}}$)

The worst-case value of the overall deterministic component is the algebraic sum of the maximum values of the *non-exclusive* deterministic individual components. Indeed, in evaluating the overall contribution, attention must be paid to the fact that some components are mutually exclusive. For instance, a nonzero inclination generates a yaw-like depointing when the satellite passes the nodes of the orbit and turns into a depointing induced by the north–south latitudinal displacement when the satellite is 90° away from the orbit nodes (the vertex of the orbit). These two depointing components are mutually exclusive as they do not occur at the same time. Therefore, one should calculate the overall components along each axis of the deterministic depointing due to the satellite motion within the SK box at node as follows:

$$\Delta\theta_{\mathrm{SK},x,\mathrm{node}} = \Delta\theta_{i,x} + \Delta\theta_{\mathrm{EW},x} + \Delta\theta_{\mathrm{e},x}$$
$$\Delta\theta_{\mathrm{SK},y,\mathrm{node}} = \Delta\theta_{i,y} + \Delta\theta_{\mathrm{EW},y} + \Delta\theta_{\mathrm{e},y}$$

and at vertex as follows:

$$\Delta\theta_{\mathrm{SK},x,\mathrm{vertex}} = \Delta\theta_{\mathrm{NS},x} + \Delta\theta_{\mathrm{EW},x} + \Delta\theta_{\mathrm{e},x}$$
$$\Delta\theta_{\mathrm{SK},y,\mathrm{vertex}} = \Delta\theta_{\mathrm{NS},y} + \Delta\theta_{\mathrm{EW},y} + \Delta\theta_{\mathrm{e},y}$$

The *random* components arise from satellite altitude movement, deformation due to mechanical and thermal constraints, and so on. In the following, the considered depointing components are those induced by rotation about:

— The roll axis $\Delta\theta_{\mathrm{R}}$
— The pitch axis $\Delta\theta_{\mathrm{P}}$
— The yaw axis $\Delta\theta_{\mathrm{Y}}$

Assuming the individual random components are independent, the σ value of the overall random depointing component along any axis is obtained as the square root of the sum of the σ^2 values of all individual random depointing components.

The 3σ value of any depointing component can be considered proportional to the 3σ value parameter, characterising the cause for depointing. Indeed, the magnitude of the satellite movement is small enough for the earlier approximations to apply, and proportionality between a depointing component and the parameter measuring the cause for depointing can be assumed.

Along each x and y axis, the 3σ value of the overall random component generated by attitude control (AC) errors, $\Delta\theta_{\mathrm{AC}}$, is given by:

$$\Delta\theta_{\mathrm{AC},x} = [\Delta\theta_{\mathrm{R},x}^2 + \Delta\theta_{\mathrm{P},x}^2 + \Delta\theta_{\mathrm{Y},x}^2]^{1/2}$$

$$\Delta\theta_{\mathrm{AC},y} = [\Delta\theta_{\mathrm{R},y}^2 + \Delta\theta_{\mathrm{P},y}^2 + \Delta\theta_{\mathrm{Y},y}^2]^{1/2}$$

where $\Delta\theta_{\mathrm{R},x}$, $\Delta\theta_{\mathrm{P},x}$, $\Delta\theta_{\mathrm{Y},x}$ are the 3σ values of the depointing components along the x axis and $\Delta\theta_{\mathrm{R},y}$, $\Delta\theta_{\mathrm{P},y}$, $\Delta\theta_{\mathrm{Y},y}$ are the 3σ values of the depointing components along the y axis. These depointing components can be calculated using Eqs. (9.33a), (9.33b), (9.35a), (9.35b), and (9.37a) and (9.37b), considering that ε_{R}, ε_{P}, and ε_{Y} are the 3σ values for the satellite attitude control errors.

Finally, the *overall* depointing components along each axis, $\Delta\theta_x$ and $\Delta\theta_y$, are obtained by adding the worst-case value of the overall deterministic component to the 3σ value of the random component.

Assuming that the depointing due to eccentricity results only from the radial displacement (the longitudinal displacement is accounted for by the SK box size), the component expressions at node are:

$$\Delta\theta_{x,\mathrm{node}} = \Delta\theta_{\mathrm{AC},x} + \Delta\theta_{\mathrm{SK},x,\mathrm{node}}$$

$$\Delta\theta_{y,\mathrm{node}} = \Delta\theta_{\mathrm{AC},y} + \Delta\theta_{\mathrm{SK},y,\mathrm{node}}$$

and at vertex are:

$$\Delta\theta_{x,\mathrm{vertex}} = \Delta\theta_{\mathrm{AC},x} + \Delta\theta_{\mathrm{SK},x,\mathrm{vertex}}$$

$$\Delta\theta_{y,\mathrm{vertex}} = \Delta\theta_{\mathrm{AC},y} + \Delta\theta_{\mathrm{SK},y,\mathrm{vertex}}$$

Combining the overall x and y components. The depointing due to all satellite motions (attitude and station keeping), $\Delta\theta_m$, is obtained from the relevant combination of the overall x and y components. From earlier, there are two values for the overall components: at nodes and at vertex. Then the depointing $\Delta\theta_{\mathrm{m}}$ to be considered is given by:

$$\Delta\theta_{\mathrm{m}} = \max[\Delta\theta_{\mathrm{m,node}}\Delta\theta_{\mathrm{m,vertex}}]$$

where $\max[X,Y]$ represents the larger of X or Y.

The simplest approach to obtain the depointing due to satellite motion, $\Delta\theta_m$, is to consider the square root of the sum of the squares of the overall $\Delta\theta_x$ and $\Delta\theta_y$ components:

$$\Delta\theta_{\mathrm{m}} = [(\Delta\theta_x)^2 + (\Delta\theta_y)^2]^{1/2}$$

By using the component values $\Delta\theta_x$ and $\Delta\theta_y$ as defined earlier, which correspond to a probability of 99.73% of not being exceeded, there is a greater probability that the overall depointing will not exceed this value. This approach may be too pessimistic.

Estimating a depointing value with a given probability of not being exceeded is not a simple matter; if the two components are independent and have a Gaussian distribution with zero mean and the same variance σ^2, it is known that the distribution of the depointing angle $\Delta\theta_{\mathrm{m}}$ is a Rayleigh distribution. The value of depointing that would not be exceeded with a given probability can be obtained from the Rayleigh cumulative distribution function:

$$F(\Delta\theta_{\mathrm{m}}) = 1 - \exp[-(\Delta\theta_{\mathrm{m}})^2/\sigma^2]$$

For instance, a depointing angle of $3:44\sigma$, where σ is the standard deviation of each of the two components, is not exceeded with a probability of 99.73%. If the two components have different means and a common variance, the distribution of the depointing angle is a Rice distribution.

In practice the two components have different means and variances, and there is no general expression for the cumulative distribution function.

9.7.5.4 Total pointing error

The total pointing error $\Delta\theta$ should take into consideration the depointing $\Delta\theta_m$, due to the satellite motion, and the depointing $\Delta\theta_{bor}$, associated with the initial boresight misalignment and deformations due to mechanical and thermal constraints.

Indeed, the direction of the actual boresight can differ from the nominal direction, as defined in the antenna reference frame, either as a result of initial misalignment when mounting the antenna on the satellite or measuring the radiation pattern, or deformation of the antenna reflector. Such deformation may be caused by temperature differences, depending on the variable orientation of the sun.

The initial boresight misalignment is a random variable when one considers the manufacturing of many satellites. However, for a given satellite, its value is fixed, and the resulting depointing can be considered as a deterministic component of the overall pointing error. Therefore, at worst, considering that its unknown orientation could be aligned with that of $\Delta\theta_m$, its value should be added to $\Delta\theta_m$:

$$\Delta\theta = \Delta\theta_m + \Delta\theta_{bor}$$

Example 9.1 An example calculation is now given in order to illustrate the earlier derivations. This example concerns a geostationary satellite with the following parameters:

— Satellite attitude control accuracy (3σ values):

$$\varepsilon_R = 0.05°, \varepsilon_P = 0.03°, \varepsilon_Y = 0.5°$$

— Orbit inclination: $i = 0.07°$
— Orbit eccentricity: $e = 5 \times 10^{-4}$
— SK box: NS = EW = ± 0.1°
— Nominal location the antenna is aiming at:
l = =45°N
L (relative longitude with respect to the satellite) = 60°
— Initial boresight misalignment: $\Delta\theta_{bor} = 0.03°$

The SK box is large enough to accommodate the daily orbital displacement with some margin. The maximum latitudinal displacement due to inclination = $i = \pm 0.07°$; the maximum longitudinal displacement due to eccentricity = $\pm 114\, e = \pm 114 \times 5 \times 10^{-4} = \pm 0.06°$.

The values given in Table 9.3 are computed from the equations given in Sections 9.7.4 and 9.7.5:

— The set of angles (α, β), (α^*, β^*), and (θ,φ), any of which can be used to define the nominal direction in which the antenna should point
— The individual components of depointing along the x- and y-axes

Recall that the depointing due to eccentricity $\Delta\theta_e$ results from only the radial displacement. Indeed, the longitudinal displacement is $\pm 114e = \pm 0.06°$ and is less than the EW SK box size. The overall components of the depointing due to the satellite motion within the SK box are:at node:

$$\Delta\theta_{SK,x,node} = \Delta\theta_{i,x} + \Delta\theta_{EW,x} + \Delta\theta_{e,x} = -0.0023°$$

$$\Delta\theta_{SK,y,node} = \Delta\theta_{i,y} + \Delta\theta_{EW,y} + \Delta\theta_{e,y} = 0.0051°$$

at vertex:

$$\Delta\theta_{SK,x,vertex} = \Delta\theta_{NS,x} + \Delta\theta_{EW,x} + \Delta\theta_{e,x} = -0.0011°$$

$$\Delta\theta_{SK,y,vertex} = \Delta\theta_{NS,y} + \Delta\theta_{EW,y} + \Delta\theta_{e,y} = 0.0016°$$

Table 9.3 Satellite antenna pointing angle and depointing components under the conditions of Example 9.1

$\alpha = 5.59°$	$\beta = 6.42°$
$\alpha^* = 5.55°$	$\beta^* = 6.45°$
$\theta = 8.5°$	$\varphi = 40.9°$
$\Delta\theta_{R,x} = 0.0498°$	$\Delta\theta_{R,y} = 0.0006°$
$\Delta\theta_{P,x} = 0.0003°$	$\Delta\theta_{P,y} = 0.0298°$
$\Delta\theta_{Y,x} = -0.0483°$	$\Delta\theta_{Y,y} = 0.0554°$
$\Delta\theta_{NS,x} = -0.0056°$	$\Delta\theta_{NS,y} = 0.0011°$
$\Delta\theta_{i,x} = -0.0067°$	$\Delta\theta_{i,y} = 0.0078°$
$\Delta\theta_{EW,x} = 0.0011$	$\Delta\theta_{EW,y} = -0.0056°$
$\Delta\theta_{e,x} = 0.0034°$	$\Delta\theta_{e,y} = 0.0029°$

The standard deviation of depointing due to attitude control is obtained as the square root of the sum of the σ^2 values of the individual components. The corresponding 3σ values are:

$$\Delta\theta_{AC,x} = [\Delta\theta_{R,x}^2 + \Delta\theta_{P,x}^2 + \Delta\theta_{Y,x}^2]^{1/2} = 0.0694°$$

$$\Delta\theta_{AC,y} = [\Delta\theta_{R,y}^2 + \Delta\theta_{P,y}^2 + \Delta\theta_{Y,y}^2]^{1/2} = 0.0629°$$

The overall x and y components of depointing are:at nodes:

$$\Delta\theta_{x,node} = \Delta\theta_{AC,x} + \Delta\theta_{SK,x,node} = 0.067°$$

$$\Delta\theta_{y,node} = \Delta\theta_{AC,y} + \Delta\theta_{SK,y,node} = 0.068°$$

at vertex:

$$\Delta\theta_{x,vertex} = \Delta\theta_{AC,x} + \Delta\theta_{SK,x,vertex} = 0.068°$$

$$\Delta\theta_{y,vertex} = \Delta\theta_{AC,y} + \Delta\theta_{SK,y,vertex} = 0.061°$$

A pessimistic value of the depointing is:at nodes:

$$\Delta\theta_{m,node} = [\Delta\theta_{x,node}^2 + \Delta\theta_{y,node}^2]^{1/2} = 0.096°$$

at vertex:

$$\Delta\theta_{m,vertex} = [\Delta\theta_{x,vertex}^2 + \Delta\theta_{y,vertex}^2]^{1/2} = 0.092°$$

The worst case is: $\Delta\theta_m = \max[\Delta\theta_{m,\,node}, \Delta\theta_m, \text{vertex}] = 0.096°$

An optimistic approach would be to consider the depointing as equal to the value of the larger overall component:

$$\Delta\theta_m = 0.068°$$

If the overall x and y components are considered independent, zero-mean random variables and assumed to represent the 3σ values for the depointing components, then the standard deviation of each overall component is $\sigma \cong \Delta\theta_x/3 \cong \Delta\theta_y/3 = 0.023°$. Hence the depointing corresponding to a probability of 99.73% of not being exceeded is 3.44 σ, i.e. $\Delta\theta_m = 0.08°$.

The total pointing error is obtained by adding, to these Du_m values, the depointing associated with the initial boresight misalignment, $\Delta\theta_{bor} = 0.03°$. That is,

$$\Delta\theta = \Delta\theta_m + \Delta\theta_{bor} = 0.08° + 0.03° = 0.11°$$

if the overall x and y components are considered as independent, zero-mean random variables.

9.7.5.5 *Antenna provided with a pointing control system*

In order to limit the pointing error, the antenna can be equipped with a system to control the antenna pointing direction to a beacon located on the ground. An error angle measuring device located on the satellite antenna determines the deviation of the antenna pointing direction with respect to the direction of the beacon on the ground. The error signals generated are used to control the antenna pointing mechanism (APM).

The principles of operation of error angle detectors are described in Section 8.3.7.6. A system with several sources easily integrates into the network of radiating elements of a multisource antenna. A mode-extraction system is more practical when the antenna contains only a single source. According to the performance of the error angle detector and the dynamics of the pointing system, the pointing accuracy of an antenna provided with a control system is between 0.1° and 0.03°.

9.7.6 Conclusion

As two-thirds of the surface of the globe is submerged, global coverage is not well suited to service land stations only. A reduction of coverage area leads to an increase of the gain of the antenna

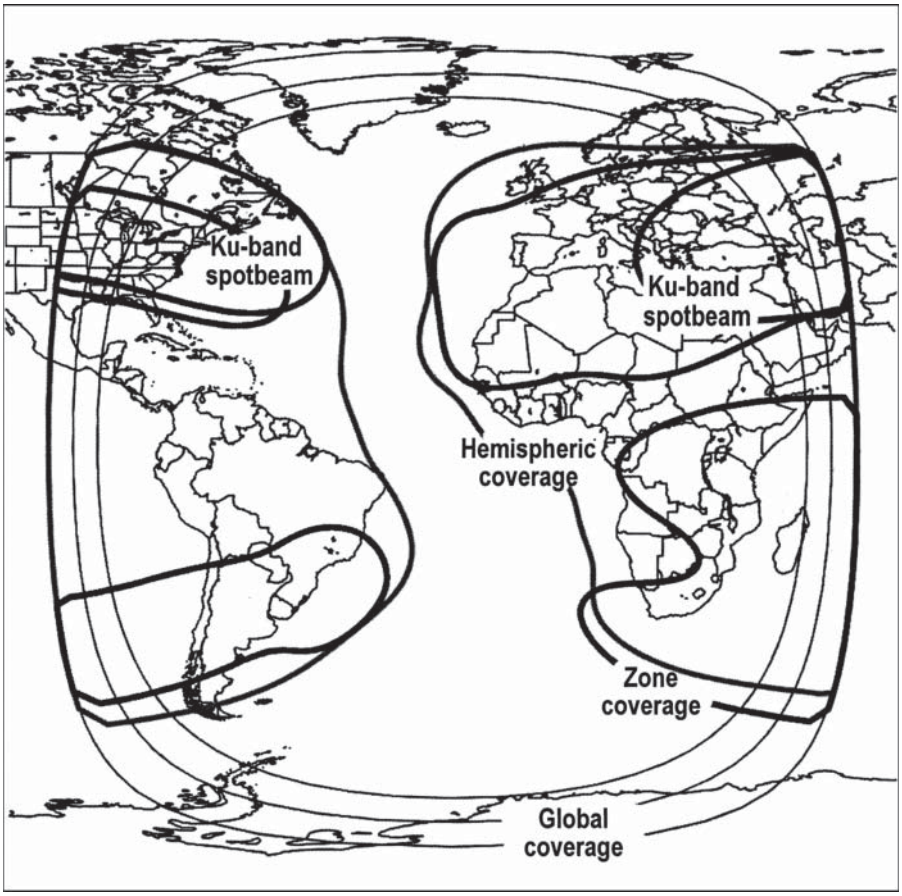


Figure 9.30 Typical Intelsat satellite coverage: global, hemispheric, zone, and spotbeam.

that provides this coverage. Furthermore, limitation of coverage to the region to be served makes it possible to reuse frequency with space diversity; two antenna beams sufficiently separated can use the same frequency with reduced mutual interference.

For international links, it is thus preferable to limit the coverage to continents (coverage of a hemisphere), regions (regional or zonal coverage), or even to zones of small extent dispersed at several points on the globe (multiple-point coverage). In the context of a national network, the coverage is limited to the country's territory (national coverage). These zones must of course be within the coverage represented in Figure 9.26 in the case of a geostationary satellite.

For example, Intelsat satellites, in addition to global coverage, provide reduced coverage (hemispheric and zones) for links at frequencies of 6/4 GHz and spot-beam coverage for zones of small extent between which a large amount of traffic flows (14/11 GHz). Such coverages are illustrated in Figure 9.30.

9.8 ANTENNA CHARACTERISTICS

9.8.1 Antenna functions

The main functions of satellite antennas are as follows:

— To *collect the radio waves* transmitted, in a given frequency band and with a given polarisation, by ground stations situated within a particular region on the surface of the earth
— To *capture as few undesirable signals* as possible (i.e. signals that do not match the characteristics stated earlier – they may be from a different region or not have the specified values of frequency or polarisation)
— To *transmit radio waves*, in a given frequency band and with a given polarisation, to a particular region on the surface of the earth
— To *transmit minimum power outside* the specified region

The link budget between the satellite and the ground depends on the EIRP. For an available transmission power P_T, the EIRP increases with the gain G_T of the transmitting antenna. Similarly on the uplink, a high G/T for the satellite requires a high value of receiving antenna gain.

A high value of antenna gain is obtained with a directional antenna. The required directivity depends on the mission to be performed – global coverage of the earth, zone, or spot coverage. Obtaining high directivity in association with good fitting of the beam to the geometrical contour to be covered permits frequency reuse by space diversity and hence more efficient use of the spectrum.

This reuse of frequencies requires antennas with reduced side lobes in order to limit interference. The ITU-R provides a *reference mask for the antenna radiation pattern* which is presented in Figure 9.31 (ITU-R Rec. S.672 – 'Satellite antenna radiation pattern for use as a design objective in the fixed-satellite service employing geostationary satellites, version 4', published in September 1997). A standard radiation pattern mask requires a well-defined centre axis for the beam as the gain variations are defined versus the off-axis angle θ. This is not the situation when shaped beams are considered (see Section 9.8.6). The mask proposed by the ITU-R defines the required gain decrease as a function of the angular distance from the EOC.

In Figure 9.31, region *a* corresponds to the part of the main lobe outside the coverage, the typical gain variation being expressed as:

$$G(\theta) = G_{max} - 3(\theta/\theta_0)^2 \quad \text{(dBi)}$$

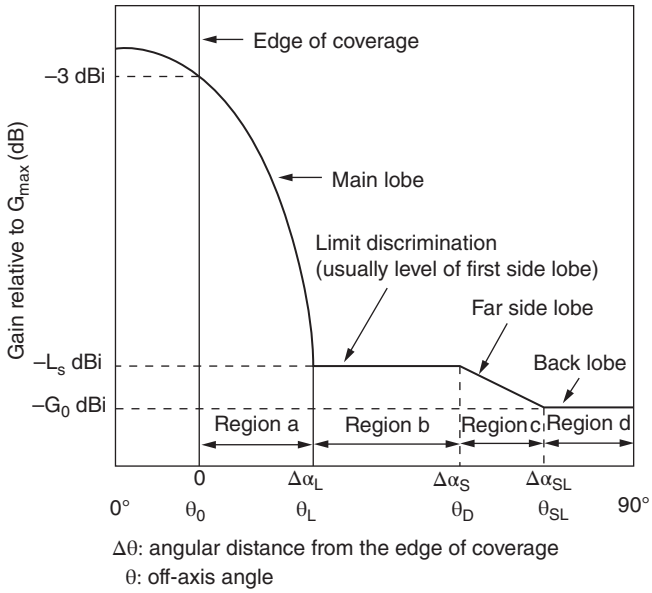


Figure 9.31 Reference limits for the antennas of a fixed service satellite. Source: reproduced from ITU-R Rec. S.672 with the permission of the ITU.

Region *b* is such that the discrimination is large enough to allow satellites operating at the same orbital location to provide coverage. The limit discrimination $-L_s$ could be about −20 to −30 dBi. Region *c* incorporates the far side lobes. Within region *d* (back lobe), gain is equal to $-G_0 = 0$ dBi.

Frequency reuse is also achieved by using orthogonal polarisation. A high value of polarisation isolation is thus necessary to limit interference.

In summary, the important characteristics of the antenna subsystem are:

— Conformity of the beam to the region to be covered
— An antenna radiation pattern with reduced side lobes
— High isolation between orthogonal polarisations
— Accurate beam pointing

Coverage and minimum beamwidth are closely associated with the satellite attitude and orbit-control procedure. Narrow beams and strict pointing specifications can require the use of an active antenna pointing system.

9.8.2 The RF coverage

Having determined the *geometrical contour* from the set of reference points for which a RF performance objective must be satisfied when depointing is taken into account, it is necessary to define the antenna beam that enables this objective to be achieved.

This depends on the nature of the specified objective: usually this means achieving at least a given EIRP for *transmit coverage* and G/T for *receive coverage* specified. The beam that maximises the gain at the specified points at the edge of coverage is then sought. In this case it should be noted that, even if the antenna gain is the same for the points specified at the edge of coverage, the

power received by the stations located at these points differs from one to the other. The distance to the satellite and the elevation angle vary with the station considered; these lead to variations of free space loss and atmospheric attenuation, respectively.

Hence, if optimisation of *power flux* within a given geometrical contour is the specified objective, it is necessary to weight each coverage reference point with a coefficient that represents the relative attenuation variations. Definition of the beam is more complex, and its angular width can differ significantly from that obtained from the geometrical contour.

Various types of antenna beam are used to illuminate the earth region within the geometrical contour:

— A beam of circular cross-section
— A beam of elliptical cross-section
— A shaped beam
— Multiple beams

The antenna beam does not always perfectly encompass the geometrical contour. The beam is characterised in different planes by its N dB beamwidth, which is defined by the solid angle at the edge of which the gain has fallen by N dB with respect to the maximum gain (at boresight). Its representation on the map gives the RF coverage or *footprint* of the beam (curves of equal gain).

The form of footprint obtained depends on the chosen representation. Hence, a beam of circular cross-section appears as an ellipse when represented on a plane if the axis of the beam is not perpendicular to the plane. In particular, with representation on a plane tangential to the earth at the sub-satellite point, an antenna beam of circular cross-section is represented by a circle if the boresight coincides with the direction of the earth centre. For a geostationary satellite, the angle between the boresight and the direction normal to the plane is at most 8.7° and the distortion is small, on the order of 1%.

On the other hand, the representation defined from the true view angles faithfully reproduces the form of the antenna beam independently of the boresight direction and the altitude of the satellite.

9.8.3 Circular beams

The cross-section of the beam is circular. It is the same as the radiating aperture of the antenna, which is usually reflecting.

9.8.3.1 RF coverage at 3 dB

The angular 3 dB beamwidth is taken as the angle $u_{3\,\mathrm{dB}}$ subtending the geometrical contour. The on-axis antenna gain in this case is equal to:

$$G_{\max} = 48\,360\eta/\theta_{3\mathrm{dB}}^2 \tag{9.47}$$

and the gain on the contour (edge of coverage [EOC]) is thus:

$$G_{\mathrm{eoc}} = G(\theta_{3\mathrm{dB}}/2) = G_{\max}/2 = 24\,180\eta/\theta_{3\mathrm{dB}}^2$$

where η is the efficiency of the antenna and $\theta_{3\,\mathrm{dB}}$ is expressed in degrees.

In the earlier expressions, the coefficient k that relates the 3 dB beamwidth to the ratio $\lambda = D(\theta_{3\,\mathrm{dB}} = k\lambda/D)$ is taken to be 70 (for a reflector antenna, k varies between 57 for uniform

illumination and 80 when the main lobe of the source is entirely intercepted by the reflector). Considering the maximum achievable value for h to be 0.75, the maximum gain at the edge of a beam of angular width $\theta_{3\,\mathrm{dB}}$ is given by:

$$G_{\mathrm{eoc}}(\mathrm{dBi}) = G(\theta_{3\mathrm{dB}}/2)_{\mathrm{dBi}} = 42.5 - 20\log\theta_{3\mathrm{dB}}\ (\mathrm{dBi}) \tag{9.48}$$

Variations of antenna gain and hence EIRP and G/T in the RF coverage are thus limited to 3 dB.

9.8.3.2 RF coverage providing maximum gain on the geometrical contour

The curves in Figure 9.32 show how the gain along the boresight and the gain in a given direction u_0 with respect to the boresight vary as a function of the ratio D/D_0, where D is the diameter of the antenna and D_0 is the particular value corresponding to a beam of 3 dB angular width such that $\theta_{3\,\mathrm{dB}} = 2\theta_0$, $D_0 = k\lambda/\theta_{3\,\mathrm{dB}} = k\lambda/2\theta_0$ [HAT-69]. It appears that for D between D_0 and 1.3 D_0, the gain in the u_0 direction remains greater than its initial value and passes through a maximum. The on-axis gain itself increases by more than 2 dB.

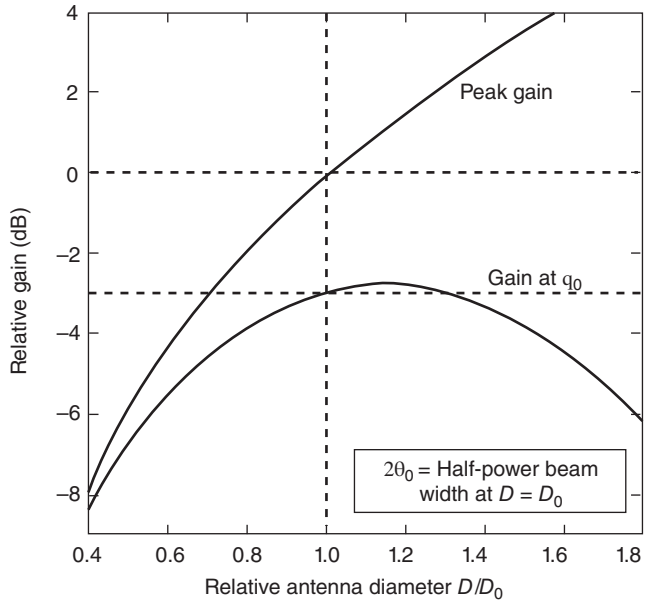


Figure 9.32 Variation of on-axis gain and gain in the θ_0 direction as a function of D/D_0 ($2\theta_0 = \theta_{3\,\mathrm{dB}}$ for $D = D_0$).

An antenna exists, therefore, where the gain in the θ_0 direction with respect to the antenna axis is maximised. For this antenna, the gain in the direction θ_0 is N dB lower than the on-axis gain. Taking $\theta_0 = \theta_{N\,\mathrm{dB}}/2$, where $\theta_{3\,\mathrm{dB}}$ is the N dB antenna beamwidth, the value of N can be determined as a function of the $\theta_{3\,\mathrm{dB}}$ beamwidth.

Assuming a parabolic gain variation about the axis (Eq. (5.5)):

$$\Delta G = 12(\theta/\theta_{3\mathrm{dB}})^2$$

hence for $\theta = \theta_0 = (\theta_{N\text{dB}}/2)$ and $\Delta G = N(\text{dB})$:

$$N = 12[(\theta_{N\text{dB}}/2)/\theta_{3\text{dB}}]^2 = 3(\theta_{N\text{dB}}/\theta_{3\text{dB}})^2$$

Hence:

$$\theta_{N\text{dB}} = \theta_{3\text{dB}}\sqrt{(N/3)} = k(\lambda/D)\sqrt{(N/3)} \tag{9.49}$$

The gain on the axis of the beam as a function of N and θ_N dB is given by:

$$G_{\max} = \eta(\pi D/\lambda)^2 = \eta[(k\pi)^2/3][N/(\theta_{N\text{dB}})^2] \tag{9.50}$$

The gain $G(\theta_{N\,\text{dB}}/2)$ at the edge of a beam of angular width $2\theta_0 = \theta_{N\,\text{dB}}$ corresponding to an N dB gain fallout thus has a value:

$$G(\theta_{N\text{dB}}/2)_{\text{dB}} = (G_{\max})_{\text{dB}} - N = 10\log[\eta(k\pi)^2/3] + 10\ \log[N/(\theta_{N\text{dB}})^2] - N$$

Hence:

$$G(\theta_{N\,\text{dB}}/2)_{\text{dB}} = 10\log[\eta(k\pi)^2/3] + 10\ \log\ N - 20\log\theta_{N\,\text{dB}} - N$$

To obtain the maximum gain at the edge of a beam of fixed angular width $2\theta_0 = \theta_{N\,\text{dB}}$, it is necessary to maximise $10\log[\eta(k\pi)2/3]$ and $10\log N - N$. Maximisation of $10\log N - N$ leads to:

$$[10/(N\ \ln\ 10)] - 1 = 0,\ \ \text{hence}\ N = 10/\ln 10 = 4.34$$

The gain is thus maximised on the geometrical contour when the angle subtending the contour is $\theta_{4.3\,\text{dB}}$.

It remains to maximise $10\log[\eta(k\pi)^2/3]$ in order to obtain the highest possible gain. When a reflector antenna is used, the efficiency η of the antenna depends in particular on the efficiency of illumination η_{i} of the reflector; this in turn determines the factor k that defines the 3 dB angular width (k and η_{i} vary in opposite directions).

The value currently used for k is 70, and this corresponds to a source radiation pattern that illuminates the edge of the reflector with a relative level with respect to the centre of around −12 dB; the efficiency of illumination η_{i} is thus on the order of 0.75. On the other hand, the factor k is maximum and approximately equal to 80 for a source radiation pattern whose main lobe is entirely intercepted by the reflector. The efficiency of illumination η_{i} is no greater than 0.6 under these conditions. Nevertheless, the latter scheme maximises the product $\eta_{\text{i}}k^2$, to a value on the order of 3800, and consequently maximises the quantity $10\log[\eta(k\pi)^2/3]$.

Assuming that the antenna efficiency η coincides with the efficiency of illumination η_{i} (no ohmic losses), the maximum gain at the edge of a beam of angular width $2\theta_0$ corresponding to an N dB gain fallout with $N = 4 : 3$ dB thus has a value:

$$G\left(\frac{\theta_{N\,dB}}{2}\right)_{\text{dB}} = 10\log\left(\frac{3800\pi^2}{3}\right) + 10\log 4.3 - 4.3 - 20\log\theta_{N\text{dB}}$$

Hence:

$$G(\theta_{4.3\text{dB}}/2)_{\text{dB}} = 43 - 20\log\theta_{4.3\text{dB}}\ \ (\text{dB}) \tag{9.51}$$

The gain on the axis is 4.3 dB greater.

Under these conditions, the diameter of the reflector obtained from Eq. (9.49) is given by:

$$D = \lambda\left(\frac{k}{\theta_{N\,\text{dB}}}\right)\sqrt{(\text{N}/3)} = 95(\lambda/\theta_{4.3\text{dB}})$$

Recall that, in the case where the angular width is that corresponding to a 3 dB gain fallout and the illumination efficiency is maximum with a value of k on the order of 70, the diameter is given by $D = 70(\lambda/\theta_{3\,dB})$.

Maximisation of the gain at the edge of a beam of specified angular width $2\theta_0$, obtained by choosing a relative gain level at the edge of −4.3 dB with respect to the on-axis gain and an illumination efficiency of 0.6, leads to an increase of the reflector diameter of around 35%. This increase in diameter entails an increase in mass on the order of 50%. The gain at the edge is then around 0.5 dB greater than when the relative gain at the edge is −3 dB with respect to the on-axis gain. The on-axis gain itself is around 1.8 dB greater.

9.8.4 Elliptical beams

A narrow beam of elliptical cross-section provides greater flexibility for matching the geometrical contour. The beam is characterised by two angular widths $\theta_{AN\,\mathrm{dB}}$ and $\theta_{BN\,\mathrm{dB}}$ that correspond to the major axis A and the minor axis B of the ellipse, and the orientation of the ellipse with respect to the reference frame (Figure 9.33).

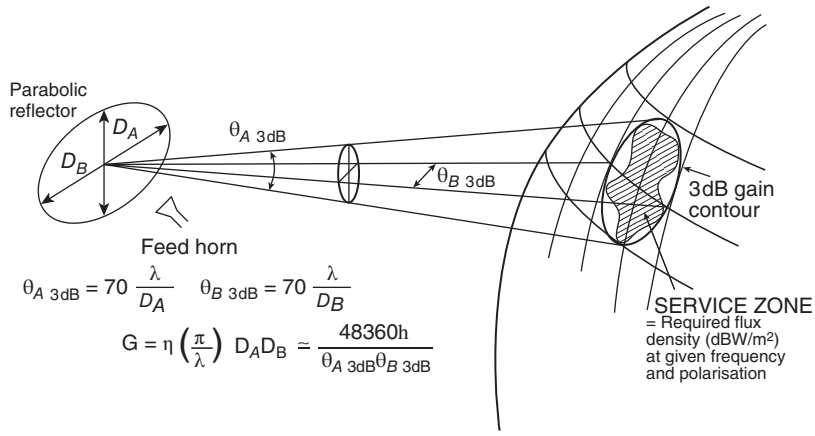


Figure 9.33 Characterisation of an elliptical antenna beam.

9.8.4.1 RF coverage at 3 dB

The on-axis gain is:

$$G_{\max} = 48\,360\eta/(\theta_{A3\,\mathrm{dB}}\theta_{B3\,\mathrm{dB}}) \tag{9.52a}$$

and the gain on the contour (edge of coverage) is:

$$G_{\mathrm{eoc}} = G_{\max}/2 = 24\,180\eta/\theta_{A3\mathrm{dB}}\theta_{B3\,\mathrm{dB}}$$

where η is the efficiency of the antenna and $\theta_{A3\,\mathrm{dB}}$ and $\theta_{B3\,\mathrm{dB}}$ are expressed in degrees. The angles $\theta_{A3\,\mathrm{dB}}$ and $\theta_{B3\,\mathrm{dB}}$ are related to the corresponding diameters of the radiating aperture by $\theta = 70(\lambda/D)$ (Figure 9.33).

9.8.4.2 RF coverage at 4.3 dB

It is possible, as for the beam with a circular cross-section, to define a RF coverage that maximises the gain on the geometrical contour. The on-axis gain is:

$$G_{\max} = 69\ \ 320\eta/(\theta_{A4.3\text{dB}}\theta_{B4.3\text{dB}}) \tag{9.52b}$$

and the gain on the contour is:

$$G_{\text{eoc}} = G(\theta_{4.3\text{dB}}/2) = 25754\eta/\theta_{A4.3\text{dB}}\theta_{B4.3\text{dB}}$$

where η is the antenna efficiency and $\theta_{A4.3\,\text{dB}}$ and $\theta_{B4.3\,\text{dB}}$ are expressed in degrees.

9.8.4.3 Optimisation

In the general case, the reference points of the service zone contour define a region located away from the longitude of the satellite. It is then necessary to determine the parameters of an ellipse that corresponds to a 3 dB coverage (or another value) and maximises the gain at the reference points at the EOC. The parameters of the ellipse that characterise the footprint of the beam are the major axis A, the minor axis B, the inclination (tilt) of the major axis T, and the position (X_0,Y_0) of the centre of the ellipse with respect to the sub-satellite point in a true view representation (Figure 9.34). An optimisation procedure can involve choosing four extreme points on the coverage and finding the ellipses of minimum angular width that pass through these points and cover all the other points. A constraint on the minimum value of the ellipticity A/B of the beam can be introduced to take account of the feasibility of the antenna.

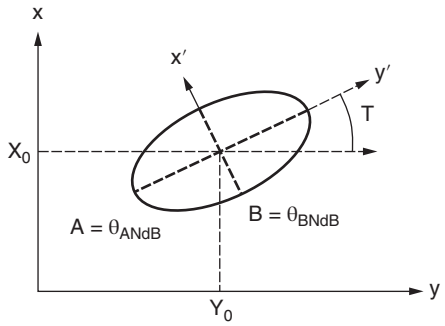


Figure 9.34 The defining parameters of elliptic coverage.

9.8.5 The influence of depointing

The loss of gain associated with depointing depends on how the performance objectives are specified.

9.8.5.1 Performance specified in terms of minimum EIRP over a region

The geometrical contour takes account of the pointing error of the antenna beam; a circle of radius equal to the pointing error is centred on each reference point of the service zone contour using true view angle representation.

Taking pointing errors into account for circular or elliptical beams leads to a broadening by twice the pointing error (Figure 9.35). Broadening the antenna beam that just covers the zone to be served by twice the depointing is equivalent to defining the geometrical contour by assigning a circle of uncertainty with a radius equal to the depointing to all reference points.

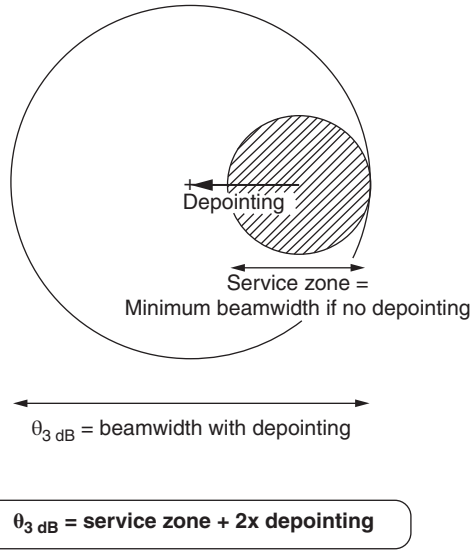


Figure 9.35 Extension of a beam on account of depointing.

Considering an elliptical beam and assuming that a 3 dB coverage that takes account of the depointing Du is used, the antenna gain on the coverage axis is (Eq. (9.52a)):

$$G_{\max} = 48\ \ 360\eta/(\theta_{A3\mathrm{dB}}\theta_{B3\mathrm{dB}})$$

If the antenna beam just covered the region to be served (not taking depointing into account), the on-axis gain would be:

$$G'_{\max} = 48\ \ 360\eta/[(\theta_{A3\ \ \mathrm{dB}} - 2\Delta\theta)(\theta_{B3\mathrm{dB}} - 2\Delta\theta)]$$

The loss of gain ΔG due to broadening of the beam to take account of depointing is thus:

$$\begin{aligned}\Delta G = G_{\max}/G'_{\max} &= [(\theta_{A3\mathrm{dB}} - 2\Delta\theta)(\theta_{B3\mathrm{dB}} - 2\Delta\theta)]/(\theta_{A3\mathrm{dB}}\theta_{B3\mathrm{dB}}) \\ &= (1 - 2\Delta\theta/\theta_{A3\mathrm{dB}})(1 - 2\Delta\theta/\theta_{B3\mathrm{dB}})\end{aligned}$$

Hence, in decibels:

$$\Delta G(\mathrm{dB}) = 10\log(1 - 2\Delta\theta/\theta_{A3\mathrm{dB}}) + 10\log(1 - 2\Delta\theta/\theta_{B3\mathrm{dB}})$$

The earlier loss of gain results from a broadening by $2\Delta\theta$ of the 3 dB beamwidth to take account of depointing. To keep the required value of EIRP at the edge of coverage, the transmitter power must be increased by $-\Delta G$(dB).

If $\Delta\theta$ is small compared with $\theta_{A3\,\mathrm{dB}}$ and $\theta_{B3\,\mathrm{dB}}$, this gives:

$$\begin{aligned}\Delta G(\mathrm{dB}) &= -(10/2.3)[(2\Delta\theta/\theta_{A3\ \mathrm{dB}}) + (2\Delta\theta/\theta_{B3\ \mathrm{dB}})] \\ &= -8.7[(\theta_{A3\ \mathrm{dB}} + \theta_{B3\ \mathrm{dB}})/(\theta_{A3\ \mathrm{dB}}\theta_{B3\ \mathrm{dB}})]\Delta\theta \end{aligned} \tag{9.53}$$

With a circular beam (and coverage to 3 dB), $\theta_{A3\,\mathrm{dB}} = \theta_{B3\,\mathrm{dB}} = \theta_{3\,\mathrm{dB}}$. This gives:

$$\Delta G(\mathrm{dB}) = -1.74(\Delta\theta/\theta_{3\ \mathrm{dB}}) \quad (\mathrm{dB}) \tag{9.54}$$

It should be noted that this expression differs from the expression $\Delta G(\mathrm{dB}) = -12(\Delta\theta/\theta_{3\,\mathrm{dB}})^2$, which characterises the loss of gain due to depointing in the vicinity of the boresight of the antenna. The problem posed here is different; it is to evaluate the loss of gain due to broadening of the beam by $2\Delta\theta$ to ensure coverage of a service zone with a minimum EIRP. Taking $\Delta\theta$ equal to $(\theta_{3\,\mathrm{dB}}/10)$, the loss is −1.7 dB.

9.8.5.2 Performance specified in terms of required EIRP in a given direction

In this case, the service must provide a required EIRP at a specified point (a single receiving station) situated nominally on the axis of the antenna beam. Depointing $\Delta\theta$ of the satellite antenna causes a loss of EIRP as a function of the depointing components $\Delta\theta_A$ and $\Delta\theta_B$ with respect to the axes y' and x', which define the antenna coverage (see Figure 9.34):

$$G(\Delta\theta) = G_{\max} - 12[(\Delta\theta_A/\theta_{A3\mathrm{dB}})^2 + (\Delta\theta_B/\theta_{B3\ \mathrm{dB}})^2] \quad (\mathrm{dBi}) \tag{9.55}$$

Hence there is a loss of gain for depointing with components $\Delta\theta_A$ and $\Delta\theta_B$, which has a value

$$\Delta G(\mathrm{dB}) = -12[(\Delta\theta_A/\theta_{A3\mathrm{dB}})^2 + (\Delta\theta_B/\theta_{B3\ \mathrm{dB}})^2] \quad (\mathrm{dB})$$

With a circular beam, $\theta_{A3\,\mathrm{dB}} = \theta_{B3\,\mathrm{dB}} = \theta_{3\,\mathrm{dB}}$ and $\Delta\theta = \sqrt{\Delta\theta_A^2 + \Delta\theta_B^2}$. This gives loss of gain:

$$\Delta G(\mathrm{dB}) = -12(\Delta\theta/\theta_{3\ \mathrm{dB}})^2 \quad (\mathrm{dB}) \tag{9.56}$$

This loss of gain results from variation of the satellite antenna pointing direction due to the effect of the motion of the satellite. To provide the specified EIRP in the direction of the earth station situated nominally on the antenna axis, the transmitter power must be increased by $-\Delta G$(dB).

Example 9.2 Consider an antenna of 3 dB width equal to 1°. Depointing is estimated at 0.3°. In the first case (Eq. (9.54)), the loss of gain is −5.2 dB. In the second case, (Eq. (9.56)), the loss of gain is −1.1 dB.

The example shows that the power margin to be provided is greater, for the same resulting depointing, in the case of a service that provides minimum EIRP over a region (such as direct television broadcasting or very small aperture terminal [VSAT]) than in the case of a service to a given earth station. If this margin is deemed to be too large in view of the value of depointing, it would be advisable to provide an antenna beam pointing control system that limits the variation of antenna coverage.

9.8.6 Shaped beams

The elliptical beam is a first step towards matching the antenna radiation pattern to the service zone. However, most often the service zone is not elliptical, and perfect matching is not achieved. This both leads to interference outside the specified service zone and a loss of gain with respect to the maximum which is theoretically possible over coverage of a given angular area.

9.8.6.1 Gain limit over a given service zone

The theoretical limit of gain G_{lim} over a service zone of complex shape is obtained by considering an ideal lossless antenna whose beam conforms exactly to the solid angle Ω (in steradian) that subtends the service zone (the gain is zero outside the service zone). This gain is, by definition, equal to:

$$G_{\text{lim}} = 4\pi/\Omega \tag{9.57}$$

The solid angle can be approximated by the angular area (in radian^2) of the service zone. By considering a beam of circular cross-section and angular width 2θ, the corresponding solid angle Ω is equal to $2\pi\,[1-\cos\theta]$, and the angular area S has a value $\pi\theta^2$. In the case of global coverage of the earth by a geostationary satellite ($2\theta = 17.4°$), the approximation leads to an error that is less than 2×10^{-3}.

9.8.6.2 Determination of the angular area

The angular area S of the service zone is defined by a set of n points that constitute a polygon and can be calculated from the true view coordinates X and Y of the points defined by Eq. (9.24). The angular area S of the polygon is given by $S = \Sigma_n S_{\text{i}}$, where the area S_{i} is the algebraic area defined by the line joining point P_{i} to point $P_{\text{i+1}}$ (Figure 9.36a). This gives:

$$S = \sum_n \left\{ \frac{1}{2}[(X)_i + (X)_{i+1}][(Y)_i - (Y)_{i+1}] \right\} \quad (\text{degree}^2) \tag{9.58}$$

The pointing error is represented by a disc of uncertainty of radius equal to the depointing $\Delta\theta$ centred on each vertex of the polygon (Figure 9.36b). Since the sum of the exterior angles of a polygon is equal to 2π, the increase of area ΔS to take the pointing error into account is given by:

$$\Delta S = P\Delta\theta + \pi\Delta\theta^2 = \Delta\theta(P + \pi\Delta\theta) \quad (\text{degree}^2) \tag{9.59}$$

where P is the perimeter of the polygon that represents the service zone and is determined by:

$$P = \sum_n \{[X_{i+1} - X_i]^2 + [Y_{i+1} - Y_i]^2\}^{1/2} \quad (\text{degree}^2) \tag{9.60}$$

9.8.6.3 Beam-shaping techniques

Beam shaping can be obtained using two different methods whose principles are given here. Implementation is discussed in Section 9.8.9.

The first method consists of modifying the power distribution within a beam generated by a single source. Shaping of the beam is achieved by modifying the contour of the aperture or the profile of the reflector (see Section 9.8.9.2). Whatever technique is used, the shape of the beam can only be modified by mechanical changes, which cannot be carried out straightforwardly while the satellite is in orbit.

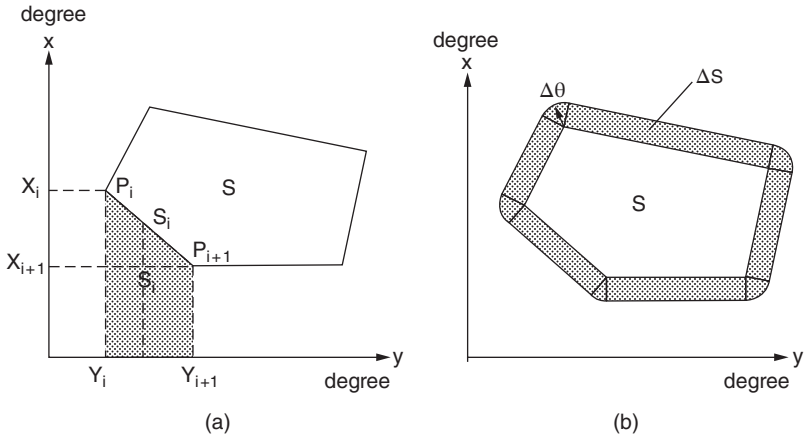


Figure 9.36 Angular area S of the service zone: (a) no pointing error; (b) ΔS increase due to depointing $\Delta\theta$.

With the second technique, shaping of the beam is obtained by combining the radiation of several elementary beams. These beams are generated by several radiating elements that are excited by coherent signals having a given amplitude and phase distribution imposed by a BFN (Figure 9.37). The array of radiating elements can be located at the focus of an antenna reflector or lens. It can also generate the antenna beam directly (a direct radiating array antenna; see Section 9.8.9.4).

The latter technique enables a beam of any shape to be obtained, and its gain distribution over the RF coverage can be adapted on demand. Since the resultant beam is obtained by combining elementary beams of smaller angular width, it is possible to obtain a gain close to the gain limit G_{lim} over a large part of the service zone; a decrease occurs only at the edge of the service zone. Furthermore, the decrease outside the service zone is rapid. So, even in the case of illumination of a service zone of simple geometric shape (e.g. circular), the compound beam has a definite advantage (in terms of gain over the service zone and reduction of interference outside it) over illumination by a single beam.

Another major advantage lies in the possibility of modifying the RF coverage of the antenna by controlling the amplitude and phase distribution of the radiating elements. This possibility can be effective even when the satellite is in orbit by having the BFN with elements that can be controlled by telecommand.

The disadvantages of this technique are the added complexity and mass of the antenna and the increase in the size of the radiating aperture (due to generation of individual beams that are narrower than the angular width of the service zone).

9.8.7 Multiple beams

Unlike the preceding coverages, which use a single beam in a given frequency band with a given polarisation, multiple-beam coverage implies generation of several beams that may be in different frequency bands and have different polarisations.

9.8.7.1 Multiple separate beams

The service zone consists of a set of geographical regions that are separated from each other. These regions are of simple geometric form in true view angle representation and are illuminated

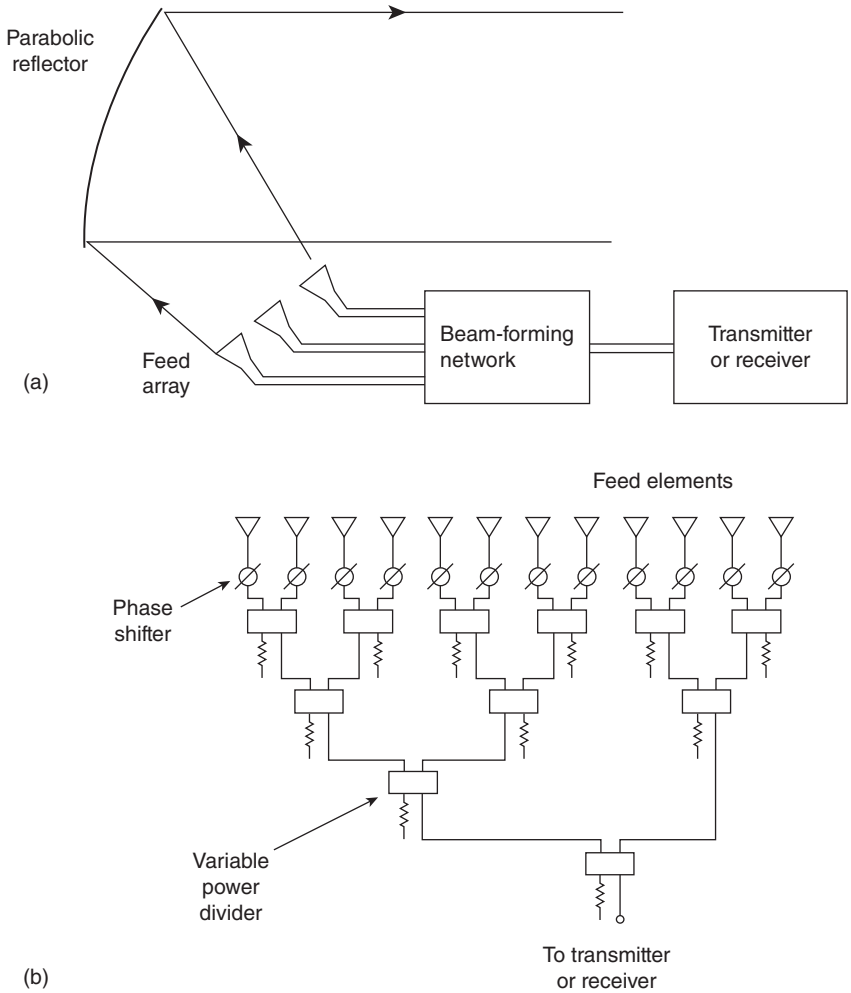


Figure 9.37 Shaped beam antenna using an array of radiating elements.

by beams of narrow circular cross-section. The regions could correspond to large towns between which it is required to establish high-capacity links. The beams can thus share the same frequency bands when their angular separation is sufficient. The use of orthogonal polarisation enables the isolation between links to be increased, should the angular separation be too small. Figure 9.38 illustrates this concept in the context of coverage of regions of Europe with large communications requirements.

9.8.7.2 Contiguous beams

Service within a given geometrical contour can be provided by a set of narrow contiguous beams rather than a single beam (Figure 9.39). Since each of the beams is narrower than a beam that

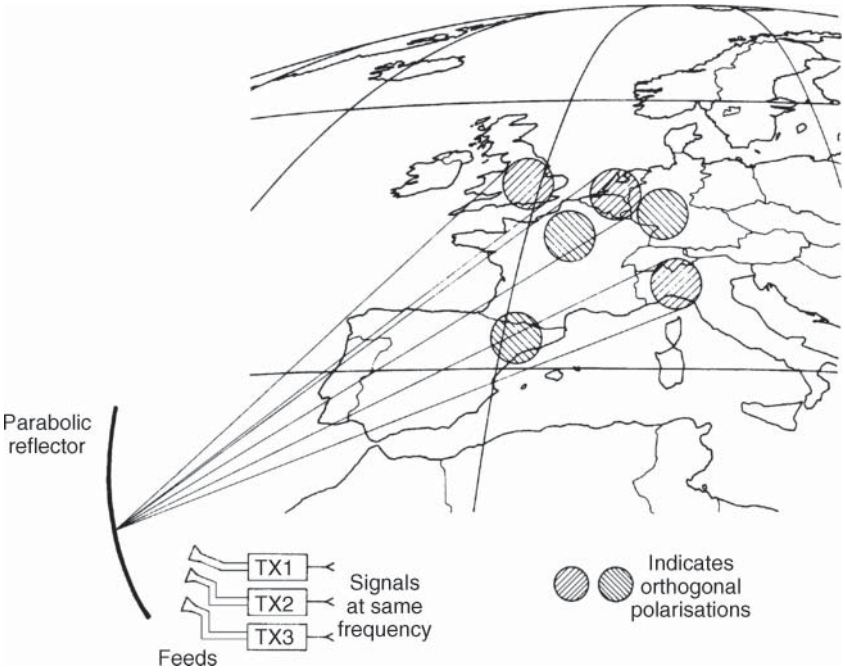


Figure 9.38 Separate multiple beams.

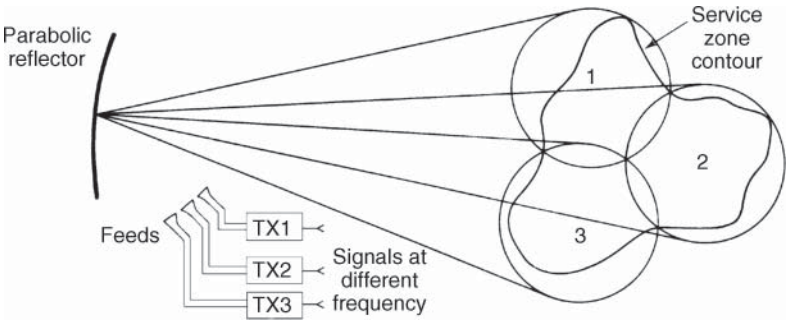


Figure 9.39 Contiguous multiple beams.

would cover the whole of the region to be served, the corresponding gain is higher. It is thus conceivable to use earth stations with small-diameter antennas.

Since the beams partially overlap, the frequencies used must differ from one beam to the other. The beams thus share the total available system bandwidth as conditioned by the Radio Regulations. The capacity per beam is thus limited more severely when the number of beams is large. Another disadvantage of this concept is associated with the fact that the information transmitted differs from one beam to the other. To ensure interconnectivity, routing of the carriers between beams must be considered. This is specific to multibeam satellite networks and is discussed in Section 7.4.

9.8.7.3 Beam lattice

Combining multiple beams and frequency reuse produces a lattice coverage, where a basic pattern formed from a cluster of beams using a set of different frequencies is regularly repeated over the service zone (Figure 9.40).

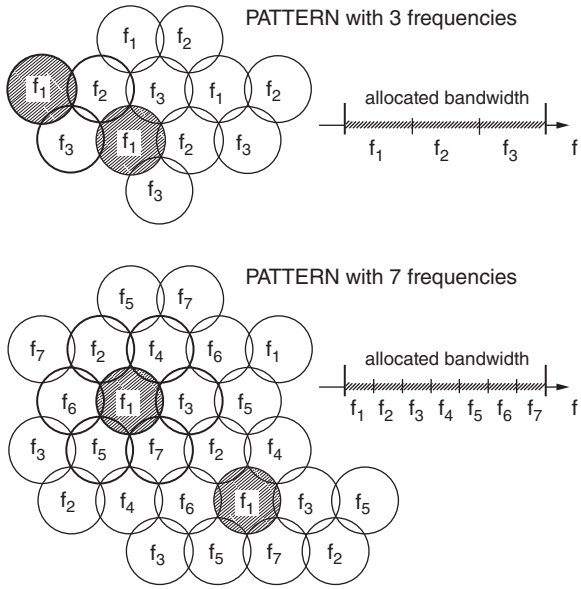


Figure 9.40 Lattice coverage with (a) a three-frequency pattern; (b) a seven-frequency pattern.

Figure 9.40 shows the variation of angular distance between beams that reuse the same frequency band as a function of the number of frequencies used. The angular distance determines the amount of co-channel interference. Figure 9.41 shows how interference occurs on a particular beam in a three-frequency lattice. The greatest interference occurs at the edge of the beam; here the level of the interfering signal is highest and the level of the wanted signal is lowest due to the decrease of gain at the edge. The contributions of the six beams of the adjacent patterns that share the same frequency must be considered.

By using a larger number of beams (and frequencies) in the basic pattern, the angular distance between beams using the same frequency is increased; this leads to a decrease of interference in the system. On the other hand, the usable bandwidth, and hence the capacity per beam, consequently decreases. An example of European coverage by a three-frequency beam lattice is given in Figure 9.42.

9.8.8 Types of antenna

The types of antenna used differ in accordance with the principle used to control the satellite attitude. A simple method of providing attitude stabilisation consists of causing the satellite to rotate about an axis perpendicular to the plane of the orbit (spin stabilisation; see Section 10.2.6).

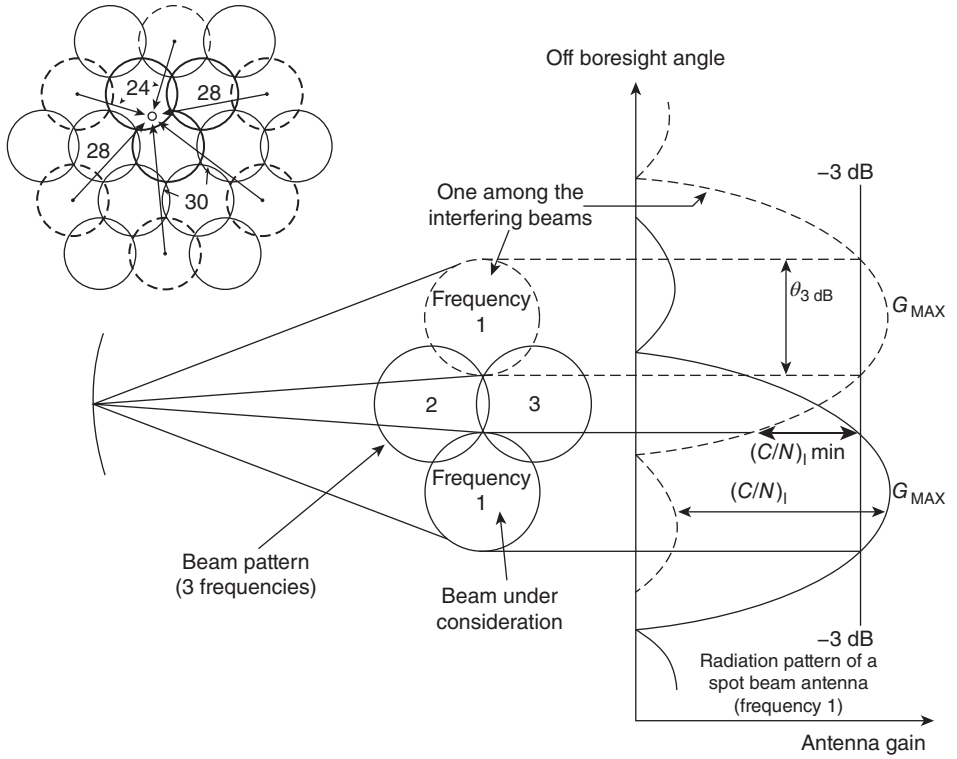


Figure 9.41 Interference within a lattice.

The antennas can be mounted directly on the rotating satellite or on a platform that maintains a constant orientation with respect to the earth. The attitude of the satellite as a whole can also be controlled in order to maintain a fixed orientation with respect to the earth (three-axis stabilisation; see Section 10.2.7).

In the case where the antennas are mounted on a platform that is rotating with respect to the earth, the antenna must have a toroidal radiation pattern or generate rotation of the pattern in such a way as to compensate for that of the platform.

Nowadays, communications satellites consist of a platform that supports the payload and whose attitude is stabilised with respect to the earth. The radiating aperture of the antenna thus maintains a fixed orientation with respect to the direction of the earth centre.

9.8.8.1 Antennas with a toroidal radiation pattern

For a spin-stabilised satellite, the simplest antenna generates a radiation pattern of revolution about the axis of rotation. To ensure global coverage, the beamwidth of the toroidal pattern is on the order of 17°. The antenna gain is only a few decibels.

A toroidal pattern can be obtained by means of a set of linear wires (wire antennas). This procedure was used on the first operational satellites; for Intelsat I and II, for example, the antenna gain is from 4 to 5 dB receiving and around 9 dB transmitting.

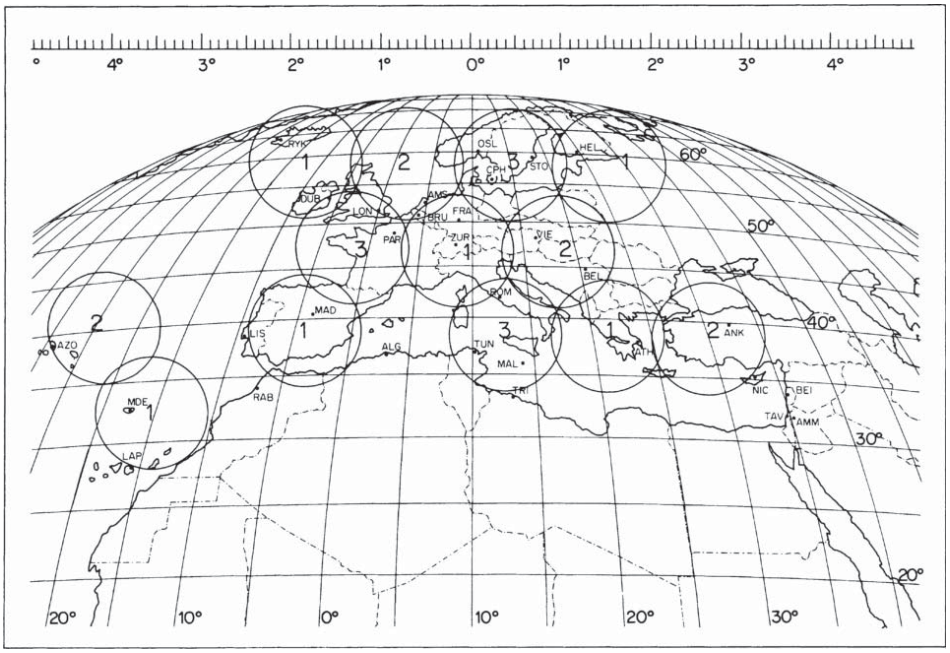


Figure 9.42 European coverage with a three-frequency lattice.

9.8.8.2 Despun antennas

To increase the antenna gain, it is necessary to concentrate the beam on the region to be covered and then to ensure that its orientation remains fixed with respect to the earth. The antenna beam thus turns in the opposite direction to the rotation of the satellite; the antenna is said to be *despun*.

Mechanical despun antennas. This approach consists of rotating the antenna assembly about the axis of rotation of the satellite, by means of an electric motor, in such a way as to keep the antenna axis pointing towards the earth [DON-69]. The presence of bearings, for which lubrication is difficult, and rotating couplings between the antenna and the radio equipment limit reliability and degrade performance.

Electronic despun antennas. Electronic scanning provides an elegant solution to overcome the problems associated with mechanical despun antennas. The antenna consists of a set of radiating elements mounted on a cylinder. These radiating elements are fed sequentially with a phase that varies as a function of the rotation of the satellite. The drawback of this type of antenna, other than the losses in the feeders to the radiating elements, lies in the amplitude and phase discontinuities that appear in the antenna radiation pattern during successive switchings. This type of antenna is used on the Meteosat satellite, for example.

9.8.8.3 The stabilised platform

The satellite consists of a platform on which the antennas and the repeaters are mounted. This platform maintains a fixed orientation with respect to the earth.

In the case of a spin-stabilised satellite, this platform consists of the upper part of the satellite that is driven in contra-rotation with respect to the lower part, which itself rotates about an axis perpendicular to the plane of the orbit (a *dual-spin satellite*). This approach permits installation of high-performance antennas and avoids the problems of rotating couplings between the antennas and the radio equipment. However, the problems associated with the presence of a mechanical bearing (such as lubrication and mechanical friction that disturbs the gyroscopic effect) and slipping contacts to transfer electrical energy remain. The Intelsat VI satellite is an example of this type of architecture.

The three-axis stabilised satellite itself forms the platform on which the antennas are mounted. Greater freedom is thus provided for mounting large antennas.

Whatever the type of attitude control of the stabilised platform, if the antenna mounting is rigid, the pointing accuracy of the antennas is that of the attitude stabilisation (down to 0.05°). Greater pointing accuracy requires the use of systems that control antenna pointing using a beacon on the ground.

9.8.9 Antenna technologies

The frequency bands used by communications satellites are such that the wavelength is small compared with the mechanical size of the antenna. The antennas used are of the radiating aperture type – horn, reflecting, lens, and array antennas.

9.8.9.1 Horn antennas

The horn antenna is one of the simplest types of directional antenna. It is well suited to, and widely used for, global coverage of the earth. A 3 dB beamwidth of 17.5° is obtained at 4 GHz from a horn whose aperture diameter is 30 cm.

A beam of smaller width would require a horn with a larger aperture and proportionally greater length, thereby making installation on the satellite difficult. Furthermore, the horn antenna has poor side-lobe characteristics. These characteristics are improved by corrugation (annular discontinuities) of the interior of the horn. The length of the horn can be reduced by using an excitation system employing a microstrip antenna.

Horns are, however, currently used as a primary source in reflecting antennas.

9.8.9.2 Reflector antennas

This type of antenna is the most commonly used to obtain spot beams or shaped beams. The antenna consists of a parabolic reflector illuminated by one or more radiating elements located at the focal point.

The technique of reflector manufacturing usually consists of bonding two carbon fibre layers impregnated with resin on each side of a core of aluminium honeycomb. This technique allows excellent results to be obtained in terms of profile realisation accuracy, dimensional stability, and rigidity in spite of the mechanical and thermal constraints. Reflection losses are low, less than 0.1 dB in Ku band.

It is possible to modify the pointing direction of the beam in orbit by telecommand by providing the antenna with a control device for the mechanical orientation of the reflector. With a multisource antenna, pointing can also be achieved by modifying the phase distribution of the radiating element feed.

Two-reflector mounting. A two-reflector mounting in which the main reflector is illuminated by an auxiliary reflector that is itself illuminated by the radiating element or elements (a Cassegrain or Gregorian mounting, according to whether the auxiliary reflector is hyperbolic or parabolic) can also be used. A two-reflector mounting, on account of the compactness of the antenna obtained, has an advantage with respect to mechanical mounting of the antenna on the satellite. In certain cases, it also facilitates antenna design (e.g. for shaped beams).

Offset mounting. Symmetrical mountings suffer from blocking of the aperture by the radiating elements or the auxiliary reflector and their supports; this leads to a degradation of efficiency and an increase in the level of side lobes. Use of a portion of the reflector that is offset with respect

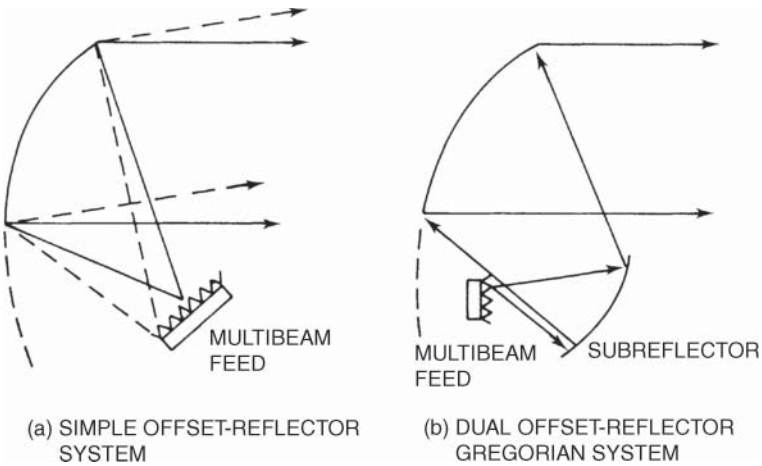


Figure 9.43 Reflector antennas with offset mounting: (a) simple reflector; (b) dual reflector (Gregorian).

to the principal axis of the parabola avoids blocking the aperture of the paraboloid (an offset mount). Offset illumination can be used with a one- or two-reflector mounting (Figure 9.43). Offset mounting also permits easier integration of the antenna onto the satellite, particularly with large reflectors that require folding of the antenna for the launch phase.

When the offset reflector is illuminated by a linearly polarised wave, an orthogonally polarised component originates in the reflector due to the asymmetry with respect to the antenna axis. Offset mounting is thus characterised by a low value (on the order of 20–25 dB) of antenna polarisation isolation. When circular polarisation is used, beam squinting is also observed.

Reflector contour shaping. A circular antenna reflector generates a beam of circular cross-section. Simple beam shaping is achieved by modifying the shape of the reflector contour. Hence an elliptical reflector generates a beam of elliptical cross-section (see Figure 9.33). In practice, applications are limited to that case. Indeed, a reflector of excessively complicated form would lead to difficulty in correctly matching the illumination pattern of the primary source, since this would involve low illumination efficiency and a high level of side lobes.

Multifeed antennas. By locating an array of radiating elements at the focus of the antenna, it is possible to obtain either a shaped beam or multiple beams. If the array of radiating elements is fed from the same signal with a particular amplitude and phase distribution, a shaped beam is obtained. This distribution is obtained by means of a set of phase shifters, couplers, and power splitters – the BFN.

Independently fed sources permit generation of separate beams that are characterised by their frequency and polarisation. A multibeam coverage such as that illustrated in Figure 9.38 can thus be obtained.

This technique is also used to generate lattice coverage. The size of the source array increases with the size of the lattice and sources located at the edge of the array are situated far from the focal point; this leads to degradation of the corresponding radiation pattern. When the number of beams becomes large, a reduction in the number of sources can be obtained by sharing the sources among several beams; a beam in a given direction is obtained by an appropriate amplitude and phase distribution from several sources.

Reflector surface shaping. The beam shape can be adjusted by shaping the reflector surface, departing from a strict parabolic profile. For instance, it is possible to use a reflector of circular form whose profile is parabolic in one plane and cylindrical in the other. The beam obtained in this way is no longer of circular cross-section but is approximately elliptical. The aperture remains circular, and this facilitates optimisation of the illumination.

Furthermore, the profile of the reflector can deviate from the parabolic shape, whatever the considered plane. A circular reflector whose profile deviates from parabolic at the edges enables the relative gain at the coverage boundary to be increased; in this way, gain variations within the coverage are limited.

Finally it is possible to synthesise a reflector profile that enables a beam to be obtained whose spatial power distribution corresponds to that required to ensure illumination of the service zone. These synthesis techniques are complex, but they lead to more effective antennas, as only one source is used instead of several. The drawback is that the defined beam pattern is given and cannot be changed during the satellite lifetime, even though some experiments (on the ground) of variable mechanical control of the shape of the reflector have been attempted. A significant reduction in mass of the antenna is then achieved, which makes this solution very attractive today and widely used.

Dual-grid antenna. To obtain an antenna radiation pattern having high polarisation isolation, one approach is to use a reflector consisting of a grid: that is, an array of conductors parallel to the required linear polarisation. When the grid is illuminated by a radio wave, only the

component of the electric field parallel to the grid is reflected. Current can flow only along the conductors, and the field component orthogonal to the grid cannot exist. The antenna displays a high cross-polarisation discrimination (typically 40 dB). Two separate antennas whose grids are perpendicular can be used to generate two beams with linear orthogonal polarisation.

In order to reduce bulk and mass, the dual-grid antenna concept has been developed and used on Eutelsat II, for example. The antenna consists of two reflectors with offset feed mounted one behind the other with a slight bias so that their foci are located at two different points (Figure 9.44). To illuminate each reflector, it is thus easy to locate the radiating elements operating with a given polarisation at the corresponding focus. The front reflector is formed from a material that is transparent to radio waves on which an array of conductors is arranged parallel to the electric field of the waves generated by the associated illuminating sources. These waves are thus reflected by the first reflector. The waves radiated by the other source array have orthogonal polarisation. Hence they pass through the front reflector and are reflected by the rear one before again passing through the front reflector. The rear reflector can be either a grid (whose orientation is orthogonal to the first) or a conventional reflector. Even if a component with polarisation orthogonal to the nominal polarisation is generated due to offset mounting, this orthogonal component, which is parallel to the grid of the front reflector, is blocked at the rear of the grid.

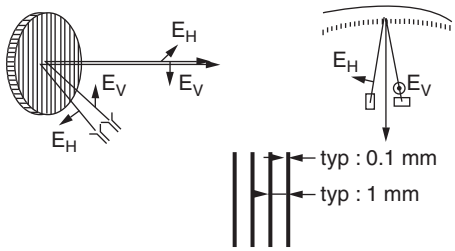


Figure 9.44 Gridded reflector antenna organisation.

The front reflector, and the structure between the two reflectors, must be transparent to radio waves but capable of withstanding mechanical and thermal stresses. For instance, Kevlar layers can be bonded on each side of a honeycomb core of the same material.

The array of conductors can be embedded in the composite material and produced either by chemical etching of a conducting deposit or by mechanically cutting a copper deposit formed on the Kevlar layer before bonding to the honeycomb. For a Ku-band antenna, typical dimensions are on the order of 0.1 mm for the conductor width with spacing on the order of a millimetre. The polarisation isolation obtained with this type of antenna exceeds 35 dB.

Dichroic reflectors. A dichroic surface is reflecting to radio waves within a given band of frequencies and transparent outside this band. To obtain such a surface, an array of dipoles whose dimensions are characteristic of the frequency to be reflected is arranged on a substrate that is transparent to electromagnetic waves.

By realising the auxiliary reflector of two-reflector antennas using this technique, the antenna has two focal points that depend on the frequency of operation. This permits the same reflector to be used in two different frequency bands; the difficulty associated with mounting a matched source for each frequency band at the focus is thereby resolved. Figure 9.45 shows an example where the dichroic reflector is reflecting in Ka band and transparent in Ku band. Furthermore, a surface that is polarisation selective can be used to generate two different foci according to the polarisation of the wave in Ku band.

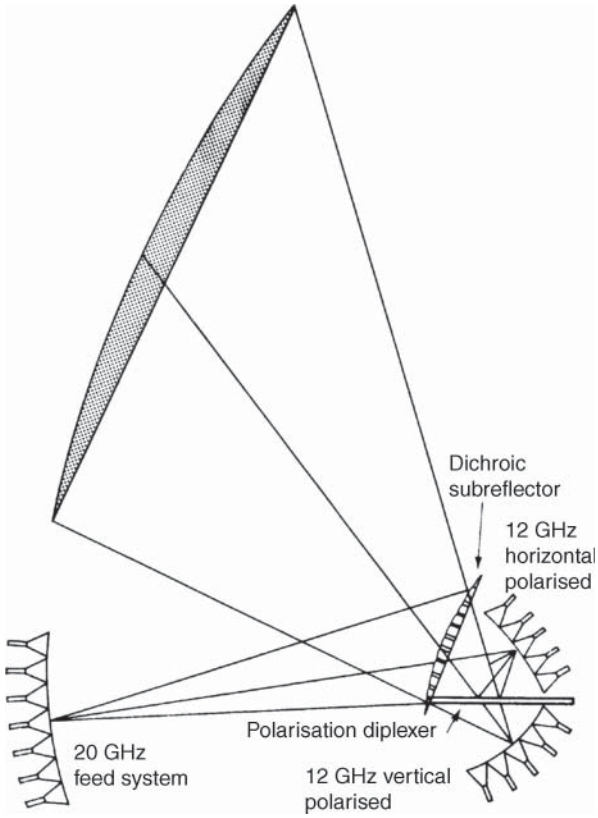


Figure 9.45 Dual frequency antenna with a dichroic surface.

9.8.9.3 Lens antennas

Antennas of this kind associate one or more radiating elements with a 'lens' that focuses the radiated electromagnetic energy. Lens antennas have the advantage with respect to symmetrical reflecting antennas of having the source array situated behind the radiating aperture, and this eliminates blocking of the beam. This characteristic is particularly useful when a large set of sources associated with a high-performance (and hence bulky) BFN is required to support the creation of a large number of multiple beams or high-performance beam forming, for example.

The principle of the lens is to produce a propagation delay that is maximum along the axis and reduces towards the periphery where it becomes zero. The spherical wave generated by the source is thus transformed into a plane wave. Several approaches to realisation of the lens can be envisaged:

— A homogeneous dielectric material. The lens obtained in this way has the advantage of a wide passband but has a high mass.
— An assembly of metallic waveguides (a stepped or zoned lens) whose length is arranged in such a way as to produce the required phase advance to transform the incident spherical wave into a plane wave (Figure 9.46) [SCO-76]. These lenses are light but have a relatively narrow bandwidth (on the order of 5%).

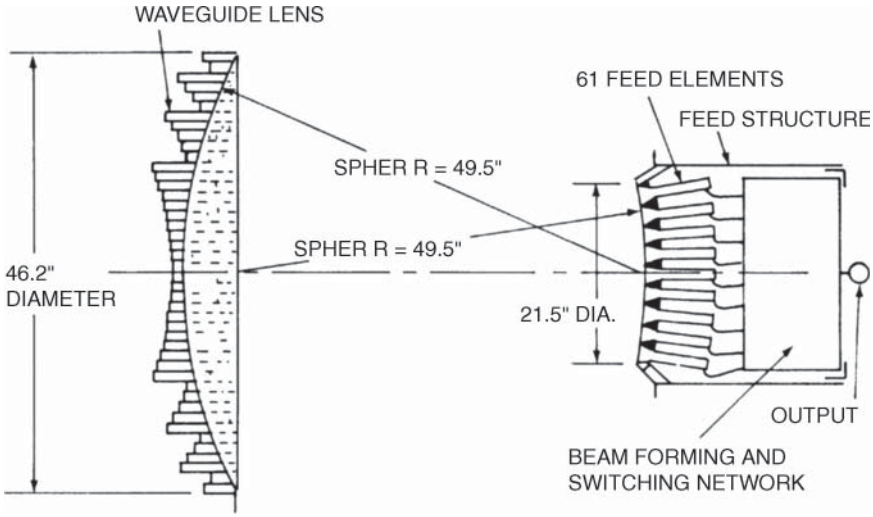


Figure 9.46 Lens antenna formed from an assembly of waveguide sections.

— An assembly of delay lines terminated by radiating elements. The bandwidth of the 'bootlace' lens is wide and the weight is intermediate between waveguide lenses and dielectric lenses.

Lens antennas suffer from a large mass and bulk. Their application seems to be reserved at present for military satellites where the capacity for dynamic reconfiguration enables an antenna radiation pattern to be obtained that has zero gain in any particular direction in order to protect against jamming. The American satellite DSCS III (defense satellite communication system) has been equipped with a receiving antenna that produces 61 beams of 2° and two transmitting antennas each generating 19 beams.

9.8.9.4 Array antennas

An array antenna uses a large number of radiating elements distributed over the area that constitutes the radiating aperture. The overall radiation pattern results from a combination in amplitude and phase of the waves radiated by the array of elements. In principle, operation of the array is similar to that of a source array located at the focus of a reflecting antenna. The difference lies mainly in the number of radiating elements and the surface area; these are determined by the required gain and width of the antenna beam that is radiated directly by the array. The radiating elements can be horns, dipoles, resonant cavities, printed elements, etc.

Properties of array antennas. The distance between the radiating elements is typically on the order of $0:6\ \lambda$. The radiation pattern is adjusted by modifying the phase and amplitude of the signal feed to the radiating elements by means of controllable power dividers and phase shifters.

For example, by feeding all the radiating elements in phase with the same amplitude, the beam obtained has characteristics similar to those of a beam generated by a reflecting antenna with uniform illumination; the maximum gain is proportional to $(\pi D/\lambda)^2$, and the 3 dB beamwidth is on the order of λ/D in radians: that is, around $60\ \lambda/D$ in degrees. By attenuating the amplitude on the periphery of the radiating aperture, the side-lobe level is reduced and the beamwidth increased. On the other hand, the on-axis gain decreases.

By feeding the elements with a phase that varies linearly from one element to the next from one edge of the array to the other, an inclination of the phase plane with respect to the surface of the array can be introduced, and this modifies the orientation of the beam.

Feeding an array antenna. With a conventional array antenna, the antenna input power is provided by a conventional power amplifier. Of course, on account of the law of reciprocity, the antenna operates in a similar manner on reception, and a low noise amplifier is connected at the output of the beam forming array.

The antenna efficiency is determined by the amplitude weighting at the edge of the array and the ohmic losses in the power splitters and phase shifters (from one to several dB depending on complexity). The ohmic losses in the power distribution constitute a critical parameter.

A shaped beam is obtained by feeding the radiating elements with a particular amplitude and phase distribution of the power available at the antenna input. Dynamic control of the beam is obtained by using controllable power dividers and phase shifters.

9.8.9.5 Active antennas

With active antennas, amplifier modules directly power the radiating elements. It is a more advanced phase array antenna in which each antenna element has its own transmit and receive units, all controlled by the computer. According to the total power to be radiated, the power available per amplifier module, and the number of radiating elements, an active module can be connected to a single radiating element or a small group of radiating elements.

In the case of an antenna performing both transmitting and receiving functions, the active module also includes the LNA and transmit–receive separation, which is performed at the module input by a circulator.

An active antenna constitutes in principle a directly radiating array. It is, however, conceivable to combine the active array with a one- or two-reflector mounting. Use of a reflector mounting enables a large radiating aperture to be obtained without the size of the array becoming prohibitive (and causing problems of folding, etc.). For example, by illuminating an auxiliary parabolic reflector, whose focal point coincides with that of the main reflector, with the near field of the array, a magnified image of the array is obtained (Figure 9.47). The choice between an array antenna with direct radiation and an array with illumination by reflector depends on considerations associated with the number of beams, the state of the art of the technology, the power used, etc.

Beam shaping. The beam-shaping elements (attenuators and phase shifters) form an integral part of the active modules. They are located upstream of the power amplifying elements and downstream of the low noise amplifiers. Figure 9.48 shows the block diagram of the Ku-Band active transmit antenna that was designed for the French Technological Satellite Program Stentor [ALB-03]. The antenna has 3 RF input accesses and radiates via the 48 radiating elements. The BFN, 48 SSPAs, filters and radiating elements, RF and TM/TC/DC harnesses, and passive calibration hardware are supported by a composite structural carbon sandwich panel and composed of a radiating panel mounted on the satellite earth-panel. The radiating panel interfaces mechanically with the platform through five mounting feet and thermally by means of thermal control hardware in order to transfer heat dissipation to the spacecraft radiating surfaces. Coupled thermal control is realised in two ways – normally it uses capillary pumped loop (CPL) technology; the redundant thermal control hardware makes use of variable conductance heat pipes (VCHP). Other equipment of the active antenna is implemented inside the spacecraft: the preamplifier unit, antenna controller (CCU), ultra-stable oscillator (USO), and calibration unit

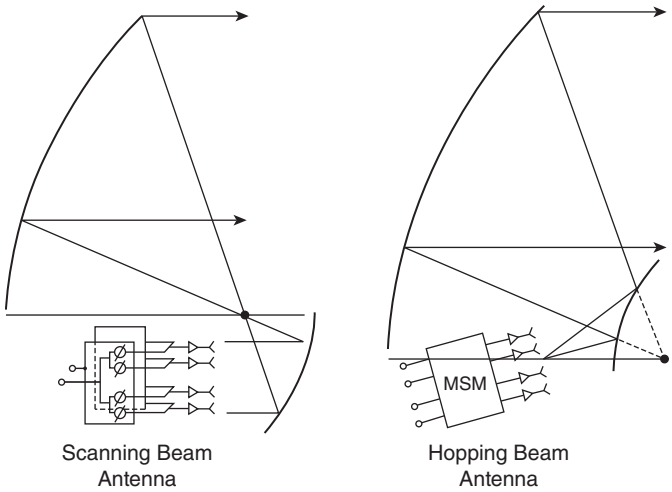


Figure 9.47 Antenna combining a phased array with reflectors. Source: reproduced from [SOR-88] with the permission of the AIAA.

(TXCAL). Beam shaping can be realised either at the operating frequency of the antenna or at IF. In the latter case, frequency conversion is performed between the beam forming elements that define the amplitude and phase distribution and the radiating elements. The latter approach facilitates realisation of the elements, particularly when the antenna operates at high frequency.

Advantages of active antennas. The advantages lie in:

— The use of low-power, solid-state amplifiers that enable good linearity to be obtained
— The high reliability associated with realising a function by parallel connection of a large number of identical elements (failure of elements leads to progressive degradation with little noticeable effect on performance)
— The possibility of reproducing components

Active antennas also permit high EIRPs to be obtained. For a given antenna (as defined by the coverage), limitation of the radiated power results from technical limitations of the power of a single amplifier per channel (around 250 W with a TWT). With an active antenna, the radiated power depends on the power of the active module and the number of radiating elements.

The disadvantages of the active antenna arise mainly in connection with the losses in the modules (ohmic losses and the limited efficiency of amplifiers). Losses between the amplifier module and the radiating element are limited due to their proximity.

Technology of active antennas. Modern techniques of high-frequency circuit integration, particularly those of MMICs, permit realisation of active modules of small bulk and low mass. These modules can incorporate beam-shaping devices (controllable attenuators and phase shifters). It is also possible to realise beam forming using optical techniques.

The radiating elements themselves can be realised using microwave IC technologies on dielectric substrates. An array of printed radiating elements (a patch) can be produced by etching a conducting layer that has been deposited on a dielectric substrate over a ground plane (the microstrip techniques). The problems of printed antennas lie in the narrowness of the passband. Furthermore, the array radiates in a direction perpendicular to the plane of the printed antenna, and this poses difficulties in integrating the active modules and the supply to the radiating elements. One solution to the narrowness of the passband is to feed the printed element via

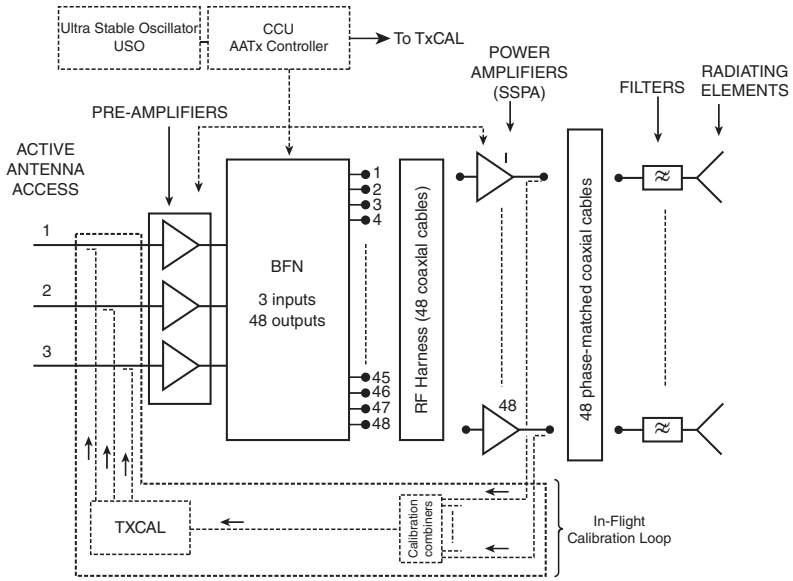


Figure 9.48 Organisation of the active transmit antenna of the STENTOR satellite.

electromagnetic coupling (electromagnetically coupled patch [EMCP]). The passband can attain 15% of the operating frequency.

Radiating elements realised from slotted lines can also be considered. A slotted line consists of two parallel conductors on the same face of a dielectric substrate deposited on a ground plane. To cause radiation, the interval between the two conductors is modified as the extremity of the substrate is approached. Different types of slot antenna on dielectric substrates (tapered slot antennas [TSAs]) are obtained according to the profile used (Figure 9.49):

— *Vivaldi antenna*: The width of the slot varies exponentially.
— *Linearly tapered slotline antenna (LTSA)*: The width of the slot varies linearly.
— *Constant-width slot antenna (CWSA)*: Discontinuity in the slot width.

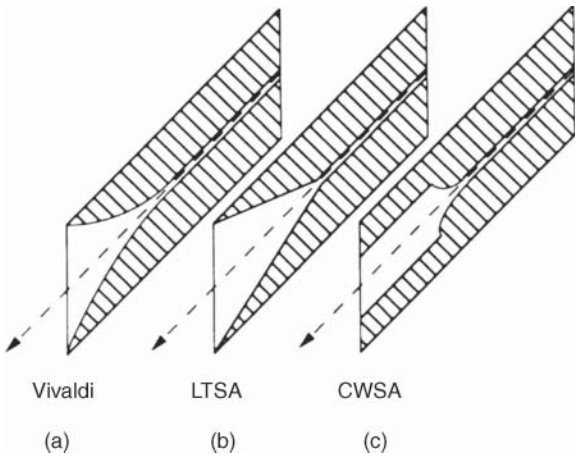


Figure 9.49 Types of tapered slot antennas: (a) Vivaldi; (b) LTSA; (c) CWSA.

Slot antennas on dielectric substrates belong to the class of travelling wave antennas. Propagation of the wave occurs along the slot, and the antenna radiates with linear polarisation in the direction of the extremity of the slot, in the plane parallel to the substrate. It is thus simple to locate various radiating elements side by side to produce an array. Furthermore, each slot can be fed from an active module integrated on to the same substrate.

9.8.9.6 Large deployable antennas

To provide a gain large enough to serve hand-held mobile terminals in L and S bands, large-diameter (up to 15–20 m) antennas are launched on board geostationary satellites. Different solutions can be considered for deployable reflectors: umbrella-like structures, inflatable antennas, etc. The so-called AstroMesh deployable reflector embodies a new concept for deployable space structures: a pair of ring-stiffened, geodesic truss domes, in which the ring is a truss deployed by a single cable (Figure 9.50) [THO-01] and [MAR-09].

Another challenge for these antennas is the design of the feed systems to feed unfurlable reflectors. Candidate solutions for the BFN technology include the semi-active antenna concept with amplifiers powering feeds via Butler-like matrices. Power levels at the amplifiers are kept close to nominal to ensure optimum power efficiency, and phase-only control of signals at the amplifier inputs allows shaping and pointing of the beams [ROE-95].

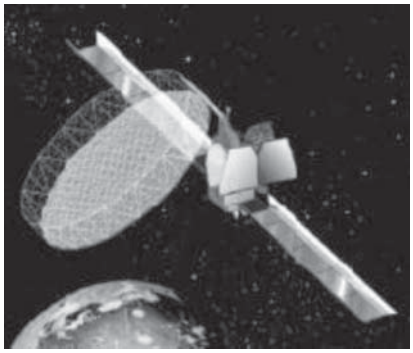


Figure 9.50 Artist's view of a geostationary satellite with a 12 m unfurlable antenna. Source: courtesy of Northrop Grumman.

9.9 CONCLUSION

Development of antenna technology has led to the realisation of antennas with shaped beams and multiple beams that can be mobile. These antennas permit frequency reuse using space and/or polarisation diversity. The search for high directivity leads to the use of radiating apertures of large dimensions. Large reflectors can be deployed in orbit. Increased directivity is directly associated with increased pointing accuracy. Limited precision of attitude and orbit control and pointing uncertainty require the use of a pointing control system.

For repeaters, technological progress has reduced mass and power consumption while enabling more sophisticated functions to be realised. Hybrid and MMIC technologies have now been space qualified and employed.

The use of resonators enables a large reduction of the size and mass of multiplexing filters. Although SSPAs are used at L and C bands, TWTs remain the preferred choice when high power and high RF/DC efficiency are required. A particularly large effort is being devoted to an increase of efficiency in order to reduce the power consumption and heat dissipation of both tube amplifiers and solid-state amplifiers. Both TWTA and SSPA technologies have been used for satellite payload [LOH-15]. SSPA is considered advantageous in term of mass and cost at low frequency but not as efficient as TWTA at high frequency. It is expected that gallium nitride (GaN) will be able to deliver improved performance compared to gallium arsenide (GaAs).

Finally, the complexity of antennas and repeater equipment poses numerous problems of analysis, interference reduction, and integration. Simulation software tools are used to optimise the radiation patterns of antennas and the performance of microwave circuits, and computer-aided design facilitates mechanical integration. Economical manufacture of these to achieve simulation and analytical performance is still a significant challenge.

REFERENCES

[ALB-03] Albert, I., Chane, H., and Raguenet, G. (2003). *The STENTOR active antenna: design, performances and measurement results*. Paper presented at the IEEE International Symposium on Phased Array Systems and Technology.

[ASS-81] Assal, F., Betaharon, K., Zaghloul, A., and Gupta, R. (1981). Wideband microwave switch matrix for SS-TDMA systems. In: *ICDSC 5th International Conference on Digital Satellite Communications*, Genoa, 421–427. IEEE.

[AUB-92] Auboin, J. (1992). Second generation DBS satellite TWTs. In: *AIAA 14th International Communication Satellite Systems Conference*, Washington, DC, 133–141.

[BAU-85] Bauer, R., Steiner, W., and W uerscher, W. (1985). Method and instrumentation for the precise measurement of satellite transponder saturation point. *International Journal of Satellite Communications* **3**: 265–270.

[BEN-86] Benet, C.A. and Dewell, R.D. (1986). Antenna beam pointing error budget analysis for communications satellites. *Space Communication and Broadcasting* **4** (3): 205–214.

[BER-71] Berman, A. and Mahle, C.E. (1970). Non linear phase shift in travelling-wave tubes as applied to multiple access communications satellite. *IEEE Transactions on Communication Technology* **18**: 37–48.

[BOS-04] Bosh, E. and Fleury, G. (2004). Space TWTs today and their importance in the future. Paper 3259, presented at the AIAA 14th International Communication Satellite Systems Conference, Monterey.

[CAM-90] Cameron, R.J., Tang, W.C., and Kudsia, C.M. (1990). Advances in dielectric loaded filters and multiplexers for communications satellites. In: *AIAA 13th Communication Satellite Systems Conference*, Los Angeles, 264–273.

[CAR-08] Caron, M. and Huang, X. (2008). Estimation and compensation of amplifiers gain and phase mismatches in a multiple port amplifier subsystem. Presentation at the ESA Workshop on Advanced Flexible Telecom Payloads, ESTEC, The Netherlands.

[DON-69] F. Donnelly, Graunas, R., and Killian, J. et al. (1969) The design of the mechanically despun antenna for the Intelsat III communications satellite, *IEEE Transactions on Antennas and Propagation*, **17** (4), pp. 407–415. IEEE.

[FUE-73] Fuenzalida, J.C., Shimbo, O., and Cool, W.L. (1973). Time domain analysis of intermodulation effects caused by nonlinear amplifiers. *COMSAT Technical Review* **3** (1): 89–143.

[HAT-69] Hatch, G.W. (1969). Communications subsystem design trends for the DSC program. *IEEE Transactions on Aerospace and Electronic Systems* **5** (5): 724–730.

[HOE-86] Hoeber, C.F., Pollard, D.L., and Nicholas, R.R. (1986). Passive intermodulation product generation in high power communications satellites. In: *AIAA 11th International Communication Satellite Systems Conference*, San Diego, 361–375.

[JON-08] Jones, T. et al. (2008). Payload architectures and hardware developments for flexible multi-beam GEO communication systems. Presentation at the ESA Workshop on Advanced Flexible Telecom Payloads, ESTEC, The Netherlands.

[KOV-91] Kovac, R., Lee, M., Miller, N., et al. (1991). SAW-based IF processors for mobile communications satellites. Presentation at the IAF Congress, Montreal.

[LOH-15] Lohmeyer, W.Q., Aniceto, R.J., and Cahoy, K.L. (2015). Communication satellite power amplifiers: current and future SSPA and TWTA technologies. *International Journal of Satellite Communications and Networking* **34** (2): 95–113.

[MAR-09] Marks, G., Keay, E., Kuehn, S., et al. (2009). Performance of the AstroMesh deployable mesh reflector at Ka-band frequencies and above. Presentation at the Ka and Broadband Communications, Navigation and Earth Observation Conference, Calgari, Italy.

[MOA-86] Moat, R. (1986). ACTS baseband processor. In: *IEEE GLOBECOM 86*, Houston, 16.4.51–16.4.56.

[MOR-88] Moreli, G. and Matitti, T. (1988). The Italsat satellite program. In: *AIAA 12th Communication Satellite Systems Conference*, Arlington, 112–122.

[NAD-88] Naderi, M. and Kelly, P. (1988). NASA's advanced communications technology satellite (ACTS): an overview of the satellite, the network, and the underlying techniques. In: *AIAA 12th Communication Satellite Systems Conference*, Arlington, 204–224.

[PEN-84] Pennoni, G. (1984). A TST/SS-TDMA telecommunications system: from cable to switchboard in the sky. *ESA Journal* **8**: 151–162.

[PRI-93] Pritchard, W., Suyderhoud, H., and Nelson, R. (1993). *Satellite Communication Systems Engineering*, 2e. Prentice Hall.

[ROE-95] Roederer, A.G. (1995). Semi-active multimatrix reflector antennas. *Electromagnetics* **15** (1).

[SAG-87] Saggese, E. and Speziale, V. (1987). In-orbit testing of digital regenerative satellite: the Italsat planned test procedures. *International Journal of Satellite Communications* **5** (2): 183–190.

[SAL-81] Saleh, A. (1981). Frequency independent and frequency dependent nonlinear models of TWT amplifiers. *IEEE Transactions on Communications* **29** (11): 1715–1720.

[SCO-76] Scott, W.G., Luh, H.S., Smith, T.M., and Grace, R.H. (1976). Development of multiple beam lens antennas. In: *AIAA, 6th Communications Satellite Systems Conference*, Montreal, 76–250.

[SEY-06] Seymour, C.D. (2006). Development of high power solid state power amplifiers. Presentation at the AIAA International Communications Satellite Systems Conference (ICSSC), San Diego.

[SOR-88] Sorbello, R. (1988). Advanced satellite antenna developments for the 1990s. In: *AIAA 12th International Communication Satellite Systems Conference, Arlington, VA, March, Paper AIAA-1988-873*, 652–659.

[TAN-90] Tang, W.C. and Kudsia, C.M. (1990). Multipactor breakdown and passive intermodulation in microwave equipment for satellite applications. Presentation at the IEEE Military Communications Conference MILCOM 90, Monterey.

[THO-01] Thomson, M. (2001). Flight heritage for the AstroMesh™ deployable reflector. Paper 300, presented at the AIAA 19th Communication Satellite Systems Conference, Toulouse.

[VOI-08] Voisin, P. et al. (2008). Flexible communication payloads: a challenge for the industry and a new perception of solutions for operators. Presentation at the ESA Workshop on Advanced Flexible Telecom Payloads, ESTEC, The Netherlands.

10 THE PLATFORM

A communications satellite incorporates various subsystems with distinct functions. It is customary to distinguish the communication payload (or communication modules: antennas and repeater), as discussed in Chapter 9, from the platform (also called the *bus* in some literature) or service module that supports and powers the payload. As shown in Figure 10.1, the payload typically consists of antennas, repeaters, and other communication equipment forming the satellite communication module; and the service module has a propulsion tank, a solar panel, batteries, avionics, a liquid apogee engine, etc.

The organisation of a communications satellite platform is determined mainly by the following:

— The requirements of the communications payload
— The nature and effects of the space environment
— The performance of launchers, and the constraints that they impose

The communications mission conditions the design of the payload; as far as the platform is concerned, this design results in requirements such as the electrical power to be provided; the payload mass that can be accommodated; the antenna pointing accuracy; the thermal power to be extracted; the space required for equipment mounting; the number of telemetry, tracking, and command (TTC) channels; and so on.

The performance of the platform with respect to its ability to accommodate large payloads is typically illustrated by a domain in *payload mass* versus *payload direct current (DC) power* reference axes. An example is given in Figure 10.2 for Alphabus, the large platform being developed in Europe for communications satellites (see Section 10.7). In the figure, different performance domains are shown, depending on the type of propulsion system, which conditions, for a given mission life, the amount of propellant to be embarked (Section 10.3), and therefore the mass left available for the communications payload considering a given satellite mass at launch (i.e. capability of the selected launch vehicle). Recently, the European Space Agency (ESA) has developed a new satellite platform in the Neosat programme, which comprises two platform lines: Eurostar Neo by Airbus, and Spacebus by the Thales Alenia Space (TAS), for future satellite communication systems and networks.

The nature and effects of the space environment affect orbit control, subsystem organisation, and the choice of materials and components. Launchers impose mechanical constraints on the structure of the satellite. Their performance limits the mass that can be injected into orbit and influences the specification of the propulsion subsystem. Furthermore, the limited enclosed volume within the fairing may require the solar panels and antennas to be folded.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

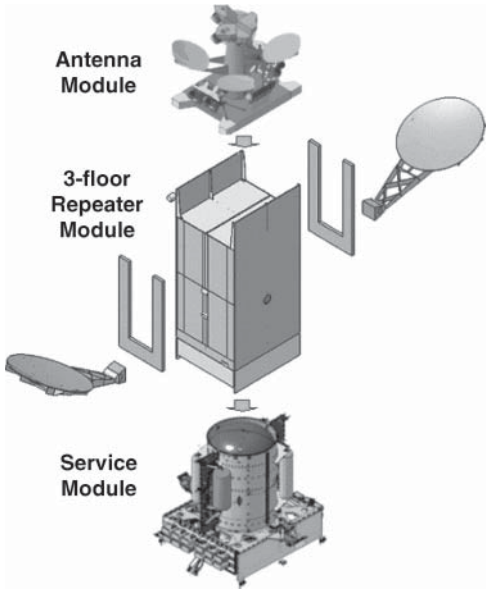


Figure 10.1 Geostationary platform configuration. Source: courtesy of EADS Astrium and Thales Alenia Space.

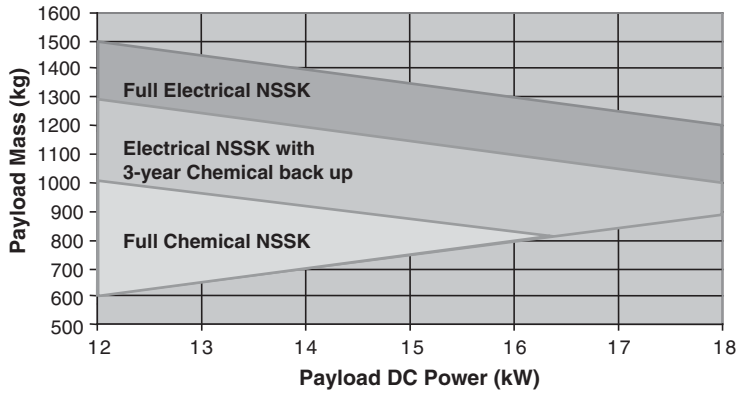


Figure 10.2 Alphabus payload mass vs payload DC power domain (15-year life). Source: reproduced from [BER-07] with permission.

The basic information necessary to determine the required electrical and thermal power to be dissipated by the payload has been provided in Chapter 9. The launch procedures and the performance of launchers are presented in Chapter 11. Finally, the nature and effects of the environment are the subjects of Chapter 12.

10.1 SUBSYSTEMS

A list of the platform subsystems is given in Table 10.1. The functions to be provided and the most significant characteristics are indicated. Three common characteristics are not indicated but are essential and should be emphasised:

— Minimum mass
— Minimum consumption
— High reliability

Each subsystem is specified and designed for the particular mission to be fulfilled, taking into account these three criteria, the technology used, and the characteristics of other subsystems. The performance and specification of a particular subsystem depend on the presence of other subsystems, and this influences the interfaces between subsystems. Each interface is itself defined by numerous characteristics, of which the most typical are given in Figure 10.3. Particular attention must be paid to the problems of electromagnetic compatibility (EMC) and the numerous items of radio-frequency (RF) equipment operating in different frequency bands at different power levels.

In this chapter, the various subsystems are examined in order to clarify their essential characteristics.

Table 10.1 A list of example Eurostar platforms

	Eurostar 2000	Eurostar 2000+	Eurostar 3000	Eurostar 3000LX	Eurostar Neo
Launch mass (10^3 kg)	2.3	3.4	5.7	6.4	5.7–6.4
Maximum payload mass (kg)	400	550	1200	1200	1200
Spacecraft (kW)	2–4	4–8	8–14	14–20	7–25

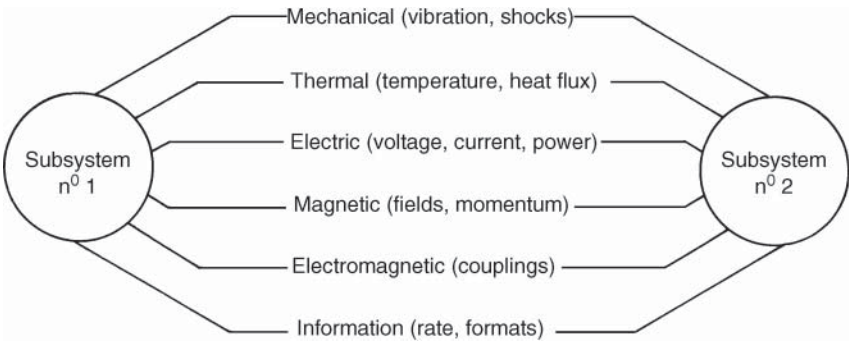


Figure 10.3 Interfaces between two subsystems.

10.2 ATTITUDE CONTROL

Motion of the satellite can be resolved into motion of its centre of mass in the earth-centred coordinate system and motion of the body of the satellite about its centre of mass. The motion of the centre of mass is characteristic of the satellite orbit, and its control is discussed in Chapter 2. Motion of the body of the satellite about its centre of mass is determined by the evolution of the attitude.

The attitude of the satellite is represented with respect to the *yaw*, *roll*, and *pitch* axes of a local coordinate system (Figure 10.4). This coordinate system is centred on the centre of mass of the satellite; the *yaw axis* points in the direction of the centre of the earth; the *roll axis* is in the plane of the orbit, perpendicular to the first and oriented in the direction of the velocity; and the *pitch axis* is perpendicular to the other two (and, hence, perpendicular to the orbit) and oriented in such a way that the coordinate system is regular (towards the south for a geostationary satellite). In the nominal attitude configuration, the axes of the satellite-fixed coordinate system are, in principle, aligned with the axes of the local coordinate system. The attitude of the satellite is represented by the angles of rotation about the various axes between the local coordinate system and the satellite-fixed coordinate system.

Maintaining attitude is fundamental for the satellite to fulfil its function. The accuracy and reliability of this subsystem determine the performance of most of the other subsystems; for example, narrow beam antennas and solar panels must be suitably oriented.

10.2.1 Attitude control functions

The role of attitude control usually consists of maintaining the mechanical axes in alignment with the local coordinate system to an accuracy defined by the amplitude of rotation about each of the axes (the value of amplitude corresponds to a given probability of remaining within range). To be specific, typical ranges are ±0.05° for roll and pitch and ±0.2° for yaw for a geostationary satellite.

In certain cases, a constant bias or a particular law of progression about one or more axes may be required in accordance with the requirements of the mission and the particular orbit concerned.

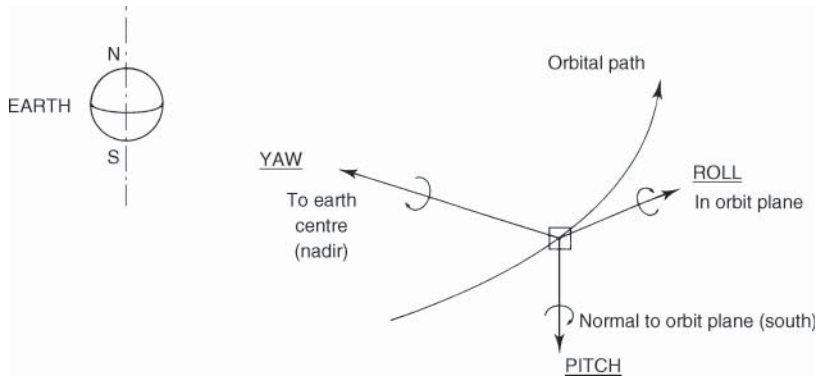


Figure 10.4 Local coordinate system: definition of yaw, roll, and pitch.

Maintaining attitude requires two functions:

— A steering function causes the part of the satellite that must be oriented towards the earth to turn about the pitch axis in order to compensate for the apparent movement of the earth with respect to the satellite. For a geostationary satellite, this rotation is made at constant velocity equal to one revolution per day (0.25° per minute).
— A stabilisation function compensates for the effects of attitude-disturbing torques. The disturbing torques are created by gravitational forces, solar radiation pressure, and interaction between current loops and the terrestrial magnetic field (see Chapter 12). These natural disturbing torques are small (on the order of 10^{-4}–10^{-5} N m). In contrast, disturbing torques created by lack of alignment of the thrust of the orbit control actuator with respect to the position of the centre of mass can be large (on the order of 10^{-2}–10^{-3} N m).

In the past, *passive attitude control* has been considered. It involves using the effects of natural torques to maintain the required attitude. For example, use of the gravity gradient, which tends to align the axis of lowest inertia with a local vertical (see Section 12.2.1.1), has been investigated [MOB-68]. The accuracy obtained (at best, a few degrees) is incompatible with the pointing requirement of telecommunications satellites, which use narrow-beam antennas requiring precise pointing.

Consequently, *active attitude control* is used. The process generally involves:

— Measurement of the attitude of the satellite with respect to external references
— Determination of the attitude with respect to the local coordinate system
— Computation of the actuator commands
— Execution of the corrections by means of actuators mounted on the satellite
— Evolution of the attitude in accordance with the satellite dynamics under the effect of actuating and disturbing torques

The system can operate in a closed loop on board the satellite; that is, the actuator control is directly generated by on-board equipment as a function of the outputs of attitude sensors. The attitude sensor outputs may also be transmitted to the ground on the telemetry (TM) channels and the actuators operated by the telecommand (TC) channels to restore the attitude with evaluation on the ground of the required corrective actions. Closing of the control loop on the ground is possible only if dynamic progression of the satellite attitude is slow. In practice, depending on the techniques used and the axes considered, a combination of these two principles is often used.

Although active attitude control is realised, it is useful to take advantage of natural effects as follows:

— Gyroscopic stiffness obtained by the creation of angular momentum on board the satellite
— Control torques using magnetic coils interacting with the terrestrial magnetic field
— Active torques by taking advantage of solar radiation pressure

Use of natural effects results in greater operational flexibility of attitude control and a reduction of the quantity of consumable propellant to be embarked on the satellite.

10.2.2 Attitude sensors

These sensors measure the orientation of the satellite axes with respect to external references (such as the earth, the sun, or the stars) or the progression of the orientation with time

(gyrometers). Their essential characteristic is accuracy. It depends not only on the procedure used but also on the alignment errors of the detector with respect to the body of the satellite.

The sensors most used on board geostationary communications satellites are solar detectors, terrestrial horizon detectors, and gyrometers. For certain applications, star sensors widen the range of possibilities. Finally, it is possible to use a RF beacon or a laser to obtain a measure of attitude.

10.2.2.1 Sun sensors

Sun sensors use photovoltaic elements that produce a current when illuminated by the sun. They measure one or two angles between their mounting base and incident sunlight. The accuracy obtained for measurement of the angle between the direction of the sun and an axis related to the satellite is on the order of 0.01°.

10.2.2.2 Earth sensors

The earth surrounded by its atmosphere appears as a spherical black body at a temperature of 255 K when its radiation is measured in the infrared absorption band of carbon dioxide (14–16 μm). As seen from space, the image of the earth contrasts strongly with respect to the background plane, which has a temperature around 4 K. Measurement of infrared radiation, the emission of which is approximately uniform over the whole surface of the earth, by means of a thermosensitive element (such as a bolometer, thermocouple, or thermopile) permits the contour of the terrestrial globe to be detected.

Scattering of reflected solar light (the earth's albedo) can also be used; this is detected by means of photoelectric cells or phototransistors. The measurements are corrupted due to the difficulty in separating it from the tropopause. An accuracy on the order of 0.05° is obtained.

10.2.2.3 Star sensors

An image of a given portion of the sky provides a map of the stars whose relative positions are detected and compared with a reference map. About 10 stars are tracked simultaneously. The measurement accuracy is high (on the order of 10^{-3}°). There is a danger of saturation by the sun, earth, or other bright sources, which is prevented by use of light baffles. Sensors are available with digital processors and software incorporating star catalogues as well as recognition and tracking algorithms. The software enables three-axis attitude and angular velocity determination whatever the orbit is. New star trackers are based on active pixel sensor technology, offering low mass, low power, improved capabilities, and flexibility. Thanks to these devices, gyroless satellite attitude control can be considered.

10.2.2.4 Inertial units

Inertial units use accelerometers to detect translational motion of the satellite or gyrometers to measure the angular velocity about one axis. They are subject to drift and bias errors. Mechanical devices have a limited lifetime (about 10 000 hours) that prohibits their continuous use for conventional missions of 10 years. Gyrolasers overcome these limitations.

10.2.2.5 RF sensors

These sensors depend on measurement of the characteristics of radio waves transmitted to the satellite by ground radio beacons. They permit measurement of the angle between the antenna axis on board the satellite and the desired direction of the beacon. Rotation about the boresight (the yaw angle) is difficult to evaluate. By measuring the rotation of a polarised wave from a single radio beacon, a value of the yaw angle is obtained. Unfortunately, the orientation of the polarisation is affected by Faraday rotation, and the accuracy of the yaw angle is on the order of 0.5°. On the other axes, the accuracy could be as small as 0.01°.

10.2.2.6 Laser detectors

The use of a laser beam instead of a RF beacon has been considered for determining the orientation of the satellite. The expected accuracy is 0.006° for roll angles and 0.6° for the yaw angle. One of the major problems lies in attenuation of the laser beam by clouds.

10.2.3 Attitude determination

The purpose of attitude determination is to determine the orientation of the satellite in the local coordinate system, as defined in Figure 10.4. For a rotating sensor, the line of sight (LoS) of a particular object defines a cone whose axis is the axis of rotation and whose vertex half angle is the angle between the direction of the object and the axis of rotation. Rotation of the sensor results either from rotation of the satellite on which the sensor is mounted or from a scanning device associated with the sensor.

The earth and sun are privileged objects. Determination of the direction of the centre of the earth may be achieved by using a detector with a narrow field of view, which, in the course of rotation of the satellite, sweeps out a cone that intersects the terrestrial surface (Figure 10.5). Scanning the illuminated region of the earth produces a signal whose duration permits the *nadir angle* to be determined: that is, the angle between the axis of rotation and the axis joining the satellite and the centre of the earth (the nadir axis).

A solar detector permits measurement of the vertex half angle of the second cone associated with the axis joining the satellite and the sun. The axis of rotation of the sensor is situated at one of the intersections of the two cones defined by the two observations (Figure 10.6). Selection of one of the two intersections requires a third measurement or some a priori knowledge of the orientation of the satellite.

The method presented here permits determination of the direction of the axis of rotation in space, knowing the relative position of the satellite with respect to reference objects. This assumes, therefore, that the orbit of the satellite and the position of the satellite in the orbit have been exactly determined.

As far as a geostationary satellite is concerned, once the satellite is in position in its nominal orbit, the requirements of the mission (such as pointing the antennas towards the earth) no longer require determination of the attitude in space, only the orientation of the satellite with respect to the earth.

An earth sensor thus readily permits attitude determination with respect to the roll and pitch axes. The first generations of geostationary satellites made use of static horizon detectors using thermopiles; these provide a signal that is related to the angle between the direction of the centre of the earth and the direction of the optical axis of the sensor. Today measurements are obtained by mechanical scanning sensors (Figure 10.7). The advantage lies in the wider field of view, which means the sensor can be used for a large range of satellite altitudes.

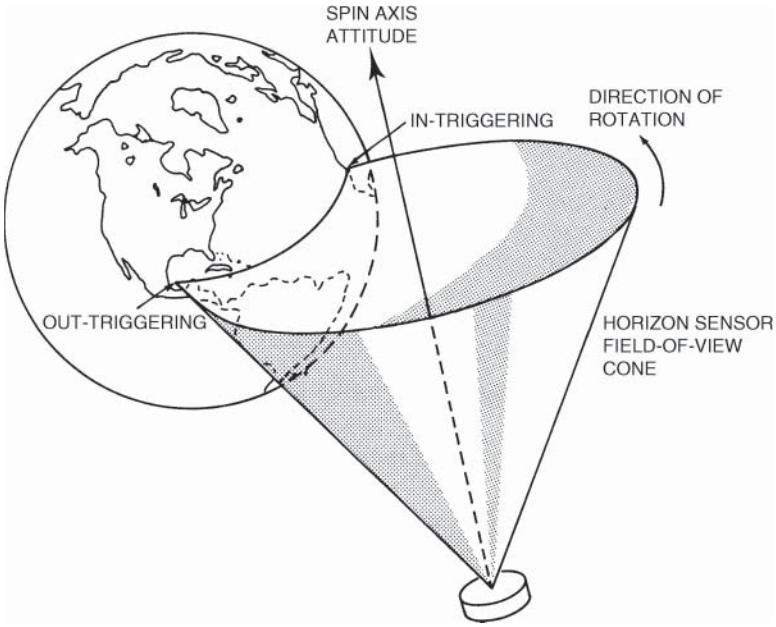


Figure 10.5 Earth contour detection. Source: reproduced from [WER-78] with the permission of Kluwer Academic Publishers.

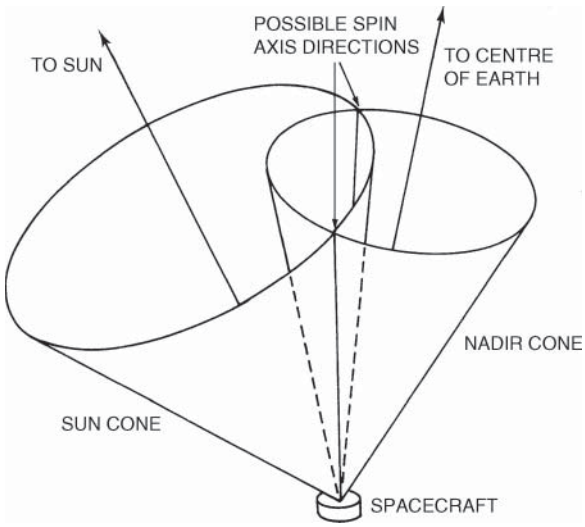


Figure 10.6 Determination of the two possible sensor spin axis directions. Source: reproduced from [WER-78] with the permission of Kluwer Academic Publishers.

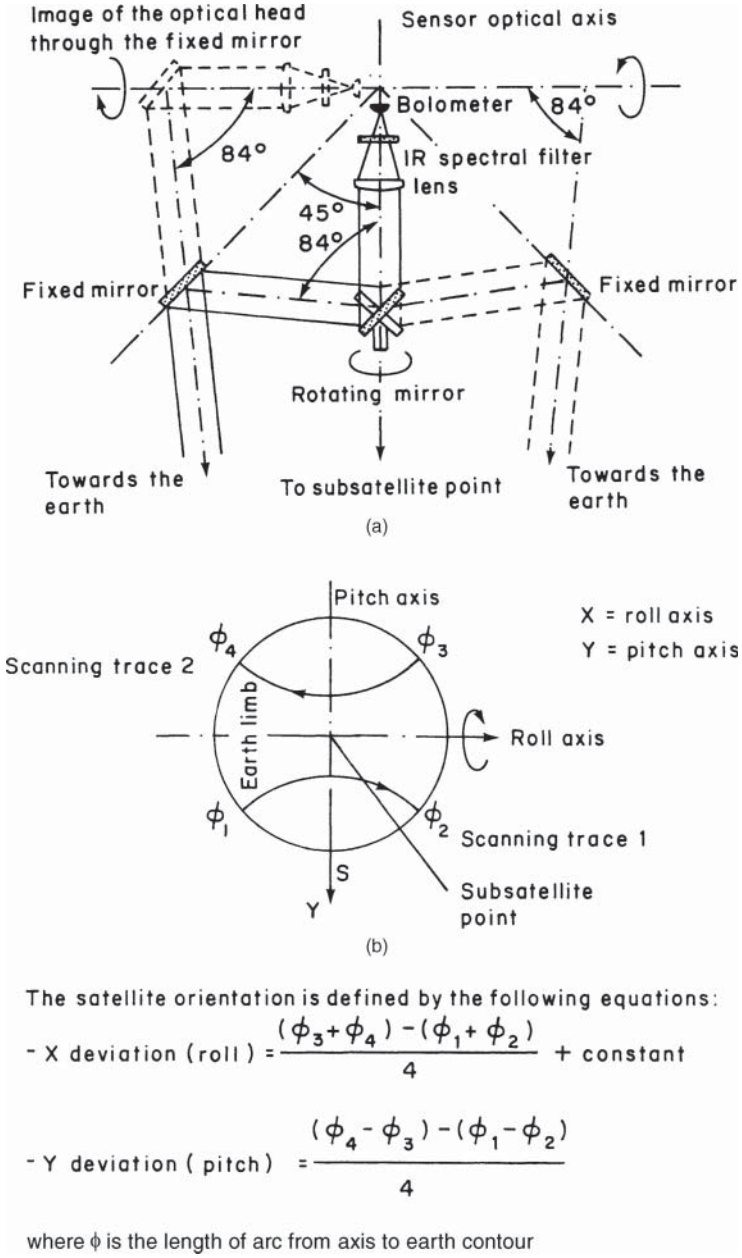


Figure 10.7 Scanning sensor: (a) principle; (b) definition of measured parameters X and Y.

The circular form of the earth's image does not permit the pointing error about the yaw axis to be obtained directly. One approach is to combine a measurement made with a sun or star sensor. Otherwise, in the case where dynamic progression is slow, knowledge of the orientation about the yaw axis can be estimated from a measurement of roll performed six hours earlier, as the directions in space of the roll and yaw axes interchange every six hours as a consequence of the rotation of the satellite about the earth. The use of specific measurement is, therefore, not indispensable as long as the disturbing torques remain small. When large disturbing torques are generated (such as those caused by the use of thrusters for orbit control), a knowledge of the progression of the orientation in yaw is obtained by using an additional sensor, such as a gyrometer. Another approach consists of making directly three-axis attitude determination with a star sensor. Years ago, star sensor technology was reserved for specific applications requiring very high accuracy because of the large cost and large mass; but this technology is now affordable for commercial telecommunications satellites, thanks to significant reductions in cost, mass, and power consumption.

10.2.4 Actuators

Modification of the attitude is generally obtained by generating a torque which, taking into account the dynamics of the particular satellite, causes an angular acceleration, or velocity, about an axis. The attitude control actuators, therefore, have the purpose of generating torques. Various types are available as follows.

10.2.4.1 Angular momentum devices

Angular momentum devices include reaction wheels and gyroscopes that exploit the principle of conservation of angular momentum.

With a *reaction wheel*, variation rate $\mathrm{d}\omega/\mathrm{d}t$ of the velocity of rotation ω of the wheel, of moment of inertia I, causes the angular momentum $H = I\omega$ to be modified, and a torque T aligned with the axis of the wheel is generated:

$$T = \mathrm{d}H/\mathrm{d}t = I\mathrm{d}\omega/\mathrm{d}t \quad (\mathrm{Nm}) \tag{10.1}$$

With a *gyroscope,* the wheel rotates at constant velocity and is gimballed about one or two axes. Any commanded change in the orientation of the moment of inertia causes generation of a torque T equal to the subsequent variation rate $\mathrm{d}H/\mathrm{d}t$ of the angular momentum vector. The limited lifetime of such steering devices makes this type of device very little used for active torque generation.

Angular momentum devices are particularly suitable for maintaining attitude when the satellite is subject to cyclic disturbing torques. Disturbing torques with nonzero mean values (caused, for example, by the effect of solar radiation pressure), and disturbing torques of excessive amplitude can require a compensating angular momentum variation that exceeds the limits for the velocity of rotation of the wheel or the orientation of the gyroscope. It is then necessary to provide an unloading torque by means of another torque generator (a thruster, for example).

10.2.4.2 Thrusters

Thrusters produce reaction forces on the satellite by expelling material (propellant) through nozzles. The force obtained is a function of the quantity of material (mass) ejected per unit time

dm/dt and depends on the specific impulse I_{sp} of the propellant used (see Section 10.3):

$$F = gI_{sp}(dm/dt) \text{ (N)} \tag{10.2}$$

where g is the normalised terrestrial gravitational constant ($g = 9:807\,\text{m s}^{-2}$).

The torque obtained depends on the length l of the lever arm with respect to the centre of mass of the satellite:

$$T = Fl \text{ (N m)} \tag{10.3}$$

The torques to be applied are on the order of 10^{-4}–10^{-1} N m. With a lever arm of 1 m, the thrusts are thus on the order of 10^{-4}–10^{-1} N. In the interests of simplicity and reducing mass, these thrusters are generally part of the thruster assembly that provides orbit control after installation in orbit (see Chapter 11). These thrusters often provide thrusts greater than the values given earlier (from a few newtons to tens of newtons). Smaller thrusts, which can be varied, are obtained by using thrusters in an on–off mode of operation with a variable duty cycle.

10.2.4.3 Magnetic coils

Magnetic coils with n turns of area S create a magnetic moment $M = nIS$ (Am^2) when fed with a current I. This magnetic moment can generate a torque T by interaction with the terrestrial magnetic field B:

$$\mathbf{T} = \mathbf{M} \times \mathbf{B} \text{ (N m)} \tag{10.4}$$

For geostationary satellites, the value of the terrestrial magnetic field is small, typically $1 \times 10^{-7}\,\text{Wb m}^{-2}$ (see Chapter 12). The torque obtained with a coil of $n = 500$ turns, with area $S = 4\,\text{m}^2$ and current $I = 1\,\text{A}$, is at most $2 \times 10^{-4}\,\text{N m}$. This can compensate for some of the disturbing torques exerted on the satellite.

10.2.4.4 Solar sails

Solar radiation pressure (see Chapter 12) applied to a surface of sufficient size is capable of generating non-negligible torques. In general, these are disturbing torques that are countered by designing the satellite so that the apparent surface in the direction of the sun is symmetrical with respect to the centre of mass. The most significant surfaces are those of the solar generators. For instance, a satellite stabilised in three axes has two symmetrical panels that are aligned with the pitch axis. By modifying the apparent surfaces of the two panels of the solar generator, it is possible to generate torques about two axes (Figure 10.8). A torque about an axis in the direction of the sun is created by introducing a symmetrical bias between the normal to each panel and the direction of the sun (the windmill effect). A further torque can be obtained about an axis perpendicular to the direction of the sun in the plane of the orbit by means of a bias of one panel alone (asymmetry of the apparent surfaces).

The bias introduced must remain limited (to about 10° maximum) in order to avoid excessive reduction of the flux captured by the solar generator. The effect of the solar sails is increased by adding surfaces that are oblique with respect to the panels at the extremities of the solar generators. The torques obtained are sufficient to compensate for most disturbing torques (except those induced during station-keeping manoeuvres).

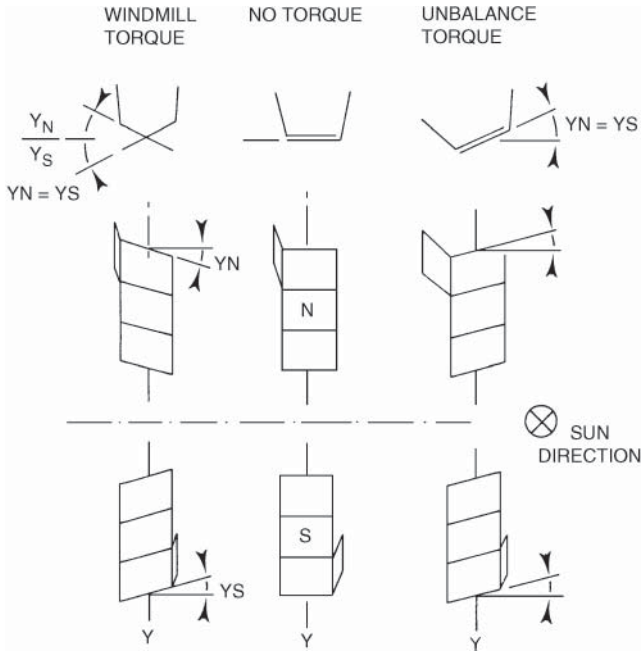


Figure 10.8 Solar sailing.

10.2.5 The principle of gyroscopic stabilisation

Gyroscopic stabilisation is obtained by creating an angular momentum on board the satellite. By virtue of the principle of conservation of angular momentum, the orientation of the angular momentum tends to remain fixed in inertial space (gyroscopic stiffness). By choosing an angular momentum aligned with the pitch axis, the pitch axis benefits from gyroscopic stiffness, which preserves a fixed orientation in space in spite of movement of the satellite in its orbit. Hence movements about the roll and yaw axes are limited.

The benefit of gyroscopic stabilisation is better appreciated if the effects of disturbing torques on a satellite are compared with and without angular momentum generation. Figure 10.9 shows the effect of a disturbing torque T_d, which is exerted about the Z axis of a satellite whose mechanical axes x, y, and z are initially aligned with the X, Y, and Z axis of a reference frame:

— The satellite *without angular momentum* starts to rotate about the z axis with a *constant angular acceleration* $d\Omega/dt$ given by:

$$d\Omega_z/dt = T_d/I_z \text{ (rad/s}^2\text{)} \tag{10.5}$$

where I_z is the moment of inertia of the satellite about the z axis.

— The satellite *with on-board angular momentum* H about the y axis as a consequence of the gyroscopic effect rotates about the x axis at a *constant angular velocity* Ω_x given by:

$$\Omega_x = T_d/H \text{ (rad/s)} \tag{10.6}$$

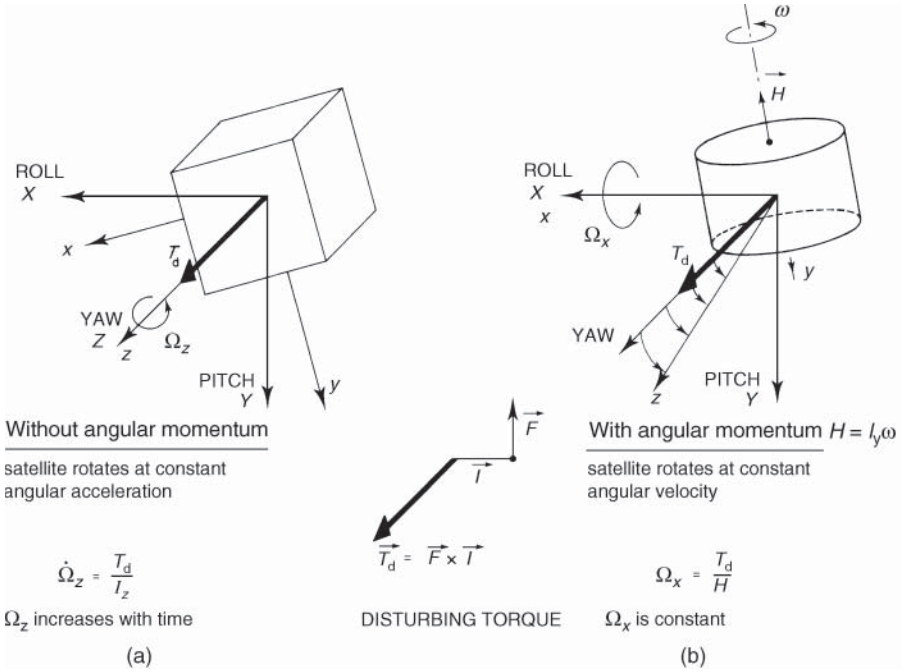


Figure 10.9 Action of a disturbing torque on a satellite: (a) without on-board momentum; (b) with on-board momentum.

Example 10.1 The time required to obtain an attitude pointing error of 0.1° is calculated, assuming a disturbing torque of $T_d = 1 \times 10^{-5}$ N m and a satellite moment of inertia I about each of the axes equal to 1000 m^2 Kg.

In the case of a satellite without angular momentum, the angular acceleration is:

$$d\Omega_z/dt = T_d/I_z = 1 \times 10^{-5}/1000 = 1 \times 10^{-8} \text{rad/s}^2$$
$$= (360/2\pi)10^{-8} = 5.73 \times 10^{-7} \text{degree/s}^2$$

The rotation θ_z about the z axis is at constant angular acceleration, hence $\theta_z = (1/2)(d\Omega_z/dt)t^2$ and $t = [2\theta_z/(d\Omega_z/dt)]^{1/2} = [0.2/5.73 \times 10^{-7}]^{1/2} = 590\,\text{s} = 9.8\,\text{min}$. After 9.8 minutes, the satellite, which was initially aligned with the reference frame, reaches the permitted pointing error limit, and a corrective action must be performed.

In the case of a satellite with angular momentum of $H = 100$ N m s about the y axis, rotation about θ_x the x axis occurs at a constant angular velocity given by:

$$\Omega_x = T_d/H = 1 \times 10^{-5}/100 = 1 \times 10^{-7} \text{rad/s}$$
$$= (360/2\pi)10^{-7} = 5.73 \times 10^{-6} \text{degree/s}$$

Since the angular velocity is constant, $\theta_x = \Omega_x t$ and $t = [\theta_x/\Omega_x] = [0.1/5.73 \times 10^{-6}] = 17452\,\text{s} = 290\,\text{min} = 4.8\text{h}$. In this case, 4.8 hours are required for a satellite, which was initially aligned with the reference frame, to reach the permitted pointing error limit; a longer time is, therefore, available before corrective action must be performed.

An angular momentum H of 100 N m s can be obtained in either of the following ways:

— By rotation of the entire satellite (spin stabilisation) about the y axis with a velocity ω_y such that:

$$H = \omega_y I_y, \text{ hence}$$

$$\omega_y = H/I_y = 100/1000 = 0.1\text{rad/s} \cong 0.95\text{rpm(revolutions per minute)}$$

— By means of a momentum wheel mounted within the satellite

The basic design consists of a highly reliable bearing unit, a spoked flywheel mass, and a DC motor in a vacuum-tight, evacuated housing. The heavy flywheel forming the rotor of an electric motor is rotating at high speed (Figure 10.10). The wheel typically contains built-in wheel drive electronics. Depending on the angular momentum to be obtained (say 15–70 N m s), the mass may be between 5 and 10 kg and the speed of rotation between 5000 and 20 000 rpm. To limit friction torques, the wheel may be suspended on magnetic bearings.

10.2.6 Spin stabilisation

Spin stabilisation was used on early communications satellites, such as numerous US national and export satellites, most based on the Boeing 376 platform (Figure 10.11), and Intelsat VI. The satellite is given a rotation movement (spin) of several tens of revolutions per minute about one of the principal axes of inertia. This is a simple process that benefits from the properties of the gyroscope but has the disadvantage of either leading to a rotating antenna with low-gain antenna radiation pattern or necessitating contra-rotation of the antenna or the supporting platform (see Section 9.8). In the absence of disturbing torques, the angular momentum H maintains a fixed direction with respect to an absolute reference frame. For a geostationary satellite, the axis of rotation is thus always parallel to the line of the poles (the pitch axis).

Oscillations of the axis of rotation about the direction of the angular momentum H (nutation) arise when the moment of inertia about the axis of rotation is not sufficiently large with respect to that about the other perpendicular axes. These oscillations must be damped by internal dissipation of kinetic energy (nutation damping) or actively controlled by using thrusters, in which case the system has a tendency to instability (when the moment of inertia about the axis of rotation is equal to or smaller than that about the other axes). This situation arises with highly elongated satellites of which a large part (despun shelf accommodating the communication payload) does not participate in the creation of angular momentum (dual-spin stabilisation). This was the case, for example, with the Intelsat VI satellite.

Disturbing torques have two effects: they reduce the velocity of rotation about the stabilised axis, and they cause a stabilised axis pointing error. It is, therefore, necessary to maintain the velocity of rotation (for example, by means of thruster 1 in Figure 10.12) and correct the pointing error.

When the component of the disturbing torque perpendicular to the axis of rotation is constant, the pointing error consists of a constant velocity drift about an axis perpendicular to the axis of the torque. Correction requires application of a torque that cancels the drift. In general, the correcting torque is applied periodically as soon as the pointing error reaches the maximum tolerated; a thruster, such as thruster 2 in Figure 10.12, is used. It operates by means of impulses in synchronism with the velocity of rotation of the satellite.

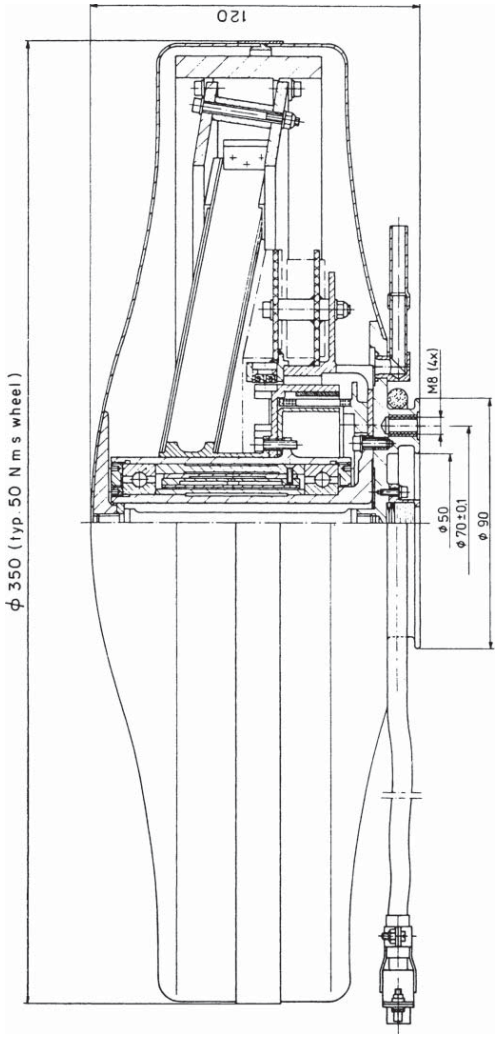


Figure 10.10 Inertia wheel. Source: reproduced with the permission of Teldix GmbH.

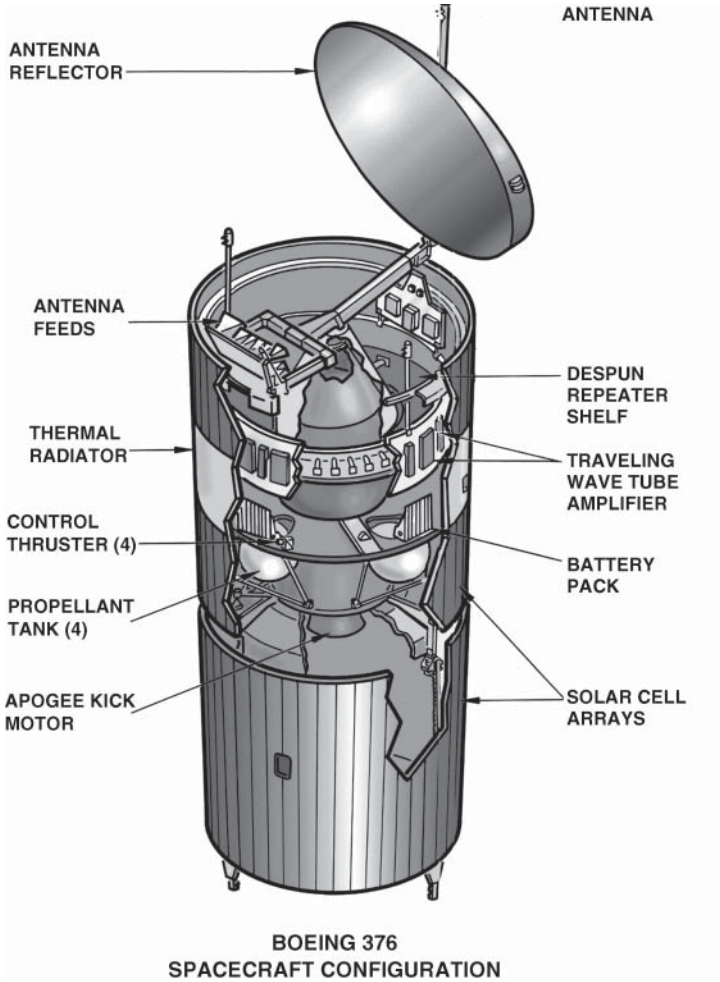


Figure 10.11 Spin stabilisation: the Boeing HS 376 spacecraft.

10.2.7 Three-axis stabilisation

The term *three-axis stabilisation* denotes an attitude-control system in which the body of the satellite maintains a fixed orientation with respect to the local coordinate system. It should be noted that spin-stabilised satellites are also, strictly, described as stabilised in three axes since active control of the attitude about the three reference axes is provided. The nomenclature is thus a little restrictive.

Since the body of the satellite maintains a fixed orientation with respect to the earth, mounting of large antennas is facilitated. Furthermore, it is simple to install unfolding solar panels that are aligned with the pitch axis and rotate about this axis in order to follow the apparent daily movement of the sun about the satellite.

The daily rotation of the satellite body about the pitch axis no longer provides sufficient gyroscopic rigidity to combat disturbing torques. It is, therefore, necessary to design a rapid

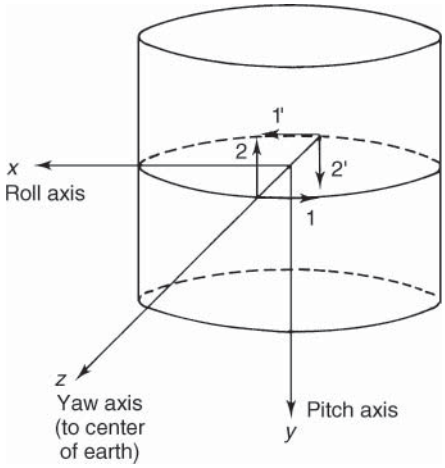


Figure 10.12 Actuator implementation on a spinning satellite: (1) spin speed control; (2) attitude control about the roll and yaw axes.

dynamic attitude control system using actuators that permit flexible and precise generation of correcting torques. Another technique is to re-establish gyroscopic rigidity by means of one or more flywheels (inertia wheels) mounted on board the satellite, thereby providing the satellite with on-board angular momentum. This technique is the most widely used for communications satellites.

10.2.7.1 Satellite with on-board angular momentum (one wheel)

The satellite (Figure 10.13) contains a momentum wheel whose axis is aligned with the pitch axis in the nominal attitude configuration. The angular momentum generated about this axis provides gyroscopic rigidity that tends to keep the mechanical axis of the satellite on which the wheel is mounted and the pitch axis coincident with the local coordinate system. Furthermore, by varying the velocity of the wheel slightly about its nominal value, correcting torques along the pitch axis are easily generated.

Attitude control in normal mode is as follows. Pitch and roll angles are measured by one or more earth sensors operating in the infrared. The gyroscopic stiffness conveyed by the momentum wheel enables movement about the roll and yaw angles to be limited. Pitch control is realised by exchanging angular momentum between the wheel and the body of the vehicle (that is, by varying the rotation rate of the wheel). Roll control is obtained by using an actuator (such as a thruster, a magnetic coil, or solar sails) that generates a torque about this axis. The gyroscopic stiffness avoids measurement of the yaw angle. Indeed, in the course of the orbit, there is an interchange between the roll axis and the yaw axis every six hours, and this permits yaw control using the measurement of roll that was available six hours earlier.

During highly disturbed phases (for example, during station-keeping operations), roll and yaw control are realised separately on each axis. Variations of yaw angle are then measured by a specific sensor such as an integrating gyrometer, sun sensor, or star sensor. The pointing accuracy obtained with a stabilisation system of this type is on the order of 0.03° for the roll angle, 0.02° for the pitch angle, and 0.3° for the yaw angle.

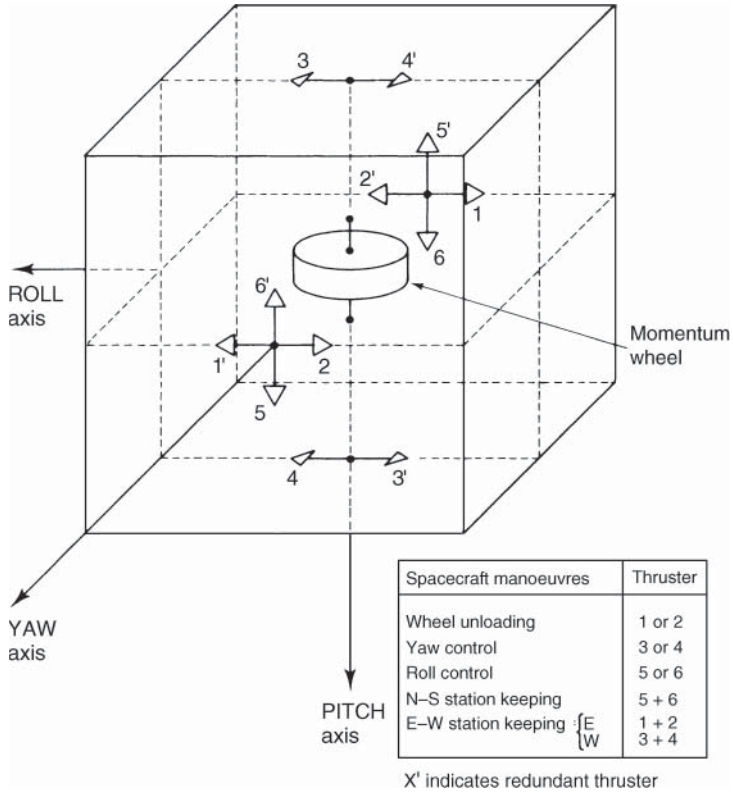


Figure 10.13 Attitude control of a body-fixed satellite by a single momentum wheel.

Pitch attitude control. Let T_d be a disturbing torque acting on the pitch axis. T_d is then aligned with the pitch axis. The moments of inertia of the satellite and the flywheel are I_S and I_W, respectively, and ω is the velocity of rotation of the wheel. If ϕ is the angle of the satellite about the pitch axis, the following can be written (angular momentum conservation):

$$I_S\ddot{\phi} + I_W\dot{\omega} = T_d \ \text{(N m s)} \tag{10.7}$$

As the satellite must rotate with a constant angular velocity of one revolution per day about the pitch axis, $\dot{\phi} = 0.25°\ \text{min}^{-1}$ and $\ddot{\phi} = 0$. Hence:

$$\dot{\omega} = T_d/I_W \ \text{(rad/s)}$$

If T_d is constant, the wheel must be accelerated by means of its drive motor. When the maximum or minimum velocity ω_M is reached, the wheel must be *unloaded*; that is, its velocity must be brought back to the nominal value by applying a torque opposite to the reaction torque. This torque is produced by thrusters (one for each correction direction) oriented perpendicularly to the pitch axis of the satellite (thrusters 1 and 2 in Figure 10.13).

If, at time $t = 0$, the wheel is rotating with nominal velocity ω_0, it will reach a velocity ω_M under the influence of torque T_d after a time t_1 such that:

$$(\omega_M - \omega_0)/t_1 = T_d/I_W, \ \text{hence}\ t_1 = \Delta H/T_d \ \text{(s)} \tag{10.8}$$

where ΔH is the difference between the angular momentum H of the wheel at its mean velocity and at its extreme velocity ω_M.

The unloading torque T_u must be equal to the torque produced by the deceleration of the wheel in order not to disturb the attitude of the satellite. Then the desaturation time t_u is such that

$$T_u t_u = T_d t_1, \text{ hence } t_u = t_1(T_d/T_u) \text{ (s)} \tag{10.9}$$

Control in roll and yaw. The control principles are the same for both axes. Stabilisation about these axes is the gyroscopic stabilisation provided by the inertia wheel. As the angular momentum of the satellite is negligible compared with that of the wheel, a disturbing torque T_d aligned with one axis (the yaw axis, for example) causes a rotation at constant velocity Ω_x about the orthogonal axis (the roll axis). If $H = I_W \omega_m$ is the angular momentum of the wheel (ω_m being the lowest value of the velocity of rotation of the wheel and therefore corresponding to the most unfavourable case), the drift velocity is $\Omega_x = T_d/H$. For a maximum pointing error ε, the time between two corrections is:

$$t_2 = \varepsilon/\Omega_x \text{ (s)} \tag{10.10}$$

The correcting torque T_c, opposed to the disturbing torque T_d, is generated by actuators. An example of pairs of thrusters (one per direction of correction) appears in Figure 10.13; thrusters 3 and 4 are for correction about the yaw axis, and thrusters 5 and 6 are for correction about the roll axis. Each thruster acts for a time t_c such that:

$$T_c t_c = T_d t_2, \text{ hence } t_c = t_2(T_d/T_c) \text{ (s)} \tag{10.11}$$

where T_c is the correcting torque exerted by the thruster.

Required mass of propellant. If F is the thrust, I_{sp} is the specific impulse, and t_c is the acting time, the mass of propellant m is given by (see Section 10.3.1)

$$m = Ft_c/gI_{sp}$$

Over a period of one year, the number of thruster operations is $365 \times 24 \times 3600/t_1$ for unloading, $365 \times 24 \times 3600/t_2$ for corrections about the yaw axis, and the same number for corrections about the roll axis. The total cumulative time of operation is:

$$t = 365 \times 24 \times 3600[(t_u/t_1) + 2(t_c/t_2)] \text{ (s)}$$

The annual mass of propellant is:

$$m = 31.5 \times 10^6 (F/gI_{sp})[(t_u/t_1) + 2(t_c/t_2)] \text{ (kg)} \tag{10.12}$$

Example 10.2 Consider a satellite with the configuration of Figure 10.13 (thrusters on the periphery). Its characteristics are as follows:

— Thrusters: thrust $F = 0.5$ N; specific impulse $I_{sp} = 290$ s; lever arm length $l = 0:75$ m
— Flywheel: nominal velocity = 7500 rpm; nominal angular momentum $H = 50$ N m s; permissible variation of angular momentum $\Delta H = \pm 5$ N m s

The disturbing torques (T_d) considered are those that appear at times other than the station-keeping corrections; they are assumed to be constant and equal about each axis: $T_d = 5 \times 10^{-6}$ N m. Attitude control must be to within 0.1°.

Pitch control. The time t_1 between two unloading operations is:

$$t_1 = \Delta H / T_d = 1 \times 10^6 \text{ s (11.6 days)}$$

The unloading torque per pair of thrusters is $T_u = 2Fl = 0.75$ N m. The unloading time t_u is:

$$t_u = t_1(T_d/T_u) = 6.7\text{s}$$

Yaw (or roll) control. The rotation velocity about the yaw axis is $\Omega_z = T_d/H = 1 \times 10^{-7}$ rad s^{-1}. The time t_2 between two corrections is $t_2 = \varepsilon/\Omega_z = 1.7 \times 10^4$ s (4.7 hours). The correcting torque per pair of thrusters is $T_c = 2Fl = 0.75$ N m. The operating time t_c is:

$$t_c = t_2(T_d/T_c) = 0.12 \text{ s}$$

Mass of propellant (10 years). For a lifetime of 10 years, the mass of propellant is:

$$m = 31.5 \times 10^6 (10F/gI_{sp})[(t_u/t_1) + 2(t_c/t_2)] = 2.3 \text{ kg}$$

A margin should be provided to take into account variations of thruster operation, hold-on propellant (unusable residual propellant in the propulsion system), and so on.

Solar sail torqueing. Drift of the orientation of the angular momentum of the inertia wheel, under the effect of disturbing torques about the pitch and roll axes, can be continuously compensated for by means of a torque generated by appropriate control of the orientation of solar generators to which flaps have been added in order to reinforce the solar sail effect (see Figure 10.8). This permits use of propulsion systems that eject mass to be limited, and hence the quantity of propellant to be loaded on to the satellite and the operational constraints associated with the use of thrusters are reduced. By way of example, the Eurostar platform is designed with this type of attitude control about the roll and yaw axes. Control about the pitch axis is still realised by control of the velocity of the inertia wheel.

10.2.7.2 Satellite with on-board angular momentum (several wheels)

With a single inertia wheel, the axis of this wheel must be coincident with the pitch axis in the nominal attitude configuration in order to maintain a fixed orientation in space during rotation of the satellite in its orbit, i.e. control with zero degrees of freedom (0 DOF). It is not, therefore, feasible to introduce a bias on one of the axes in order to change, for example, the direction of the antenna boresight, as this requires modifying the orientation of the angular momentum along the orbit that translates into propellant consumption.

Control of the orientation of the spacecraft body is facilitated by using two or three wheels whose axes are inclined with respect to the pitch axis (Figure 10.14). The direction of the resulting angular momentum depends on the relative velocities of rotation of the wheels and can thus be modified by adjusting their velocities. Depending on the number of wheels, one or two degrees of freedom (1 DOF or 2 DOF) are introduced into the attitude control. An additional wheel for redundancy can also be added. A configuration with two wheels in a V configuration (at angles of ±20°) about the satellite pitch axis in the pitch–yaw plane and one orthogonal to the pitch axis for redundancy provides an additional degree of freedom. Indeed, the satellite could be rotated at virtually no cost about the roll axis by exchanging angular momentum between the two V-mounted wheels. This allows operation of the satellite into an inclined orbit while maintaining the appropriate pointing of the antennas by rotating the satellite to compensate the north–south displacement of the coverage (the Comsat manoeuvre).

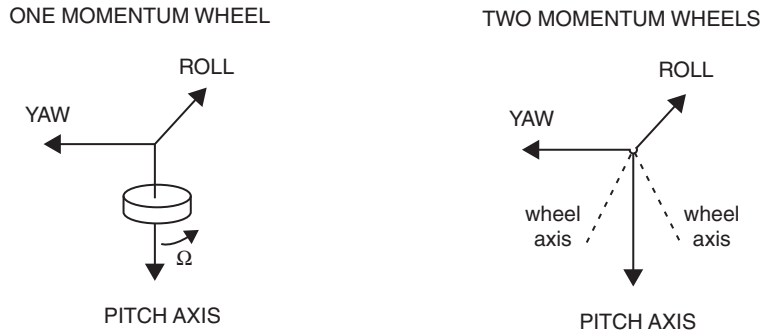


Figure 10.14 Angular momentum generated by one or several momentum wheels.

10.2.7.3 Satellite without on-board angular momentum

In spite of the constraints imposed by the high dynamics of satellite attitude variations that require continuous control, it is feasible not to take advantage of the gyroscopic rigidity provided by installation of one or more inertia wheels on board the satellite. This permits considerable freedom in respect of satellite attitude and allows the orientation to be continuously modified, for example to shift the RF coverage of the antennas or to compensate for the effects of pointing displacement caused by a non-nominal orbit (e.g. nonzero inclination).

In this context, it is useful to use actuators that are capable of generating finely modulated control torques. Three reaction wheels arranged along the three principal axes of the satellite allow disturbing torques to be compensated for by exchanging angular momentum between the body of the vehicle and each of the three wheels. This exchange is obtained by continually controlling the velocity of rotation of the wheels from attitude information using an on-board computer (OBC) (Figure 10.15). The gyroscopic rigidity introduced by the wheels is, in this case, a secondary, parasitic, effect that remains small on average; the mean velocity of rotation of the wheels is nominally close to zero.

The principle presented here for control about the pitch axis remains useful with reaction wheels on each of the three axes, but formulating complete equations for the system is much more complex as a consequence of gyroscopic coupling terms due to interaction of the movement of the satellite and the wheels. When the disturbing torque about a given axis has a nonzero mean value, compensation for the torque is accompanied by a continuous increase of the velocity of the wheel. When the maximum velocity is reached, it is necessary to unload the wheel by compensating the reaction torque generated by electrical braking by means of an external torque (generated for instance by a thruster). In Figure 10.15, it can be seen that unloading operations are performed when the angular momentum of the wheel exceeds $\pm 80\%$ of its maximum angular momentum. Unloading reduces the angular momentum of the wheel to $\pm 10\%$ of its maximum value. The example considered is that of the Japanese Broadcasting Satellite for Experimental Purposes (BSE), launched in 1978. The sixth Japanese Experimental Test Satellite (ETS-VI) and the Olympus satellite of the ESA were also of the zero-momentum type with three reaction wheels plus a redundant one.

Zero-momentum, three-axis stabilisation has rekindled interest in recent years for multimission satellites because of the full flexibility it offers in the orientation of the satellite. Examples are the medium size Star Bus spacecraft developed by Orbital Corporation and the large telecommunications platform Alphabus developed jointly by TAS and Airbus Defence and Space, formally European Aeronautic Defence and Space Company (EADS) Astrium.

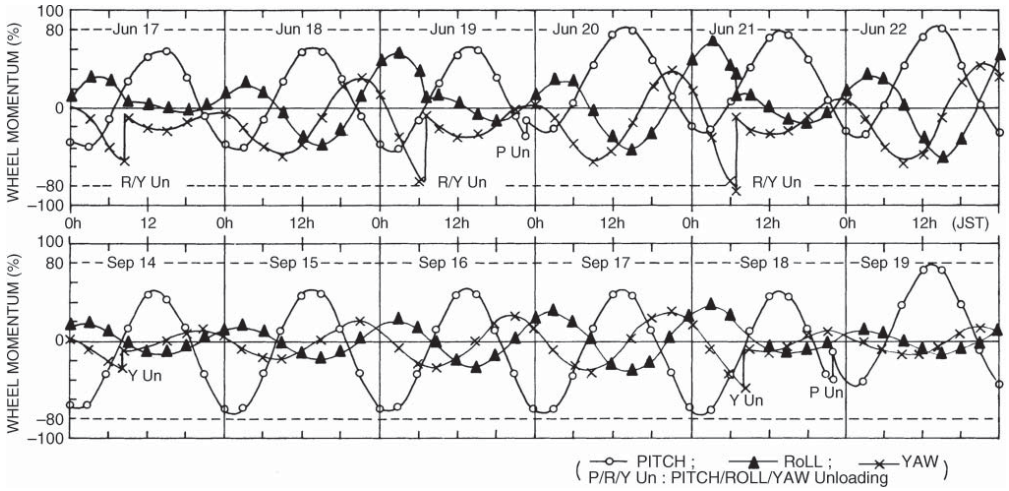


Figure 10.15 Variation of the angular momentum of the three reaction wheels (BSE satellite).

10.3 THE PROPULSION SUBSYSTEM

The role of the propulsion subsystem is mainly to generate forces that act on the centre of mass of the satellite. These forces modify the satellite orbit, either to ensure injection into a predetermined orbit or to control drift of the nominal orbit. The propulsion system also serves to produce torques to assist the attitude control system. The forces generated by the propulsion units are reaction forces resulting from the expulsion of material.

10.3.1 Characteristics of thrusters

There are two classes of thruster to be considered:

— Low-power thrusters, from a few millinewtons to a few newtons, which are used for attitude and orbit control in reaction control system (RCS).
— Medium- and high-power thrusters, from several hundreds of newtons to several tens of thousands of newtons, which are used for orbit changes during the launch phase. Depending on the type of launcher used, these thrusters form the *apogee kick motor* (AKM) or the *perigee kick motor* (PKM).

The specific characteristics of attitude and orbit control thrusters of RCS are:

— Low thrust levels (several tens of millinewtons to about ten newtons)
— A large number of operating cycles of limited duration (a few hundreds of milliseconds to a few hours)
— A cumulative operating time of several hundreds or thousands of hours
— A lifetime of greater than 15 years

10.3.1.1 Velocity increments

The law of conservation of momentum can be written:

$$M\,\mathrm{d}V = v\,\mathrm{d}M \ (\mathrm{N\,s}) \tag{10.13}$$

It expresses the fact that between time t and time $t + \mathrm{d}t$, a satellite of initial mass M moving with velocity V has lost mass $\mathrm{d}M$ and increased its velocity by $\mathrm{d}V$. The velocity of ejection of the mass $\mathrm{d}M$ with respect to the satellite is v. Integrating between time t_0 (satellite mass = $M + \mathrm{m}$) and time t_1 (satellite mass = M) gives:

$$\Delta V = v \log[(M + m)/M] \ (\mathrm{m/s}) \tag{10.14}$$

where m is the mass of material ejected and M is the mass of the satellite at the conclusion of the manoeuvre.

10.3.1.2 Specific impulse

The velocity increment obtained depends on the nature of the material ejected (the propellant) and the velocity of ejection v. The choice of propellant used is influenced by the ease of obtaining a high ejection velocity. Propellants are characterised by a parameter called the specific impulse I_{sp}.

The specific impulse is the impulse (force × time) transferred during a time dt by unit weight of propellant consumed during this time interval:

$$I_{sp} = F\,\mathrm{d}t/g\mathrm{d}M = F/[g(\mathrm{d}M/\mathrm{d}t)]\ \text{(s)} \tag{10.15}$$

where $g = 9.807\ (\mathrm{m\,s^{-2}})$ is the terrestrial gravitational constant.

The specific impulse is thus also the thrust per unit weight of propellant consumed per second. As dM/dt is the mass flow rate ρ of propellant ejected,

$$I_{sp} = F/\rho g\ \text{(s)} \tag{10.16}$$

Equation (10.13) can also be written $M\,\mathrm{d}V/\mathrm{d}t = v(\mathrm{d}M/\mathrm{d}t)$; that is, $F = v\rho$, hence:

$$I_{sp} = v/g\ (s) \tag{10.17}$$

The specific impulse is thus expressed in seconds. In certain cases, the specific impulse is defined as the impulse delivered per unit mass of propellant used and can thus be expressed in different units depending on the system used: N s/kg or lbf s/lbm (1 lbm = 0.4536 kg, 1 lbf = 4.448 N, 9.807 lbf s/lbm = 1 N s/kg). The advantage of expressing the specific impulse in seconds is that the unit is universally used (I_{sp}(s) = I_{sp}(lbfs/lbm) = (1/9.807) I_{sp} (Ns/kg)).

10.3.1.3 Mass of propellant for a given velocity increment

Combining Eqs. (10.14) and (10.17) gives:

$$\Delta V = (gI_{sp})\log[(M+m)/M] = gI_{sp}\log[M_i/M_f]\ \text{(m/s)} \tag{10.18}$$

where M_i is the initial mass and M_f is the final mass after combustion of the propellant.

The mass of propellant m necessary to provide a given ΔV to a satellite of mass M_f after combustion of propellant characterised by its specific impulse I_{sp} is obtained from:

$$m = M_f[\exp(\Delta V/gI_{sp}) - 1]\ \text{(kg)} \tag{10.19}$$

The mass of propellant m necessary to provide a given ΔV can also be expressed as a function of the initial mass M_i before combustion of the propellant:

$$m = M_i[1 - \exp(-\Delta V/gI_{sp})]\ \text{(kg)} \tag{10.20}$$

10.3.1.4 Total impulse time of operation

The total impulse I_t communicated to the system by ejection of a mass m of propellant is obtained by integrating the elementary impulse $F\,\mathrm{d}t$ over the time of operation. Assuming the specific impulse to be constant over the time of operation, this gives:

$$I_t = gmI_{sp}\ \text{(N s)} \tag{10.21}$$

The time of operation T depends on the thrust F. Assuming the mass flow rate ρ to be constant, Eqs. (10.16) and (10.17) lead to:

$$T = gmI_{sp}/F = I_t/F\ \text{(s)} \tag{10.22}$$

Table 10.2 Specific impulses for types of propulsion

Type of propulsion	I_{sp} (s)
Cold gas (nitrogen)	70
Hydrazine	220
Heated hydrazine	300
Bi-propellant	290[a]
	310[b]
Electrical	1000–10 000
Solid	290

[a]Blow down operation mode (i.e. pressurant gas is stored in the same tank as propellant and pressure decreases as propellant is consumed).
[b]Regulated pressure operation mode (i.e. a regulator maintains a constant gas pressure).

10.3.1.5 Chemical and electrical propulsion

Two classes of propulsion system exist:

— *Chemical propulsion* has a thrust level between 0.5 N and several hundreds of newtons for propulsion with liquid propellants, and from hundreds to tens of thousands of newtons for propulsion with solid propellants.
— *Electrical propulsion* can deliver a thrust on the order of up to 100 millinewtons.

The specific impulses depend on the propellant used and the type of thruster (Table 10.2).

10.3.2 Chemical propulsion

The principle of chemical propulsion consists of generating gases at high temperature by chemical combustion of liquid or solid propellants. These gases are accelerated by the nozzle.

10.3.2.1 Solid propellants

Solid propellant motors are reserved for generating velocity increments for initial injection into orbit. These motors can be used once only and develop large thrusts (from hundreds to tens of thousands of newtons). The specific impulse obtained is on the order of 295 s. A description of these motors, together with their characteristics, is given in Chapter 11.

10.3.2.2 Cold gas

Cold gas propulsion consists of releasing a gas stored under pressure in a reservoir through a nozzle. The material used, depending on its nature and the pressure, can be in a liquid state (examples are freon, propane, and ammonia) or a gaseous state (such as nitrogen) in the reservoir. These systems are characterised by relative simplicity, low thrusts, and small specific impulses (less than 100 seconds). They were used mainly on the first satellites and are still of interest whenever problems of thermal control and pollution associated with hot gas systems arise.

10.3.2.3 Mono-propellant hydrazine

A hot gas at a temperature of around 900 °C composed of ammonia, nitrogen, and hydrogen is obtained by catalytic decomposition of hydrazine, which is then released through a nozzle (Figure 10.16). The catalyst is a metal (iridium) and is designed in such a way that the contact area is as large as possible within a small volume (small spherical granules are used). The performance of the propellant depends on the temperature of the catalyst and of the hydrazine.

The thrust obtained is limited by the quantity of hydrazine, which can be decomposed in unit time (a function of the area available to the reaction) and is typically on the order of 0.5–20 N.

The specific impulse is on the order of 220 seconds and depends on the operating conditions of the propellant (such as whether it is starting from cold or hot and whether it is operating in continuous or pulse mode). In particular, operation in pulse mode has a low performance as a consequence of the relatively long time required to establish the thrust.

10.3.2.4 Bi-propellant propulsion

Bi-propellant systems use an oxidant–fuel pair that has the property of spontaneous ignition (hypergolic propellants) when they come into contact in the combustion chamber, in order to produce hot gases for release through the nozzle.

The most commonly used pair consists of nitrogen tetroxide (N_2O_4) as the oxidant and monomethylhydrazine (CH_3NHNH_2, or MMH) as the fuel. The gas produced is a mixture of water, nitrogen, carbon dioxide, carbon monoxide, and hydrogen.

The *mixture ratio* in the thruster is an important parameter on which the performance depends. It is defined as the ratio of the mass of oxidant to the mass of fuel flowing per unit time. For the mixture considered, the optimum ratio is on the order of 1.6. This value is also the ratio of the density of the two propellants. This property means the volume of the reservoirs used to store the propellants on board the satellite is the same for the two propellants; this facilitates integration of the reservoirs and limits development costs.

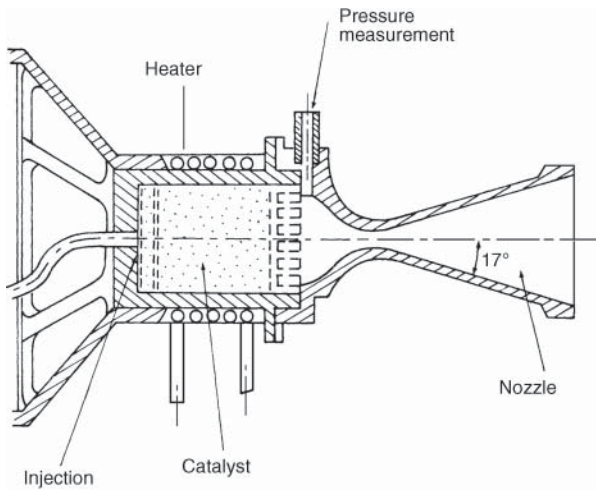


Figure 10.16 Hydrazine thruster.

Realisable thrusts range from tens of newtons for motors used for orbit control to several hundreds, even thousands, of newtons for the motors used for injection into orbit (see Section 11.1.5).

The specific impulse obtained is between 290 and 310 seconds; it depends on the mass flow rate, and hence the thrust concerned, and the supply pressure of the propellants in the combustion chamber (see Section 10.3.2.5). The useful performance obtained by virtue of the high value of specific impulse is affected by the greater dry mass of the propulsion system due to the duplication of valves, filters, reservoirs, pipework, etc. The overall balance becomes useful only for satellites with a sufficiently high mass in orbit (on the order of one ton, if only the orbit control propulsion system is considered). On the other hand, bi-propellant propulsion is particularly useful in connection with the unified propulsion concept (see Section 10.3.4.3).

10.3.2.5 Operation of liquid propellant propulsion

Liquid propellant propulsion systems contain reservoirs to store the propellants; a pressurisation system to drive the propellants from the reservoirs; pipework on which the filters, valves, pressure tappings, filling, and draining orifices, etc. are mounted; and the thrusters themselves (see Figure 10.20, Section 10.3.4.3). Precise thermal control maintains the temperature of the various internal parts within a range that is often narrow; the difference between the freezing and boiling temperatures of propellants can be small. Thermal control is ensured by use of both insulating material and electrical heaters as far as the reservoirs, pipework, and valves are concerned and by heat sinks that ensure extraction of the heat generated from the combustion chamber and the motor nozzles into the satellite structure.

The propellants are stored in one or, more often, several reservoirs arranged in such a way that the position of the centre of mass of the satellite varies as little as possible as the reservoirs become exhausted. Positioning is also influenced by consideration of the relative values of the moments of inertia about the various axes, particularly with satellites stabilised by rotation either in the transfer phase or in the normal mode. The problem of sloshing of the propellants always arises for satellites that contain a phase of attitude control by rotation; this can compromise control stability particularly when large quantities are embarked. The use of reservoirs containing energy dissipating devices permits any oscillation that would otherwise arise to be rapidly damped.

Injection of propellants into the motors under a given pressure required for correct operation is ensured by a reservoir pressurising device. The simplest system is to incompletely fill the reservoirs (by r%) in order to reserve space for a pressurising gas that forces the propellants out of the reservoir. As the reservoir becomes exhausted, the initial pressure decreases (this is described as *blow down*). The ratio of the initial pressure P_i to the final pressure P_f is equal to the ratio of the final volume V_f to the initial volume V_i available to the gas:

$$P_i/P_f = V_f/V_i = 1/(1 - r) \tag{10.23}$$

This pressure variation ratio (the *blow down ratio*), equal to the inverse of the complement of the filling coefficient r is limited to a maximum value that depends on the type of propulsion used. The limiting value is on the order of four for hydrazine mono-propellant propulsion (a maximum filling coefficient of 3/4) and two for bi-propellant systems.

To increase the filling coefficient of the reservoir, it is possible to store the pressurising gas in a separate reservoir. The gas can be stored there either with the required initial pressure, which simply represents an increase of the total volume, or with a much higher pressure; in which case it feeds the reservoir by way of a pressure-reducing value. The pressure-reducing valve provides regulation of the gas pressure on the propellants, and this ensures that the thrusters operate under conditions for obtaining the best performance. The disadvantage is that the lifetime of the

pressure regulator is limited, and it is not feasible to operate the system continuously throughout the lifetime of the satellite. Operation under constant pressure in thus reserved for the orbit injection phase, particularly with systems with unified propulsion (see Section 11.1.4.2), and possible repressurising of reservoirs in the course of the life of the satellite when, during normal operation in blow down mode, the pressure becomes too low.

Finally, the problem of separation of liquid and gas in the reservoirs arises. In fact, the pressurising gas and the propellant constitute an emulsion due to the absence of gravity on board the satellite. It must be ensured that only the liquid escapes through the duct that feeds the motors. For rotation-stabilised satellites, positioning the duct on the periphery of the satellite permits separation of the gas and liquid by means of the artificial gravity caused by the rotation of the satellite. This gravity does not exist with three-axis stabilised satellites. It is necessary to separate the liquid and gaseous phases mechanically within the reservoir. A polymer membrane can be used with propellants that are not very corrosive (hydrazine). A metallic membrane (bellows) can be used for small reservoirs.

On the other hand, the use of a membrane is not possible with corrosive propellants (nitrogen peroxide) over long lifetimes. The use of surface tension forces that exist at the interface between a liquid and a solid surface can ensure that only the liquid is present within a network of fine cavities (a sieve or metallic sponge of porous material) that feeds the duct. A bubble trap blocks possible bubbles in the pipework that may form when the satellite is subjected to large accelerations.

10.3.2.6 Location of thrusters

As far as the thrusters used for attitude and orbit control are concerned, the number of thrusters and their location are dictated by various considerations. The forces to be generated are as follows:

— *Parallel to the orbit*: Thrust is exerted in the plane of the satellite orbit and serves to control the semi-major axis and eccentricity of the orbit (to maintain longitude) (i.e. thruster 1 controlled by pulses phased with the rotation of the satellite in Figure 10.12; thrusters 1 and 2 or 1′ and 2′ used simultaneously in Figure 10.13). The thrust also serves to unload the inertia wheel for body-fixed satellites with on-board momentum wheels (1 and 1′ or 2′ and 2 used simultaneously) and to maintain the velocity of rotation of the satellite when it is spin stabilised (1 and 1′ in Figure 10.12).
— *Perpendicular to the orbit*: Thrust is exerted along the pitch axis. It serves to correct the inclination (thrusters 2 and 2′ in Figure 10.12; 5 and 6 or 5′ and 6′ used simultaneously in Figure 10.13) and to modify the orientation of the north–south axis (2 and 2′ controlled by pulses phased with the rotation of the satellite in Figure 10.12; 5 and 5′ or 6 and 6′ in Figure 10.13).

The choice of thruster position must be made in accordance with the nature of the mechanical effects to be obtained (torques about the centre of mass or forces acting on the centre of mass).

The gas jet at the output of the nozzle is characterised by its angular width. This jet must not strike parts of the satellite since this would cause problems of deviation of the jet and hence deviation of the direction of the thrust in addition to thermal problems. One approach, when the thrust to be generated must be parallel to one surface of the satellite, is to mount the nozzle with a given inclination (10–15°) with respect to the surface. The interaction between the jet and the surface is thus reduced, to the detriment of the efficiency of the thruster in the required direction and the generation of an orthogonal thrust component. Furthermore, the location of the thrusters

must take into account the problems of pollution of sensitive surrounding surfaces (such as solar cells, radiating surfaces, sensors, etc.).

A compromise between the various requirements is sought by attempting to minimise the required number of thrusters. The ease of integration (which consists, for example, of locating several mutually pre-aligned thrusters on the same plate) also arises in the criteria to be considered.

10.3.3 Electric propulsion

Electric propulsion involves the use of an electrostatic or electromagnetic field to accelerate and eject ionised material. Electric propulsion is an advanced technology in comparison with chemical propulsion. It is characterised by low thrusts (less than 0.1 N) with a high specific impulse (1000–10 000 seconds). A notable reduction can thus be achieved in the mass of propellants to be embarked in comparison with chemical technologies. On the other hand, the operating times are much greater in view of the low thrust. Electric propulsion, above all, requires a large amount of electrical power. In the specification of a system, therefore, not only the specific impulse but also the specific power must be considered; the latter is equal to the ratio of the electric power to the thrust. It is on the order of 25–50 W mN^{-1} depending on the type of thruster.

Various electric propulsion techniques have been developed; these include electrothermal propulsion (resistojet and arcjet), plasma propulsion, and ionic propulsion [HUM-95].

10.3.3.1 Resistojet propulsion of hydrazine thrusters

In order to increase the velocity of ejection and hence the specific impulse of a hydrazine thruster, the gas obtained after catalytic decomposition may be superheated to a temperature on the order of 2000 °C before release through the nozzle. The superheating is provided electrically in a heat exchanger.

The specific impulse obtained is on the order of 300 seconds, which is more than 20% greater than that of hydrazine; this leads to an equivalent reduction in the quantity of propellant to be loaded on to the satellite to ensure provision of a given velocity increment. The disadvantages lie in the high electrical consumption of each motor (several hundreds of watts), the limited thrusts obtained (0.5 N), problems with the behaviour of materials at high temperatures and, consequently, reliability.

10.3.3.2 Arcjet propulsion

A low-power arcjet [MES-93] thruster consists basically of an anode, made out of materials capable of withstanding high temperatures, such as pure tungsten or tungsten–rhenium alloy, which serves as chamber, throat, and expansion nozzle. The cathode is usually made from thoriated tungsten and has the shape of a rod with a conical tip.

The propellant gas (argon, ammonia, or catalytically decomposed hydrazine) is fed into the arc chamber and is heated by an arc discharge.

Low-power arcjet thrusters operating at input powers on the order of 1 kW can offer considerable advantages, both in terms of increased payload capability and extended lifetime, over chemical thrusters in performing station-keeping manoeuvres of medium to small geostationary satellites.

Arcjet thrusters, besides their inherent simplicity, can use hydrazine as a propellant allowing a high level of commonality with the other elements of the spacecraft propulsion subsystem.

10.3.3.3 Plasma propulsion

Pulsed plasma thrusters. The thruster is a form of capacitor using a rod of Teflon placed between two electrodes. This capacitor is fed by an electric generator and charges until the high voltage causes a spark to flash across the surface of the rod. A layer of the material is ionised, and the plasma is accelerated by the self-generated electromagnetic field [FRE-78]. Once the capacitor is discharged, it recharges until the next discharge occurs. Wear of the Teflon rod is compensated for by advancing the rod under spring pressure.

The technique is simple and does not require a neutralising device since the plasma ejected is electrically neutral. On the other hand, problems of pollution and EMC are not negligible. This type of thruster was used both experimentally and operationally on US Lincoln Experimental Satellite Series (LES-6, LES-8, LES-9) in the 1970s. The specific impulses obtained were between 1000 and 5000 seconds.

Stationary plasma thruster (SPT). This type of thruster (Figure 10.17) was developed in the 1960s by Russian scientists and engineers [MOR-93]. The central spike forms one pole of an electromagnet and is surrounded by an annular space, around which is the other pole of the electromagnet, with a radial magnetic field in-between. A hollow cathode provides a source of electrons. The electrons migrate from the cathode to the anode and are trapped by a radial magnetic field generated by the outer and inner solenoids. This orbital rotation of the electrons is a circulating Hall current. The propellant, such as xenon gas, is fed through the anode. The trapped high-energy,

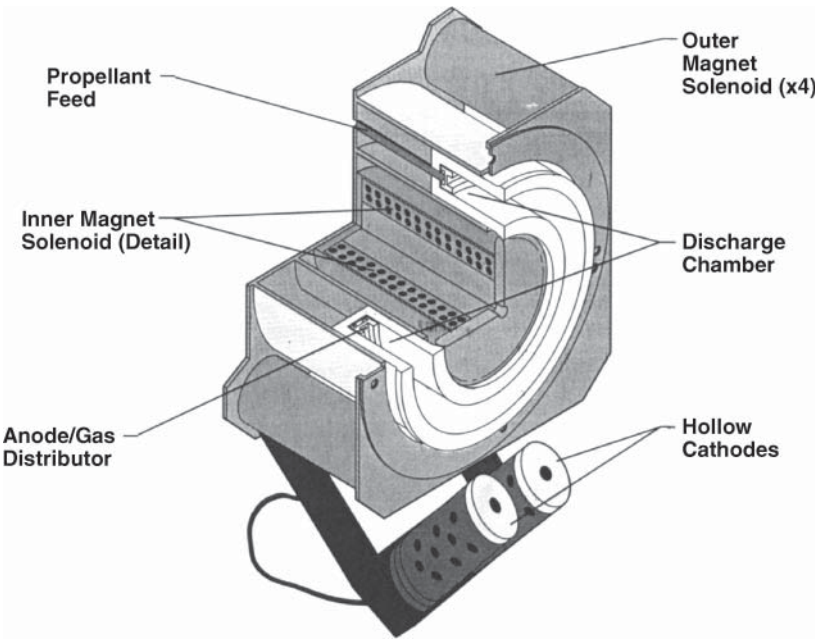


Figure 10.17 Stationary plasma thruster (SPT).

circulating electrons collide with atoms of xenon gas distributed by the anode ring. As most electrons are trapped in the Hall current, they have a long residence time inside the thruster and are able to ionise almost all (~90%) of the xenon propellant. The ions produced in the collision process are accelerated out of the discharge chamber by the electric field induced by the heterogeneous electron density, and create thrust. Upon exiting, the ions pull an equal number of electrons with them from the cathode, creating a plume with no net charge. The axial magnetic field is designed to be strong enough to substantially deflect the low-mass electrons, but not the high-mass ions, which have a much larger gyroradius and are hardly impeded. Some (30%) electrons, because of instabilities, are freed from the magnetic field; they drift towards the anode and do not produce thrust. The thruster energetic efficiency is around 63%. Another issue with SPTs is the significant beam divergence and variation in thrust direction.

Hundreds of SPTs have flown, first on Russian spacecraft and then on Western satellites after re-engineering of the motor by an international group led by Snecma of France (note: Snecma was renamed Safran Aircraft Engines in 2016) and in collaboration with Fakel of Russia. This technology is especially appealing for station-keeping, repositioning of satellites, and orbit raising (see Section 10.3.4.4).

The typical characteristics of a SPT (Snecma SPT 1350) are as follows:

— Propellant: xenon
— Thrust: 88 mN
— Specific impulse: 1650 seconds
— Electric power: 1350 W
— Mass flow rate: 5.3 mg s^{-1}
— Design total impulse: 3000000 N s
— Design cycles: 8200
— Thruster mass (with Xenon flow rate control): 5.3 kg
— Thruster dimensions: 15 cm × 22 cm × 12 : 5 cm

10.3.3.4 Ionic propulsion

In an ion thruster, charged particles (ions) are accelerated by an electric field. The ionised material is a heavy metal, which is in liquid state at the storage temperature in order to facilitate feeding to the thruster; examples are mercury, xenon, and caesium. It is necessary to neutralise the beam by ejecting the same quantity of charge of the opposite sign in order to avoid raising the satellite to an excessive potential with respect to the surrounding medium. This is achieved by means of an electron gun (a neutraliser). Various types of thruster have been developed; they differ in the technique used to obtain the ions from the metallic atoms as follows.

Prior ionisation (Figure 10.18). The electrons are extracted from the atoms in an ionisation chamber after vaporisation of the propellant by electric heating. The ions created in this way are then accelerated by a grid raised to a high negative voltage. The specific impulses obtained are on the order of 2000–3000 seconds, and thrusts are on the order of 2–20 mN for a corresponding electric power consumption of 60–600 W.

The electrons are extracted either by means of electron bombardment of a cloud of atoms by an electron gun or excitation by an induced RF field of several hundreds of kHz. Motors using electronic bombardment are also called *Kaufman motors* and examples are the UK-10 (18 mN thrust) motors developed by Marconi (now Astrium UK), 2 and 20 mN motors developed by Mitsubishi, and 18 and 25 mN Xenon Ion Propulsion System (XIPS) developed by Boeing. Motors using RF fields include the RIT-10 (15 mN thrust) developed by DASA, now Astrium in Germany. The electron-bombardment ion thruster assembly (EITA) system based on UK-10 and the RF

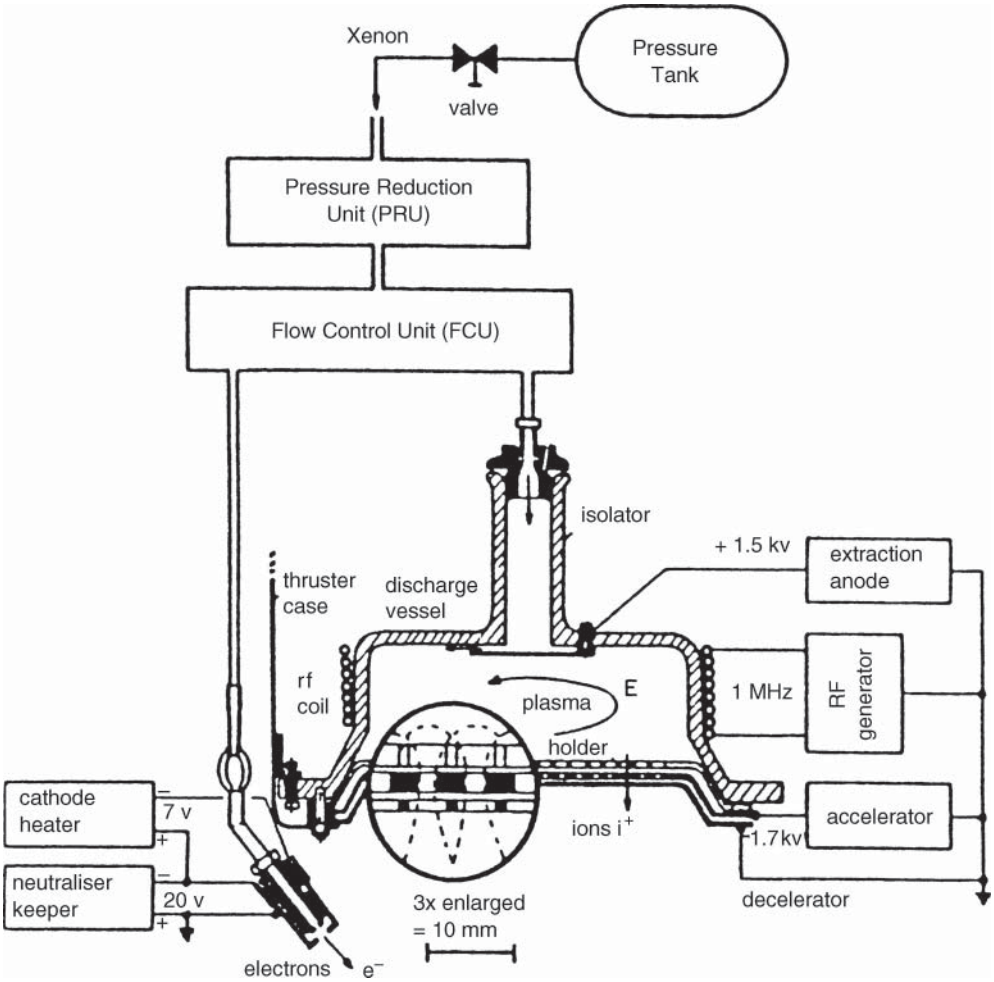


Figure 10.18 The principle of an ionic thruster.

ion thruster assembly (RITA) based on RIT-10 have been flown on the ESA Artemis spacecraft. Table 10.3 compares the characteristics of the UK-10 and RIT-10 motors.

The 25 cm XIPS thruster, manufactured by L-3 Communications, Electron Technologies, is used on the Boeing 702 communications satellites series. The 25 cm ion thruster consists of a cylindrical plasma discharge chamber, discharge hollow cathode, three-ring magnetic cusp plasma confinement, and neutraliser hollow cathode. The three-grid ion accelerator utilises domed molybdenum grids with approximately 11 000 apertures to produce the high-purveyance xenon ion beam. The 25 cm thruster is designed to operate at two power levels. The high-power mode operates at 4.5 kW of input power to produce a 1.2 kV, 3A ion beam. In this mode, the thruster produces 165 mN thrust at a specific impulse of 3500 seconds. The high-power mode is used exclusively for the orbit-insertion phase. Nearly continuous operation in the high-power mode has been achieved for 500–1000 hours. The requirements are dependent on the launch vehicle and satellite. The low-power mode, with a thruster input power of 2.2 kW, is used for station-keeping.

Table 10.3 Ion thruster characteristics

Characteristic	EITA/UK-10/T5	RITA/RIT-10
Thrust level	18 mN	15 mN
Exhaust velocity	40 869 m s^{-1}	46 800 m s^{-1}
Voltage	1100 V	1500 V
Beam current	329 mA	234 mA
Mass flow rate	0.55 mg s^{-1}	0.46 mg s^{-1}
Specific impulse	3200 s	3400 s
Total input power	476 W	459 W
Efficiency	55%	51%

Table 10.4 Typical parameters of the L-3 Communications 25 cm XIPS thruster

Parameter	Low-power station-keeping	High-power orbit raising
Active grid diameter (cm)	25	25
Average I_{sp} (seconds)	3400	3500
Thrust (mN)	79	165
Total power consumed (kW)	2.2	4.5
Mass utilisation efficiency (%)	80	82
Typical electrical efficiency (%)	87	87
Beam voltage (V)	1215	1215
Beam current (A)	1.45	3.05

In this mode, the thruster produces nominally 79 mN of thrust with an I_{sp} of 3400 seconds. Typical performance of the 25 cm thruster is summarised in Table 10.4.

Field emission (Figure 10.19). This process permits both ionisation and acceleration of the ions to be obtained. Two plates, of which one side has a bevelled section with a very sharp edge, are assembled in such a way that the extremities of the bevels coincide. The liquid metal progresses by capillarity to the extremity. A slotted electrode raised to a high potential with respect to the plates (about 10 kV) creates an intense local electric field. The field is sufficiently large to extract electrons from the propellant atoms, and the ions generated in this way are directly accelerated by the electric field. This type of motor was developed, in particular, by SEP (now Safran Aircraft Engines) under an ESA contract.

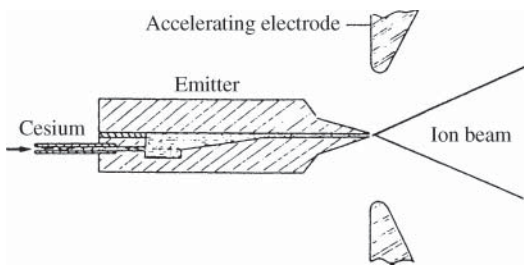


Figure 10.19 Field emission ionic thruster.

The specific impulses obtained are very high, between 8000 and 10 000, and thrusts on the order of 10 mN can be obtained by parallel operation. Electrical consumption, on the other hand, is relatively high, on the order of kW.

10.3.3.5 Implementation aspects

With electric propulsion, high specific impulses enable the mass of propellant embarked to be reduced in comparison with chemical thrusters. On the other hand, electric thrusters require additional electrical energy, which can lead to an increase in the mass of the solar generators. In total, they are advantageous only for satellites with a long lifetime (greater than seven years).

Arcjet propulsion suffers from a larger demand for power consumption and a limited lifetime compared with the increasing lifetime required by communications satellites.

With ionic and plasma propulsion, the width of the ion beam at the thruster output is large, about 40°. To avoid interaction of the jet with the surface of the satellite when the force produced must be parallel to the surface, it can be necessary to incline the thrust axis with respect to the body, and this leads to a loss of thrust efficiency. Moreover, the direction of thrust varies with time and consequently disturbing torques are generated. The problem can be resolved by mounting the motor on a movable plate, which permits the disturbing torques to be minimised after calibration in orbit. Finally, it is necessary to take into account the low thrust in determining correction strategies (see Section 10.3.4.4).

10.3.4 Organisation of the propulsion subsystem

The organisation of the propulsion subsystem varies in accordance with the types of propulsion used.

10.3.4.1 Solid apogee kick motor associated with hydrazine propulsion

This combination was widely used for communications satellites until the mid 1980s. It involves equipping the satellite with a solid propellant AKM, which is used only for injection into orbit, and a mono-propellant hydrazine propulsion system, which is used for attitude and orbit control. The system benefits from relative simplicity and remains of interest for small satellites from a mass balance point of view. There are several disadvantages:

— The very high thrust of the solid motor requires spin-attitude stabilisation during the manoeuvre. This prohibits deployment of appendages (such as antennas and solar panels) that are unsuited to support the transmitted acceleration in a transfer orbit.
— The solid motor can be ignited only once and is not duplicated.
— The velocity increment provided is not adjustable once the motor is integrated into the satellite. Possible differences between the nominal transfer orbit and the transfer orbit into which the launcher has injected the satellite cannot be compensated.
— The specific impulses are not among the highest.
— It is necessary to install two different propulsion systems.

10.3.4.2 Solid apogee kick motor associated with bi-propellant propulsion

The use of bi-propellant in place of hydrazine permits a mass gain for large satellites (above 1200 kg). The bi-propellant propulsion system is, however, more complex, and the usefulness of this combination remains limited. It is preferable to consider the unified propulsion concept.

10.3.4.3 Unified bi-propellant propulsion

All the propellants required for injection into orbit and attitude and orbit control are stored in a single set of reservoirs. The most used propellants remain monomethyl hydrazine and nitrogen peroxide.

A high-thrust apogee motor (400 N, for example) is used for injection into orbit. However, in view of the size of the velocity increment to be provided and the limited thrust, the manoeuvre requires, in general, several firings (burns) to avoid a loss of efficiency (see Section 11.1.3.5). A set of thrusters of lower thrust is used for attitude and orbit control.

The assembly is fed by one or more pairs of reservoirs that are pressurised by helium stored in a separate reservoir (Figure 10.20). Operation of the system is as follows:

— During the launch phase, the propellant reservoirs are isolated from the thrusters and the helium reservoir by closed pyrotechnic valves.
— Once in transfer orbit, these pyrotechnic valves are opened, and this permits the helium to pressurise the propellants under constant pressure by means of a regulator. A set of electric valves provides a feed to the apogee motor for the various manoeuvres; operation under constant pressure ensures a maximum specific impulse on the order of 320 seconds. When the satellite is in the final orbit, the apogee motor is totally isolated from the rest of the subsystem by a pyrotechnic valve. The helium reservoir is also isolated from the propellant reservoirs by closure of valves under electrical control. The supply to the attitude and orbit control thrusters is then realised in blow down mode, but the pressure variation remains small since a large proportion of the propellants is consumed during the apogee manoeuvres. The specific impulse obtained is on the order of 290 seconds.

There are many advantages:

— Dividing the apogee manoeuvre into several burns permits accurate calibration and control of injection.
— When the transfer orbit is nominal, the propellants normally provided to allow for deviation from this orbit and not used by the apogee motor are available for attitude and orbit control, and this enables the expected lifetime to be increased.
— Integration is facilitated by the presence of a single system.

The specific impulse is greater than that of other usable propellants, but the dry mass of the propulsion system is greater. However, the benefit remains to the advantage of unified propulsion, even for small satellites. For example, a satellite of 1240 kg in transfer orbit supports an additional 55 kg of payload with a unified system in comparison with a conventional solid apogee motor plus hydrazine (Table 10.5) [MOS-84]. The benefits increase for large satellites.

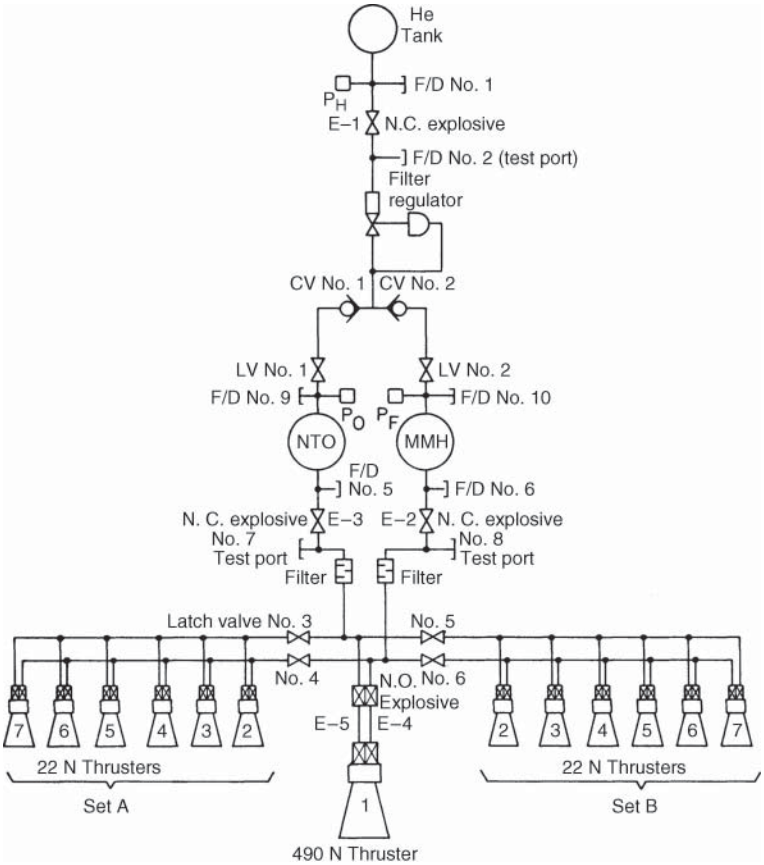


Figure 10.20 Unified bi-propellant propulsion.

10.3.4.4 Unified bi-propellant propulsion combined with electric propulsion

It is useful to consider the use of electric propulsion to provide north–south orbit control of the satellite. For a 10-year lifetime, this control requires around 47 m s^{-1} for a geostationary satellite, which represents more than 90% of the required velocity increments. Use of a high-specific-impulse propellant thus permits a large reduction of the global mass of the propulsion subsystem for missions with long lifetimes (greater than 10 years).

The low thrust of electric propulsion requires special procedures. Only correction of the long-term component of drift of the inclination of the orbital plane is generally considered (see Section 2.3.4.5).

A typical strategy is as follows: in order to compensate for the permanent drift in orbit inclination, the manoeuvre is performed daily (except during eclipse periods, discussed shortly). Because of the low thrust, a few hours (two to four) are necessary each time.

The electric power required for operation of the electric thruster is readily available at the beginning of life (BOL) without the need to overdimension the solar generators. The efficiency of solar cells decreases with time spent in orbit as a consequence of degradation due to high-energy

Table 10.5 Unified bi-propellant versus solid AKM and mono-propellant AOCS mass comparison (1240 kg spacecraft in transfer orbit)

	Bi-propellant system (kg)	Solid AKM and mono-propellant	Advantage (kg)
Expendables:			
Apogee manoeuvre (310 versus 285 I_{sp})	538	573	35
On-orbit needs (288 versus 220 I_{sp})	110	144	34
End-of-life mass:	592	523	69
Propulsion inerts:			
Helium	1.6	0.2	−1.4
AKM burn-out	–	34	34
Propulsion system	60	20	−40
Residuals	10	1.4	−8.6
Net spacecraft mass:	520.4	467.4	53

radiation (see Section 10.4). The generator is thus dimensioned to provide the nominal power at end of life (EOL), and excess power (30%) is available at the BOL.

When fighting against the drift in inclination with daily manoeuvres, one overcompensates the inclination drift so as to reach the maximum acceptable value of inclination just before the eclipse season starts (in the opposite direction to the one the natural drift leads to). Then the inclination evolves freely during the equinox season until the beginning of a new period of corrections (a total of 275 manoeuvres per year). Therefore north–south correction manoeuvres during the two periods of 45 days of eclipse around the equinoxes are avoided. Hence the part of the solar generator that is used to recharge the batteries during the eclipse periods (used only to provide trickle charge outside these periods) is available to provide the power required for operation of the electric thrusters. The use of electric propulsion, therefore, has a limited influence on the dimensioning and mass of the electrical supply system.

The problem of uncertainty of the direction of the thrust vector of electric motors can be resolved by mounting the motor on a platform that permits orientation of the thrust direction. Calibration, however, requires accurate measurement of attitude variation during manoeuvres, particularly about the yaw axis. In view of the length (several hours) of the manoeuvres, use of a conventional integrating gyrometer is not possible due to the drift. The use of a solar sensor is one possible solution.

10.3.5 Electric propulsion for station-keeping and orbit transfer

The previous section underlined the benefits of electric propulsion for station-keeping, in connection with the use of chemical propulsion for geostationary transfer orbit (GTO) to geostationary earth orbit (GEO) orbit transfer. The high specific impulse of electric propulsion compared to chemical propulsion makes it appealing also for orbit transfer. However, its much lower thrust imposes long thrust durations. Different thrust strategies have been investigated in view of the spacecraft operation constraints. The smallest overall duration of the orbit transfer is obtained by continuous thrust, but this does not minimise the propellant consumption.

Further launch mass saving can also be obtained from considering an all-electric propulsion satellite. For a 15-year lifetime, the all-chemical propulsion satellite is as much as twice as heavy

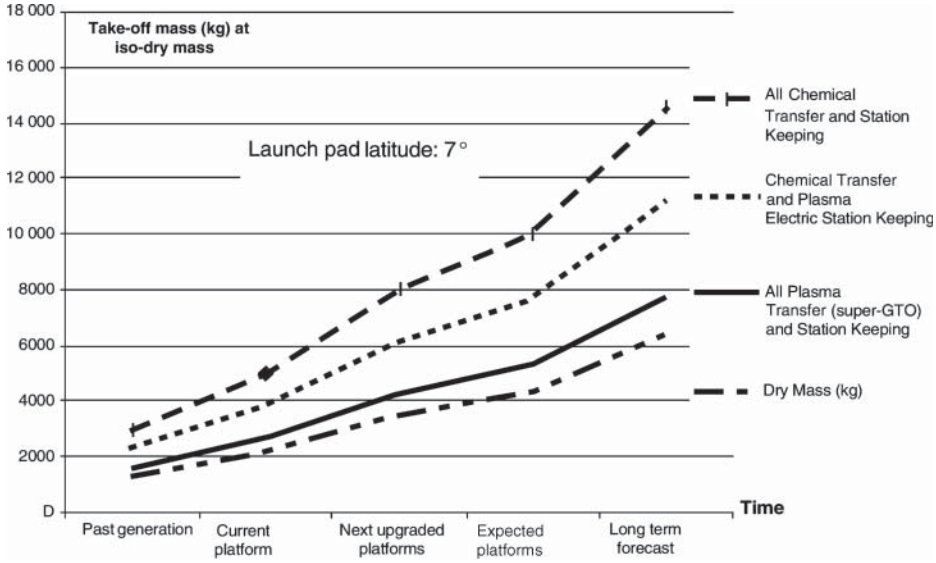


Figure 10.21 Mass at launch depending on spacecraft propulsion technology.

as an all-electrical propulsion satellite. Figure 10.21 shows the mass at launch of a satellite with respect to the expected evolution of platform technology depending on the type of propulsion. The bottom curve shows the expected dry mass evolution with time and the other curves relate to the different combinations of propulsion types. Figure 10.21 refers to a transfer orbit inclination of 7° (Ariane launch). The higher the inclination, the larger the difference between electrical and chemical propulsion.

The use of electric propulsion for orbit raising was pioneered by Boeing on the 702-Spacecraft (Boeing Space Systems BSS-702 Bus) using the 25 cm (XIPS-25) first launched on 22 December 1999 (Galaxy XI) and launched on 1 June 2017 (ViaSat-2). It includes two fully redundant subsystems with two ion thrusters and a power processor each. After launch, the thrusters are first used in a high-power orbit-insertion mode that requires nearly continuous operation of two of the thrusters for times of 500–1000 hours, depending on the launch vehicle and satellite weight. This mode utilises about 4.5 kW of bus power to produce 165 mN thrust at a specific impulse of about 3500 seconds. Once orbit insertion is completed, each of the four thrusters is fired once daily for an average of about 45 minutes in a low-power, 2.3 kW mode for station-keeping. In this mode, the beam voltage is kept the same, and the discharge current and gas flow are reduced to generate a 79 mN thrust at a specific impulse of 3400 seconds. The ion thrusters are also used for any optional station change strategies and are ultimately used for deorbiting at the EOL satellite's lifetime.

10.4 THE ELECTRIC POWER SUPPLY

In view of limitations in mass and volume, the electric power supply of a satellite poses one of the most restricting problems. The increase of effective isotropic radiated power (EIRP) necessary for the use of small earth stations means the electric power required is over 10 kW for communications satellites, particularly those intended for broadcasting of television or for mobile and

personal communications. The electric power to be provided is directly related to the RF power of the amplifiers in the payload as a function of their efficiency.

The electric power supply subsystem consists of:

— A *primary source* of energy that converts energy available in another form into electrical energy (for civil applications, it consists of a solar generator)
— A *secondary source* of energy (such as a battery of electrochemical accumulators) that is substituted for the primary energy source when this cannot fulfil its function, for example in an eclipse period
— *Conditioning* (regulation and distribution) and *protection* circuits

10.4.1 Primary energy sources

The only external source is solar radiation. On-board sources of energy (nuclear piles or combustible materials) are not, at present, technologically satisfactory for performing the mission of a geostationary communications satellite. However, during the first minutes following injection into the transfer orbit phase, on-board electrochemical accumulators, which subsequently constitute the secondary energy source, play the role of a primary energy source.

10.4.1.1 Characteristics of solar radiation

The characteristics of solar radiation are presented in Section 12.3. The normalised solar flux at a distance of 1 AU is 1353 W m^{-2}. However, the value resulting from in-space measurements is on the order of 1370 W m^{-2}. The solar flux captured by a surface perpendicular to the plane of the equator evolves in the course of the year as a function of variation of the earth–sun distance and the declination of the sun with respect to the equatorial plane (see Figure 12.3).

The sun can be considered to be a black body at 6000 K, and spectral radiation is maximum around 0.5 µm; 90% of the power radiated is concentrated between 0.3 and 2 : 5 mm (see Figure 12.2).

10.4.1.2 Solar cells

Solar cells operate according to the principle of the photovoltaic effect (the appearance of a voltage at the connections to a *p*–*n* junction subjected to a photon flux).

Current–voltage characteristic. An example of the current I_c versus voltage V_c characteristic as a function of the load due to the circuit fed by the cell is represented in Figure 10.22 for a 2 cm by 2 cm silicon cell. The incident solar flux is assumed to be normal to the surface and equal to the normalised value (1353 W m^{-2}). It is of course necessary to take into account the angle between the normal to the surface and the direction of the sun; the flux actually captured is a function of the cosine of this angle (for angles which are not too large, i.e. less than 45°).

Maximum power is obtained when the product $V_c I_c$ is maximum: that is, in the region of the 'knee' in Figure 10.22. The maximum power and open-circuit voltage depend on temperature (the open circuit voltage falls by 50% if the temperature rises from 27 to 150 °C).

Conversion efficiency. The BOL typical efficiency is about 15% at the point of maximum power of a conventional silicon cell subjected to solar radiation above the atmosphere at a temperature of 27 °C. The efficiency decreases under the effect of radiation; a decrease of 30% in 10 years is typical for a satellite in geostationary orbit. The magnitude of the degradation depends on the type of orbit concerned, the mean solar activity during the period concerned, and the occurrence

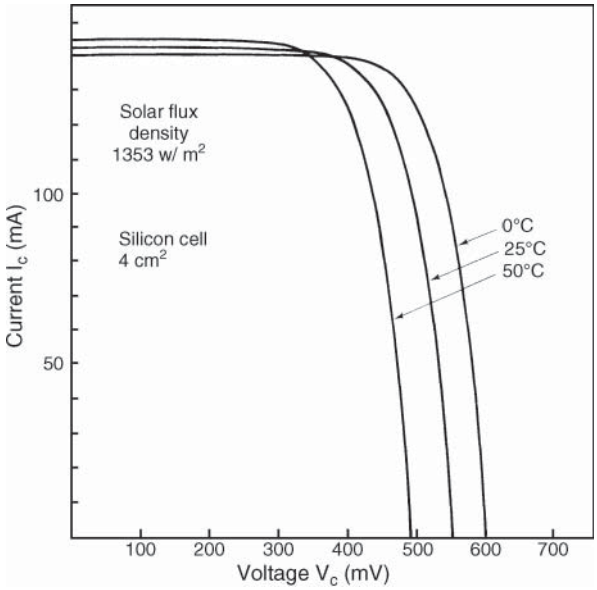


Figure 10.22 Typical current–voltage characteristic of a silicon solar cell.

of solar flares (see Chapter 12). Dimensioning of the solar generator must allow for degradation of the initial efficiency over the anticipated lifetime.

In order to limit degradation, the cell is protected by a cover that is transparent to the longer wavelengths for which the sensitivity of the cell is greatest but capable of attenuating the damaging part of the radiation. This cover glass is realised in quartz or fused silica.

Technology. Silicon solar cells have been used for many years. Constant progress has enabled the efficiency of the cells to be increased and the mass decreased.

Silicon cells are realised in a relatively thick monocrystalline chip of thickness 50–200 μm; the thinner, the lighter (20–60 mg cm^{-2}). The increase of efficiency in the course of time (from less than 10% in the 1960s to around 18% at present) has been obtained by anti-reflecting surface treatment in order to favour penetration of solar light, the use of a reflecting deposit on the back face in order to make photons that have not given up their energy during initial passage through the cell pass through it again (back surface reflector [BSR]), and with extremely thin absorber (ETA) silicon solar cells.

The use of gallium arsenide (GaAs) enabled higher values of efficiency, about 20%, to be obtained, but the difficulty in fabricating GaAs cells that are competitive in cost to silicon prevented large-scale application in satellites. This has changed with the development of techniques for growing and doping layers of GaAs that have been epitaxially grown on germanium (Ge) substrate. Since the GaAs/Ge cells are more resistant to the damage caused by high-energy particles from the sun than are silicon cells, solar panels do not require as many additional cells to meet EOL power requirements. The success of the epitaxial GaAs over Ge process has led to the extension of this process to the design and fabrication of multi-junction, or cascade, cells. These cells are composed of several layers of III-V compound materials, such as GaAs, GaInP, GaInAsP, and GaSb grown epitaxially on Ge. Multi-junction solar cells are composed of several junctions in series; each junction is defined in a different semiconductor layer and possesses a different band gap, and is therefore tuned to a different wavelength segment of the

solar spectrum. The materials are arranged such that the band gaps of the junctions become progressively narrower from the top junction to the bottom junction. Thus high-energy photons are absorbed in the top junction, generating electron–hole pairs, and less energetic photons pass through to the lower junctions where they are absorbed and generate additional electron–hole pairs. The current generated in the junctions is then collected at ohmic contacts formed at the top and bottom of the solar cell.

As an example, triple junction (TJ) solar cells with n–on–p polarity built on a 140-μm uniform thickness Ge substrate offers up to 29.9% efficiency with a solar cell mass of 84 mg cm^{-2}. On top of the greater efficiency, the cells show an excellent radiation resistance with the ratio of remaining power to initial power P/Po = 0.89 further to exposure of 5×10^{14} e cm^{-2}, 1 MeV energy electron fluence. This translates into limited overdimensioning of the solar generator at the BOL.

The use of nanomaterials may allow a significant increase in efficiency. Theoretical studies show that cells with double intermediate electronic bands created by layers of nanometre-sized semiconductor crystals (quantum dots) inserted into the i region of an ordinary p–i–n junction solar cell offer a conversion efficiency of 71%. Alternative technologies, such as thin film cells, exist. Thin-film cells have low efficiency (about 10%) but are produced inexpensively today for terrestrial applications with a mass specific power on the order of five times greater than crystalline technology. Finally, quantum dots and other nanomaterials have also recently been shown to provide dramatic improvement in the performance of thin-film photovoltaics and hybrid inorganic/organic conductive polymer-based solar cells. In the future, the use of nanomaterials will permit the development of viable thin-film solar arrays for space and, ultimately, the production of these arrays out of lightweight, flexible, polymer-based materials.

Table 10.6 shows typical characteristics of silicon, single-junction GaAs/Ge, multi-junctions (various types of double and triple junctions), and thin-film solar cells.

Table 10.6 Typical characteristics of solar cell technologies

Cell type	Efficiency, BOL 28 °C		Efficiency, EOL 1E15, 60 °C		Cell weight (kg m^{-2})
	%	KW m^{-2}	%	KW m^{-2}	
Si (200 μm)	12.6	0.170	8.7	0.118	0.464
Si (67 μm)	15.0	0.203	9.2	0.124	0.156
Si (100 μm) with diode	17.3	0.234	12.5	0.169	0.230
GaAs/Ge (137 μm)	19.6	0.265	14.7	0.199	0.720
DJ cascade (137 μm)	21.8	0.295	18.1	0.245	0.720
TJ standard (140 μm)	26.0	0.352	21.0	0.284	0.840
TJ improved (140 μm)	29.9	0.393	25.1	0.340	0.840
Thin film	12.6	0.170	9.5	0.128	0.100

BOL, beginning of life; EOL, end of life (for 1E15 1 MeV equivalent electron fluence); DJ, double junctions; TJ, triple junctions; solar flux, 135.3 mW cm^{-2}.

10.4.1.3 Solar generator (solar panels)

The solar generator consists of several thousands of cells interconnected in order to deliver the power *P* required. They are bonded to panels that provide the necessary rigidity and thermal regulation. The filling efficiency *f*, which characterises the ratio of the area occupied by the cells to the total area of the panel, is on the order of 90%.

Interconnection of cells. The cells are connected in series and in parallel in order to deliver a voltage V of several tens of volts (up to 100 V) and a current I of several tens of amps. The voltage V to be delivered determines the number of cells to be connected in series; if V_c is the cell voltage corresponding to the chosen operating point (on the order of 0.5 V for silicon, 1 V for GaAs, 2.4 V for triple junctions), the number of cells in series is equal to V/V_c.

The number of branches in parallel depends on the current $I = P/V$ to be delivered; if I_c is the current corresponding to the chosen operating point (for example, on the order of 0.15A for a cell of 4 cm^2), the number of branches to be connected in parallel is equal to I/I_c.

This basic organisation is modified in order to minimise the consequences of cell breakdown and the effect of shadow (due to the satellite body or antennas on the solar panel). An open-circuit breakdown of one cell in a branch leads to loss of the whole branch since electrical continuity is no longer provided. This can be avoided by connecting groups of cells in parallel. In contrast, a cell in one branch becoming short-circuited means the electromotive force of this branch becomes less than that of all the others. The current distribution is, therefore, unbalanced with a danger of insulation breakdown due to excessive local thermal dissipation. A diode in series with each branch of cells enables the defective branch to be isolated. The generator thus consists of small groups of cells arranged in parallel and in series; the choice of a series–parallel combination is such that it maximises the overall reliability, taking into account the relative failure rates associated with cell failure in a short- or open-circuit state (see Chapter 13).

A non-illuminated cell in a branch behaves as a load for the other cells. The current that passes through it can involve an excessive thermal dissipation, which leads to insulation breakdown. Protection is provided by placing parallel diodes on one or more cells along the branch. Figure 10.23 shows the principles of solar cell arrangement.

Dimensioning. The power P_c delivered by a solar cell (of area s) is expressed by:

$$P_c = \phi\, es(1 - l) \text{ (W)} \tag{10.24}$$

where ϕ is the solar flux captured by the cell (W m^{-2}), e is the efficiency of the cell (e.g. 17% at BOL for a Si cell), s is the area of the cell (m^2), and l are losses (%) due to cover, cabling, etc. (a typical value is 10–15%).

The solar flux captured by the cell depends on the illumination conditions and is obtained from the nominal solar flux $W = 1370$ W m^{-2}, the distance d to the sun, and the angle θ between the normal to the cell and the direction of the sun:

$$\phi = W(a^2/d^2)\cos\theta \text{ (W/m}^2\text{)} \tag{10.25}$$

where a is the mean sun–earth distance = 1 AU. Variations of the ratio a^2/d^2 as a function of the date in the year are given in Section 12.3.1. The efficiency of the cell depends on the degradation caused by high-energy radiation. The cell manufacturer provides values of efficiency for various values of 1 MeV equivalent electron fluence (electrons cm^{-2}). The actual efficiency can be determined from the estimated dose accumulated during the time spent in orbit. In the absence of precise data, the degradation of efficiency can be modelled to a first approximation by an exponential law for a satellite in geostationary orbit. For example, the following is suggested for silicon cells:

$$e_{EOL} = e_{BOL}[\exp(-0.043T)] \tag{10.26}$$

where T is the time spent in orbit in years.

The surface area A of the solar panel required to generate a given power P is given by:

$$A = (P/P_c)s/f = ns/f \text{ (m}^2\text{)} \tag{10.27}$$

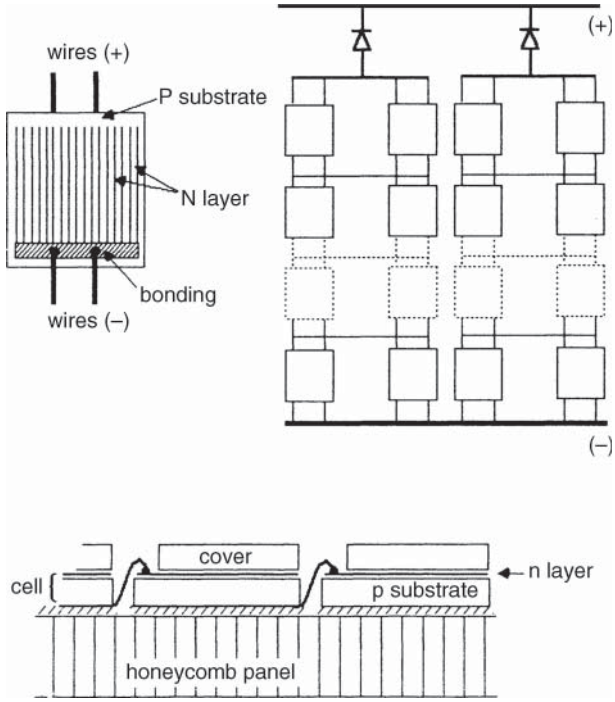


Figure 10.23 Arrangement of a solar generator.

where P_c is the power delivered per cell (this depends on the illumination conditions), n is the number of cells required, and f is the filling efficiency (85–95%), and s is the area of Si cells in m^2.

The power nP_c delivered by the solar generator varies with time. The requirements of the satellite also vary with time. Dimensioning of the solar generator must, therefore, be performed for the worst-case conditions.

In the case of a geostationary satellite, the power delivered by the generator is lowest at the summer solstice. At the equinoxes, the power is greater but also serves to recharge the battery since the satellite enters eclipse periods. It is necessary to check that the dimensioning provides the required power at all times.

Spin-stabilised satellites. On spin-stabilised satellites, the solar panels form the exterior envelope of the body of the satellite. Additional cylindrical panels can be deployed after launching to increase the useful surface area (for example, the Boeing HS 376 spacecraft, Figure 10.11). HS376 was introduced as a communication satellite bus by Hughes Space and Communications Company in 1978, then become Boeing BSS-376 on 12 April 1985, and the last launch was on 27 September 2003.

The number of cells required is large since they are not all illuminated by the sun at the same time. By considering a cylinder with its axis normal to the direction of radiation, the variation of incidence between cells located on the illuminated side leads to a surface area that must be $\pi/2$ times greater than that of a plane-perpendicular surface for the same power. Including the surface in shadow, the number of cells to be installed is thus π times greater than if the panels were flat.

In practice, the different operating conditions of the cells limit this factor to a value of between 2 and 2.5. In the course of rotation, following passage into shadow, the mean operating

temperature of the cells is lower and the efficiency is higher. Furthermore, degradation due to solar radiation is appreciably less.

Three-axis stabilised satellites. Various types of solar panel can be used with three-axis stabilised satellites:

— Flexible panels, which are rolled up in a storage container during launching and unrolled in orbit by a deployable mast.
— Semi-rigid hinged panels, which are folded in concertina fashion in a storage container during launching and extended in orbit by a deployable mast.
— Rigid panels of large dimensions (e.g. 3.9×2.3 m, conditioned by the size of the launch vehicle fairing), joined in groups of three or five to constitute a solar generator wing by a hinge arrangement that permits folding for launching. Deployment in orbit is achieved by means of a set of springs, cables, pulleys, and velocity regulators in order to ensure coordinated movement of the panels without shocks.

Once deployed, the solar generator wings are rotated in order to follow the apparent movement of the sun about the satellite. For a geostationary satellite, the generator wings are aligned with the pitch axis and rotation occurs daily.

Operation of an orientation device requires:

— Solar sensors with electronic measurement and control circuits
— A drive motor with sliding contacts to transfer the current to the satellite (called a bearing and power transfer assembly [BAPTA] or solar array drive electronics [SADE]; Figure 10.24).

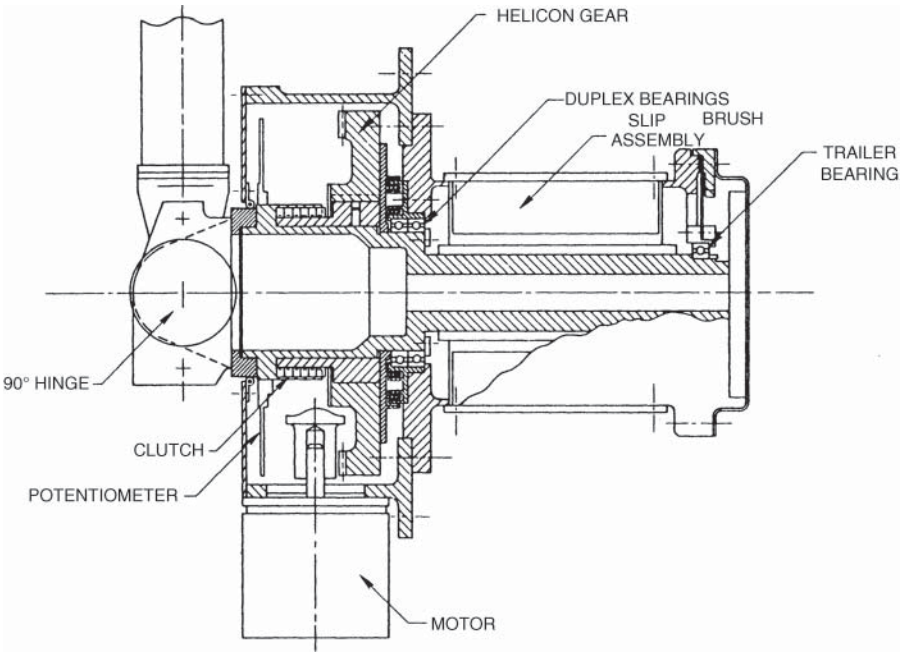


Figure 10.24 Solar generator bearing and power transfer assembly.

Specific performance. Assuming best use of the cells, the mass balance of flat solar panels compared with panels mounted on the body of a spin-stabilised satellite remains in favour of flat solar arrays. By way of indication, the panels of solar cells mounted on the body of a spin-stabilised satellite enable 30–35 W m^{-2} and 8–12 W kg^{-1} to be obtained, and the specific mass is on the order of 3–5 kg m^{-2}. For flat solar panels, the performance is on the order of 200 W m^{-2} and 40 W kg^{-1} for silicon, and 370 W m^{-2} and 50 W kg^{-1} for GaAs.

An increase in the efficiency of solar panels along with reduction of the weight can be achieved by means of *light concentrators*: Fresnel optics mounted onto the substrate concentrate the incoming sunlight onto the solar cells. Associated with multifunction cells, the concentrator solar array can achieve up to 40% efficiency. Another option makes use of reflective panels mounted at an angle on each side of the solar panel to increase its lighting. This solution has been considered by Boeing for the 702 satellite series. It is based on the Boeing Space Systems (BSS-702) bus that the first launch is on 22 December 1999 for Galaxy XI and last launched on 1 June 2017 for ViaSat-2.

10.4.2 Secondary energy sources

The secondary energy source stores energy from the primary energy source when it is operational and returns the stored energy when the primary energy source ceases to operate. Electrochemical batteries are the most appropriate means of doing this. They play a particularly important role in the case of communications satellites for which operation during an eclipse is imposed by the availability objectives usually specified. Recall that, for a geostationary satellite, eclipses occur on 90 days per year and have a duration that can be as long as 70 minutes (see Chapter 2).

The following qualities are sought:

— Adequate lifetime, which depends on the depth of discharge (DOD) and the temperature
— High specific energy in terms of Wh/kg

10.4.2.1 Battery cell parameters

The characteristic parameters of a battery cell are as follows:

— *Capacity C (Ah)*: Product of the current drawn and the time of use
— *Specific energy (Wh kg^{-1})*: The energy stored per unit mass
— *Mean discharge voltage V_d (V)*: Dependent on the intensity of the discharge current
— *DOD*: The percentage of the stored energy that is effectively used at the end of the longest period of use without recharging
— *Charge efficiency η_{ch}*: The ratio of the energy stored to the energy consumed for recharging
— *Discharge efficiency η_d*: The ratio of recovered energy to that part of the stored energy that has been used

Figure 10.25 displays a graph of the typical variation of the number of cycles as a function of the DOD.

The DOD is a parameter to be defined by the user. The choice of the DOD is dictated by the expected lifetime of the battery or, more precisely, by the number of charge and discharge cycles to be obtained. As the DOD decreases, the number of charge and discharge cycles that the battery can support increases.

A higher specific energy permits the use of a lighter battery for the same electrical power delivered during an eclipse. The larger DOD permitted with a NiH_2 or Li-ion battery than with a

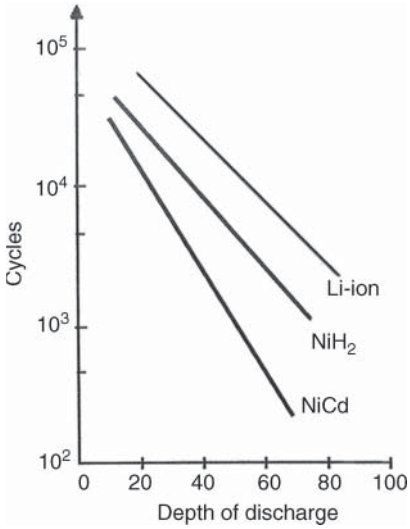


Figure 10.25 Typical variation of the number of cycles as a function of the depth of discharge (at 25 °C).

NiCd battery for the same number of cycles (Figure 10.25) also contributes to the mass reduction. For example, for a mission in geostationary orbit, the maximum DOD for a NiCd battery is 50%; while for a NiH_2 or Li-ion battery the DOD can reach 80%.

10.4.2.2 Dimensioning

Dimensioning the battery in terms of the energy to be provided takes into account the capacity of cells available from manufacturers. The energy E_c recovered from a cell of capacity C as a function of the parameters defined earlier is given by:

$$E_c = CV_d \text{DOD}\eta_d \quad \text{(W h)} \tag{10.28}$$

The battery consists of n cells in series, and the energy recovered is thus equal to nE_c. The number of cells n in series is chosen in such a way that nV_d is just greater than the voltage V to be obtained during discharge:

$$n = \text{integer} \geq V/V_d \tag{10.29}$$

One cell is often added to provide the proper voltage even in the case of short-circuit breakdown of a battery cell.

Let P be the power to be provided for the duration of an eclipse T_{ecl} (hour). The energy that the battery must have delivered after time T_{ecl} is thus given by:

$$E = PT_{ecl} \quad \text{(Wh)}$$

The capacity C of a battery element under these conditions is given by:

$$C = PT_{ecl}/nV_d \text{DOD}\eta_d \quad \text{(A h)} \tag{10.30}$$

The calculation rarely leads to a capacity value corresponding to that of a cell available from manufacturers. It is, therefore, necessary to choose cells of a capacity slightly greater

than that required ($E_c > E$). This leads to a battery that is overdimensioned with respect to the requirements, thereby imposing a mass penalty. If the voltage to be delivered is not imposed, for instance, when a regulated bus is considered (see Section 10.4.3.3), the number of cells n can be altered to provide the nominal power for an optimum battery mass. The energy to be delivered during the eclipse may be split between several batteries (typically two) to ensure some form of redundancy and facilitate integration in the satellite (see Section 10.4.3.6).

10.4.2.3 Technologies

Nickel–cadmium (NiCd) cells have been used since the advent of communications satellites for the storage of electrical energy. NiCd cells have an anode (negative) in cadmium hydroxide and a cathode (positive) in nickel hydroxide, immersed in an alkaline solution (electrolyte) comprising potassium, sodium, and lithium hydroxides. The cells deliver a voltage of 1.2 V during discharge. NiCd batteries have a legendary reputation for robustness, reliability, and service life.

Used for the first time on the NTS 2 satellite in 1974, nickel–hydrogen (NiH_2) cells have replaced NiCd as a result of their higher specific energy and greater lifetime. NiH_2 cells are based on hydrogen gas acting on a carbon electrode (using a design derived from fuel cell technology) plus a nickel hydroxide cathode (positive). The electrolyte uses potassium hydroxide, and the separator is in a zirconium ceramic material. NiH_2 cells deliver a voltage of 1.2 V. The cells are ovoid-shaped with an inconel casing. The internal pressure can reach 70 bars at the end of charge. NiH_2 cells benefit from the lightness of hydrogen and give 50% more energy per unit mass than a NiCd equivalent.

The NiH_2 battery system offers significant advantages, both at the battery level and at the spacecraft power system level, over NiCd batteries. They have an excellent cycling capability. NiH_2 batteries offer superior abuse tolerance to both overcharge and overdischarge due to the unique electrochemistry inherent in the system. This simplifies on-orbit battery operation and reduces on-board spacecraft power system electronics for battery charge control.

Lithium-ion (Li-ion) batteries allow battery weight reduction up to 50% compared to the other existing space qualified technologies. Li-ion electrochemistry involves the use of lithium insertion compounds. In a Li-ion cell, the negative electrode (anode) is graphite and the positive electrode (cathode) is a lithium-bearing metal compound. Li-ion cells have an exceptional cycling aptitude owing to the stable electrode structure: charging and discharging involve exchange of lithium ions between the electrodes via the electrolyte. Because of the high output voltage (up to 4.2 V), a non-aqueous electrolyte is used, mainly comprising a mixture of organic carbonates. Various active materials can be used for the positive electrode: lithium cobalt oxide, lithium nickel oxide, lithium aluminium oxide, lithium manganese oxide, or lithium iron phosphate. Figure 10.26 shows Li-ion cells and battery packaging. Li-ion cells offer several advantages:

— High specific energy (up to 175 Wh kg^{-1}), i.e. more than twice that of NiH_2 (60 Wh kg^{-1})
— Smaller and lighter than conventional space battery technology (50% weight reduction compared to NiH_2)
— Low thermal power and high efficiency, thus allowing a reduced size of radiators and solar panels
— High charge retention, which translates into less launch-pad workload
— Open circuit voltage stability versus cell temperature
— No memory effect
— State of charge directly related to voltage, providing an effective energy gauge
— Modular approach providing flexibility for battery system design

Li-ion cells allow parallel cell connections without any protecting device. Cells adapt to a common voltage or state of charge even though their capacities may differ. However the specific characteristics of Li-ion cells require well-adapted battery management and control systems.

Other electrochemical combinations can be considered for particular applications; examples are rechargeable silver–zinc batteries, used on Ariane, and silver–hydrogen batteries (Ag H_2) that provide high specific energies but a limited lifetime, for low-orbit satellites. Table 10.7 summarises the performance of various types of electrochemical cell (a KOH solution is diluted potassium hydroxide); the performance indicated corresponds to a DOD of 100%. Table 10.8 compares the performance of the three types of cell of interest for satellite use. Sodium cells are also a candidate for space applications but require high operating temperatures (350 °C).

Table 10.7 Characteristics of battery cells

Type of cell	Electrolyte	Nominal cell voltage (V)	Energy density (W h/kg)	Temperature range (°C)	Cycle life at levels of depth of discharge		
					25%	50%	75%
NiCd	KOH solution	1.25	25–30	−10 to +40	20 000	3000	800
NiH_2	KOH solution	1.30	50–70	−10 to +40	15 000	>2000	1000
Li-ion	Non-aqueous	3.6	120–175	0 to +40	>60 000	>10 000	>1500
AgCd	KOH solution	1.10	60–70	0 to +40	3500	750	100
AgZn	KOH solution	1.50	120–130	+10 to +40	2000	400	75
Pb-Acid	Diluted sulphuric acid	2.10	30–35	+10 to +40	1000	700	250

Table 10.8 Advantages of lithium-ion cells compared to other available technologies

	NiCd	NiH_2	Li-ion	System impact of Li-ion
Energy density ($Wh\,kg^{-1}$)	30	70	165	Weight saving
Energy efficiency (%)	72	70	96	Reduction of charge power
Thermal power (on a scale of 1–10)	8	10	3	Reduction of radiator, heat pipe sizes
Self discharge (%/day)	1	10	0.3	No trickle and simple management at launch pad
Temperature range (°C)	0–40	−20 to 30	0–40	Easy management
Memory effect	Yes	Yes	No	No reconditioning
Energy gauge/monitor	No	Pressure	Voltage	Better observation for states of charge
Charge management	Constant current	Constant current	Constant current then constant voltage	Weight saving
Modularity	No	No	Yes	Ability to put cells in parallel

(a)

(b)

Figure 10.26 (a) Lithium-ion cells; (b) cells packaged in a battery module. Source: courtesy of SAFT.

10.4.2.4 Operation of the battery

Battery charging. Battery charging is generally performed at constant current, the current I_{ch} being such that:

$$I_{ch} = C/10 \text{ to } C/15 \quad \text{(A)} \tag{10.31}$$

where C is the cell capacity expressed in A h.

The recharge time T_{ch} is a function of the energy $E = PT_{ecl}$ (W h) provided during the eclipse, the battery voltage V_{ch} during charging, the charging current I_{ch}, and the charging efficiency η_{ch} (0.75–0.9):

$$T_{ch} = PT_{ecl}/I_{ch}V_{ch}\eta_{ch} = C\,\text{DOD}/I_{ch}\eta_{ch} \quad \text{(h)} \tag{10.32}$$

It must be verified that the charge time is less than the time between eclipses (22.8 hours for a geostationary satellite). The power required for the recharge is given by:

$$P_{ch} = I_{ch}V_{ch}/\eta_{reg} \quad \text{(W)} \tag{10.33}$$

where η_{reg} is the efficiency of the charge regulator. In the absence of a regulator (an unregulated bus), $\eta_{reg} = 1$.

Care must be taken not to overcharge the battery. Protection against overcharging can be provided by limiting the voltage at the end of charging: that is, by terminating the charge at constant voltage. Another charging technique consists of operating the end of charge at a given temperature. This technique has the advantage of a better charging efficiency, better precision of the end of charge, and an increase in the lifetime of the battery.

Once the battery is charged, a continuous current on the order of $C/75$ to $C/50$ (trickle charge) compensates for the self-discharge of the battery, if the cell technology requires it (NiH_2).

Battery discharge. The battery discharge current can be relatively large and is limited by heating problems associated with the internal resistance. A value on the order of $C/2$ to $C/5$ is a good compromise that avoids an excessively rapid fall of voltage at the end of discharge. The mean voltage during discharge depends on the current; it is on the order of 1.2 V per cell for NiCd, 1.3 V per cell for NiH_2 (Figure 10.27), and 3.6 V for Li-ion. At the end of the discharge, the voltage

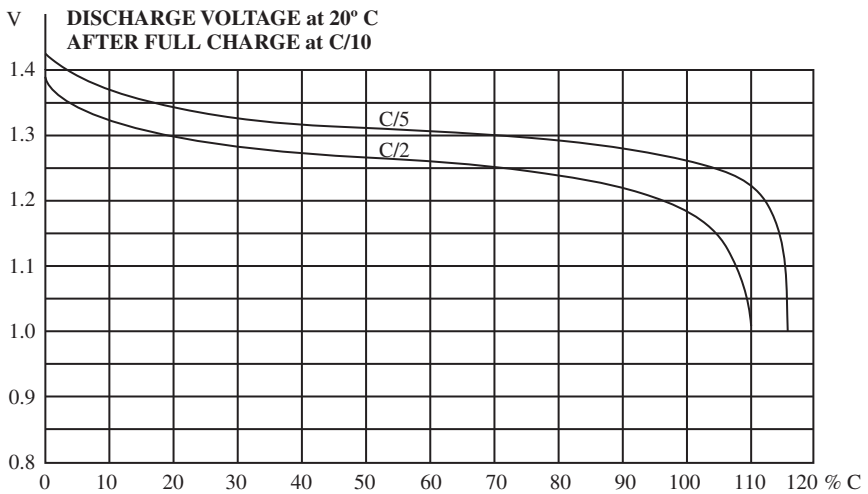


Figure 10.27 Discharge characteristic of a NiH_2 cell.

decreases rapidly. It is then necessary to stop the discharge (at a minimum voltage on the order of 0.7 V) in order to avoid polarity inversion.

Operating temperature; reconditioning. The temperature range within which the battery elements must be maintained is narrow, typically 0–15 °C. Too low a temperature affects the performance; a high temperature limits the lifetime. Furthermore, the battery gives off heat during discharge and cools during charging. Thermal control is thus designed so that the battery can radiate heat and electric heaters limit the fall of temperature during recharging.

After a long period of non-use (with the battery fully charged but not used), it is necessary to perform reconditioning in order to restore the nominal performance of the battery (memory effect). The battery is first completely discharged at low current. The battery is then fully recharged at low current; one or more charge–discharge cycles can be added. Note that Li-ion batteries do not require reconditioning.

10.4.3 Conditioning and protection circuits

The voltage delivered by the primary energy source (the solar generator) depends on the operating point. The power available depends on the incident solar flux and the degree of degradation of the solar cells due to irradiation.

During an eclipse, the temperature of the solar generator decreases and can fall to −180 °C. After the eclipse, the voltage delivered by the cells is around 2.5 times the nominal value that corresponds to the equilibrium temperature (on the order of a few tens of degrees in the illuminated phase). As far as the battery is concerned, the voltage delivered by a cell depends strongly on the state of charge. The battery terminal voltage can vary from 15% to 30% between the start and end of an eclipse depending on the DOD realised.

It is, therefore, necessary to provide conditioning circuits for the electrical energy intended for distribution to the equipment and to provide battery charging. The circuits will be associated with those for control and protection. Ohmic losses must, of course, be minimised in power distribution; evaluation of the losses associated with the various stages of regulation and distribution, to which ohmic losses are added, shows that, in certain cases, a third of the power generated by the solar generator is effectively used by the payload.

10.4.3.1 Unregulated bus

The schematic arrangement is presented in Figure 10.28. The solar generator feeds the platform equipment directly and the payload by way of a distribution unit. The operating point is defined by the intersection of the current–voltage (I_G, V_G) characteristic of the generator and that representing the equipment load (a hyperbola representing operation at constant power P_E). Figure 10.29 illustrates these characteristics and the chosen operating point (N).

The battery is connected through a switch either to a recharging solar generator in series with the main generator (out of eclipse) or to the supply bus during eclipse periods. Out of eclipse, the bus voltage varies in accordance with the evolution of the characteristics of the solar generator and the power consumed.

When the satellite is in eclipse, the battery, which then provides the power, determines the potential V_B of the bus, which feeds the equipment (point B; the battery has a characteristic of the voltage source type). The bus voltage decreases with time during the discharge.

The *unregulated bus* has the advantage of simplicity and hence good reliability. On the other hand, the equipment connected to the bus is subjected to large voltage variations (10–40%). Some equipment may be designed to accept such variations of supply voltage (e.g. equipment to supply travelling wave tubes [TWTs]), others require a voltage converter and regulator at the

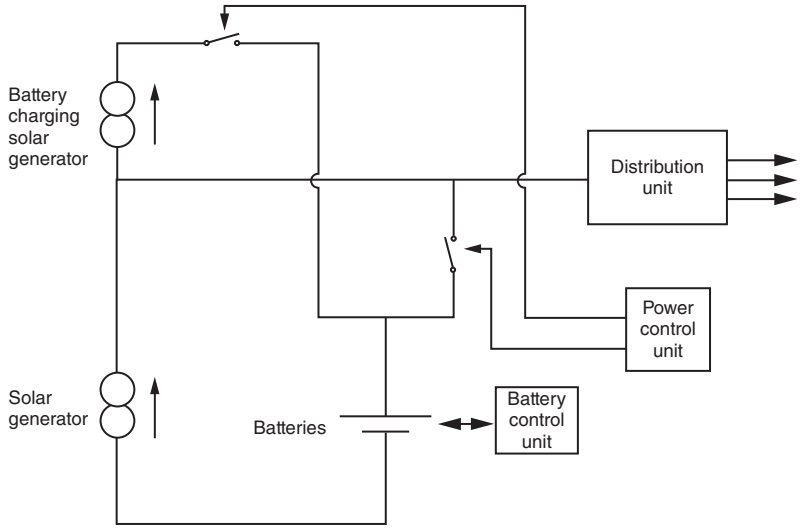


Figure 10.28 Unregulated bus power supply.

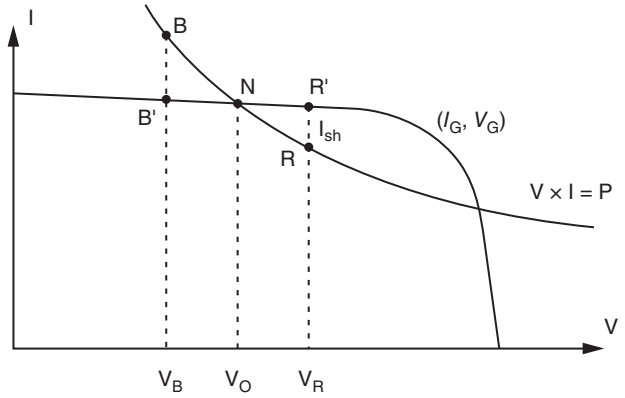


Figure 10.29 Operating points with the various types of voltage regulation bus.

point where the electrical energy is distributed to the equipment concerned (for both the payload and the platform).

10.4.3.2 Sun-regulated bus

In order to limit voltage variations for most of the time, voltage is regulated during periods of sunlight. The organisation of the power system is presented in Figure 10.30. The solar generator feeds the equipment at constant voltage by means of a voltage regulator. Outside eclipses, the equipment supply voltage is thus kept constant (equal to V_R) within a range of a few per

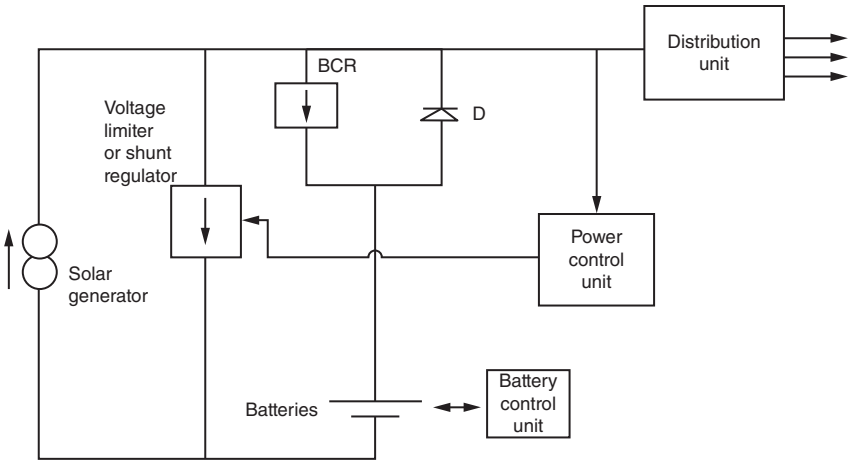


Figure 10.30 Sun regulated bus power supply (battery charge regulator [BCR]).

cent depending on the performance of the regulator. Two operating points are now defined in Figure 10.29: that of the solar generator, point R, and that of the load, point R′. The segment RR′ represents the current I_{sh} shunted through the regulator.

A battery charge regulator (BCR) connected to the bus provides recharging and maintains constant battery current outside eclipses (the bus voltage is greater than the battery voltage). When the satellite is in eclipse, the battery is directly connected to the bus, which feeds the equipment (point B). The bus voltage imposed by the battery voltage V_B then changes with the discharge of the battery.

The *sun-regulated bus* remains relatively simple in concept and in operational use. The battery supply to the bus is provided without intervention (due to the diode) when the bus voltage falls below the battery voltage following a temporary increase of demand or entry of the satellite into eclipse. Voltage variations outside periods of eclipse are limited. On the other hand, the equipment connected to the bus is still subjected to variations of battery voltage during discharge.

10.4.3.3 Regulated bus

Voltage regulation during sunlight and eclipse periods is obtained by decoupling the battery from the bus by means of a battery discharge regulator (BDR) (Figure 10.31). During sunlight, a voltage regulator maintains constant the voltage of the solar generator and the bus to which the equipment is connected. During an eclipse, the battery provides the power to the bus by way of the BDR, which keeps the bus voltage constant and equal to V_R (point R in Figure 10.29).

The advantages of operating equipment at constant voltage are obtained to the detriment of system complexity and hence a potential reduction in reliability. Problems of EMC can arise when the equipment is connected directly to the bus. Depending on the impedance of the system, variation of the current consumed by equipment may lead to voltage variations that can cause coupling between equipment; this applies particularly with operation in time division multiple access (TDMA) mode, which involves peak current demands by the amplifiers at the frame rate, should all time slots not be occupied by traffic bursts.

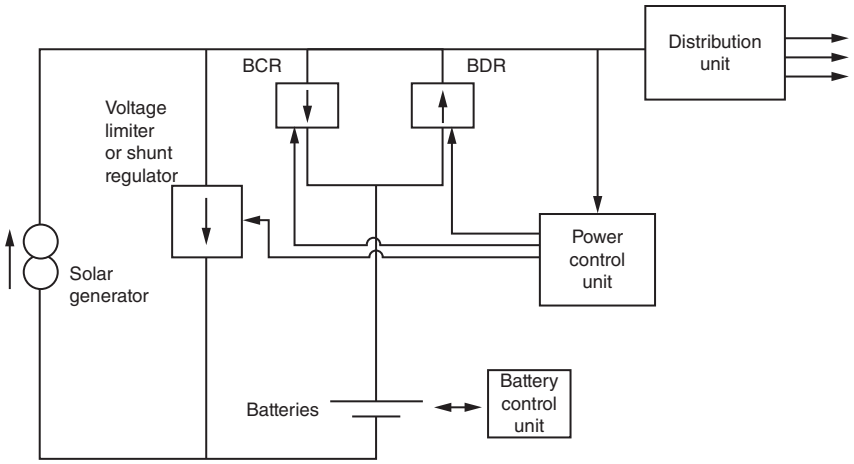


Figure 10.31 Regulated bus power supply (battery charge regulator [BCR]; battery discharge regulator [BDR]).

10.4.3.4 Other organisations

The three organisations presented here are not the only possible ones, and other combinations can be conceived. Hence, it is possible to use one or more dedicated elements of the solar generator for battery recharging instead of taking power from the main bus.

Another solution involves use of the battery on exit from an eclipse as a buffer to limit the voltage increase with an unregulated bus; the voltage increase is due to the excess power delivered by the solar generator at low temperature.

It is also possible to define an architecture where the solar generator, the battery, and the equipment are permanently in parallel. The battery thus feeds the equipment automatically on entry into an eclipse. It recharges itself on exit from the eclipse by imposing its voltage on the bus. Once charged, the voltage is relatively constant and the battery plays the role of a buffer. A shunt resistance connected in parallel with the battery by means of a switch enables any excess current to be absorbed.

A voltage regulator of the shunt type is usually used since, on the one hand, this type of regulator is well suited to the behaviour of a source of the current generator type and, on the other hand, it avoids any voltage drops between the generator and the equipment. However, due to the appearance of semiconductor components having a low voltage drop in the on state (such as HEXFET transistors), the use of regulators of the series type can be considered. The principle of the two types is illustrated in Figure 10.32.

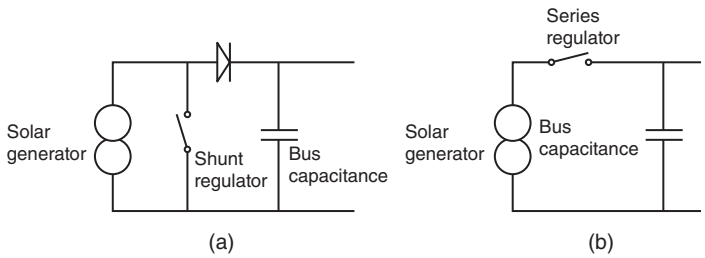


Figure 10.32 Organisation of (a) shunt and (b) series regulators.

10.4.3.5 Comparison of the various architectures

In addition to the advantages and weaknesses indicated previously, unregulated systems suffer during eclipse from a lock-out phenomenon on the battery voltage, which necessitates overdimensioning the solar generator in order to escape from the resulting stable state without interruption of the power supplied to the equipment.

The reason for this lock-out is as follows: for unregulated and sun-regulated organisations, on passing through an eclipse, the solar generator characteristic disappears from Figure 10.29 and the operating point of the load moves from point R or R′ towards point B. At the end of the eclipse, the solar generator characteristic reappears and the operating point of the solar generator is then point B′. The solar generator thus provides a power $V_B I_{B'}$. The balance of the load power P is borne by the battery (BB′V_B). This situation corresponds to a stable state with no exit (in order to be able to disconnect the battery) except by cancelling this balance of power. This means displacing the operating point of the generator from point B′ to point B and hence the availability of a generator capable of providing a larger power than that strictly necessary out of eclipse, as defined by the operating points N and R.

One criterion in choosing between the various organisations is also the overall mass and efficiency of the power supply subsystem. A comparative study of the performance of the subsystem concerned for various satellites using different organisations indicates that the use of a *regulated bus* provides the best performance in terms of global mass and efficiency. Even if the regulated bus requires additional equipment, its mass is compensated for by the elimination of converters and regulators in the distribution of power to equipment. The operating point of the solar generator is optimised as a consequence of decoupling of the battery and operation at constant voltage.

Furthermore, with a regulated bus, the battery voltage is no longer directly determined by the bus voltage and can thus be chosen to optimise the number of battery cells in accordance with available capacities. The battery voltage can be chosen to be above or below the bus voltage. The regulated bus architecture is the most efficient in terms of mass for a high power level, above several kW.

It is preferable to recharge the battery with a dedicated section of the solar generator instead of extracting the power required for recharging from the bus by means of a BCR. Below about 5 kW, it is preferable to choose a battery voltage less than the bus voltage whatever the type of battery. Above about 5 kW, a battery voltage higher than the bus permits a reduction of mass.

Finally, selecting a sufficiently high nominal bus voltage, which leads to lower currents for a given power, improves the system efficiency by reducing ohmic losses. Typical voltages range from 42 to 50 V (commonly used in the past) to above 100 V when high powers are required.

10.4.3.6 Redundancy, protection, and distribution circuits

Redundancy of the main elements of the electric power supply subsystem (the solar generator and the battery) is impossible on account of the mass and large bulk. To minimise the chance of catastrophic breakdown and loss of the whole subsystem due to the failure of one element, the subsystem on a communications satellite is generally organised into two separate supply branches on which the satellite equipment is distributed. Each of these two branches is fed by one wing of the solar generator and associated with a separate battery. Interconnection facilities are provided for the various elements in case of breakdown of one of them. The secondary equipment (regulators, switches, and control devices) are replicated.

The power supply subsystem also incorporates protection circuits (for example, battery discharge limiters to avoid polarisation inversion of the cells). In this way, certain precautions can be envisaged: for example, allowance for failure of one cell of a battery (such as using reverse

diodes or a parallel relay to ensure electrical continuity in case of open-circuit). These solutions are to be used with care; the addition of extra elements may lead to a larger overall probability of breakdown than that of the element that the device is intended to protect.

Finally, the subsystem contains regulation and conversion circuits that are designed to supply equipment with regulated voltages. These circuits must have the highest possible efficiency, low mass, and high reliability. The high efficiency is obtained by using chopping techniques of various kinds to raise or lower the voltage.

10.4.4 Example calculations

The power to be provided at the EOL (10 years) is 2000 W (spin-stabilised satellite) or 6000 W (three-axis stabilised satellite). This power is to be provided under the least favourable conditions: that is, at the summer solstice when the solar flux is given by $\phi = 1370 \times 0.89 = 123\,\mathrm{W\,m^{-2}}$ (see Figure 12.2).

10.4.4.1 Spin-stabilised satellite

A cylindrical satellite is covered on the external lateral surface with silicon solar cells of surface area $s = 4\,\mathrm{cm^2}$. In the course of rotation, the cells successively face the sun and cold space. Hence their mean temperature remains low, on the order of 10 °C. At 10 °C, the efficiency of the cells is $e = 14\%$. The degradation of efficiency at EOL is estimated at 22%. The various losses (due to cell protection windows, cabling, etc.) are included in a factor $l = 0.9$. The filling factor of the panels is $f = 0.85$.

As a consequence of the cylindrical form and rotation, the form factor F, which characterises the ratio of the actual surface on which the cells are mounted to the equivalent surface normal to the axis of the cylinder, is equal to π.

The power at EOL is given by:

$$P = (1 - 0.22)\phi\,\mathrm{elns}/F$$

The number of cells n required is thus:

$$n = \frac{2000\pi}{(1 - 0.22) \times (1370 \times 0.89) \times 0.14 \times 0.9 \times (4 \times 10^{-4})} = 131082$$

The required surface A taking into account the filling factor f is equal to

$$A = ns/f = 131\,082 \times 4 \times 10^{-4}/0.85 = 62\,\mathrm{m^2}$$

which represents a cylinder of diameter 3.5 m and height 5.6 m.

By estimating the mean mass per cell (including cabling, mounting, and the protecting window) at 0.8 g and a support of mass 1.6 kg m^{-2}, the total mass is $131\,082 \times 0.8 \times 10^{-3} + 62 \times 1.6 = 204$ kg. The specific power is thus $2000/204 = 9.8\,\mathrm{W\,kg^{-1}}$.

A power requirement of 6000 W would result in a cylinder volume that could not be accommodated under the launch vehicle fairing.

10.4.4.2 Three-axis stabilised satellite

The satellite contains two rectangular orientable solar panels covered with cells of $2 \times 4 = 8\,\mathrm{cm^2}$. The panels are aligned with the pitch axis and can be oriented about this axis. For the sake of comparison with the spin-stabilised spacecraft, the same type of cells is considered. As the cells continuously face the sun, degradation is greater and equal to 28% in 10 years. The mean

temperature is also higher, and this leads to a lower efficiency at the BOL, which thus has a value $e = 13\%$. The losses are slightly greater due, for example, to the shadow of the antenna farms on the panels, hence $l = 0.88$. The filling factor f is equal to 0.90.

The power at the EOL is given by:

$$P = (1 - 0.28)\phi\, e \ln s$$

Aiming at EOL power of 6000 W, the number of cells n required is thus:

$$n = \frac{6000}{(1 - 0.28) \times (1370 \times 0.89) \times 0.13 \times 0.88 \times (2 \times 4 \times 10^{-4})} = 74687$$

The required surface area A taking into account the coefficient of filling f is equal to $74\,678 \times 8 \times 10^{-4}/0.90 = 66.4\,\text{m}^2$.

Estimating the mean mass per cell at 1.2 g and a support of mass $0.6\,\text{kg m}^{-2}$, the total mass is $74\,678 \times 1.2 \times 10^{-3} + 66.4 \times 0.6 = 129.5\,\text{kg}$. The specific power is thus $6000/129.5 = 46.3\,\text{W kg}^{-1}$.

Considering modern triple-junction cells with BOL efficiency of 28.5% and taking into account a degradation factor for 10 years life of about 15% (less sensitivity to high energy particles) will lead to a significant reduction in size of the solar array.

10.4.4.3 The battery

Assume the power to be provided on board the geostationary satellite under consideration during the longest eclipse is $P = 4000$ W. NiH_2 batteries of capacity $C = 93$ Ah are used with a DOD of 80% to guarantee a lifetime of 10 years; the mean voltage during discharge is $V_d = 1.3$ V per cell, and the discharge efficiency is $\eta_d = 0.95$.

Since the duration of the longest eclipse is 1.2 hours, the energy to be provided during the eclipse is $E = 1.2\ P$ (W h). The energy E_c provided by a battery cell of capacity C is given by $E_c = CV_d\text{DOD}\eta_d$ (W h). The number of cells required is thus:

$$n = 1.2P/CV_d\text{DOD}\eta_d = 1.2 \times 4000/93 \times 1.3 \times 0.8 \times 0.95 = 52\,\text{cells}$$

To increase the reliability of the system, the battery is divided into two sections of 26 cells each. The nominal voltage during discharge is 33.8 V. To avoid loss of a cell by short-circuit, one cell more than is necessary is added to each half battery. Also, diodes are placed at the terminals of each battery to ensure electrical continuity in case of open-circuit breakdown.

If the exact mass of a battery is not known, a rapid estimate of the mass can be obtained from typical values of specific energy per unit mass E_m of the technology used.

The energy to be stored is $E_s = 1.2\ P/\text{DOD}\eta_d = 6315$ W h. The mass of the battery is thus given by:

$$M(\text{kg}) = E_s/E_m = 1.2P/\text{DOD}\eta_d E_m$$

Assuming a specific energy E_m equal to $55\,\text{W h kg}^{-1}$, the mass of the battery is estimated as $M = 115$ kg.

10.5 TELEMETRY, TRACKING, AND COMMAND (TTC) AND ON-BOARD DATA HANDLING (OBDH)

Telemetry, tracking, and command (TTC) [ETSI-17; ETSI-18]] deals with:

— Receiving control signals from the ground to initiate manoeuvres and to change the state or mode of operation of equipment

— Transmitting results of measurements, information concerning satellite operation, the operation of equipment, and verification of the execution of commands to the ground
— Enabling measurement of the ground–satellite distance, and possibly the radial velocity, in order to permit location of the satellite and determination of orbit parameters

On-board data handling (OBDH) is often associated with the TTC subsystem. OBDH includes all housekeeping, data processing and formatting, together with data traffic and time management on board the satellite.

Telemetry and telecommand links are low-bit rate links, a few kilobits per second at most. This differs for scientific satellite telemetry (such as observation of the earth), for which the data rates are much greater, typically a few tens of megabits per second.

One of the major characteristics required of TTC links is availability. Ensuring availability of the TTC links is fundamental for diagnostics in case of breakdown and for performing corrective actions.

The necessary reliability is obtained by means of suitably replicated transmitting and receiving equipment (transponders). This equipment is associated with one or more antennas that have a radiation pattern such that the gain is as constant as possible, or at least greater than a minimum value, throughout most of the space around the satellite. This permits links to be established whatever the attitude of the satellite.

10.5.1 Frequencies used

The TTC links should nominally be operated within the space operation service (SOS) frequency bands. The frequencies normally used are in S band as follows:

— The uplink in the band: 2025–2120 MHz
— The downlink in the band: 2200–2300 MHz

It is clear that the available bandwidth (on the order of 100 MHz) is insufficient to accommodate all the modulated carriers from the various satellites in orbit. Also, these bands are reserved for operations associated with injecting the satellite into orbit and an emergency mode in case of a problem in normal mode in the operational phase. During these phases, the orientation of the satellite attitude with respect to the earth is arbitrary. An omnidirectional antenna is thus indispensable.

In order to avoid congestion in these bands, the TTC links in normal mode are routed through the satellite payload. They therefore use the frequency bands corresponding to the nominal satellite service: the fixed satellite service (FSS), for example.

The problem associated with the use of nominal frequency bands is that usually the payload antennas are directional and, in the case of a pointing error due to an attitude control problem, the TTC links are interrupted. It is therefore necessary for the TTC links to be rerouted via the transmitting and receiving system in S band; this should be automatic since access to the satellite from the ground is no longer possible. This function is realised by means of a selecting device that is activated by command from the ground once the satellite is on station in the nominal configuration and routes the service links via the satellite payload by means of a priority relay. The device then monitors the level of the received telecommand carrier. As long as the level remains within predefined limits, the detector output signal keeps the relay energised. In the case of a pointing error, or a problem on the uplink of the nominal mission, the relay switches to the released position that corresponds to routing via the S band repeaters (Figure 10.33). Since the replicated repeaters are associated with one or more antennas that have a quasi-omnidirectional

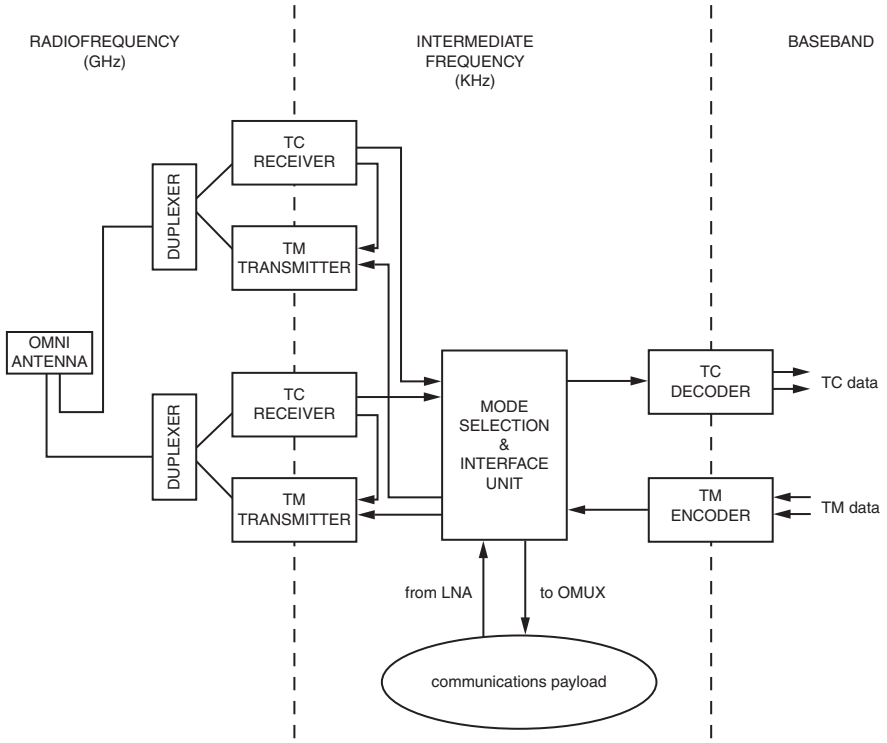


Figure 10.33 Block diagram of a centralised TTC subsystem.

radiation pattern, the TTC links continue to be transmitted even if the attitude is highly perturbed.

However, the nominal frequency band of the communications payload can also be used to handle TTC data during the injection of the satellite in orbit. The same TC receiver and TM transmitter are then used during injection and the operational life of the satellite. To obtain omnidirectional coverage during injection, several wide beamwidth antennas on different sides of the satellite are usually used. Once on station, the input of the TC receiver and the output of the TM transmitter are then connected to the main antennas of the communications payload.

10.5.2 The telecommand links

The telecommand (TC) links are provided by a carrier whose frequency depends on the band used and is phase or frequency modulated by a subcarrier at a few kilohertz (e.g. 8 kHz). Digital phase modulation of this subcarrier by the data has previously been realised. The bitstream (often NRZ-L formatted) has a data rate ranging from some hundreds of bit/s to several kbit/s, depending on the application. On account of the low bit rate, the use of a subcarrier enables the useful spectrum (that of the data) to be separated from the carrier itself.

The commands to be transmitted are regulating commands to either adjust a parameter on board the satellite to a particular value (for example, the helix current of a TWT) or load registers

in a computer or memory with binary system commands (for example, opening or closing a relay, 0 and 1 for open and closed).

Commands, depending on the particular mode selected, can be:

— Executed immediately after reception
— Stored in memory and executed on reception of a specific command
— Stored in memory and executed when activated at a given time determined by the time management system on board the satellite or when activated by a signal produced by one of the subsystems on board the satellite

Repetition enables the integrity of the received words to be verified before they are used. One of the important characteristics of the telecommand link is security. It is fundamental for the survival of the satellite that it really is the selected command that is executed.

Various precautions are taken, such as error-correcting coding of the data words, repetition for verification and detection of possible differences, deferred execution of commands, etc. With deferred execution of commands, the command is detected on board the satellite, stored in memory, retransmitted to the ground by telemetry for verification, and executed only after authentication by an execution command sent on the telecommand link.

Finally, precautions are also taken to make the system insensitive to signals transmitted by intruders (anti-jamming); these include narrow-band reception, input limiters, insensitivity to nonstandard signals, and possible encryption on the link.

Use of a *spread-spectrum link* permits the problems of both interference between systems and protection from undesired signals to be solved and also makes efficient use of the frequency bands through the possibility of multiple access to the same band.

10.5.3 Telemetry links

Telemetry (TM) links are also provided by a carrier that is phase or frequency modulated by a subcarrier at a few kilohertz (e.g. 40.96 kHz). The digital data phase modulates this subcarrier. The data rate ranges from a few tens of bits/s to a few kilobits/s.

The data to be transmitted may consist of analogue information signals, corresponding, for example, to the results of measurements, digital words (the value in a register or the output of an encoder), or binary system states (0 or 1, relay open or closed). Analogue information signals are sampled, quantised, and encoded with a number of bits depending on the required resolution and the range of amplitude variation of the signal. A clock is necessary to discretise the analogue information. In accordance with the dynamic behaviour of the information transmitted, sampling is not performed at the same rate for all signals; with respect to a basic cycle, some signals are undersampled (sampled at multiples of the cycle) and others are oversampled (sampled several times per cycle).

The data are obtained in either of the following ways:

— Directly from satellite equipment and conditioned (analogue–digital conversion, formatting, etc.) in the telemetry encoder
— At the output of a processing unit on a local on-board network to which the various satellite equipment has access (see Section 10.5.5)

10.5.4 Telecommand (TC) and telemetry (TM) message format standards

There is a need for message format standards in order to ensure compatibility of the satellite telecommand decoding and telemetry encoding systems with the control ground stations and the ground data-handling and -processing systems. In addition, a standard common to the various space agencies and satellite operators allows a given party to operate their satellite using the facilities of another body if required. Finally, the purpose of a standard is to unify data interfaces and data-handling operations in ground systems.

Two main types of standard have been elaborated:

— The pulse code modulation (PCM) standards were published in the 1970s and, as an example, defined for ESA the PSS-45 and PSS-46 documents for command and telemetry, respectively (note that the former ESA PSS documents have now been replaced by the ECSS documents available at https://ecss.nl).
— The packet standards originated from the recommendations of the Consultative Committee for Space Data Systems (CCSDS, www.ccsds.org) and are available in a series of reports, such as CCSDS 100.0-G-1 for telemetry (Telemetry Summary of Concept and Rationale. Silver Book. Issue 1, December 1987) and CCSDS 200.0-G-6 for telecommand (Telecommand Summary of Concept and Rationale. Green Book. Issue 6, January 1987).

To keep up with the requirements of space technologies and applications, both ESA and CCSDS have revised their documents.

The ESA procedures, specifications, and standards (PSS) documents concerning space data communications have now been replaced ECSS standards and are available at https://ecss.nl. The ECSS documents have three types: standard, technical, and memoranda. These are grouped into five branches: S for system, M for management, Q for product assurance, E for Engineering, and U for sustainability. Each document is coded ST for standard, AS for adopted as standard from another standards organisation, HD for handbook, or TM for technical memo.

For example, ECSS-E-ST-70-41C, 'Telemetry and telecommand packet utilisation' (15 April 2016), is an ECSS document of the Engineering (E) branch, Standard (ST), Number 70–41, Issue C (a two-number code for a specific requirement and one number for a generic requirement).

Here are some more examples:

- ECSS-E-ST-50C: Communications (31 July 2008)
- ECSS-E-ST-50-01C: Space data links – Telemetry synchronisation, and channel coding (31 July 2008)
- ECSS-E-ST-50-03C: Space data links – Telemetry transfer frame protocol (31 July 2008)
- ECSS-E-ST-50-04C: Space data links – Telecommand protocols, synchronisation and channel coding (31 July 2008)
- ECSS-E-ST-70-41C: Telemetry and telecommand packet utilisation (15 April 2016)

The CCSDS has classified its documents according to the following areas:

- Space Internetworking Services
- Mission Ops. and Information Management Services
- Spacecraft Onboard Interface Services

- System Engineering
- Cross Support Services
- Space Link Services

All publications of the documents are colour coded: blue for Recommended Standards, magenta for Recommended Practices, green for Informational Reports, orange for Experimental, and yellow for Records, as well as silver for Historical together with Numerical Numbers. CCSDS Experimental Specifications are used for research and development and as such are not considered standards track documents, but may be rapidly transferred onto the standards track should the requirement emerge. Patent licencing may apply to various documents.

Here are some additional examples:

- CCSDS 130.0-G-3: Overview of Space Communications Protocols. Green Book. Issue 3. July 2014.
- CCSDS 130.1-G-2: TM Synchronisation and Channel Coding – Summary of Concept and Rationale. Green Book. Issue 2. November 2012.
- CCSDS 130.2-G-3: Space Data Link Protocols – Summary of Concept and Rationale. Green Book. Issue 3. September 2015.
- CCSDS 131.0-B-3: TM Synchronisation and Channel Coding. Blue Book. Issue 3. September 2017.
- CCSDS 132.0-B-2: TM Space Data Link Protocol. Blue Book. Issue 2. September 2015.
- CCSDS 230.1-G-2: TC Synchronisation and Channel Coding – Summary of Concept and Rationale. Green Book. Issue 2. November 2012.
- CCSDS 231.0-B-3: TC Synchronisation and Channel Coding. Blue Book. Issue 3. September 2017.
- CCSDS 232.0-B-3: TC Space Data Link Protocol. Blue Book. Issue 3. September 2015.

All the up-to-date documents can be found on the websites for ECSS (https://ecss.nl) and CCSDS (https://public.ccsds.org/Publications). Here, only the principles and generic issues are discussed.

10.5.4.1 PCM standards

Telecommand. The telecommand message is organised in frames preceded by a group of bits for acquisition of synchronism. Each frame consists of words of several bits. The length of the frame depends on the applied standard. For example, in the ESA standard, the frame consists of 96 bits (Figure 10.34). The first word, of 16 bits, constitutes an address and synchronisation word specific to the destination decoder. It is followed by a mode-selection word of four bits, which is immediately repeated. The mode-selection word is used to indicate the type of command transmitted in three words of 12 bits; the three words are repeated once and contain the data. The frame terminates with a repetition of the address and selection word (80 bits between the two words). The 12-bit words used to carry the data may contain data coded in 8 bits, which has been extended to 12 bits by means of an error-correcting code.

Telemetry. The telemetry message is organised in frames, and a group of frames constitutes a *format*. Each frame consists of words and starts with a synchronising code; the first frame contains a format-identification word. Frames are identified by a counter. In the ESA standard, the format consists of 16 frames, and each frame contains 48 words. The eight-bit words constitute the data; this may be part of a data word for data whose resolution requires coding in more than eight bits (the data word is then shared among several eight-bit words) or a block of data that requires only a single bit, such as the state of a relay.

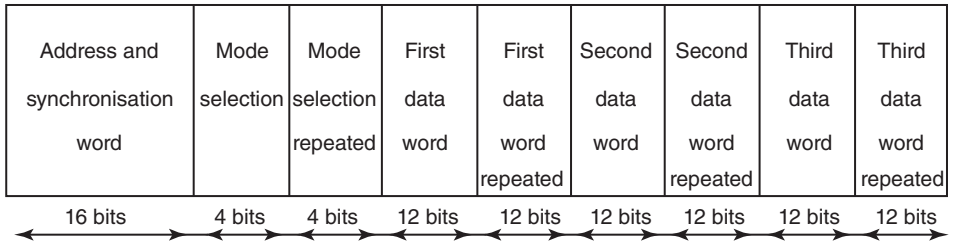


Figure 10.34 Command message frame in the ESA standard.

10.5.4.2 Packet standards

The CCSDS is an international organisation; its members are the various national space agencies around the world. From the beginning of the 1980s, CCSDS developed a series of technical recommendations for the standardisation of common data systems functions (radio frequency and modulation, packet telemetry, telemetry channel coding, and packet telecommand).

Telecommand. The introduction of more capable microprocessor-based spacecraft subsystems results in an increase in spacecraft autonomy and complexity, requiring data systems with high-throughput needs. Common requirements for greater telecommand capability and efficiency with reduced costs are addressed by the CCSDS telecommand concept. The proposed layer structure allows a complex spacecraft commanding procedure to be decomposed into sets of simple functions. Within each layer, the functions exchange information using standard data-formatting techniques and standard protocols. Three main data formats are defined:

— User data packets (TC packets)
— The TC transfer frame to reliably transport TC packets (or segmented packets)
— The telecommand link transmission unit or command link transmission unit (CLTU), which encapsulates the channel-coded transfer frames

Figure 10.35 shows the layer structure for telecommand data organisation (after CCSDS 232.0-B-3).

Telemetry. Packet telemetry is a concept that facilitates the transmission of space-acquired data from sources to ground users in a standardised and highly automated manner. Packet telemetry has a layer structure, where each layer implements the different functions to allow the multiplexing of various kinds of telemetry data onto a single physical RF channel. Two main data structures are defined in packet telemetry:

— A *source packet* encapsulates a block of source data, which may include ancillary data and which may be directly interpreted by the ground user processor. The source packet header contains an identifier used to route the packet to its destination sink, and also information about the length, sequence, and other characteristics of the packet (Figure 10.36).
— The *transfer frame* has a fixed length for a given mission or satellite, and it embeds the source packets to provide reliable and error-controlled transfer through the transmission media. The transfer frame header permits the ground system to route the source packets to their intended destination.

Virtual channelisation is used in the multiplexing mechanism that allows the various sources that generate packets to be given virtually exclusive access to a physical channel by assigning

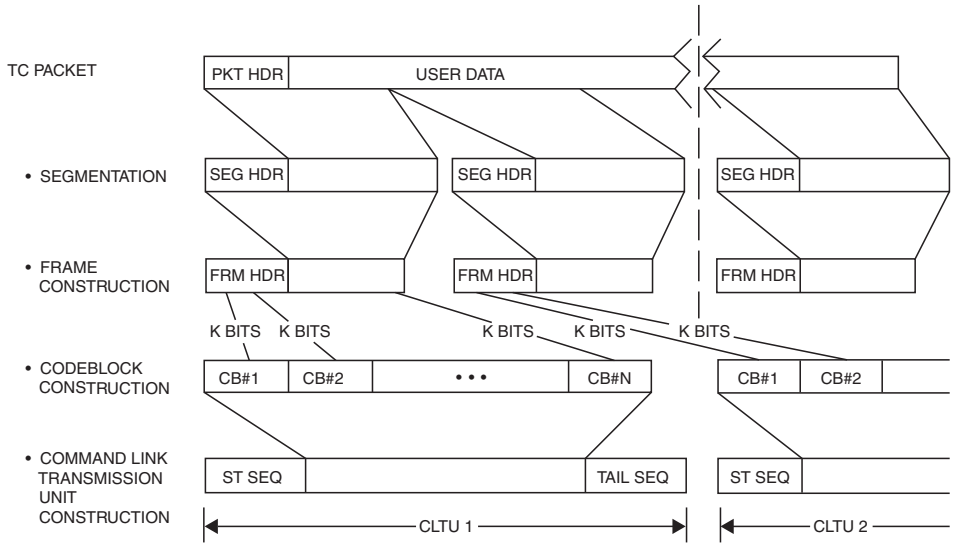


Figure 10.35 Telecommand data structures. Source: reproduced courtesy of CCSDS.

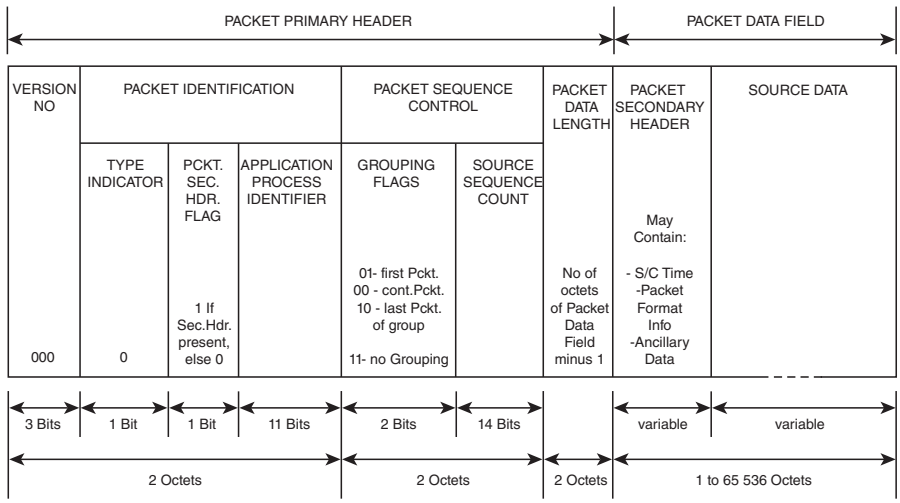


Figure 10.36 Telemetry source packet format. Source: reproduced courtesy of CCSDS.

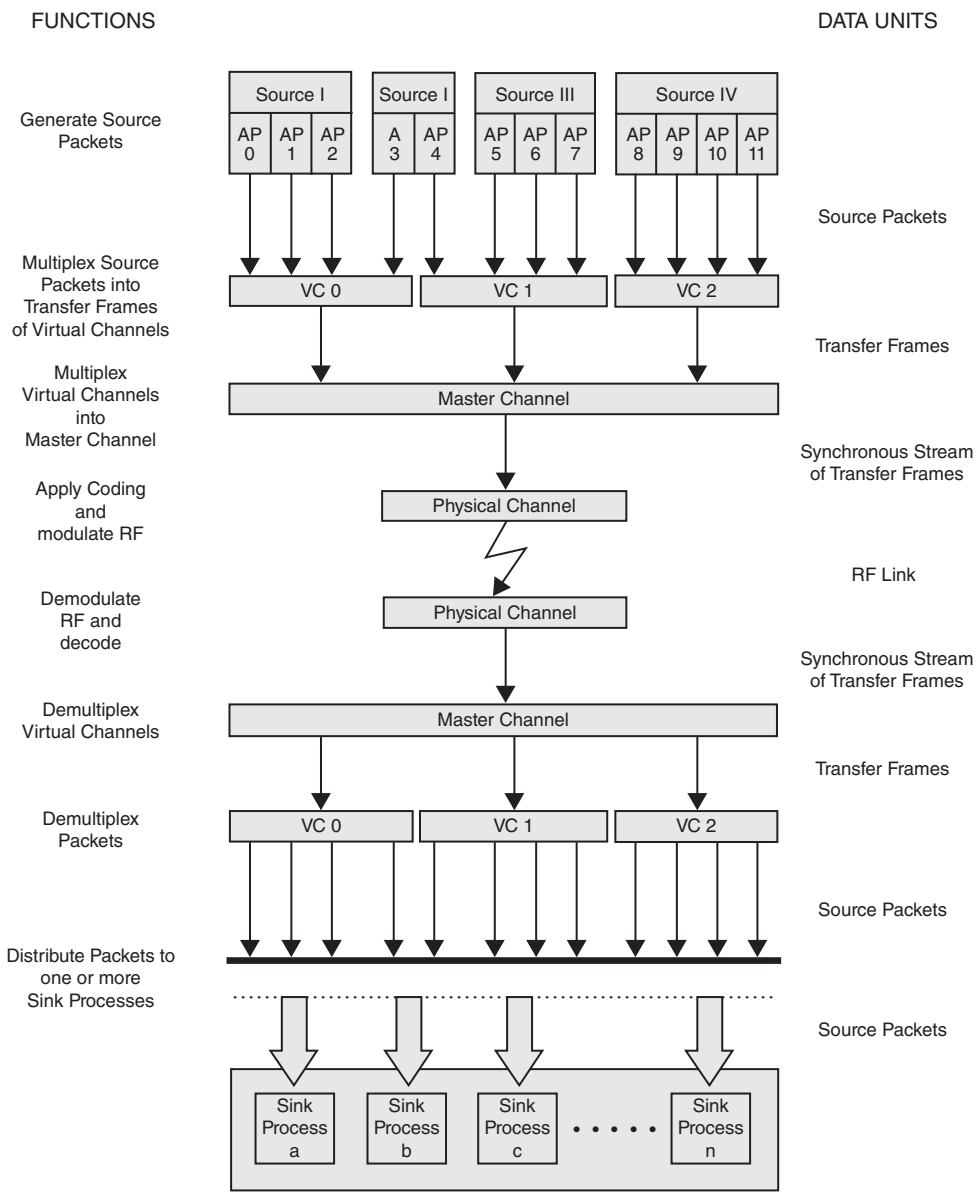


Figure 10.37 Telemetry data flow. Source: reproduced courtesy of CCSDS.

them transmission capacity on a frame-by-frame basis. So a virtual channel is a given sequence of transfer frames, which are assigned a common identification code, enabling unique identification of all transfer frames belonging to that sequence. Figure 10.37 illustrates the overall telemetry data flow through this layered packet telemetry structure.

Recommendations for advanced orbiting systems (AOSs). CCSDS decided to improve its conventional packet standard recommendations to provide a more diverse and flexible set of data

handling services to meet the needs of AOSs. Typical AOSs include manned and man-tended space stations, unmanned space platforms, free-flying spacecraft, and advanced space transportation systems. Many of these systems need services to concurrently transmit multiple classes of digital data (including audio and video) through space–ground, ground–space, and space–space data channels at relatively high combined data rates, and the conventional concepts of telemetry and telecommand become blurred. Instead, the forward and return space links become the vehicles for extensive two-way interchange of many different classes of digital message traffic between ground and space. Some more details on AOS links can be found in the following recommendations.

— CCSDS 700.0-G-3: Advanced Orbiting Systems, Networks, and Data Links: Summary of Concept, Rationale, and Performance. Green Book. Issue 3. November 1992.
— CCSDS 732.0-B-3: AOS Space Data Link Protocol. Blue Book. Issue 3. September 2015.

To handle different classes of data that share a single link, various transmission schemes (e.g. asynchronous, synchronous, isochronous) are provided, as are different user data-formatting protocols (e.g. bitstreams, octet blocks, and packetised data) and different grades of error control. Capabilities are included to run commercially derived ground network protocols into the space segment. The recommendations therefore provide a space-adapted analogue of the terrestrial concept of an integrated services digital network (ISDN), recognised as the Consultative Committee for Space Data Systems Principal Network (CPN). New protocols have been developed including:

— CCSDS 732.1-B-1: Unified Space Data Link Protocol. Blue Book. Issue 1. October 2018.
— CCSDS 734.0-G-1: Rationale, Scenarios, and Requirements for DTN in Space. Green Book. Issue 1. August 2010.

Space Communications Protocol Specification (SCPS). In order to facilitate compatibility with terrestrial networks and to reduce space system development costs, CCSDS started to adopt the Internet protocol (IP), the de facto terrestrial standard, on top of the currently developed CPN application-specific integrated circuits (ASICs). As shown in Figure 10.38, this SCPS reuses most of the IP with some modifications for the space environment (for example, document CCSDS 714.0-B-2, SCPS Transport Protocol, Blue Book, 2006). This aims at facilitating the internetworking between spacecraft and terrestrial networks for the next generation of space missions.

10.5.5 On-board data handling (OBDH)

OBDH deals with:

— Command processing: decoding, validation, acknowledgement, and execution (immediate or deferred) of command signals
— Acquisition, compression, coding, and formatting of telemetry information
— Processing of data: relating to the on-board management subsystem itself (such as, time and configuration) and on demand from the satellite subsystems
— Storage of data: telemetry data, modes, and software
— Synchronisation, data timing, and traffic management: on-board time management, distribution of on-board timing and clock signals to subsystems, dating of events and measurements, and management of traffic between subsystems

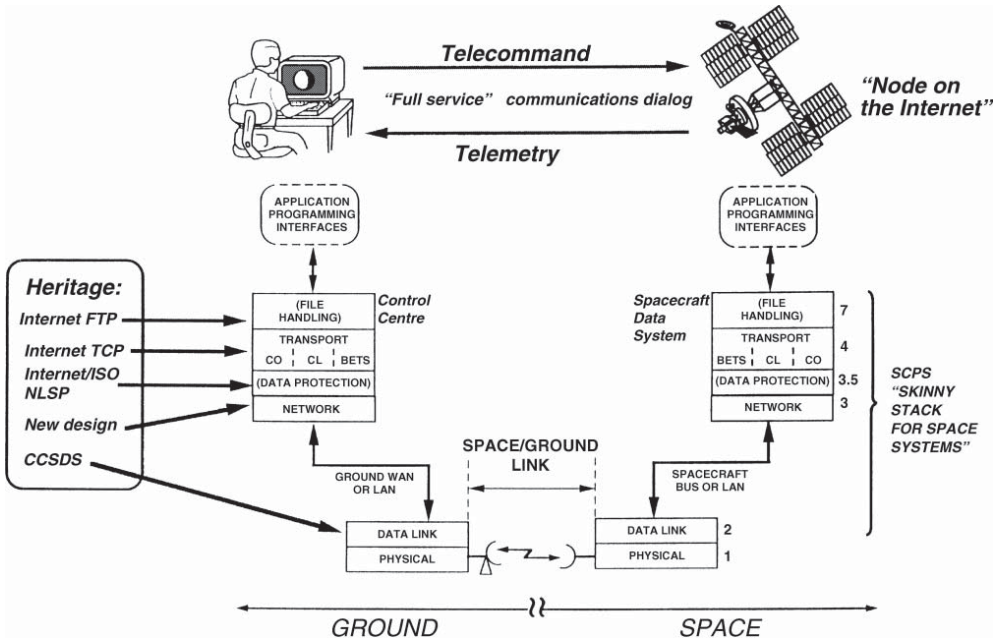


Figure 10.38 Adaptation of Internet protocols to the Space Communications Protocol of CCSDS.

— Monitoring and control: acquisition and analysis of monitored and diagnostic parameters, decision taking (e.g. changing into survival mode and reconfiguration), generation, and execution of appropriate commands.

Depending on the type of satellite and its complexity, these functions may not all be required. In particular, on-board management of the first communications satellites was rudimentary in respect of the small number of TTC channels that communicated with the ground (a few tens). The number of channels was subsequently increased (to several hundred at the end of the 1970s), but on-board management functions remained limited and could be satisfied by two specific pieces of equipment: the telecommand decoder and the telemetry encoder. During the 1980s, the increase in size and complexity of satellites has led to an increase in the number of telecommand and telemetry channels (4400 for Intelsat VI). At the same time, microprocessors and high-capacity memories, which can be used in a space environment, have been developed. These developments have led to organisation of the OBDH in association with the telecommand and telemetry subsystem in a modular form based on a data-transfer bus.

10.5.5.1 Centralised architecture

The OBDH can be restricted to decoding telecommand signals and encoding telemetry signals. The interfaces with this equipment consist of a subcarrier modulated by a bitstream from or to the respective RF equipment (the telecommand receiver and the telemetry transmitter) and the telecommand and telemetry signals to or from the satellite equipment on as many links as there are signals. The block diagram of this architecture is given in Figure 10.33.

The functions to be realised are limited mainly to processing of command signals (decoding, validation, and execution) and telemetry signals (acquisition, formatting, and dating):

— The telecommand decoder detects bits after recovery of the bit rate (primary synchronisation), separates the various format components (such as the address and the mode) from the data (secondary synchronisation), and validates and transmits the execution commands after demultiplexing the data on the various equipment channels.
— The telemetry encoder performs analogue-to-digital conversion of analogue telemetry signals, multiplexes the different channels and generates the data format by adding identification and synchronisation bits, and generates and modulates the subcarrier with the bitstream to obtain the intermediate frequency (IF) signal, which in turn modulates the telemetry carrier at the TM transmitter.

The increase in the number of TTC channels leads to increased complexity of this equipment. Furthermore, it is necessary to route the various electrical signals separately from the satellite equipment to the TTC subsystem. This leads to bulky cabling with a high mass (13 km and 130 kg of cables on Intelsat VI). This architecture is thus unsuitable for modern satellites, which require a large number of TTC channels as a consequence of the complexity of their subsystems.

10.5.5.2 Modular architecture

The system organisation uses a decentralised architecture with a communication bus between the various pieces of data processing and handling equipment (Figure 10.39). The various modules are the command decoder, the central terminal unit (CTU), the data bus, and the remote terminals.

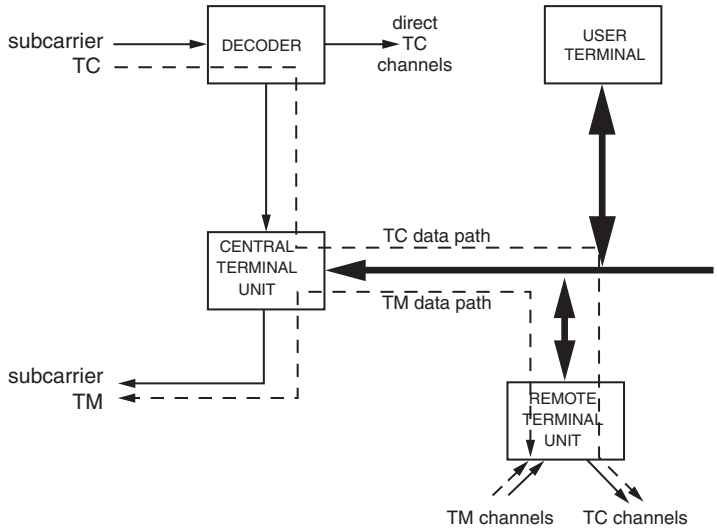


Figure 10.39 Modular architecture for data management.

The telecommand decoder. As with centralised architecture, this restores the bitstream and then separates the various components of the data format. Commands are divided into two categories: those processed by the CTU for transmission on the data bus to the appropriate terminal, and those of high priority that are processed by the decoder (direct TC channels).

The CTU. This equipment performs many functions:

— Handling the command data concerned and distributing it to the equipment via the bus
— Handling the data traffic on the bus, defining on-board time and distributing it to users
— Generating terminal interrogations for acquisition of telemetry data
— Multiplexing data from remote terminals via the bus and that which it generates directly and formatting it to produce the telemetry bitstream
— Modulating the subcarrier with the bitstream to obtain the IF signal destined for the telemetry transmitter

If the need arises, this equipment can monitor critical parameters and take appropriate decisions (such as reconfiguration, changing to emergency mode, and so on). The CTU is associated with an OBC that provides the required processing capacity. This computing power can be made available to the various satellite subsystems to perform processing in time sharing on data relating to a subsystem (centralised processing).

Data bus. This carries data and clock signals between the CTU and the various units; in certain cases, it also carries power signals to activate relays (to connect or disconnect equipment). Date-exchange management is governed by protocols.

Remote terminal unit (RTU). On activation by the CTU, the RTU performs the following functions:

— Acquiring command data and interrogations on the bus
— Directing commands to the equipment concerned in the form of electrical signals that are user specific
— Delivering clock signals available on the bus to the users

— Acquiring telemetry signals from users on the various channels, performing analogue-to-digital conversion and encoding if necessary, and transmitting telemetry data on the bus (on request from the CTU)

RTUs are used to connect simple equipment, which does not have internal control capacity (examples are relays and heaters for temperature control), to the bus. More sophisticated equipment already has computing capability and is thus capable of controlling exchanges when connected directly to the data bus. Remote terminals and users can be provided with computing capacity known as intelligent terminal units (ITUs), which permits data processing to be performed locally (decentralised processing, as opposed to centralised processing where the computer is associated with the CTU).

Some subsystems that consist of several units dispersed about the satellite may have their own data bus, which then accesses the main data bus by way of a remote interface unit.

10.5.5.3 Data bus interfaces and protocols

In order to allow the use of standardised equipment, the interfaces between units and the communication protocols have been the subject of standardisation. Two main standards are used for the local housekeeping data network on board a communications satellite: the ESA OBDH standard (refer to ECSS-E-ST-50-13C: Interface and communication protocol for MIL-STD-1553B data bus onboard spacecraft, 15 November 2008) and the US MIL-STD-1553 B standard (also refer to MIL-STD-1553: Tutorial and Reference, by Alta Data Technologies, 5 January 2015).

OBDH standard. The data bus consists of an interrogation line and a response line (full duplex) on shielded twisted pairs. The standard defines the coding of the bits, the data structure, the protocols, the user interface, etc. The general characteristics of the standard are:

— Compatible with the PCM/TC ESA and PCM/TM ESA standards
— Two-wire, full-duplex bus protocol (interrogation and response)
— Up to 31 users connected to the bus
— Response time on the bus $<140\,\mu s$
— Data rate on the bus ≤ 500 kbps
— Bus length < 20 m
— Distribution of one to five clocks
— No possibility of interruption of the CTU by users (CTU bus management is by polling)
— No internal redundancy or possibility of standard redundancy
— Up to 2×48 direct ON/OFF commands provided by the decoder
— 19 useful bits per 32-bit packet on the bus
— On-board clock: 4–6 MHz, stability 10^{-6} seconds per year

1553-B standard. The data bus consists of a single line (half duplex) on a shielded twisted pair; it provides a single data path between the bus controller (BC) and all the associated remote terminals (RTs). The 1553 protocol ensures that only one station is transmitting at any given time, thanks to the bus monitors. The standard defines the coding of the bits, the data structure, the protocol, the user interface, etc. The general characteristics of the standard are:

— Single-wire, half-duplex bus
— Asynchronous transmission
— Three types of unit connected to the bus: the bus controller, remote terminals, bus monitors
— Up to 31 users connected to the bus

— Data rate up to 1 Mbit second
— Three word formats: command word, data word, status word
— Two terminal operating modes: transmit and receive
— Up to 30 subaddresses for a given terminal address
— Up to 32 data words: 20-bit words carrying 16 information bits
— Biphase (Manchester) waveform coding

10.5.5.4 Satellite control

Satellite control operations consist, in some cases, of elaborating commands in order to execute actions in accordance with information available on board the satellite.

With the rudimentary on-board management architectures of first-generation satellites (see Section 10.5.5.1), the information from telemetry channels is grouped in the control station where the context is analysed, and the decisions taken are transmitted on the telecommand channels. This mode of operation is quite suitable for geostationary communications satellites, which are continuously visible from the control station and whose management is simple (they have few TTC channels).

The existence of computing power on board the satellite enables information processing and direct generation of appropriate commands on board the satellite to be considered. This permits the load on the control station to be lightened and the availability of the satellite to be increased by automatic reconfiguration in case of breakdown; for orbiting satellites, complex control actions can be contemplated even when they are invisible from earth stations.

A control hierarchy is thus always instituted; the first level corresponds to operations and major events that are still processed by the control station; the second level corresponds to events and operations directly processed on board the satellite. A third level can be introduced into the hierarchy when the subsystems themselves are capable of controlling their mode of operation and reconfiguring in case of breakdown (a decentralised processing capacity in the intelligent remote terminals).

10.5.6 Tracking

10.5.6.1 Distance (range) measurement

Distance measurement is performed by means of specific subcarriers that modulate the telecommand carrier, are coherently demodulated in the receiver, and are then used to modulate the telemetry carrier. Comparison of the initial phase of the signals with the phase of the demodulated signals on the ground enables the round-trip time to be obtained. This time, from which the precisely known time delay in the receiving equipment is deducted, permits the distance to the satellite to be calculated.

Various approaches are possible, depending on the nature of the subcarrier; these include fixed frequency (tone), variable frequency, modulation by a pseudorandom number (PN) sequence, etc. The tone system is currently used. In this case, the subcarrier is a sinusoidal wave of fixed frequency f. Measurement of the phase shift between the transmitted and received tones, which is a function of the distance R from the station to the satellite (a round-trip trajectory of $2R$), enables this distance to be determined:

$$\Delta\phi = 2\pi f(2R)/c \tag{10.34}$$

where c is the velocity of light.

Table 10.9 Distance ambiguity ΔR as a function of tone frequency f

f	100 kHz	20 kHz	4 kHz	800 Hz	160 Hz	32 Hz	8 Hz
ΔR (km)	1.5	7.5	37.5	187.5	937.5	4687.5	18 750

As phase shift is measured modulo 2π, the measurement is the same for all values of R such that $2\pi f(2R)/c = 2k\pi$ where k is an integer. The distance ambiguity ΔR corresponds to the distance obtained for $k = 1$. Table 10.9 gives the distance ambiguity ΔR as a function of the frequency of the tone used. Observe that, for a geostationary satellite, the frequency of the tone must be at most 8 Hz for the measurement to be made without ambiguity.

The choice of tone frequency is thus guided by conflicting considerations. A high frequency (100 kHz) is necessary to ensure accuracy of phase measurement; conversely, the frequency must be low enough for the wavelength to be long enough with respect to the distance to be measured for there to be no ambiguity.

The difficulty is avoided by transmitting two tones simultaneously; these are a major tone at 100 kHz, which permits good measurement accuracy; and a minor tone, obtained by division of the major tone (and hence in phase with it), which enables the ambiguity to be resolved. The procedure is as follows: the major tone at 100 kHz is first transmitted with the first minor tone at 20 kHz (division by 5). On reception, the minor tone is compared with five signals separated in phase by $2\pi/5$, which are obtained by division by five of the received major tone. Only one of these signals is in phase with the received minor tone and is selected (Figure 10.40a). A replica of the received minor tone at 20 kHz is obtained, but with the phase accuracy of a 100 kHz tone. This signal serves in turn to create five signals at 4 kHz, also differing in phase by $2\pi/5$, which are compared in turn with the new minor tone at 4 kHz, which has replaced that transmitted at 20 kHz (the major tone remains continuously transmitted to ensure continuity of the replica retrieved on reception). Only one of the five signals is in phase and is selected in its turn. The process is repeated with minor tones of 800 Hz, 160 Hz, 32 Hz, and finally 8 Hz. In this way, a signal at 8 Hz is obtained by successive division of the major tone at 100 kHz and whose phase relationship is known (Figure 10.40b). Comparison with the transmitted minor tone at 8 Hz thus permits determination of the distance with the accuracy obtained for a 100 kHz tone.

Figure 10.41 shows an example of the spectrum of the signal that modulates the telemetry carrier (subcarrier modulated by the data at 40.96 kHz) or the telecommand carrier (subcarrier modulated by the data at 8 kHz). The minor tones are transposed in frequency between 16 and 20 kHz to reduce the required bandwidth (Table 10.10).

Measurement accuracy depends on the tone frequency, the received signal-to-noise ratio, the stability of the transit time in the satellite equipment, and variations of propagation time in the ionosphere. For a tone of frequency f, the root mean square value of the distance error is given by:

$$S_{err} = \frac{c}{4\pi f}\left(\frac{k}{\sqrt{S/N}}\right) \tag{10.35}$$

where $[k/\sqrt{(S/N)}]$ is the quadratic mean value of the phase error, k is a constant depending on the structure of the receiver used, and S/N is the signal-to-noise ratio at the phase detector input. Variations of tropospheric propagation time can cause an error of between 0 and 300 m (at 2 GHz), but this can be estimated separately. Distance measurement can be performed with an error on the order of a few tens of metres.

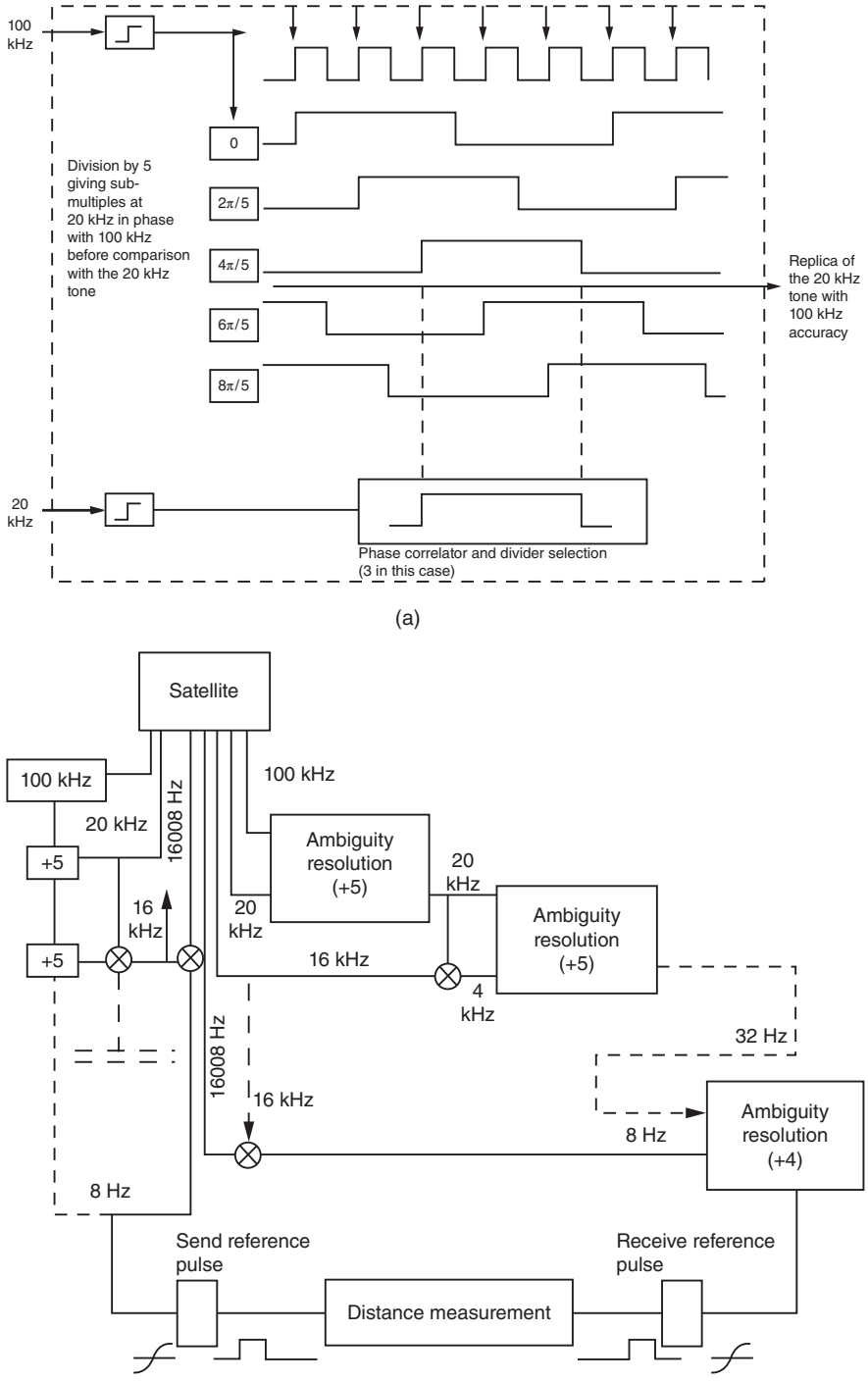


Figure 10.40 Principle of tone range measurement.

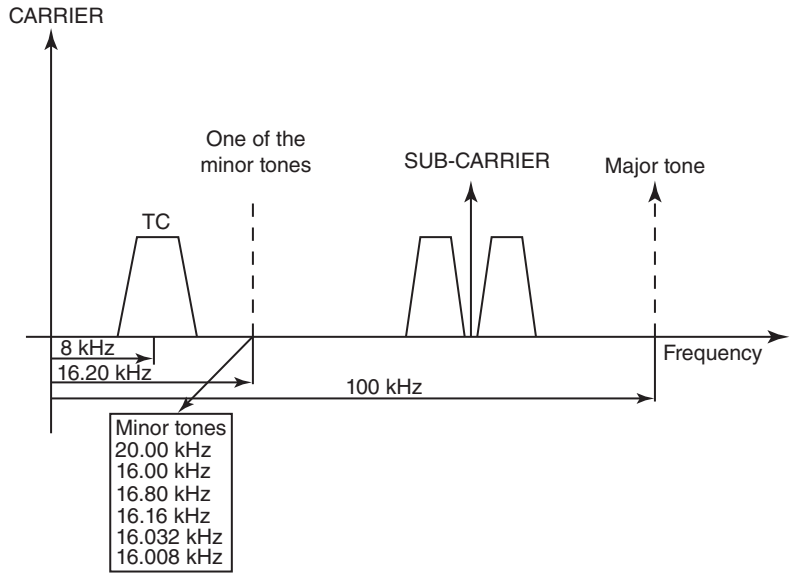


Figure 10.41 Spectrum of the telemetry carrier modulating signal.

Table 10.10 Major and minor tones

Major tone	Divider	Virtual minor tone (Hz)	Transmitted tone (Hz)
100 kHz	5	20 000	20 000
	5	4000	16 000
	5	800	16 800
	5	160	16 160
	5	32	16 032
	4	8	16 008

10.5.6.2 Measurement of radial velocity (range rate)

Radial velocity can be obtained by measurement of the Doppler effect. It is necessary to ensure frequency and phase coherence at the transponder between the downlink and uplink carriers. The nominal frequency f_d of the downlink carrier is such that:

$$f_d/f_u = 240/221 \tag{10.36}$$

where f_u is the nominal frequency of the uplink carrier. If the satellite is given a velocity V_r with respect to the control station, the frequency received on board is equal to:

$$f_u^* = f_u[1 + (V_r/c)] \text{ (Hz)} \tag{10.37}$$

where c is the velocity of light.

The frequency retransmitted on the downlink is obtained by multiplication by the ratio 240/221 and the frequency received by the control station is equal to:

$$f_d^* = (240/221)f_u[1 + (V_r/c)]^2 \text{ (Hz)} \tag{10.38}$$

Taking account of the very small value of V_r with respect to c, this gives:

$$f_d^* = (240/221)f_u[1 + (2V_r/c)] \tag{10.39}$$

The radial velocity is thus obtained as a function of the frequency difference Δf between the received frequency f_d^* and the nominal frequency f_d on the downlink ($f_d = (240/221)f_u$):

$$V_r = -(c/2)(221/240)\Delta f/f_u \ (m/s) \tag{10.40}$$

Measurement of the radial velocity requires operation of the transponder in coherent mode that can be different from the normal non-coherent mode where the downlink carrier is obtained from an on-board oscillator. Mode selection is achieved by telecommand.

10.6 THERMAL CONTROL AND STRUCTURE

The purpose of thermal control is to maintain the satellite equipment within the temperature ranges that enable it to operate satisfactorily, by providing its nominal performance, and avoiding any irreversible deterioration when it is not operating. This also applies to the structure of the satellite, which must remain within a mean temperature range in order to minimise deformation and guarantee precise alignment of attitude stabilisation sensors and antennas.

10.6.1 Thermal control specifications

Thermal control must be optimised with respect to the constraints of both the operational and transfer phases. These constraints are very different (due to different orbits and attitudes, the state of the apogee motor, etc.).

The objectives of thermal control are thus to maintain the equipment within specified temperature ranges; these differ for equipment when operating and when on standby. Furthermore, the behaviour of the equipment may differ depending on whether it is operating or at rest; in operation, it usually generates heat that the thermal control must remove; when at rest, the equipment must, in certain cases, be heated in order to avoid an excessively low temperature. Finally, the maximum values of temperature gradients (with respect to time) must also be considered.

10.6.1.1 Specified temperature ranges

The temperature ranges to be maintained differ greatly from one piece of satellite equipment to another. Examples are:

— Antenna: −150 °C to +80 °C
— Electronic equipment: −30 °C to +55 °C (on standby); +10 °C to +45 °C (operating)
— Solar generator: −160 °C to +55 °C
— Battery: −10 °C to +25 °C (on standby); + 0 °C to +10 °C (operating)
— Solar sensor: + 30 °C to +55 °C
— Propellant reservoir: + 10 °C to +55 °C
— Pyrotechnic unit: −170 °C to +55 °C

These temperature ranges are those that the equipment may be expected to encounter once in orbit. This implies that the equipment has been designed to operate at, or withstand, more

extended temperature ranges than those stated. In particular, a specified range within which the nominal performance of the equipment must be maintained is defined by adding modelling errors to the limits of the estimated temperature range. A still wider range within which the equipment must not suffer irreversible degradation is also defined.

10.6.1.2 Characteristics of the space environment

The characteristics of the space environment are presented in Chapter 12. As far as thermal control is concerned, the most useful characteristics are recalled in Figure 10.42. It must be remembered that the satellite is subject to the effects of three radiation sources (the sun, the earth, and the terrestrial albedo) that have different spectral distributions and geometric forms and are absorbed differently by the surface of the satellite. Eclipses and variations of attitude and distance modify the illumination conditions over the course of time. Cold space absorbs all radiation from the satellite. The ambient vacuum prevents convection.

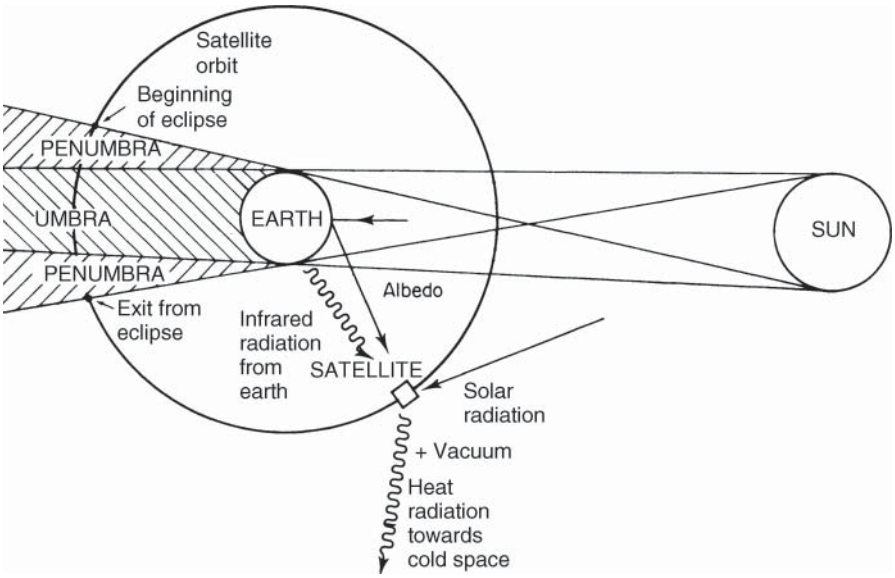


Figure 10.42 Space environment characteristics of importance to thermal control.

10.6.1.3 The principle of thermal control

The mean temperature of a piece of equipment is the result of an equilibrium between the heat generated internally, the heat absorbed and radiated by the surfaces of the unit, and the heat received or removed by conduction through the mechanical mounting of the equipment.

The mean temperature of the satellite is the result of equilibrium between the heat generated internally, the heat absorbed by the surfaces of the satellite, and the heat radiated by the surfaces of the satellite. Thermal control, therefore, consists of:

— Adjusting the thermal conductivities between the various parts of the satellite, either to favour heat exchanges by conduction between one point and another (by choice of material, surface

and section, use of heat pipes, etc.) or to limit exchanges (by use of insulating materials, etc.)
— Making use of the thermo-optical properties (such as emissivity and absorption) of surfaces to favour, for example, the extraction of heat by radiation while minimising that captured using optical solar reflectors (OSRs)
— Providing local sources of heat if necessary (electric heaters)
— Arranging for certain surfaces to be able to radiate to remote space without constraint in order to lower the temperature (infrared detectors)

Thermal control is realised either passively or actively with a preference for passive control wherever possible in view of its simplicity, low cost, and reliability.

10.6.2 Passive control

Passive control is based on the absorptive and emissive properties of surfaces. *Absorptivity* α is defined as the ratio of the power absorbed by unit surface area to the incident power. *Emissivity* ε is defined as the ratio of the power radiated per unit surface area to the power that would be radiated by unit surface area of a black body. Recall that the power radiated per unit surface area of an ideal black body (W/m^2) is σT^4, where T is the temperature of the black body (K) and s $\sigma = 5.67 \times 10^{-8}$ Wm^{-2} K^{-4} is the Stefan–Boltzmann constant (see Section 12.3).

Depending on the material used, the values of absorptivity α and emissivity ε vary between zero and unity. For a given material, the ratio α/ε is of prime importance in determining the mean temperature of a surface exposed to the sun.

10.6.2.1 Types of surface

Various types of surface are used:

— White paint absorbs infrared radiation (terrestrial flux) and reflects solar flux. It is a cold surface in sunlight (−150 °C to −50 °C) since the ratio α/ε is small (ε can reach 0.9, α is on the order of 0.17).
— Black paint absorbs all wavelengths but is also characterised by a high emissivity ($\varepsilon = 0:89$). The absorptivity is also high ($\alpha = 0:97$). The temperature in sunlight is greater than 0 °C.
— Aluminium paint absorbs little and emits little. The emissivity is low (ε on the order of 0.25) and so is the absorptivity (α on the order of 0.25). The equilibrium temperature in sunlight is close to 0 °C On the other hand, as the emissivity is less than that of black paint, an aluminium covering is warmer in shadow than a black covering.
— Polished metal absorbs the visible part of the solar spectrum (solar absorbers) but reflects infrared radiation. These surfaces are warm in sunlight (50 °C to 150 °C) since the ratio α/ε is high (for example, for gold $\varepsilon = 0:04$ and $\alpha = 0:25$).

The values of the thermo-optical properties of claddings are affected by uncertainties and variations due to characterisation errors, problems of reproducibility during fabrication, sensitivity to contamination, and degradation due to the effect of the space environment.

In order to limit the exchanges, a surface can be isolated by using *superisolating padding* or *multilayer isolation* (MLI). This padding consists of several sheets of plastic (Mylar) aluminised on both faces and separated by a material of low conductivity (Dacron mesh). The exterior layer is aluminised only on the internal face and consists of either Kapton (with a golden appearance)

for temperatures that do not exceed 150 °C or titanium for high temperatures (up to 400 °C). The conductance is on the order of $0:05\,\mathrm{W\,m^{-2}\,K}$.

10.6.2.2 Radiating surfaces

For communications equipment that dissipates heat (such as power amplifiers), radiators having a very low absorptivity to emissivity ratio are used. These surfaces are, therefore, capable of efficiently radiating the heat generated while absorbing the least possible solar radiation. These radiators are strips of silica silvered on the reverse side as OSRs or produced from sheets of plastic material (Teflon, Kapton, or Mylar) with a deposit of silver or aluminium on the reverse face as *second surface mirror* (SSM).

A quick estimate of the surface area S required to extract the thermal power dissipated by the payload of a communications satellite can be obtained from considering that when the equilibrium temperature T (K) is reached, the sum of the thermal power P (heat in watts) and the power absorbed from the sun is equal to the power radiated by the surface considered:

$$P + \alpha\phi S = \varepsilon \upsilon S \sigma T^4 \tag{10.41}$$

where α and ε are the absorptivity and emissivity of the radiator, ϕ is the flux ($\mathrm{W\,m^{-2}}$) received from the sun and depends on the earth–sun distance and the angle of incidence, S is the radiating surface area ($\mathrm{m^2}$), and υ is the view factor of the radiating surface.

The view factor is the complement of the percentage of the 2π steradians of space above the surface that is obstructed (occulted) by obstacles (such as the solar generator). For a three-axis stabilised satellite, the radiating surfaces are mounted on the north and south sides of the satellite. The solar generators that are mounted in alignment with the north–south axis mask part of the space for the radiators. Typically, the coefficient υ is between 0.85 and 0.9. This coefficient applies only to the corresponding term in the radiated power. For the absorbed power, having regard to the angle of incidence, almost all of the surface participates in capturing solar flux.

Taking an equilibrium temperature of several tens of degrees (30–40 °C), the surface area S is calculated for the most unfavourable case: that is, at the EOL when α reaches its highest value as a consequence of degradation and at the time of year when the flux captured as a function of the orientation of the surface is maximum (see Section 12.3.1).

For a surface oriented perpendicularly to the equatorial plane:

$$\phi = \cos\partial \times d^{-2} \times 1370\,\mathrm{W/m^2}$$

where ∂ is the declination and d the distance (in astronomical units (AU)) of the sun. The most unfavourable case occurs just before the spring equinox, when, from Figure 12.3, $\partial = 4:3°$ and $d^{-2} = 1:011$, from which $\phi = 1:008 \times 1370\,\mathrm{W\,m^{-2}}$.

For a surface oriented parallel to the equatorial plane:

$$\phi = \cos(90° - \partial) \times d^{-2} \times 1370\,\mathrm{W/m^2}$$

The most unfavourable conditions arise:

— Just before the winter solstice for a surface situated on the south side of the satellite for which, from Figure 12.3, $\partial = 23:5°$ and d $-2 = 1:033$, from which:

$$\phi = 0.412 \times 1370\,\mathrm{W/m^2}$$

— At the summer solstice for a surface situated on the north side of the satellite for which, from Figure 12.3, $\partial = 23.5°$ and d−2 = 0 : 965, from which:

$$\phi = 0.385 \times 1370\ \mathrm{W/m^2}$$

It is clear that the values of ϕ are smaller for surfaces parallel to the equatorial plane. This justifies mounting radiating surfaces on the north and south panels of three-axis stabilised satellites.

When the radiating surface is not illuminated by the sun (for example, at the equinoxes and the summer solstice for a surface on the south side of the satellite), the equilibrium temperature of the surface (of a size already determined for the most unfavourable case) is obtained by putting $\phi = 0$ in Eq. (10.41). This temperature should not fall below the temperature range of equipment attached to the surface.

10.6.2.3 Example calculation for a three-axis stabilised satellite

The communications payload requires a DC power of 5400 W, and the RF to DC power efficiency is 50%. The following characteristics are assumed:

— Heat to be dissipated = 5400 × 0 : 5 = 2700 W (assumed to be equally shared between the north and south sides, i.e. 1350 W per side).
— α of the radiator after combustion of the apogee motor and 10 years in orbit = 0.17 (value at launch = 0.04, after combustion of the apogee motor = 0.07, absolute degradation of 0.01 for each year in orbit).
— ε of the radiator = 0.75.
— v is taken equal to 1.
— Equilibrium temperature of the radiator $T = 32\,°\mathrm{C} = 305\,\mathrm{K}$.

From Eq. (10.41):

$$S = P/(\varepsilon v \sigma T^4 - \alpha\phi)$$

The area of the surface to be mounted on the south side is:

$$S = 1350/(0.75 \times 5.67 \times 10^{-8} \times 305^4 - 0.17 \times 0.412 \times 1370) = 4.96\ \mathrm{m^2}$$

For the north side:

$$S = 1350/(0.75 \times 5.67 \times 10^{-8} \times 305^4 – 0.17 \times 0.385 \times 1370) = 4.85\ \mathrm{m^2}$$

The equilibrium temperature of the radiating surface of the south side at the summer solstice is given by:

$$T = (P/S\varepsilon\sigma)^{1/4} = (1350/4.96 \times 0.75 \times 5.67 \times 10^{-8})^{1/4} = 282\mathrm{K} = +10°\mathrm{C}$$

All this assumes that the heat dissipated by the radiating equipment is uniformly distributed on the surface and hence the temperature of the radiating surface is the same at all points. Since the heat is generated by devices (such as TWTs) that have a relatively small mounting area, it is necessary to distribute the heat from the equipment across the surface. Heat pipes (see Section 10.6.3) are used.

10.6.3 Active control

This is used to complement passive control and may consist mainly of:

— Electric resistance heaters, controlled by thermostats or by command
— Movable shutters (Maltese cross or louvres), more or less covering the radiating surface and controlled by a temperature transducer (a bimetallic strip) or by command
— Heat pipes (which can also be classed as passive thermal control) that transfer heat from hot points to radiators under a small temperature difference by means of successive vaporisation and condensation of a fluid at the two extremities of a tube (Figure 10.43)

Heat pipes provide a large capacity for heat transfer due to the high values of latent heat of the fluids used. They are used to distribute the heat dissipated by equipment on the radiating

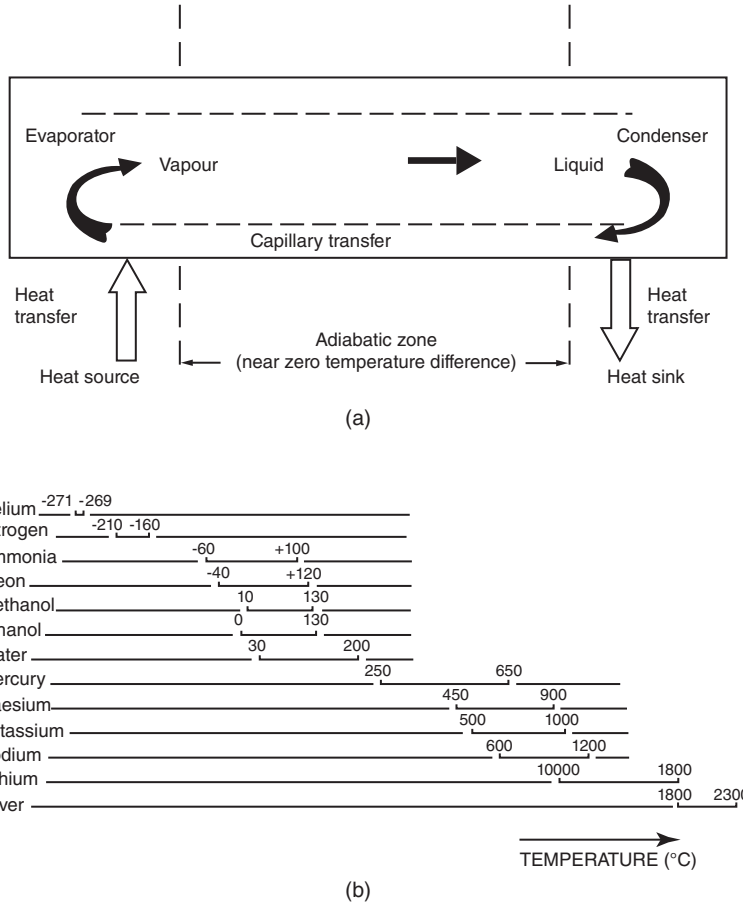


Figure 10.43 Heat pipe: (a) principle of operation; (b) temperature operating range.

surface. They can be buried in the honeycomb structure that forms the panel and are also used to connect the north and south sides to balance the panel temperatures at the solstices.

10.6.4 Structure

The functions of this subsystem can be classified as mechanical, geometric, and other functions.

10.6.4.1 Mechanical functions

— Supporting the on-board equipment, particularly during the launching phases when the mechanical constraints imposed by the launcher are highest (see Figure 12.7)
— Permitting the various separations and deployments that change the satellite from its launch phase configuration to its operational configuration, and accepting the forces acting during these operations (such as deployment of solar generators and antennas)
— Providing the satellite with the required rigidity (launcher and appendices decoupling)
— Permitting handling of the satellite on the ground

10.6.4.2 Geometric functions

The geometric functions are related to the requirements of surface form and volume of the satellite. They are:

— To provide sufficient mounting surface for the satellite equipment (such as transponders and antennas)
— To reserve sufficient volume between the satellite and the fairing to accommodate folded appendices (such as antennas and solar generator)
— To provide sufficient accessibility of equipment during integration of the satellite
— To guarantee precise and stable location of equipment, particularly sensors and antennas
— To provide sufficient space for the radiating surfaces, which must be conveniently situated on the satellite with respect to the mounting surfaces provided for power amplifiers
— To provide a launcher and fairing interface

10.6.4.3 Other functions

The following functions are not directly mechanical or geometric:

— To provide a reference potential for the equipment
— To guarantee the same potential at different parts of the satellite in order to avoid uncontrolled electrical discharges
— To satisfy the requirements for thermal control (such as the value of thermal conductance between different points and support for isolating materials)
— To protect components from radiation and high energy particle flux

10.6.4.4 Materials used

The essential qualities of this subsystem are a trade-off between lightness and resistance to deformation. Since the forces along the principal axis of the satellite are, in principle, the greatest (this

depends on the configuration of the satellite on the launcher and the type of apogee motor), an architecture based on a central tube that contains the solid apogee motor, or the propellant reservoirs in the case of unified propulsion, is often adopted (Figure 10.1).

Current techniques permit the structure mass not to exceed about 5% of the total mass of the satellite. This is achieved by using alloys of aluminium and magnesium, honeycomb panels, bonded assemblies, and composite materials based on carbon fibres (for solar panels and antenna towers). The use of beryllium is limited on account of its prohibitive cost.

10.6.5 Conclusion

The preceding sections have defined the objectives and techniques of thermal control and its impact on the structure of the satellite. The impact of the choice of attitude stabilisation procedure on the design of the thermal control subsystem and the general organisation of the structure must be emphasised. The type of propulsion system (solid or liquid propellant apogee motor) also influences the architecture.

The problems are very different for a spin-stabilised satellite and a three-axis stabilised satellite. For example, rotation of a spin-stabilised satellite ensures uniform exposure conditions for the lateral faces of the satellite body to the various sources of radiation; this is not the case for three-axis stabilised satellites. Organisation of the structure is also influenced by rotational symmetry. The main problem with spin stabilisation resides in the limited surface to mount equipment, including solar cells and optical solar radiators. It means that only limited capacity communications missions (around 1200 W total RF power) can be aimed at. This value is significantly below today's typical mission requirements.

The area available for mounting radiating surfaces on a communications satellite is one of the factors that determines the maximum thermal power the satellite can radiate. Because of the spacecraft dimensions imposed by the launch vehicle fairing, the limited space for installing these surfaces is a limit imposed on the electrical power (and therefore the capacity) the satellite can deliver for a given RF/DC efficiency of the payload amplifiers. Use of amplifiers of higher efficiency permits this limit to be exceeded. Another option is to increase the available surface by considering deployable solar radiators connected to the source of heat by means of flexible heat pipes.

10.7 DEVELOPMENTS AND TRENDS

The development of satellite communications has been considerable during the last 30 years. This has resulted in an increase in the size and power of satellites. The Intelsat I satellite had a mass of 40 kg at the BOL, and the electrical power available on board was only 33 watts. At the end of the 1980s, Intelsat VI satellites weighed 2500 kg BOL with a power of 2200 W. This corresponds to a GTO mass of 4170 kg [NEY-90]. Then, in response to the increasing communications requirements during recent years, larger and larger communications satellites were considered. At the turn of the century, satellites launched for TV broadcasting and mobile applications had typically a 3500–5000 kg GTO mass and a 2200–3000 kg BOL mass, and delivered 8–12 kW DC power.

The aerospace industry aims to increase mass (up to 7–10 t) and power (up to 20 kW) for future high-data-rate multimedia satellites. At the same time, technological progress has permitted better use of the mass and power installed in orbit; a more sophisticated payload has provided higher ratios of communications capacity to mass and power, high-efficiency electrical supply systems (solar cells and distribution), high specific energy (batteries), higher-performance propulsion systems requiring less mass of propellant for a given velocity increment, and so on.

On the other hand, progress in technology (higher solar cell and battery efficiency, lightweight structure and antennas, higher propellant specific impulse, higher TWT efficiency, etc.) has allowed large increases in the electric power (and communications mission capacity) with limited increase in satellite mass.

Several classes of communications satellite have developed:

— Small satellites, with a BOL mass of less than 1000 kg, were able to satisfy domestic services during the 1980s.
— Medium satellites, with BOL mass in the range 1000–3000 kg, are well suited to various applications.
— Large satellites, of BOL mass greater than 3000 kg, are used for specific applications requiring a high electrical power (such as direct television broadcasting, mobile services, VSAT, multimedia, etc.).

The large majority of these are geostationary satellites, and this orbit continues to be of major use for communications services. Non-geostationary orbits (NGSOs) are also of interest for mobile satellite services (see Section 2.2).

In contrast, microsatellites (mass less than 100 kg) can provide applications such as electronic mail, data collection, and paging.

As far as medium and large satellites are concerned, the current most useful trade-off between organisation and technology for the various platform subsystems is as follows:

— Attitude control by three-axis stabilisation, using angular momentum whose orientation may or may not be controllable, associated with compensation for disturbing torques by solar wing or magnetic coil
— Unified bi-propellant propulsion system for injection into orbit and orbit control in the operational phase, with use of electric propulsion for north–south control
— Power supply system with deployable solar generator (and use of GaAs cells) and NiH_2 and Li-ion batteries
— Power distribution by regulated bus (with the possible use of a series regulator in place of the conventional shunt regulator)
— Management of telecommand, telemetry, and tracking according to a modular decentralised architecture based on a data-exchange bus
— Thermal control using networks of heat pipes to optimise the use of radiating surfaces with potential use of deployable radiators
— A structure using composite materials such as carbon fibre

These technologies are used on current platforms provided by the various manufacturers, such as Eurostar (Astrium), HS 602 (Boeing Company), A2100 (Lockheed-Martin), Spacebus (TAS), 1300 family (Space Systems Loral), etc. Experimental platforms, such as Stentor (Centre National d'Etudes Spatiales [CNES]) and ETS-VIII (Japan), incorporated advanced technologies.

The new multipurpose Alphabus platform is targeted at the high-power payload telecommunications satellite market [BER-07]. In its upper range, it will allow customers to take full benefit of the capabilities of the new generation of 5-m fairing commercial launchers, with respect to both payload volume and launch mass. At the lower end of its range, compatibility with a 4-m fairing remains achievable. A wide range of commercial payloads to provide TV broadcast, multimedia, Internet access, and mobile or fixed telecommunications services can be accommodated on the Alphabus platform.

The Alphabus contract covers the development, and qualification of a complete product line, with the following nominal capabilities:

— *Lifetime*: 15 years
— *Payload power*: 12–18 kW (conditioned power)
— *Satellite mass*: Up to 8.1 t (at launch)
— *Payload mass*: Up to 1200 kg
— *Typical payload capacity*: Up to 200 transponders, equivalent to more than 1000 TV channels (SDTV) and more than 200 000 audio channels

The Alphabus product line is designed for future growth and will be compatible in its extended version with higher payload power (up to 22 kW), higher payload dissipation, and higher payload mass (up to 1400 kg).

The Alphabus product line has the following key features:

— Structure: Central tube and additional carbon and aluminium panels
  - Section: 2800 mm × 2490 mm
  - Launcher interface: 1666 mm
— Chemical propulsion:
  - 500 N apogee engine and 16 × 10 N RCT thrusters
  - Two propellant tanks (max 4200 kg of bi-propellant)
  - Helium tanks (2 × 150 l)
— Electrical propulsion:
  - Xenon tanks (Max 350 kg)
  - PPS 1350 thrusters on thruster orientation mechanisms
— Power generation and distribution:
  - Two GaAs solar array wings with four to six panels
  - Power supply and power distribution offering both 100 V and 50 V regulated buses
  - Modular Li-ion battery
— Modular concept comprising an antenna module for easier antenna accommodation and efficient assembly and test
— Attitude and orbit control (AOCS)
  - Gyros
  - Star and sun sensors
  - Reaction wheels
— Data handling through a 1553 bus for payload

With the Alphabus product line, European industry extends its telecommunications satellite range significantly beyond the capabilities of the existing platforms, such as Eurostar E3000/Eurostar Neo [AIR-19] and Spacebus 4000/Spacebus Neo [THA-19], with respect to both maximum payload power and mass. This development has been initiated, by ESA and CNES, as a coordinated European response to the increased market demand for larger telecommunication payloads for new broadband, broadcasting, and mobile communications services.

Based on their extensive experience in this field, Airbus Space and TAS led, as co-prime contractors, a European Alphabus industrial team. The Alphabus platform has been marketed by these companies to complement their portfolio into this extended payload range. Now they have developed new platforms – Eurostar Neo and Spacebus Neo – for new satellites.

REFERENCES

[AIR-19] Airbus Space. (2019). Eurostar Series: Eurostar E3000 and Surostar NEO. https://www.airbus.com/space/telecommunications-satellites/eurostar-series.html.

[BER-07] Bertheux, P. and Roux, M. (2007). The Alphabus product line. Paper B2.4.06, presented at the International Astronautics Conference, IAC-07.

[ETSI-17] ETSI. (2017). Satellite earth stations and systems (SES); radio frequency and modulation standard for telemetry, command and ranging (TCR) of communications satellites. EN 301 926 V1.3.1.

[ETSI-18] ETSI. (2018). Satellite earth stations and systems (SES); technical analysis for the radio frequency, modulation and coding for telemetry command and ranging (TCR) of communications satellites. TR 103 956 V1.1.1.

[FRE-78] Free, B.A., Guman, W.J., Herron, G., and Zafran, S. (1978). Electric propulsion for communications satellites. In: *AIAA 7th CSSC*, San Diego, 746–758. AIAA.

[HUM-95] Humble, R., Henry, G., and Larson, W. (1995). *Space Propulsion Analysis and Design*. McGraw Hill.

[MES-93] Messerschmid, E.W., Zube, D.M., Kurtz, H.L., and Mesiger, K. (1993). Development and utilization objectives of a low power Arcjet for the P3D (Oscar) satellite. Paper 93–056, presented at the 23rd International Electric Propulsion Conference.

[MOB-68] Mobley, F.L. (1968). Gravity gradient stabilization result from the Dodge Satellite. Paper 68–460, presented at AIAA, San Francisco.

[MOR-93] Morozo, A. (1993). Stationary plasma thruster (SPT); Development steps and future perspectives. Presentation at the 30th International Electric Propulsion Conference (IEPC), Florence, Italy.

[MOS-84] Moseley, V.A. (1984). Bipropellant propulsion systems for medium class satellites. Presentation at the 10th Communication Satellite Systems Conference.

[NEY-90] Neyret, P., Dest, L., Hunter, E., and Templeton, L. (1990). The Intelsat VII spacecraft. In: *AIAA 13th International Communication Satellite Systems Conference*, Los Angeles, May, 95–110. AIAA.

[THA-19] Thales Alenia Space. (2019). Spacebus NEO: a flexible, competitive platform, compatible with all launchers. https://www.thalesgroup.com/en/activities/space/telecommunications.

[WER-78] Wertz, J.R. (1978). *Spacecraft Attitude Determination and Control*. Kluwer Academic Publishers.

11 SATELLITE INSTALLATION AND LAUNCH VEHICLES

In the previous chapters, the system was assumed to be in its nominal operating configuration with the satellite in its orbit. The types of service offered, the communications techniques used, orbits, and system components have been successively examined. System installation, which determines successful commissioning of the system, now remains to be described. The specific functions involved in launching satellites to orbits to be performed are presented in this chapter. Some characteristics of launch vehicles are also described. It can be seen that in the last 10 years, significant progress has been made in launching vehicle technologies, so that much larger spacecraft can be launched into space more quickly and economically. But there are also growing concerns that too many spacecraft have been sent to space: the number is increasing significantly, and action is required before space becomes so polluted that future space missions are endangered.

11.1 INSTALLATION IN ORBIT

11.1.1 Basic principles

Installation in orbit consists of positioning the satellite into its nominal orbit from a launching site on the surface of the earth. A launch vehicle, which may have various associated auxiliary propulsion systems, is used to inject the satellite into an intermediate orbit called the *transfer orbit*. The procedure using a transfer orbit is based on the so-called Hohmann transfer, which enables the satellite to move from a low-altitude circular orbit to a higher-altitude circular orbit with a minimum expenditure of energy [HOH-25]. The first velocity increment changes the low-altitude circular orbit into the transfer orbit, which is an elliptical one whose perigee altitude is that of the circular orbit (the velocity vector before and just after the velocity increment is perpendicular to the radius vector of the orbit), and the altitude of the apogee depends on the magnitude of the applied velocity increment. A second velocity increment at the apogee of the transfer orbit enables a circular orbit to be obtained at the altitude of the apogee (the velocity vector just before and after the velocity increment is perpendicular to the radius vector of the corresponding orbit).

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

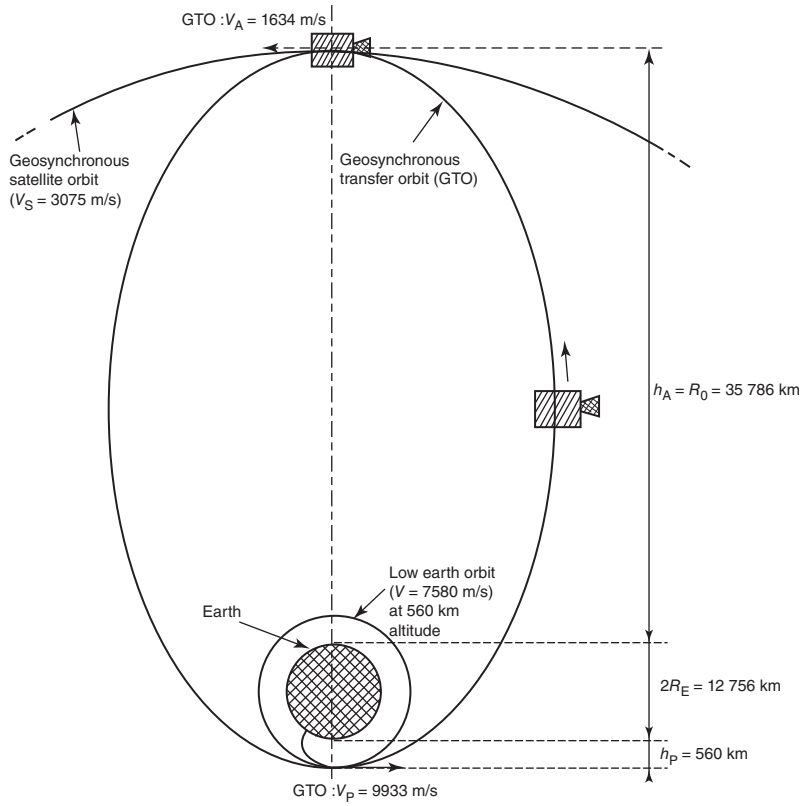


Figure 11.1 Geosynchronous transfer orbit (GTO) from a low earth orbit to a geosynchronous orbit.

Figure 11.1 illustrates this procedure for a geostationary satellite, which will serve as a reference in the following part of this chapter (most communications satellites are geostationary today; there will be more low earth orbit [LEO] and medium earth orbit (MEO) satellite constellations developed for the near future). The equatorial circular orbit at an altitude of 35 786 km is reached by way of a transfer orbit into which the satellite has been injected by the launching system. Circularisation of the orbit is achieved by means of a velocity impulse provided at the apogee of the transfer orbit.

Additional details may be provided depending on the type of launcher; the procedure is covered by one of the following methods:

— From a LEO, the satellite may be injected into the transfer orbit by means of a specific propulsion system (the *perigee stage* or the *perigee motor* depending on whether the system is independent of the satellite or an integral part of it). This procedure was used when launching a satellite with the American Space Shuttle or Space Transportation System (STS) or the Titan launch vehicle (a family of US rockets used during 1959–2005). A second velocity increment must be provided at the apogee to circularise the orbit, either by an independent *transfer stage* or by an *apogee motor* integrated into the satellite.
— The satellite may be directly injected into the geostationary transfer orbit (GTO). The launcher must communicate the appropriate velocity to the satellite at the perigee of an elliptic orbit

whose perigee altitude is that of the injection point (injection at the perigee) and whose apogee altitude is that of the geostationary satellite orbit. This is the procedure used by most conventional launchers, such as Ariane (a family of European launching vehicles in operation since 1980) and Atlas (a family of missiles and space-launching vehicles that have been developed since the late 1950s). A velocity increment must be provided at the apogee of the transfer orbit by the satellite apogee motor to circularise the orbit.

— Finally, the launcher itself can inject the satellite into geostationary earth orbit (GEO). The launcher successively provides the velocity increments required to cause the satellite (and the last stage of the launcher) to move into the transfer orbit and the velocity increment to circularise the orbit. This procedure is used by a few conventional launchers such as the Proton (a family of Russian rockets developed in 1965 and replaced by new Angara family since 2018).

Precise determination of the transfer-orbit parameters requires trajectory tracking on several successive orbits. In order to avoid excessive perturbations of successive orbits by atmospheric drag, the selected altitude of the perigee must not be below 150 km. It is generally from 200–600 km.

11.1.2 Calculation of the required velocity increments

11.1.2.1 Orbit velocity

The relation $V^2 = 2\,\mu/r - \mu/a$ permits the velocities at the perigee and apogee of the transfer orbit to be calculated (see Section 2.1.4.2), where:

— a is the semi-major axis of the ellipse.
— μ is the earth's gravitational constant ($\mu = 3.986 \times 10^{14}\ \mathrm{m^3\ s^{-2}}$).
— r is the distance from the centre of the earth to the point concerned on the ellipse, which moves with velocity V. The semi-major axis of the ellipse has a value given by:

$$a = [(h_\mathrm{P} + h_\mathrm{A})/2] + R_\mathrm{E} \tag{11.1}$$

where h_P and h_A are the altitudes of the perigee and the apogee and R_E is the terrestrial radius equal to 6378 km. Taking 560 km for the altitude of the perigee (note: Ariane 5 and Ariane 6 have set a reference orbit for the altitude of the perigee as 250 km) [AR5-16]; [AR6-18]:

$$a = (560 + 35786)/2 + 6378 = 24551\mathrm{km}$$

Hence:

— At the perigee: $r_\mathrm{P} = 6938$ km, $V_\mathrm{P} = 9933\ \mathrm{m\,s^{-1}}$
— At the apogee: $r_\mathrm{A} = 42\,164$ km, $V_\mathrm{A} = 1634\ \mathrm{m\,s^{-1}}$

11.1.2.2 Direct injection of the satellite into the transfer orbit

Most conventional launchers inject the satellite at the perigee of the transfer orbit. The launcher should thus convey the satellite to the required altitude h_P with a velocity vector parallel to the surface of the earth (which is perpendicular to the radius vector so that the injection point is the perigee).

The required injection velocity at the perigee is given by:

$$V_P = \sqrt{[(2\mu/(R_E + h_P)) - (\mu/a)]} \quad (\text{m s}^{-1}) \tag{11.2}$$

11.1.2.3 Coplanar velocity increments

If it is assumed that successive orbits are in the same plane, the velocity increment required to transfer from one orbit to another is equal to the difference between the velocity of the satellite in the final orbit and the velocity in the initial orbit.

For circularisation at the apogee of the transfer orbit, since the velocity of the geostationary satellite is 3075 m s^{-1}, the velocity increment to be provided is given by:

$$\begin{aligned} \Delta V &= 3075 - 1634 \text{ m s}^{-1} \\ &= 1441 \text{ m s}^{-1} \end{aligned} \tag{11.3}$$

The velocity increment at the perigee from a low-altitude circular orbit is calculated in a similar manner.

11.1.3 Inclination correction and circularisation

In the previous discussion, all orbit changes have taken place in the same plane. If the final orbit is the geostationary satellite orbit, the initial orbits must be in the equatorial plane. What are the parameters that determine the inclination given to orbits by the launcher? What must be done if the inclination of the transfer orbit is not zero?

11.1.3.1 Minimum inclination of the initial orbit provided by the launcher

The launch vehicle takes off from a launch base *M* and follows a trajectory in a plane that contains the centre of the earth and is characterised by the angle *A* (the launch azimuth) between the vector U, projection on the horizontal plane of the velocity vector V with a direction towards the north. The components of the *unit vectors* along OM and U are (Figure 11.2):

Unit vectors along	OM	U
Components along Ox :	$\cos l$	$-\sin l$
Components along Oy :	0	$\cos(90° - A) = \sin A$
Components along Oz :	$\sin l$	$\cos l$

where l is the latitude of the launching base. The unit vector along the direction of the vector product OM∧U is perpendicular to the plane of the orbit and has a component along Oz given by:

$$\cos i = \sin A \cos l \tag{11.4}$$

where i is the inclination of the plane containing the trajectory of the launcher.

The inclination i of this plane is thus greater than or equal to the latitude l of the launching base if the trajectory of the launch vehicle is planar; this is the usual procedure since every manoeuvre that changes the plane induces mechanical constraints and an additional expenditure of energy.

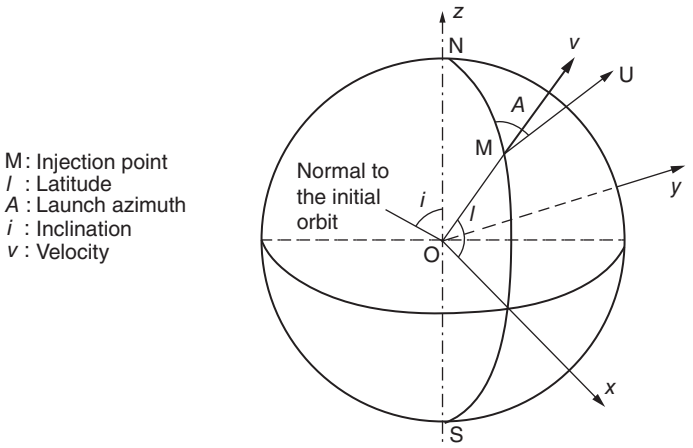


Figure 11.2 Launch azimuth A, launch-pad latitude l, and transfer orbit inclination i.

The minimum inclination equal to the latitude of the launchpad is obtained for a launch azimuth $A = 90°$ that is for a launch towards the east. A launch towards the east also enables the greatest benefit to be taken of the velocity introduced into the trajectory by the rotation of the earth. The velocities calculated for the orbits are absolute velocities in a reference fixed in space. At the instant of take-off, since the launch vehicle (and the satellite) are coupled to the rotating earth, they benefit in the plane of the trajectory from a velocity V_l induced by the rotation of the earth and equal to:

$$V_l = V_E \sin A \cos l \ (\text{m s}^{-1}) \tag{11.5}$$

where $V_E = \Omega_E R_E = 465\,\text{m s}^{-1}$ is the velocity of a point on the earth's equator, with the angular velocity of the rotation of the earth $\Omega_E = 2\pi/86\,164\,\text{rad s}^{-1}$ or $\Omega_E = 360°/86\,164\,\text{deg s}^{-1}$ (see Section 2.1.5) and $R_E = 6378$ km, the mean equatorial radius. It should be noticed that the velocity induced by the rotation of the earth can, in certain cases, be a disadvantage, particularly when polar orbits or an inclination greater than 90° (retrograde orbits) are to be obtained. It is then necessary to provide additional energy, which becomes greater as the latitude of the launch base becomes less.

Without a manoeuvre to change the plane, zero inclination would therefore require a launch base situated on the equator. Furthermore, the velocity component induced by rotation of the earth would then be a maximum. If the latitude of the launching base is not zero, the inclination of the orbit obtained is no longer zero, and an inclination correction manoeuvre must be performed. For example, for a launch from the Kennedy Space Center (KSC) at Cape Canaveral in Florida, Eastern Test Range (ETR), latitude 28°, the orbit cannot be inclined at less than 28°. For a launch from the base at Kourou in French Guyana, latitude 5.3°, the inclination cannot be less than 5.3°.

11.1.3.2 Inclination correction and circularisation procedure

Consider the transfer orbit into which the satellite is placed by a conventional launcher. The plane of the transfer orbit is defined by the centre of the earth and the velocity vector at a given instant; the inclination of this orbit is defined by the angle between the plane of the orbit and the equatorial plane (see Section 2.1.4).

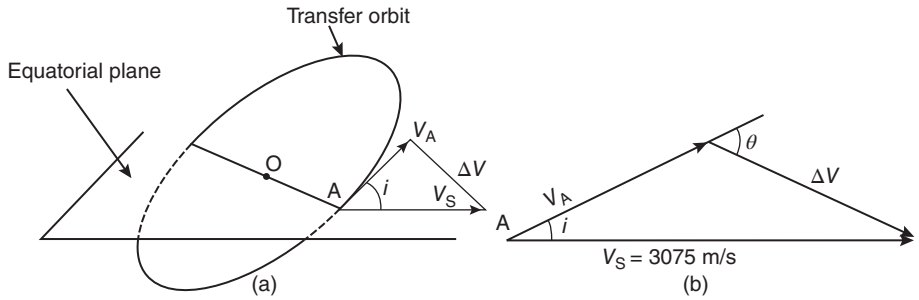


Figure 11.3 Inclination correction: (a) transfer orbit plane and equatorial plane; (b) required velocity increment (value and orientation) in a plane perpendicular to the line of nodes.

Inclination correction – that is, transferring the satellite from the plane of the transfer orbit to the plane of the equator (Figure 11.3a) – requires a velocity increment to be applied as the satellite passes through one of the nodes of the orbit such that the resultant velocity vector V_S is in the plane of the equator (Figure 11.3b). As the final orbit should be within the equatorial plane – that is, of zero inclination – the manoeuvre must be performed at the nodes. (This is not always the case when north–south station-keeping inclination corrections are performed, as described in Section 2.3.4.5; the latter can be performed away from the nodes since the orbit after correction is not necessarily of zero inclination.)

However, in some special cases (such as satellites without north–south control and satellites awaiting operational service in an inclined parking orbit; see Section 2.3.4.5), the intended orbit is not of zero inclination, and the right ascension of the ascending node is imposed. It is then necessary to change the inclination of the initial orbit by a given amount while displacing the position of the ascending node; the velocity increment is no longer applied at the node, and a special procedure is then used [SKI-86].

For a given inclination correction, the velocity impulse ΔV to be applied increases with the velocity of the satellite. The correcting operation is more economic when this velocity is low. The correction is thus performed at the apogee of the transfer orbit at the same time as circularisation. For this, the following conditions are required:

— The *perigee–apogee* line (the apsidal line) should be in the equatorial plane: that is, coincident with the line of nodes. This implies that injection at the perigee of the transfer orbit occurs on crossing the equatorial plane.
— The *apogee of the transfer orbit* should be at the altitude of the geostationary satellite orbit.
— The *thrust direction of the apogee motor* should have a correct orientation with respect to the satellite velocity vector in the plane perpendicular to the local vertical. As the apogee motor is mounted rigidly along a mechanical axis of the satellite, the orientation of this axis must be stabilised during the manoeuvre.

Taking into account the fact that V_S is nearly twice V_A, the geometry of Figure 11.3b shows that θ is approximately $2i$. For Cape Canaveral, θ is approximately 56°; for Kourou, since the nominal inclination of the Ariane transfer orbit is approximately 7°, θ has a value of around 14°. The exact value of θ is determined from:

$$\theta = \text{arc sin}\,(V_S \sin i / \Delta V)\,(\text{rad}) \tag{11.6}$$

where ΔV is the total velocity increment to be applied for circularisation and inclination correction by an amount Δi. The value of this velocity increment is given by:

$$\Delta V = \sqrt{(V_S^2 + V_A^2 - 2V_A V_S \cos\ i)}\ (\text{m s}^{-1}) \tag{11.7}$$

where V_S is the velocity of the satellite in the final circular orbit (equal to 3075 m s^{-1} for the geostationary satellite orbit) or by:

$$\Delta V = \left[\frac{\mu K}{(R_E + h_P)}\left(1 + \frac{2k}{K+1} - 2\sqrt{\frac{2k}{k+1}}\cos i\right)\right]^{\frac{1}{2}}\ (\text{m s}^{-1}) \tag{11.8}$$

where $K = (R_E + h_P)/(R_E + h_A)$, with:

h_P = altitude of the perigee
h_A = altitude of the apogee (equal to R_0 = 35 786 km)
R_E = 6378 km, the mean equatorial radius
$\mu = 3.986 \times 10^{14}$ m^3 s^{-2}, the earth's gravitational constant

Assuming that the altitude of the perigee is 560 km, $K = 0.165$, and the expression for ΔV reduces to:

$$\Delta V = \sqrt{[12.125 - 10.05\cos i]}\ (\text{km s}^{-1}) \tag{11.9}$$

It can be seen that, for i greater than 70°, the inclination correction requires an impulse greater than that required for circularisation.

11.1.3.3 Procedures based on three impulses

The Hohmann procedure is optimum for transfer between coplanar circular orbits using two velocity increments. When the ratio between the radius of the final and initial orbits is large (greater than 12), the bi-elliptical procedure, which uses three velocity increments, is more economical [MAR-79]. This also holds for smaller radius ratios (on the order of six, which are typical of GTOs) when the inclination change to be achieved exceeds 40°. Such a procedure can be considered for launch sites at high latitude; this consists of injecting the satellite into a transfer orbit whose apogee altitude is greater than that of the geostationary satellite orbit (Figure 11.4) (*supersynchronous transfer orbit*). At the apogee of this transfer orbit, a manoeuvre corrects the inclination and increases the altitude of the perigee to that of the geostationary satellite orbit. This manoeuvre requires a velocity increment less than that required for the same operation at the geostationary satellite altitude since the velocity of the satellite is lower. A final velocity increment reduces the apogee altitude to that of geostationary satellites. This procedure also reduces the propellant mass needed for apogee boost, thus increasing the lifetime of the satellite (more fuel available for station keeping) or the available payload mass, at the expense of an increase of the required launcher performance.

11.1.3.4 Procedure from an initial inclined circular orbit

When the launcher delivers the satellite into a low-altitude circular orbit, if a change of inclination is necessary, two approaches are possible with a two-manoeuvre-impulse strategy:

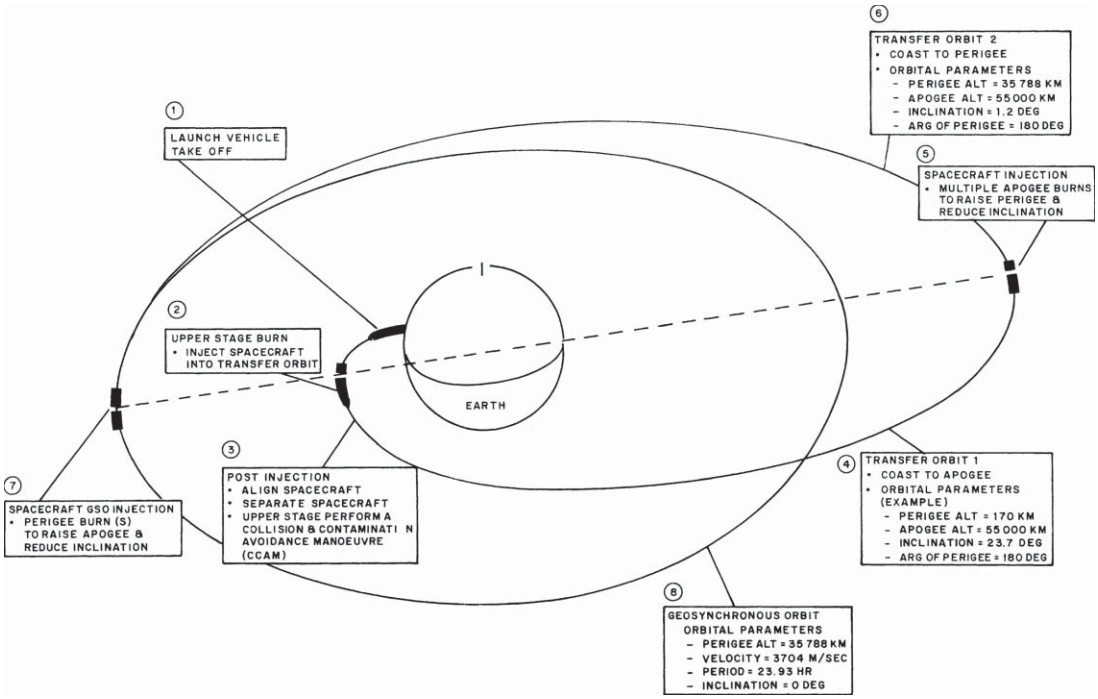


Figure 11.4 Supersynchronous transfer orbit ascent profile. Source: reproduced from [WHI-90] with the permission of the American Institute of Aeronautics and Astronautics.

— The satellite is first injected into a transfer orbit with an inclination equal to that of the initial orbit, and then the orbit is circularised and the inclination is corrected at the same time at the apogee of the transfer orbit. This method may be preferable since the velocity is lower at the apogee than at the perigee and inclination correction at the former requires less propellant. The velocity increment ΔV_P to be provided at the perigee is given by:

$$\Delta V_P = V_P - V_I (\text{m s}^{-1}) \tag{11.10}$$

where V_P is the velocity at the perigee of the transfer orbit given by Eq. (11.2) and V_I is the velocity in the initial circular orbit of altitude h_P. V_I is given by $\sqrt{[\mu/(R_E + h_P)]}$.

— The inclination correction can be shared between the velocity increment at the perigee and the velocity increment at the apogee. This procedure is called the *generalised Hohmann method*; it enables the total required velocity increment $\Delta V + \Delta V_P$ to be minimised by varying the magnitude of the inclination correction Δi_P obtained during the first velocity increment at the perigee. The velocity increment to be provided at the perigee then has a value given by:

$$\Delta V_P = \sqrt{(V_I^2 + V_P^2 - 2V_I V_P \cos \Delta ip)} (\text{m s}^{-1}) \tag{11.11}$$

The velocity increment to be provided at the apogee is evaluated using Eq. (11.7), where the inclination change is given by $\Delta i_A = i_I - \Delta i_P$, where i_I is the inclination of the initial orbit.

In the case of an initial orbit of altitude 290 km and inclination 28.5°, optimisation shows that the total velocity increment is minimised by reducing the inclination by 2.2° on injection at the perigee of the transfer orbit, which thus has an inclination of 26.3° (the altitude of the apogee is 35 786 km).

11.1.3.5 Non-impulsive velocity increments

The procedures described assume impulsive velocity increments are applied at specific points in the orbit (*impulsive* indicates short duration with respect to the period of the orbit). Example 11.1 shows that the quantity of propellant required for these manoeuvres is large and it is therefore necessary for the motor thrust to be high so that the combustion time is short (see Section 10.3 for the relations between the various magnitudes). With solid propellant motors (see Section 11.1.4.1), the impulse assumption is justified (the thrust is of several tens of thousands of newtons with a combustion time of only tens of seconds).

On the other hand, with bi-propellant motors (see Section 11.1.4.2), the thrust is limited to a few hundreds of newtons (typically 500 N), and the burn time for an inclination correction and circularisation manoeuvre at the apogee of the transfer orbit can be on the order of 100 minutes. During the burn, the satellite moves significantly in the orbit and therefore does not remain in the vicinity of the apogee; this reduces the efficiency of the manoeuvre. This loss of efficiency causes an additional quantity of propellant to be consumed in comparison with the quantity that would be required by an impulsive manoeuvre [ROB-66].

A first step in reducing the loss of efficiency is obtained by igniting the motor before the satellite reaches the apogee in such a way that the combustion time extends over a section of the orbit that is symmetrical with respect to the apogee.

Two techniques allow the loss of efficiency to be further reduced in order to approach that of an impulsive manoeuvre:

— Control of the direction of the motor thrust in such a way that it always remains parallel to the orbit as does the velocity of the satellite
— Subdividing the velocity increment into several burns

In connection with thrust orientation, two techniques can be considered during apogee motor operation:

— Conservation, with inertial axes, of the thrust direction, which remains fixed in space
— Orientation of the thrust direction with displacement of the satellite

Fixed orientation of the thrust direction in space is easily obtained by spinning the satellite about the axis along which the motor is mounted. The axis will previously have been oriented in the required direction, as defined by the angle θ given by Eq. (11.6), in the plane perpendicular to the radius vector at the apogee. At some distance from the apogee, the orientation of the motor thrust is, therefore, not that of the satellite velocity vector, and the efficiency of the thrust is reduced (the effective thrust is the actual thrust multiplied by the cosine of the angle between the thrust vector and the velocity vector).

Orientation of the thrust so that it remains aligned with the velocity vector requires active control of the satellite attitude in accordance with a particular control law. The efficiency of the manoeuvre is then increased and reaches around 99.5% with respect to an impulsive manoeuvre.

Furthermore, an efficiency approaching that of an impulsive manoeuvre is obtained by subdividing the velocity increment required for circularisation and inclination correction into several manoeuvres. Passage from the transfer orbit to the geostationary satellite orbit is thus performed by means of several increases of perigee altitude using short-duration burns as the satellite passes through the apogees of the resulting intermediate orbits. Figure 11.5 illustrates this procedure.

The advantages of multiple burns are as follows:

— Since the portion of the orbit the thrust exerted on is reduced on each manoeuvre, the efficiency of the manoeuvre is higher.
— The motor thrust can be calibrated during the first operation, thereby permitting more precise subsequent burns.
— Optimisation of successive burns can be performed by taking into account errors and variations of previous thrusts.
— By varying the amplitude and rate of burns, it is possible to combine the operations of circularisation and inclination correction with those of positioning the satellite in orbit at its station longitude (see Section 11.1.7.2). This permits the quantity of propellant consumed for these operations to be minimised.

Between each burn, the satellite performs at least two revolutions in orbit so that the orbit parameters may be determined precisely.

The number of burns is chosen in such a way that loss of efficiency with respect to an impulsive burn is limited as far as possible. With a two-burn procedure, the duration of each burn remains long and efficiency is limited. A three-burn procedure enables an efficiency of 99.75% to be achieved while allowing flexibility in the split of the total velocity increment among the three burns; this facilitates accommodation of various constraints, such as the solar aspect and visibility of stations, in the optimisation process [POC-86]. For more than three burns, the operational constraints of the control centre become prohibitive. In distributing the total velocity increment among the burns, the third one is generally chosen to be the smallest in order to reduce errors; this enables deviation from the desired orbit to be minimised. In contrast, depending on the optimisation process used and the constraints taken into account, the values of the first two increments may either be decreasing (e.g. 57% and 36% [RAJ-86]) or increasing (e.g. 33% and 45% [POC-86]).

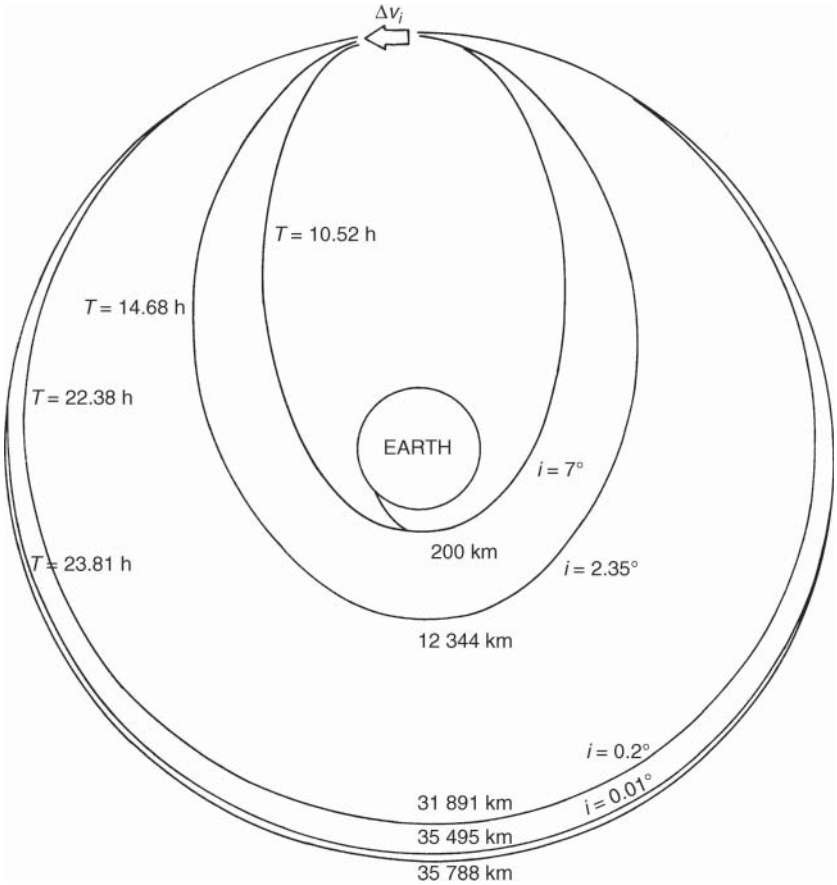


Figure 11.5 Geostationary orbit injection with apogee multiple burn strategy.

11.1.3.6 Increase of velocity at the perigee

The use of re-ignitable bi-propellant motors permits use, if necessary, of the motor for the first time on passing through the perigee to increase the velocity of the satellite (*perigee velocity augmentation* [PVA] manoeuvre).

This manoeuvre overcomes a possibly limited performance of the launcher with respect to the required launch mass of the satellite. A launcher in a given configuration is capable of placing a certain mass in an elliptical transfer orbit of nominal altitude at the apogee. If the mass is greater, the velocity on injection at the perigee will be less than the nominal velocity, and this will result in an altitude of the apogee, which is lower than the nominal one. It is then possible to use the satellite motor to provide a velocity increment that compensates for the inadequate velocity at the perigee and hence restores the altitude of the apogee to its nominal value. This is detrimental to the satellite propellant budget but can prove useful after overall optimisation to avoid the use of a more powerful (and hence more expensive) launcher when the mass of the satellite to be put into orbit slightly exceeds the nominal performance of the launcher. The performance of launchers is effectively discretised with steps that can be large (several hundreds of kilograms) depending

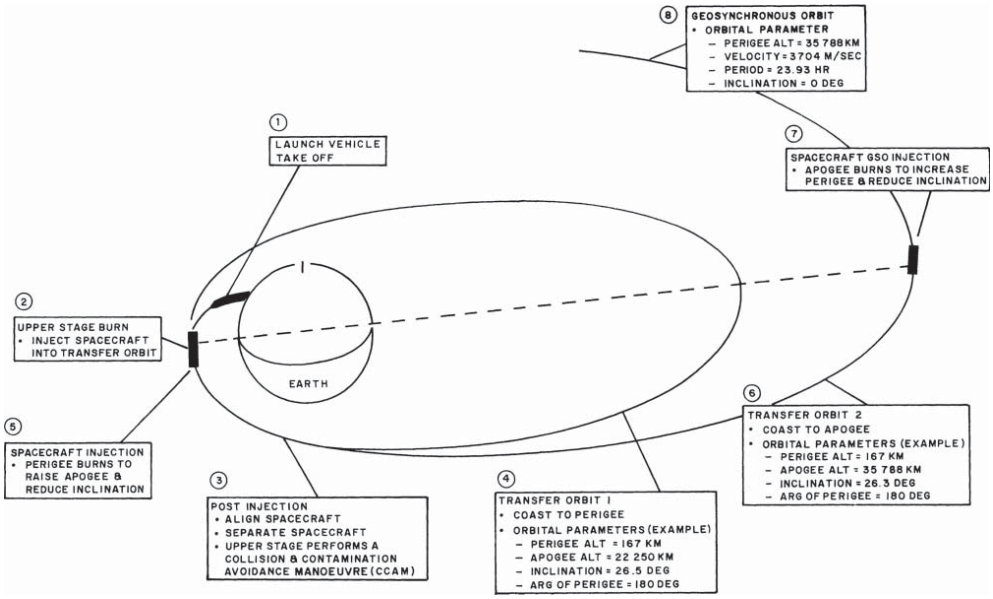


Figure 11.6 Geostationary transfer orbit ascent profile with perigee velocity augmentation. Source: reproduced from [WHI-90] with the permission of the American Institute of Aeronautics and Astronautics.

on the type or version. The procedure is illustrated in Figure 11.6 with the Delta launch vehicle [WHI-90]. Development of the family of Delta rockets for ballistic missiles started in 1960s, and they have been developed further for space-launching vehicles. The Delta Heavy IV family has been in service since 21 December 2004.

11.1.3.7 Continuous velocity increments

Electric propulsion provides a high specific impulse but a low thrust (see Section 10.3.3). It was initially implemented for station keeping, and later was also considered for orbit injection in view of the appealing perspective of saving propellant mass. However, the low thrust translates into a longer duration of the orbit-transfer manoeuvre. Several strategies can be envisaged, depending on whether the objective is to minimise the transfer duration or the propellant mass consumption. Investigations indicate that the continuous electric thrust strategy is best suited to minimising the transfer orbit duration, possibly preceded by a series of manoeuvres using chemical propulsion.

The duration of the orbit transfer between GTO and GEO can be reduced by more than 40% (at given thrust level) at the expense of a loss of the manoeuvre efficiency (67% instead of 80%). The total number of orbits needed for the transfer is reduced by more than a factor of three. This is an important result, because the effect of the radiation of the Van Allen belts can be minimised when the number of orbits crossing the proton belts is minimised.

Moderate orbit transfer duration as well as a low number of Van Allen belt crossings can be obtained when the launch vehicle has injected the satellite into a supersynchronous transfer orbit with apogee altitude of about 60 000 km. Then the satellite is gradually circularised into the geostationary orbit. Launchers with low launchpad latitude require less velocity increment and therefore lower propellant mass consumption along with lower electrical energy consumption, and a shorter transfer orbit duration compared to a launch from higher latitudes.

11.1.4 The apogee (or perigee) motor

The velocity increments required for changes of orbit, when not provided by the launcher, are provided by either the apogee or perigee motor, depending on the situation. Definitions of the characteristic magnitudes of thrusters and the relations between the mass of propellant and the velocity increment provided are explained in Section 10.3.

11.1.4.1 Solid propellant thrusters

Solid propulsion has been for many years a conventional technology for satellite apogee motors and the motors of transfer stages that are attached to them. A solid propellant motor consists of a solid mixture of oxidiser and fuel in a case of titanium or composite material (an epoxy-impregnated, Kevlar-wound shell) and a nozzle through which gases are expelled (Figure 11.7). The nozzle is usually realised in a composite material where a carbon fibre skeleton or substrate serves as a reinforcement and a carbon matrix ensures binding of the fibres. This material can support very high temperatures (3500 °C) and has good mechanical resistance and low density.

The propellant grain uses a fuel with high percentages of polybutadiene carboxide or hydroxide to which additives are added. Aluminium powder is used as an oxidiser. Thermal protection is deposited on the inner surface of the case before melting the grain in order to protect the case

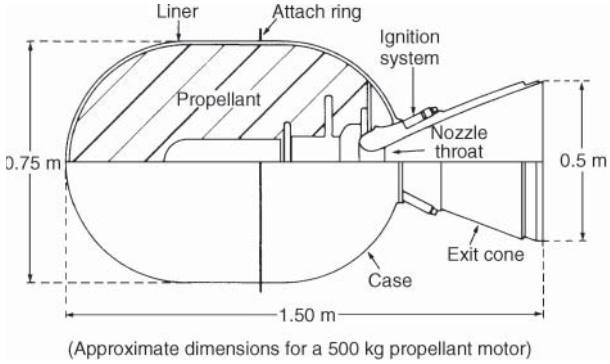


Figure 11.7 Construction of a solid propellant motor.

during combustion. The grain after solidification is machined in such a way as to provide a duct in which the combustion propagates. The form given to this duct determines the variation with time of the surface of the propellant that is available for combustion, and this in turn conditions the thrust profile of the motor. The mixture is ignited by an electrically controlled igniter located either at the extreme rear (opposite to the nozzle) or close to the throat of the nozzle.

For a case of a given size, it is possible to vary the quantity of grain contained within limited proportions. This unloading factor, on the order of 10–15%, enables the velocity increment that the motor provides to be adjusted when manufacturing the motor. The specific impulse of solid propellant motors is on the order of 290 seconds.

Examples of these motors are given in Table 11.1. Most have been developed for use on missiles.

These motors are used as apogee or perigee motors, depending on their performance in accordance with the combined (satellite–motor) mass and the velocity increment to be imparted to it. Note that the large acceleration (tens of g) due to the quick burning of a large amount of propellant generating very high thrust (tens of kN) induces significant mechanical stresses on the satellite, in particular if the satellite is large and has many appendices. Liquid propulsion with lower thrust is preferred today for large communications satellites. Solid motors are used as strap-on boosters instead, as the upper stage of conventional launchers, or to propel transfer stages.

11.1.4.2 Bi-propellant liquid motors

These motors use hypergolic (spontaneous ignition on contact) liquid propellants that generate hot gases by combustion; these expand in a convergent–divergent nozzle. Various fuel–oxidiser combinations can be used (Table 11.2, from [PRI-86]).

For satellite motors, a much-used combination is nitrogen tetroxide (N_2O_4 or NTO) as the oxidiser and monomethylhydrazine ($CH_3{\cdot}NH{\cdot}NH_2$ or MMH) as the fuel; this enables combustion temperatures on the order of 3000 °C to be obtained. Thrusts are on the order of 500 N.

The specific impulse is on the order of 310–320 seconds under nominal conditions of mixture ratio and propellant supply pressure to the motor. The *mixture ratio* is defined as the ratio of the mass of oxidiser to the mass of fuel injected into the motor combustion chamber in unit time. This ratio influences the specific impulse of the motor and must be optimised to obtain the maximum value of impulse as a function of the chemical composition of the propellants. The mixture ratio

Table 11.1 Characteristics of various solid propellant thrusters used as perigee and apogee motors

Name	Mass (kg)	Propellant (max.) (kg)	Impulse (10 N s)	Max. thrust (empty) (N)	I_{sp} (s)*
MAGE 1	368	335	0.767	28 500	287.6
MAGE 1S	447	410	1.168	33 400	290.7
MAGE 2	529	490	1.410	46 700	293.8
STAR 30B	537	505	1.460	26 825	293.1
STAR 30C	621	585	1.645	31 730	285.2
STAR 30E	660	621	1.780	35 365	290.1
STAR 31	1 398	1 300	3.740	95 675	293.2
STAR 48	2 115	1 998	5.695	67 820	290.0
STAR 62	2 890	2 740	7.820	78 320	291.2
U.T. SRM-2	3 020	2 760	8.100	260 750	303.6
Aerojet 62	3 605	3 310	9.310	149 965	286.7
STAR 63E	4 422	4 059	11.866	133 485	298
STAR 75	4 798	4 563	13.265	143 690	296.3
Aerojet 66	7 033	6 256	17.596	268 335	286.7
Minuteman III	9 085	8 390	23.100	206 400	280.6
U.T. SRM-1	10 390	9 750	28.100	192 685	295.5

*I_{sp} is sometimes expressed in Ns/kg or lbf s/lbm. The equivalences are (1 lbm = 0.4536 kg; 1 lbf = 4.448 N); I_{sp} (s) = I_{sp} (lbf sec/lbm) = I_{sp} (N s/kg) × 1/9.807.

value is on the order of 1.6. The supply of propellants under pressure is achieved by means of a pressurising gas (helium) stored under high pressure (200 bar) in a reservoir. This gas, after passing through a pressure regulator, forces the propellant from the satellite tanks into the motor under constant pressure (10–14 bar) during operation.

The use of electric pumps to feed the motor has been considered. Pumps enable the mass of the propulsion system to be reduced due to elimination of the pressurising helium reservoir and reduction of the mass of the propellant tanks. The latter need no longer support the motor supply pressure but only the pump feed pressure; this is obtained from helium, which is stored under low pressure either directly in the propellant tanks or in a small auxiliary tank. The electric power required to operate the pumps can be provided by the solar generator and the satellite batteries or by separate batteries.

The use of bi-propellant propulsion for the apogee motor (and possibly the perigee motor) leads naturally to the concept of a unified propulsion system (see Section 10.3.4.3); this permits the mass to be reduced in comparison with the combination of a solid propellant apogee motor and a propulsion system for attitude and orbit control using catalytic decomposition of hydrazine. Furthermore, the excess propellant not used during a nominal apogee thrust can be used for orbit control, thereby permitting an increase in the lifetime of the satellite.

For the transfer stages, it is advantageous to use motors of higher thrust. Thrusts on the order of a few thousand newtons (3000–12 000 N) can be obtained with an MMH/NTO combination that provides a specific impulse I_{sp} of about 320 seconds.

For these transfer stages, cryogenic propulsion using a combination of liquid oxygen (LOX) and liquid hydrogen (LH2) can also be considered, and this enables much greater specific impulses, on the order of 470 seconds, to be obtained with the disadvantage of more sophisticated technology.

11.1.4.3 Hybrid propellant motors

Hybrid thrusters can be considered with a view to their use as re-ignitable motors for the transfer stage between a low-altitude circular orbit and the geostationary satellite orbit. These consist of motors whose propellants are in different physical phase: the oxidiser is liquid, but the fuel is solid. The advantages of such a system are relatively simple design and hence moderate cost, flexibility that includes the possibility of extinguishing and reigniting the motor (a limited number of times), and high performance (I_{sp} on the order of 295 seconds). Engineering difficulties lie mainly in the problems of thermal transfer (radiation cooling) due to the long combustion times.

11.1.4.4 Electric thrusters

Electric thrusters provide a high specific impulse, which allows the use of less mass. The technology of electric thrusters is discussed in Section 10.3.3. The benefit is discussed in Section 10.3.5. The drawback is the low thrust, which translates into a longer duration of the transfer manoeuvre. Specific strategies have to be implemented (see Section 11.1.3.7).

Example 11.1 Mass of Propellant Required for Circularisation of the Transfer Orbit (Bi-Propellant Apogee Motor)

The assumption of a transfer orbit with zero inclination is made initially so that the effect of inclination on the operational mass of the satellite can be indicated in a subsequent section. The

Table 11.2 Oxidiser/fuel combinations for bi-propellant propulsion

Fuel	Oxidiser	Specific impulse (s)
(C) Hydrogen (H_2)	(C) Oxygen (O_2)	430
(L) Kerosene (RP-1)	(C) Oxygen	328
(L) Hydrazine (N_2H_4)	(C) Oxygen	338
(L) UDMH*	(C) Oxygen	336
(C) Hydrogen	(C) Fluorine (F_2)	440
(L) Hydrazine	(C) Fluorine	388
(L) 0.5 UDMH–0.5 N_2H_4	(C) Fluorine	376
(L) Hydrazine	(L) Nitrogen tetroxide (N_2O_4)	314
(L) MMH[b] (CH_3NHNH_2)	(L) Nitrogen tetroxide	328
(L) Aerozine (AZ50)	(L) Nitrogen tetroxide	310
(L) UDMH†	(L) Nitrogen tetroxide	309
(L) 0.5 UDMH–0.5 N_2H_4	(L) Nitrogen tetroxide	312
(L) UH25‡	(L) Nitrogen tetroxide	320
(L) 0.5 UDMH–0.5 N_2H_4	(L) Nitric acid	297
(L) Pentaborane	(L) Nitric acid	321

(C) means cryogenic; (L) means liquid at ambient temperature.
*UDMH: unsymmetrical dimethylhydrazine.
†MMH: monomethylhydrazine.
‡UH25: 25% hydrazine hydrate and 75% UDMH.
Source: reproduced from [PRI-86] with permission.

altitude of the perigee of the transfer orbit is 560 km, and the velocity increment to be provided at the apogee is $\Delta V = 1441\ \mathrm{m\,s^{-1}}$ (Eq. (11.3)). A satellite of mass $M_{\mathrm{GTO}} = 3000$ kg in the transfer orbit (GTO) uses a bi-propellant, liquid apogee motor that delivers a specific impulse of 310 seconds and a thrust of 400 N. The mass m of propellant required for circularisation is deduced from Eq. (10.20):

$$m = M_{\mathrm{GTO}}\left(1 - e^{-\frac{\Delta V}{g\,I_{sp}}}\right) = 3000 \times \left(1 - e^{-\frac{1441}{9.81\times 310}}\right) = 1132\ \mathrm{kg}$$

Taking into account the dry mass of the reservoirs, the motor, the pipework, and accessories, the mass of the propulsion system for the apogee manoeuvre is about 1250 kg. The mass of the satellite in geostationary orbit is equal to $3000 - 1132 = 1868$ kg.

The thrust F of the apogee motor is 400 N. From Eq. (10.16), the mass flow rate of the propellant is given by

$$\rho = \frac{F}{g\,I_{\mathrm{sp}}} = \frac{400}{310 \times 9.81} = 0.13\ \mathrm{kg\ \ s^{-1}}$$

and the duration of combustion is:

$$t = m/\rho = 1132/0.13 = 8700\ \mathrm{s} = 145\ \mathrm{min.}$$

During this period, which is long in comparison with the period of the orbit, the satellite moves with respect to the apogee, and it is necessary to subdivide the manoeuvre into several burns so that its efficiency is not excessively reduced.

The maximum acceleration Γ to which the satellite is subjected is given by:

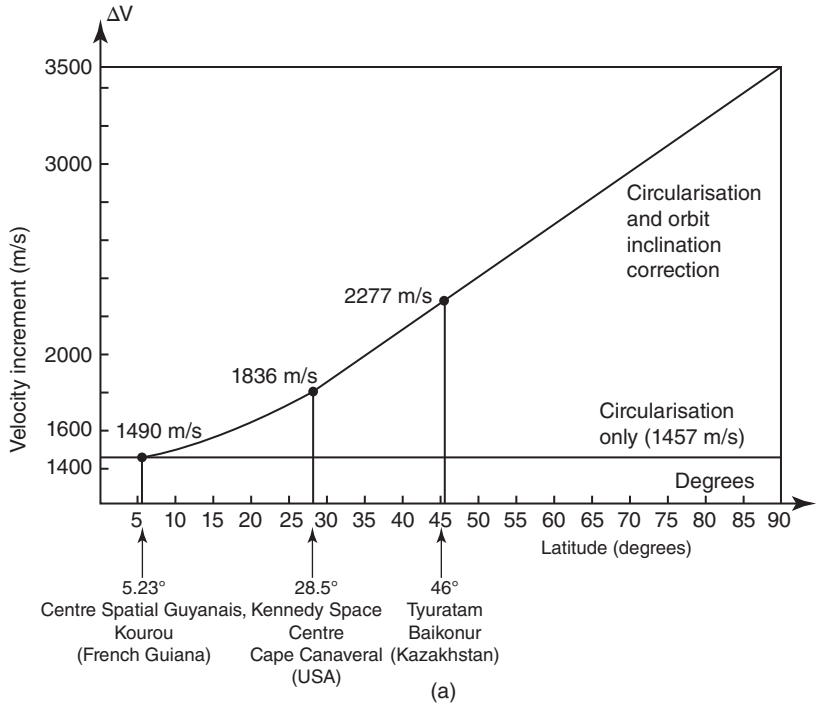
$$\Gamma = F/M = 400/1868 = 0.21\ \mathrm{m\ \ s^{-2}}$$

where M is the mass of the satellite at the end of combustion of the motor.

Compared to solid propulsion, with a motor of specific impulse $I_{sp} = 295$ s delivering a thrust of 95 000 N, the mass of propellant required is $m = 1177$ kg, the mass flow rate is $\rho = 32.8\ \mathrm{kg\,s^{-1}}$, and the combustion time is $t = 36$ s. The thrust can thus be considered an impulse. Considering a motor case of 98 kg, the mass of the propulsion system is $1177 + 98 = 1275$ kg. The acceleration imparted to the satellite reaches $95\,000/(3000 - 1177) = 52\ \mathrm{m\,s^{-2}}$ at the end of the combustion.

11.1.4.5 Influence of the latitude of the launch site on the mass

Assuming an inclination (i) of the transfer orbit equal to the latitude of the launch site (launch towards the east), the curve representing ΔV as a function of the latitude of the launch site is plotted in Figure 11.8a. Table 11.3 gives the required velocity increments and the corresponding mass of propellant for various launch sites and a satellite of mass $M_{GTO} = 3000$ kg equipped with a motor of specific impulse 310 seconds to inject the satellite into GEO from the apogee of a (200–35 786 km) transfer orbit (typical perigee altitude ranges from 200–600 km). Figure 11.8b illustrates the percentage mass reduction of the satellite in orbit as a function of the latitude of the launch site. In the case of a launch site at 28.5° latitude (Cape Canaveral), the mass lost with respect to an equatorial launch site is on the order of 12%; this emphasises the advantage of a launch site situated close to the equator such as Kourou.



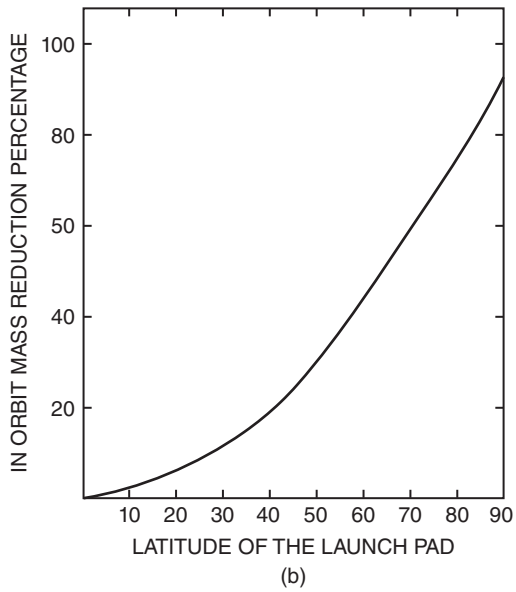


Figure 11.8 Influence of the launch site latitude: (a) required velocity increment for circularisation and correction of transfer orbit inclination against launch site latitude (perigee altitude 200 km) with a launch towards the east, ensuring an orbit inclination equal to the latitude of the launch site; (b) influence of the latitude of the launch site on the satellite mass at the beginning of life (mass at launch, 3000 kg; specific impulse, 310 seconds).

Table 11.3 Influence of the latitude of the launch site

	Kourou (France)	Cape Canaveral (USA)	Baikonur (Kazakhstan)
Latitude	5.23°	28.5°	46°
ΔV ($m\,s^{-1}$)	1490	1836	2277
Propellant mass (kg)	1162	1360	1581
Loss with respect to Kourou (kg)	0	198	419
Usable satellite mass (kg)	1838	1640	1449

11.1.5 Injection into orbit with a conventional launcher

Installing a geostationary satellite into orbit using a launcher with several stages requires three phases (Figure 11.9).

11.1.5.1 Launch phase

From launcher take-off to injection at the perigee of the GTO, the actions are as follows:

1. Increase the altitude in order to achieve the altitude of the perigee.
2. Drop the fairing after passing through the dense layers of the atmosphere.

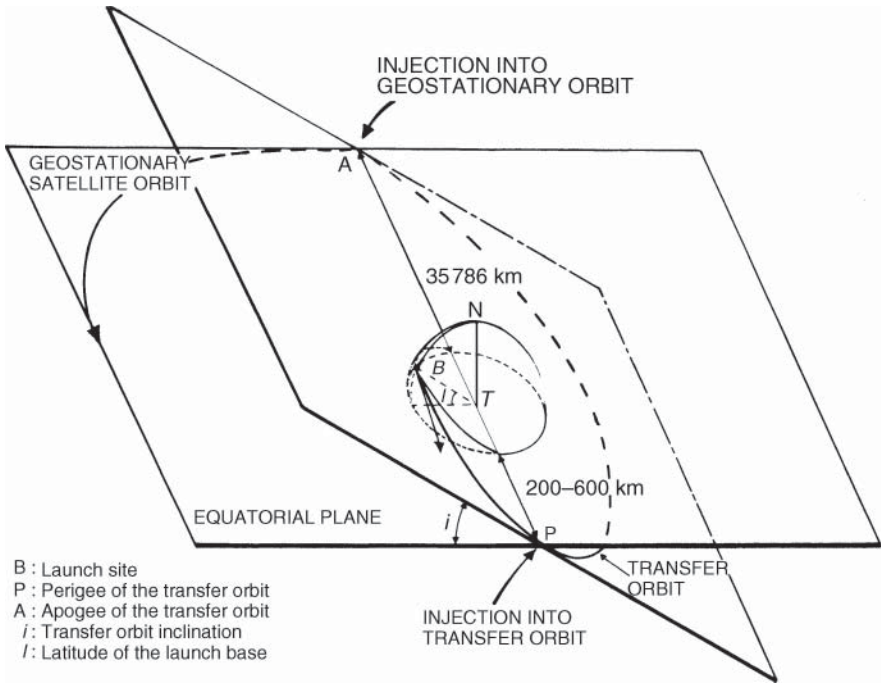


Figure 11.9 Sequence for launch and injection into transfer and geostationary orbits with an expendable launch vehicle.

3. Bring the last stage–satellite assembly onto a trajectory that intersects the equatorial plane parallel to the surface of the earth with the required velocity on passing through the equatorial plane (the perigee of the transfer orbit).

Figure 11.10 illustrates a possible strategy. This strategy – used by, for example, Delta and Atlas launchers – contains an intermediate coasting phase during which propulsion is stopped. The last stage–satellite assembly is then oriented and spun in order to maintain this orientation during ignition of the last stage. This ballistic phase is made necessary by the long distance to be covered from the launch base (Cape Canaveral) to passing through the equatorial plane; this does not permit continuous thrust of the motors on the trajectory.

In the strategy used by Ariane, the distance to be covered is shorter and the thrust is continuous (with the exception of the separation of the stages). As a consequence of the motor thrust on each phase of the flight, a given acceleration is conveyed to the launch vehicle at each instant as a function of the mass, and the velocity reached is determined by the length of the trajectory. Optimisation of performance leads to a trajectory that rises above the perigee altitude, and injection is performed at a point on the transfer orbit beyond the perigee (guidance having orientated the velocity vector in the appropriate direction).

It should also be noted that the argument of the perigee of the transfer orbit to be obtained is not 180° but about 178°. This is to take into account the drift of the argument of the perigee by 0.817° d^{-1} (see Section 2.3.2.3). It will cause the orbit to rotate in its plane in such a way that at the time of the nominal manoeuvre at apogee number 6 (after 5.5 orbits: that is, about two days),

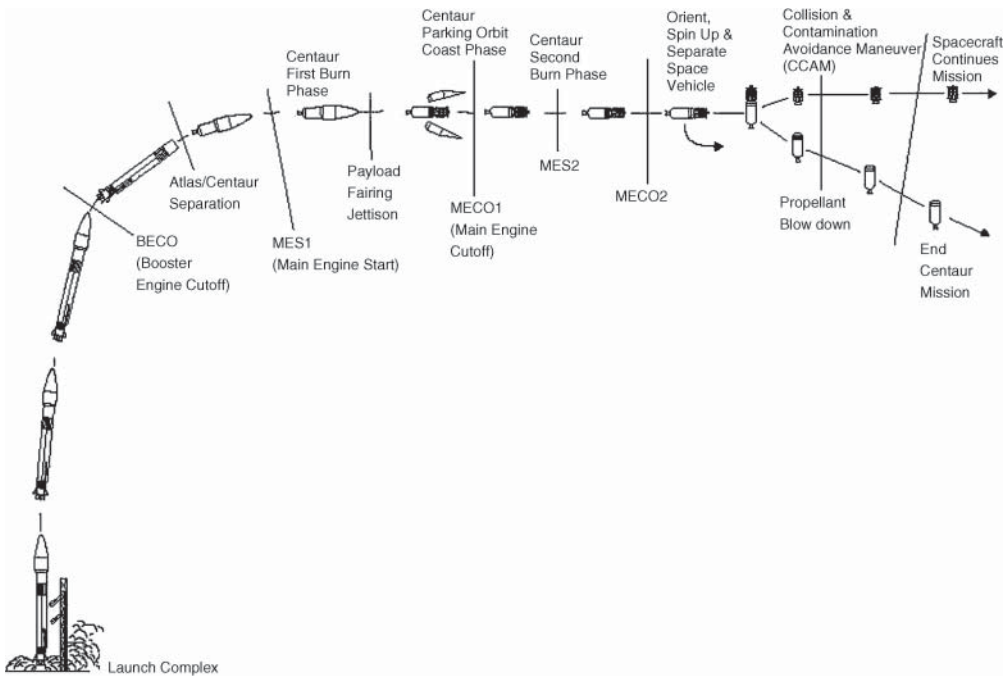


Figure 11.10 Mission profile of an expendable launch vehicle (Atlas) with a coasting phase for a typical GTO mission.

it will coincide with the ascending node and hence will be in the equatorial plane (the argument of the perigee is then equal to 180°).

11.1.5.2 Transfer phase

The transfer phase starts with injection of the composite satellite–launcher final stage and terminates with injection into the quasi-geostationary satellite orbit at the apogee of the transfer orbit. In this phase, the actions are as follows:

1. Separate the satellite and the final stage.
2. Determine orbit parameters.
3. Measure the satellite attitude.
4. Correct the satellite orientation in view of the apogee manoeuvre.

The orientation may be maintained either by causing the satellite to rotate (spin stabilisation) or by active attitude control using sensors and actuators (three-axis stabilisation).

It is important to obtain a transfer orbit whose parameters are close to those of the nominal orbit; the altitude of the apogee must be that of the geostationary satellite orbit, and the apsidal line must be in the equatorial plane. If not, then the satellite, after injection into the quasi-geostationary satellite orbit, will have to correct the nonzero eccentricity and inclination of the orbit using its actuators, hence using on-board propellant, which reduces the satellite lifetime.

11.1.5.3 Positioning phase

This phase starts with injection into the quasi-geostationary satellite orbit at the apogee of the transfer orbit and terminates with positioning of the satellite at the chosen station in the geostationary satellite orbit.

11.1.5.4 Other procedures

Other procedures for installation into orbit that permit the launch vehicle to move out of the plane defined by the launch azimuth and the latitude of the launch base (*dog leg* manoeuvres) can be envisaged. In this way, by reigniting the last stage in flight, the Proton launcher permits direct injection of the satellite into the geostationary satellite orbit. This approach avoids the need for an apogee motor and thus permits launching of satellites with larger useful mass.

11.1.6 Injection into orbit from a quasi-circular low altitude orbit

The operations involved in putting a geostationary satellite into orbit from a quasi-circular low-altitude orbit (see Figure 11.11) differ from those described earlier due to the inability of the launcher to inject the satellite directly into the transfer orbit. This is the situation with launches using the Space Shuttle, which has a circular nominal orbit of altitude 290 km and inclination 28.5°. It is also the case with two-stage launchers, such as the Titan launch vehicle (a family of US rockets from 20 December 1959–19 October 2005), which delivers its payload into an elliptical orbit (148 km × 259 km) with an inclination of 28.6°.

The changes with respect to the operations described earlier are as follows.

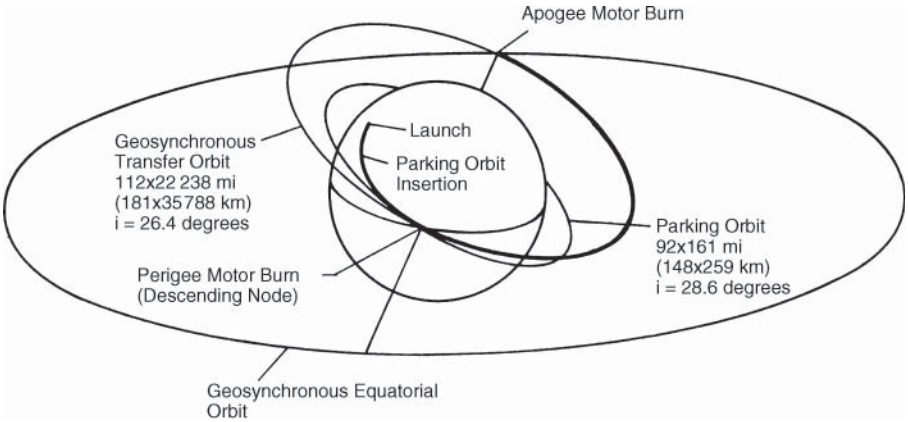


Figure 11.11 Geostationary orbit injection sequence from a low-altitude parking orbit.

11.1.6.1 Launch phase

The launch phase does not lead directly to injection of the satellite at the perigee of the transfer orbit but places the launch vehicle in a circular or slightly elliptical orbit of nonzero inclination.

11.1.6.2 Injection into the transfer orbit

After separation from the launcher into a particular attitude configuration, a velocity increment is imparted to the satellite in order to inject it at the perigee of the transfer orbit. This increment is provided by the perigee motor on passing through the equatorial plane.

The perigee motor can be integrated into the satellite or can be part of an auxiliary stage to which the satellite is attached. An integrated perigee motor can be implemented for this use (usually a solid fuel motor), with possible separation after use. This motor (in the case of a bi-propellant motor) can also be reused to provide part or all of the velocity increment at the apogee.

An auxiliary stage can merely fulfil the injection function at the perigee (the perigee stage, usually based on a solid fuel motor) or it can fulfil both the functions of injection at the perigee and injection at the apogee (the transfer stage). The transfer stage can use solid fuel motors (two are needed), bi-propellant propulsion with a reusable motor, or a combination of two technologies – a bi-propellant apogee stage associated with a separable solid fuel perigee stage.

During injection into the transfer orbit, orientation of the perigee motor thrust is ensured either by causing the satellite to rotate (spin stabilisation) or by active attitude control using sensors and actuators (three-axis stabilisation). Attitude control of the satellite–stage combination can be ensured by the satellite attitude control subsystem (in the case of an integrated motor) or by a subsystem that is specific to the auxiliary stage.

11.1.6.3 Transfer and positioning phases

These phases are similar to those described earlier, knowing that certain operations may possibly be realised by the transfer stage.

11.1.7 Operations during installation (station acquisition)

Installation consists of transferring the satellite from the transfer orbit provided by the launch vehicle to the longitude of its station. Installation thus covers the transfer and positioning phases specified previously. It also covers the operations of configuring the satellite to perform the mission for which it has been designed.

11.1.7.1 Choice of apogee where the manoeuvre is performed

The choice is influenced by various considerations:

— The satellite must perform at least one or two transfer orbits to permit orbit determination with adequate accuracy.
— The satellite must not remain too long in the transfer orbit since the space environment there is special and different from that of the final orbit for which the satellite has nominally been designed. In particular, the satellite is subjected to more numerous eclipses and on each orbit passes through a region of trapped particles (the Van Allen Belt; see Section 12.4). Finally, the electrical energy resources, which consist of batteries and possibly a partially deployed solar generator, are limited.
— Ignition of the apogee motor must be performed within the visibility of at least two control stations in order to increase the chances of success of this critical manoeuvre by means of redundancy.
— The apogee chosen must be close to the longitude of the desired station. After the apogee manoeuvre, the satellite is quasi-geostationary, close to the position of the apogee where the manoeuvre was performed.

Figure 11.12 shows the track of the satellite in the transfer orbit (typical of a launch by Ariane) and indicates the position of successive apogees. In this way, the number of the transfer orbit on which the manoeuvre must be performed in order to satisfy the constraints of visibility and proximity to the station is determined. For the launch of a satellite intended to be positioned in the vicinity of the Greenwich meridian, the nominal choice is Apogee 4 with a second chance at Apogee 6 in the case of a problem. These apogees are visible from the control stations at Toulouse and Kourou.

11.1.7.2 Drift orbit

Due to variations of the parameters of the transfer orbit and the apogee manoeuvre, and also in order to permit the satellite to achieve the longitude of its station, the orbit obtained after the apogee manoeuvre is never exactly the geostationary satellite orbit. A residual nonzero inclination and eccentricity exist, and the semi-major axis differs from that of the synchronous orbit, and this leads to drift of the satellite. The drift orbit attained in this way must be corrected by means of low-thrust thrusters from the satellite orbit control system in such a way that, after a certain time (several days), the satellite reaches the intended station longitude on an orbit of the desired eccentricity and inclination (not necessarily zero; see Section 2.3.4.5).

When a multiple-burn procedure is used, the drift orbit is generally included in the optimisation of the injection procedure, which takes the satellite to its station longitude by taking into account inaccuracy of the transfer orbit provided by the launcher (see Section 11.1.3.5).

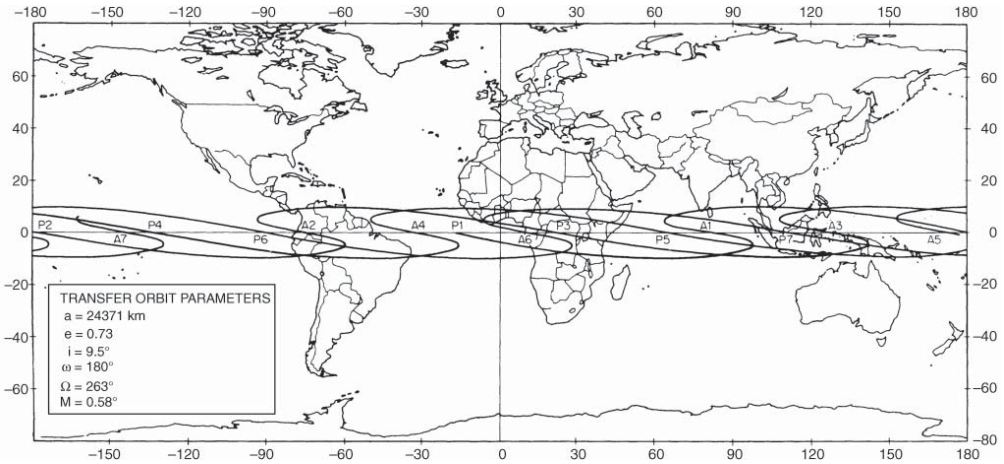


Figure 11.12 Satellite track during transfer orbit (Ariane launch) showing successive locations of the apogee (A) and perigee (P).

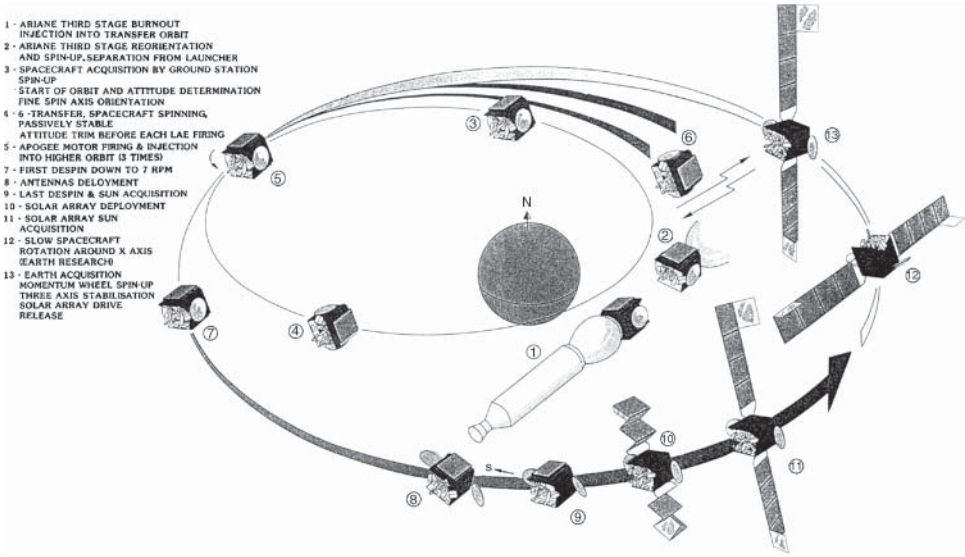


Figure 11.13 Example operations during transfer and drift orbits.

In addition to corrections of the orbit, the main operations to be performed are as follows (Figure 11.13):

1. Reduction of the velocity of rotation in the case of attitude control by spin during the apogee thrust.
2. Acquisition of the sun by the solar sensors (already achieved during the transfer orbit with three-axis control).
3. Attitude control.
4. Deployment of solar panels and activation of the panel rotation system to follow the apparent movement of the sun.
5. Acquisition of the earth (using an infrared sensor).
6. Establishment of the desired attitude using information from the earth and solar sensors.
7. Activation of the momentum wheels.

The positioning phase lasts for several days after the apogee manoeuvre.

11.1.7.3 Satellite test and acceptance

Once the satellite is at the intended station longitude in its nominal orbit, station-keeping is activated. The operational life of the satellite starts when various satellite subsystem tests have been performed and their performance has been evaluated (in-orbit testing [IOT]). This test and acceptance phase lasts for several weeks.

11.1.8 Injection into orbits other than geostationary (non-GEO orbits)

In the preceding discussion, the final orbit to be attained was geostationary, which is the orbit used by the majority of communications satellites. However, as seen in Chapter 2, other orbits

have useful properties. A large numbers of LEO satellite constellations have been developed in the recent years and will be launched into space from 2020. The procedures used depend on the type of orbit to be attained.

11.1.8.1 Injection into polar orbit

Polar orbits are of interest for communications as the satellites are visible for a given period of time from any location at the surface of the earth. A constellation of satellites with appropriate phasing (several satellites on the same orbit and different orbit planes with evenly distributed right ascension values) allows continuous coverage of the earth (e.g. the Iridium system). Some applications of sun-synchronous orbits (SSOs) have been proposed. Injection of the satellite into an orbit with the desired inclination is achieved by choosing the launch azimuth as a function of the latitude of the launch site according to Eq. (11.4). For low-altitude orbits (several hundreds of kilometres), direct injection of the satellite into the final orbit is possible. For high-altitude circular orbits (several thousands of kilometres), an intermediate transfer orbit must be used following the principle described in Section 11.1.1.

11.1.8.2 Injection into inclined elliptic orbits

The procedure used depends on the mass of the satellite to be placed in orbit and the capacity of the launcher.

For a satellite of high mass using the full capacity of the launcher, an orbit of the desired inclination is obtained by adjusting the launch azimuth in the context of a dedicated launch. Depending on the characteristics of the final orbit, the satellite is injected into either a transfer orbit or the final orbit.

For a satellite of limited mass that does not use the full capacity of the launcher, sharing of this capacity among several satellites permits the launching cost to be reduced. It is, however, unlikely that the intended orbit of both satellites carried by the launcher will have the same desired inclination; most opportunities for multiple launching that arise concern missions in GTO. It is thus necessary to consider procedures that permit the inclined elliptic orbit to be joined from the standard GTO provided by the launcher, by modifying the inclination and raising the altitude of the perigee and the apogee.

For example, to obtain an orbit of the Tundra type (see Section 2.2.1.2), procedures with two or three velocity increments can be considered as follows:

— With the two-velocity increment procedure, the first impulse modifies the inclination from that of the transfer orbit (e.g. 6°) to 63 : 4°, and a second impulse raises the perigee altitude to 22 000 km.
— In the three-increment procedure, the first impulse at the perigee of the transfer orbit greatly increases the altitude of the apogee (for example, to 100 000 km). At the apogee, the second impulse modifies the inclination at least cost due to the low velocity of the satellite (see Section 11.1.3.3). The third impulse enables the final orbit to be attained.

Optimisation shows that the total velocity increment to be provided is less with the three impulse procedure. To place a satellite in an orbit of the Tundra type from the standard GTO transfer orbit, the total increment to be provided is on the order of 2300 m s^{-1} instead of 2500 m s^{-1} with the two-impulse procedure.

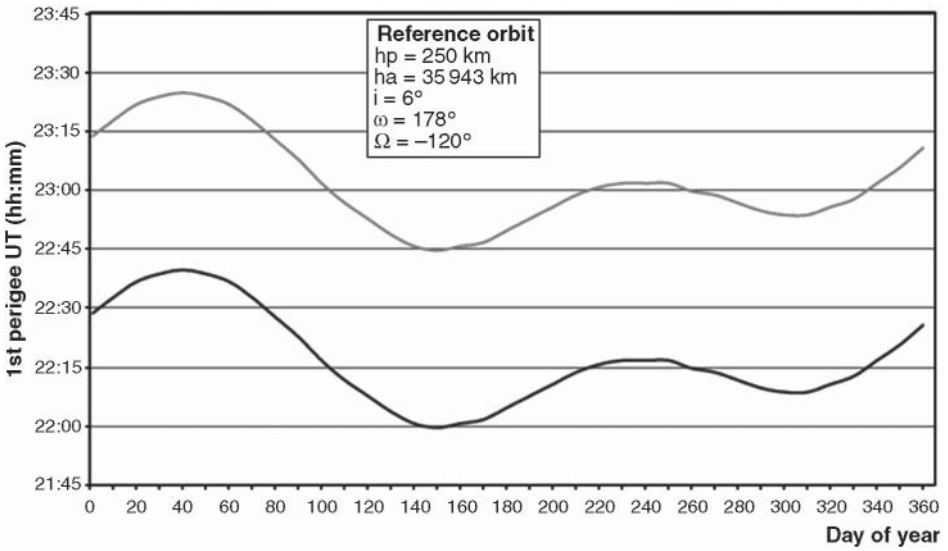


Figure 11.14 Typical launch window with Ariane 5 dual launch.

11.1.9 The launch window

The launch window specifies the time periods when launching a satellite is possible, taking into account the following associated constraints:

— Enabling determination of the attitude with the required accuracy. This relates to permissible ranges of the direction angles of the sun and the earth with respect to the satellite axes, considering that the earth, the satellite, and the sun must not be aligned.
— Avoiding saturation of the sensors or disappearance of references during the apogee manoeuvres. This relates to the position of the satellite with respect to the sun and the position and duration of eclipses.
— Ensuring an electric power supply. This relates to the position of the satellite with respect to the sun and the position of eclipses.
— Guaranteeing thermal control. This relates to the position of the satellite with respect to the sun and the number and duration of eclipses.
— Being within radio visibility of the control station during the critical phases. This relates to satellite-to-station geometry and solar radio disturbances.

From a combination of the various constraints, possible time slots for launching the satellite, or launch windows, can be deduced for all days of the year. In the case of a double launch, the launch window must satisfy the constraints imposed by installation of both satellites. Figure 11.14 illustrates such a launch window for a double launch (constraints should be fulfilled for both satellites) with the Ariane 5 launch vehicle [AR5-16].

11.2 LAUNCH VEHICLES

To be free of the constraints imposed by the only country selling launchers at the start of the 1980s (the United States, with the Delta and Atlas Centaur launchers), some industrialised countries

such as Russia, China, Japan, and Europe, all developed their own launcher programmes that enable satellites to be placed in geostationary orbit, followed by other countries such as Brazil, India, Israel, and South Korea.

At the same time, the United States was engaged in a programme whose objective was to develop a recoverable and reusable launcher (the Space Shuttle's first flight was 12 April 1981), with a view to reducing the cost of installation into orbit and with the idea of abandoning the production of conventional launchers. The Challenger Space Shuttle catastrophe on 28 January 1986, which led to immobilisation on the ground of US civil and military satellites for more than two and a half years, caused a revision of this policy. The production and development of conventional launchers was consequently restarted. Furthermore, the Reagan administration decided that use of the Space Shuttle would be reserved for government missions. It also proposed that companies developing conventional launchers should commercialise their launching services, and the installations at various launch sites that are government property (Air Force and NASA) were put at their disposal [STA-88]. The last flight of the Space Shuttle was on 21 July 2011.

In 1983, before the end of the Cold War, the Soviet Union proposed the service of its Proton launcher to the Western world to launch the Inmarsat II satellites.

Table 11.4 summarises the main characteristics of launchers that are operational, or nearly so, starting from the first decade of the twenty-first century.

Finally, there has been growing demand for suitable launchers for small satellites since 2000. They include adaptors for powerful launchers (Ariane structure for auxiliary payloads [ASAP], APEX with Ariane), the development of dedicated launchers (Pegasus, Taurus), and numerous other projects whose completion has been hectic. Great capabilities of space launchers have been developed quickly since then.

11.2.1 Brazil

The Brazilian Space Agency (AEB in Portuguese) was created in 1994 with the purpose of launching satellites into LEO. The development of the Veículo Lançador de Satélites (VSL) 1 started in 1984 after the success of the sounding rockets Sonda 3 and 4. Based on the Sonda 3 and 4, the 19.5-meter three-stage, solid propellant vehicle must be capable of placing a 120 kg satellite into LEO. While working on the VSL-1, the AEB is also developing the VSL-2 (Alfa project), a medium-sized launch vehicle capable of placing satellites into GTO.

11.2.2 China

In 1986, China decided to commercialise its Long March launchers. The launch vehicles are designed by the China Aerospace Science and Technology Corporation (CASC) and commercialised by the Great Wall Industry Corporation. A series of Long March launch vehicles have been developed for missions to LEO, GTO, SSO, and other orbits. The characteristics of the main launch vehicles are given in Table 11.5 for the earlier development of the launching vehicles. For comparisons with other countries in the recent development, refer to Table 11.4.

The LM-2C launch vehicle is a two-stage launch vehicle mainly used to launch recoverable satellites since its first flight on 9 September 1982. The maximum diameter of its core stage is 3.35 m, and the LEO payload capacity is 3366 kg. An increase in the performance of the LM-2C by the addition of four strap-on boosters (LM-2E) raises the launcher's performance to 9.5 tons in LEO and 3.5 tons in GTO.

The LM-3A launch vehicle is a large, three-stage, liquid-propellant launch vehicle with an LH2–LOX third stage. It is 52.52 m long and 3.35 m in diameter and has a GTO capability of 2650 kg. LM-3A made its maiden flight on 8 February 1994, sending two satellites into orbit

Table 11.4 Operational launch vehicles around 2010

Launch vehicle	First launch (success)	Lift-off mass (tons)	Size (m)	Launch site	LEO performance			GTO performance	
					Altitude (km)	Mass (kg)	Inclination (degree)	Mass (kg)	Inclination (degree)
CHINA									
LM-2E	1982	460	49.7	Xichang	200	9 500	28.2°	3 500	28.2°
LM-3B	1990	425	54.8	Xichang	200	14 000	28.2°	5 100	28.2°
LM-3B(A)	2002	670	62	Xichang	—	—	—	7 000	28.2°
LM-4B	1999	264	45.8	Xichang	750	2 200	98°	—	—
LM-5	2013	643	60	Wenchang	200	25 000	52°	14 000	28.2°
LM-6	2015	103	29	Taiyuan	SSO:700	1 080	—	—	—
LM-7	2016	594	53	Wenchang	SSO:700	13 500	—	5 500	—
LM-11	2015	58	20.8	Jiuquan	LEO: — 700	700	—	—	—
					SSO:700	350	—	—	—
CIS									
Cosmos 3	1962	120	32	Plesetsk	400	1 400	63°	—	—
Soyuz Fregat	2000	305	46.6	Baikonur	1 400	4 000	51.8°	—	—
Soyuz/ST	2004	305		Baikonur	800	2 900	98°	—	—
Proton M Breeze	2001	690	53	Baikonur	200	21 000	51.6°	GTO: 5 500 GEO: 2 920	51.6°
Zenit-2	1985	445	57	Baikonur	200	13 740	51.4°	—	—
Zenit-3SL	1999	462.2	59.6	Offshore	?	6 100	?	GTO: 6 000 GEO: 1 840	?
Dnepr	1999	209	34.3	Baikonur/Yasny	200	4 500	46.2°	—	—
Rockot	2000	107	29	Plesetsk	400	1 900	63°	—	—
Angara 1.1	2010	149	34.9	Plesetsk	—	2 000	63°	—	—
Angara 1.2/A5	2014	171.5–790	64	Plesetsk	—	3 800–245 000	63°	GTO: 5 400–7 500 GEO: 5 000	63°

(*Continued*)

Table 11.4 (*continued*)

Launch vehicle	First launch (success)	Lift-off mass (tons)	Size (m)	Launch site	LEO performance			GTO performance	
					Altitude (km)	Mass (kg)	Inclination (degree)	Mass (kg)	Inclination (degree)
EUROPE									
Ariane 44 L	1988	470	58.4	Kourou	800	6 500	98.6°	4 900	7°
Ariane 5 G	1998	750	59	Kourou	800	9 500	98.6°	6 640	7°
Ariane 5 ECA	2002	780	59	Kourou	Escape	5 200	(to Mars)	10 500	7°
Ariane 5 ES-	2008	760	59	Kourou	1 000	21 000	51.6°	—	—
Vega	2012	137	30	Kourou	700	1 430	90°	—	—
					SSO:400	1 450	—	—	—
					1 500 × 200	1 963	5.4°	—	—
Ariane 62	2020	530	63		LEO	10 350	—	GTO	500
					SSO	6 450	—		
Ariane 64	2020	860	63		LEO	21 650	—	GTO	11 500
					SSO	14 900	—	GEO	500
INDIA									
PSLV	1994	294	44	Sriharikota	400	2 900	98°	450	18°
GSLV	2001	402	49	Sriharikota	—	5 000	—	2 500	18°
GSLV-II	2010	415	49	Sriharikota	—	5 000	—	2 700	—
GSLV-III	2014	640	43.4	Sriharikota	600	8 000	—	4 000	—
ISRAEL									
Shavit	1988	30	18	Palmachim	366	225	143°	—	8°
JAPAN									
MV	1997	140	30.7	Kagoshima	250	1 800	31°	—	—
H-IIA 202	2001	285	53	Tanegashima	250	10 000	30°	4 000	30°
H-IIB	2009	531	56.6	Tanegashima	—	16 500	—	8 000	—
H-3	2020	574	63	Tanegashima	—	—	—	6 500	($T = 1.5\,km\,s^{-1}$)
SOUTH KOREA									
KSLV-I	2009	140	30	Goheung	300	100	38°	—	—
KSLV-II	2017	200	47.2	Goheung	800	1 500	97°	—	—

USA									
Taurus	1994	68.4	27.9	Canaveral	400	1 300	28.5°	510	28.5°
Delta II (7925)	1994	231	30	Canaveral	185	4 970	28.7°	1 800	28.7°
Delta III	2000	301		Canaveral	185	8 290	28.7°	3 810	28.7°
Delta IV Medium	2002	256	62	Canaveral	185	8 120	28.7°	4 210	28.7°
Delta IV Med +	2003	325	68	Canaveral	185	11 475	28.7°	6 565	28.7°
Delta IV Heavy	2004	733	72	Canaveral	—	28 790	—	14 220	—
Atlas II AS	1992	237	46	Canaveral	185	8 618	28.5°	3 720	28.5°
Atlas III A	2000	218	52.8	Canaveral	185	8 660	28.5°	4 060	28.5°
Atlas III B	2002	226	55.5	Canaveral	185	10 500	28.5°	4 500	28.5°
Atlas V 401	2007	333	62.2	Canaveral	—	—	—	4 950	28.5°
Atlas V 521	2008	380	62.7	Canaveral	—	20 520	—	8 900	—
Falcon 1	2008	27	21	Omelek	185	420	—	—	—
Falcon 9	2015	549	71	Omelek	—	22 800	28.5°	8 300 (Mars: 4 020)	27°
Falcon Heavy	2018	1 420	70	Omelek	—	63 800	28.5°	26 700 (Mars: 4 020) (Pluto: 3 500)	27°
Pegasus XL	1994	23.1	17.6	Aircraft	400	550	70°	—	—
Sea Launch	1999	466	65	Offshore	—	—	—	5 250	0°

Table 11.5 The Long March family of launch vehicles

	LM-2C	LM-2E	LM-2D	LM-4	LM-3A	LM-3B	LM-3B(A)
Stages	2	2	2	3	3	3	3
Strap-on boosters	0	4	0	0	0	4	4
Length (m)	40	49.7	37.7	45.8	52.5	54.8	62
Diameter (m)	3.35	3.35	3.35	3.35	3.35	3.35	3.35
Mass (ton)	213	460	232	250	241	426	670
Lift-off thrust (MN)	2.96	5.89	2.25	2.96	2.94	6.04	9.06
Payload mass (ton)	3.3*/1.4†	9.5*/3.5†	3.7*	2.8‡	2.6†	5.2†	7†
Year of first flight	1982	1990	1992	1988	1994	1996	2002

*Low earth orbit.
†Geostationary transfer orbit.
‡Sun-synchronous orbit.

successfully. The LM-3B launch vehicle is a large launch vehicle formed by strapping four boosters onto the LM-3A and has a GTO capability of 5200 kg. It is primarily designed to launch geostationary satellites. The LM-4 launch vehicle is a three-stage liquid-propellant launch vehicle mainly used to launch satellites into SSO. It is 45.8 m long and 3.35 m in diameter. Its payload capacity to a 748 km, 98°-inclined orbit is 2790 kg. High performance and low cost are proposed, with a core stage (lower composite) 5 m in diameter. The lower composite will be a 1.5-stage rocket with all the engines fired on the ground and will be used to launch LEO payloads. By combining the lower composite with an upper stage, a 2.5-stage rocket can be formed, which will be mainly used to launch payloads to GTO and SSO. By using a different number of strap-on boosters on the basic 5 m diameter core vehicle, the new launch vehicle series will have a capability in the range 10–25 tons in LEO. A combination with an upper stage will put its GTO capability to 6–13 tons.

The main launch base is situated at Liangshan near Xichang in the province of Sichuan (latitude 28.2 °N) where the construction of a second launching assembly is planned. China also has an older site, the Jiuquan Satellite Launch Center, near Jiuquan (latitude 41 °N) in the province of Gansu at the edge of the Gobi Desert.

In October 2007, China started to build a new production facility for a Long March 5 (LM-5) vehicle near Binhai New Area in Tianjin; and a new launch facility in Wenchang on Hainan Island, South China. Tianjin has a port that allows the large rocket to be transported by sea to the launch facility on the island. With four boosters and a height of 60 m, this launcher weighs about 650 tons. It is capable of delivering 1.5–25 ton payloads to LEO or 1.5–14 ton payloads to GTO, thus providing launch performance similar to the Delta IV Heavy, Atlas V, and Ariane 5. The total investment was 4.5 billion Yuan ($657 million).

By the end of 2008, the Long March rockets had carried out 113 missions. Even if Chinese launch vehicles do not have the commercial past and the successes of US or European launchers, the recent success of sending the first Chinese man into space and unmanned vehicles landed on the moon surface have consolidated the image of the Chinese space market and attract new customers. New China will try to become a major player in world space market with the success of LM-5/6/7 and LM-11, in addition to space exploration applications and scientific research.

11.2.3 Commonwealth of Independent States (CIS)

The former USSR had various series of launchers, each containing diverse models of varying capacity and use. Several launchpads are available. The oldest, Tyuratam (or Baikonur),

46 °N latitude, is situated 370 km to the southwest of the town of Baikonur in the republic of Kazakhstan. It was used to launch the first artificial satellite, Sputnik 1.

An active complex is that of Plesetsk (near Arkhangelsk), close to the border with Finland (62.7 °N). A third complex is Kapustin Yar (48.6 °N).

11.2.3.1 Soyuz

Soyuz is part of a series, which served to launch the first Sputniks, consisting of two- and three-stage launchers assisted by four additional boosters arranged obliquely. The propellants used are kerosene and LOX (hydrazine and LOX for the third stage). These launchers are assembled horizontally.

Version A2, or the Soyuz Launcher, is much used, particularly for launching manned Soyuz and Cosmos reconnaissance vehicles.

Arianespace, together with Aerospatiale Matra of France (now EADS Launch Vehicles), the Russian space agency RKA, and the Russian space centre of Samara formed the Starsem joint venture on 17 July 1997 to market the Soyuz launch vehicle worldwide and to assist non-Russian customers to launch their payloads. Different configurations of launch vehicle are offered based on the Soyuz booster with different upper stages and size of fairings. The launcher first- and second-stage design is based on the Semyorka vehicle, developed originally to meet defence requirements. Multiple satellite payloads can be deployed. Launches are made from Baikonur, where a dedicated launchpad has been upgraded.

11.2.3.1.1 Soyuz–Ikar

Used to inaugurate Starsem commercial operations, Soyuz–Ikar uses the basic Soyuz vehicle composed of a lower portion consisting of four boosters (the first stage) and a central core stage (second stage), and an upper portion consisting of the third stage, payload adapter, and fairing. Liquid oxygen and kerosene are used as propellants for the basic Soyuz launcher.

Located beneath the payload fairing is the Ikar fourth stage, an upper stage with an in-flight restart capability of up to 50 times, permitting injection of multiple payloads into different orbits. Ikar uses unsymmetrical dimethylhydrazine (UDMH) as fuel and N_2O_4 as oxidiser (up to 900 kg propellant) and can be controlled from the ground or operated in an autonomous mode, ensuring spacecraft orientation and stabilisation upon injection into orbit. The Soyuz–Ikar vehicle of 43.4 m total height has a 4100 kg payload performance to a 450 km circular orbit inclined at 51.8°, or 3300 kg to a 1400 km circular orbit. The fairing diameter is 3.3 m.

11.2.3.1.2 Soyuz–Fregat

Designed for cost-effective solutions on missions to medium- and high-altitude earth orbits (including constellation deployment and earth escape trajectories), Soyuz–Fregat uses the basic Soyuz vehicle topped off by the Fregat upper stage, a flight-proved propulsion subsystem used on nearly 30 interplanetary spacecraft. Fregat embarks up to 5400 kg of UDMH and N_2O_4 propellant, and the single-chamber engine providing 20 kN thrust with a 327 seconds I_{sp} can be restarted as many as 20 times.

The Soyuz–Fregat launcher version has a 5000 kg payload performance to a 450 km circular orbit inclined at 51.8°, or 4000 kg to a 1400 km circular orbit.

11.2.3.1.3 Soyuz–ST

This vehicle will use an upgraded version of the basic Soyuz launcher, with improvements also introduced for the payload. The Soyuz launcher first- and second-stage engines will have redesigned combustion chamber injectors, while the third-stage structure will be reinforced and its propellant tanks enlarged. A new digital flight control and telemetry system will be integrated in the Soyuz–ST third stage, replacing the current analogue system. The new system will provide a more precise trajectory during flight and in orbital injection, and will enable a dog-leg manoeuvre capability when required to reach an extended range of inclination from fixed azimuths (±5°).

The most visible change to Soyuz–ST is the upper composite integrating a payload adapter and an Ariane 4 class fairing. The Soyuz–ST fairing will be approximately 1 m longer than the largest Ariane 4 fairing, providing the necessary volume for large satellite payloads. Soyuz–ST can be fitted with the Ikar or Fregat upper stages, or use the full payload fairing volume for the satellite payload.

Payload performance to a 450 km circular orbit inclined at 51.8° is 4900 kg. When equipped with the Fregat upper stage, the payload lift capability increases to 5500 kg to the 450 km circular orbit, 4600 kg to a 1400 km, or 2900 kg to an 800 km SSO. Soyuz can be launched from Kourou. Indeed, in 2004, the European Space Agency (ESA) confirmed that Soyuz could use the former launchpad of Ariane 4.

11.2.3.2 Cosmos

Cosmos SL-8 is a two-stage, liquid-fuelled rocket developed in the 1960s by Polyot. The first stage has two pump-fed combustion chambers and four nozzles; it is the motor of the SS-5 missile. The second stage is a re-ignitable motor with a single pump-fed main engine and an auxiliary system for orientation and attitude control. This second-stage propulsion system fires twice during its journey into space, coasting in a ballistic trajectory between the two burns.

Polyot formed a joint venture with Assured Space Access Incorporated of Arlington Virginia. Called Cosmos USA in 1995, it aimed to assist non-Russian customers in launching their payloads on Cosmos SL-8 [BZH-97]. The Cosmos rocket can be launched into orbits with inclinations of 51° from Kapustin Yar and of 66°, 74°, and 83° from the Plesetsk site. This launch vehicle is capable of placing 1110 kg into a 1000 km circular 51° inclined orbit.

11.2.3.3 Proton

The Proton is part of the D series of heavy launchers capable of placing from 17–21 tons in low orbit, depending on the configuration. The stages (two to four depending on the version) use liquid propellants. The first stage consists of a central body containing the oxidiser and six auxiliary reservoirs containing the fuel. Six motors (Glushko/GDL-OKB RD-253) are mounted at the end of the auxiliary reservoirs and develop a thrust of 150 tons. The second stage is propelled by four Kosberg/RD-010 motors, each developing a thrust on the order of 650 kN. The third stage uses a single Kosberg/JRD motor and has four vernier motors for orientation control. These stages use hydrazine and nitrogen peroxide as propellants. The launchers, assembled horizontally, are fired from Tyuratam.

Version D1e, the Proton launch vehicle, is used particularly for launching satellites into geostationary orbit. The Soviet Union proposed in 1983 to offer the services of the Proton launcher to the Western world to launch the Inmarsat II satellites. A commercial organisation, Licensintorg, was established under the aegis of the Glavkosmos agency.

The Proton launcher has four stages and is 57.2 m high. Its mass at take-off is on the order of 700 tons. The first three stages place the fourth stage and payload with maximum combined mass of 19 760 kg in a circular orbit of altitude 200 km and inclination 51.6°. The first three stages are controlled by a closed loop, triple redundant, inertial navigation unit (INU) guidance system.

11.2.3.3.1 Block-DM upper stage

The Energia-built fourth stage is actually a transfer (Block-DM) stage. It is powered by a restartable, LOX/synthetic kerosene propulsion system delivering a vacuum thrust of 84 kN with a specific impulse of 350 seconds. Altitude control is provided by small thrusters (vernier motors) that use hydrazine and nitrogen peroxide. The Block-DM inert mass at separation is approximately 2140 kg. Propellant carried is dependent on the specific mission requirements and is varied to maximise performance for the mission. The Block DM is capable of operating on-orbit for a minimum of 24 hours and is controlled by a closed-loop, triple-redundant guidance system that is commandable in flight. The Block-DM fourth stage is used to perform all subsequent manoeuvres to place the spacecraft into either GTO or GEO.

Insertion into GTO can be accomplished in one of two ways:

— Through a single-impulse transfer that allows spacecraft separation shortly after ascent
— Through a two-impulse transfer that allows the Block DM to perform most of the desired inclination change at orbit apogee

Insertion directly into GEO is accomplished through a two-impulse transfer that phases in the parking orbit or through a three-impulse transfer that makes use of an intermediate phasing orbit. Using a standard two-burn mission profile, the Block DM is capable of delivering 4350 kg to GTO. Using a three-burn injection scheme developed to support commercial missions, the Block DM is capable of delivering between 4700 and 4930 kg to GTO. The performance for direct injection into GEO is 1880 kg.

The standard commercial 4.35 m diameter fairing with an internal payload envelope 6.6 m in length (exclusive of the payload adapter) and 4.1 m in diameter provides adequate volume for most large communications spacecraft.

11.2.3.3.2 Launch sequence

In a typical Proton launch, the vehicle's six first-stage engines ignite at lift-off. Stage two ignition occurs approximately two minutes into the flight. Stage three vernier engine ignition occurs at 330 seconds, followed by separation of the second and third stages and ignition of the stage three main engine. For typical Proton missions, the first three stages inject the elements above the third stage into a 200-km circular parking orbit. The Block-DM fourth stage then performs all mission-specific manoeuvres, starting from the parking orbit. The first burn of the Block-DM engine occurs approximately 55 minutes after lift-off as the vehicle crosses the first ascending node, and lasts about 6.5 minutes. The second Block-DM burn, which places the spacecraft into its final orbit, occurs approximately 5.5 hours later at geostationary altitude, and lasts 2.5 minutes.

Lockheed-Martin Corporation of the United States, Khrunichev State Research and Rocket Space Complex, and NPO Energia of the Republic of Russia formed the Lockheed Khrunichev Energia International (LKEI) joint venture to market the Proton launch vehicle to non-Russian government customers worldwide. Through LKEI, launch services are provided by the International Launch Services (ILS) company (see Section 11.2.3.2).

11.2.3.3.3 Proton-M/Breeze-M

The Proton-M/Breeze-M system is a 'modernised' version of the Proton, capable of placing approximately 21 000 kg into LEO at 51.6°. The Khrunichev-built Breeze-M upper stage is a derivative of the flight-proven Breeze-K stage (as used on the Rockot system). The Breeze-M is powered by a single NTO/UDMH propulsion system delivering a vacuum thrust of 22 kN. The Breeze-M is composed of a central cylinder and a jettisonable external propellant tank. Inert mass of the stage at lift-off is approximately 2250 kg. Propellant carried is dependent on the specific mission requirements and is varied to maximise performance for the mission. As with the Block-DM, the Breeze-M is capable of on-orbit operation for a minimum of 24 hours and is controlled by a closed-loop, triple-redundant guidance system that is commandable in flight.

The Breeze-M fourth stage is capable of placing payloads into high-energy GTO or directly into GEO. A payload mass of approximately 5500 kg can be delivered to a GTO resulting in a 1500 m s^{-1} Delta velocity to GEO, equivalent to a Kourou-launched geostationary transfer. Performance for direct injection into GEO is 2920 kg.

For typical Proton-M/Breeze-M missions, the first three stages inject the upper composite into sub-orbital ballistic trajectory. Approximately two minutes after separation, the Breeze-M fourth stage performs a main engine burn to reach a low earth 'support' orbit inclined 52° to the equator. The second burn of the Breeze-M engine occurs approximately 55 minutes after lift-off as the vehicle crosses the first ascending node and lasts nearly 12 minutes. After one revolution in an intermediate transfer orbit, a third Breeze-M burn occurs to complete the raising of apogee to geostationary altitude. The fourth Breeze-M burn, which places the spacecraft into its final orbit, occurs approximately 5.5 hours later at geostationary altitude and lasts 10 minutes. Total launch mission duration is approximately 10 hours.

11.2.3.4 Rockot

In March 1995, Daimler-Benz Aerospace AG (DASA) of Germany (which became EADS Astrium and is now Airbus Space and Defence) and the Russian Khrunichev State Research and Production Space Center set up Eurockot Launch Services under their joint ownership.

Rockot is a three-stage liquid propellant launch vehicle, composed of a former Russian SS-19 strategic missile (first and second stage) that has been withdrawn from military use, and a flight-proven upper stage of the Breeze family, with multiple-ignition capability. The commercial Rockot-KM offered by Eurockot incorporates a new payload fairing and a modified payload interface structure so as to be able to launch large, tall spacecraft. The booster unit that provides the first and second stage of Rockot is accommodated within an existing transportation/launch container, identical to that used during the previous silo launches so as to maintain heritage. The third-stage Breeze provides the orbital capability of the launcher. The Breeze contains a multiple restartable liquid propellant main engine. The upper stage contains a modern control/guidance system.

Eurockot offers commercial launch services to LEO between 48° up to SSO by providing two launch sites, Plesetsk and Baikonur. Rockot cannot aim at the geostationary orbit from Plesetsk. However, this launch system is particularly reliable for launches of small and medium-sized spacecraft into sun-synchronous, almost polar, and highly inclined orbits. Rockot can also be used for the initial set up of constellations by launching several satellites in one launch. Earth escape and planetary missions for small payloads using an additional propulsion module can be achieved. The launches are performed above ground using the transport/launch container. The launcher rests physically on a ring at the bottom of the launch container. During lift-off, the launcher is guided by two guide rails within the launch container. The container protects

the launch table environment from the engine plumes and gases and ensures that the correct temperature and humidity are maintained during storage and operation.

The Rockot launch vehicle can place 1800 kg into 450 km, 63° inclination circular polar (1500 kg at 1000 km) or 1000 kg into a 800 km SSO. Inclined GTO and earth-escape missions can also be served with an additional commercial solid stage.

11.2.3.5 Angara

Angara is a new family of launchers under development by Khrunichev State Research and Production Space Center, dedicated to satellite launches. It is being developed to replace the Proton and was launched in 2014, with several versions putting into orbit payloads up to 24.5 tons to LEO and 7.5 tons to GTO.

The Angara family consists of three types of rocket:

— Launch vehicles of the first type (versions 1.1 and 1.2) are for launching small telecommunications satellites of up to two tons.
— Satellites of up to four tons can be put into orbit with launch vehicles of the second type (version 3).
— Rockets of the third type (versions 5 and 7) can launch over 20 tons.

The launchpad is located at Plesetsk. The ground installations have been modernised. In July 2006, the massive Angara launch platform (LP) was delivered to Plesetsk in pieces: the $14 \times 14\,m \times 5\,m$ tall launch table weighs 1.185 tons. Like other Russian launchers, Angara will be horizontally processed using rail-based transporter–erectors. The five versions of Angara are based on the same architecture.

Angara 1.1 (145 tons and 35 m in length) uses the RD191M engine burning kerosene and LOX. The second stage is the Breeze vehicle. Angara 1.1 will be able to launch 2100 kg into LEO. Version 1.2 should able to launch 3600 kg into LEO, thanks to a larger first stage. Angara 3 will use a Soyuz stage and a Block-IE upper stage. Version 5 uses additional boosters to put a payload of 24 tons into LEO. This version will have two subversions: one basic, capable of 6600 kg in GTO and 4000 kg directly in GEO; and KVSK (with a cryogenic stage), capable of 8000 kg in GTO and 5000 kg in GEO. Version 7 is now under development.

11.2.3.6 Dnepr

The Dnepr rocket is used for commercial purposes by the International Space Company (ISC) Kosmotras, a joint project between Russia, Ukraine, and Kazakhstan created in 1997. Dnepr is a three-stage vehicle of 34.3 m based on the Russian intercontinental ballistic missile (ICBM) R-36MUTTH and designed by the Yuzhnoye Design Bureau in Ukraine. The main differences between Dnepr and R-36MUTTH are the payload adapter located in the space head module and the modified flight-control unit. Moreover, two stages may be added in order to increase the capacity of the vehicle. Dnepr is fuelled by liquid propellants and is capable of placing a satellite of 4500 kg into LEO at 200 km altitude with an inclination of 46°.

Dnepr is mainly launched from Baikonur but may also be launched from the newly created Cosmodrome at the Yasny launch base (Dombarovsky), in the Orenburg region of Russia. It flew for the first time in 1999. Ten years later, Dnepr has been launched 12 times and failed only once. The ICBM used for Dnepr has been launched 160 times for military purposes with 97% reliability, promising a good future for the commercial launches of Dnepr as around 150 ICBMs are available to be converted into launch vehicles.

11.2.3.7 Zenit

As with Dnepr, Zenit is a space launch vehicle designed by the Yuzhnoye Design Bureau in Ukraine. Its development started in 1976 with two objectives: a liquid rocket booster (LRB) for the Energia rocket (Soviet launcher) and, equipped with a second stage, to be a stand-alone launch vehicle.

11.2.3.7.1 Zenit-2

This was the first Zenit to be designed. It is a two-stage vehicle fuelled by RP-1/LOX. This 57 m launcher is capable of putting a payload of 13.8 tons into LEO at an altitude of 200 km. Launched from Baikonur, it flew for the first time in 1985. Zenit-2 is reliable at 83%, as it failed 6 times in 37 launches. In 2001, Zenit-2 was the lowest-cost vehicle for achieving LEO in terms of payload weight per launch and one of the lowest in terms of total cost per launch. This launcher was used in the well-known Globalstar mission.

The Zenit-2M has an upgraded control system and modernised engines; it was first launched in 2007. Land Launch is a subsidiary of Sea Launch and commercialises the Zenit-2M under the name Zenit-2SLB; these launches are also made from Baikonur.

The Zenit-2M–Fregat is a three-stage vehicle using a Fregat (Soviet rocket) upper stage launched from Baikonur.

11.2.3.7.2 Zenit-3SL

This vehicle was developed for the Sea Launch consortium to send satellites into GTO using the Block-DM-SL upper stage, provided by Russia's Energia and used for the launchers N1 and Proton. The two other stages are those of Zenit-2; Boeing provides the fairing for protection of payload during launch. The Zenit-3SL capacities are 6.1 tons into LEO, 5.25 tons into GTO, and 1.9 tons into GEO.

As with the Zenit-2, there is a Zenit-3M composed of a Zenit-2 with the upper stage of the Zenit-3SL. The Zenit-3M is also commercialised by Land Launch under the name Zenit-3SLB. It flew successfully for the first time in April 2008 from Baikonur.

11.2.4 Europe

11.2.4.1 From Ariane 1 to Ariane 4

The Ariane family of launch vehicles has been developed by the European Space Agency under the management of the Centre National d'Etudes Spatiales (CNES).

In July 1973, during the European Space Conference, it was decided to combine the European Launcher Development Organisation (ELDO) and the European Space Research Organisation (ESRO) into a single body called the ESA. Among the objectives of the agency was that of developing a range of three-stage launchers to permit direct injection of the satellite–apogee motor combination from a transfer orbit into the geostationary satellite orbit without a ballistic phase, starting from the Kourou launch site in French Guyana. This orbit can be obtained accurately using guidance of the launch vehicle by an on-board computer that uses information provided by an inertial unit. Steering is provided by orientation of the jet of the main motors of the various stages. Furthermore, an attitude and roll control system (SCAR – systèm de contrôl d'altitude et de roulis in French) enables the third stage and the satellite to be positioned in the desired attitude before separation with, if necessary, rotation at up to 10 rpm. Economy is thus realised in

the quantity of propellant to be embarked for later corrections of the satellite orbit and attitude. This mass saving may represent one to three years of satellite lifetime. The launchers can put several independent satellites into orbit simultaneously by means of specific adaptation systems (Sylda – 'système de lancement double Ariane' means Ariane double launch system, and Spelda – 'structure porteuse externe lancement triple Ariane' means Ariane triple launch external structure).

The first firing of an Ariane 1 launcher took place on 24 December 1979. A programme of improvements by minor modifications of the launcher (such as addition of extra solid boosters, modification of the motor combustion chamber pressures, and increase of the mass of propellant) led to Ariane 3, whose first flight took place in August 1984. Table 11.6 shows the evolution of Ariane vehicles.

The Ariane 4 programme was developed as decided in January 1982. This launch vehicle was the workhorse of the Arianespace company for the 1990s; the first flight of the Ariane 4 launcher took place on 15 June 1988 and the last on 15 February 2003. All in all, Ariane 4 flew 116 missions over a period of 15 years, with only three failures (one in 1990 and two in 1994). This gives it a success rate of 97.4%. Ariane 4 placed a total of 182 satellites, weighing a total of 444 tons, into orbit for 50 customers. Over the final nine years, it carried out 74 successful launches in a row.

The performance improvement was obtained by an increase of the capacity of the first stage and the use of additional, more powerful, solid or liquid propellant boosters. As a consequence of the use of additional boosters, the nominal performance increased from 2.1 to 4.9 tons for GTO with altitudes at the perigee 200 km and apogee 35 786 km; the inclination is 7°.

A dedicated launching assembly called ensemble de lancement Ariane 2 (ELA-2) permitted the interval between two firings to be reduced to less than one month. The launch vehicle was assembled vertically in a special building and then moved on a mobile platform to the launching area, which was remote from the assembly building. Assembling of the fairing of pre-encapsulated satellites with the launch vehicle was performed in the launch area. During these last preparations, assembling of another launch vehicle on a second mobile platform could be started in the assembly building.

11.2.4.2 Ariane 5 and future Ariane 6

Since 2000, a growing number of satellites with a large mass have taken an increasing part of the market. Market studies estimated the demand for the early years of the new millennium to be up to 12 000 kg in GTO with an available diameter under the fairing on the order of 4.5 m. To remain competitive, the cost of launching must be reduced as far as possible. One cost-reduction factor involves the use of a sufficiently powerful launcher to provide simultaneous launching of several satellites. Finally, the highest possible reliability is required of a commercial launcher as the payloads become more and more costly. The Ariane 5 launcher development programme was undertaken in 1985 to fulfil these objectives.

Table 11.6 Ariane launch family performance

	Ariane 1	Ariane 2	Ariane 3	Ariane 4	Ariane 5 (G/ES)
Year of operation	1981	1984	1984	1986	1999
Geostationary transfer orbit (kg)	1 800	2 175	2 580	2 100–4 900	6 640–8 000
Low earth orbit (200 km) (kg)	4 900	5 100	5 900	8 990	20 500
Sun-synchronous orbit* (kg)	2 400	2 800	3 250	6 490	9 500
Escape (kg)	1 100	1 330	1 550	2 580	—

*800 km, 98.6° inclination.

11.2.4.2.1 Ariane 5G

This is the generic or original version of the Ariane 5 launcher, consisting of the central cryogenic main stage, two solid booster stages, and an upper stage. Using a limited number of engines, this architecture is both simple and robust and has growth potential, particularly for the upper-stage component. Figure 11.15 illustrates the configuration of Ariane 5 [AR5-16], and Table 11.7 gives the main characteristics of the basic version of Ariane 5 (Ariane 5G). Ariane 5 is launched from

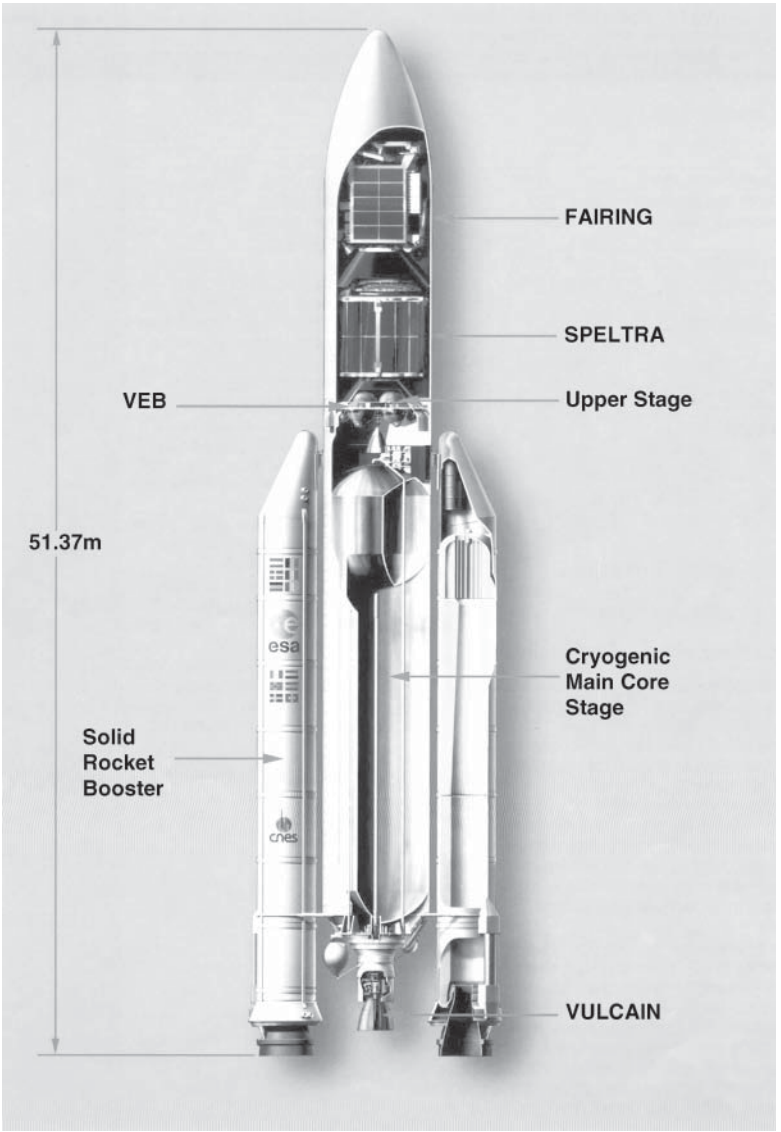


Figure 11.15 Ariane 5 configuration.

Table 11.7 General characteristics of Ariane 5

	EAP	EPC	EPS	VEB	Fairing	Launcher
Diameter (m)	3.05	5.45	5.4	5.4	5.4	—
Height (m)	31.2	30	4.5	1.6	17/12.7	51.37
Total mass (ton)	268/unit	170	10.9	1.1	2.4/1.4	725
Mass of propellant at take-off	237/unit	LOX:132 LH2:26	MMH:3.8 N2O4:5.9	Hydrazine 60 kg	—	682

a newly developed launching assembly (ELA 3) installed at the launching site at Kourou. The Ariane 5G standard performance is 6640 kg in GTO.

The cryogenic main core (EPC – etage principal cryyotechnique) of 30 m in size is powered by a Vulcain engine that provides up to 116 tons of thrust in vacuum. The Vulcain engine is ignited on the launchpad seven seconds before lift-off, allowing full monitoring of the engine during its startup and the stabilisation of thrust. It operates for a total of 589 seconds. The cryogenic propellant used in the stage is non-toxic. At the end of its flight, the main cryogenic stage reenters the atmosphere and disintegrates over the ocean.

The solid booster stages (EAP – Etage d'acceleration a poudre) propel the 725-ton Ariane 5 from the launch table with an acceleration of 0.5 G at lift-off. The boosters are 30 m tall and are loaded with 240 tons of solid propellant each. They deliver a combined thrust of 1370 tons at lift-off (more than 90% of the total launcher thrust at the start of flight). The booster burns for 130 seconds with an average thrust of 1000 tons before separating over a designated zone of the Atlantic. The storable propellant stage (EPS) is the first upper stage developed for Ariane 5. It propels the launcher payload to its final orbit and provides an accurate orbital injection. The stage carries about 10 tons of propellant (nitrogen tetroxide and monomethyl hydrazine) and delivers a thrust of about 3 metric tons.

The vehicle equipment bay (VEB) incorporates most of the avionics, including the two on-board computers for flight guidance (one prime and one backup) and the primary and backup inertial measurement units that provide guidance and attitude data to the computers. The VEB also houses the attitude control system, which supplies launcher roll control after booster separation, and three-axis control during the upper-stage burn and payload deployment manoeuvres.

11.2.4.2.2 Ariane 5G +

This was the first improved version of Ariane 5G for the EPS second stage, launched successfully three times in 2004. Compared to Ariane 5G, the changes give a lighter P2001 nozzle on the EAP boosters and modifications to the EPS upper stage and the VEB.

11.2.4.2.3 Ariane 5 GS

This is the latest version of Ariane 5G, produced after Ariane 5G + and launched for the first time in 2005. This version is based on components used in Ariane 5 ECA (evolution cryotechnique type A) and Ariane 5 ES-ATV (evolution storable – automated transfer vehicles). It includes EAP boosters with more propellant in the S1 segments, a composite VEB with electrical equipment identical to those produced for the Ariane 5 ECA, and an EPS stage loaded with 300 kg additional propellant.

11.2.4.2.4 Payload fairings and dual launches

Two payload fairing versions are available on Ariane 5, both with a useful inner diameter of 4.57 m. The short fairing version, which is 12.7 m long, can accommodate payloads more than 11.5 m high. The long fairing, which is 17 m long, can house payloads more than 15.5 m high. Figure 11.16 shows the typical payload compartment configurations.

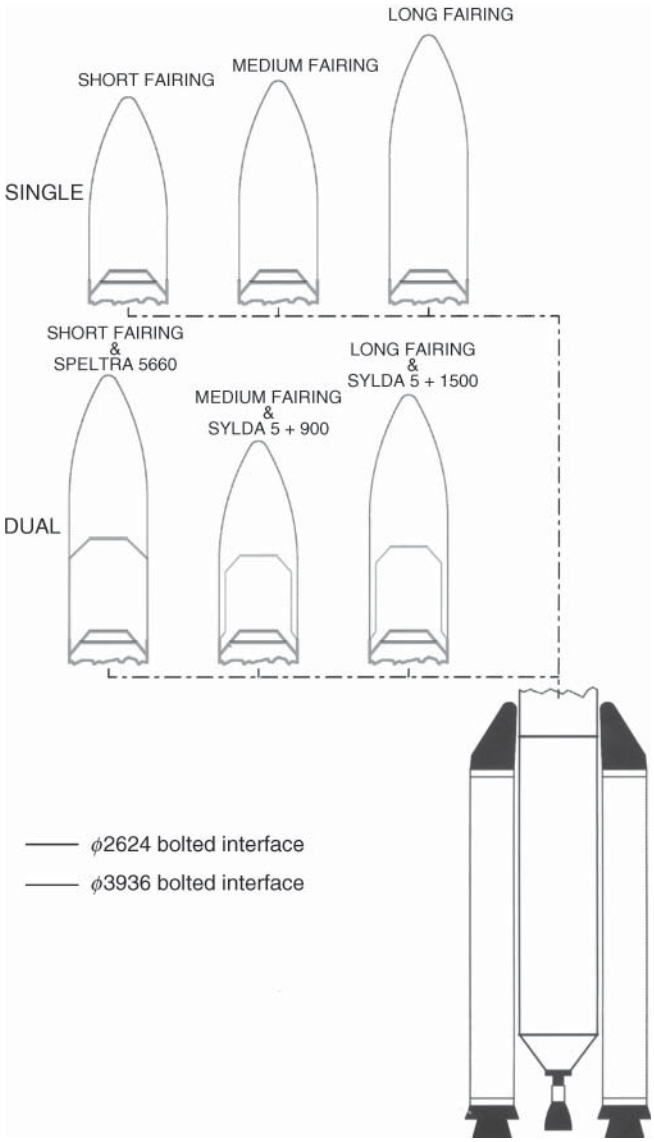


Figure 11.16 Ariane 5 typical payload compartment configurations.

The internal structure for dual launches (Sylda 5) is housed inside the fairing and allows Ariane 5 to launch two primary payloads on a single flight. It has a useful inner diameter of 4 m and exists in six versions to accommodate satellites with a maximum height of 2.9–4.4 m. Sylda mass varies from 425–500 kg.

Larger satellites can be accommodated into the external structure for dual launches (Speltra), positioned between the upper stage and the payload fairing. One satellite is accommodated inside the Speltra, while the other is mounted atop the Speltra and is enclosed in the payload fairing. The Speltra can house payloads with external diameter of 4.57 m. The standard Speltra version is 7 m high and weighs 822 kg. A short version and a stretched version of Speltra are also offered.

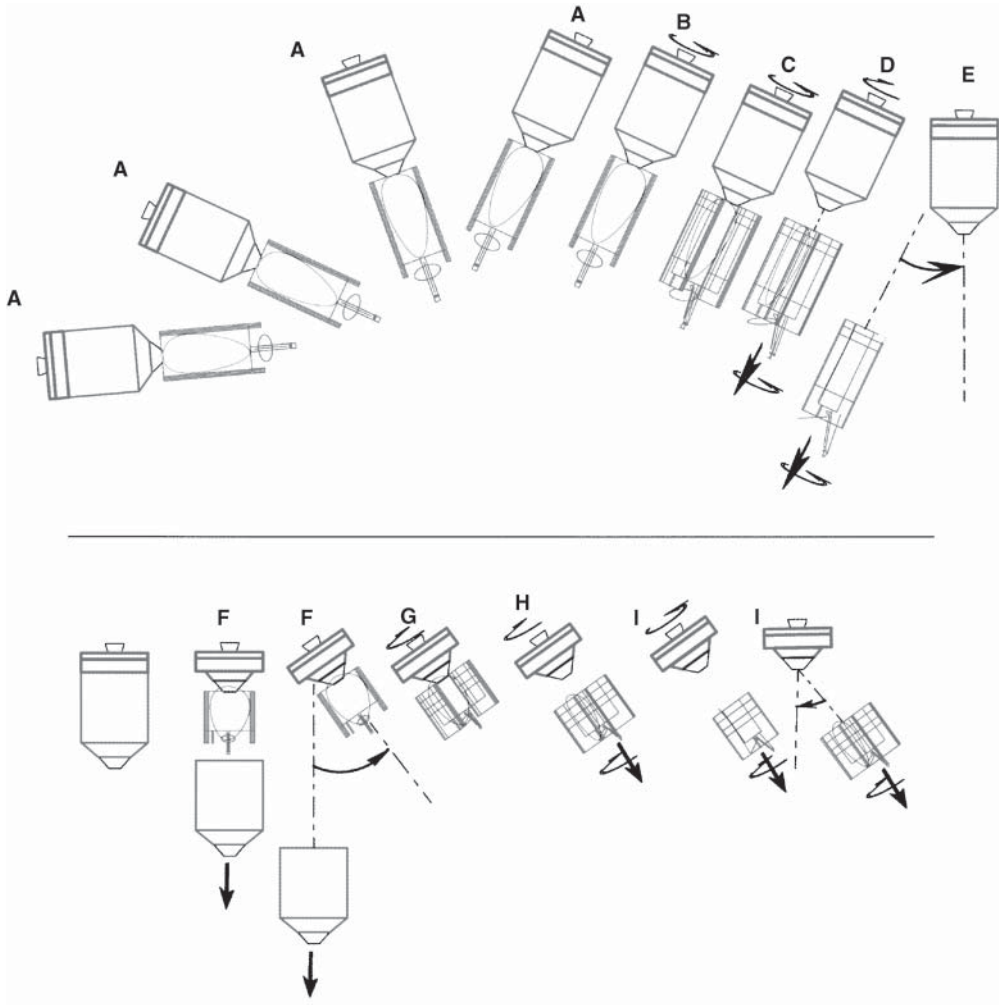
With a dual launch, satellites can be separated with different requirements in terms of orientation and spin velocity. Figure 11.17 illustrates the separation procedure.

11.2.4.2.5 Ariane 5 Plus

The Ariane 5 Plus program was the follow-on to the Ariane 5 Evolution (Ariane 5 E) effort and involves improvements to the launcher lower composite (composed of the main cryogenic stage and solid booster stages). The improvements are focused on structural weight reduction, major modifications to the solid boosters and the main cryogenic stage, as well as a thrust increase to 138 tons for the cryogenic stage Vulcain 2 main engine. An improved version of the current EPS upper stage entered service in 2002 (Ariane 5 ES). This EPS, combined with the improvements to the Ariane 5 lower composite, increased the launch vehicle payload lift capability to 7300 kg with a double payload to GTO (with Speltra) and 8000 kg in a single-satellite launch. With its in-flight restart capability and an operational time of several hours, the improved upper stage allows for more complex missions, including the deployment of constellations and scientific satellites.

One of the most important elements of the Ariane 5 Plus program was the development of two new cryogenic upper stages (ESC-A and ESC-B – Étage Supérieur Cryotechnique type A and B), which increase the launcher payload lift capacity (Ariane 5 ECA). The ESC-A upper stage is powered by the same 6.5 ton thrust HM-7B engine as used in the Ariane 4 third stage (which is designed for ignition once during flight) and carries 14 metric tons of LOX and LH2 propellant. Performance is 10 000 kg into GTO on a dual-payload mission (using the Sylda system) and 10 500 kg with a single payload. The ESC-B is loaded with 25 metric tons of LOX and LH2 to feed the new 15.5 metric ton-thrust Vinci engine using the *expander cycle*. With the capability of performing multiple restarts in ballistic flight, it improves the Ariane 5 payload capability in dual GTO launch to 11 000 kg (with Speltra) and 12 000 kg in a single-payload mission.

The main outcome of this programme is the Ariane 5 ECA (Evolution Cryotechnique type A) with a height of 59 m, a mass of 777 tons, and a capability of delivering a payload of 10.5 tons to GTO. Although Ariane 5 ECA is also capable of delivering a payload of 21 tons into LEO, the dedicated vehicle for LEO and MEO launches is Ariane 5 ES (Evolution Storable) with a payload capacity of 21 tons. Its main missions, called Ariane ES-ATV, are to launch the automated transfer vehicle (ATV) to reach the International Space Station (ISS). The first ATV (*Jules Verne*) was successfully launched on 9 March 2008, followed by *Johannes Kepler* on 16 February 2011, *Edoardo Amaldi* on 23 March 2012, *Albert Einstein* on 5 June 2013, and *Georges Lemaitre* on 22 July 2014, with 20 tons of supplies each, to the ISS. This Ariane 5 version also launched the Galileo satellites in 2013. In the ECA version, the second stage is powered by the HM-7B engine, and the first stage uses the new Vulcain 2 engine, modified from Vulcain 1 to increase the thrust by 20% up to 137 tons. The Ariane 5 ECB should further improve the GTO capacity to 12 tons, but the



A: Orientation of composite (Upper stage + VEB + payload) by attitude control system (SCA)

B and C: Spin-up by SCA

D: Separation of spacecraft

E: Spin-down and reorientation to SYLDA 5 jettisoning attitude

Note: Spacecraft separations can also be accomodated under a 3-axis stabilized mode.

F: SPELTRA jettisoning. Reorientation as requested by inner spacecraft.

G: Spin-up by SCA.

H: Separation of lower spacecraft.

I: Upper stage avoidance manœuvre (Spin down, attitude deviation by SCA and passivation).

Figure 11.17 Typical spacecraft/Speltra separation sequence for various altitude and spin requirements. Source: reproduced from [AR5-16] with the permission of Arianespace.

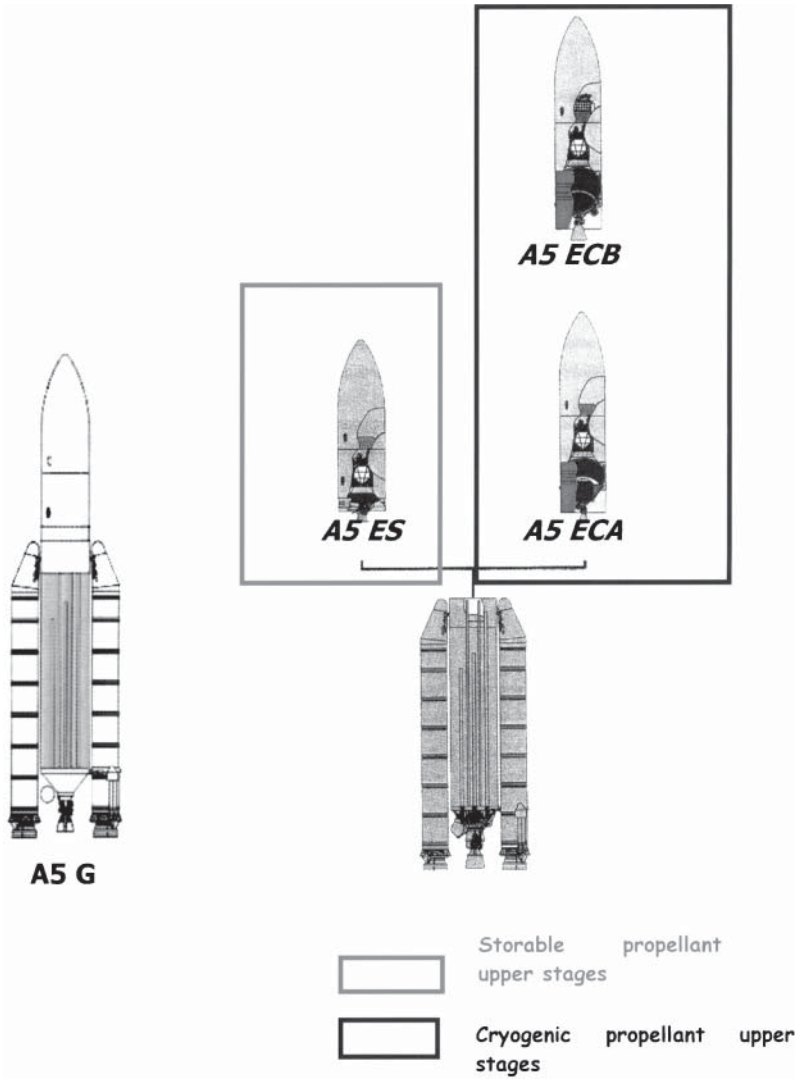


Figure 11.18 Ariane 5 launch vehicle family.

program has been put on hold for budget reasons. Figure 11.18 illustrates the Ariane 5 family [AR5-16].

11.2.4.2.6 Ariane 6

Ariane 6 is being developed for future European space launching vehicles with a plan of a first test flight scheduled for 2020 [AR6-18]. Two versions of Ariane 6 are under development:

- Ariane 62 is being developed as a small mid-lift vehicle. It weighs around 530 tons with a payload up to 5000 kg into GTO and 10 350 kg into LEO.

- Ariane 64 is being developed as a heavy-lift vehicle. It weighs around 860 tons with the capability of commercial dual-satellite launches of up to 11 500 kg into GTO and 21 500 kg into LEO.

11.2.4.3 Ariane structure for auxiliary payloads

The ASAP platform, which can be mounted on top of the upper stage as well as on the Speltra or Sylda structures, carries mini- or microsatellites as secondary payloads. When located under a primary payload, the ASAP platform accommodates up to eight microsatellites, each weighing under 120 kg. When mounted inside a dedicated Sylda structure, it can carry up to four minisatellites weighing up to 300 kg each, or two 300 kg minisatellites and six 120 kg microsatellites.

11.2.4.4 Vega

The development of Vega ('Vettore Europeo di Generatione Avanzta' in Italian means advanced generation European carrier rocket) started in 1998 with the aim of being launched from Kourou in 2009. Vega is designed to launch small satellites (300–2000 kg) in polar and LEO, at a low price compared to large launch vehicles. Its reference capability is a payload of 1500 kg at 700 km of altitude on polar orbit. It has the first launch from Kourou on 13 February 2012 with 13 successes by early 2019.

Vega is a single-body launcher of 30 m height, 137 tons at lift-off, and 3 m diameter with three solid propulsion stages (P80, Zefiro 23, and Zefiro 9) and an additional liquid propulsion upper module (AVUM) used for orbit control and satellite release. The P80 first stage (11 m long and loaded with 88 tons of solid propellant) delivers a thrust of 3040 kN. The second stage, Zefiro 23 (7.5 m in length, 24 tons of propellant), delivers a thrust of 100 tons or 1070 kN. The last stage, Zefiro 9 (3.2 m long, 10 tons of propellants), delivers a thrust of 305 kN in vacuum.

An upgraded Vega (Lyra program) is planned; it should have new third and fourth low-cost LOX–HC stages and a new guidance system. The purpose of this new Vega is to upgrade the polar orbit payload capability to 2000 kg. The Vega has been improved with further development to be used as the side booster of the Ariane 6.

11.2.5 India

In the 1970s, under the administration of the Indian Space Research Organisation (ISRO), India initiated the development of launchers for national scientific requirements. With the launching of a 35 kg satellite on 18 July 1980 using a four-stage solid propellant SLV-3 launcher (satellite launch vehicle 3), India became the sixth country to have launched a satellite by its own means. Launches are made from the Sriharikota launchpad situated 160 km north of Chennai (formerly Madras).

A performance improvement programme led to the augmented satellite launch vehicle (ASLV), which has a capacity of 150 kg in low-altitude orbit. The first launch of ASLV was made in 1987; but due to three failures in four launches, the program was stopped in 1994.

The Polar satellite launch vehicle (PSLV) is capable of placing 2900 kg in sun-synchronous polar orbit at 400 km altitude and even 3200 kg at 200 km altitude into LEO orbit. It can only launch small satellites into GTO orbit. The PSLV has four stages using solid (stages 1 and 3) and liquid (stages 2 and 4) propellants. From 1993–2008, 14 PSLVs were launched with only two failures. From 1993–May 2019, 48 PSLVs were launched, with only two failures and one partial

failure. The PSLV also has capacity to launch a group of LEO constellation satellites. Moreover, research is underway to build a three-stage PSLV, removing the second stage of the PSLV in order to send small satellites into LEO.

A cryogenic motor development programme enables the geostationary satellite launch vehicle (GSLV) to launch 2500 kg in GTO and 5000 kg in LEO. This three-stage vehicle using both solid and liquid propellants flew for the first time in 2001 from Sriharikota. Stages 1 and 2 are almost the same as for the PSLV, but stage 3 is new with a cryogenic motor furnished by Russia.

The GSLV-III has been developed, and successfully conducted the first orbital test launch on 5 June 2017 from the Satish Dhawan Space Centre, Andhra Pradesh; and also on 22 July 2019 from the second launchpad of the Satish Dhawan Space Centre in Sriharikota, India. Although keeping the same name, it is entirely different from its predecessor. Indeed, using an indigenous cryogenic stage, the GSLV-III is a two-stage vehicle designed to launch heavy satellites in GTO (4000 kg) and in LEO (8000 kg). A new programme has been considered with the universal launch vehicles (ULVs) for LEO up to 41 300 kg and GTO 16 300 kg.

11.2.6 Israel

Israel initiated space programs when the Israel Space Agency (ISA) was established on 19 September 1983. Their Shavit launcher, which means 'comet' in Hebrew, is used to launch the small Israeli satellites Ofeq in LEO. The launches are made from Palmachim Airbase situated near Rishon LeZion, the fourth city of Israel. Shavit was launched for the first time in 1988. It is a three-stage vehicle based on the Jericho II, an Israeli ballistic missile almost identical to the South African missile RSA-3. It delivers a thrust of 760 kN at lift-off, with a mass of 30 tons and a height of 18 m. The last flight was 13 September 2016. The Shavit-2 has mass 30 500–70 000 kg and height 26.4 m with the capability of delivering 350–800 kg to LEO.

11.2.7 Japan

11.2.7.1 NASDA N and H launch vehicles

The launchers for application programmes have been developed in Japan by Mitsubishi Heavy Industries on behalf of the Japanese space agency (NASDA – national space development agency). The N range of launchers consisted of N1 and N2. Model N2 was derived from the Delta 2914 launcher, the second stage being developed by Ishikawa-Harima Heavy Industries (IHI) under licence from Aerojet. This launcher was capable of placing 700 kg in GTO. The useful diameter of the fairing was 2.2 m. The H-I launcher was similar to the N2 launcher with the exception of the second stage, which became cryogenic and used the LE-5 motor of 10.5 tons thrust developed by IHI. This launcher could place 1100 kg in GTO with a fairing diameter of 2.2 m.

The H-II launcher was based on a new design incorporating a central body with two cryogenic stages and two large extra solid boosters. The launcher was 50 m high with a diameter of 4 m and a mass without payload of 260 tons. The first stage used a newly developed LE-7 cryogenic motor of around 860 kN thrust, while the second stage motor was an improved version of the LE-5. The total propellant mass (LOX and LH2) was 103 tons. The additional solid boosters consisted of four segments and were 23.4 m high with a diameter of 1.8 m, containing 118 tons of propellant for a thrust of 3160 kN. Guidance was provided by a strap-down inertia unit using gyrolasers, allowing for a coasting phase. Steering was achieved by controlling the direction of the jet from

the additional thrusters and the first and second motors. Hydrazine thrusters were used during the second stage coasting phase.

Performance was on the order of 2.2 tons in geostationary orbit, 4 tons in a GTO of 30 inclination, and 10 tons in a circular orbit of 200 km altitude. The useful volume under the fairing was 3.7 m diameter and 12 m high. The first launch took place in 1994. Commercialisation of the N2 and H-I launchers was impossible owing to the use of sections built under US licence. It is possible for the H-II family to be commercialised, but the launch site at Tanegashima (latitude 30 N) poses some operational problems on account of restrictions due to fishing.

11.2.7.2 H-IIA launch vehicle

To make commercialisation possible, a great effort was put into achieving cost reductions under an improvement programme initiated by NASDA. Today, the programme is driven by the Japan Aerospace Exploration Agency (JAXA), formed on 1 October 2003 from the merger of Japan Institute of Space and Astronautical Science (ISAS), the National Aerospace Laboratory of Japan (NAL), and NASDA. Since 2012, new legislation has been passed to extend JAXA's remit from peace purposes only to include space development, such as warning systems; and political control has been under the Prime Minister's cabinet through the Space Strategy Office.

The upgraded H-IIA programme aimed at drastic reductions in launch cost and the creation of a launch vehicle family. The main design changes from the original H-II are:

— The first stage makes use of a simplified propulsion system and an improved LE-7A engine.
— The new second stage employs separate tanks, instead of the integral tank with a common bulkhead.
— A simplified propulsion system adopts more reliable valves and an improved 137 kN thrust LE-5B engine
— A payload support structure is mounted on the second stage; this means a payload can be handled in an encapsulated state within the payload fairing in order to shorten launch operation.

The H-IIA in its standard configuration is capable of launching a two-ton class payload into GEO, the same as the H-II rocket. It can launch a three-ton class payload into GEO with the configuration augmented by a large LRB. Growth capability up to a four-ton class payload launching into GEO is also considered in the design. The first stage of the H-IIA launch vehicle consists of the first stage core equipped with the LE-7A LH2/LOX engine and two solid rockets boosters (SRB-A). The LE-7A engine is an improved LE-7 engine (developed for the first stage of the H-II launch vehicle) with 110 tons of thrust (in vacuum). The new SRB-As use a composite material motor polybutadiene composite solid propellant. Each SRB-A provides 230 tons of thrust. The LRB consists of the first stage structure and two LE-7A engines. It is used for launching three-ton class (or heavier) payloads. The second stage of the H-IIA launch vehicle is equipped with the LE-5B LH2/LOX engine. The LE-5B engine is an improved LE-5A engine (developed for the H-II second stage) and provides 14 tons of thrust (in a vacuum). The attitude control of the second stage is performed by the thrust vector control of the LE-5B engine nozzle with electrical actuation and hydrazine gas-jet reaction control systems.

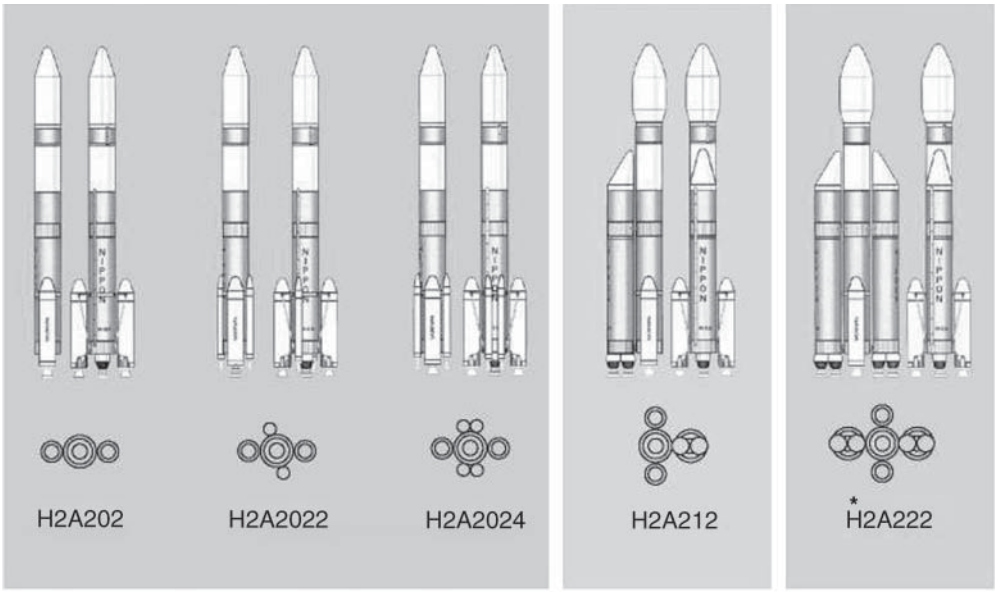
The H-IIA launch vehicle employs a stripped-down inertial guidance and control system. The system consists of the newly developed inertial measurement unit that uses four-ring laser gyros and the guidance control computer. The inertial guidance and control system enable the H-IIA launch vehicle to correct errors autonomously and maintain the planned orbit without commands from the ground stations.

11.2.7.3 The H-IIA launch vehicle family

H-IIA is designed to cope with various payload masses by attaching boosters to the first stage. The standard vehicle of the H-IIA family uses two solid rocket boosters and is called H-IIA202. The launch capability of H-IIA202 can be increased by attaching two or four solid support boosters (SSBs). With two SSBs (the launch vehicle is called H-IIA2022), the launch capability is 4.5 tons to GTO; with four boosters (H-IIA2024), the launch vehicle performance increases to 5 tons in GTO. The standard configuration (H-IIA202) and its variations with SSBs (HIIA 2022 and 2024) are called the H-IIA202 Standard Type.

Additional increase in the launch capability was expected using LRB. The LRB is about the same size as the H-IIA first stage and uses the same type of propellant tank with two sets of LE-7A motors. With an LRB (the launch vehicle is called HIIA212), the launch capability rises to 7.5 tons to GTO and 17 tons in LEO.

Further launch capability increase was envisioned by attaching two LRBs to H-IIA (Figure 11.19 shows the planned H-IIA family). The H-IIA222 would have been capable



Type	H2A202	H2A2022	H2A2024	H2A212	H2A222
Length[m]	53	53	53	53	53
Mass[ton]**	285	316	347	403	520
2nd stage	1	1	1	1	1
1st stage	1	1	1	1	1
SRB-A	2	2	2	2	2
LRB	–	–	–	1	2
SSB	–	2	4	–	–

* H2A222 Is a plan for future development.
** Not including payload mass

H-IIA Code Name Format
H2A abcd (a: Number of stages. b: Number of LRB(s). c: Number of SRB-As. d: Number of SSBs)

Figure 11.19 The H-IIA launch vehicle family.

of launching a 9.5 ton payload to GTO and 23 tons to the LEO. However, the plans for LRBs were cancelled and with them the development of H-II212 and H-IIA222 vehicles.

11.2.7.4 H-IIB and H-3 launch vehicle

The H-IIB is a two-stage launch vehicle with two LE-7A engines, a wide body (5.2 m), external tanks on the new first stage, and four SRB-A booster rockets. The second stage includes LE-5B engine as for H-IIA. In comparison with the H-IIA202, it has increased in total length from 53 to 56.6 m, in mass from 289 to 531 tons, and from 2 to 4 SRB-A (solid rocket boosters). The maximum launch capacity rises from 6 to 8 tons in GTO and 16.5 tons in LEO.

The main purpose of the H-IIB is to send the H-II transfer vehicle (HTV), weighing 16.5 tons, to bring supply goods to the ISS. H-3 has been developed since the authorisation given by the government on 17 May 2013, with a planned first launch scheduled in 2020.

11.2.7.5 GX and Epsilon launch vehicles

The GX vehicle is under development by the Galaxy Express Corporation, a joint venture between IHI Corporation (IHI), JAXA, Lockheed Martin Corporation (LM), and several other Japanese companies.

This two-stage launcher of 48 m height could launch a payload of 3.6 tons in LEO at 200 km altitude. The first stage was Atlas III, with the Russian RD-180 engine. The second stage was Japanese and developed by JAXA; so far it was the only vehicle fuelled by liquefied natural gas (primarily methane) with LOX as the oxidiser. The project was cancelled in 2011.

The Epsilon was developed starting in 2007, with its first flight on 14 September 2013. It has height 24 and 26 m for the enhanced version and mass 9000 and 95 400 kg (enhanced) to deliver a payload of 1500 kg to LEO (250 × 500 km), 700 kg to LEO (500 km), and 590 kg to SSO (500 km).

11.2.8 South Korea

The space agency of South Korea, Korea Aerospace Research Institute (KARI), was founded in 1989 to build sounding rockets. KARI has developed the Korea space launch vehicle (KSLV). The program consists of two launchers. KSLV-I (renamed Naro-1) was launched from the Naro Space Centre in Goheung on 30 January 2013.

KSLV-I is a two-stage vehicle able to put a satellite of 100 kg into LEO at 300 km altitude. The first stage, Angara UM, is a Russian Angara launcher. It uses an RD-191 engine fuelled by LOX/kerosene; it gives a thrust of 2095 kN. KSR-1 from the Korean rocket KSR is used as the second stage of the vehicle. Burning solid propellant, it delivers a thrust of 86 kN. At lift-off, the launcher weights 140 tons, is 33 m in height, and delivers a total thrust of 1910 kN.

KSLV-II will be composed of three stages: Angara UM as the first and second stage and KSR-1 as the third. This 47.2 m launcher of weight 200 tons has been designed to place a satellite of 1500 kg at 800 km altitude in LEO. KSLV-II has been planned for launch from Naro Space Center in 2021.

11.2.9 United States of America

At the beginning of the 1980s, the United States had various conventional launcher programmes of varying capacities. In parallel, a different kind of space transport system using recoverable

units was developed: the Space Shuttle. The decision by NASA in 1984 to abandon conventional launchers in favour of the Space Shuttle seemed to put an end to most programmes. The Challenger Space Shuttle accident on 28 January 1986 led, in the same year, to the United States taking the decision to reactivate conventional launcher development in order to provide a replacement for launching military satellites and to no longer use the Space Shuttle for launching commercial satellites. Hence, in 1987, American companies prepared themselves for commercial use of improved versions of their launchers after restarting the production lines and launcher component supply channels. Since 2000, there has been a lot of restructuring of the US aerospace industry, the development of new launch vehicles, and partnerships with former Eastern bloc companies to operate Russian launch vehicles. Now there are some new systems under development including a space launch system (SLS), New Glenn, interplanetary transport system (ITS), OmegA launcher, Vector-R, and Vulcan launch vehicles.

11.2.9.1 Delta

The programme for the Delta launcher, developed by McDonnell-Douglas Astronautics Company (MDAC; now Boeing), was initiated at the end of the 1950s. The first firing of a Delta launcher took place in 1960; it had a capacity of 54 kg in GTO. Successive modifications have led to the Delta 3920 PAM launcher that, in 1982, was capable of injecting 1270 kg into GTO due, in particular, to the use of the PAM upper stage as the third stage. In 1984, NASA decided to abandon the programme; but, following the unavailability of a means of launching during 1986, MDAC obtained a contract from the US Air Force in January 1987 for development of the Delta II programme (the MLVII launcher, particularly for launching the Navstar/GPS satellites). At the same time, MDAC prepared to provide commercial launching services.

To serve commercial customers, agreements were made with the US Air Force and NASA for use of two government-owned launchpads at Space Launch Complex 17, Cape Canaveral, Florida, and one pad at Space Launch Complex 2, Vandenberg Air Force Base, California.

11.2.9.1.1 Delta II

Two steps were anticipated in the development of the Delta II launcher. The first firing of launcher Delta 6925, capable of injecting 1447 kg into GTO, took place on 14 February 1989. In comparison with the 3920 PAM-D launcher, the increase (164 kg) in performance of Delta 6925 was obtained by an increase, to 96.5 tons, in the capacity of the reservoirs of the first stage, which burns kerosene and LOX, and by replacement of the nine additional Castor IV solid boosters with Castor IVA solids of higher performance. The second stage remained unchanged. The third stage consists of the upper PAM-D stage that used the STAR 48 motor.

11.2.9.1.2 Delta III

Developed to address the growing lift requirements of the commercial launch market, the Delta III provides a GTO capability of 3810 kg, twice the payload of Delta II. First launch was in August 1998. Notable features of the Delta III include a cryogenically propelled single-engine upper stage, bigger and more powerful strap-on solid rocket motors than the Delta II, and a larger composite fairing to house bigger payloads.

The Delta III first stage is powered by a Boeing RS-27A main engine and two vernier engines to control roll during main engine burn and attitude control between main engine cut-off and second-stage separation. The diameter of the booster fuel tank was increased from the Delta II

to reduce length and improve control margins. Nine 1.17-meter diameter Alliant Techsystems strap-on solid rocket motors augment first-stage performance and are directly evolved from the Delta II graphite epoxy motors (GEMs), but provide 25% more thrust. Three of the strap-ons are equipped with thrust vector control to further improve vehicle manoeuvrability and control. The Delta III second-stage Pratt & Whitney RL 108-2 cryogenic engine is derived from the RL 10 engine. Cryogenic fuels produce more energy, allowing lift of heavier payloads. The RL 108-2 also incorporates a larger exit cone for increased specific impulse and payload capability.

Delta III incorporates an inertial flight control avionics system using ring laser gyros and accelerometers to provide redundant three-axis attitude and velocity data. From the 3 m diameter composite fairing developed for Delta II, a new 4 m composite fairing has been developed. In response to industry requirements, Boeing encloses Delta III payloads within the fairing at the payload processing facility before transporting the entire package to the launchpad for launch vehicle integration.

11.2.9.1.3 Delta IV

The Delta IV launch vehicle has been developed in the framework of the Evolved Expendable Launch Vehicle (EELV) programme of the US Air Force. The EELV programme development and procurement cycle began in 1995. During the first phase, four competitors completed a 15-month contract to validate low-cost concepts. In December 1996, two contractors were selected to participate in the second phase, known as the Pre-Engineering, Manufacturing, and Development phase.

All Delta IV variants use the Boeing AS-68, LH2 and LOX burning, 2900 kN thrust engine and common booster core (CBC). The engine is 30% more efficient than conventional LOX/kerosene engines. Delta II or Delta III upper stages are added to the CBC, completing each vehicle.

The Delta IV family includes five launch vehicles: Medium, three variants of the Medium vehicle known as Medium-plus, and Heavy. Figure 11.20 illustrates the whole Delta family.

The Delta IV Medium has a performance of 4210 kg to GTO using the cryogenic second stage engine of the Delta III and the 4 m composite fairing of Delta III for payload protection. Commercial derivatives of the Medium-class launch vehicle retain the Delta IV CBC and are distinguished by the number of Alliant Techsystems solid rocket motors attached to the booster core, along with the size of the upper stages and payload fairings. The Delta IV Medium-plus family is composed of:

— Delta IV Medium-plus (4,2) with two solid rocket motors and a 4 m fairing for a GTO payload of 5845 kg
— Delta IV Medium-plus (5,2) with two solid rocket motors and a 5 m fairing for a GTO payload of 4640 kg
— Delta IV Medium-plus (5,4) with four solid rocket motors and a 5 m fairing for a GTO payload of 6565 kg

The Delta IV Heavy, with its height of 72 m and weight of 733 tons, has a capacity ranging from 4300 to 12 980 kg into GTO, and can also lift over 23 000 kg into LEO. It links three of the CBCs together for lift-off and adds a modified and enlarged Delta III upper stage engine with larger tanks for increased propellant. This launch vehicle uses the 5-m metallic fairing that Boeing manufactures for the Titan IV launch vehicle or a 5-m composite fairing derived from the 4-m

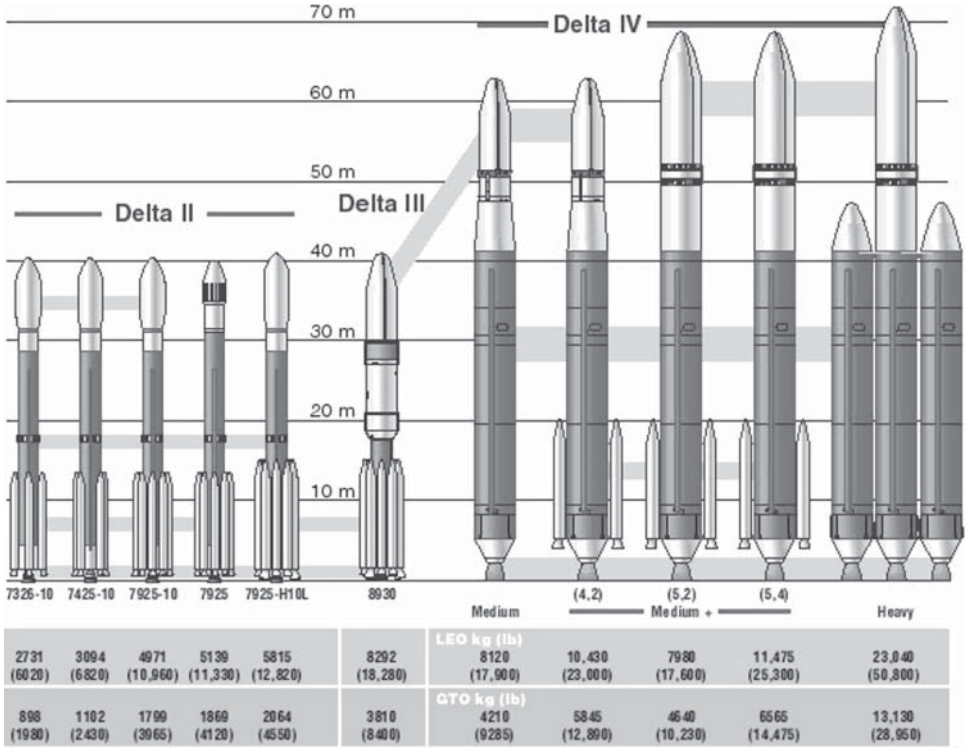


Figure 11.20 The Delta family of launch vehicles.

Delta III composite fairing. The next version of Delta IV Heavy, thanks to an increased engine thrust and the addition of new boosters, plans to deliver a 95 tons payload into LEO.

Launch preparation and testing of the Delta IV take place at Cape Canaveral and at Vandenberg Air Force Base. Both sites process rockets horizontally, away from the launchpad, to reduce pre-launch on-pad time from 24 days to 6–8 days.

The Vulcan Launcher has been developed since 2014 as the new launch vehicle, equipped with a new engine and aiming for a first launch in 2021. It has height 58.3 m and mass 546 700 kg with the capacity to deliver a payload of 34 900 kg to LEO 16 300 kg to GTO and 7200 kg to GEO.

11.2.9.2 Atlas

The Lockheed-Martin launch vehicle also has a long history [BON-82]. An Atlas launched the first communications satellite SCORE in 1958. The Centaur stage was then developed to form the two-stage Atlas/Centaur launcher. The Atlas family includes, in particular, the Atlas H, which was capable of placing 1960 kg in low orbit; and the Atlas G, which served as the first stage of the commercial Atlas/Centaur or Atlas I. This launcher was built for commercial use (the first commercial flight was on 15 July 1990) by General Dynamics to provide launching and associated services within the agreement signed with NASA and the US Air Force for use of government installations such as the 36 B and 36 A launchpads at Cape Canaveral.

11.2.9.2.1 Atlas I/II

The Atlas booster uses a Rocketdyne MA-5A stage-and-a-half propulsion system with two booster engines and one sustainer engine burning a combination of LOX and RP-1 propellant (total thrust of 2.18 MN). The four solid strap-on Castor IVA rocket boosters act to augment thrust of the booster stage in pairs. Two of the four solid rocket boosters ignite at lift-off, contributing an additional 876 kN of thrust. The second pair of solid rocket boosters fires during flight after the first pair burns out.

The Centaur upper stage is powered by two Pratt & Whitney RL10 engines burning LOX and LH2 delivering up to 190 kN of thrust. Guidance is provided by an INU located on the Centaur. This system also enables the satellite to be delivered into the transfer orbit with the desired altitude by reorientation of the Centaur after the second extinction of the motors.

Two fairings have been developed. One is called the MPF (medium payload fairing) and has an external diameter of 3.3 m (useful diameter 2.9 m); the other is called the LPF (large payload fairing) and has an external diameter of 4.2 m (useful diameter 3.65 m). The possibility of double launches has not been retained. Depending on the fairing used, the performance of the Atlas I was 2340 kg with the MPF fairing and 2220 kg with the LPF fairing. Atlas II differed from Atlas I in having elongated stages, increased first stage thrust, and a modification of the cryogenic propellant mixture of the Centaur stage. Performance was increased to 2810 kg in GTO.

11.2.9.2.2 Atlas IIAS

An increase of the capacity to 3180/3066 kg, depending on the fairing used, is achieved with Atlas IIA by increasing the quantity of propellant embarked and improvement of motor performance. A new control system is also used. By the addition of four extra Castor IV boosters to increase lift-off thrust, the payload lift capability is in the 3490–3720 kg class range to GTO with the Atlas IIAS.

11.2.9.2.3 Atlas III

To remain competitive in the commercial market, Lockheed-Martin is pursuing improvements in reliability, performance, and cost. By reducing the number of engines, recurring cost can be significantly reduced. The main modifications between the standard Atlas IIAS and the new Atlas III are replacement of the Atlas 1 and 1/2 stage including the four solid boosters, with a single-stage booster without solids; replacement of the two Centaur motors with one; and reduction of the number of staging events from four to one. The single-stage Atlas IIIA booster uses a high-performance RD-180 propulsion system produced by a US–Russian joint venture consisting of Pratt & Whitney (US) and NPO Energomash (Russia). The RD-180 burns LOX and RP-1 propellant and develops a lift-off (sea-level) thrust of 2.6 MN. The RD-180 throttles to various levels during atmospheric ascent to effectively manage the air-loads experienced by the vehicle enabling minimum Atlas vehicle and launch site infrastructure changes. Additionally, throttling results in satellite experienced flight environments that are nearly identical to Atlas IIAS. The Centaur IIIA upper stage is powered by one Pratt & Whitney RL10A engine burning LOX and LH2.

The typical Atlas launch sequence is illustrated in Figure 11.10. In a typical Atlas IIIA launch, the vehicle two RD-180 engines are ignited shortly before lift-off. Pre-programmed engine thrust settings are used during booster ascent to minimise vehicle loads by throttling back during peak transonic loads/high dynamic pressure regions while otherwise maximising

vehicle performance. Booster engine cut-off occurs approximately three minutes into flight and is followed by separation of Centaur from Atlas. The first Centaur burn lasts about nine minutes after which the Centaur and its payload coast in a parking orbit. During the first burn, approximately 10 seconds after ignition, the payload fairing is jettisoned. The second Centaur ignition occurs about 23 minutes into the flight, continues for about 3 minutes, and is followed several minutes later by the separation of the spacecraft from Centaur.

The Atlas IIIA is capable of delivering payload system weights in the 3400–4060 kg range to GTO. The first Atlas IIIA flew successfully on 24 May 2000.

The Atlas IIIB is the next incremental enhancement to the Atlas product line. With incorporation of a 1.68-m stretched Centaur upper stage, the Atlas IIIB is capable of delivering payload system weights of 4500 kg to GTO using Centaur in a dual-engine configuration.

11.2.9.2.4 Atlas V

The Atlas V launch vehicle system incorporates a common core booster (CCB) stage, uses the Centaur upper stage and payload fairings developed for the Atlas II and Atlas III series launch systems, and could be fitted with a variable number of solid rocket boosters. For the Atlas V vehicle family, a three-digit vehicle designator identifies configuration options: the first digit of the vehicle designator signifies the payload fairing (4 or 5 m); the second digit (0–5) identifies the number of solid rocket boosters used; and the third digit identifies the number of Centaur engines employed on the Centaur upper stage.

The single-stage Atlas V booster is based on the structurally stable CCB stage and is powered by the RD-180 propulsion system developing 3.8-MN of thrust. Throttling is used late in the ascent to manage vehicle acceleration. Atlas V flight environments are similar to those found on Atlas II and Atlas III series vehicles. The CCB stage is scarred for the addition of solid rocket boosters. The Atlas V solid rocket boosters by Aerojet each have a fuelled mass of approximately 46 260 kg and develop a thrust in excess of 1.11 MN.

The Atlas V 400, configured with an extended-length LPF and a single-engine Centaur, is capable of delivering a payload system weight of 4950 kg to GTO. It had its first launch in 2002.

The Atlas V 500 series launch vehicle extends the capability of the Atlas V with the addition of a 5-m diameter (4.6-m usable diameter) payload fairing (with two length options) and solid rocket boosters. The Atlas V 500 family is capable of delivering a payload weight of between 3950 and 8650 kg to GTO. The use of two CCBs as strap-on boosters to the central core results in the Atlas V HLV configuration (heavy lift vehicle). Each booster is powered by a single RD-180 engine. The expected performance is about 6350 kg directly in GEO and 19 tons in LEO. It had its first launch in 2003.

In 2004, US Congress passed legislation restricting the purchase and use of the Russia-supplied engines on Atlas V; US engines have to be used to continue its service. The United Launch Alliance (ULA) started the Vulcan Centaur launcher to replace Atlas V and Delta IV launch vehicles (see Section 11.2.9.1).

11.2.9.3 Ares

Ares is the new family of NASA launchers developed within the framework of Project Constellation. A major objective of this programme is a new lunar mission planned for 2020. The Ares I and V models will be designed in parallel. Ares IV, which should have been an intermediary launcher between the two other versions, was finally dropped.

Unlike the Space Shuttle, where both crew and cargo are launched simultaneously on the same rocket, the project has two separate launch vehicles: Ares I for the crew and Ares V for the cargo. This will allow each rocket to be designed specifically for a precise objective.

Ares I is the crew launch component of the project, with a payload capability of about 25 tons into LEO. The Ares I rocket is specifically designed to launch the Orion crew vehicle. Orion is a crew capsule, similar in design to the Project Apollo capsule, to transport astronauts to the ISS, the Moon, and possibly Mars. The first stage of Ares I is a more powerful and reusable solid fuel rocket derived from the current Space Shuttle solid rocket booster. The upper stage is propelled by one J-2X rocket engine fuelled by LH2 and LOX.

Ares V is the cargo launch component of the project, with a height of 94 m and a diameter of 5.5 m. It will launch the Earth Departure Stage and Altair lunar lander. The Ares V will complement the Ares I. The Ares V, a two-stage vehicle, will be able to carry about 188 tons to LEO and 71 tons to the Moon. The first stage of the launcher is made of six RS68 engines (LOX/LH2) under a tank similar to Ares I. These engines are similar to those used for Delta IV. The second stage has a height of 32 m and a diameter of 5 m. It is based on a J-2X engine, derived from the launcher Saturn 1B.

11.2.9.4 Falcon

In June 2000, the billionaire Elon Musk created Space Exploration Technologies Corp (SpaceX) in order to develop a launcher able to significantly reduce launch costs. Three vehicles (Falcon 1, 5, and 9) are considered to cover all orbits and a maximal payload. The Falcon can be launched at Vandenberg AFB in California and on the isle of Omelek (Marshall Islands) in the Pacific.

11.2.9.4.1 Falcon 1

This is a small launcher (21 m and 27 tons) with two stages. The first aluminium-built stage is reusable. The second shows a good trade-off between mass and mechanical resistance thanks to a mix of aluminium and lithium. The two engines were designed by SpaceX: a Merlin motor for the first stage and a Kestrel motor for the second stage, both based on the Pintle injector (used in the 1970s with Apollo). The payload capacity is 670 kg on LEO and 430 kg on SSO.

The first flight of the Falcon 1 occurred on 24 March 2006. It ended with a failure due to a fuel line leak and subsequent fire. The launch took place from Omelek Island. The second test flight was originally scheduled for January 2007 but was delayed due to problems with the second stage. After an aborted attempt, the rocket was launched on 21 March 2007 with a DemoSat payload for DARPA and NASA that was not able to reach the desired orbit. Finally, the first successful mission of Falcon 1 occurred on 28 September 2008. Falcon 1 v1.0's first flight was on 4 June 2010 and v1.1's on 29 September 2013. Both v1.0 and v1.1 have now retired to be replaced by the new generations.

Falcon 5. Plans were based on a two-stage launch vehicle, using five Merlin engines for the first stage and one for the second. The Falcon 5 has been cancelled.

11.2.9.4.2 Falcon 9

This vehicle uses the same engines, electronic systems, and control management as Falcon 1. Falcon 9 uses nine Merlin motors in parallel. Two versions of this vehicle are offered to the customer now:

— *Falcon 9 full thrust*: Partially reusable with height 71 m and weight 549 054 kg, capable of delivering a payload up to 22 800 kg to LEO (28.5°), 5000 kg in GTO (27°), and 4020 kg to Mars. The first launch was on 22 December 2015.
— *Falcon 9 heavy:* Partially reusable with height 70 m and weight 1 420 788 kg, capable of delivering a payload up to 63 800 kg to LEO (28.5°), 26 700 kg in GTO (27°), 16 800 kg to Mars, and 3500 kg to Pluto. The first launch was on 6 February 2018.

11.2.9.5 Sea Launch

The Sea Launch system is a marine-based launch operation concept that provides a number of advantages over traditional satellite launch systems with land-based launch sites. This launch system is operated by the Sea Launch Limited Partnership, which includes the Boeing Commercial Space Company, the Russian space company RSC Energia, the Ukrainian aerospace organisation KB Yuzhnoye/PO Yuzmash, and the marine operators Kvaerner Group of London.

The Sea Launch system consists of the Zenit-3SL rocket and launch platform, the assembly and command ship (ACS), and the Home Port facility in Long Beach, California. The launch platform is a self-propelled, semi-submersible platform originally built for off-shore oil drilling. The platform includes systems for conducting launch vehicle erection, fuelling, and launch operations. The ACS is the command centre for launch operations and also provides facilities for integration of the payload with the Zenit-3SL prior to transfer of the launch vehicle to the launch platform.

Spacecraft processing occurs in facilities in southern California. The vessel sails from the home port to launch sites in the Pacific, tailored to cater to a wide range of orbits. Upon arrival at the selected site, the platform is submerged to its launch position and the vehicle is rolled out and erected. The launch is controlled from the command ship, with communication links available to satellite manufacturer and customer sites. The command ship provides accommodation for the launch team members and houses the automated launch control centre.

11.2.9.5.1 Launch Vehicle

The Zenit-3SL launch vehicle is based on the proven technologies of the land-based Zenit-2 (KB Yuzhnoye/PO Yuzmash) and the Block-DM upper stage (RSC Energia). The Zenit launch vehicle is the most modern heavy-lift launch vehicle developed by the former Soviet Union (see Section 11.2.3.7). It has fully automated pre-launch servicing, is integrated horizontally, and is transported to the launchpad and erected by a single transport unit [ROS-94]. Zenit uses LOX and kerosene. Stage I is powered by a single turbo-pump RD-170 engine with four thrust chambers, delivering a sea-level thrust of 7240 kN. Stage II is powered by an RD-120 main engine with a single thrust chamber and an RD-8 vernier with four thrust chambers, which produce a total vacuum thrust of 912 kN.

The Block-DM upper stage was developed as the fourth stage of the Proton launch vehicle. The Block DM-SL is a restartable upper stage; it can be fired up to seven times during a mission, operates on LOX and kerosene, and produces 790 kN of thrust. Three-axis stabilisation of the Block-DM during coast periods is provided by two vernier engines.

11.2.9.5.2 Launch platform

The launch platform, adapted from an existing semi-submersible oil rig, provides transportation of the integrated launch vehicle (ILV) to the launch site and automated launch support

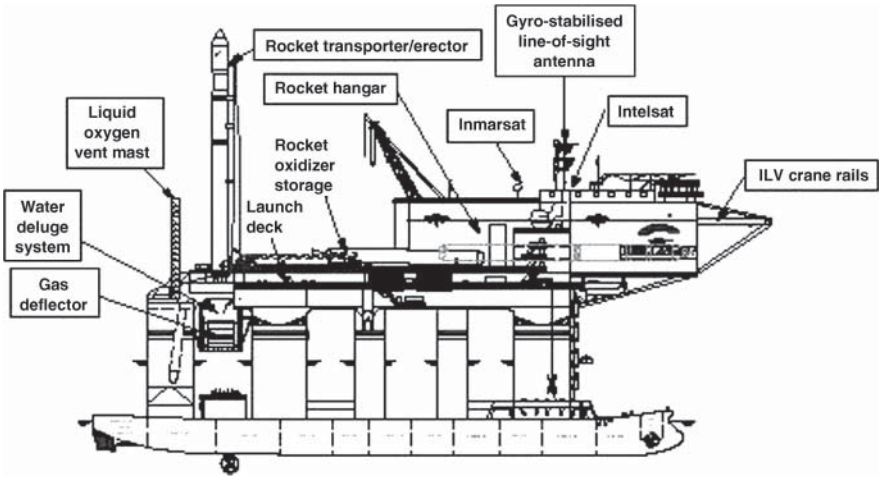


Figure 11.21 Launch platform and Zenit-3SL.

equipment, including the launchpad (Figure 11.21). The launch platform is self-propelled and rides catamaran style on a pair of pontoons. Once at the launch location, the pontoons are submerged by ballasting to achieve a stable launch position. The platform has an overall length of approximately 133 m. Its overall displacement is approximately 26 360 tons.

11.2.9.5.3 Assembly and command ship

The assembly and command ship illustrated in Figure 11.22 provides three primary functions for Sea Launch operations:

— Facility for assembly, processing, and checkout of the launch vehicle
— Mission control centre to monitor and control all operations
— Accommodation for marine and launch crews

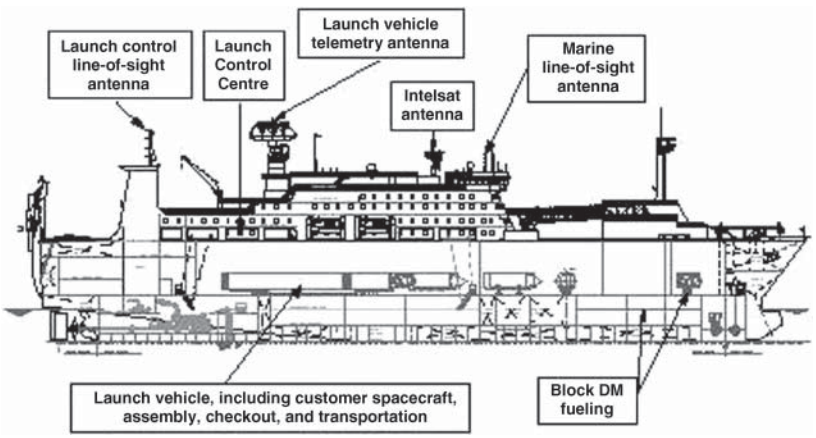


Figure 11.22 Assembly and command ship (ACS).

11.2.9.5.4 Performance

The performance capability is 5250 kg to a standard GTO 200 35 786 km; 0° inclination [LOC-01]. A wide range of injection orbits and inclinations are possible due to the restart capability of the Block DM-SL. The equatorial location of the standard launch site eliminates the need for plane-change manoeuvres to reach a 0° orbital inclination. In terms of spacecraft mass to final orbit, this advantage means a Sea Launch 5250 kg GTO payload capability would be equivalent to approximately 6000 kg of payload launched from Cape Canaveral. Another advantage of the standard launch site location is that the spacecraft is launched where it is largely unconstrained by terrain overflight and fishing vessel clearance considerations and can be launched to any desired azimuth. In addition, the combination of benign weather, inherent platform stability, and an active levelling system make it possible to launch year round. Additional performance capability increases up to 6000 kg in GTO are being implemented, together with an increase to 4.57 m in the diameter of the payload fairing.

11.2.9.5.5 Launch operations

The launch vehicle and spacecraft integration activities are performed at the Home Port for the Sea Launch system in Long Beach, California. The customer spacecraft is encapsulated and then integrated to the Zenit-3SL, forming the ILV. The ILV is loaded onto the launch platform (LP) for transportation to the launch site. The standard launch site location is on the equator in the Pacific Ocean at coordinates 0 °N, 154 °W. This site was chosen for the performance advantage of launching at 0° latitude, favourable weather conditions, absence of land and major shipping lanes, and the operational advantages of close proximity to Christmas Island, which provides a retreat in case of medical emergencies and an alternative delivery location for personnel and supplies. Nominal transit time from Home Port to the launch site is 11 days. From encapsulation through launch, the PLU maintains a climate-controlled environment for the spacecraft. Upon arrival at the launch site, the LP transitions from transit draft, 7.5 m, to launch draft, 22.5 m, using a system of ballasting tanks and pumps to displace 18 000 tons of water. The ballasting operation takes approximately six hours, during which time the ILV remains horizontal on the transporter/erector in the hangar. Once the LP reaches launch draft, the station-keeping system using the global positioning system (GPS) inputs maintains the LP heading within three degrees and station within 150 m. A fine levelling system maintains a level attitude within an estimated accuracy of about 1° by pumping water between the three tanks located in the columns of the LP, with pumping capability up to 1500 m^3 h^{-1} of water.

Approximately 24 hours prior to launch, the fully automated transporter/erector (T/E) rolls out of the hangar, carries the unfuelled ILV to the launchpad, and erects it to the launch position. The launch may be accomplished in significant wave heights up to 2.5 m, and most of the waves in the launch region are of low period and small wave height.

11.2.9.5.6 Land Launch

This subsidiary of Sea Launch conducts commercial launches from the Baikonur Cosmodrome thanks to a Zenit-3SLB rocket. This version of Zenit-3SL has been modernised (using a lighter payload fairing) in order to insert the payload directly into a geosynchronous orbit, rather than leaving it in a geosynchronous transfer orbit. The first launch was made with success on 28 April 2008, when a Zenit-3SLB was used to place AMOS-3 (Israeli communication satellite with 15 Ku/Ka transponders) into a geosynchronous orbit.

11.2.9.6 Pegasus

Pegasus is the first winged launcher, developed by Orbital Sciences and Hercules. This vehicle is launched from a plane at 12 000 m altitude flying at Mach 0.8. It bypasses weather conditions because the plane is above the clouds and does not require ground installations, which are very expensive.

Pegasus exists in two versions (length about 17 m, mass from 18 500 to 23 130 kg for the XL vehicle). The payload capacity of both launchers is 400 kg. Since 1990, 43 missions have been launched with 3 failures, 2 partial successes, and 38 full successes.

Several projects are associated with Pegasus. Indeed, Pegasus components have been the basis of other Orbital Sciences launchers. For instance, the launcher JXLV is a version of Pegasus with one stage, which has been used for launching the NASA X-43 Scramjet (supersonic combustion Ramjet). Taurus and Minotaur, also developed by Orbital Sciences, are two other examples of launchers derived from Pegasus. They are used for military applications. Minotaur is the result of the transformation of the two first stages of a Minuteman II rocket, with the third and fourth stages derived from Pegasus. It is also called the OSP Space Launch Vehicle and can deliver from 340 to 580 kg into LEO depending on inclination and final altitude.

11.2.10 Reusable launch vehicles

A reusable launch vehicle (RLV) is a system with the ability to be launched several times. Although this has not been demonstrated by NASA with the partially reusable Space Shuttle, the objective of reusable systems is to provide low-cost launches and wide access to space.

However, RLVs need specific technologies to be developed. An important component is a shield able to support the heat created by friction when re-entering the atmosphere. Heat shields are typically made of ceramic or carbon-carbon tiles. The maintenance of the RLVs will be an essential part of the reliability and the cost of the launcher. Different concepts are being investigated:

— Horizontal or vertical take-off
— One or several stages
— Horizontal or vertical landing
— Different types of propellant

With the aim of reducing both the cost of launch and the environmental impact, several countries are developing reusable or partially RLVs. Falcon 1 is partially reusable, as it reuses one of its two stages; it made its first successful flight in 2008.

The launchers in development are:

— Silver Dart from PlanetSpace (USA), a two-stage vehicle expected to launch vertically and land horizontally on an aircraft runway.
— Falcon 9 from SpaceX (USA). The rockets have been partially reused. It has started the first step towards fully reusable space launchers in the future.
— K-1 from Kistler Aerospace (USA), aiming to reach a geostationary orbit.
— Hopper from ESA (Europe).
— Avatar RLV from ISRO (India), a single-stage aerobic hypersonic vehicle that takes off and lands horizontally.

11.2.11 Cost of installation in orbit

It is difficult to specify the cost of launching a satellite, since the cost depends on the type of service provided, the performance of the launcher, the commercial policy of the organisation that sells the service, etc. A rounded order of magnitude is 100 million euros for a launch capacity on the order of 5000 kg at take-off (about 3100 kg in geostationary orbit): that is, about 20 000 euros per kilogram at take-off or about 30 000 euros per kilogram in geostationary orbit.

There is harsh competition between the half-dozen launch companies able to place today's large communications satellites (of a few tons) into orbit. However, cost comparisons between one launcher and another is not easy, since it is not sufficient to compare the cost and capacity placed in orbit. For the same capacity, launchers differ in a large number of characteristics – inclination of the transfer orbit, accuracy with which the orbit is obtained, useful volume under the fairing, static and dynamic mechanical constraints (such as longitudinal and transverse acceleration, vibration, shock, noise spectrum), thermal constraints, interfaces, etc. These characteristics can have a very significant impact on the system design, satellite lifetime, and hence the overall cost of the system.

With modern high-capacity launchers that are capable of multiple launches, the issue of sharing the cost of launching among users is also to be considered. The price is usually invoiced in ratio to the mass to be carried, taking into account the constraint imposed by the adaptors.

REFERENCES

[AR5-16] Arianespace. (2016). Ariane 5 users' manual, issue 5, revision 2.

[AR6-18] Arianespace. (2018). Ariane 6 user's manual, issue 1, revision 0.

[BON-82] Bonesteel, M.M. (1982). Atlas and Centaur adaptation and evolution—27 years and counting. In: *IEEE International Conference on Communications, Philadelphia*, 3F.2.1–3F.2.8. IEEE.

[BZH-97] Bzhilianskaya, L. (1997). Russian launch vehicles on the world market: a case-study of international joint ventures. *Space Policy* **13** (4): 323–338.

[HOH-25] Hohmann, W. (1925). *Die Errichbarkeit der Himmelsk€ orper*. Munich: Odelbourg.

[LOC-01] Locke, S. (2001). Key design and operation aspects of the Sea Launch system. Paper 047, presented at the AIAA 19th Communication Satellite Systems Conference, Toulouse.

[MAR-79] Marec, J.P. (1979). *Optimal Space Trajectories*. Elsevier.

[POC-86] Pocha, J.J. and Webber, M.C. (1986). Operational strategies for multi-burn apogee manoeuvres of geostationary spacecraft. *Space Communication and Broadcasting* **4** (3): 229–233.

[PRI-86] Pritchard, W.L. and Sciulli, J.A. (1986). *Satellite Communications—Systems Engineering*. Prentice Hall.

[RAJ-86] Rajasingh, C.K. and Leibold, A.F. (1986). Optimal injection of TVSAT with multi-impulse apogee manoeuvres with mission constraints and thrust uncertainties. In: *Mecanique Spatiale pour les Satellites Geostationnaires, Colloque CNES*, 493–510. Cepadues.

[ROB-66] Robbins, H.M. (1966). An analytical study of the impulsive approximation. *AIAA Journal* **4** (8): 1417–1423.

[ROS-94] Rossie, J. and Forrest, J. (1994). Zenit at Baikonur: unique automated launch system. *Spaceflight* **36**: 326–327.

[SKI-86] Skipper, J.K. (1986). Optimal transfer to inclined geosynchronous orbits. In: *Mécanique Spatiale pour les Satellites Géostationnaires*, 71–84. Cepadues.

[STA-88] Stadd, A. (1988). Status and issues in commercializing space transportation. Presentation at the AIAA 12th Communications Satellite Systems Conference, Arlington, Virginia.

[WHI-90] White, R. and Platzer, M. (1990). ATLAS family update. Paper 90–0827, presented at the AIAA 13th Communication Satellite Systems Conference, Los Angeles.

12 THE SPACE ENVIRONMENT

This chapter discusses how the space environment affects the design and operation of the satellite during its *lifetime* in orbit. Of relevance are the following:

— Absence of atmosphere (vacuum)
— Gravitational and magnetic fields
— Meteorites and debris
— Radiation sources and sinks
— High-energy particles

The particular environment during *injection* of the satellite into orbit (acceleration, vibration, noise, and depressurisation) should also be considered.

The effects of the environment on the satellite are principally as follows:

— *Mechanical*, consisting of forces and torques that are exerted on the satellite and modify its orbit and attitude
— *Thermal*, resulting from radiation from the sun and earth absorbed by the satellite and energy radiated towards cold space
— *Degradation of materials* subjected to the action of radiation and high-energy particles

12.1 VACUUM

12.1.1 Characterisation

Vacuum is an inherent feature of the space environment. The molecular density diminishes nearly exponentially with altitude; the variation law depends on the latitude, time of day, solar activity, etc. At 36 000 km (the geostationary satellite altitude), the pressure is less than 10^{-13} torr (millimetres of mercury). 1 torr = 1/760 atm (101 325 Pa) = 101 325/760 Pa = 133.32 Pa.

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

12.1.2 Effects

12.1.2.1 Mechanical effects

The effect of atmospheric drag due to an imperfect vacuum has been considered in Section 2.3.1.4. The altitude of the apogee of an elliptic orbit tends to decrease, as does the altitude of a circular orbit. Above 400 km, atmospheric drag can be considered to be negligible.

12.1.2.2 Effects on materials

In vacuum, materials sublime and outgas; the corresponding *loss of mass* depends on the temperature (for example: 10^3 Å/year at 110 °C, 10^{-3} cm/year at 170 °C and 10^{-1} cm/year at 240 °C for magnesium), where 1 Å = 10^{-10} m. As temperatures greater than 200 °C are easy to avoid, and on the condition that excessively thin skins are not used, these effects are not important. The possibility of condensation of gases on cold surfaces is more serious (it can cause short circuits on insulating surfaces and degrade thermo-optical properties); it is thus necessary to avoid the use of materials that sublime too easily, such as zinc and caesium. Furthermore, polymers tend to decompose into volatile products.

On the other hand, a major advantage of vacuum is that metals are preserved from the effects of corrosion.

The surfaces of certain materials, particularly metals, when subjected to high-pressure contact, have a tendency to diffuse into each other by a cold welding process; the result is a large frictional force on bearings and the moving mechanisms (for example, the deployment of solar generators and antennas). It is therefore necessary to keep moving parts in sealed pressurised enclosures and use lubricants having a low rate of evaporation and sublimation. Special materials (e.g. ceramic and special alloys such as stellite) are also used for bearing manufacture.

12.2 THE MECHANICAL ENVIRONMENT

12.2.1 The gravitational field

12.2.1.1 The nature of the gravitational field

The satellite is, above all, subjected to the earth's gravitational field, which primarily determines the movement of the centre of mass of the satellite. This gravitational field has asymmetries, due to the non-spherical and inhomogeneous nature of the earth, which cause perturbations of the orbit. Perturbations also result from the gravitational fields due to the attraction of the sun and moon. These gravitational fields have been described in Chapter 2.

12.2.1.2 The effect on the orbit

The asymmetry of the earth's gravitational field and the attraction of the sun and moon cause perturbations of the Keplerian orbit of the satellite as defined by the attraction of the earth when assumed to be spherical and homogeneous. These perturbations lead to variation with time of the parameters that define the orbit (see Section 2.3).

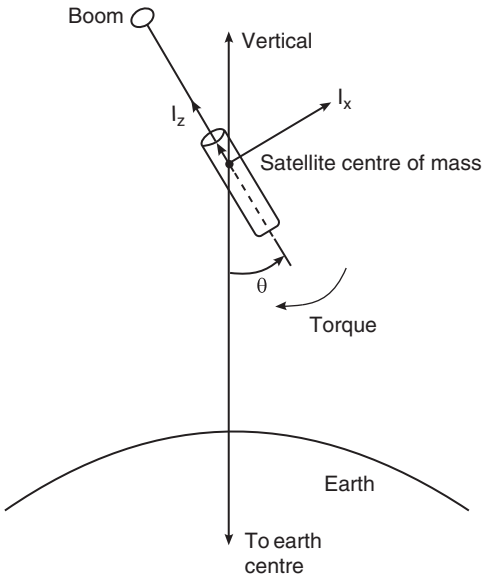


Figure 12.1 Gravity gradient generated torque.

12.2.1.3 Effect on the satellite attitude

The strength of the earth's gravitational field varies with altitude so that parts of the satellite that are more distant from the centre of the earth are less attracted than the nearer parts. Since the result of this gravity gradient does not pass through the centre of mass of the satellite, a torque is created, as shown in Figure 12.1.

The earth's gravity gradient has the effect of aligning the axis of lowest inertia of the satellite along the local vertical. Assuming that the z axis is an axis of symmetry of the satellite, the corresponding torque is given by:

$$T = 3(\mu/r^3)(I_z - I_x)\theta \quad \text{(Nm)} \tag{12.1}$$

where μ is the attraction constant of the earth, r is the distance of the satellite from the centre of the earth, I_z is the moment of inertia about the z axis, I_x is the moment of inertia about an axis perpendicular to the z axis (smaller than I_z), and θ is the angle, assumed to be small, between the z axis and the direction of earth centre.

This torque, which can be used to stabilise satellites in low orbit, is too small to use for stabilising geostationary satellites. Then the design should be such as to make its effects negligible. It is sufficient to make I_x and I_z not greatly different from each other. For example, with $I_z = 180\,\text{m}^2\,\text{kg}$ and $I_x = 100\,\text{m}^2\,\text{kg}$, the maximum torque is $T = 2.2 \times 10^{-7}\,\text{N m}$ for θ less than 10°.

12.2.1.4 Lack of gravity

As the terrestrial attraction is in equilibrium with the centrifugal force, the various parts of the satellite are not subject to gravity. This is particularly important for liquid propellants, which

cannot be extracted by gravity from the reservoirs in which they are stored. It is necessary to install a pressurising system with artificial separation of the liquid and gas either by means of a membrane or by using the properties of surface tension forces (see Section 10.3.2.5).

12.2.2 The earth's magnetic field

12.2.2.1 Characterisation of the terrestrial magnetic field

The terrestrial magnetic field H can be considered to be that of a magnetic dipole of moment $M_E = 7.9 \times 10^{15}$ Wb m. This dipole makes an angle of 11.5° with the axis of rotation of the earth. It thus creates an induction **B**, which has two components as follows:

— A normal component:

$$B_N = (M_E \sin\theta)/r^3 \qquad (\mathrm{Wb/m^2}) \tag{12.2a}$$

— A radial component:

$$B_R = (2M_E \cos\theta)/r^3 \qquad (\mathrm{Wb/m^2}) \tag{12.2b}$$

where r is the distance of the point concerned from the centre of the earth and θ is the angle between the radius vector and the axis of the dipole (using the polar coordinates of the point concerned in the reference system associated with the dipole).

For a geostationary satellite, the normal component varies between 1.03×10^{-7} and 1.05×10^{-7} Wb m^{-2}, and the radial component within $\pm 0.42 \times 10^{-7}$ Wb m^{-2}. The component perpendicular to the equatorial plane is virtually constant and equal to 1.03×10^{-7} Wb m^{-2}.

12.2.2.2 The influence of the terrestrial magnetic field

The terrestrial magnetic induction **B** exerts a torque **T** on a satellite of magnetic moment **M** such that:

$$\mathbf{T} = \mathbf{M} \wedge \mathbf{B} \qquad (\mathrm{Nm}) \tag{12.3}$$

For a geostationary satellite, the component of induction perpendicular to the equatorial plane, although the largest and most constant, produces the smallest long-term effect. The corresponding torque is in the plane of the equator; and since the satellite performs one complete rotation about its axis parallel to the axis of the poles per day, the average torque is zero.

The overall magnetic moment of a satellite results from remanent moments, moments due to electric currents in the cabling, and induced moments proportional to the earth's magnetic field. These moments can be reduced or compensated for before launching so that the torque due to the earth's magnetic moment on the ground does not exceed 10^{-4} N m. As the magnetic field is inversely proportional to the cube of the distance from the centre of the earth, the torque in the geostationary satellite orbit is divided by $(42\,165/6378)^3 = 289$. It thus becomes equal to 3.5×10^{-7} N m. In practice, the launching conditions modify some of the settings made on the ground; a safe design should therefore introduce some margin and consider a torque equal to 10^{-6} N m for the disturbing torque due to the earth's magnetic field when dimensioning the satellite attitude control system.

The earth's magnetic field can also be used in an active manner to generate satellite attitude-control torques by using appropriate actuators (magnetic coils; see Section 10.2.4).

12.2.3 Solar radiation pressure

The solar radiation pressure on a surface element of area dS has two components, one normal and the other tangential to the surface. They both depend on the angle of incidence θ of the solar radiation on the surface, measured with respect to the normal, as well as on the coefficient of reflectivity of the surface ρ and the intensity of the solar flux W (see Section 2.3.1.3). The effect of these forces on the movement of the centre of mass has also been discussed in Chapter 2. The resultant of the forces exerted on all the surface elements dS of the satellite does not in general coincide with its centre of mass. This results in a torque that perturbs the satellite attitude.

Each elemental force is proportional to $W\cos\theta$. The torque thus depends on the orientation of the sun with respect to the satellite. For geostationary satellites, the direction of the sun makes an angle between 66.5° and 113.5° with the axis perpendicular to the equatorial plane (the pitch axis). The torque causes a drift of the orientation of the north–south axis of the satellite. The torques due to solar radiation pressure, which are disturbing torques, can also be used in an active manner to participate in satellite attitude control (see Section 10.2.4).

12.2.4 Meteorites and material particles

The earth is surrounded by a cloud of meteorites (scrap material, rocks, pebbles, etc.) whose density becomes lower as the altitude increases. At geostationary satellite altitude, their velocity varies from several kilometres per second to several tens of kilometres per second. The commonest meteorites have masses between 10^{-4} and 10^{-1} g. The flux N of particles of mass equal to or greater than m per square metre per second can be estimated from the following equations:

— For $10^{-6}\ \text{g} < m < 1\ \text{g}$:

$$\log_{10}N(\geq m) = -14.37 - 1.213\log_{10}m$$

— For $10^{-12}\ \text{g} < m < 10^{-6}\ \text{g}$:

$$\log_{10}N(\geq m) = -14.34 - 1.534\log_{10}m - 0.063(\log_{10}m)^2 \tag{12.4}$$

12.2.4.1 Probability of impact

The motion imparted to the satellite by impact with a meteorite can be evaluated in statistical terms: that is, by the probability of meteorites of a given mass colliding with the satellite and by the resulting magnitude of the motion transferred. Collisions between the satellite and meteorites are assumed to occur randomly and to be modelled by a Poisson distribution. The probability of having n impacts with particles of mass between m_1 and m_2 on a surface S during time t is given by:

$$P(n) = [(Sft)^n \exp(-Sft)]/n! \tag{12.5}$$

where f is the flux of particles of mass between m_1 and m_2 such that $f = N(>m_1) - N(>m_2)$ with $N(>m)$ given by (12.4), S is the exposed surface area (m^2), and t the exposure time (seconds).

12.2.4.2 The effect on materials

Meteorite impact causes an erosion of around 1 Å per year at the geostationary satellite altitude. For the heaviest meteorites, these impacts can cause perforation of metal sheets, if too thin, which

could be disastrous for the survival of the satellite. Protection is possible by using screens consisting of several superimposed sheets of metal. The outer sheets fragment the meteorites, and subsequent ones halt the debris.

12.2.5 Torques of internal origin

Relative movement of the antennas, solar panels, and fuel causes torques that are exerted on the main body of the satellite. Furthermore, maintaining satellites in a stationary position requires periodic application of forces that act on the centre of mass of the satellite.

The satellite contains propellant reservoirs that empty in the course of the mission, and it is impossible to have a centre of mass that is firmly fixed with respect to the satellite body and hence with respect to the jets. Also, during integration of the satellite, mounting and alignment of the jets are subject to some inaccuracy. The correcting forces required to maintain position will, therefore, not be applied exactly at the centre of mass; a torque that disturbs attitude maintenance will arise during these corrections.

By way of example, considering thrusters with thrust of 2 N and a maximum displacement of the centre of mass of 5 mm, the value of the disturbing torque is $C_p = 10^{-2}$ N m.

12.2.6 The effect of communication transmissions

Electromagnetic radiation from the antennas creates a pressure that can be non-negligible if the transmitted power is high. For a radiating antenna, the force F produced is:

$$F = -(\mathrm{d}m/\mathrm{d}t)c = -\mathrm{EIRP}/c \qquad (\mathrm{N}) \tag{12.6}$$

where (EIRP) is the effective isotropic radiated power (W) and c is the speed of light (m/s).

For example, for a satellite with a 1 kW EIRP, the force F is 0.3×10^{-5} N. If the lever arm is 1 m, the torque is 3×10^{-6} N m.

The perturbation is large only in the case where the transmitted power is large and concentrated into a narrow beam; it is then necessary for the antenna axis to pass through the centre of mass or to provide two antennas whose axes are symmetrical with respect to the centre.

12.2.7 Conclusions

The satellite is subjected to perturbations that modify its nominal orbit and create torques that disturb the attitude. For a geostationary satellite, it was shown in Chapter 2 that the attraction of the sun and moon causes a variation of the inclination of the plane of the orbit on the order of 1° per year. The asymmetry of the terrestrial potential causes a longitude drift. Solar radiation pressure modifies the eccentricity of the orbit. Table 12.1 summarises the orders of magnitude of the disturbing torques in respect of attitude.

12.3 RADIATION

The energy radiated by a body depends on its temperature T and its emittance ε. The Stefan–Boltzmann law defines the *radiance* M of a body, which is the power (W) radiated per surface element $S(\mathrm{m}^2)$ as follows:

$$M = \varepsilon\sigma T^4 (\mathrm{W/m^2}) \tag{12.7}$$

Table 12.1 Attitude disturbing torques for a geostationary satellite

Origin	Moment of torque	Comments
Station keeping	10^{-2}	Only during corrections
Radiation pressure	5×10^{-6}	Continuous except during eclipses
Magnetic field	10^{-6}	Daily mean is less
Gravity gradient	10^{-7}	Continuous

where $\sigma = 5.67\times10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ is the Stefan–Boltzmann constant. For a black body, $\varepsilon = 1$. The *emittance* of a body is the ratio of the radiance of this body to the radiance of a black body at the same temperature.

Planck's law expresses the *spectral radiance* L_λ of a black body, which is the power per unit wavelength and per unit solid angle emitted in a given direction by a surface element of a black body, divided by the orthogonal projection of this surface element onto a plane perpendicular to the direction considered:

$$L_\lambda = C_{1\mathrm{L}}\lambda^{-5}[\exp(C_2/\lambda T) - 1]^{-1}(\mathrm{W/m^3sr}) \tag{12.8}$$

where $C_{1\mathrm{L}} = 1.19\times10^{-16}\,\mathrm{W\,m^2\,sr^{-1}}$ and $C_2 = 1.439\times10^{-2}\,\mathrm{m\,K}$.

A functional form of Planck's law, known as Wien's law, relates the wavelength λ_m for which the spectral radiance of the black body is a maximum to its temperature T:

$$\lambda_\mathrm{m}T = b$$
$$L_\mathrm{m}/T^5 = b' \tag{12.9}$$

where $b = 2.9\times10^{-3}\,\mathrm{m\,K}$ and $b' = 4.1\times10^{-6}\,\mathrm{W\,m^{-3}\,K^{-5}\,sr^{-1}}$.

Space radiates as a black body at a temperature of 5 K. It behaves like a cold sink with an absorptivity of 1; all the thermal energy emitted is completely absorbed.

The radiation received by the satellite arrives principally from the sun and the earth.

12.3.1 Solar radiation

Figure 12.2 shows the spectral irradiance of the sun as a function of wavelength for a surface located at a distance of 1 astronomical unit (AU), i.e. the mean sun–earth distance, from the sun. It can be seen that the sun behaves as a black body at a temperature of 6000 K. Of the power radiated, 90% is situated in the band 0.3–2.5 μm with a maximum in the region of 0:5 μm.

The incident flux is about $1370\,\mathrm{W\,m^{-2}}$ on a surface normal to the radiation located at 1 AU from the sun. It is the flux to which an artificial satellite of the earth is subjected for about 10 days of the year after the spring equinox (see Figure 2.5). The flux varies in the course of the year as a function of variation of the earth–sun distance (the earth–satellite distance being assumed to be negligible), and this variation is illustrated in Figure 12.3b. The power received by a surface element of the satellite depends on the orientation of this surface with respect to the direction of the incident radiation, which itself varies as a function of the declination of the sun. For a satellite in an orbit in the equatorial plane, if a surface perpendicular to the equatorial plane permanently orientated in the direction of the sun is considered, the incident flux is multiplied by the cosine of the declination d; the result is presented in Figure 12.3c. If a surface parallel to the plane of the equator is considered, the incident flux is multiplied by the sine of the declination and depends on the orientation (north or south) of the surface. The received power is zero at the equinoxes

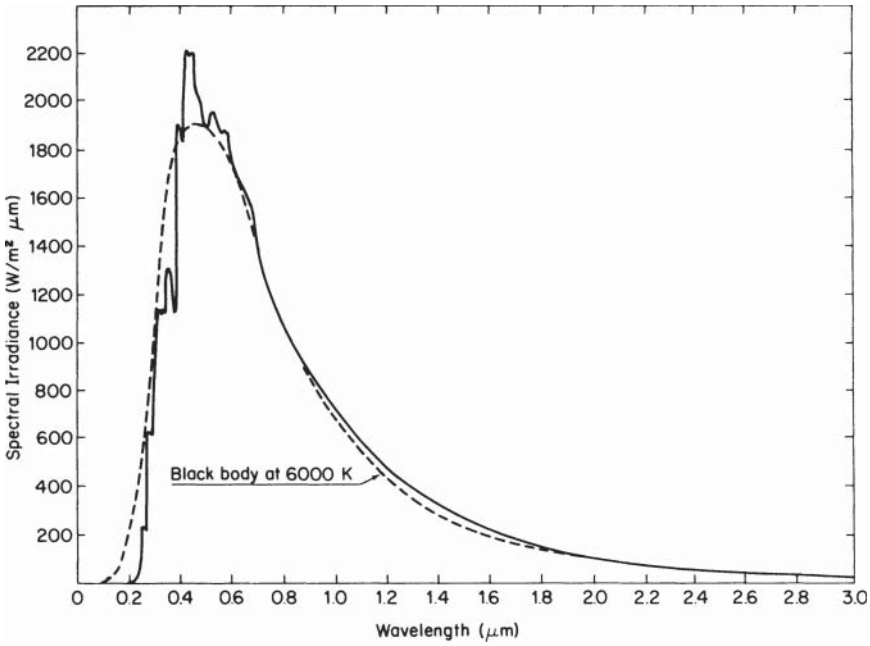


Figure 12.2 Spectral irradiance of the sun.

and during the six months of spring and summer for a surface oriented towards the north, while it is zero during the six months of autumn and winter for a surface oriented towards the south.

The apparent diameter of the sun viewed from the earth is 32 min or around 0.5°.

12.3.2 Earth radiation

Earth radiation results from reflected solar radiation (the albedo) and its own radiation. The latter corresponds reasonably to that of a black body at 250 K: that is, the irradiance is maximum in the infrared band at 10–12 μm. For a geostationary satellite, the total flux is less than 40 W m^{-2} and is thus negligible compared with that provided by the sun.

12.3.3 Thermal effects

The satellite faces in sight of the sun warm under the effect of solar radiation while the faces turned towards distant space become colder. Exchanges of heat by *conduction* and *radiation* thus occur (the vacuum prevents exchanges by convection). If the satellite retains a fixed orientation with respect to the sun, it establishes an equilibrium between the power absorbed from the sun and the heat radiated. The mean temperature results from the following thermal balance:

$$P_S + P_I = P_R + P_A \tag{12.10}$$

where P_S is the power absorbed from the direct solar flux ($P_S = \alpha W S_a$, with W the solar flux, S_a the apparent surface, and α the absorptivity), P_I is the internal power dissipated, P_R is the power radiated, and P_A is the power stored (or returned) by exchanges during temperature variation.

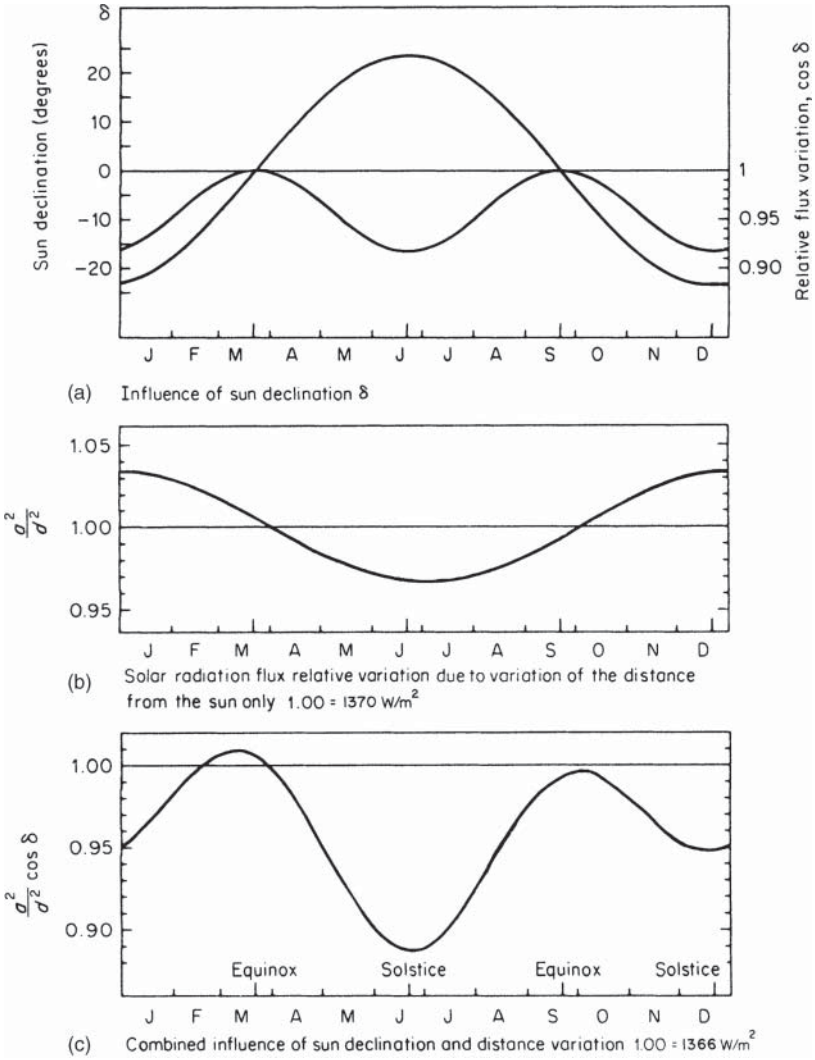


Figure 12.3 Solar radiation flux variations on a north–south sun-facing surface of a geostationary satellite. $a = 1$ IAU (semi-major axis of the earth orbit), d = sun to earth distance.

For a perfectly conducting passive sphere of radius r at the equilibrium temperature T:

$$P_S = \alpha W \pi r^2$$

$$P_I = 0$$

$$P_A = 0$$

$$P_R = \varepsilon \sigma T^4 4\pi r^2$$

P_R is the power σT^4 radiated by the total surface $4\pi r^2$ of emittance ε. The equilibrium temperature is:

$$T = [(\alpha W)/(4\varepsilon\sigma)]^{1/4} \tag{12.11}$$

Table 12.2 Equilibrium temperature of a perfectly conducting inert sphere in space at a distance of 1 AU from the sun

Surface	Absorptivity (α)	Emissivity (ε)	α/ε	T (°C)
Cold: white paint	0.20	0.80	0.25	−75
Medium: black paint	0.97	0.90	1.08	+11
Hot: bright gold	0.25	0.045	5.5	+155

Absorptivity a is the ratio of solar energy absorbed to solar energy received, and emissivity ε is the ratio of thermal flux emitted to that which would be emitted by a black body raised to the same temperature.

where σ is the Stefan–Boltzmann constant (5.67×10^{-8} W m^{-2} K^{-4}). The equilibrium temperature of this spherical inert satellite ($P_R = 0$) depends only on the thermo-optical properties of the exterior surface: that is, essentially, its colour. Table 12.2 shows the equilibrium temperature for various surfaces. It can vary from −75 to +155 °C.

The satellite equipment operates satisfactorily over a narrower range of temperatures: for example, from 0 to +45 °C. The surfaces will thus be chosen and combined judiciously to satisfy these conditions (see Section 10.6). As solar cells cover larger surfaces of the satellite, their thermo-optical properties are very important. Their absorptivity is between 0.7 and 0.8 (in open circuit) and emissivity is between 0.80 and 0.85.

12.3.4 Effects on materials

Radiation in the ultraviolet with spectrum from 100 to 1000 Å causes ionisation in materials. This has the following effects:

— Increase in the conductivity of insulators and modification of the absorptivity and emissivity coefficients of thermo-optical surfaces
— Decrease of the conversion efficiency of solar cells with time spent in orbit (the order of magnitude is a 30% decrease for silicon cells and about 20% for GaAs cells after 10 years)

At wavelengths greater than 1000 Å, solids can be excited; polymers are discoloured, and their mechanical properties are weakened.

Above 3000 Å, the effects on metals and semiconductors are practically zero.

12.4 FLUX OF HIGH-ENERGY PARTICLES

12.4.1 Cosmic particles

Cosmic particles are charged particles that consist mainly of high-energy electrons and protons; they are emitted by the sun and various sources in space. The density and energy of these particles depend on the following:

— Altitude
— Latitude
— Solar activity
— Time

12.4.1.1 Cosmic radiation

Cosmic radiation consists mainly of protons (90%) and some alpha particles. The corresponding energies are in the gigaelectron-volt range, but the flux is low, on the order of 2.5 particles cm^{-2} s^{-1}.

12.4.1.2 Solar wind

Solar wind consists mainly of protons and electrons of lower energy. The mean density of protons during periods of low solar activity is on the order of 5 protons cm^{-3} escaping from the sun at velocities around 400 km s^{-1}. The mean flux corresponding to the level of the earth's orbit is 2×10^8 protons cm^{-2} with a mean energy of several kiloelectron-volts. According to solar activity, this flux can vary by a factor of 20. During periods of intense solar activity, solar eruptions occur more frequently and liberate proton fluxes with energies between several MeV and several hundreds of MeV. On rare occasions, with a periodicity on the order of several years, the proton energy can reach GeV. From $E = mc^2$, we have $1\,eV\,c^{-2} = (1.602176 \times 10^{-19}\,C)\,1\,V/(3 \times 10^8\,m\,s^{-1})^2 = 1.78 \times 10^{-36}$ kg.

12.4.1.3 The Van Allen belts

As solar wind particles are charged, they interact with the terrestrial magnetic field and are trapped in the so-called Van Allen belts. Figure 12.4 illustrates the trapped particle flux profiles as a function of altitude and energy.

For *electrons*, there are inner and outer belts. The boundary between them is 2–3 earth radii.

The high-energy *protons* of the Van Allen belts are contained within a distance equal to 4 earth radii with a maximum concentration around 1.5 and 2 earth radii (see Figure 12.4).

The geostationary satellite orbit (at 6.6 times the terrestrial radius) is within the electron outer belt and outside the belt of trapped protons. Geostationary satellites are thus mainly affected by electrons of the outer belt and high-energy protons generated by solar flares. The flux of these protons depends on the magnitude of solar activity, which determines the occurrence of ordinary and extraordinary solar flares. An extraordinary solar flare will almost certainly occur during the 10–15 years of the usual lifetime of satellites.

The various particle fluxes per cm^2 per year are given in Table 12.3 for the geostationary orbit. Figure 12.5 shows the cumulative dose for a geostationary satellite after 12 years in orbit as a function of the orbital location considering a Si detector at the centre of a spherical shaped aluminium shielding 10 mm thick. The dose is the amount of energy absorbed per unit of mass of the considered matter (100 rad = 1 Gray = 1 J kg^{-1}).

12.4.2 Effects on materials

When subjected to charged particles, metals and semiconductors undergo excitation of the electron levels of the atoms. Plastics are ionised, and insulating minerals undergo both effects.

Solar flares in particular affect the minority carriers in semiconductors, the optical transmission of glasses, and certain polymers. The active components of electronic circuits in the satellite equipment can be protected against these effects by appropriate shielding (Figure 12.6). Equipment cases that contain sensitive components are produced in cast aluminium with wall thicknesses on the order of a centimetre. The principal effects of high-energy

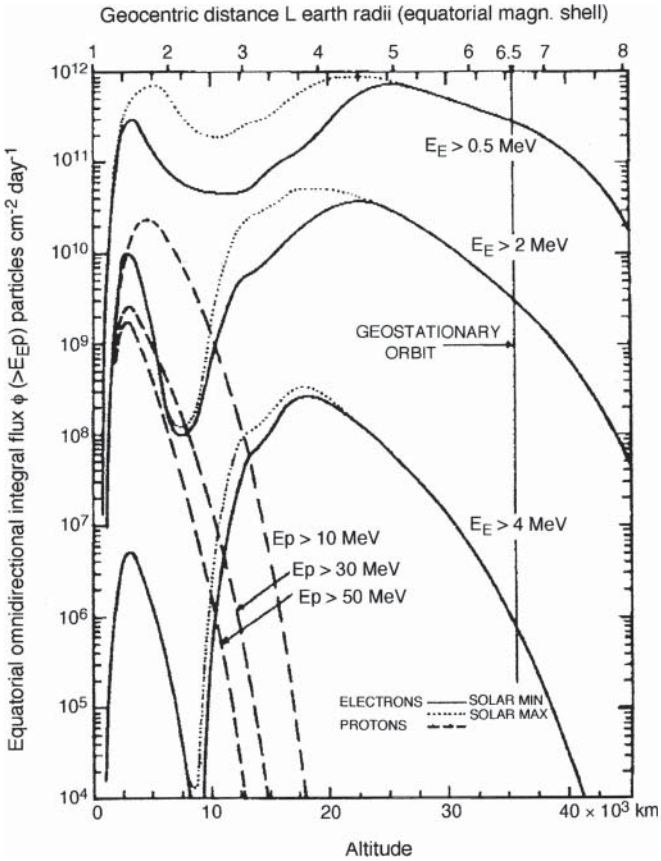


Figure 12.4 Earth's trapped electron and proton radial flux profiles. Source: reproduced from [CRA-94] with permission; © 1994 Elsevier.

Table 12.3 Total particle flux (number/cm^2 year) for the geostationary satellite orbit

Nature of particle	Low solar activity	Intense solar activity
Trapped electrons ($E > 0{:}5$ MeV)	$\sim 10^{14}$	$\sim 10^{14}$
Trapped protons	Negligible	Negligible
High-energy solar protons ($E > 40$ MeV)	$\sim 10^{7}$	$\sim 10^{10}$

particles appear as degradation of the performance of solar cells that are directly exposed to the flux and modification of the thermo-optical characteristics of surfaces that affect thermal control.

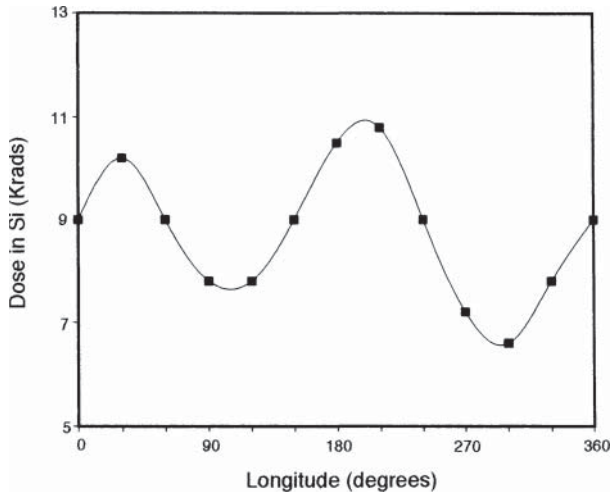


Figure 12.5 Cumulative dose after 12 years for a geostationary satellite as a function of longitude. Source: reproduced courtesy of Astrium.

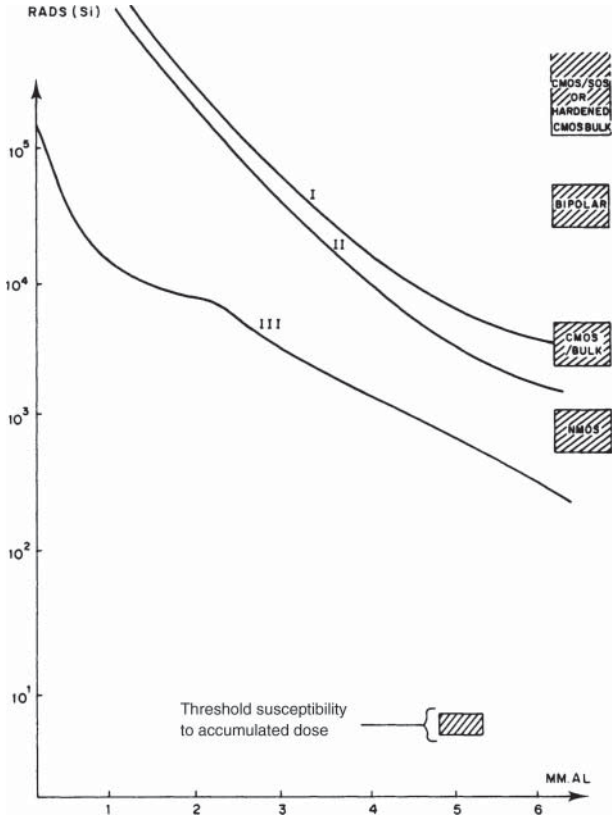


Figure 12.6 Radiation hardness of various integrated circuit technologies as a function of shield thickness. I Geostationary orbit during seven years; II geostationary orbit during three years; III low polar orbit during two years.

12.5 THE ENVIRONMENT DURING INSTALLATION

Installation – injection of the satellite into operational orbit – is preceded by two phases (see Chapter 11) during which the environment deviates somewhat from that described for the nominal orbit, particularly in the case of the orbit of geostationary satellites. These phases are as follows:

— The *launch phase* up to injection into the transfer orbit, with a duration of tens of minutes.
— The *transfer phase* during which the satellite describes elliptical orbits whose apogee is at the altitude of the final orbit, for example a 580 km × 35 786 km orbit for the case of launching a geostationary satellite by Ariane 5. This transfer phase lasts for several tens of hours.

12.5.1 The environment during launching

A fairing protects the satellite from aerodynamic heating as it passes through the dense layers of the atmosphere. Heating of the fairing has a negligible effect on the satellite.

The most important constraints are longitudinal and transverse accelerations and vibrations, and shocks communicated by the launcher during ignition of the motors and during propulsion phases. Acoustic noise under the fairing while passing through the atmosphere is also very high. The characteristics of these various excitations are given in the user manual for the launcher. An example is given in Figure 12.7 [AR5-16].

12.5.2 Environment in the transfer orbit

The attitude of satellite in the transfer orbit is usually spin-stabilised, and its configuration is different from its operational configuration since the apogee motor is full, the solar panels and antennas are folded, and so on.

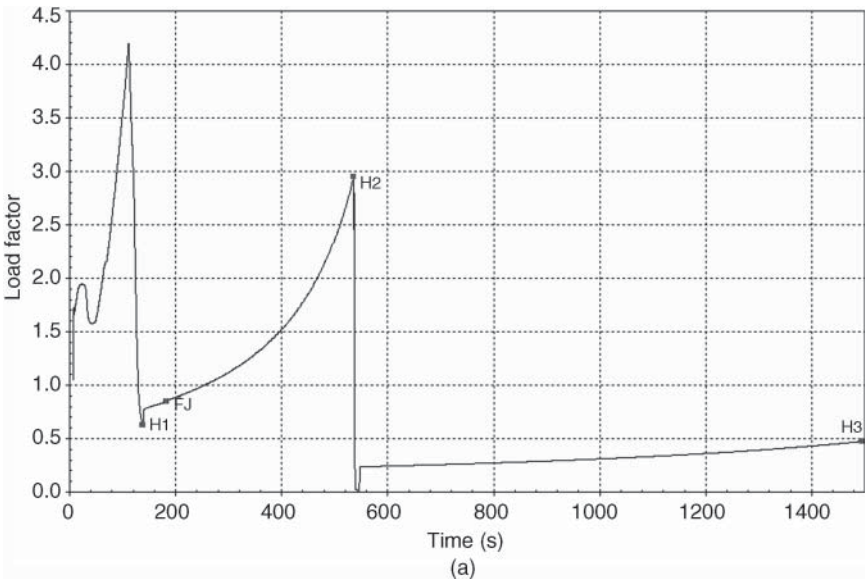


Figure 12.7 (a) Typical longitudinal static acceleration (Ariane 5). (b) Acoustic noise spectrum under the fairing. (OASPL means overall acoustic sound pressure level.)

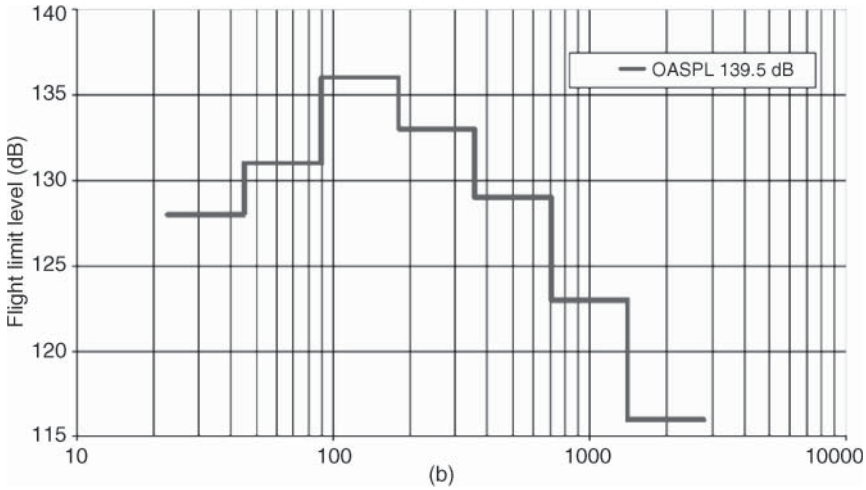


Figure 12.7 (*Continued*)

The environment and the effects discussed in the preceding sections are applicable, with the following two differences:

— Thermal effects that depend on the earth, its own radiation and albedo, with eclipses every 10 hours.
— At the perigee, atmospheric drag is not negligible, and the braking effect causes a reduction of altitude at the apogee.

REFERENCES

[AR5-16] Arianespace. (2016). Ariane 5 users' manual, issue 5, revision 2.
[CRA-94] Crabb, R. (1994). Solar cell radiation damage. *Radiation Physics and Chemistry* **43** (2): 93–103.

13 RELIABILITY AND AVAILABILITY OF SATELLITE COMMUNICATIONS SYSTEMS

The *reliability* of a system is defined by the probability of correct operation of the system during a given lifetime. The reliability of a complete satellite communications system depends on the reliability of its two principal constituents – the satellite and the ground stations.

The *availability* is the ratio of the actual period of correct operation of the system to the required period of correct operation. The availability of a complete satellite communications system depends not only on the reliability of the constituents of the system but also on the probability of successful launching, the replacement time, and the number of operational and backup satellites (in orbit and on the ground).

Availability of the ground stations depends not only on their reliability but also on their *maintainability*. For the satellite, availability depends only on reliability, since maintenance is not envisaged with current techniques.

13.1 INTRODUCTION TO RELIABILITY

13.1.1 Failure rate

For complex equipment such as that of a satellite, two types of breakdown can occur:

— Coincidental breakdowns
— Breakdowns resulting from usage (for example, wear of mechanical devices such as bearings and degradation of the cathodes of travelling wave tubes [TWTs]) and exhaustion of energy sources (such as the propellant required for station keeping and attitude control)

The instantaneous failure rate $\lambda(t)$ of a given piece of equipment is defined as the limit, as the time interval tends to zero, of the ratio of the number of pieces of equipment that fail in the time interval concerned to the number of pieces of equipment in a correct operating state at the start of the time interval (a large number of identical pieces of equipment are assumed to operate at the same time).

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

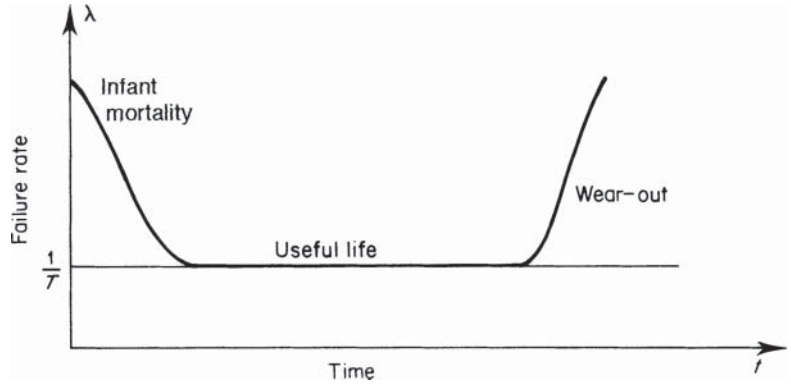


Figure 13.1 Failure rate versus time (bathtub curve).

The curve illustrating the variation of failure rate with time often has the form shown in Figure 13.1 (the *bathtub* curve), particularly for electronic equipment. Initially, the failure rate decreases rapidly with time. This is the period of early or infant failures. Subsequently, the failure rate is more or less constant. Finally, the failure rate increases rapidly with time; this is the wear-out period.

For space equipment, failures due to 'infant mortality' are eliminated before launching by means of special procedures (*burn-in*). Hence, during the period of useful life, most of the electronic and mechanical equipment has a constant failure rate λ. The instantaneous failure rate is thus often expressed in failures in time (FIT): the failure rate as the number of failures in 10^9 hours.

13.1.2 The probability of survival, or reliability

If a piece of equipment has a failure rate $\lambda(t)$, its probability of survival from time 0 to t, or reliability $R(t)$, is given by:

$$R(t) = \exp\left[-\int_0^t \lambda(u)\mathrm{d}u\right] \tag{13.1}$$

This expression is of a general form that is independent of the law of variation of failure rate $\lambda(t)$ with time.

If the failure rate λ is constant, the expression for the reliability reduces to:

$$R(t) = \mathrm{e}^{-\lambda t} \tag{13.2}$$

For a satellite, the designed maximum satellite lifetime U can be defined as the time interval at the end of which the service is no longer provided, usually due to exhaustion of the propellants. After time U, the probability of survival is zero. The curve in Figure 13.2 illustrates the variation of satellite reliability; the reliability is higher when λ is small.

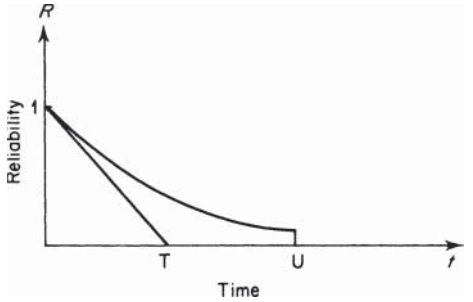


Figure 13.2 Reliability versus time.

13.1.3 Failure probability or unreliability

13.1.3.1 Unreliability F(t)

The unreliability or probability of having the system in a failed state at time t (failure has occurred between 0 and t) is the complement of the reliability:

$$R(t) + F(t) = 1 \tag{13.3}$$

13.1.3.2 Failure probability density f(t)

The failure probability density is the instantaneous probability of failure and is expressed as the derivative with respect to time of the unreliability:

$$f(t) = \mathrm{d}F(t)/\mathrm{d}t = -\mathrm{d}R(t)/\mathrm{d}t \tag{13.4}$$

The probability of failure occurring during a time interval t is thus:

$$F(t) = \int_0^t f(u)\mathrm{d}u \tag{13.5}$$

The failure rate $\lambda(t)$ is related to the failure probability density $f(t)$ by:

$$\lambda(t) = f(t)/R(t) \tag{13.6}$$

If the failure rate λ is constant, $f(t) = \lambda e^{-\lambda t}$.

For a satellite of maximum mission life U, the failure probability density as a function of time is given in Figure 13.3.

13.1.4 Mean time to failure (MTTF)

The mean time to failure (MTTF) is the mean time T of the occurrence of the first failure after entering service. It is obtained from the failure probability density using:

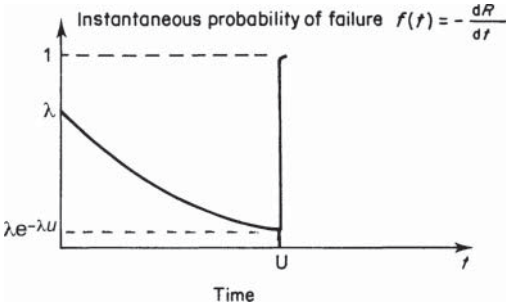


Figure 13.3 Instantaneous probability of failure versus time.

$$T = \int_0^\infty tf(t)\,\mathrm{d}t = \int_0^\infty R(t)\,\mathrm{d}t \tag{13.7}$$

If the failure rate λ is constant, $T = 1/\lambda$.

For equipment that is repaired after the occurrence of a failure, the mean time between failures (MTBF) is defined in a similar manner.

13.1.5 Mean satellite lifetime

In the case of a satellite of designed maximum lifetime U whose instantaneous probability of failure is given by Figure 13.3, the mean satellite lifetime τ can be considered as the sum of two integrals. The second is a delta function normalised so that the probability of failure in a period of time of infinite duration is equal to 1. The mean lifetime τ can be written:

$$\tau = \int_0^U t\lambda e^{-\lambda t}\mathrm{d}t + e^{-U/T}\int_U^\infty t\delta(t-U)\mathrm{d}t \tag{13.8}$$

Hence:

$$\tau = T(1 - e^{-U/T}) \tag{13.9}$$

The mean satellite lifetime t depends on the MTTF, $T = 1/\lambda$, defined for a constant failure rate λ. The ratio τ/T is the probability of failure during the maximum satellite lifetime U.

Table 13.1 gives the mean satellite lifetime t for an MTTF of 10 years as a function of the maximum satellite lifetime U.

Table 13.1 Mean satellite lifetime for MTTF of 10 yr

Max. satellite lifetime (yr)	Mean satellite lifetime (yr)
$U = T/3 = 3.3$	$\tau = 0.28\ T = 2.8$
$U = T/2 = 5$	$\tau = 0.39\ T = 3.9$
$U = T = 10$	$\tau = 0.63\ T = 6.3$
$U = 2\,T = 20$	$\tau = 0.86\ T = 8.6$
$U = 3\,T = 30$	$\tau = 0.95\ T = 9.5$

13.1.6 Reliability during the wear-out period

Components prone to wear-out, such as bearings, thrusters, and vacuum tube cathodes, have failures at end of life whose probability density can be modelled by a normal distribution (the failure rate is no longer constant, and failures are no longer accidental). The failure probability density is thus of the form:

$$f(t) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \tag{13.10}$$

where μ is the mean component lifetime and σ the standard deviation. The reliability becomes:

$$R(t) = 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \mathrm{d}t \tag{13.11}$$

A hybrid reliability can be defined as the product of the reliability considering only wear-out and the reliability, which characterises random failures. Equipment is generally designed in such a way that the lifetime determined by wear-out, μ, is long compared with the maximum satellite lifetime, U.

For components prone to wear-out, probability laws other than the normal distribution, such as the Weibull distribution, for example, are also used to model the occurrence of failures. For the Weibull distribution, the expressions for failure probability density $f(t)$ and reliability $R(t)$ are given by:

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t-\gamma}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{t-\gamma}{\alpha}\right)^{\beta}\right] \tag{13.12}$$

and

$$R(t) = \exp\left[-\left(\frac{t-\gamma}{\alpha}\right)^{\beta}\right] \tag{13.13}$$

where α, β, and γ are fitting parameters.

To model failure due to wear-out, the rate of which increases with time, the parameter β is greater than 1 ($\beta = 1$ corresponds to a constant failure rate and $\beta < 1$ corresponds to a decreasing failure rate that can model early failures).

13.2 SATELLITE SYSTEM AVAILABILITY

Availability A is defined as A = (required time – down time)/(required time), where required time is the period of time for which the system is required to operate and down time is the cumulative time the system is out of order within the required time.

To provide a given system availability A for a given required time L, it is necessary to determine the number of satellites to be launched during the required time L. The number of satellites to be launched will affect the cost of the service.

The required number of satellites n and the availability A of the system will be evaluated for two typical cases for which t_R is the time required to replace a satellite in orbit and p is the probability of a successful launch.

13.2.1 No backup satellite in orbit

13.2.1.1 Number of satellites required

As the mean lifetime of a satellite is τ, it will be necessary to put $S = L/\tau$ satellites into orbit on average during L years. As the probability of success of each launch is p, it will be necessary to attempt $n = \mathrm{S}/p$ launches, and the number of satellites n required is thus:

$$n = \frac{L}{pT[1 - \exp(-U/T)]} \tag{13.14}$$

13.2.1.2 System availability

If it is assumed that satellites close to their end of maximum lifetime U are replaced soon enough so that, even in the case of a launch failure, another launch can be attempted in time. The unavailability of the system at this time is small compared with the unavailability due to random failures.

During its maximum lifetime U, the probability that a satellite fails in a random manner is $P_\mathrm{a} = 1 - \mathrm{e}^{-U/T}$. In L years, there are S replacements to be performed, of which $P_\mathrm{a} \times S$ are for random failures. Each replacement requires a time t_r if it succeeds and, on average, a time t_r/p. The mean duration of unavailability during L years is $P_\mathrm{a} S t_\mathrm{r}/p = L t_\mathrm{r}/pT$. The mean unavailability (breakdown) rate is:

$$B = \frac{t_r}{pT} \tag{13.15}$$

and the availability $A = 1 - B$ of the system is thus:

$$A = 1 - \frac{t_r}{pT} \tag{13.16}$$

13.2.2 Backup satellite in orbit

By assuming, pessimistically but wisely, that a backup satellite has a failure rate λ and a maximum lifetime U equal to that of an active satellite, it is necessary to launch twice as many satellites during L years than in the previous case:

$$n = \frac{2L}{pT[1 - \exp(-U/T)]} \tag{13.17}$$

Taking account of the fact that t_r/T is small, the availability of the system becomes:

$$A = 1 - \frac{2t_r^2}{(pT)^2} \tag{13.18}$$

13.2.3 Conclusion

Table 13.2 provides three examples in which the time required for replacement t_r is 0.25 year and the probability p of a successful launch is 0.9. To obtain a high availability A, it can be seen that the MTTF of the satellite is significantly more important than its maximum lifetime U.

Without a backup satellite, the service is not provided, in the examples in the table, for a mean of $(1 - 0.972) \times 120 = 34$ or $(1 - 0.986) \times 120 = 1.7$ months in 10 years according to the predicted

Table 13.2 Examples of availability and number of satellites to be launched according to designed maximum lifetime and MTTF

Designed maximum lifetime (U)	5 yr	7 yr	10 yr
MTTF (T)	10 yr	20 yr	20 yr
Mean lifetime (τ)	3.9 yr	5.9 yr	7.9 yr
Probability or failure during life ($P_f = \tau/T$)	0.393	0.295	0.395
Time to replace (t_n)	0.25 yr	0.25 yr	0.25 yr
Probability of launch p success	0.9	0.9	0.9
No spare			
Annual launch rate: n/L	0.28	0.19	0.14
Availability (A)	0.972	0.986	0.986
One in-orbit spare			
Annual launch rate: n/L	0.56	0.38	0.28
Availability (A)	0.9985	0.9996	0.9996

lifetime. To limit the unavailability to one month implies an availability of at least 99.2%, and this requires the presence of a backup satellite in orbit, four to six replacement launches, and an MTTF of at least 10 years (10^5 hours). Satellites must thus be designed with a failure rate less than 10^{-5} per hour (10^4 Fit).

13.3 SUBSYSTEM RELIABILITY

Calculation of the reliability of a system is performed from the reliability of the system elements. As far as the satellite is concerned, except in the special case where elements in parallel can independently fulfil a particular mission, most subsystems are essentially in series from the point of view of reliability. This indicates that correct operation of each subsystem is indispensable for correct operation of the system.

13.3.1 Elements in series

13.3.1.1 Reliability

When elements are in series from the reliability point of view, the overall probability of correct operation is obtained by taking the product of the reliabilities of the elements. With n elements in series, the overall reliability R of the system can thus be written:

$$R = R_1 R_2 R_3 \ \ldots \ R_n \tag{13.19}$$

The reliability of a system containing four elements in series, where each has a reliability of 0.98, is thus $0:98^4 = 0:922$.

13.3.1.2 Failure rate

The overall failure rate λ of the system is obtained by adding the failure rates λ_i of each of the constituents if these are constants. The overall failure rate is thus constant and the MTTF is $1/\lambda$.

A communications satellite includes about 10 subsystems (see Chapters 8–10), and the mean failure rate per subsystem must be less than 10^{-7} hours (10^2 FIT). To obtain this reliability, some equipment must be provided with partial or total redundancy.

13.3.2 Elements in parallel (static redundancy)

13.3.2.1 Reliability if one element out of n *is sufficient*

For elements in parallel in the reliability sense, the failure probability of the ensemble is the product of the probability of failure of each of the elements:

$$F = F_1 F_2 F_3, \ \ldots \ , F_n \tag{13.20}$$

The reliability of the ensemble is given by: $R = 1 - F$. This relation is valid if correct operation of the ensemble is ensured with a single element out of the n.

13.3.2.2 Reliability if k *out of* n *elements are necessary*

If it is necessary to have k out of the n identical elements for correct operation, the various cases that correspond to correct operation must be analysed in order to evaluate the reliability. It can be shown that the probability p_k of having k elements out of the n in good order is given by the expansion of $(p+q)^n$ (the binomial rule), where p is the probability of correct operation of an element and q is that of not functioning ($p + q = 1$).

Example 13.1 A system consists of two elements of reliability R_i in parallel; correct operation is obtained if one of the two elements is in good order. The binomial rule gives $(R_i + F_i)^2 = R_i^2 + 2R_iF_i + F_i^2$.

Correct operation is obtained if both pieces of equipment are operating (reliability R_i^2) or if one has failed and the other is operational, or the inverse (reliability $2R_iF_i$). The reliability R of the system is thus:

$$R = R_i^2 + 2R_iF_i = R_i^2 + 2R_i(1 - R_i) = 2R_i - R_i^2$$

If the failure rates are constant and equal to λ, this becomes:

$$R = 2\mathrm{e}^{-\lambda t} - \mathrm{e}^{-2\lambda t}$$

13.3.2.3 Failure rate

The overall failure rate is obtained from the ratio $f(t)/R$, where the failure probability density is calculated from R using Eq. (13.4). Assuming the failure rates λ_i to be constant, it is found that the overall failure rate is a function of time and hence the overall failure rate is not constant.

Example 13.2 With two identical elements in parallel, it is found that:

$$\lambda = \frac{2\lambda_i(1 - \mathrm{e}^{-2\lambda_i t})}{2 - \mathrm{e}^{-2\lambda_i t}}$$

As time tends to infinity, λ tends to λ_i.

13.3.2.4 Mean time to failure (MTTF)

The mean time of occurrence of the first failure is calculated from the reliability using Eq. (13.7). If the failure rates λ_i of the n elements in parallel are constant and identical and if correct operation is ensured with a single element, the MTTF of the overall system can be put in the form:

$$\text{MTTF} = \text{MTTF}_i + \text{MTTF}_i/2 + \text{MTTF}_i/3 + \cdots + \text{MTTF}_i/n \tag{13.21}$$

where $\text{MTTF}_i = 1/\lambda_i$ is characteristic of one element.

Example 13.3 With two identical elements in parallel, it is found that:

$$\text{MTTF} = 1/\lambda_i + 1/2\lambda_i = 3/2\lambda_i = 1.5\,\text{MTTF}_i$$

The MTTF is increased by 50% by the parallel connection of the two pieces of equipment.

13.3.3 Dynamic redundancy (with switching)

13.3.3.1 The Poisson distribution

Consider a system constituted, in the reliability sense, of m normally active elements in parallel, where n elements can, in turn, be placed in parallel to replace a failed main element. The failure rate λ_i of each of the elements is constant and the same for each. The reliability R of the system for m elements in a correct operational state is given by the Poisson distribution:

$$R = \mathrm{e}^{-m\lambda_i t}[1 + m\lambda_i t + (m\lambda_i t)^2/2! + \cdots + (m\lambda_i t)^n/n!] \tag{13.22}$$

The MTTF is given by:

$$\text{MTTF} = [(n + 1/m)\text{MTTF}_i] \tag{13.23}$$

where $\text{MTTF}_i = 1/\lambda_i$ is characteristic of one element.

These general expressions assume failure rates that are constant and identical for the equipment while in service. They also assume that the backup equipment is in good operational order at the time when it replaces a failed piece of principal equipment. The reliability of the switching devices is also assumed to be equal to 1.

It is useful to be able to consider equipment failure rates that differ and may be modified for backup equipment that is in stand-by or operational mode. Since general expressions are either complex or impossible to establish, various special cases are presented next for particular examples.

13.3.3.2 Redundancy with different failure rates that depend on the operational state

The considered system consists of a principal element and a backup element that can replace it. The failure rate of the principal element is λ_p; the failure rate of the backup element is λ_r when the element is inactive and λ_s when it is operating. The reliability of the subsystem is evaluated by considering the various probabilities of failure and correct operation.

The probability of correct operation of the principal element between time 0 and t is $\mathrm{e}^{-\lambda_p t}$. The probability of failure of the principal element at time t_f (with $t_\text{f} < t$) is $\lambda_\text{p}\mathrm{e}^{-\lambda_\text{p} t_\text{f}}$. The probability of the backup element being in good order between 0 and t_f is $\mathrm{e}^{-\lambda_\text{r} t_\text{f}}$. The probability of correct operation of the backup element between time t_f and time t is $\mathrm{e}^{-\lambda_\text{s}(t-t_\text{f})}$.

Correct operation of the system at time t is thus ensured if the principal element is operating at time t (reliability R_p) or if, after failure of the principal element at time t_f, the backup element is operating at time t (reliability R_s). For the backup element to be in good order at time t, it must operate correctly between time 0 and time t_f and between t_f and time t.

The reliability R is thus equal to $R_p + R_s$ with:

$$R_p = e^{-\lambda_p t}$$

$$R_s = \int_0^t (\lambda_p e^{-\lambda_p t_f})(e^{-\lambda_r t_f})(e^{-\lambda_s (t-t_f)}) dt_f$$

$$= \lambda_p e^{-\lambda_s t} \int_0^t e^{-(\lambda_p + \lambda_r - \lambda_s) t_f} dt_f$$

$$= \frac{\lambda_p e^{-\lambda_s t}}{\lambda_p + \lambda_r - \lambda_s} \left[1 - e^{-(\lambda_p + \lambda_r - \lambda_s) t}\right]$$

The reliability of the system with redundancy is thus given by:

$$R = e^{-\lambda_P t} + \frac{\lambda_p}{\lambda_p + \lambda_r - \lambda_s} \left(e^{-\lambda_s t} - e^{-(\lambda_p + \lambda_r) t}\right) \tag{13.24}$$

Calculation of the MTTF T gives:

$$T = \text{MTTF} = (1/\lambda_p) + \lambda_p/\lambda_s(\lambda_p + \lambda_r) = T_p + [(T_s T_r)/(T_p T_r)] \tag{13.25}$$

where T_p; T_r; T_s are the MTTFs of the principal equipment, the backup equipment when inactive, and the backup equipment when operating, respectively.

A *particular case*. Consider a system having the same failure rate λ_p for the operational units and a zero failure rate for the inactive units ($\lambda_r = 0$). The expression for the reliability becomes:

$$R = e^{-\lambda_P t} + \lambda_p t\, e^{-\lambda_P t}$$

This expression can be obtained directly from the Poisson distribution. The mean time of occurrence of the first failure under these conditions is:

$$T = \text{MTTF} = 2/\lambda_p$$

The mean time of occurrence of the first failure is thus doubled by a redundancy of the 1/2 type (one active unit for two installed units).

13.3.3.3 Equipment redundancy taking account of the reliability of the switching element

The reliability will be evaluated for the example of a subsystem consisting of two units, of which one is the principal and the other the backup. The two cases to be considered are where the switching element is, and is not, necessary for operation of the principal element. The units have the same failure rate λ_p.

In the first case, the switch or switches are used to route the signals to the principal unit or to the backup unit (Figure 13.4a). These switches are thus, from the reliability point of view, in series with the duplicated equipment. The reliability of the ensemble is thus equal to the product

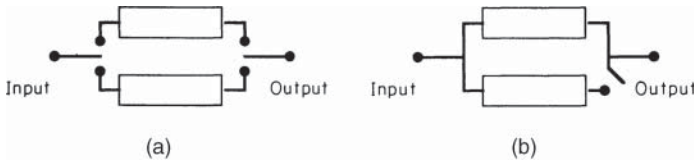


Figure 13.4 Equipment with 1/2 redundancy.

of the reliability of the switches (R_{sw} for a switch) and the reliability of the duplicated equipment that is obtained from the Poisson distribution:

$$R = R_{sw}^2[e^{-\lambda_p t}(1 + \lambda_p t)] \tag{13.26}$$

In the second case, the principal element is accessible without passing through the switch, which is used only to connect the backup unit in parallel with the principal element at time of failure (Figure 13.4b). Hence the reliability R_{sw} of the switch arises only in the backup branch. The reliability of the system is thus obtained by considering that correct operation is obtained at time t if the principal equipment is operational at this time, or if, after failure of the principal equipment at time t_f, the backup equipment operates correctly between t_f and t, on condition that the switch operates correctly. Calculation of the reliability gives:

$$R = e^{-\lambda_p t} + R_{sw}\lambda_p t e^{-\lambda_p t} \tag{13.27}$$

If the mean failure rate of the switch λ_{sw} is constant, $R_{sw} = e^{-\lambda_{sw} t}$ and that reliability becomes $e^{-\lambda_p t} + \lambda_p t e^{-(\lambda_p + \lambda_{sw})t}$. The mean time T of occurrence of the failure is given by:

$$T = \text{MTTF} = (1 + \lambda_p/\lambda_p + \lambda_{sw})^2 \tag{13.28}$$

If λ_{sw} is equal to 0, MTTF again has a value $2/\lambda_p$.

Example 13.4 Redundancy of TWTs in the Channelised Part of a Payload

Consider the channelised part of the communication payload of a satellite, where two channels share three TWTs and the associated preamplifiers (2/3 redundancy). Access to the amplifiers is by way of two switches, a switch with two inputs and three outputs ($S_{2/3}$) for the input and a switch with three inputs and two outputs ($S_{3/2}$) for the output (Figure 13.5a). In normal operation, two of the amplifiers are active and the third is in standby. The failure rate of an active equipment is λ_p. The failure rate of a standby equipment is λ_r (due to keeping it prewarmed, for example).

The probability of correct operation of both channels at time t is obtained by considering the equivalent block diagram from the reliability point of view; this is given in Figure 13.5b. The reliability of the system is equal to the product of the reliability of the switches R_{sw} and that of the set of amplifiers R_A.

The reliability R_A of the set of amplifiers is evaluated by considering the various probabilities of failure and correct operation. The probability of correct operation of both principal units between time 0 and t is $e^{-\lambda_p t}$. The probability of failure of a principal unit at time t_f (with $t_f < t$) is $\lambda_p e^{-\lambda_p t_f}$. The probability of correct operation of the other principal equipment between time 0 and t is $e^{-\lambda_p t}$. The probability of the backup element being in good order between 0 and t_f is $e^{-\lambda_r t_f}$. The probability of correct operation of the backup element between time t_f and t is $e^{-\lambda_p (t - t_f)}$.

Correct operation of the system at time t is thus ensured if:

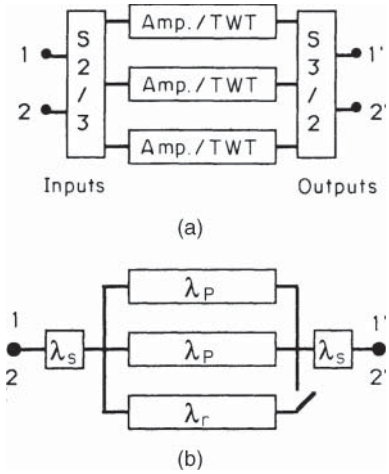


Figure 13.5 Equipment with 2/3 redundancy.

— Both principal units are in good order at time t
— Or if, after failure of a principal unit at time t_f, the backup element is in good order between 0 and t_f and operates correctly between t_f and t, knowing that the other principal unit continues to operate up to time t
— Or in the corresponding configuration to the previous one in case of failure of the other principal unit

The reliability R_A is thus given by:

$$R_A = e^{-2\lambda_P t} + 2e^{-\lambda_P t} \int_0^t \left[\left(\lambda_P e^{-\lambda_P t_f} \right) \left(e^{-\lambda_r t_f} \right) \left(e^{-\lambda_P (t - t_f)} \right) \right] dt_f$$

Hence:

$$R_A = e^{-2\lambda_P t} \left[1 + \left(2\lambda_P / \lambda_r \right) \left(1 - e^{-\lambda_r t} \right) \right] \tag{13.29}$$

Notice that if the failure rate of the standby equipment is zero ($\lambda_r = 0$), the reliability is obtained directly from the Poisson distribution:

$$R_A = e^{-2\lambda_P t}(1 + 2\lambda_P t)$$

The probability of correct operation of both channels at time t including the reliability of the switches is thus given by:

$$R = R_{S2/3} R_A R_{S3/2}$$

Numerical example:

Amplifier failure rate $\lambda_p = 2300$ Fit (2300×10^{-9} per h)
Switch failure rate $\lambda_{sw} = 50$ Fit

The expected lifetime is 10 years.
After 10 years:

The reliability R_A of the set of amplifiers is $R_A = 0.937\,65$.
The reliability R_{sw} of a switch is $R_{SW} = e^{-\lambda_{sw} t} = 0.995\,63$.
The reliability R of the ensemble is thus $R_A R_{SW}^2 = 0.929\,47$.

In comparison, the reliability of a non-duplicated amplifier is $e^{-\lambda_P t} = 0.817\,52$.

13.3.4 Equipment having several failure modes

Some equipment and elements have several modes of failure: for example, short circuit and open circuit for diodes, capacitors and so on. The consequences of a failure on the operation of the system concerned are not the same following a failure of one type or the other. The consequences also depend on the system architecture.

With a structure containing n elements in series, a failure of the open-circuit type, characterised by a probability of failure F_O for one element, involves failure of the ensemble. The probability of failure of the ensemble is thus $1 - (1 - F_O)^n$. On the other hand, with a failure of the short-circuit type, characterised by a probability of failure F_C, failure of the ensemble requires failure of all the elements. The corresponding probability of failure of the ensemble is thus $(F_C)^n$. The reliability R of the series structure is thus $R = (1 - F_O)^n - (F_C)^n$, and the series structure is robust with respect to failures of the short-circuit type.

When n elements in parallel are associated with failures of the open-circuit type, characterised by a probability of failure F_O for one element, failure of the ensemble requires failure of all the elements. The probability of failure of the ensemble is thus $(F_O)^n$. On the other hand, failure of the short-circuit type, characterised by a probability of failure F_C, involves that of the ensemble. The probability of failure of the ensemble is thus $1 - (1 - F_C)^n$. The reliability R of the parallel structure is thus $R = (1 - F_C)^n - (F_O)^n$, and the parallel structure is robust with respect to failures of the open-circuit type.

For each of these structures, there is an optimum number of elements that enables maximum reliability to be achieved. If protection against both types of failure is required simultaneously, more complex structures must be used, such as series–parallel or parallel–series types. These procedures are used for wiring the solar cells of the power generator, for example.

13.4 COMPONENT RELIABILITY

Certain subsystems, such as the payload, contain several hundreds of components. To obtain failure rates of one or two per 100 000 hours, each component must not exceed a failure rate on the order of 1 per 10 million hours.

During the design of the satellite, after the constraints have been analysed, provisional examination of the reliability enables redundancy arrangements to be defined together with the quality level of components and equipment.

13.4.1 Component reliability

Information on the failure rates of various types of component is available from the manufacturers, which have the results of component tests under particular environmental conditions. The data provided in documents published by various organisations, such as report MIL-HDBK-217C [DoD-91], can also be used. Table 13.3 gives the orders of magnitude of the failure rates of various components. The least reliable components are TWTs and components

Table 13.3 Typical failure rates of components for space applications, expressed in FIT (failure rate at 75% off-loading ratio – where applicable)

Resistors	
Solid carbon	5
Metallic film	5
Wirewound	10
Potentiometers	200
Capacitors	
Solid carbon	3
Polycarbonate	3
Mylar	5
Paper	20
Solid tantalum	20
Variable	20
High voltage	100
Silicon diodes	
Switching	4
Standard	10
Power	20
Zener	50
Detector/mixer	100
Filter sections	
Hybrid	25
Passband	10
Couplers	10
Circulators	10
Connectors	1
Transistors (planar, silicon)	
Standard	10
Switching	10
HF	20
Power	50
Integrated circuits	
Digital (bipolar)	10
Analogue	20
FET IC	
1–10 gates	100
11–50 gates	500
TWT	150
Transformers	200
Power	30
Signal	10
Inductors	
Power	20
Signal	10
Quartz crystals	80
Relays	400

with moving parts (such as rotating bearings, relays, and potentiometers). The most reliable components are passive ones such as resistors, capacitors, switching diodes, and connectors.

The failure rate of a component can be greatly reduced by an appropriate choice of loading ratio (derating). The mean power dissipated by the components is chosen to be a small percentage of the nominal power specified by the manufacturer. For example, a resistor capable of dissipating 1 W would be chosen for a resistance that must dissipate 300 mW; the loading ratio is thus 30%. In any case, operation should not be close to the specified parameter limits. The junction temperature of transistors must not exceed a specified value (typically 105 °C).

The same principle is applied to the maximum values of voltages, currents, etc. that components and equipment must be capable of withstanding. Wear-out of elements is thus reduced according to a power law as a function of reduced loading. Applicable loading ratios appear in the preferential lists of components to be used in priority according to the order of the components appearing in the list.

13.4.2 Component selection

Components are chosen after a functional examination of the equipment on preferential lists established by space agencies such as the European Space Agency (ESA/SCCG Space Component Coordination Group), Centre National d'Etudes Spatiales (CNES) (CNES/QFT/IN-0500), NASA, etc. Special procedures are followed to ensure the manufacturing quality of the component chosen, and the constancy of its properties from one sample to another and with time (these include purchase and acceptances specifications, qualifications of batches and component acceptance) [ESA-18a, ESA-18b, ESA-18c].

When a component that does not appear in the preferential lists is necessary, qualification of the component is performed using the same specifications as those which are supplied. It includes two main phases:

— The evaluation phase
— The qualification phase

13.4.2.1 Evaluation

This phase includes the following:

— Inspection of the manufacturing facilities
— Detailed examination of the production line of the component concerned
— Evaluation tests to the limits of the components
— Examination of the manufacturing and monitoring documentation (the process identification document [PID])

The evaluation phase concludes with an evaluation report and a final review of the documentation. When this phase is completed in a satisfactory manner, the qualification phase is entered.

13.4.2.2 Qualification

This phase includes the following activities:

— Manufacture of the components that constitute the qualified batch
— 100% testing at the end of production and selection
— Qualification testing of a sample of the batch

If the results are satisfactory, qualification is declared, and a qualification certificate is delivered to the manufacturer. The qualified product is then entered in the preferential list, such as the Qualified Products List (QPL) of the ESA.

Qualification is valid for a fixed period. After this period, the validity can be extended on condition that batch tests are, or have been, performed and the PIDs have not been changed.

13.4.3 Manufacture

Having selected the components, the equipment manufacturing specifications must be defined. Technical design takes into account the constraints of performance, weight, volume, etc. and constraints specific to the space environment. The manufacturing specifications include the choice of wiring process (printed circuit or otherwise), the type of solder, the form of enclosure or protective cladding, etc.

Manufacturing quality control aims at verifying that manufacturing specifications are actually observed during the various stages and the components used are actually those that have been specified.

13.4.4 Quality assurance

Quality assurance is indispensable and complementary to security and reliability. More precisely, quality assurance ensures that a number of objectives and tasks relating to the requirements of the space project become facts.

The main elements of a quality assurance programme are as follows.

13.4.4.1 Quality of the pre-project studies and definition

During the analysis phases, quality assurance consists of:

— Verifying the conformity of the design documentation (plans and specifications) with the requirements of the programme
— Verifying the conformity of the definition with general quality rules and the particular requirements of the project
— Ensuring that the requirements of reliability, security, and quality are taken into account

13.4.4.2 Design quality

At the design level, conformity of all designs, tests, and detailed specifications to the quality rules and requirements of the projects must be verified. At the model testing level, conformity of the models to the design, conformity of the test conditions to the project requirements, the quality of the results and the standards must be verified.

13.4.4.3 Supply quality

The quality of supplies relies on:

— Definition of a supply specification that conforms to the reliability requirements and technical performance
— Definition of the qualification and acceptance conditions

— The choice of components and materials
— Appraisal of defective components and materials
— Definition of acceptance procedures for batches of components

13.4.4.4 Quality of manufacture

Manufacturing quality depends on:

— Definition of an industrial document that conforms to the quality rules and the requirements of the project
— Monitoring of the manufacturing, assembly, commissioning, and repair procedures
— Execution of the quality control plan during manufacture

13.4.4.5 Quality of testing

As a general rule, the quality of testing is based on:

— Optimum definition of the test programme
— A test procedure that conforms to the objective (qualification, acceptance, or development testing) and is compatible with the requirements of the project (such as constraints encountered in the course of the mission and the duration of the mission)
— The quality, reliability, and security of the test methods
— The quality of the measuring equipment (ensured by periodic checking and suitable conditions of use)
— The quality of performance of the test
— Utilisation of the results

13.4.4.6 Control of the configuration

The quality of the whole project depends on thorough knowledge of the system at a given time and hence on subsequent control of the configuration. The organisation of the system; partitioning into assemblies, subassemblies, units, components, and so on; definition of basic documents; nomenclature; continuous updating; and availability and dissemination of documentation and information are the important factors for the quality of the configuration and its control.

13.4.4.7 Nonconformity, failures, and exemptions

All recorded nonconformities and failures must be dealt with through a procedure that includes analyses, expert appraisal, statistical evaluation, repairs, exemptions, and modifications. This programme is particularly intended to identify the origin of the difficulty, the responsibility, and the solution to be used to obtain conformity of the failed element to the reference models (this relates to the specification of the identification model, acceptance and qualification) and to avoid the occurrence of a further deviation in the subsequent part of the project.

13.4.4.8 Development programme of models and mock-ups

The quality of a development programme lies in the realisation of mock-ups and models (development mock-ups, mechanical models, thermal models, identifications models, qualifications

models, and flying models). Tests demonstrate feasibility and provide inputs for optimisation of structures and mass breakdown, mechanical behaviour, thermal behaviour, and adequacy of the hardware for the constraints that will be encountered.

13.4.4.9 Storage, packaging, transport, and handling

Storage, packaging, handling, and transport conditions are the subject of a set of rules specified with the intention of maintaining the quality of the hardware regardless of the level of integration. These rules contain a number of precautions that are taken so that the hardware is not weakened by constraints for which it is not designed, and these must be considered in connection with the equipment used for packaging, handling, and transport.

Application of these principles determines the validity of operations associated with the reliability and security of space programmes.

REFERENCES

[DoD-91] Department of Defense. (1991). Reliability prediction of electronic equipment. MIL-HDBK-217F.

[ESA-18a] European Space Agency. (2018). Charter of the European Space Components Coordination. ESCC00000 issue 2.

[ESA-18b] European Space Agency. (2018). European preferred parts list. ESCC/RP/EPPL007-37.

[ESA-18c] European Space Agency. (2018). ESCC qualified parts list (QPL). ESCC/RP/QPL005-192 (REP005).

INDEX

Satellite Communications Systems: Systems, Techniques and Technology, Sixth Edition.
Gérard Maral, Michel Bousquet, and Zhili Sun.

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.