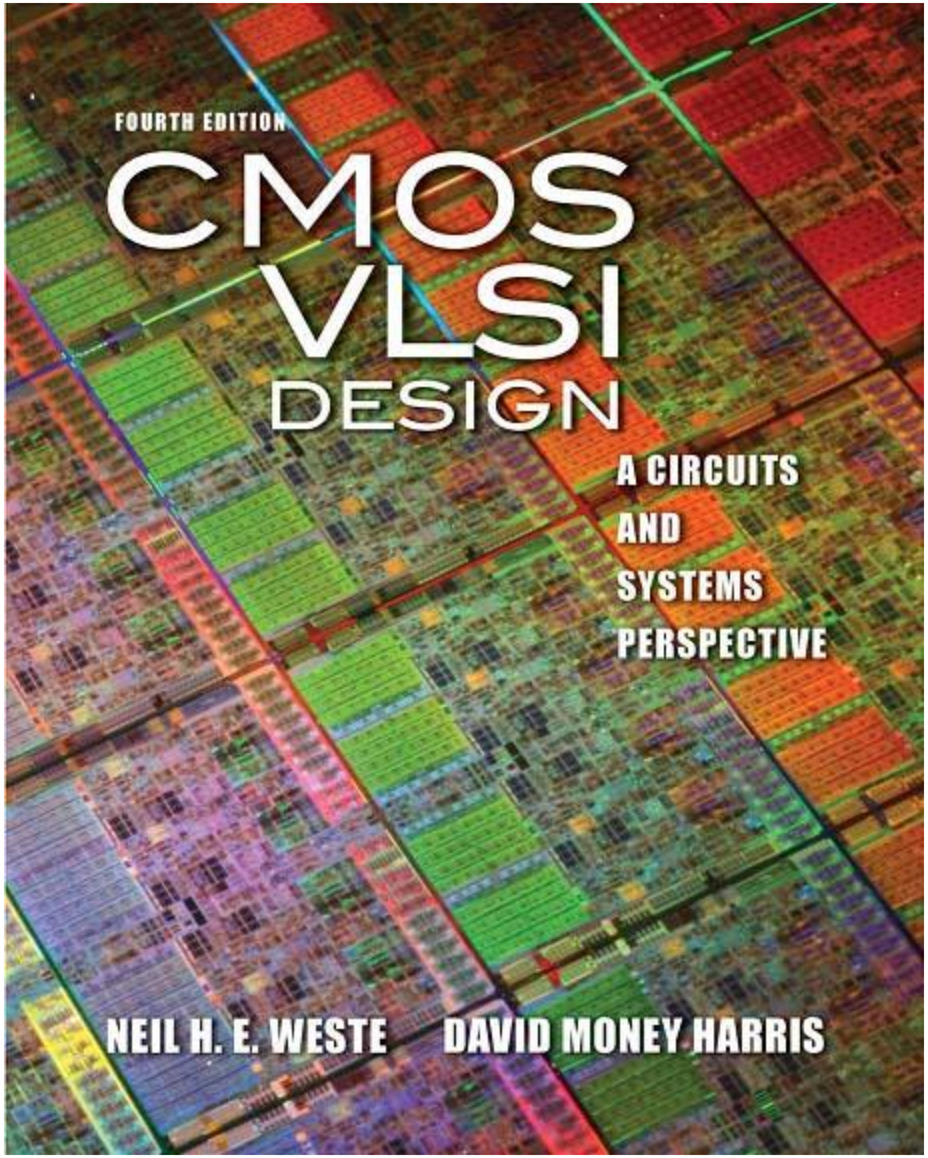


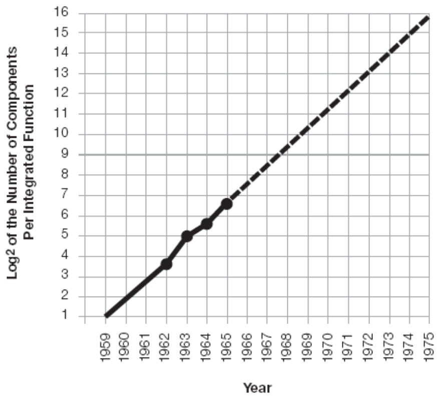
Lecture 15: Scaling & Economics

Outline

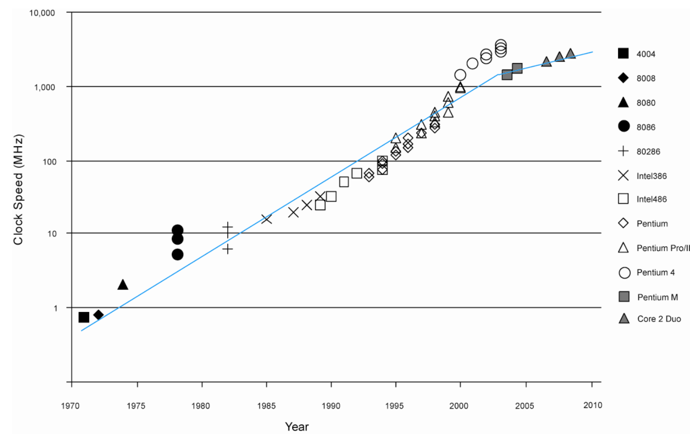
- Scaling
  - Transistors
  - Interconnect
  - Future Challenges
- Economics
- This material is from
  - Out textbook: section 7.4

Moore's Law

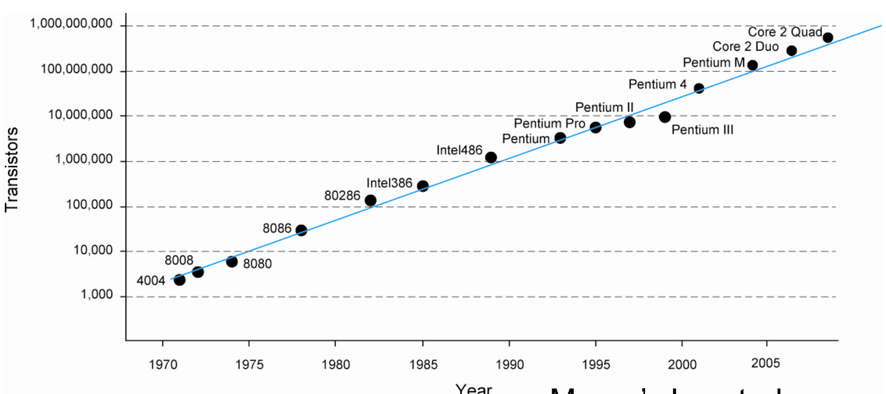
- Recall that Moore's Law has been driving CMOS



[Moore65]



Corollary: clock speeds have improved



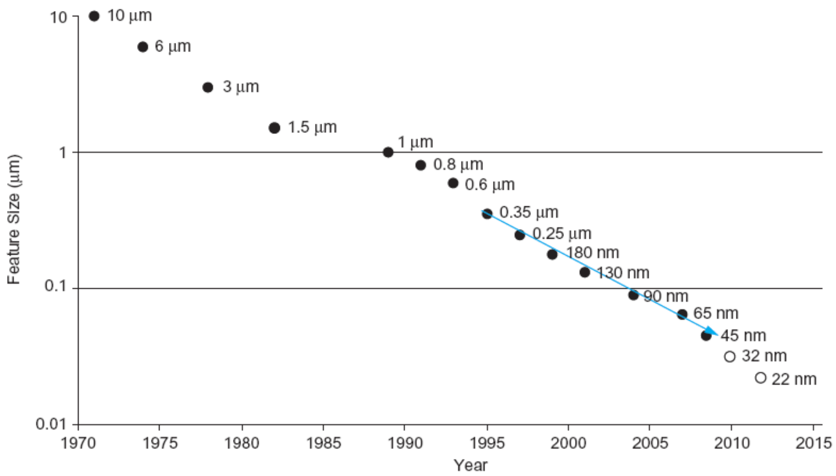
Moore's Law today

Why?

- Why more transistors per IC?
  - Smaller transistors
  - Larger dice
- Why faster computers?
  - Smaller, faster transistors
  - Better microarchitecture (more IPC)
  - Fewer gate delays per cycle

Scaling

- The only constant in VLSI is constant change
- **Feature size shrinks by 30% every 2-3 years**
  - Transistors become cheaper
  - Transistors become faster and lower power
  - Wires do not improve (and may get worse)
- **Scale factor S**
  - Typically $S = \sqrt{2}$
  - Technology nodes



Technology Scaling Methods

- Full scaling (constant-filed scaling or Dennard's scaling):
  - Scales dimensions and voltages, doping densities
  - (+) constant electrical field
  - (+) Great reduction in delay, area and power
  - (-) Changing voltages is not desirable from standard point of view
- Constant (fixed) voltage scaling:
  - scale dimensions, but not voltages
  - (+) Allows Vdd to be compatible for several process generations
  - (-) Suffers from power issues (e.g. high power density)
- Lateral scaling (gate shrink): scales only L

Device Scaling

Parameter	Sensitivity	Dennard Scaling	Constant Voltage	Lateral Scaling
Scaling Parameters				
Length: L		$1/S$	$1/S$	$1/S$
Width: W		$1/S$	$1/S$	1
Gate oxide thickness: t_{ox}		$1/S$	$1/S$	1
Supply voltage: V_{DD}		$1/S$	1	1
Threshold voltage: V_{tn}, V_{tp}		$1/S$	1	1
Substrate doping: N_A		S	S	1
Device Characteristics				
β	$\frac{W}{L}\frac{1}{t_{ox}}$	S	S	S
Current: I_{ds}	$\beta(V_{DD}-V_t)^2$	$1/S$	S	S
Resistance: R	$\frac{V_{DD}}{I_{ds}}$	1	$1/S$	$1/S$
Gate capacitance: C	$\frac{WL}{t_{ox}}$	$1/S$	$1/S$	$1/S$
Gate delay: τ	RC	$1/S$	$1/S^2$	$1/S^2$
Clock frequency: f	$1/\tau$	S	S^2	S^2
Switching energy (per gate): E	CV_{DD}^2	$1/S^3$	$1/S$	$1/S$
Switching power dissipation (per gate): P	Ef	$1/S^2$	S	S
Area (per gate): A		$1/S^2$	$1/S^2$	1
Switching power density	P/A	1	S^3	S
Switching current density	I_{ds}/A	S	S^3	S

What you should take from this table: τ,f,p, I, densities (I,P)

Gates get faster with scaling (good)

Dynamic power goes down with scaling (good)

Current density goes up with scaling (bad)

Example

A micro controller chip manufactured using 65-nm technology. The power supply for the chip is 1.25V. The chip runs at 1GHz and consumes 1W.
What is the expected speed and power if the chip is manufactured using 45-nm with constant voltage scaling.

$S = 65/45 = 1.4 = 2^{1/2}$

$Speed_{45} = S^2 * Speed_{65} = 2$ GHz

$Power_{45} = S * Power_{65} = 1.4$ W

Real Scaling (read)

- t_{ox} scaling has slowed since 65 nm
  - Limited by gate tunneling current
  - Gates are only about 4 atomic layers thick!
  - High-k dielectrics have helped continued scaling of effective oxide thickness
- V_{DD} scaling has slowed since 65 nm
  - SRAM cell stability at low voltage is challenging
- Dennard scaling predicts cost, speed, power all improve
  - Below 65 nm, some designers find they must choose just two of the three

Wire Scaling

- Wire cross-section
  - w, s, t all scale
- Wire length
  - Local / scaled interconnect
  - Global interconnect
    - Die size scaled by $D_c \approx 1.1$

Interconnect Scaling

Parameter	Sensitivity	Scale Factor
Scaling Parameters		
Width: w		$1/S$
Spacing: s		$1/S$
Thickness: t		$1/S$
Interlayer oxide height: h		$1/S$
Die size		D_c
Characteristics per Unit Length		
Wire resistance per unit length: R_w	$\frac{1}{wt}$	S^2
Fringing capacitance per unit length: C_{wf}	$\frac{t}{s}$	1
Parallel plate capacitance per unit length: C_{wp}	$\frac{w}{h}$	1
Total wire capacitance per unit length: C_w	$C_{wf} + C_{wp}$	1
Unrepeated RC constant per unit length: t_{wu}	$R_w C_w$	S^2
Repeated wire RC delay per unit length: t_{wr} (assuming constant field scaling of gates)	$\sqrt{RCR_w C_w}$	$\sqrt{S}$
Crosstalk noise	$\frac{w}{h}$	1
Energy per bit per unit length: E_w	$C_w V_{DD}^2$	$1/S^2$
Local/Semiglobal Interconnect Characteristics		
Length: l		$1/S$
Unrepeated wire RC delay	$l^2 t_{wu}$	1
Repeated wire delay	$l t_{wr}$	$\sqrt{1/S}$
Energy per bit	$l E_w$	$1/S^3$
Global Interconnect Characteristics		
Length: l		D_c
Unrepeated wire RC delay	$l^2 t_{wu}$	$S^2 D_c^2$
Repeated wire delay	$l t_{wr}$	$D_c \sqrt{S}$
Energy per bit	$l E_w$	D_c/S^2

ITRS (read)

- Semiconductor Industry Association forecast
  - Intl. Technology Roadmap for Semiconductors

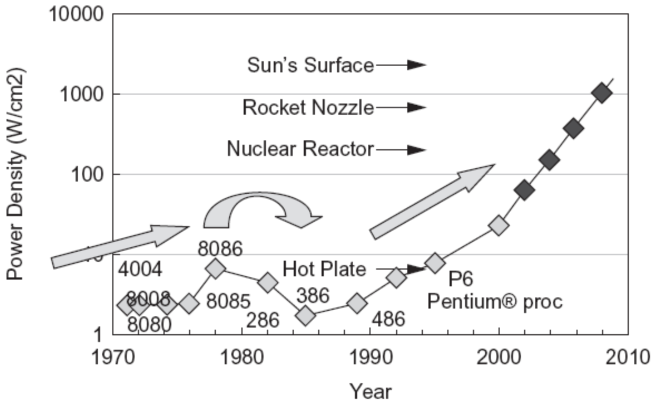
Year	2009	2012	2015	2018	2021
Feature size (nm)	34	24	17	12	8.4
L_{gate} (nm)	20	14	10	7	5
V_{DD} (V)	1.0	0.9	0.8	0.7	0.65
Billions of transistors/die	1.5	3.1	6.2	12.4	24.7
Wiring levels	12	12	13	14	15
Maximum power (W)	198	198	198	198	198
DRAM capacity (Gb)	2	4	8	16	32
Flash capacity (Gb)	16	32	64	128	256

Scaling Implications

- Improved Performance
- Improved Cost
- Interconnect Woes
- Power Woes
- Productivity Challenges
- Physical Limits

Dynamic Power (read)

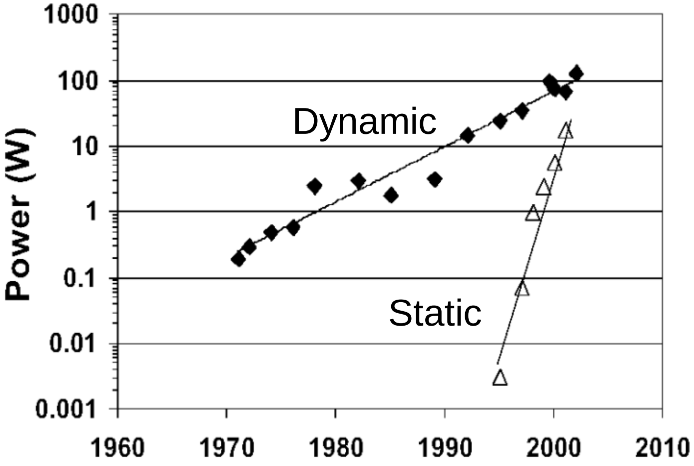
- Intel VP Patrick Gelsinger (ISSCC 2001)
  - If scaling continues at present pace, by 2005, high speed processors would have power density of nuclear reactor, by 2010, a rocket nozzle, and by 2015, surface of sun.
  - "Business as usual will not work in the future."
- Attention to power is increasing



[Intel]

Static Power (read)

- V_{DD} decreases
  - Save dynamic power
  - Protect thin gate oxides and short channels
  - No point in high value because of velocity sat.
- V_t must decrease to maintain device performance
- But this causes exponential increase in OFF leakage
- Major future challenge



[Moore03]

Physical Limits

- Will Moore's Law run out of steam?
  - Can't build transistors smaller than an atom…
- Many reasons have been predicted for end of scaling
  - Dynamic power
  - Subthreshold leakage, tunneling
  - Short channel effects
  - Fabrication costs
  - Electromigration
  - Interconnect delay
- Rumors of demise have been exaggerated

VLSI Economics (Read the rest)

- Selling price S_{total}
  - $S_{total} = C_{total} / (1-m)$
- m = profit margin
- C_{total} = total cost
  - Nonrecurring engineering cost (NRE)
  - Recurring cost
  - Fixed cost

NRE

- Engineering cost
  - Depends on size of design team
  - Include benefits, training, computers
  - CAD tools:
    - Digital front end: $10K
    - Analog front end: $100K
    - Digital back end: $1M
- Prototype manufacturing
  - Mask costs: $5M in 45 nm process
  - Test fixture and package tooling

Recurring Costs

- Fabrication
  - Wafer cost / (Dice per wafer * Yield)
  - Wafer cost: \$500 - \$3000
  - Dice per wafer: $N = \pi\left[\frac{r^2}{A} - \frac{2r}{\sqrt{2A}}\right]$
  - Yield: $Y = e^{-AD}$
    - For small A, Y $\approx$ 1, cost proportional to area
    - For large A, Y $\rightarrow$ 0, cost increases exponentially
- Packaging
- Test

Fixed Costs

- Data sheets and application notes
- Marketing and advertising
- Yield analysis

Example

- You want to start a company to build a wireless communications chip. How much venture capital must you raise?

- Because you are smarter than everyone else, you can get away with a small team in just two years:
  - Seven digital designers
  - Three analog designers
  - Five support personnel

Solution

- Digital designers:
  - $70k salary
  - $30k overhead
  - $10k computer
  - $10k CAD tools
  - Total: $120k * 7 = $840k
- Analog designers
  - $100k salary
  - $30k overhead
  - $10k computer
  - $100k CAD tools
  - Total: $240k * 3 = $720k
- Support staff
  - $45k salary
  - $20k overhead
  - $5k computer
  - Total: $70k * 5 = $350k
- Fabrication
  - Back-end tools: $1M
  - Masks: $5M
  - Total: $6M / year
- Summary
  - 2 years @ $7.91M / year
  - $16M design & prototype