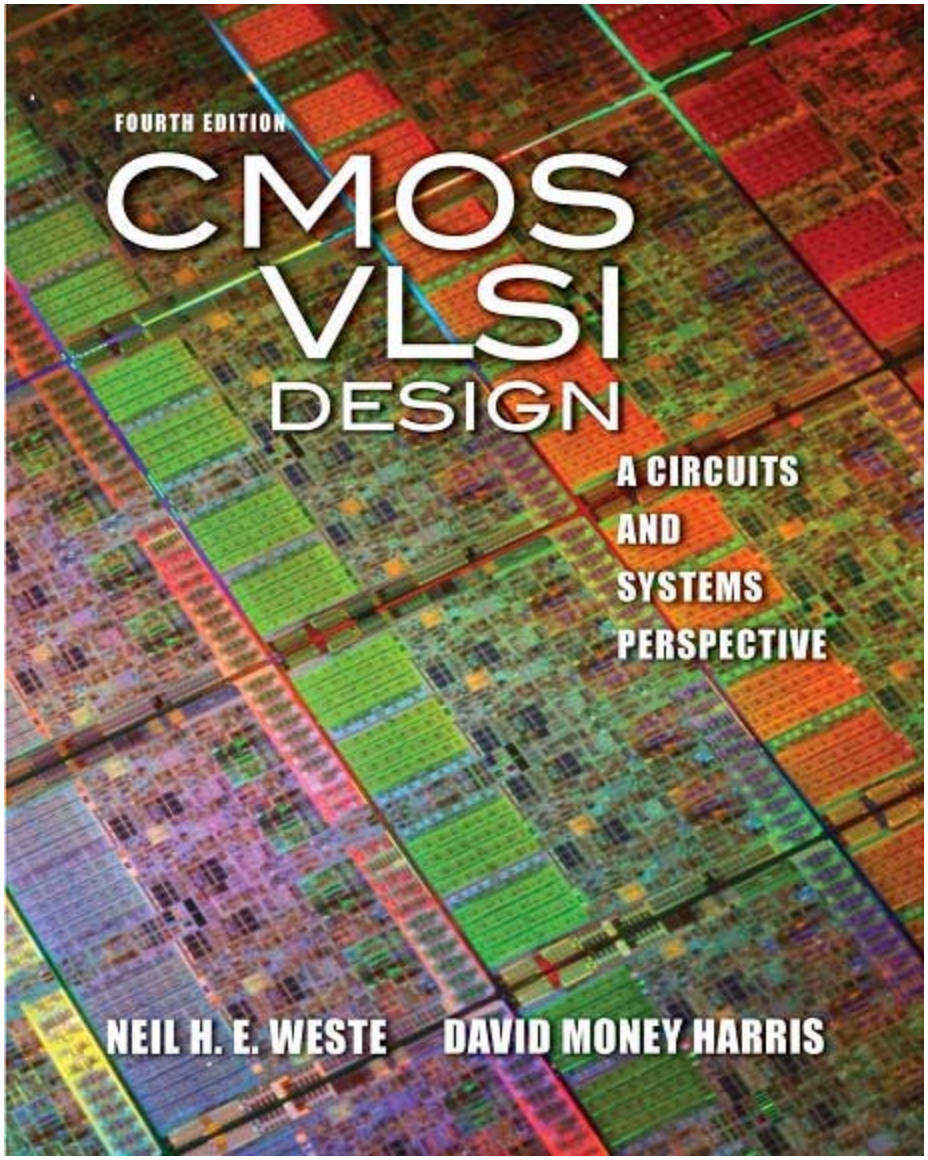


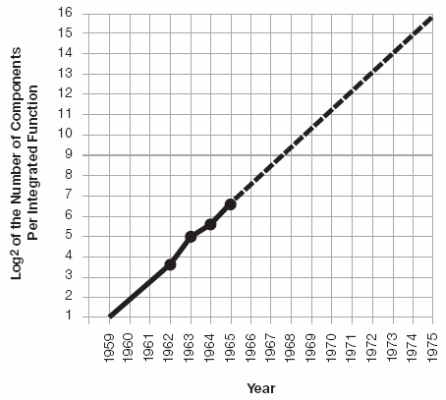
Lecture 15: Scaling & Economics

Outline

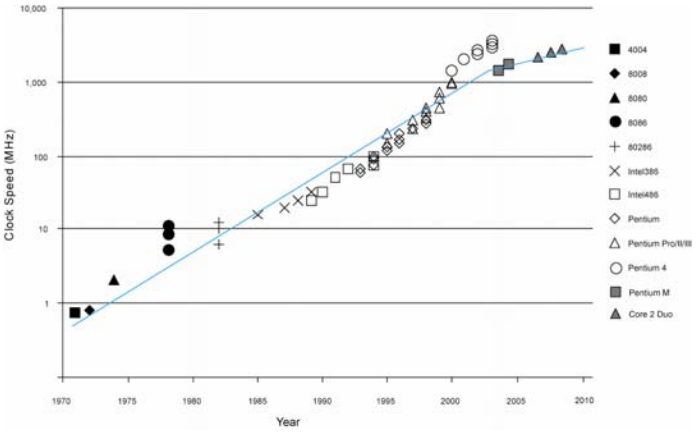
- Scaling
  - Transistors
  - Interconnect
  - Future Challenges
- Economics

Moore's Law

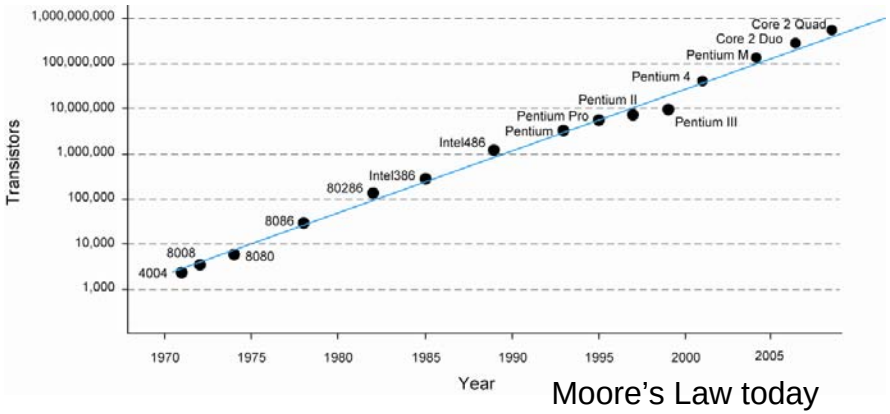
- Recall that Moore's Law has been driving CMOS



[Moore65]



Corollary: clock speeds have improved



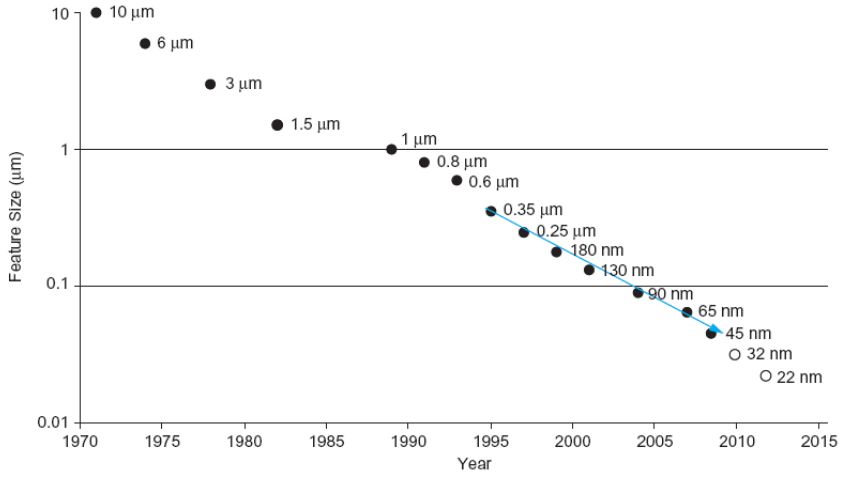
Moore's Law today

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



Dennard Scaling

- Proposed by Dennard in 1974
- Also known as *constant field* scaling
  - Electric fields remain the same as features scale
- Scaling assumptions
  - All dimensions (x, y, z => W, L, t_{ox})
  - Voltage (V_{DD})
  - Doping levels

Device Scaling

Parameter	Sensitivity	Dennard Scaling
L: Length		1/S
W: Width		1/S
t_{ox}: gate oxide thickness		1/S
V_{DD}: supply voltage		1/S
V_t: threshold voltage		1/S
NA: substrate doping		S
β	$W/(Lt_{ox})$	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S
R: effective resistance	V_{DD}/I_{on}	1
C: gate capacitance	WL/t_{ox}	1/S
τ: gate delay	RC	1/S
f: clock frequency	$1/\tau$	S
E: switching energy / gate	CV_{DD}^2	$1/S^3$
P: switching power / gate	Ef	$1/S^2$
A: area per gate	WL	$1/S^2$
Switching power density	P/A	1
Switching current density	I_{on}/A	S

Observations

- Gate capacitance per micron is nearly independent of process
- But ON resistance * micron improves with process

- Gates get faster with scaling (good)
- Dynamic power goes down with scaling (good)
- Current density goes up with scaling (bad)

Example

- Gate capacitance is typically about 1 fF/μm
- The typical FO4 inverter delay for a process of feature size *f* (in nm) is about 0.5*f* ps
- Estimate the ON resistance of a unit (4/2 λ) transistor.

Real Scaling

- 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
w: width		1/S
s: spacing		1/S
t: thickness		1/S
h: height		1/S
D_c: die size		D_c
R_w: wire resistance/unit length	1/wt	S^2
C_{wf}: fringing capacitance / unit length	t/s	1
C_{wp}: parallel plate capacitance / unit length	w/h	1
C_w: total wire capacitance / unit length	$C_{wf} + C_{wp}$	1
t_{wu}: unrepeated RC delay / unit length	R_wC_w	S^2
t_{wr}: repeated RC delay / unit length	sqrt(RCR_wC_w)	sqrt(S)
Crosstalk noise	w/h	1
E_w: energy per bit / unit length	$C_wV_{DD}^2$	$1/S^2$

Interconnect Delay

Parameter	Sensitivity	Local / Semiglobal	Global
l: length		1/S	D_c
Unrepeated wire RC delay	$l^2 t_{wu}$	1	$S^2 D_c^2$
Repeated wire delay	$l t_{wr}$	sqrt(1/S)	$D_c$sqrt(S)
Energy per bit	$l E_w$	$1/S^3$	D^c/S^2

Observations

- Capacitance per micron is remaining constant
  - About 0.2 fF/μm
  - Roughly 1/5 of gate capacitance
- Local wires are getting faster
  - Not quite tracking transistor improvement
  - But not a major problem
- Global wires are getting slower
  - No longer possible to cross chip in one cycle

ITRS

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

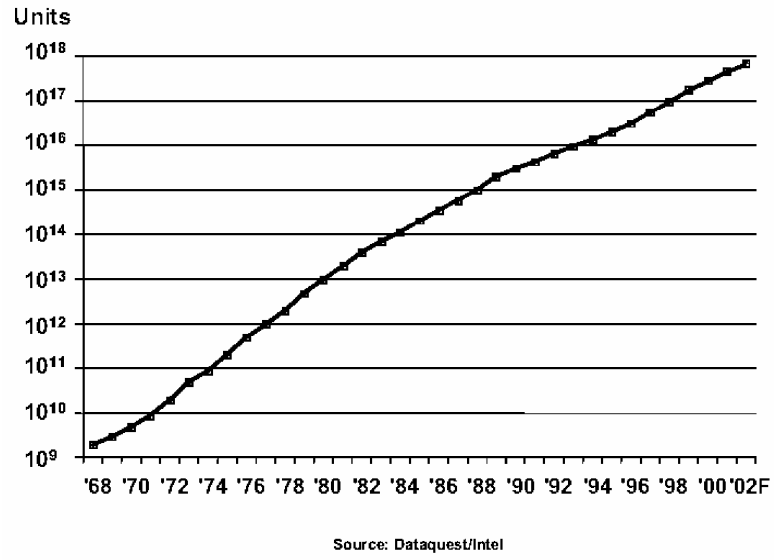
Year	2009	2012	2015	2018	2021
Feature size (nm)	34	24	17	12	8.4
L_{gate} (nm)	20	14	10	7	5
V_{DD} (V)	1.0	0.9	0.8	0.7	0.65
Billions of transistors/die	1.5	3.1	6.2	12.4	24.7
Wiring levels	12	12	13	14	15
Maximum power (W)	198	198	198	198	198
DRAM capacity (Gb)	2	4	8	16	32
Flash capacity (Gb)	16	32	64	128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

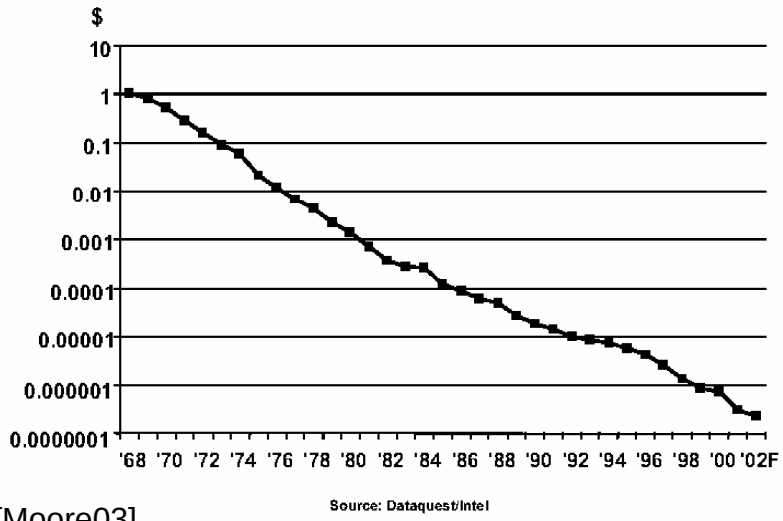
Scaling Implications

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

Cost Improvement

- In 2003, $0.01 bought you 100,000 transistors
  - Moore's Law is still going strong

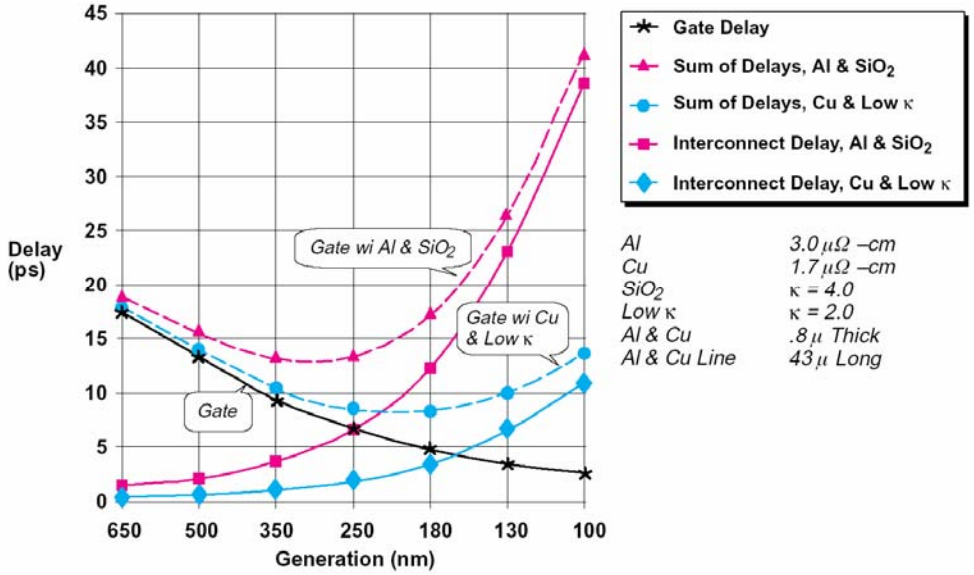




[Moore03]

Interconnect Woes

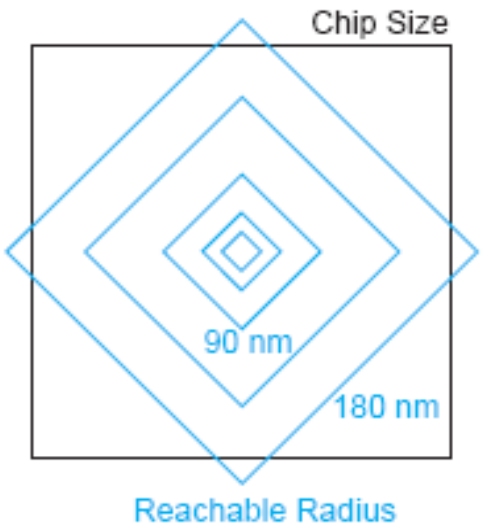
- SIA made a gloomy forecast in 1997
  - Delay would reach minimum at 250 – 180 nm, then get worse because of wires
- But…
  - Misleading scale
  - Global wires
- 100 kgate blocks ok



[SIA97]

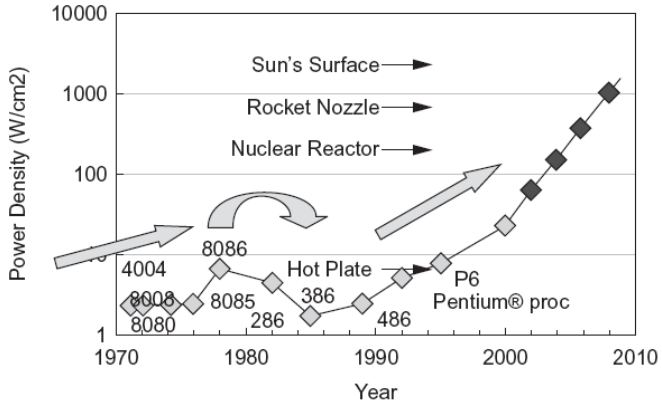
Reachable Radius

- We can't send a signal across a large fast chip in one cycle anymore
- But the microarchitect can plan around this
  - Just as off-chip memory latencies were tolerated



Dynamic Power

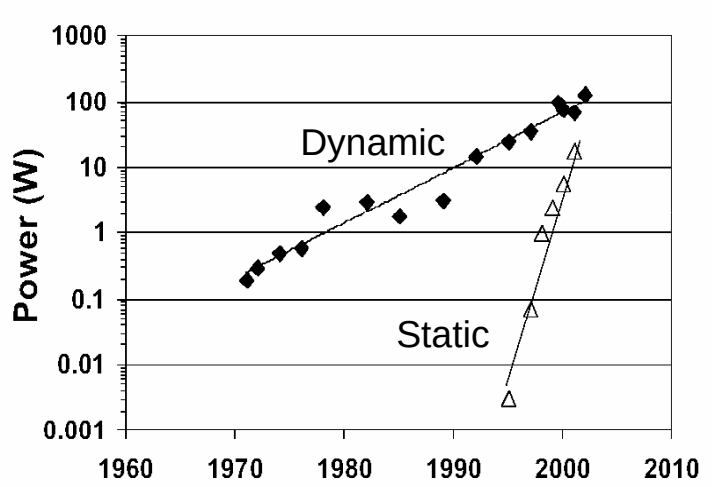
- Intel VP Patrick Gelsinger (ISSCC 2001)
  - If scaling continues at present pace, by 2005, high speed processors would have power density of nuclear reactor, by 2010, a rocket nozzle, and by 2015, surface of sun.
  - "Business as usual will not work in the future."
- Attention to power is increasing



[Intel]

Static Power

- 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]

Productivity

- Transistor count is increasing faster than designer productivity (gates / week)
  - Bigger design teams
    - Up to 500 for a high-end microprocessor
  - More expensive design cost
  - Pressure to raise productivity
    - Rely on synthesis, IP blocks
  - Need for good engineering managers

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

- 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