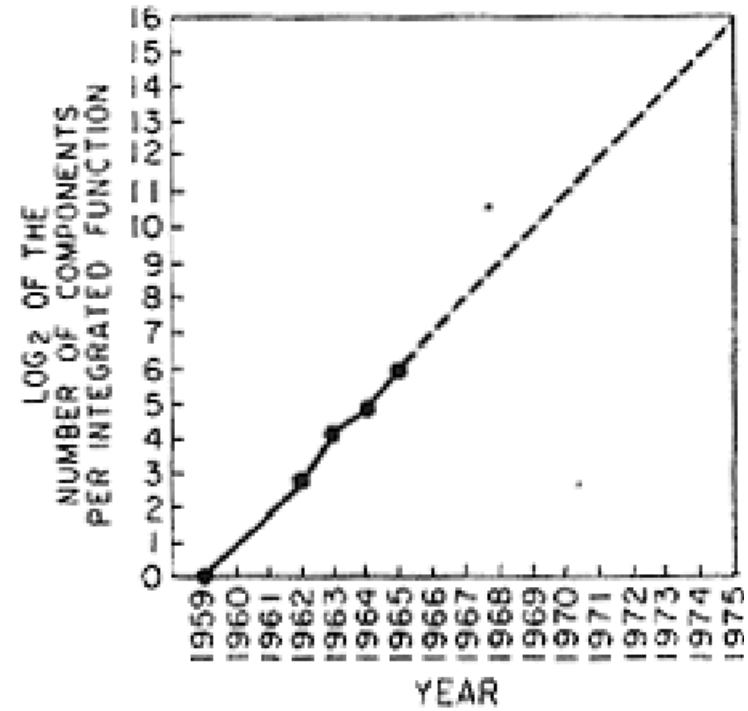
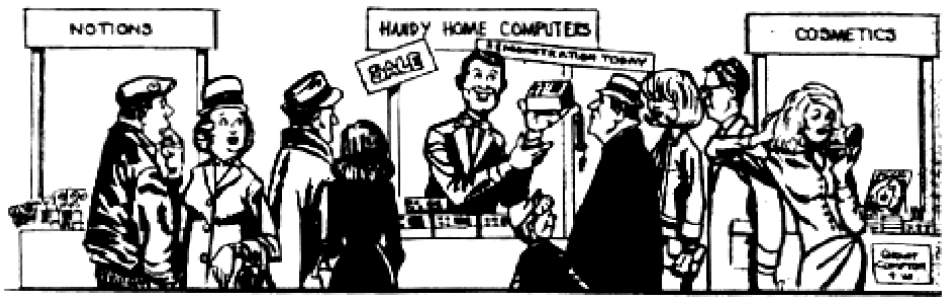
25. Scaling and Economics

- Previous Unit:
  - Clock Distribution
  - Clock Skew
  - Skew-Tolerant Static Circuits
  - Skew-Tolerant Domino Circuits
- This Unit:
  - Scaling
    - Transistors
    - Interconnect
    - Future Challenges
  - VLSI Economics

Moore's Law

- In 1965, Gordon Moore predicted the exponential growth of the number of transistors on an IC
- Transistor count doubled every year since invention
- Predicted > 65,000 transistors by 1975!
- Growth limited by power

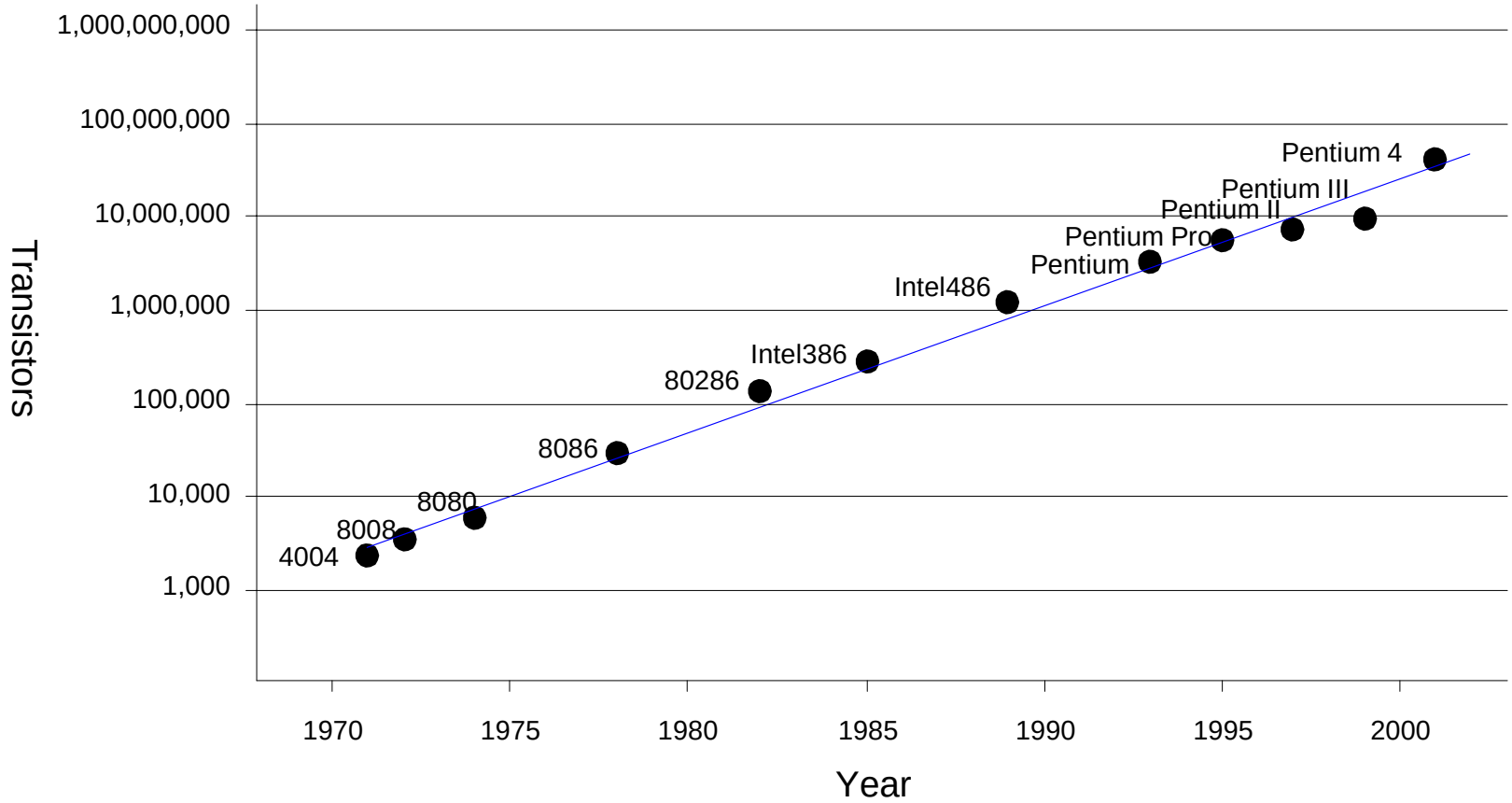




[Moore65]

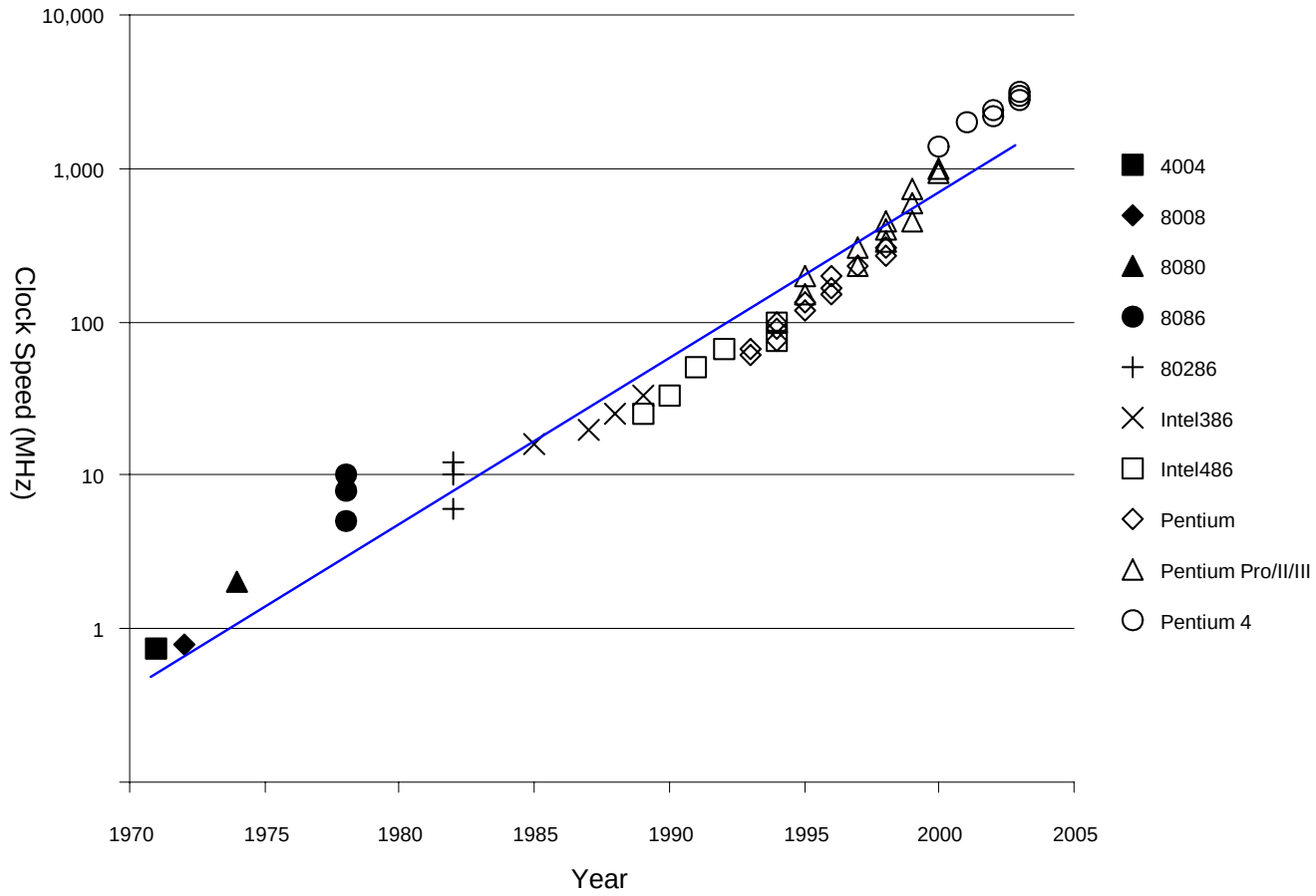
More Moore

- Transistor counts have doubled every 26 months for the past three decades.



Speed Improvement

- Clock frequencies have also increased exponentially
  - A corollary of Moore's Law



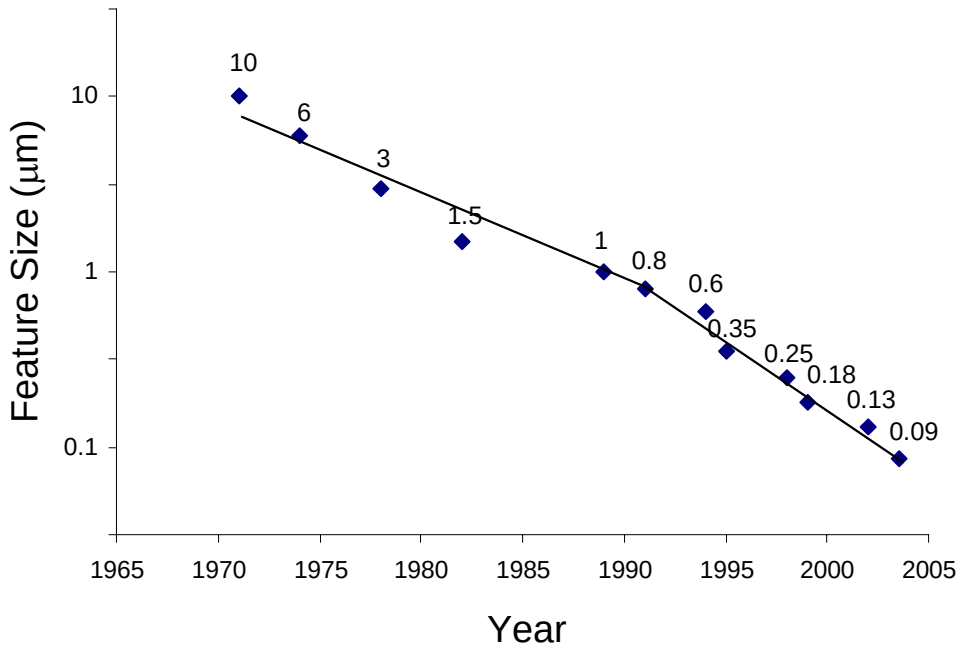
Why?

- 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
  - Wires do not improve (and may get worse)
- Scale factor S
  - Typically
  - Technology nodes

$$S = \sqrt{2}$$



Scaling Assumptions

- What changes between technology nodes?
- Constant Field Scaling
  - All dimensions (x, y, z => W, L, t_{ox})
  - Voltage (V_{DD})
  - Doping levels
- Lateral Scaling
  - Only gate length L
  - Often done as a quick gate shrink (S = 1.05)

Device Scaling

Table 4.15 Influence of scaling on MOS device characteristics

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

- Velocity saturation makes lateral scaling unsustainable

Solution

- Gate capacitance is typically about 2 fF/μm
- The FO4 inverter delay in the TT corner 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.

- FO4 = 5 τ = 15 RC
- RC = (0.5*f*) / 15 = (*f*/30) ps/nm
- If W = 2*f*, R = 8.33 kΩ
  - Unit resistance is roughly independent of *f*

Scaling Assumptions

- Wire thickness
  - Hold constant vs. reduce in thickness
- Wire length
  - Local / scaled interconnect
  - Global interconnect
    - Die size scaled by $D_c \approx 1.1$

Interconnect Scaling

Table 4.16 Influence of scaling on interconnect characteristics

Parameter	Sensitivity	Reduced Thickness	Constant Thickness
Scaling Parameters			
Width: w		$1/S$	
Spacing: s		$1/S$	
Thickness: t		$1/S$	1
Interlayer oxide height: h		$1/S$	
Characteristics Per Unit Length			
Wire resistance per unit length: R_w	$\frac{1}{wt}$	S^2	S
Fringing capacitance per unit length: C_{wf}	$\frac{t}{s}$	1	S
Parallel plate capacitance per unit length: C_{wp}	$\frac{w}{h}$	1	1
Total wire capacitance per unit length: C_w	$C_{wf} + C_{wp}$	1	between 1, S
Unrepeated RC constant per unit length: t_{wu}	$R_w C_w$	S^2	between S, S^2
Repeated wire RC delay per unit length: t_{wr} (assuming constant field scaling of gates in Table 4.15)	$\sqrt{RCR_w C_w}$	$\sqrt{S}$	between 1, $\sqrt{S}$
Crosstalk noise	$\frac{t}{s}$	1	S

Interconnect Delay

Table 4.16 Influence of scaling on interconnect characteristics

Parameter	Sensitivity	Reduced Thickness	Constant Thickness
Scaling Parameters			
Width: w		$1/S$	
Spacing: s		$1/S$	
Thickness: t		$1/S$	1
Interlayer oxide height: h		$1/S$	
Local/Scaled Interconnect Characteristics			
Length: l		$1/S$	
Unrepeated wire RC delay	$l^2 t_{wu}$	1	between $1/S$, 1
Repeated wire delay	$l t_{wr}$	$\sqrt{1/S}$	between $1/S$, $\sqrt{1/S}$
Global Interconnect Characteristics			
Length: l		D_c	
Unrepeated wire RC delay	$l^2 t_{wu}$	$S^2 D_c^2$	between SD_c^2, $S^2 D_c^2$
Repeated wire delay	$l t_{wr}$	$D_c\sqrt{S}$	between D_c, $D_c\sqrt{S}$

Observations

- Capacitance per micron is remaining constant
  - About 0.2 fF/μm
  - Roughly 1/10 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

Table 4.17 Predictions from the 2002 ITRS

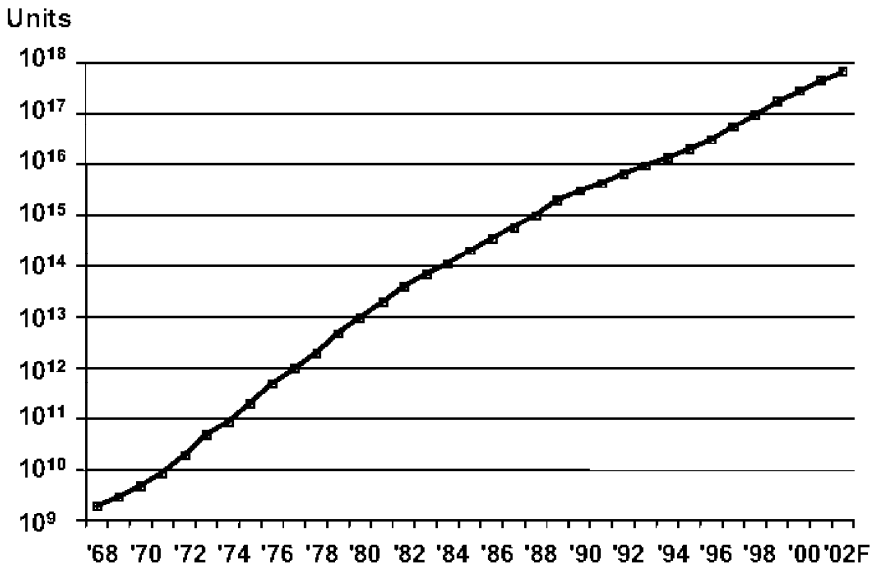
Year	2001	2004	2007	2010	2013	2016
Feature size (nm)	130	90	65	45	32	22
V_{DD} (V)	1.1–1.2	1–1.2	0.7–1.1	0.6–1.0	0.5–0.9	0.4–0.9
Millions of transistors/die	193	385	773	1564	3092	6184
Wiring levels	8–10	9–13	10–14	10–14	11–15	11–15
Intermediate wire pitch (nm)	450	275	195	135	95	65
Interconnect dielectric constant	3–3.6	2.6–3.1	2.3–2.7	2.1	1.9	1.8
I/O signals	1024	1024	1024	1280	1408	1472
Clock rate (MHz)	1684	3990	6739	11511	19348	28751
FO4 delays/cycle	13.7	8.4	6.8	5.8	4.8	4.7
Maximum power (W)	130	160	190	218	251	288
DRAM capacity (Gbits)	0.5	1	4	8	32	64

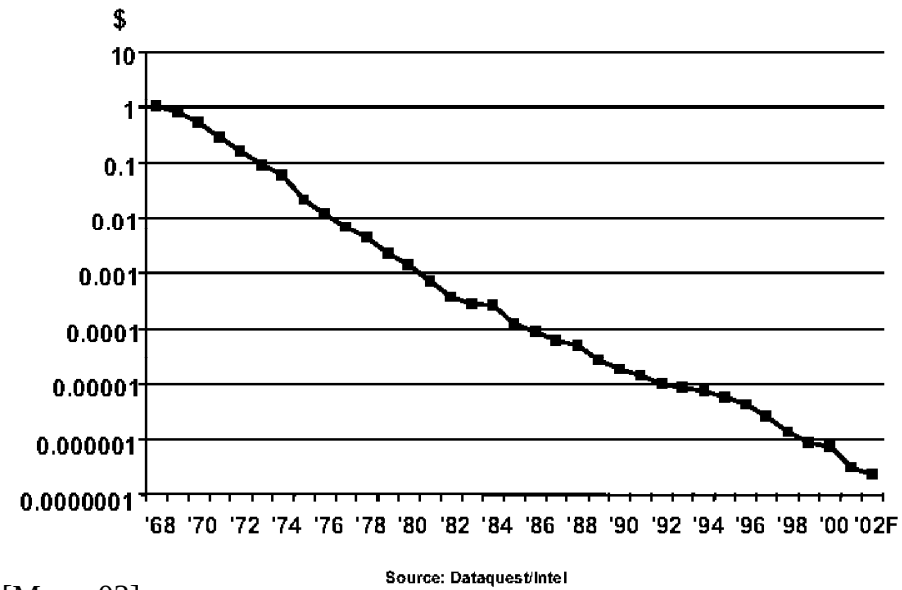
Scaling Implications

- 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

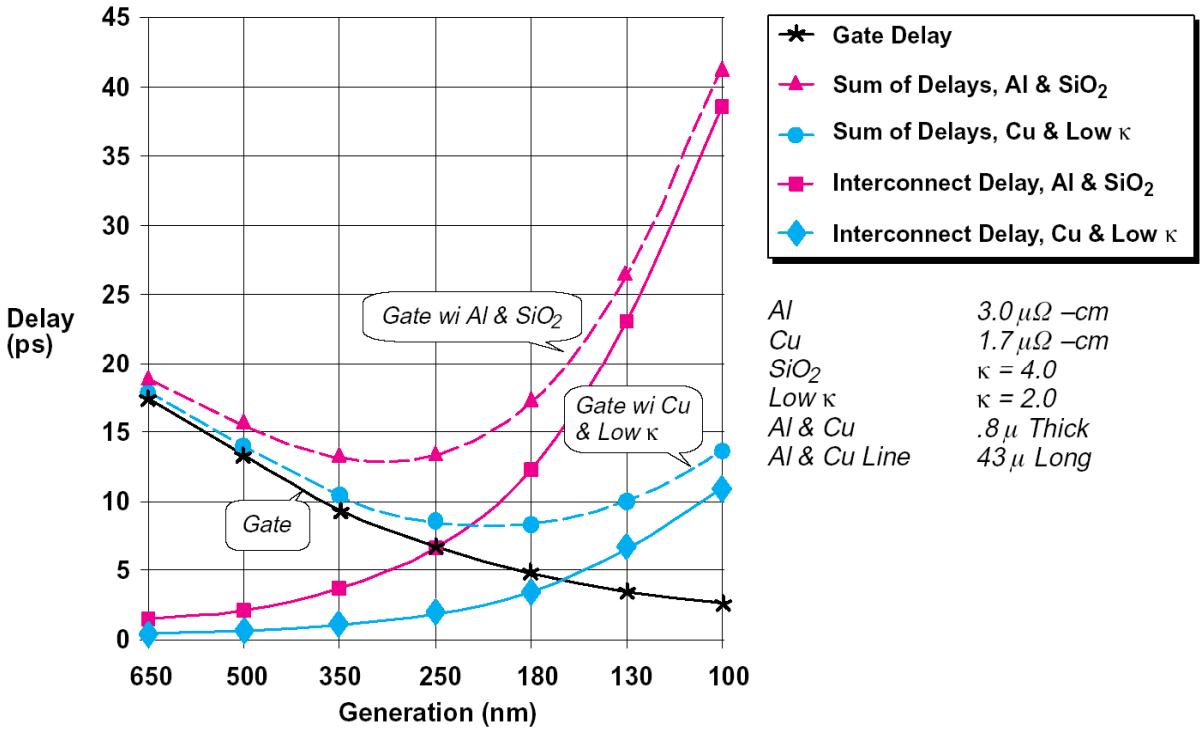




[Moore03]

Interconnect Woes

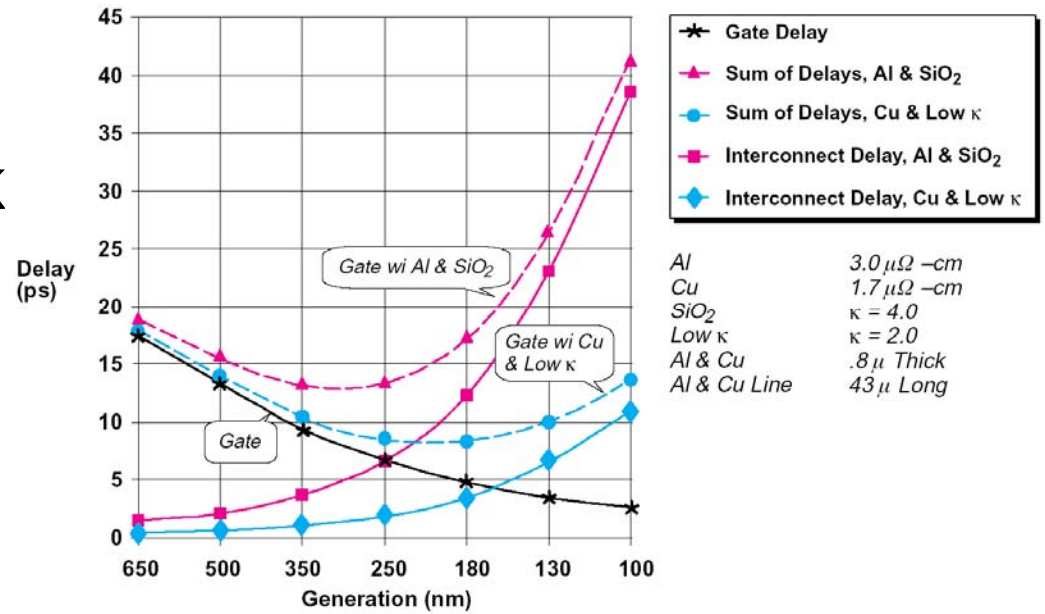
- SIA made a gloomy forecast in 1997
  - Delay would reach minimum at 250 – 180 nm, then get worse because of wires
- But…



[SIA97]

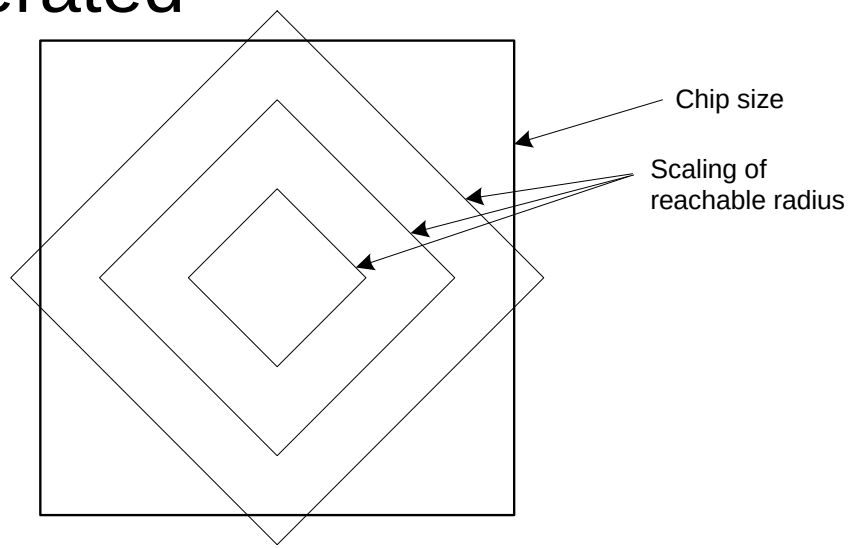
Interconnect Woes

- 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
- 100k gate blocks ok



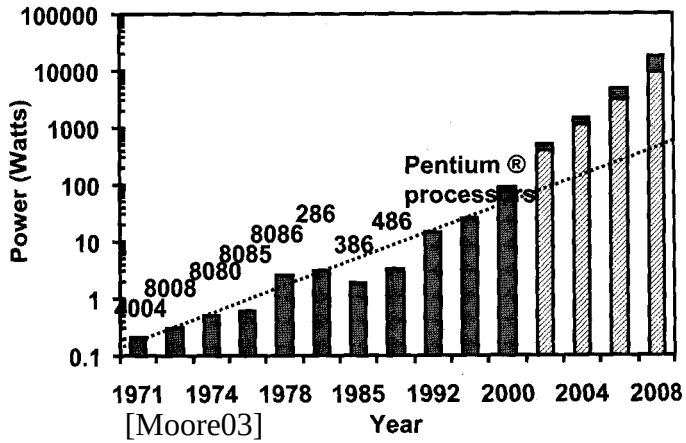
Reachable Radius

- 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

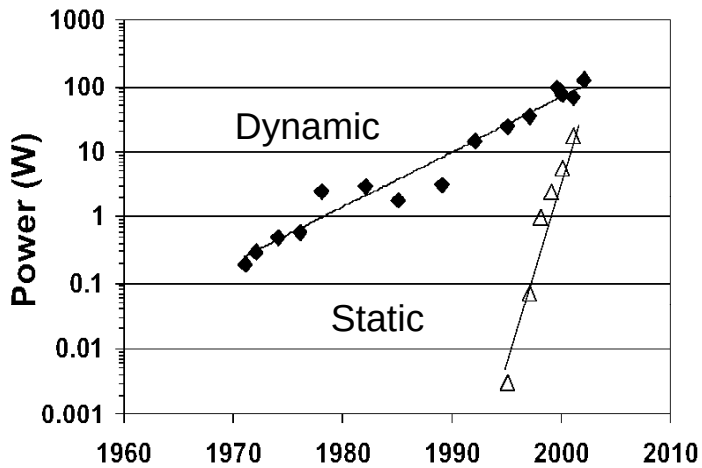
- Intel's Patrick Gelsinger (ISSCC 2001)
  - If scaling continues at present pace, by 2005, high speed processors would have power density of nuclear reactor, by 2010, a rocket nozzle, and by 2015, surface of sun.
  - "Business as usual will not work in the future."
- Intel stock dropped 8% on the next day
- But attention to power is increasing



[Moore03]

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: $500k – 1M in 130 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: $1M
  - Total: $2M / year
- Summary
  - 2 years @ $3.91M / year
  - $8M design & prototype

Cost Breakdown

- New chip design is fairly capital-intensive
- Maybe you can do it for less?