EE3230 Lecture 6: Circuit Characterization and Performance Estimation III

Ping-Hsuan Hsieh (謝秉璇)

Delta Building R908
EXT 42590
phsieh@ee.nthu.edu.tw

Outline

- Delay estimation
- Logical effort and transistor sizing
- Power dissipation
- Interconnect
- Wire engineering
- **Design margin**
- Reliability
- Scaling

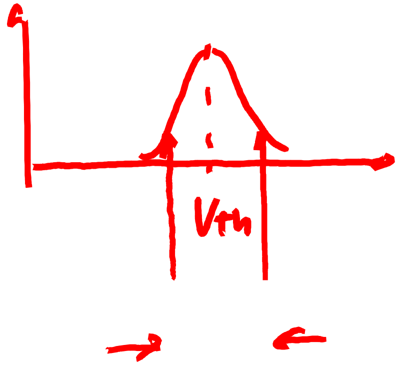
Ch7

Design Margin

- Sources of variations
  - Process variation ← distribution
  - Supply voltage
  - Temperature

Vth nominal value

3σ value

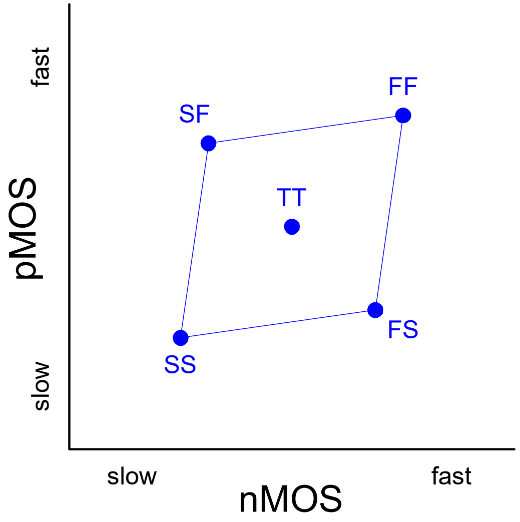


99.7%

memory的設計
必須考慮到6 or 7σ.

Process Variation

- Uncertainties in device parameters during manufacturing process
  - L_{eff}, V_{th}, t_{ox}, etc.
  - Vary around **T**ypical (nominal) values
- **F**ast
  - L_{eff}: short
  - V_{th}: low
  - t_{ox}: thin
- **S**low: opposite
- Not all parameters are independentfor NMOS and PMOS



Environmental Variation

- Supply voltage and temperature also vary
  - Across time and space
- Fast:
  - V_{DD}: ______
  - T: ______

	Corner	Voltage	Temperature
+10%	Fast	1.98V	0°C, −40°C
	Typical	V_{DD}= 1.8 V	70ºC
−10%	Slow	1.62V	100°C, 125°C

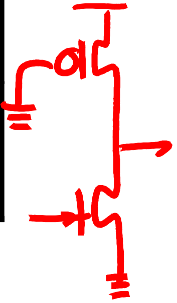
Process Corners

- Process corners describe extreme case variations
  - If a design works in all corners, it will probably work for any variation.
- Describe corner with four letters (T, F, S)
  - NMOS speed
  - PMOS speed
  - Voltage
  - Temperature

Important Corners

- Some critical corners include
  - Worst corner depending on performance of interest

Purpose	NMOS	PMOS	V_{DD}	Temp
Cycle time	S	S	S 低	S 高
Power	F	F	F 高	F 低
Subthrehold leakage	F	F	F	S 高
Pseudo-nMOS	S	F	?	?



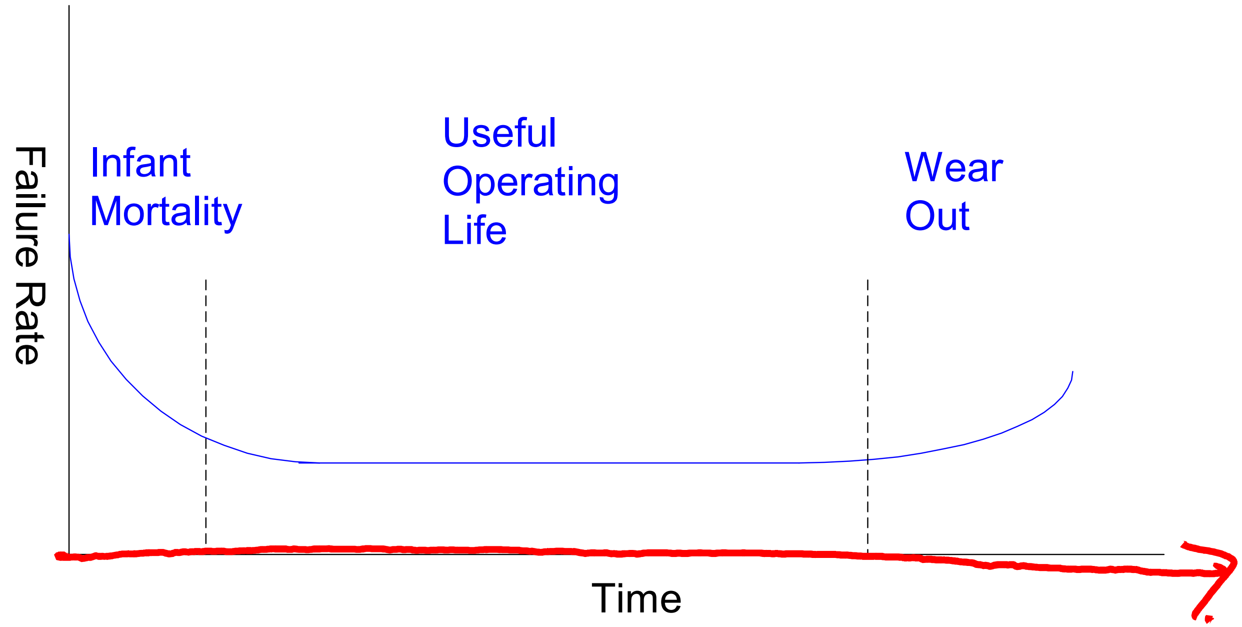
Outline

- Delay estimation
- Logical effort and transistor sizing
- Power dissipation
- Interconnect
- Wire engineering
- Design margin
- **Reliability**
- Scaling

Reliability

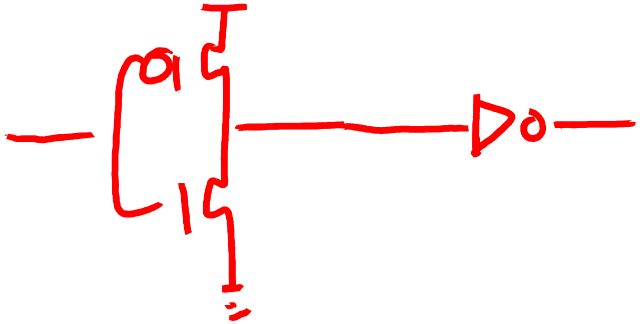
- Hard errors
- Soft errors

burn-in test



Electromigration

- “Electron wind” causes movement of metal atoms along wires
- Excessive electromigration leads to open circuits
- Most significant for unidirectional (DC) current
  - Depends on current density J_{dc} (current/area)
  - Exponential dependence on temperature
  - Black’s Equation: $MTTF \propto \frac{e^{\frac{E_a}{kT}}}{J_{dc}^{\;n}}$
  - Typical limits: J_{dc} < 1 – 2 mA / μm^2



Self-Heating

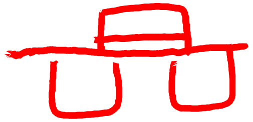
- Current through wire resistance generates heat
  - Oxide surrounding wires is a thermal insulator
  - Heat tends to build up in wires
  - Hotter wires are more resistive → slower
- Self-heating limits AC current densities for reliability

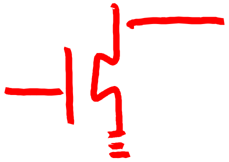
$$I_{rms} = \sqrt{\frac{\int_0^T I(t)^2 \, dt}{T}}$$

  - Typical limits: $J_{rms} < 15\ \mathrm{mA} / \mu\mathrm{m}^{\cancel{2}}$

Hot Carriers

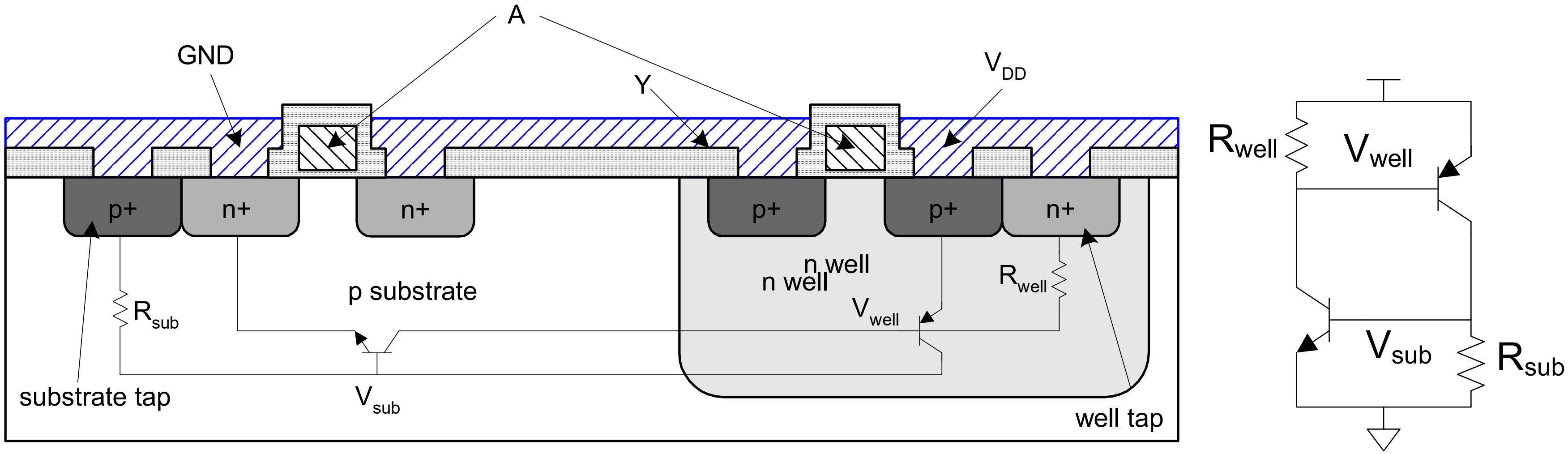
- Electric fields across channel impart high energies to some carriers
  - These "hot" carriers may be blasted into gate oxide where they become trapped
  - Accumulation of charge in oxide causes shift in V_{th} over time
  - Eventually V_{th} shifts too far for devices to operate correctly
- Choose V_{DD} to achieve reasonable product lifetime
  - Worst problems for inverters and NORs with ~~slow~~ fast input risetime and long propagation delays





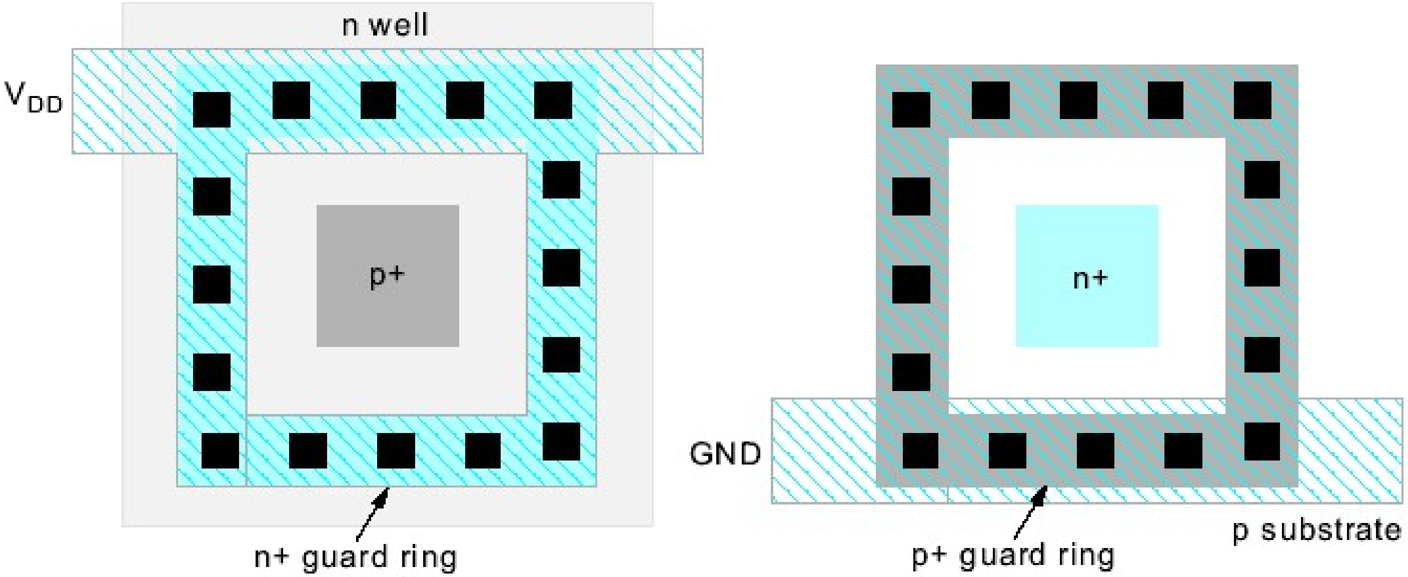
Latchup

- Positive feedback leading to V_{DD} – GND short
  - Major problem for 1970's CMOS processes before it was well understood
- Avoid by minimizing resistance of body to GND / V_{DD}
  - Use plenty of substrate and well contacts (taps)



Guard Rings

- Latchup when diffusion-to-substrate diodes could become forward-biased
- Surround sensitive region with guard rings to collect injected charge



Overvoltage

- High voltages can damage transistors
  - Electrostatic discharge
  - Oxide arcing
  - Punch-through
  - Time-dependent dielectric breakdown (TDDB)
    - Accumulated wear from tunneling currents
- Requires low V_{DD} for thin oxides and short channels
- Use ESD protection structures where chip meets real world

Summary

- Static CMOS gates are very robust
  - Will settle to correct value if you wait long enough
- Other circuits suffer from a variety of pitfalls
  - Tradeoff between performance & robustness
- Very important to check circuits for pitfalls
  - For large chips, you need an automatic checker
  - Design rules aren't worth the paper they are printed on unless you back them up with a tool

Soft Errors

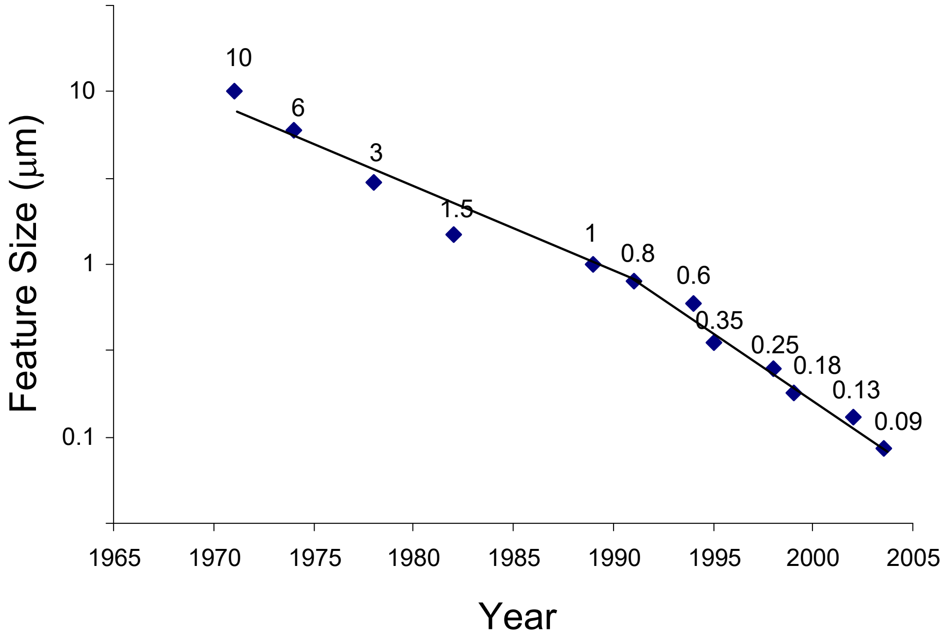
- In 1970’s, DRAMs were observed to occasionally flip bits for no apparent reason
  - Ultimately linked to alpha particles and cosmic rays
- Collisions with particles create electron-hole pairs in substrate
  - These carriers are collected on dynamic nodes, disturbing the voltage
- Minimize soft errors by having plenty of charge on dynamic nodes
- Tolerate errors through ECC (redundancy)

Outline

- Delay estimation
- Logical effort and transistor sizing
- Power dissipation
- Interconnect
- Wire engineering
- Design margin
- Reliability
- **Scaling**

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 $S = \sqrt{2}$
  - Technology nodes



Scaling Assumptions (I)

- 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

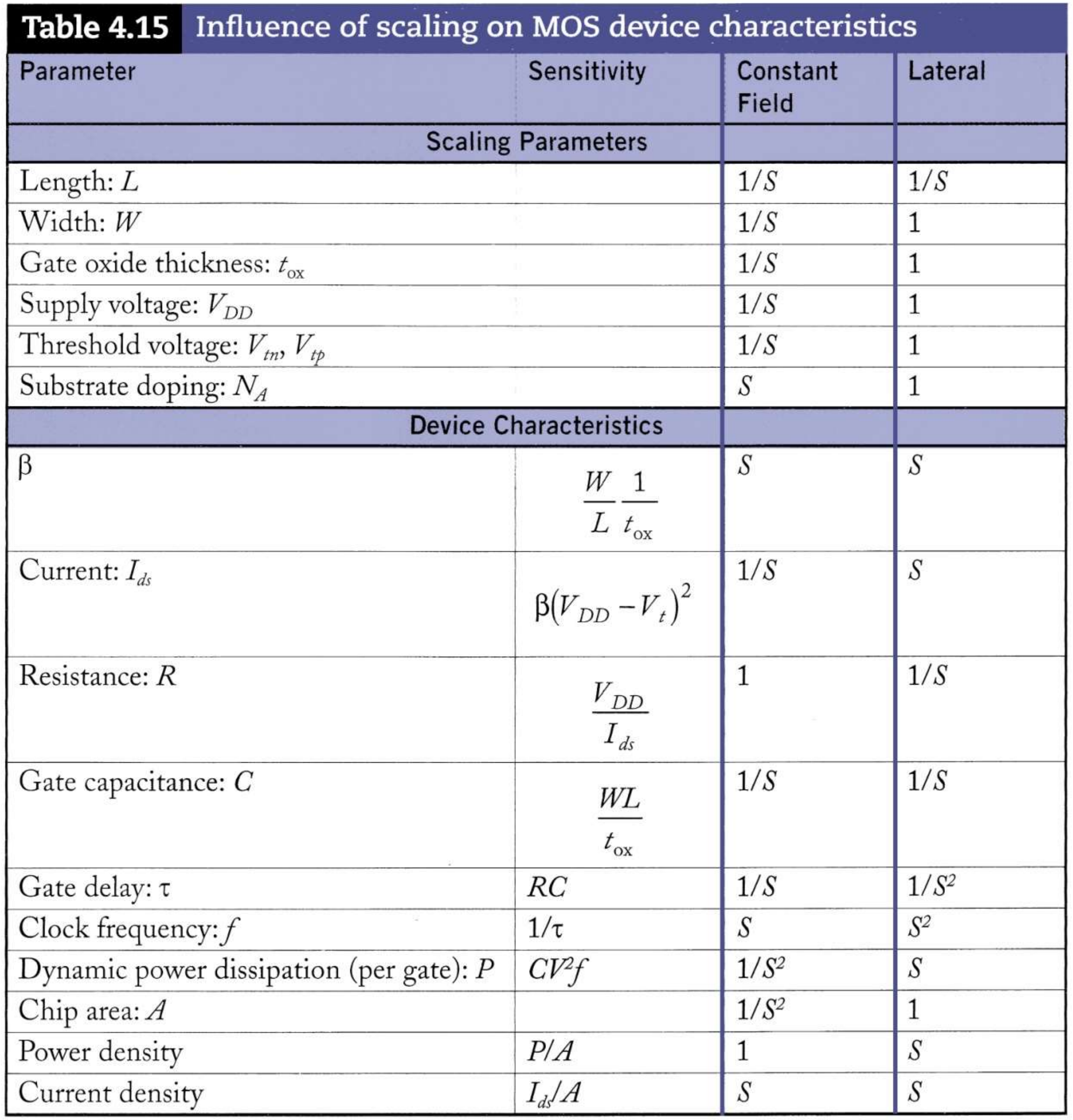


Table 4.15 Influence of scaling on MOS device characteristics

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	$\frac{W}{L}\frac{1}{t_{\text{ox}}}$	S	S
Current: I_{ds}	$\beta\left(V_{DD}-V_t\right)^2$	$1/S$	S
Resistance: R	$\frac{V_{DD}}{I_{ds}}$	1	$1/S$
Gate capacitance: C	$\frac{WL}{t_{\text{ox}}}$	$1/S$	$1/S$
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

Example

- Gate capacitance is typically about 2 fF/μm
- The FO4 inverter delay in the TT corner for a process of feature size 2λ (in nm) is about λ ps
- Estimate the ON resistance of a unit (4λ wide) NMOS transistor
- FO4 = 5 τ = 15 RC
- RC = λ/15 ps/nm
- If W = 4λ, C = 8λ*(1e-3) fF, R = 8.33 kΩ
  - Unit resistance is roughly independent of λ

Scaling Assumptions (II)

- Wire thickness
  - constant vs. reduced through scaling
- Wire length
  - Local / scaled interconnect
  - Global interconnect
    - Die size scaled by $D_c \approx 1.1$

Interconnect Scaling

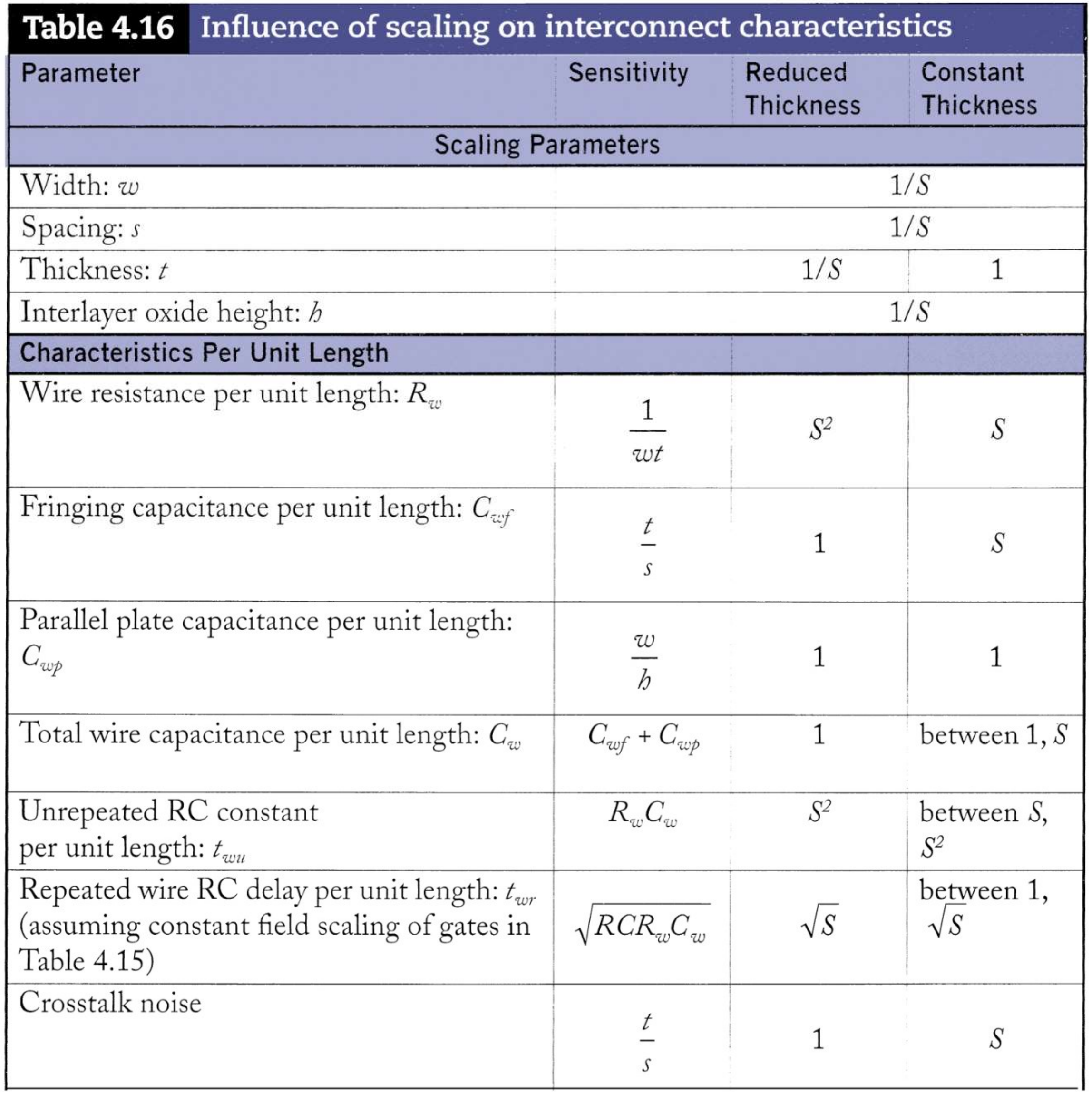


Table 4.16 Influence of scaling on interconnect characteristics

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

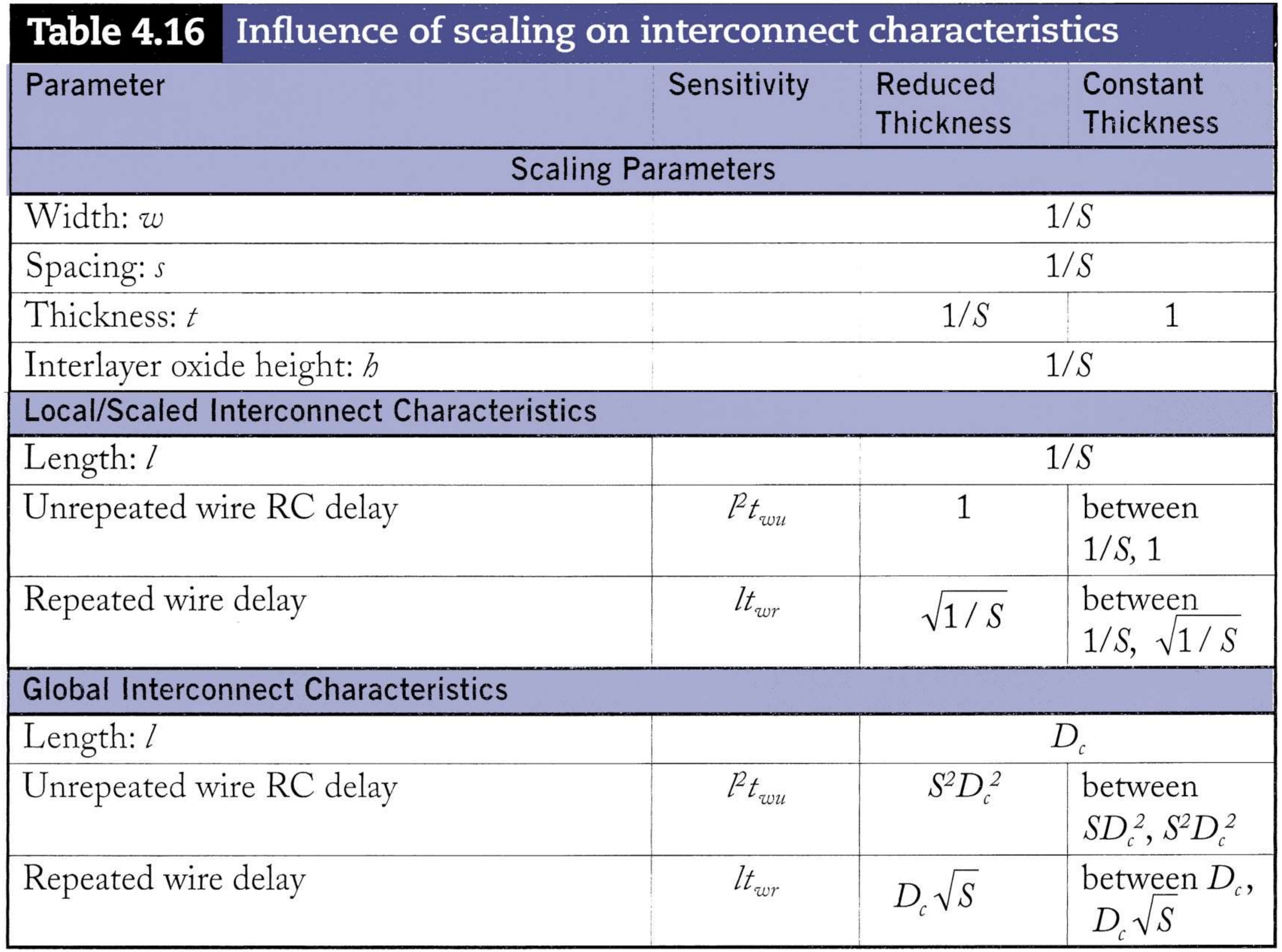


Table 4.16 Influence of scaling on interconnect characteristics

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 $S D_c^2$, $S^2 D_c^2$
Repeated wire delay	$l t_{wr}$	$D_c \sqrt{S}$	between D_c, $D_c \sqrt{S}$

Obesrvations

- Capacitance per micron of wire remains constant
  - About 0.2 *f*F/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 Roadmap

- Semiconductor Industry Association forecast
  - International Technology Roadmap for Semiconductors

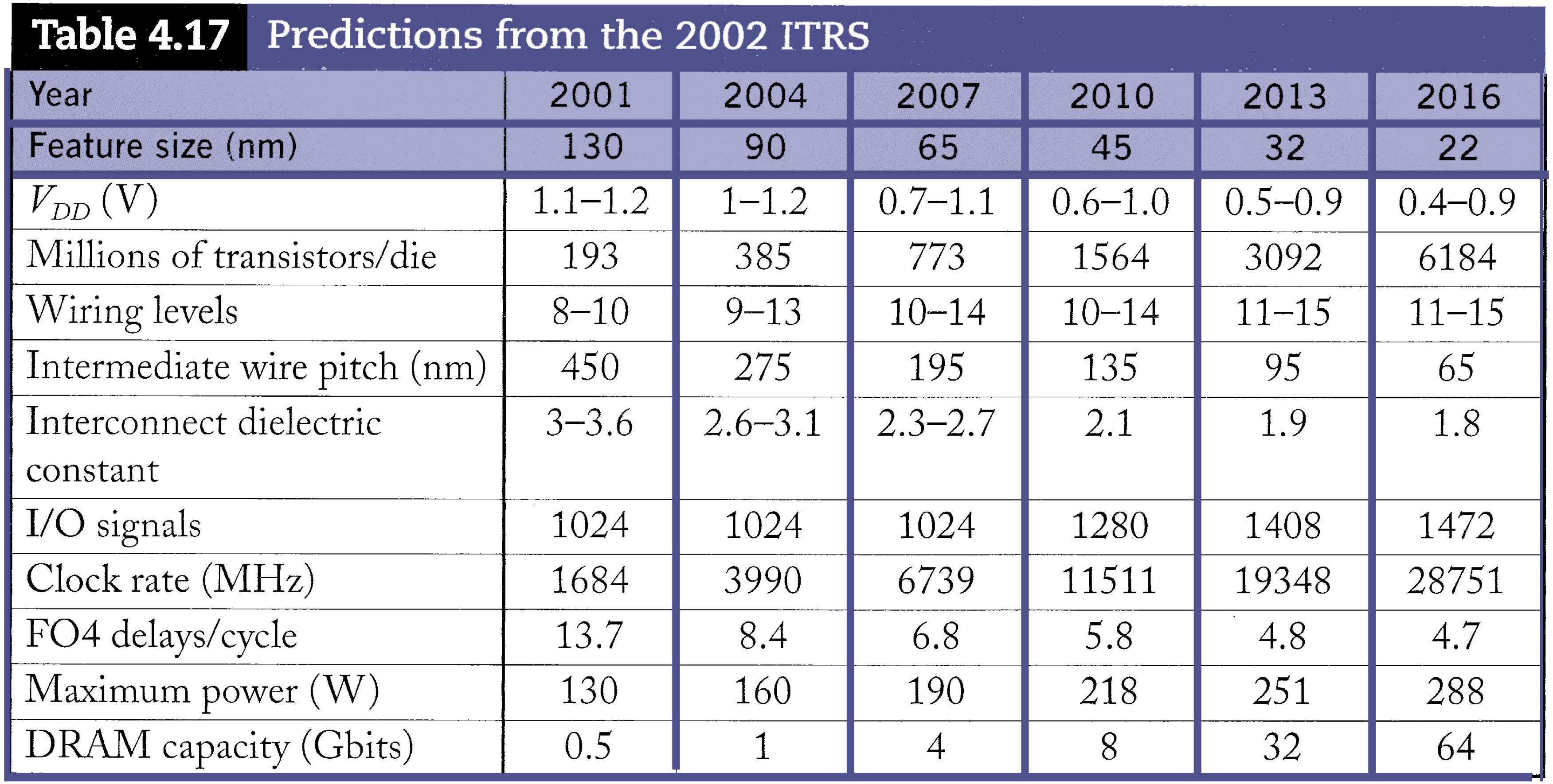


Table 4.17 Predictions from the 2002 ITRS

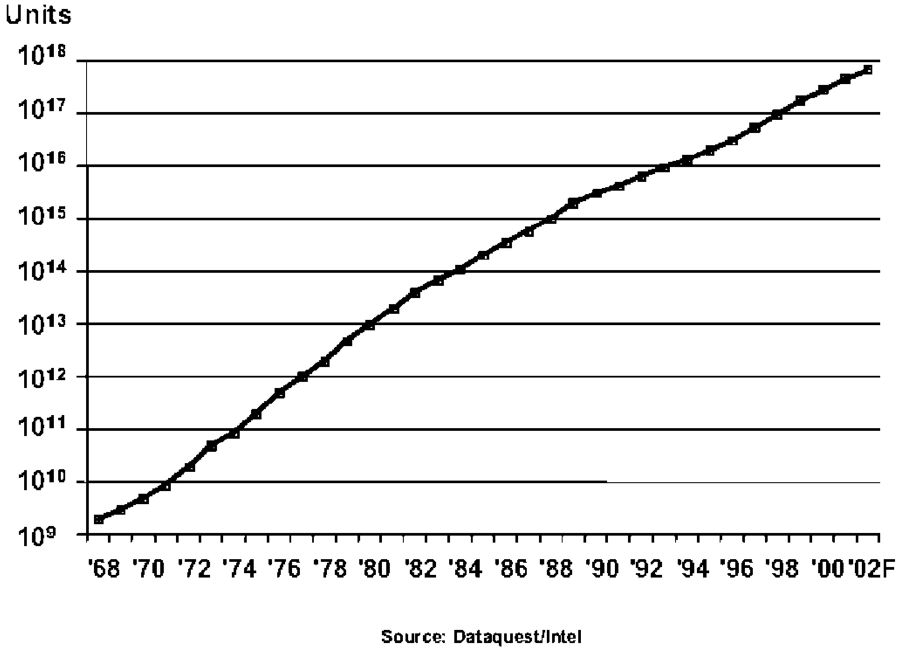
Year	2001	2004	2007	2010	2013	2016
Feature size (nm)	130	90	65	45	32	22
V_{DD} (V)	1.1–1.2	1–1.2	0.7–1.1	0.6–1.0	0.5–0.9	0.4–0.9
Millions of transistors/die	193	385	773	1564	3092	6184
Wiring levels	8–10	9–13	10–14	10–14	11–15	11–15
Intermediate wire pitch (nm)	450	275	195	135	95	65
Interconnect dielectric 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

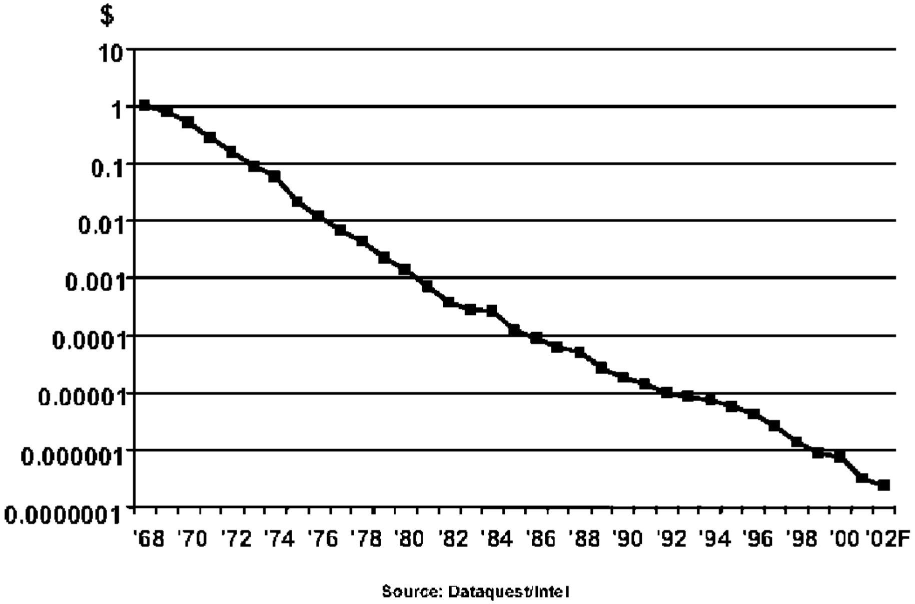
- Improved performance
- Improved cost
- Interconnect woes
- Power woes
- Productivity challenges
- Physical limits

Cost Improvement

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



Source: Dataquest/Intel

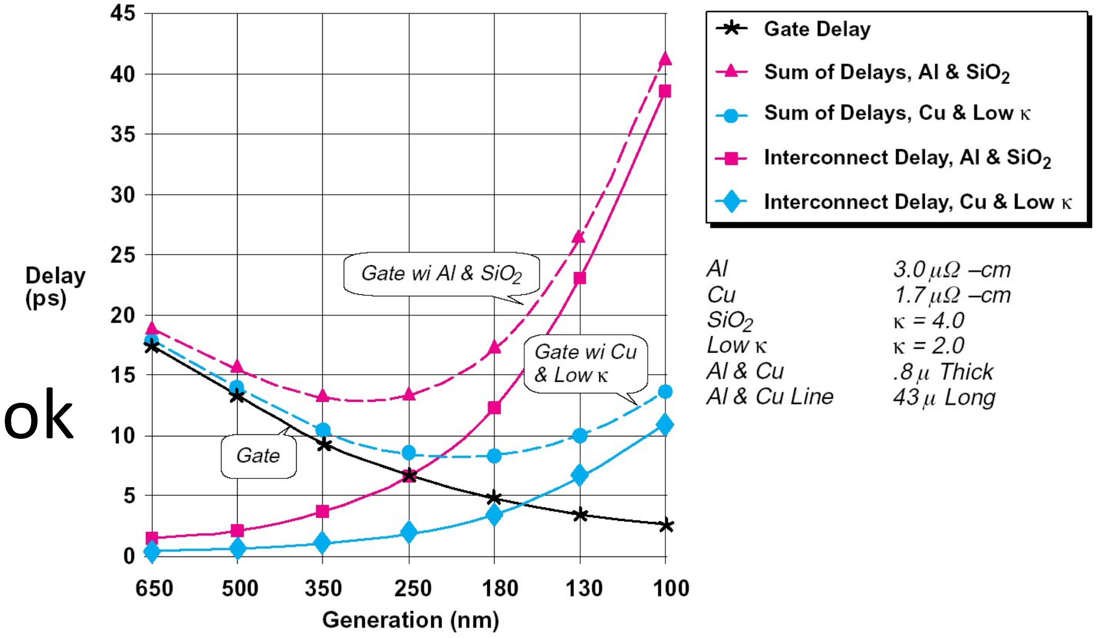


Source: Dataquest/Intel

[Moore ’03]

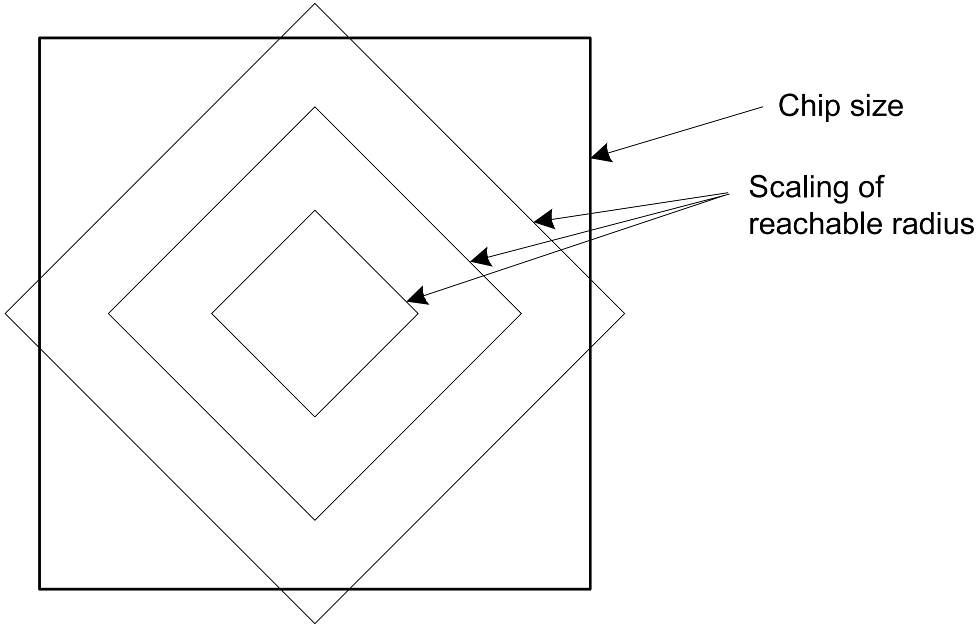
Interconnect

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



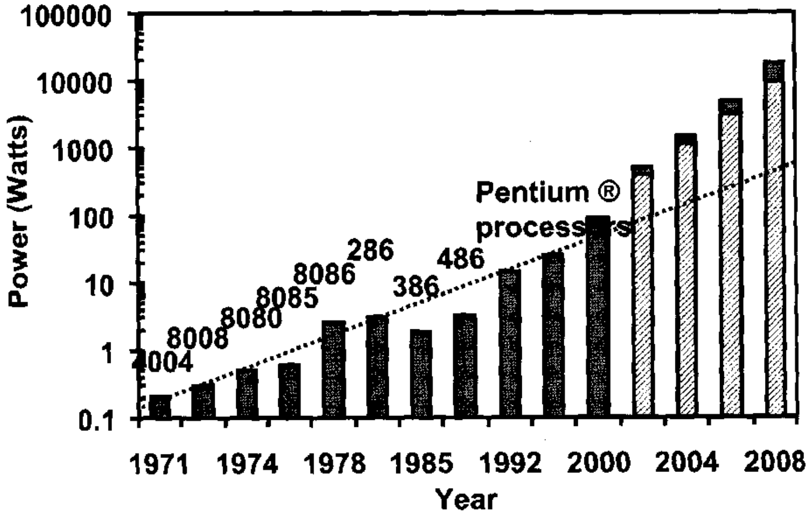
Reachable Radius

- We cannot send a signal across a large fast chip in one cycle anymore
- But micro-architect can plan around this
  - Just as off-chip memory latencies were tolerated



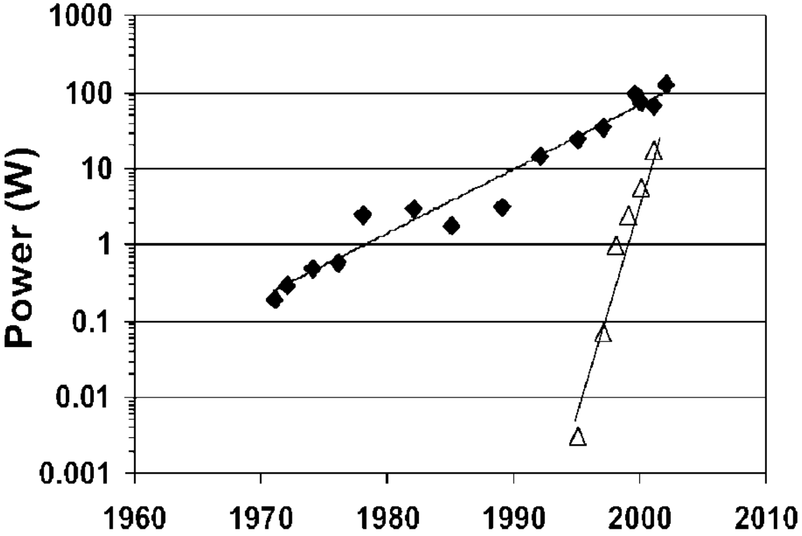
Dynamic Power Consumption

- 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"
- Intel stock dropped 8% on the next day
- But attention to power is increasing



Static Power Consumption

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



[Moore ’03]

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 Limitations

- 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