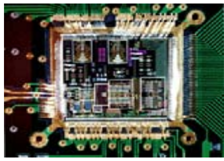
Lecture 6 : Design Margin, Reliability and Scaling

CSCE 6730
Advanced VLSI Systems

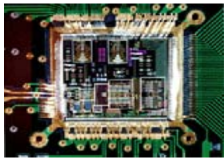
Instructor: Saraju P. Mohanty, Ph. D.

NOTE: The figures, text etc included in slides are borrowed from various books, websites, authors pages, and other sources for academic purpose only. The instructor does not claim any originality.



Lecture Outline

- Design Margin
  - Supply Voltage, Temperature, Process Variation, etc.
- Reliability
  - Electromigration, Self-heating, Hot-carriers, etc.
- Scaling
  - Transistors, Interconnect, etc.



Design Margin

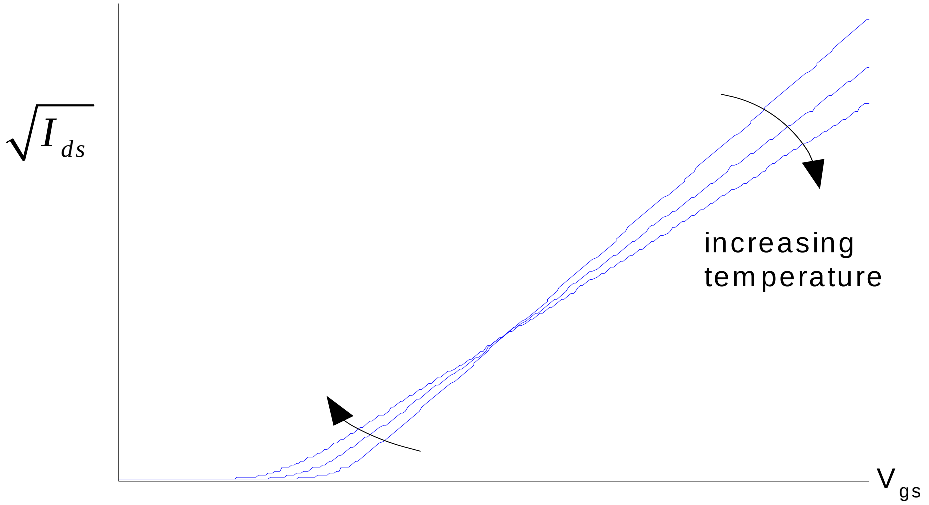
- Three different sources of variation:
  - Environmental
    - Supply voltage
    - Operating temperature
  - Manufacturing
    - Process variation
- Variations can be modeled as uniform or Gaussian distribution.
- **Objective**: Design a circuit that operates reliable over extreme ranges of the above variations.

Design Margin : Supply Voltage

- ICs are designed to operate at nominal supply voltage.
- Supply voltage may vary due to:
  - IR drop
  - L di/dt noise (self inductance)
  - M di/dt noise (mutual inductance)
- The delay in a device (t_d) that determines the maximum frequency (f_{max}) or the clock cycle time (T) is $T_d = k\, V_{dd} / (V_{dd}-V_{th})^{\alpha}$. Here, *k* and α are technology dependent constants.

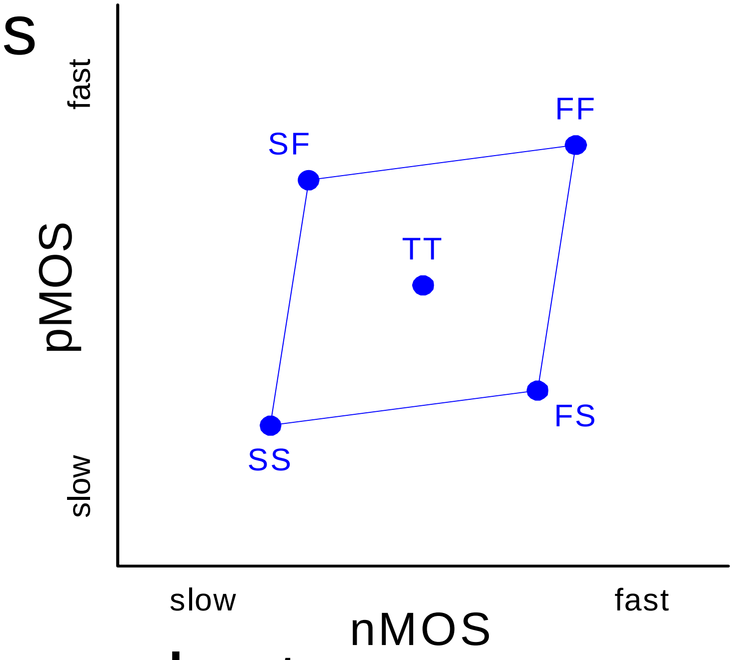
Design Margin : Temperature Sensitivity

- Increasing temperature
  - Reduces mobility
  - Reduces V_{th}
- I_{ON} decreases with temperature
- I_{OFF} increases with temperature



Design Margin : Parameter Variation

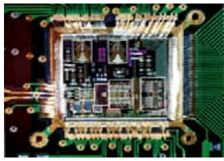
- Transistors have uncertainty in parameters
  - Process: L_{eff}, V_{th}, t_{ox} of nMOS and pMOS
  - Vary around typical (T) values
- Fast (F)
  - L_{eff}: short
  - V_{th}: low
  - t_{ox}: thin
- Slow (S): opposite
- Not all parameters are independent for nMOS and pMOS



Design Margin : Environmental Variation

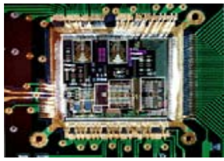
- V_{DD} and T also vary in time and space
- Fast:
  - V_{DD}: high
  - T: low

Corner	Voltage	Temperature
F	1.98 V	0 °C
T	1.8 V	70 °C
S	1.62 V	125 °C



Design Margin : Corners

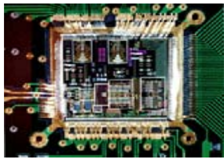
- Process corners describe worst case variations
  - If a design works in all corners, it will probably work for any variation.
- Describe corner with four letters (T, F, S)
  - NMOS speed
  - PMOS speed
  - Voltage
  - Temperature



Design Margin : Important Corners

- Some critical simulation corners include

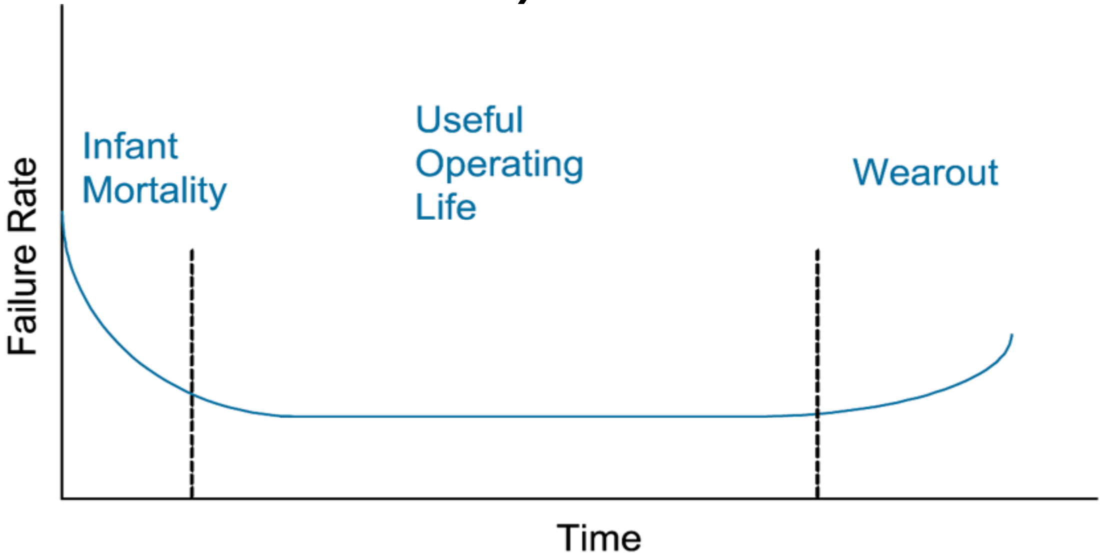
Purpose	**NMOS**	**PMOS**	**V_{DD}**	**Temperature**
Cycle time	S	S	S	S
Power	F	F	F	F
Subthrehold leakage	F	F	F	S
Pseudo-NMOS	S	F	?	?



Reliability

- Designing reliable CMOS chips are essential.
- **Mean Time Between Failure :**

  MTBF = (#devices * Hrs of Operation) / # Failures
- **Failures in Time (FIT) :** The number of failures that would occur every thousand hours per million devices, i.e. 10^9 * (failure rate / hour).



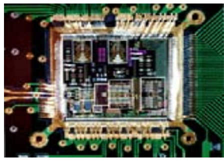
Reliability bathtub curve

Reliability : Electromigration

- Electromigration decreases reliability.
- Depends on current density.
- Occurs in wires carrying DC rather than AC, as in DC the electrons flow in a same direction.
- Mean Time to Failure :

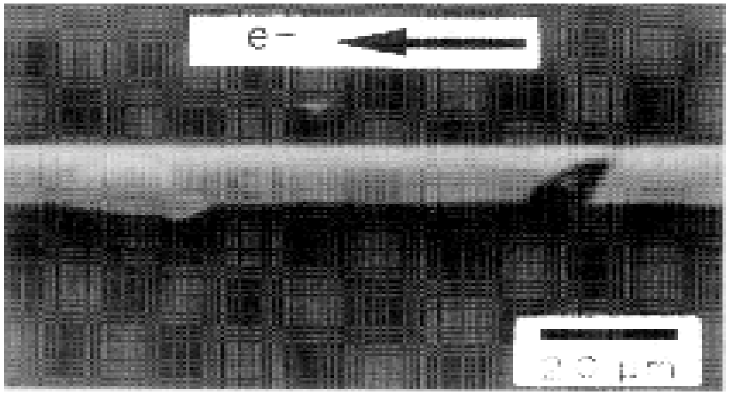
  $$\mathrm{MTF}\ \alpha\ \exp(E_a/kT)\ /\ J^{n}_{dc}$$

  Here, E_a is active energy (can be experimentally determined) and n is constant (=2).
- The electromigration DC current limits vary with materials, severe for aluminum than copper.

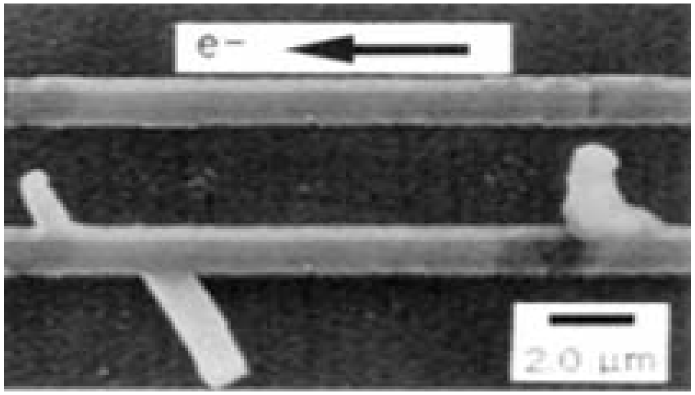


Reliability : Electromigration

- For electromigration we need a lot of electrons, and also we need electron scattering. Electromigration does not typically occur in semiconductors, but may in some very heavily doped semiconductor materials.
- Electromigration can lead to either open circuit or short circuit failure.



Open circuit failure



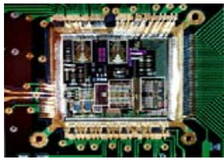
Hillocking, short circuit failure

Source: http://www.csl.mete.metu.edu.tr/Electromigration/emig.htm

Reliability : Self-heating

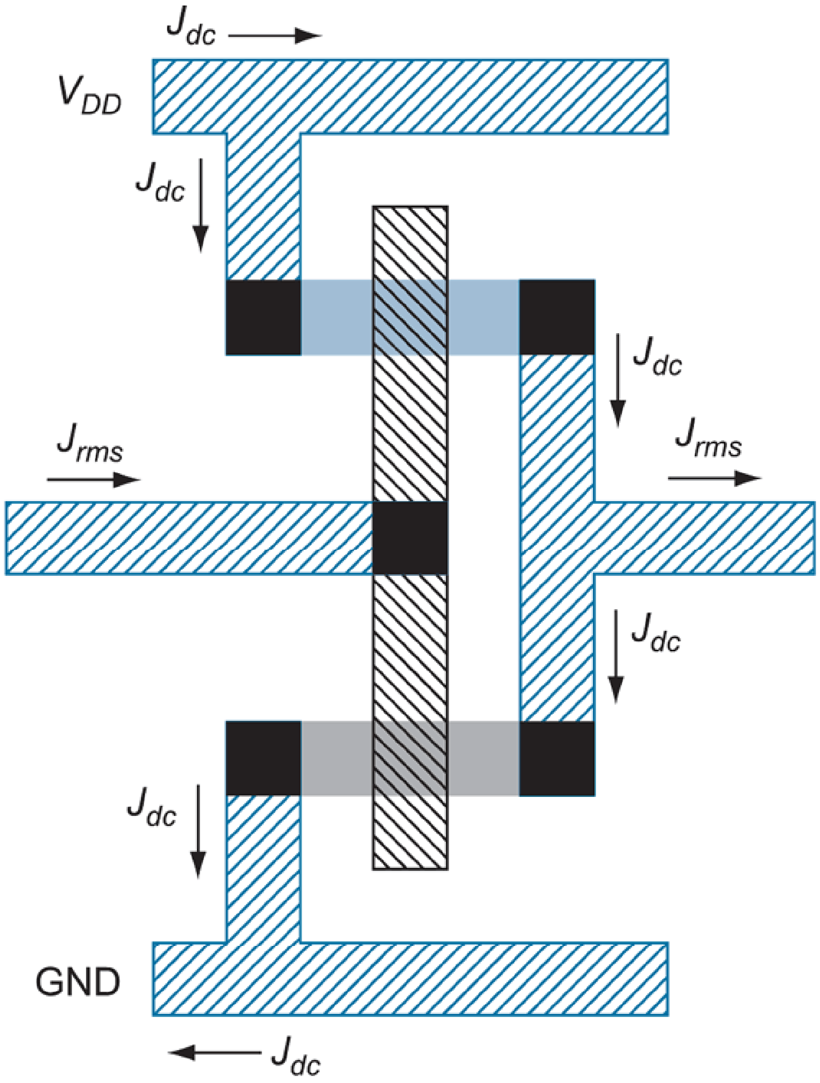
- Typically bidirectional signal line’s RMS (root mean square) current density is limited by self-heating.
- Self-heating may cause temperature-induced electromigration problems in bidirectional signal lines.
- Self-heating is more prominent for SOI processes because of poor thermal conductivity of SiO_2.
- RMS current is calculated as:

$$I_{rms} = \surd\,(\int I(t)^2 dt \,/\, T)$$



Reliability : Self-heating and Electromigration

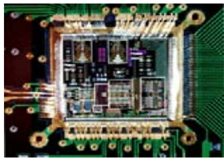
- Both DC and AC current density limit the operation.
- **DC current**: problem in power and ground lines
- **AC current**: problem in bidirectional signal lines
- Solution: widening the lines or reducing the transistor sizes, subsequently the current.



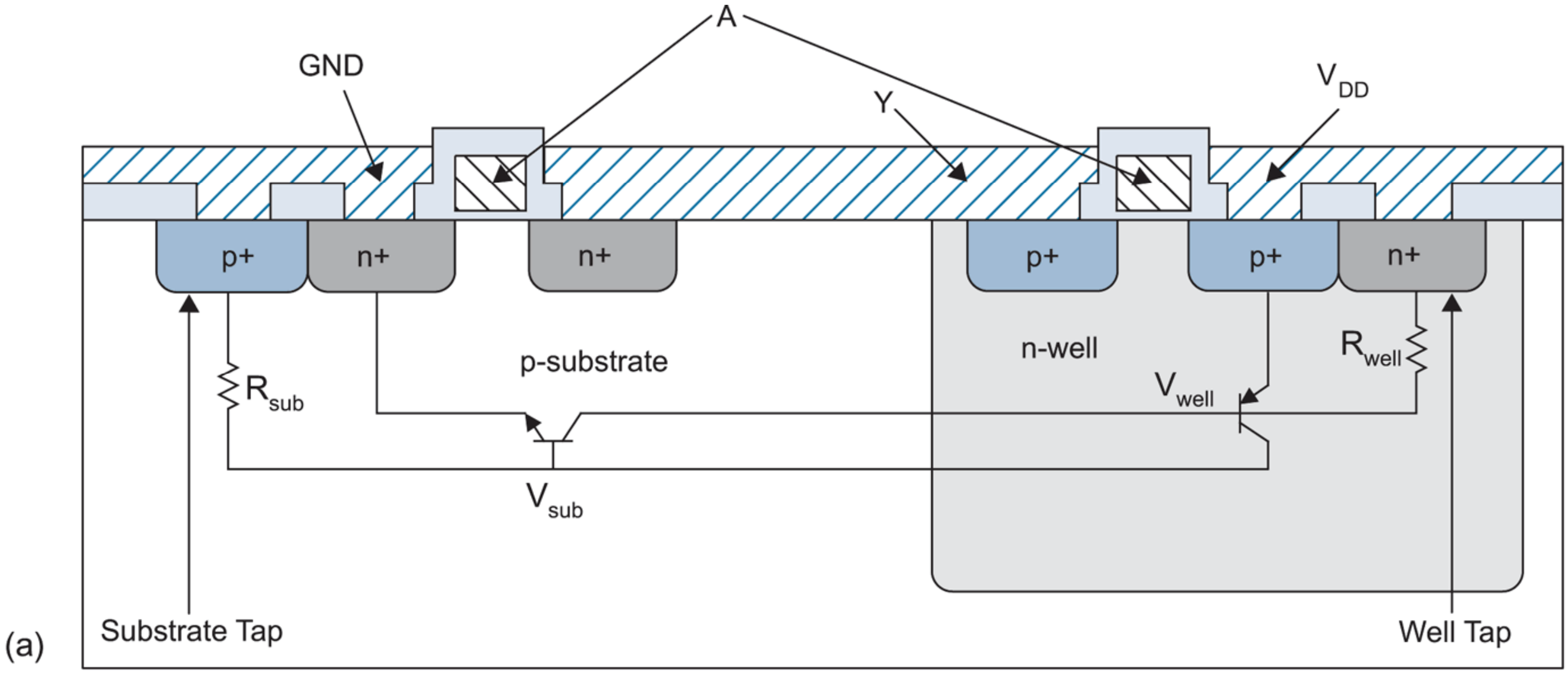
Current density limits in an inverter

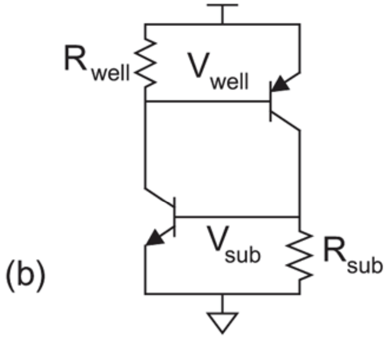
Reliability : Hot Carriers

- High-energy (hot) carriers get injected into the gate oxide and get trapped there.
- The damaged oxide change the IV characteristics:
  - Reduced current in NMOS
  - Increases current in PMOS
- Hot carriers may cause circuit wearout as NMOS transistors become too slow.
- Negative bias temperature instability (NBTI) is an similar mechanism in PMOS, where holes are trapped in oxide.
- Refer: http://www.semiconfareast.com/hotcarriers.htm



Reliability : Latch-up



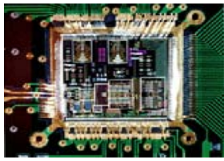


- NMOS and PMOS are formed as needed.
- In addition an NPN and an PNP transistor formed
- NPN transistor is formed between n-diffusion of NMOS, p-type substrate and n-well.
- Substrate and well provide resistance

Origin and model of CMOS latchup

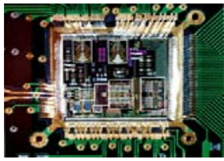
Reliability : Latch-up

- When parasitic BJT formed by the substrate, well, and diffusion turn ON, then latch-up occurs.
- This can lead to a low-resistance path between supply and ground.
- With proper process advances and layout consideration this can be avoided
  - – R_{sub} and R_{well} need to be minimized (guard rings)
- SOI processes avoid latch-up as there is no parasitic BJT.
- Processes with low voltages are less susceptible to latch-up (<0.7V complete immune).



Reliability : Over-voltage Failure

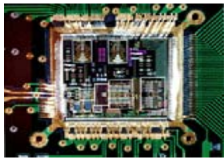
- Over-voltage problem due to :
  - Electrostatic discharge
  - Oxide breakdown
  - Punchthrough
  - Time dependent dielectric breakdown (TDDB) of gate oxide
- Electrostatic discharge (ESD) : static electricity entering the IO pad can cause transience
- Punchthrough: Higher voltages applied between source and drain lead to punchthrough when the source/drain depletion regions touch.



Reliability : Soft Errors

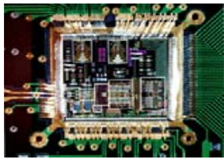
- DRAM occasionally flip value spontaneously. A soft error will not damage a system's hardware; the only damage is to the data that is being processed.
- There are two types of soft errors:
  - Chip-level soft error: Occurs when the radioactive atoms in the chip's material decay and release alpha particles into the chip. The particle can hit a DRAM cell and change it state to a different value.
  - System-level soft error: Occurs when the data being processed is hit with a noise phenomenon, typically when the data is on a data bus. The computer tries to interpret the noise as a data bit, which can cause errors in addressing or processing program code. The bad data bit can even be saved in memory and cause problems at a later time.
- When the corrupt bit is rewritten it is equal likely to experience anther error.

Source: http://www.webopedia.com/TERM/S/soft_error.html



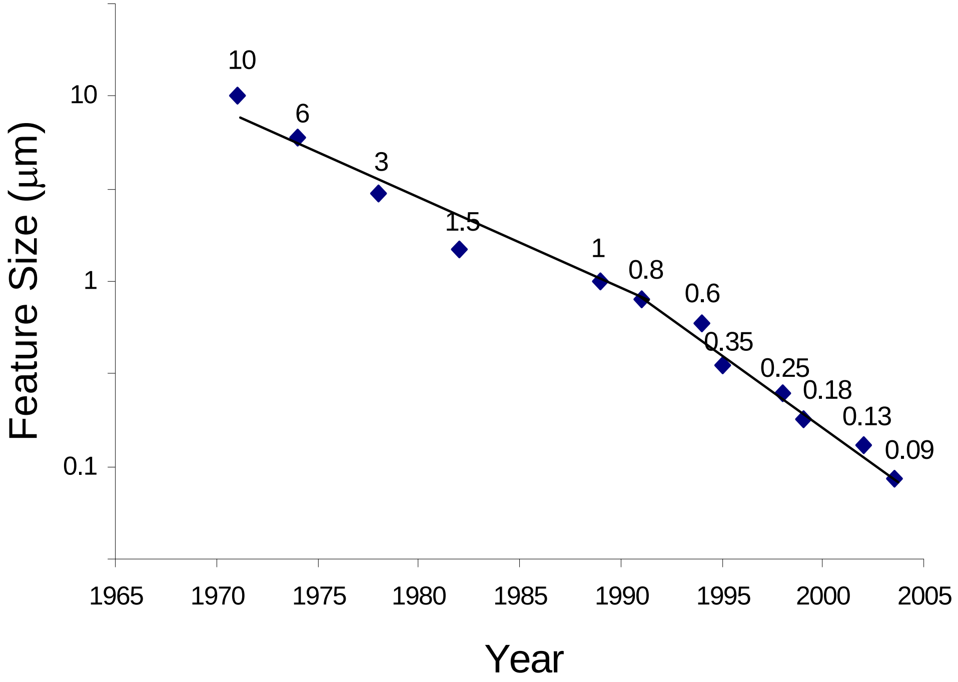
Why?

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



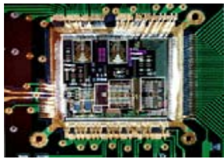
Scaling : Trend

- 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

- 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)



Scaling : Influence on MOS device

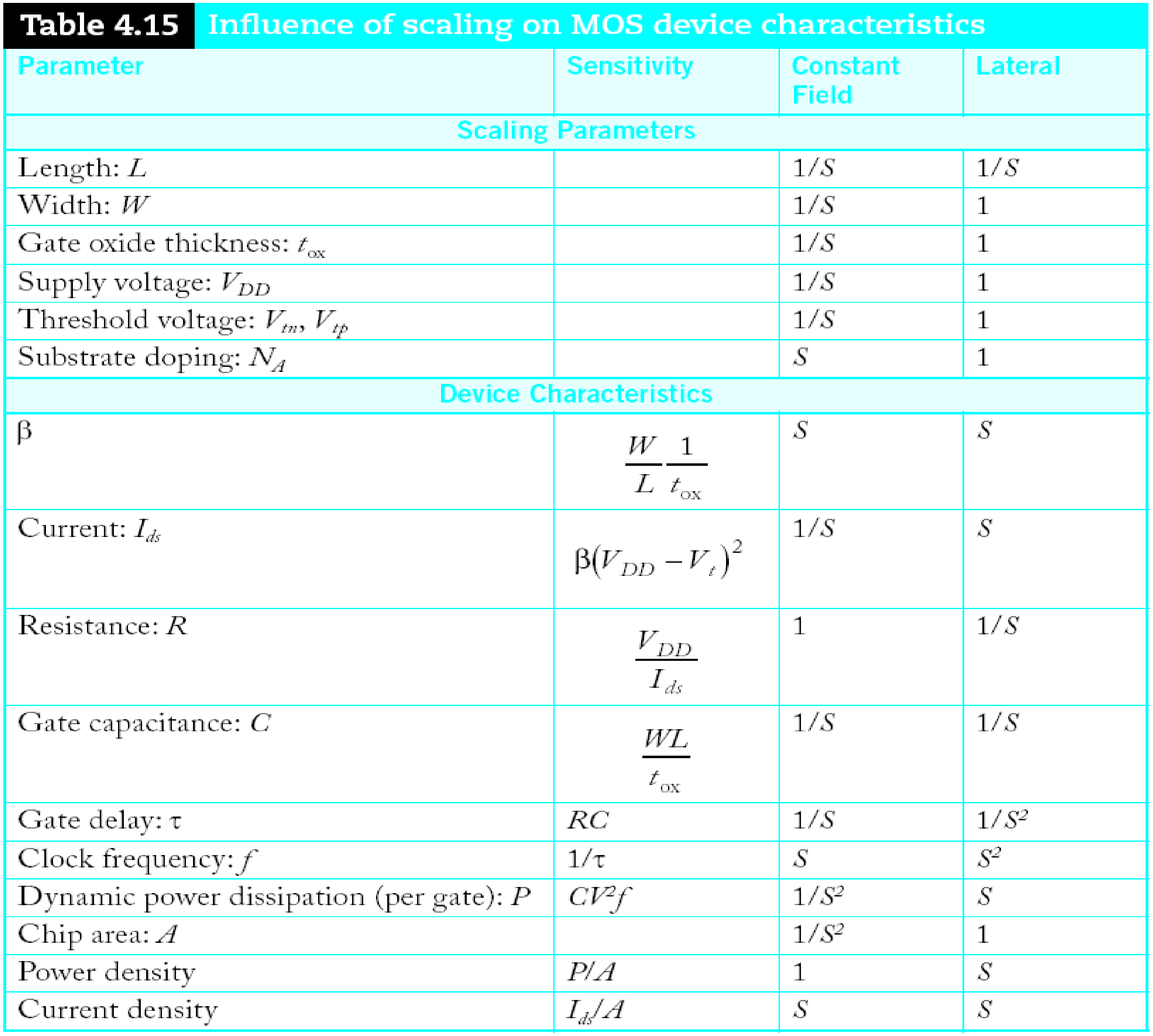
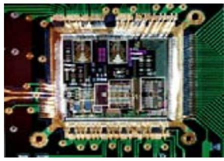


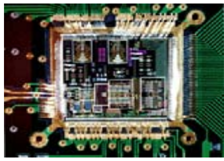
Table 4.15 Influence of scaling on MOS device characteristics

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



Scaling : Observations

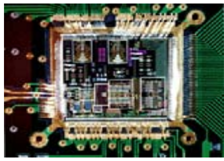
- Gate capacitance per micron is nearly independent of process
- But ON resistance * micron improves with process
- Gates get faster with scaling (good)
- Dynamic power goes down with scaling (good)
- Current density goes up with scaling (bad)
- Velocity saturation makes lateral scaling unsustainable



Scaling : Example

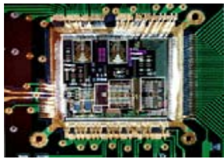
- Gate capacitance is typically about 2 fF/μm
- The FO4 inverter delay in the TT corner for a process of feature size f (in nm) is about $0.5f$ ps
- Estimate the ON resistance of a unit ($4/2\ \lambda$) transistor.

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



Interconnect 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 : Influence

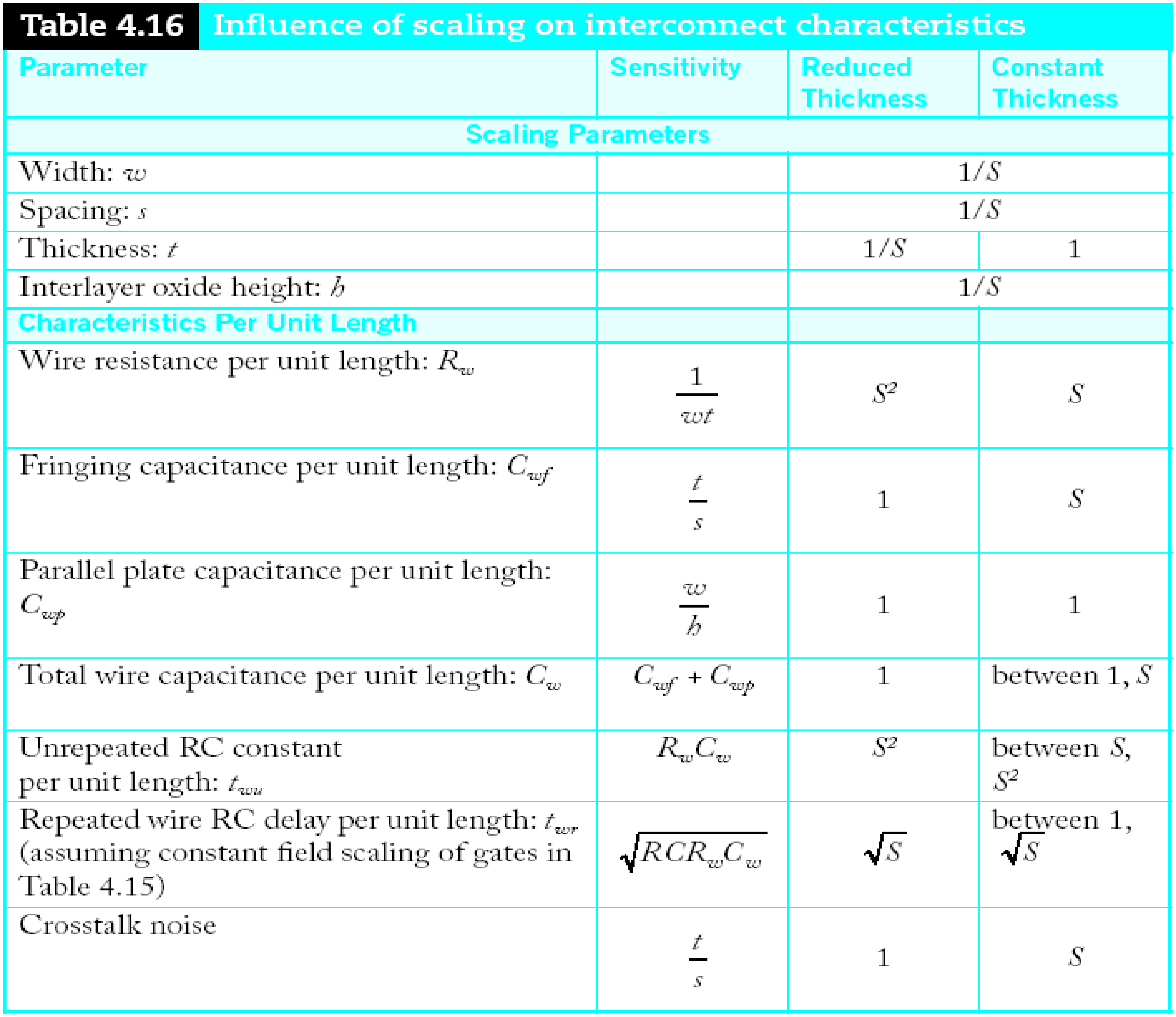
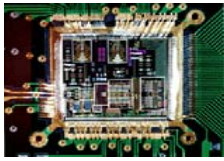


Table 4.16 Influence of scaling on interconnect characteristics

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_wC_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)	$\sqrt{RCR_wC_w}$	$\sqrt{S}$	between 1, $\sqrt{S}$
Crosstalk noise	$\frac{t}{s}$	1	S



UNIVERSITY OF NORTH TEXAS
Discover the power of ideas

Interconnect Scaling : Influence

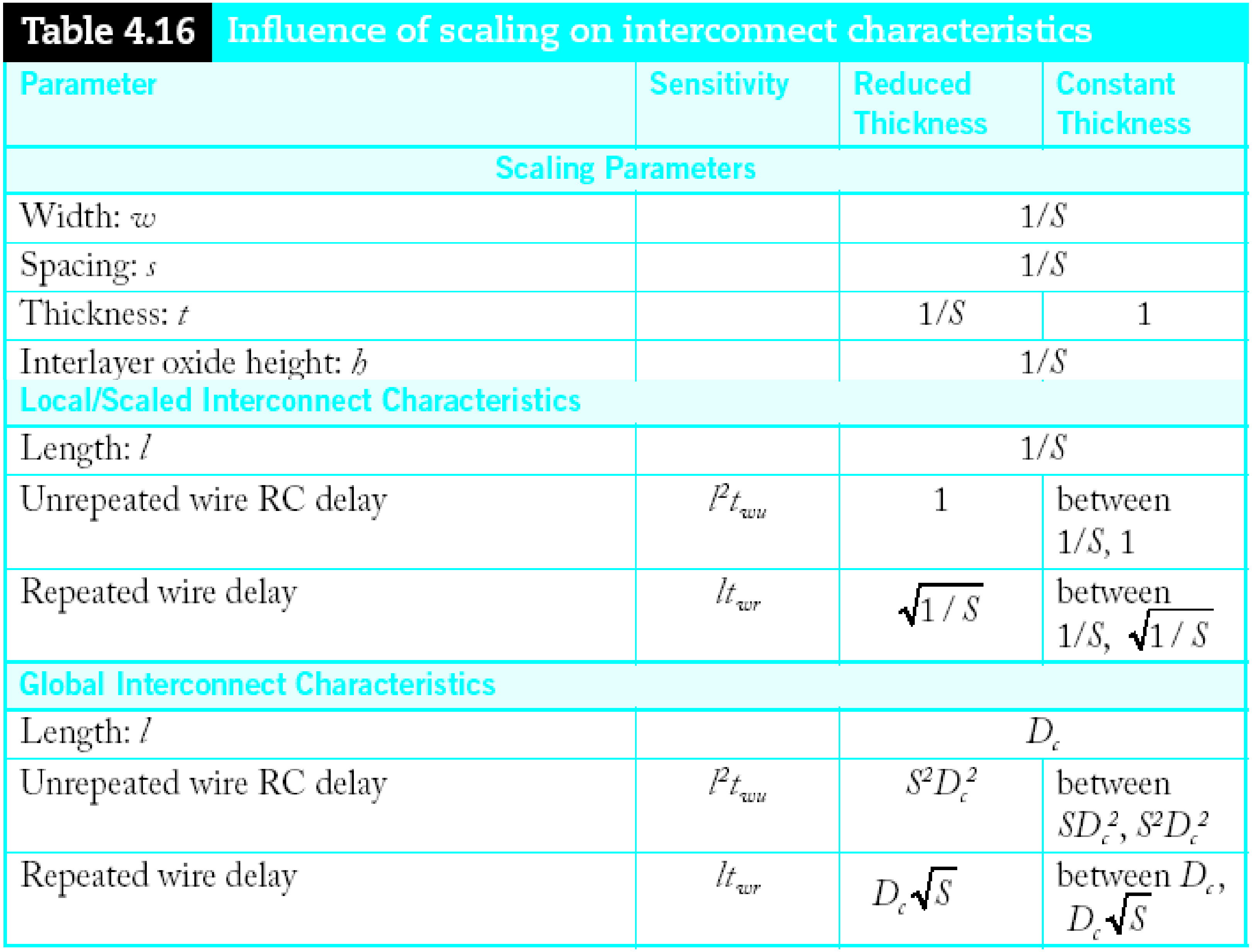
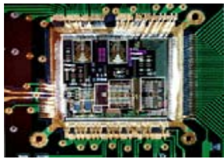


Table 4.16 Influence of scaling on interconnect characteristics

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, $D_c \sqrt{S}$

Interconnect Scaling : 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



International Technology Roadmap for Semiconductors (ITRS)

- Semiconductor Industry Association forecast

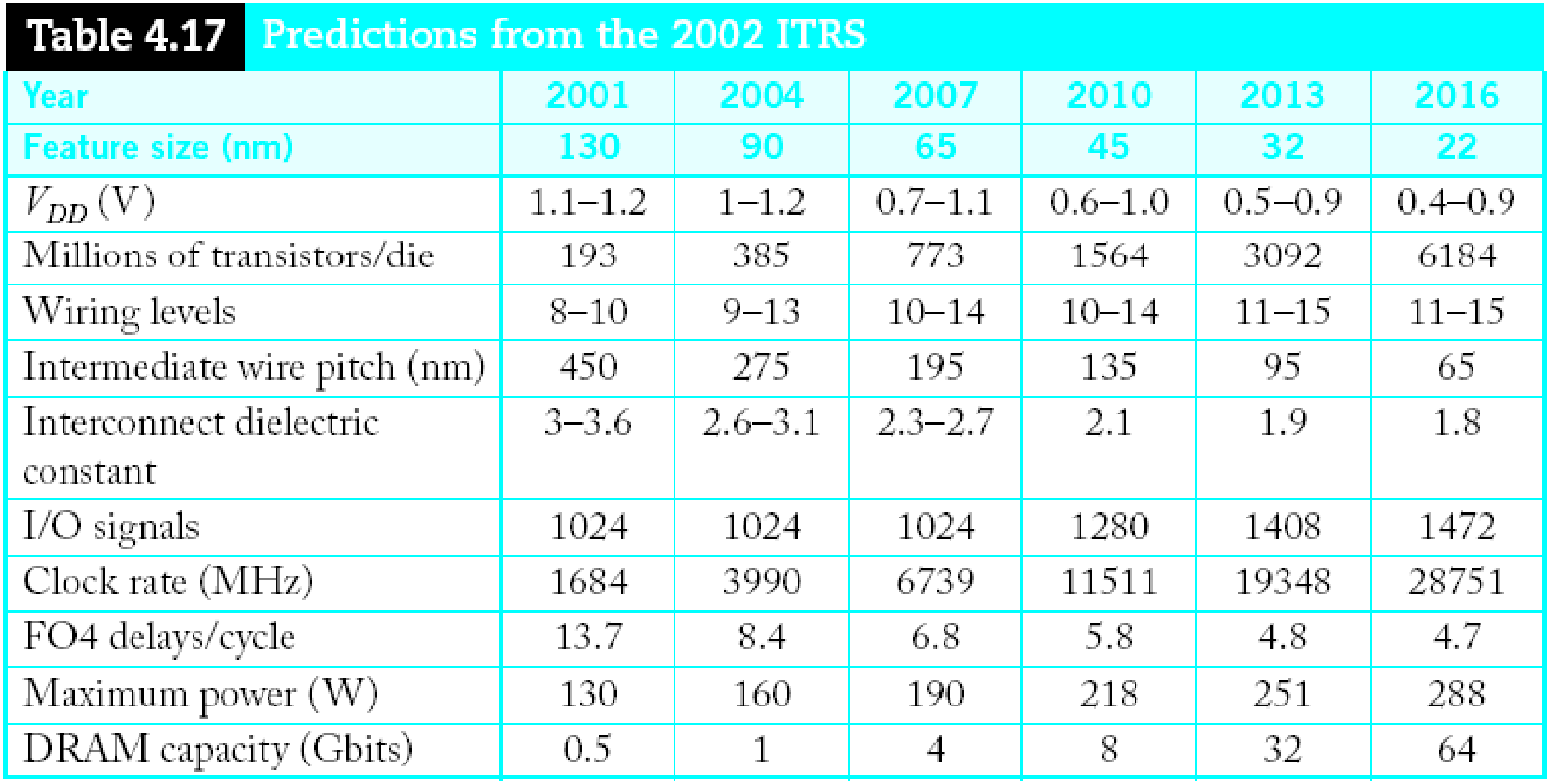
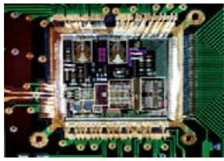


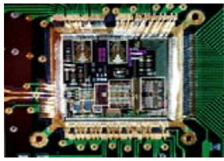
Table 4.17 Predictions from the 2002 ITRS

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



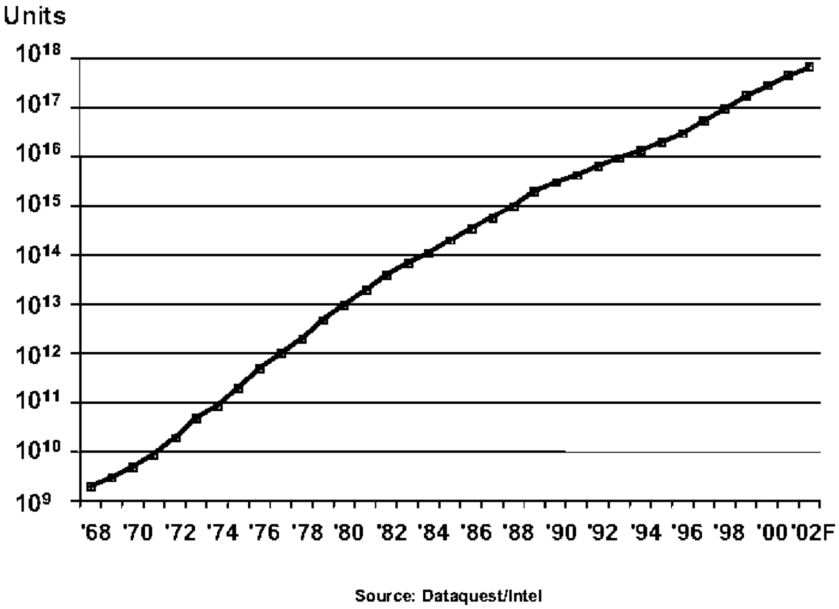
Scaling Implications

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

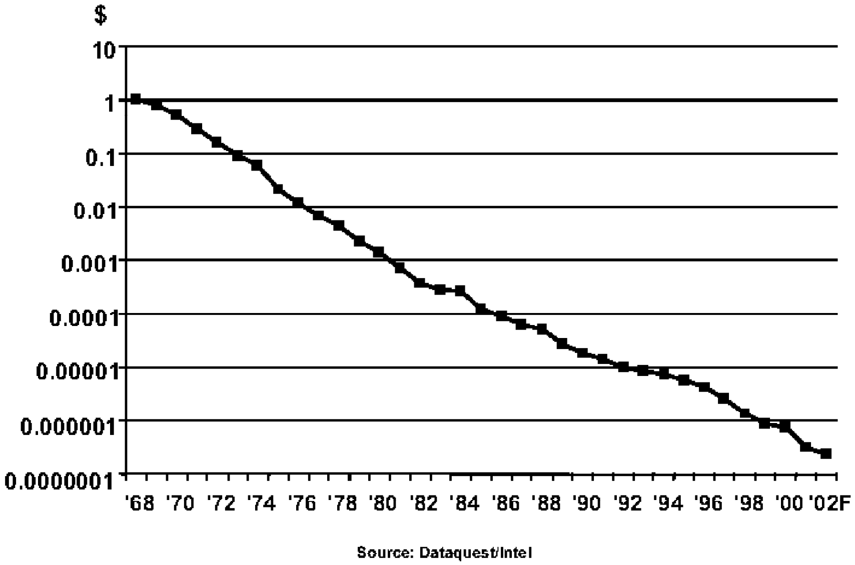


Scaling Implications : Cost Improvement

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



Source: Dataquest/Intel

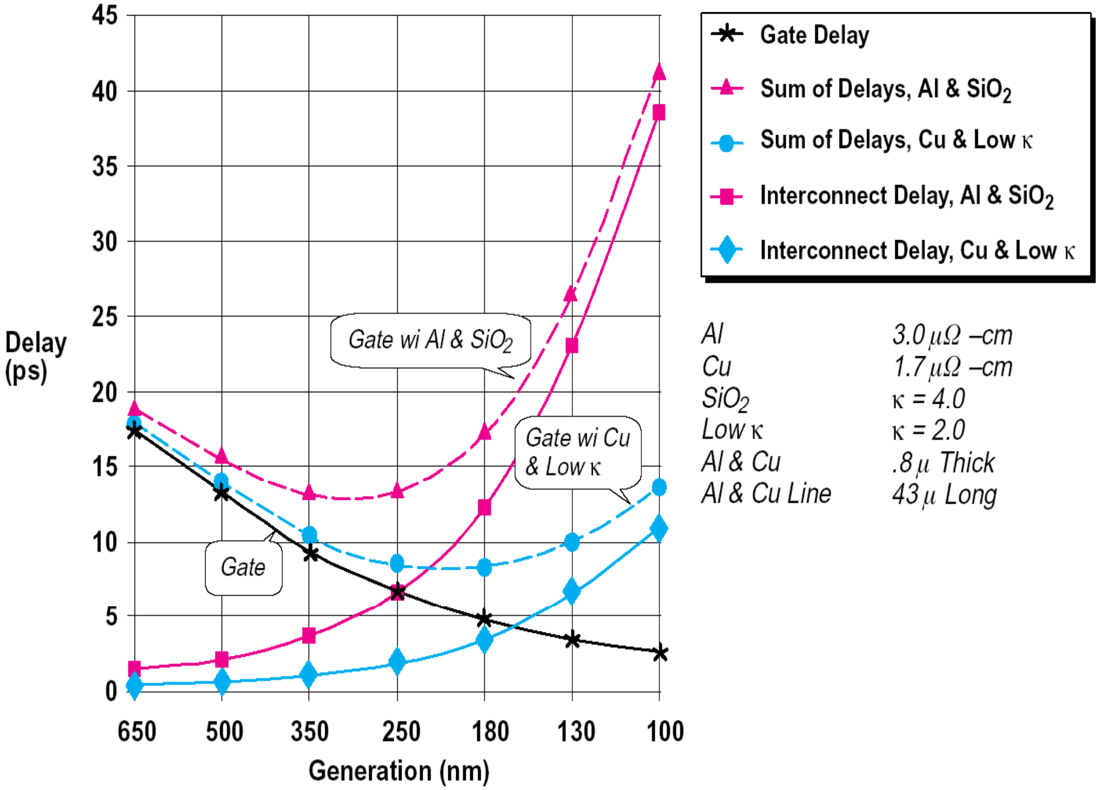


Source: Dataquest/Intel

[Moore03]

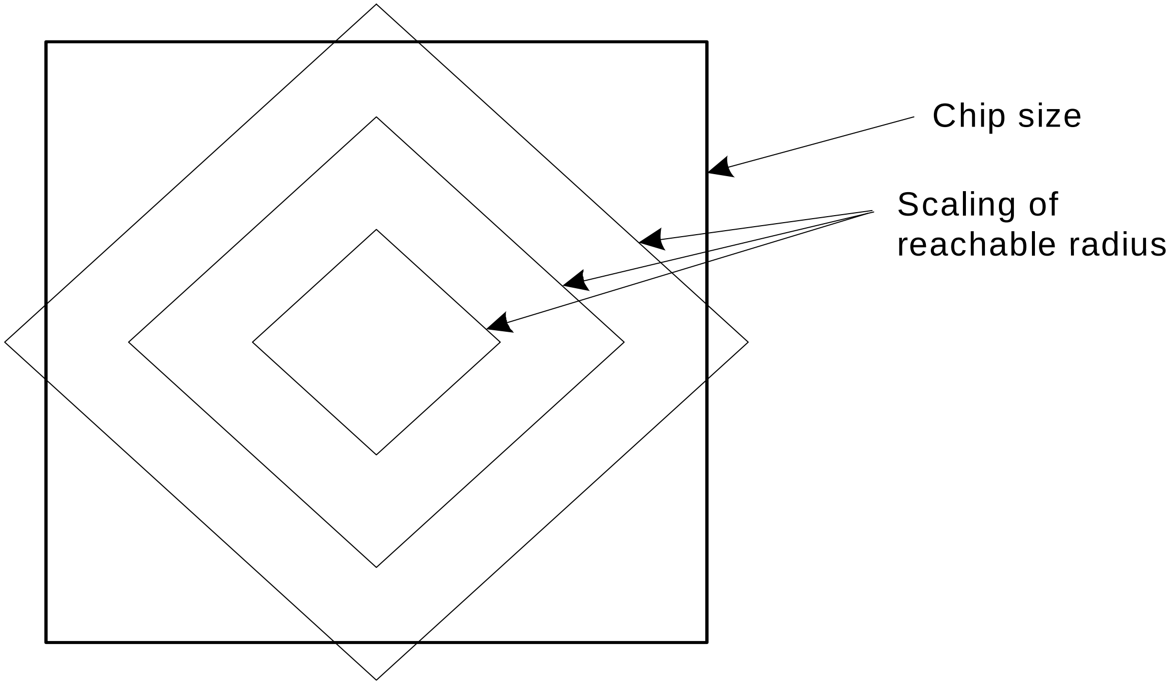
Scaling Implications : Interconnect Woes

- 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



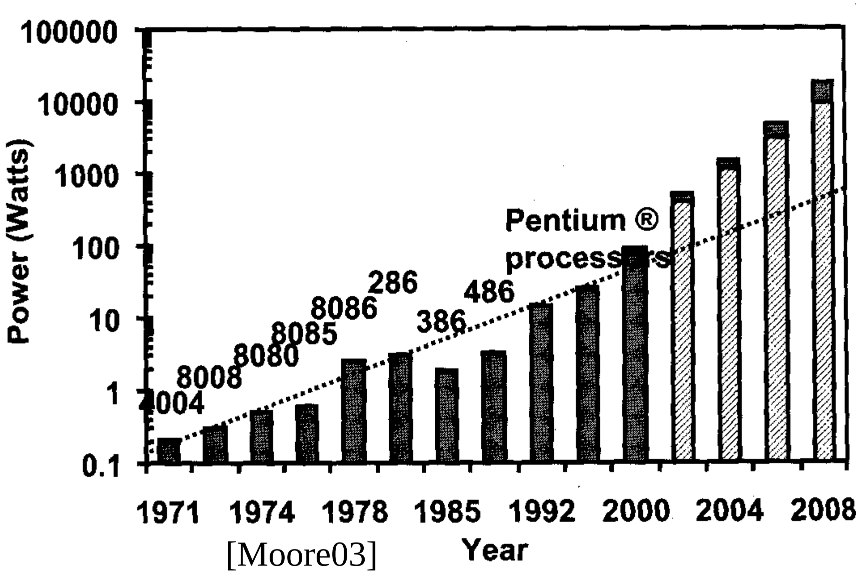
Scaling Implications : Reachable Radius

- 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

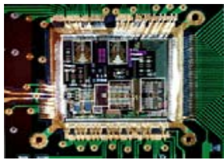


Scaling Implications : Dynamic Power

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

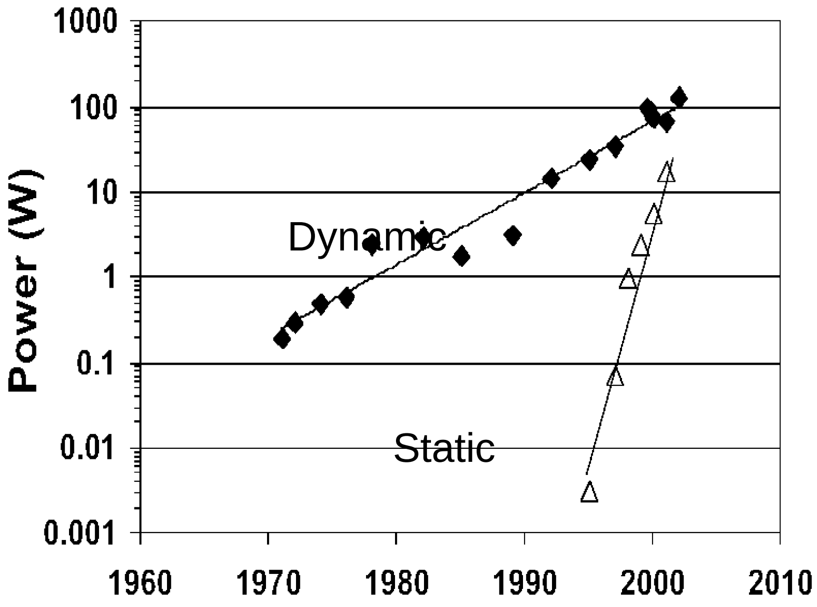


[Moore03]



Scaling Implications : Static Power

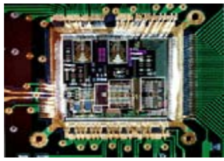
- V_{DD} decreases
  - Save dynamic power
  - Protect thin gate oxides and short channels
  - No point in high value because of velocity sat.
- V_{th} must decrease to maintain device performance
- But this causes exponential increase in OFF leakage
- Major future challenge



[Moore03]

Scaling Implications : 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



Scaling Implications : 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