Physical Design of Digital Integrated Circuits (EN0291 S40)

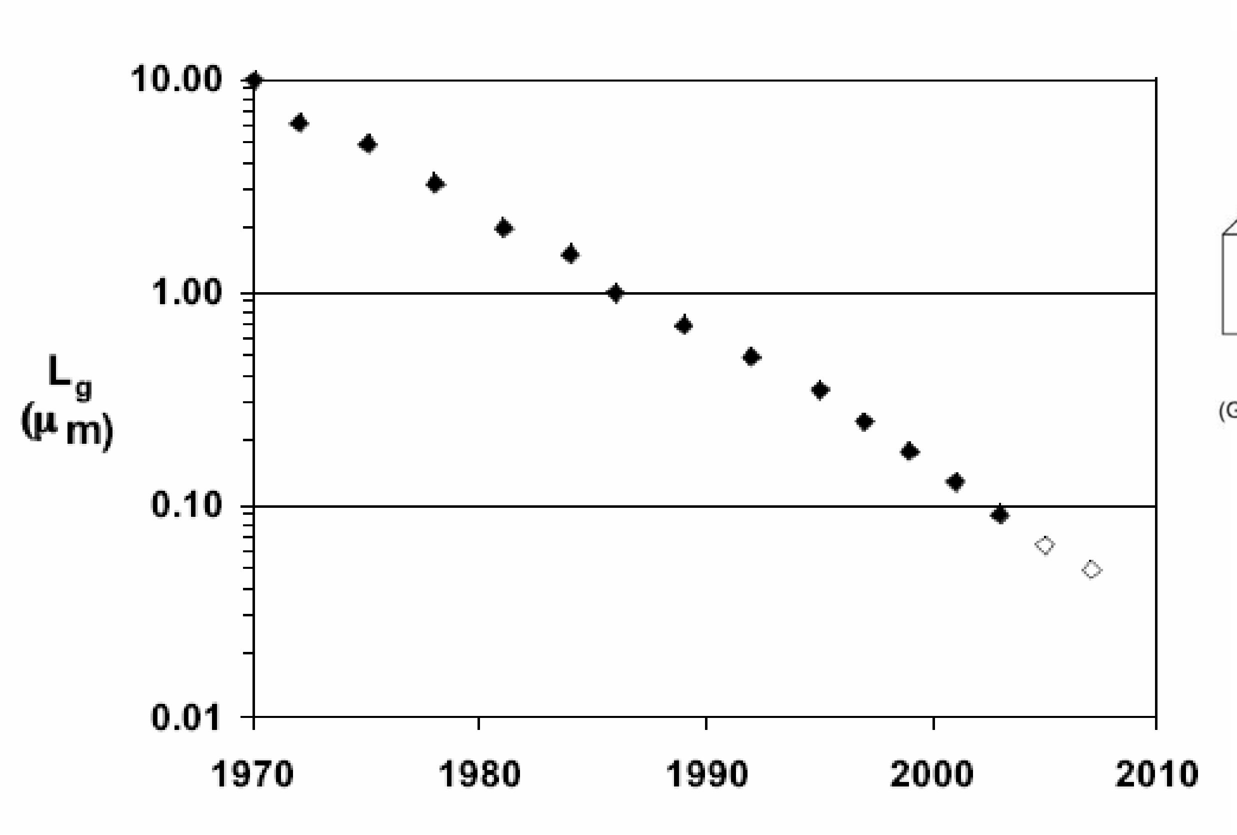
Sherief Reda
Division of Engineering, Brown University

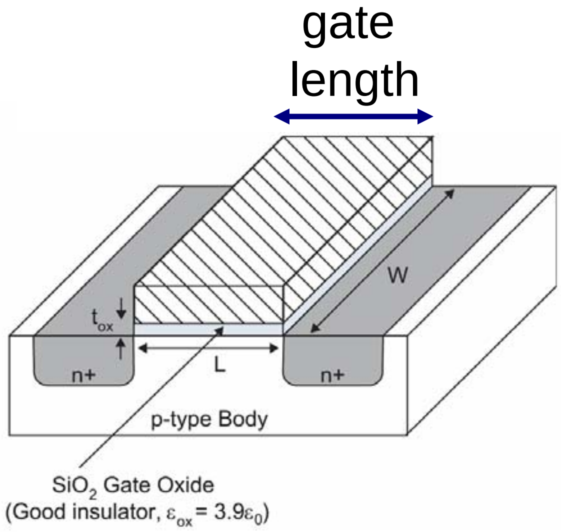
Fall 2006

Lecture 02: CMOS scaling theory

- Device scaling
- Interconnect scaling
- More implications for design and architecture
- Readings and project assignments

Minimum feature size (gate length)

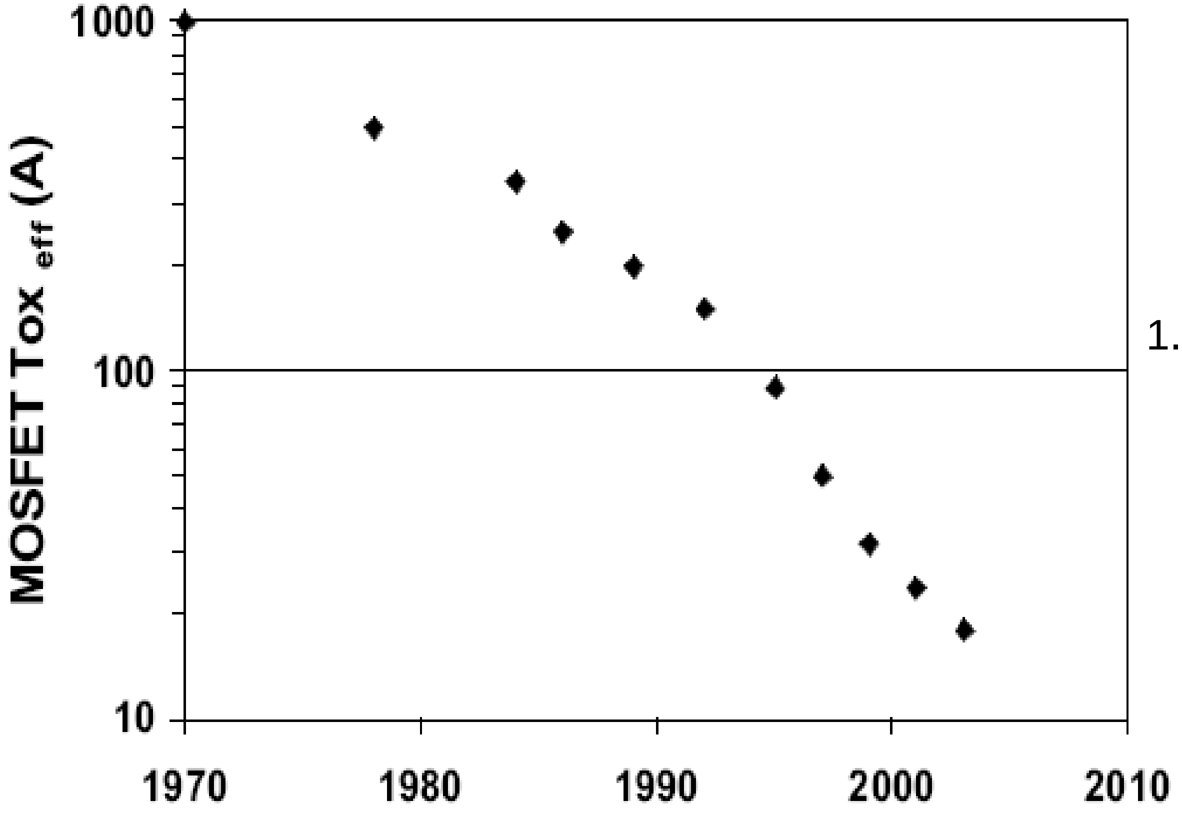


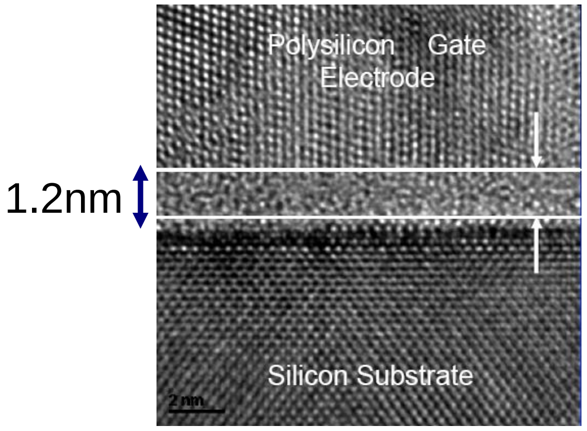


- Human hair 100um
- Amoeba 15um
- Red bood cell 7um
- HIV virus 0.1um
- Buckyball 0.001um

BROWN

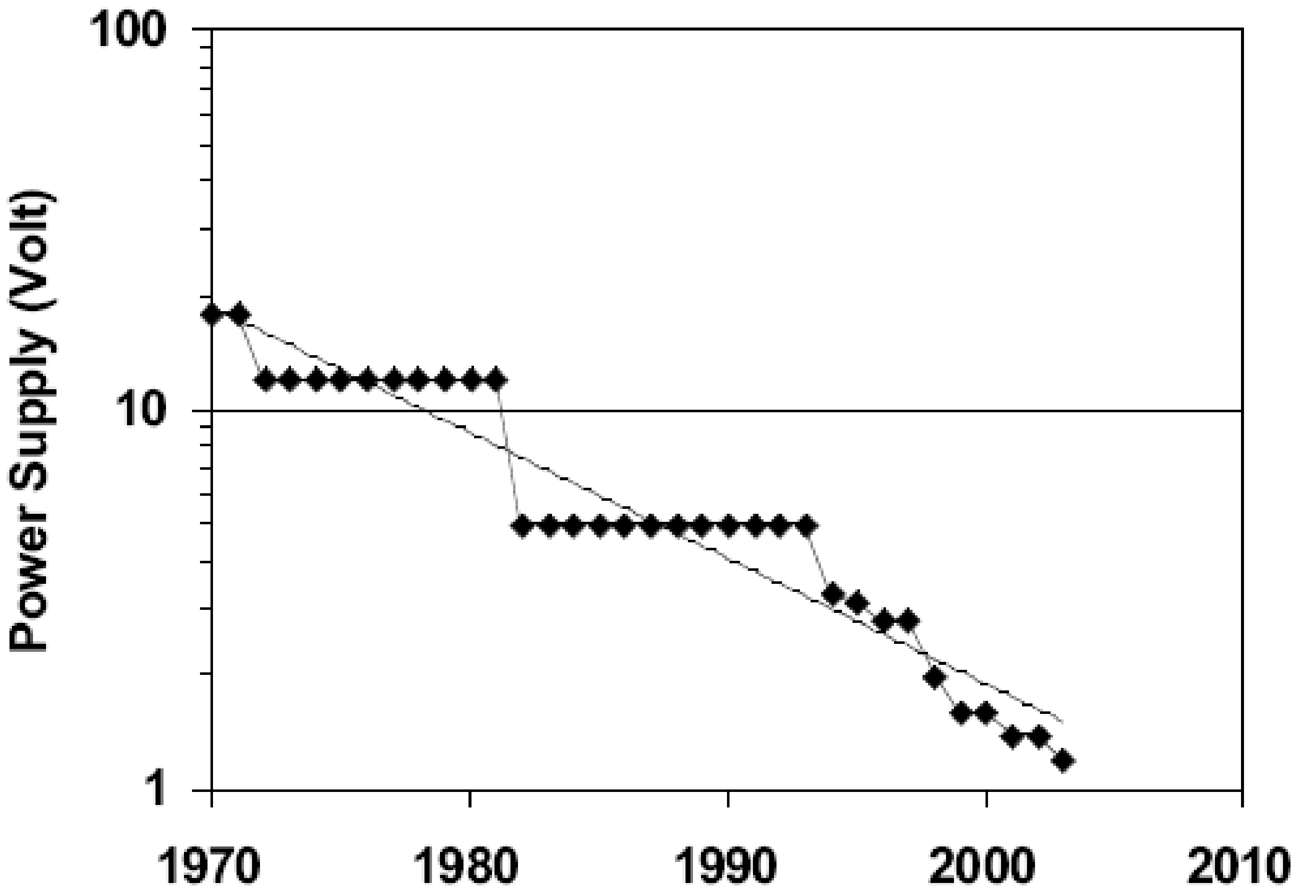
MOSFET dielectric effective thickness



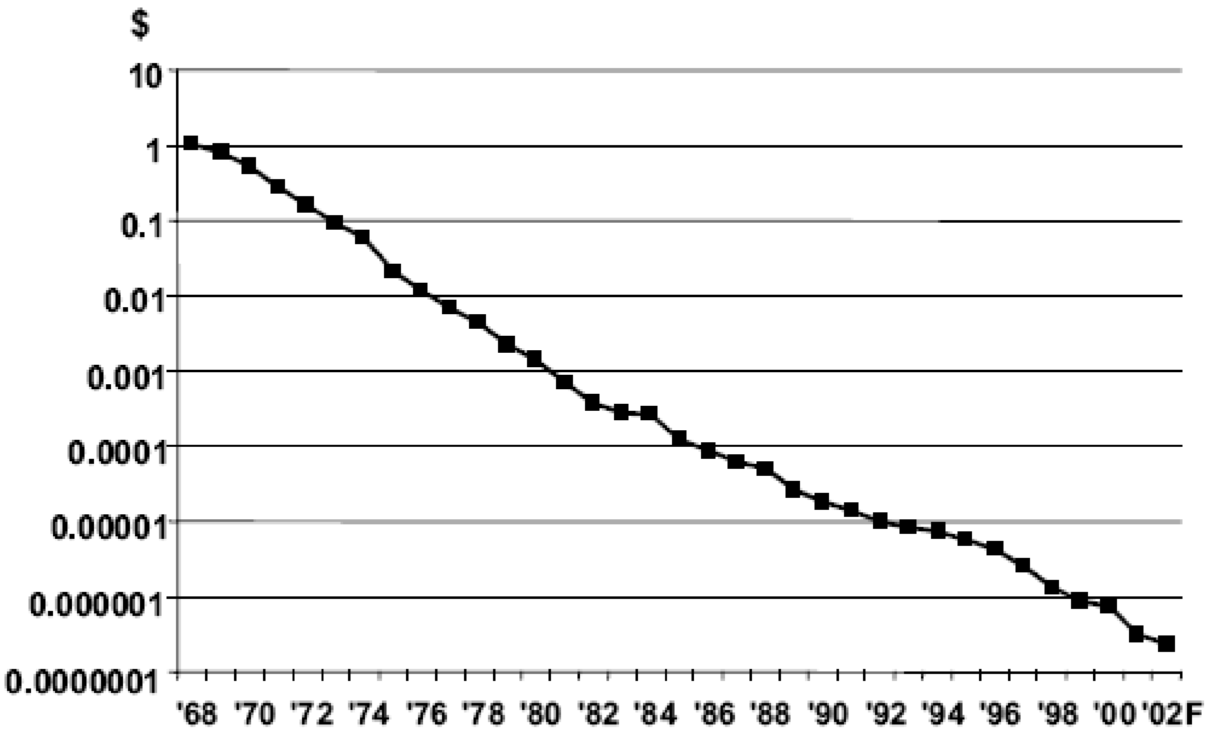


- Current oxide thickness ~ 1.0 – 2.0nm thickness → 3 – 4 atomic layers of oxide

Processor supply voltages

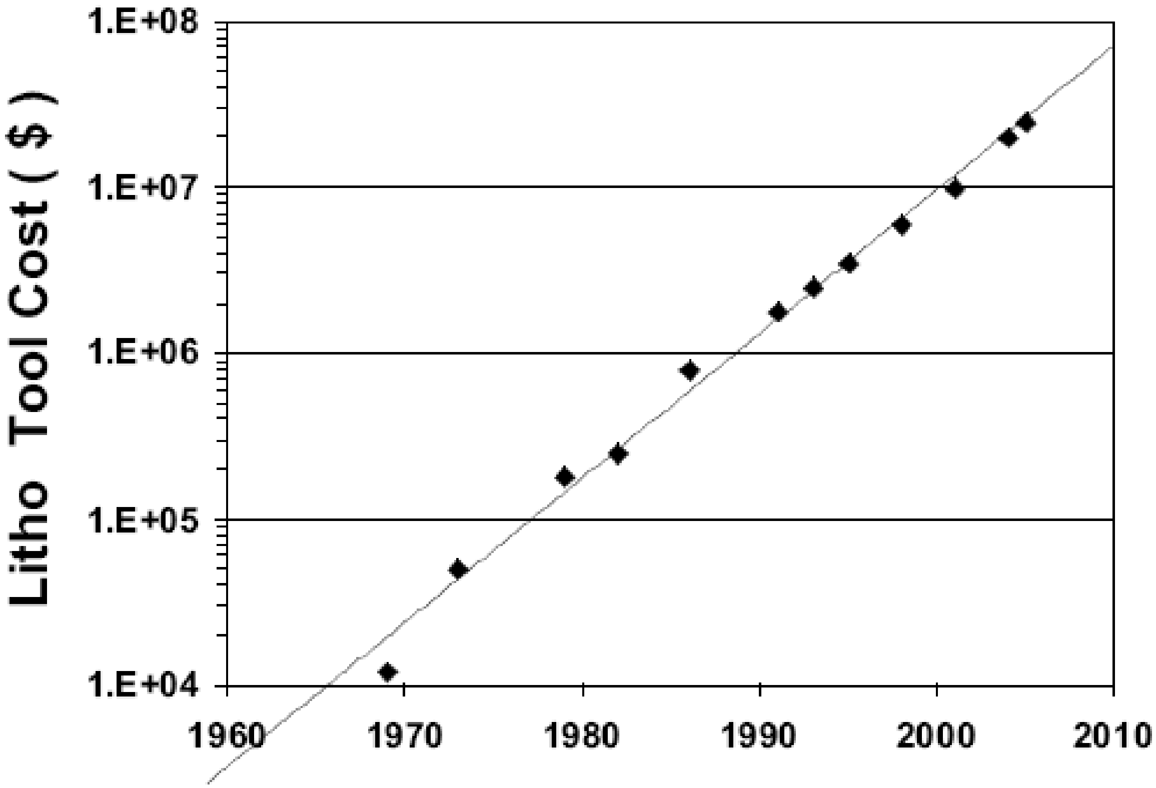


Average transistor price per year



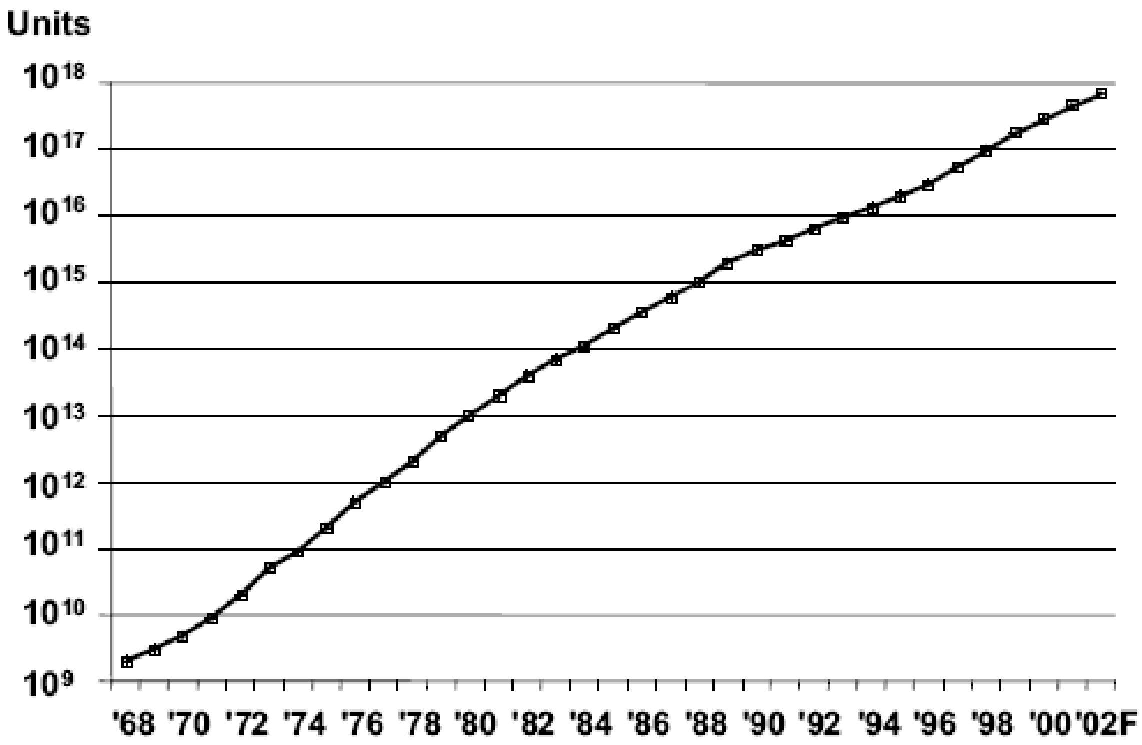
- Average price of a transistor is 0.1 micro cent!

Lithography tool cost



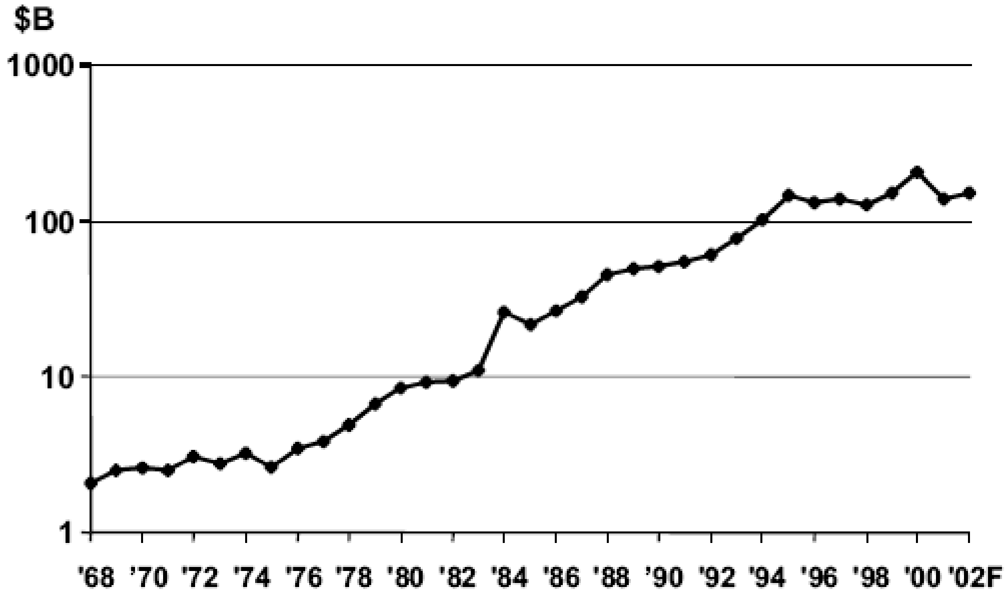
- Fewer and fewer companies can afford to have their own foundries ⇒ many turn fabless and outsource their manufacturing

Transistors shipped per year



- In 2003, Moore estimates that the number of transistors shipped is 10 quintillion – about 100 times the number of ants estimated to be stalking our planet

Worldwide semiconductor revenues

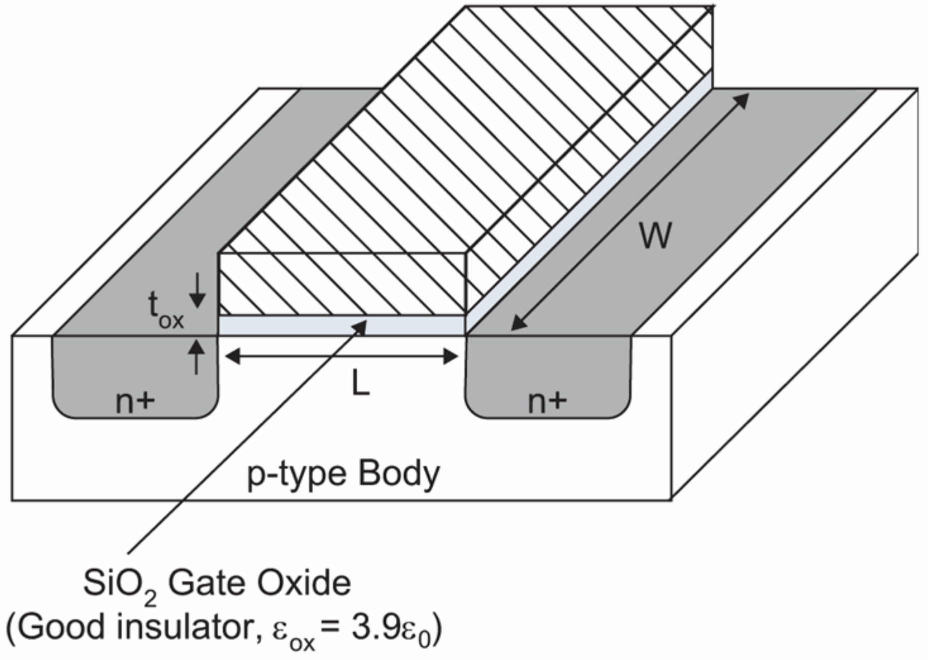


- Worldwide annual chip sales: ~$220B (opto: $17B – processors 30B – DRAM 24B – flash $16B); EDA (CAD tools) sales: ~$4B (2% of total)
- Computer/video games (software): worldwide ($19B) US ($7B)
- Box office: worldwide ~$23B US ($9B); tickets + dvds + tv rights: ~ $44B
- Worldwide total software sales $383B
- Worldwide pharmaceutical sales: $550B
- Just Exxon + Chevron 2005 total sales: ~$371B + $193B = 564B

BROWN

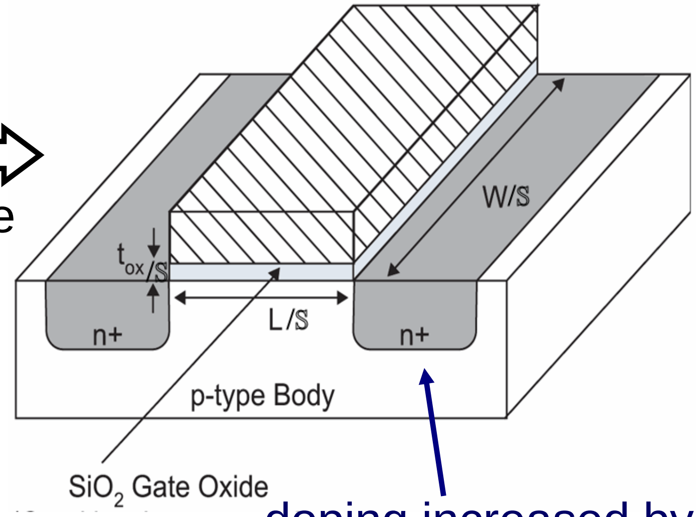
Device scaling

(very idealistic NMOS transistor)



scale

(scaled down by *L*)



doping increased by a factor of S

Increasing the channel doping density increases the channel barrier
⇒ improves isolation between source and drain during OFF status
⇒ permits distance between the source and drain regions to be scaled

Depletion width $$W_D = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} + V_{dd})}{qN_a}}$$ ⇒ scales down by *S*

Implications of ideal device scaling

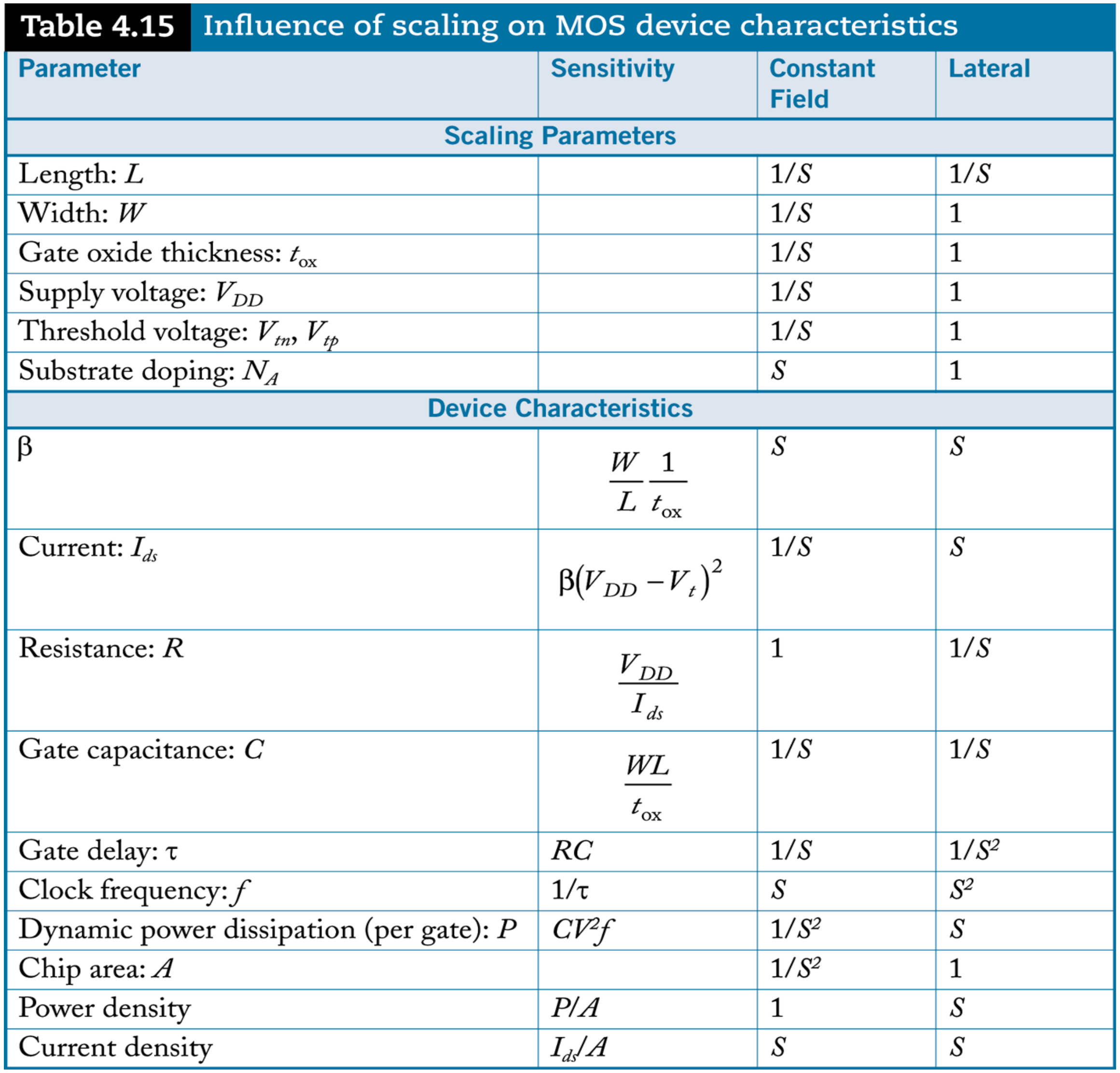
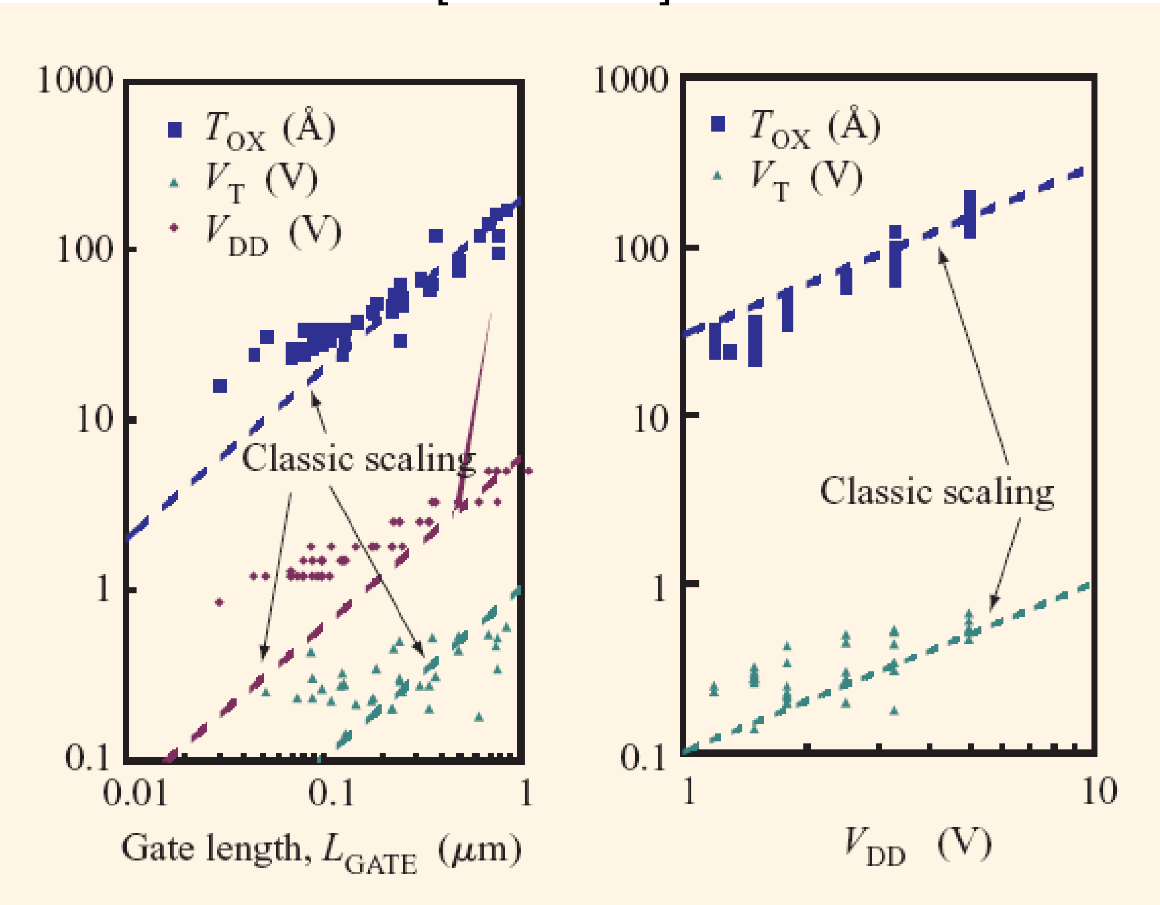


Table 4.15 Influence of scaling on MOS device characteristics

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	$\frac{W}{L}\frac{1}{t_{\mathrm{ox}}}$	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	$\frac{WL}{t_{\mathrm{ox}}}$	$1/S$	$1/S$
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

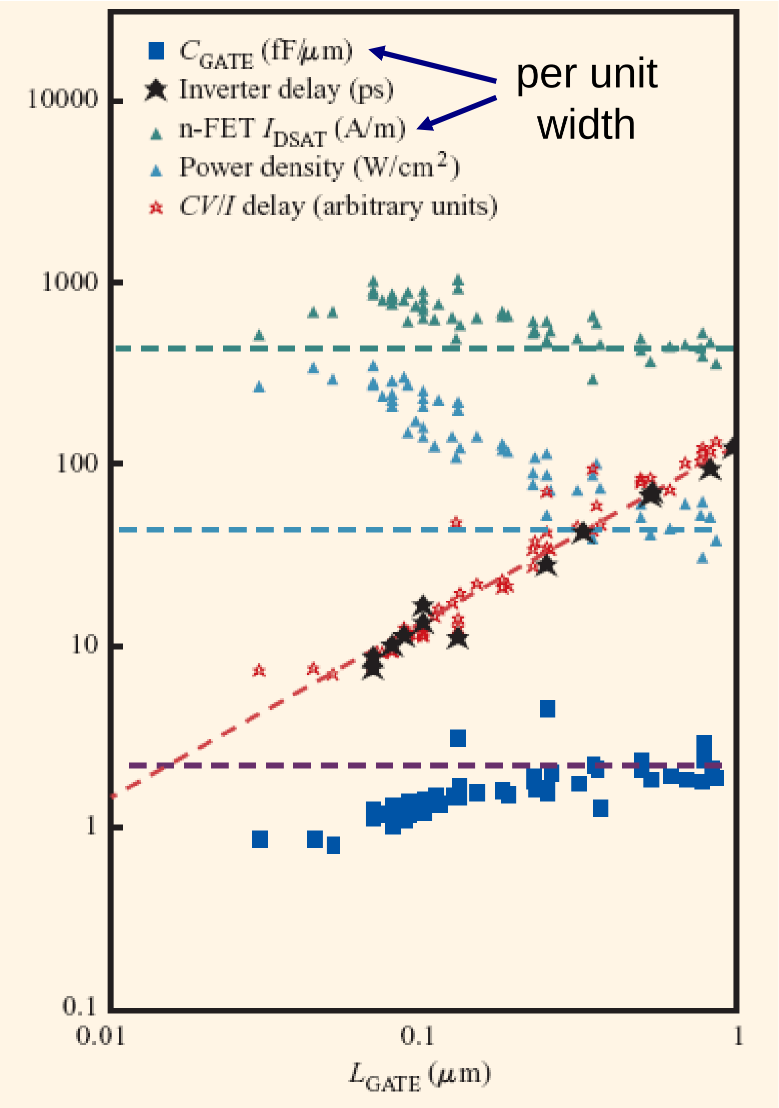
Actual scaling trends deviate from ideal

[Nowak'02]



- ➢ V_t and V_{dd} are deviating from classic scaling with respect to L_{gate}
- ➢ V_t and V_{dd} are rather following scaling trends of V_{dd}

Electrical consequences of actual scaling



Empirically, delay = C_{gate} V_{dd} / I_{SAT} follows ideal scaling

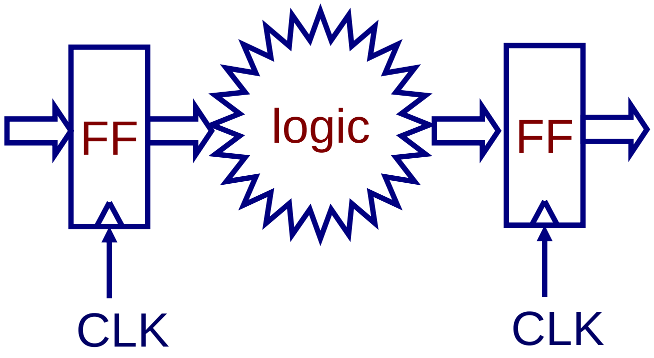
L_{gate} is dropping more rapidly than t_{ox} ⇒ more scaling down in C_{gate}

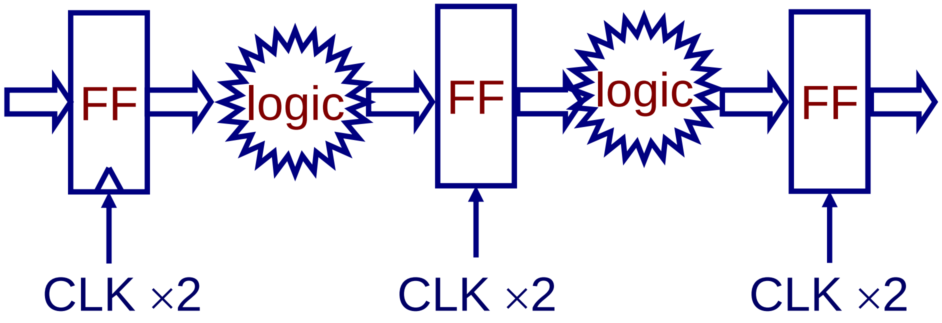
V_{dd} scales down by less than S
C_{gate} scales down by more than S
⇒ switching power (per unit area) is no longer constant ⇒ scales up

BROWN

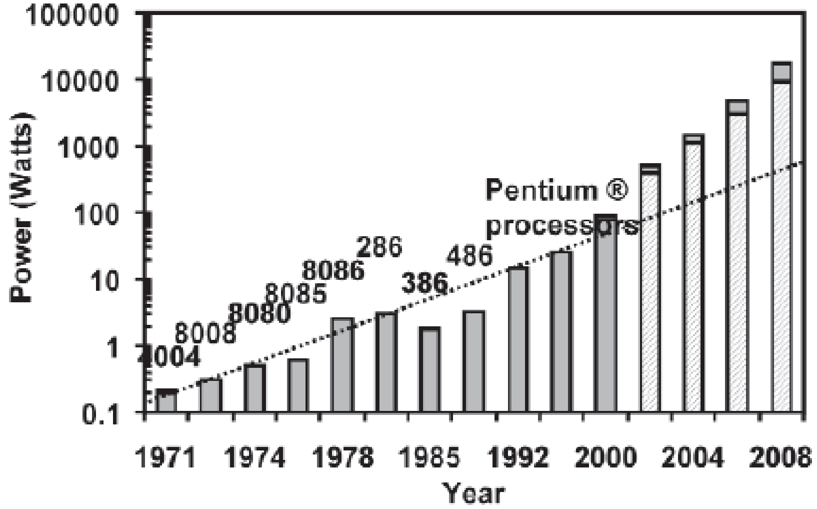
Dynamic power was further scaling up in microprocessors

Reason 1: Frequency was doubling (×2) rather than scaling by just 43% (by pursuing more pipeline stages; each stage has less logic)



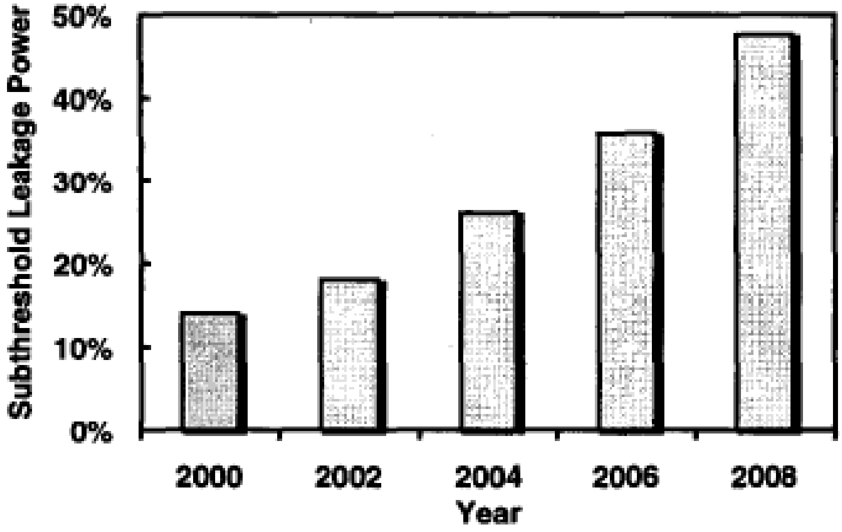


[Gelsinger'01]



Reason 2: Die sizes were also increasing in size (see slide 32)

Scaling of standby power



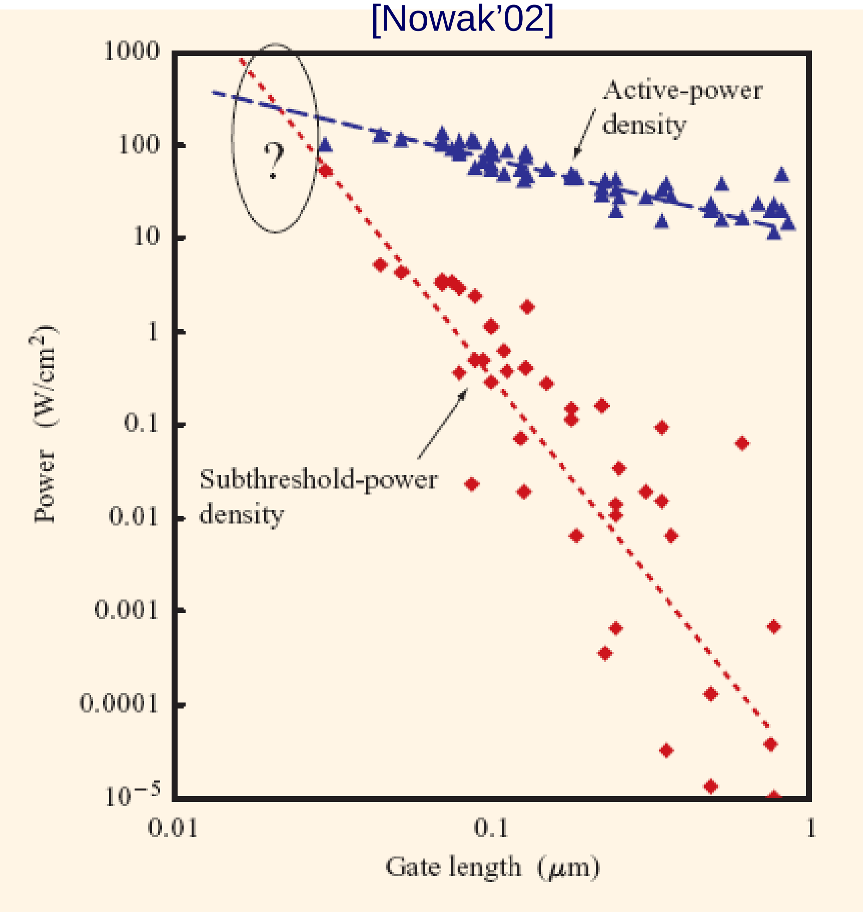
bottleneck

Standby power $P_{off} \propto \frac{1}{t_{ox}} e^{(-\frac{qV_t}{mkT})}$

- ➢ Even if V_t is kept constant after scaling, P_{off} scales up by S if t_{ox} is scaled down by S
- ➢ V_t must be scaled down if V_{DD} is scaled down (otherwise I_{SAT} is weaker and transistor is slow)
- ➢ Standby power would further increase by 10× for every 0.1V reduction of V_t

Scaling (V_{DD} & V_t) reduces dynamic power, but increases static power (per gate)

[Nowak'02]

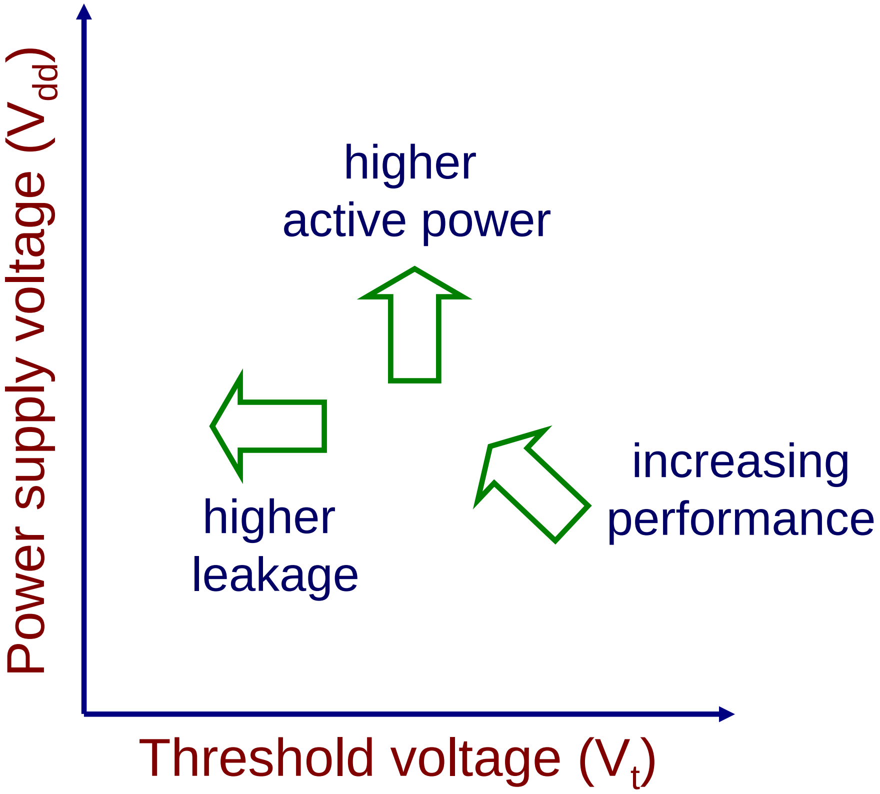


bottleneck

At 25°C, Extrapolations suggest that subthreshold power will equal dynamic power at L_{gate} = 20nm

BROWN

Power/performance tradeoffs



higher
active power

higher
leakage

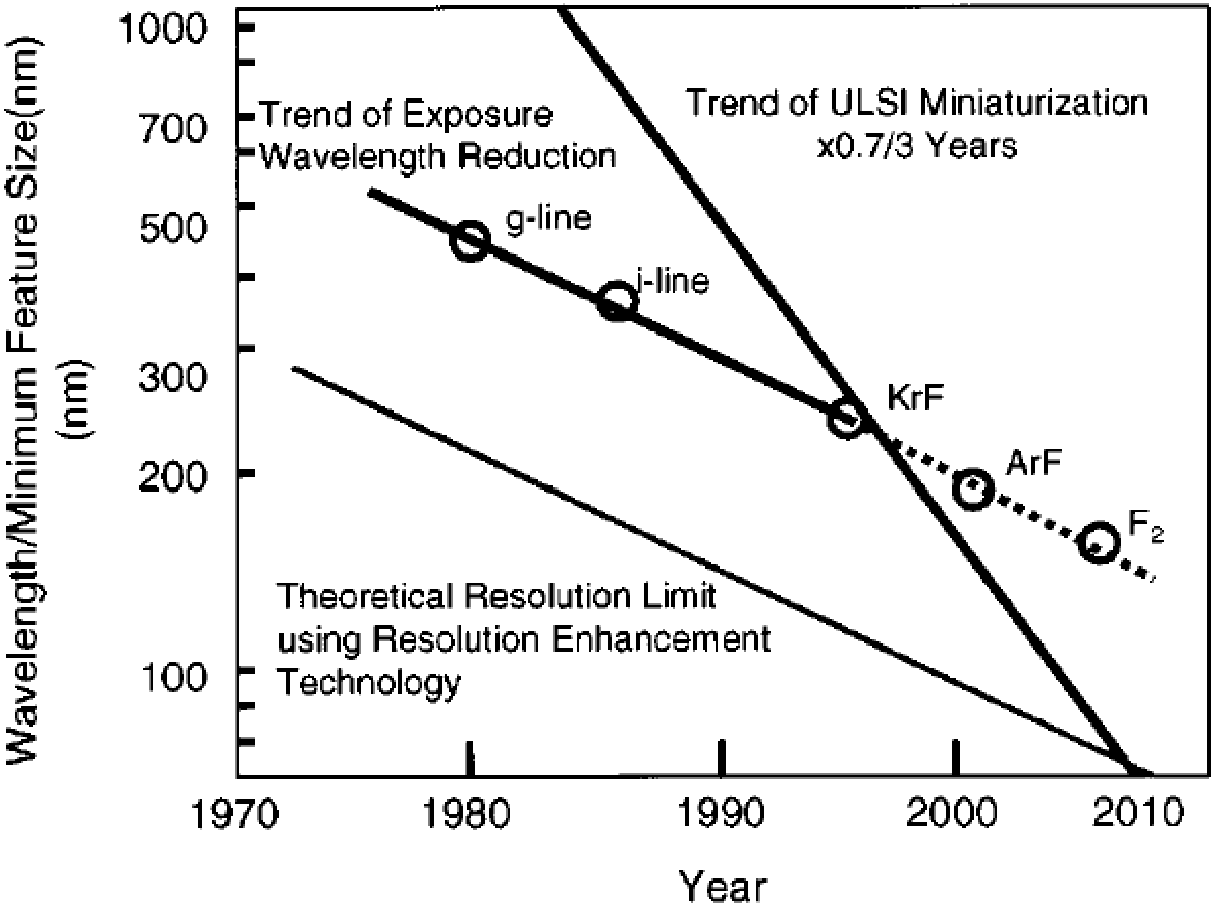
increasing
performance

Power supply voltage (V_{dd})

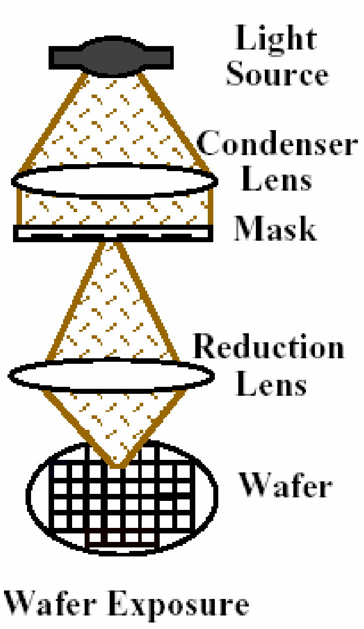
Threshold voltage (V_t)

[Taur, 01]

Scaling of lithographic light source wavelength



[source: Okazaki]

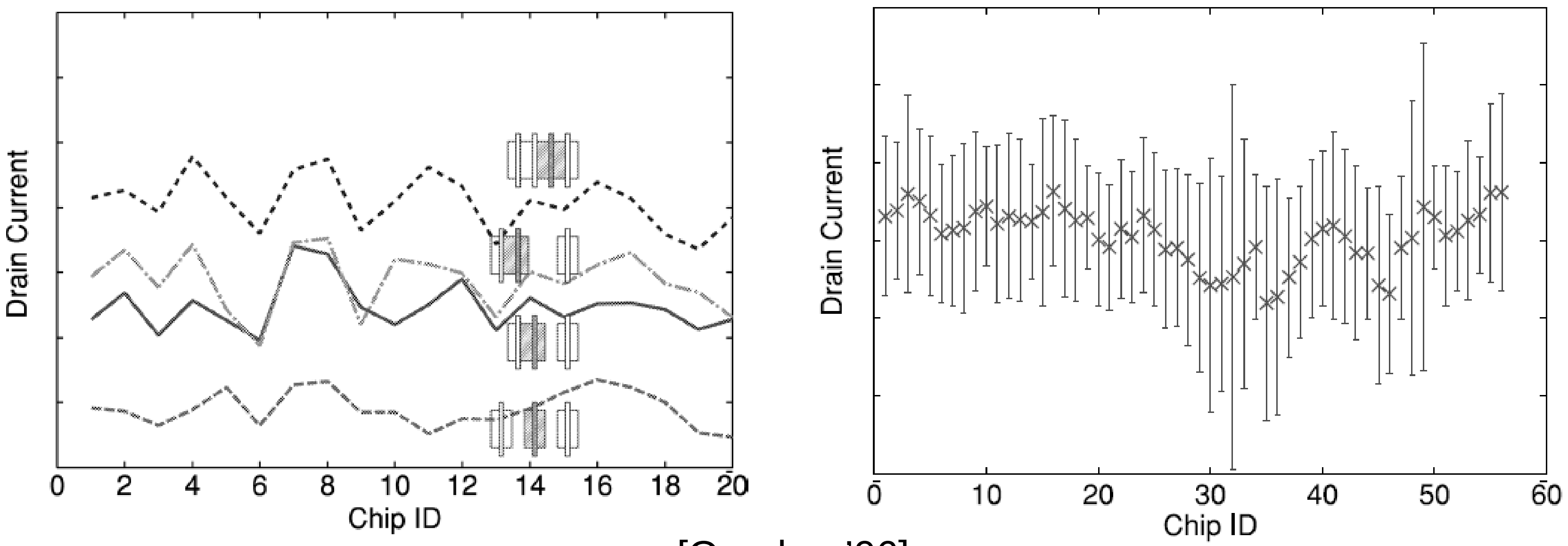


projection printing

BROWN

Scaling of variability

bottleneck



180nm

130nm

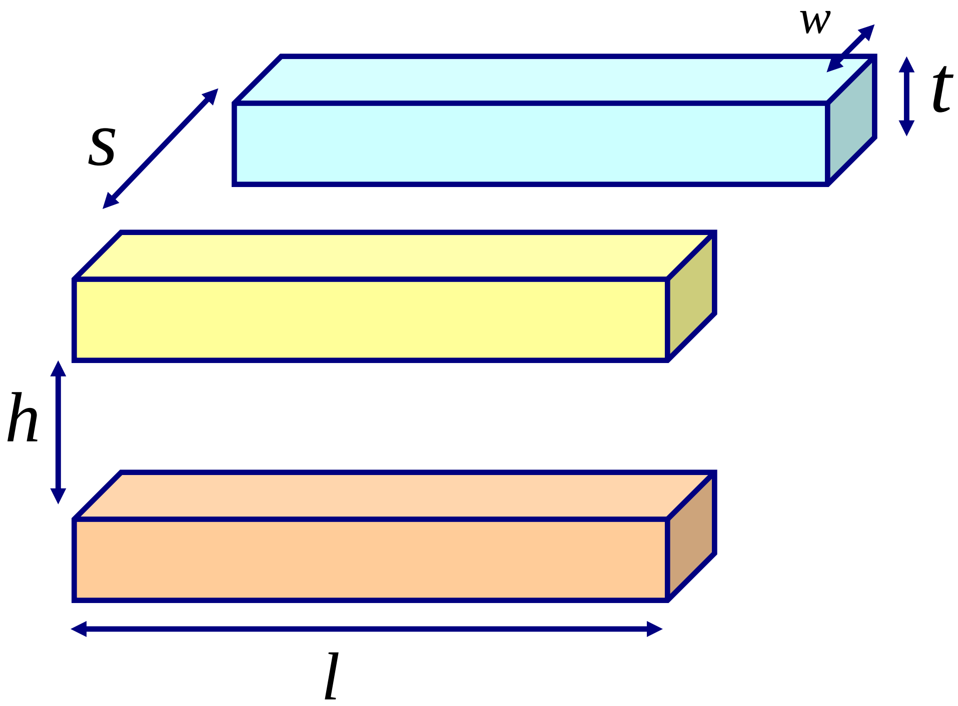
[Onodera’06]

- Die-to-die and within-die variations are getting worse but no concrete data is available

Lecture 02: CMOS scaling theory

- Device scaling
- Interconnect scaling
- More implications for design and architecture
- Readings and project assignments

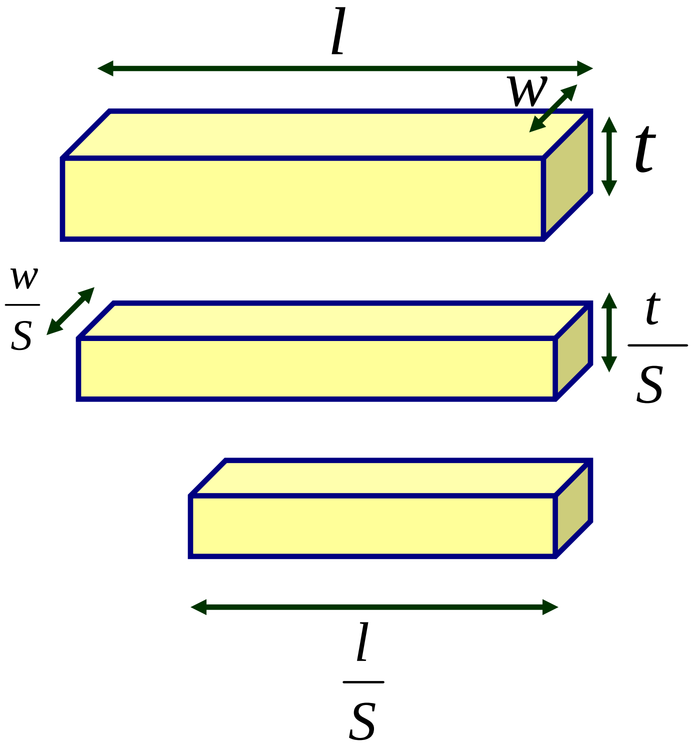
Interconnect scaling



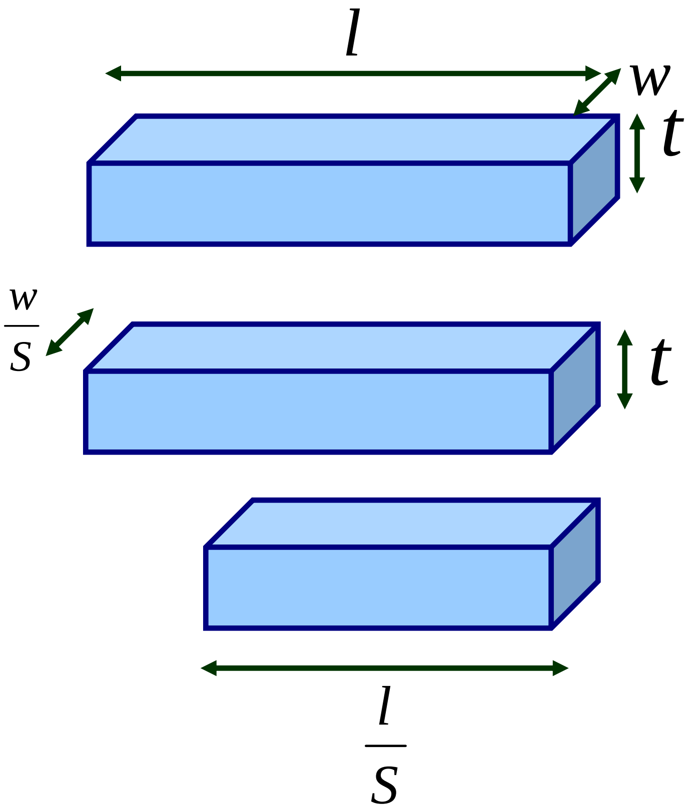
w: width of interconnect (layer dependant)
s: spacing between interconnects with same layer
h: dielectric thickness (spacing between interconnects in two vertically adjacent layers)
l: length of interconnect
t: thickness of interconnect

Constant thickness scaling versus reduced thickness scaling

reduced thickness scaling



constant thickness scaling



Implications of ideal interconnect scaling

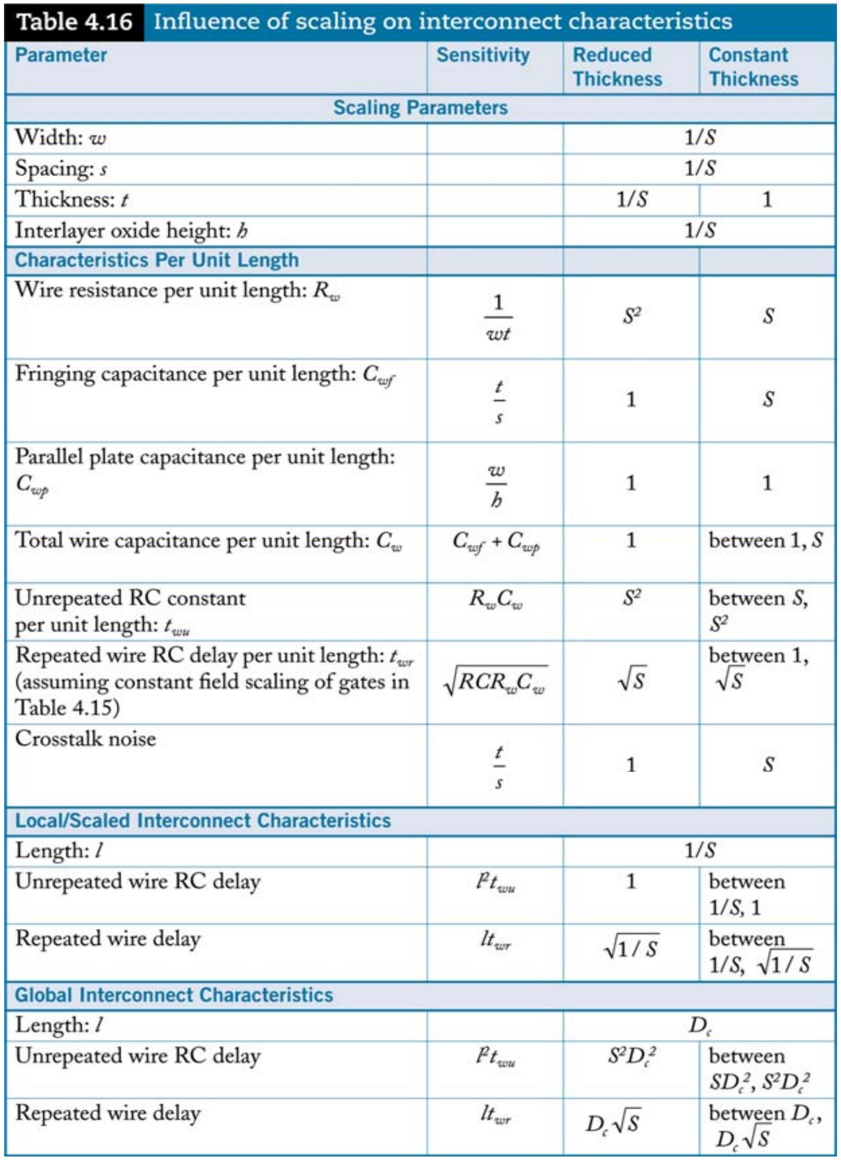
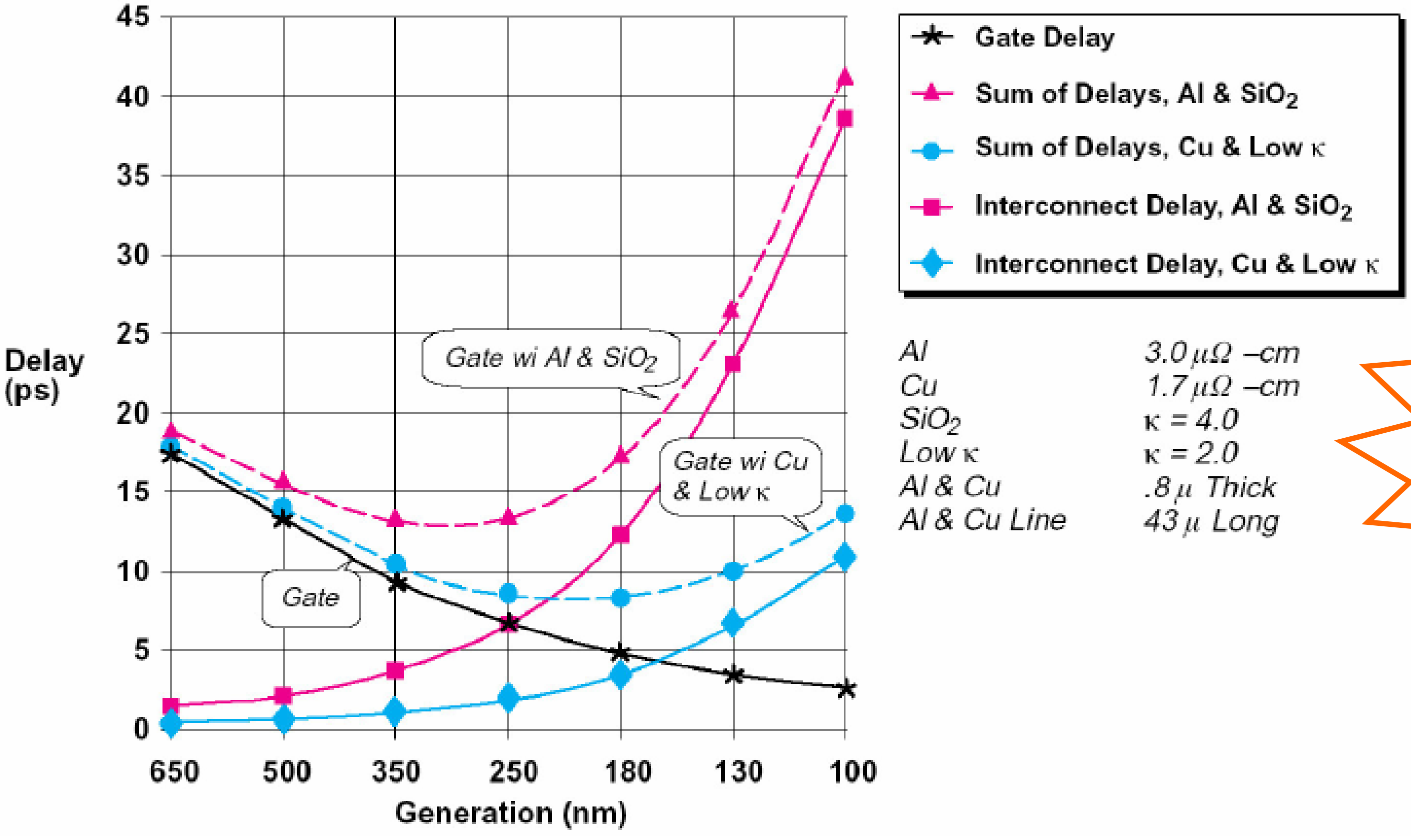


Table 4.16 Influence of scaling on interconnect characteristics

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)	$\sqrt{RCR_wC_w}$	$\sqrt{S}$	between 1, $\sqrt{S}$
Crosstalk noise	$\frac{t}{s}$	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, $S^2D_c^2$
Repeated wire delay	$l t_{wr}$	$D_c\sqrt{S}$	between D_c, $D_c\sqrt{S}$

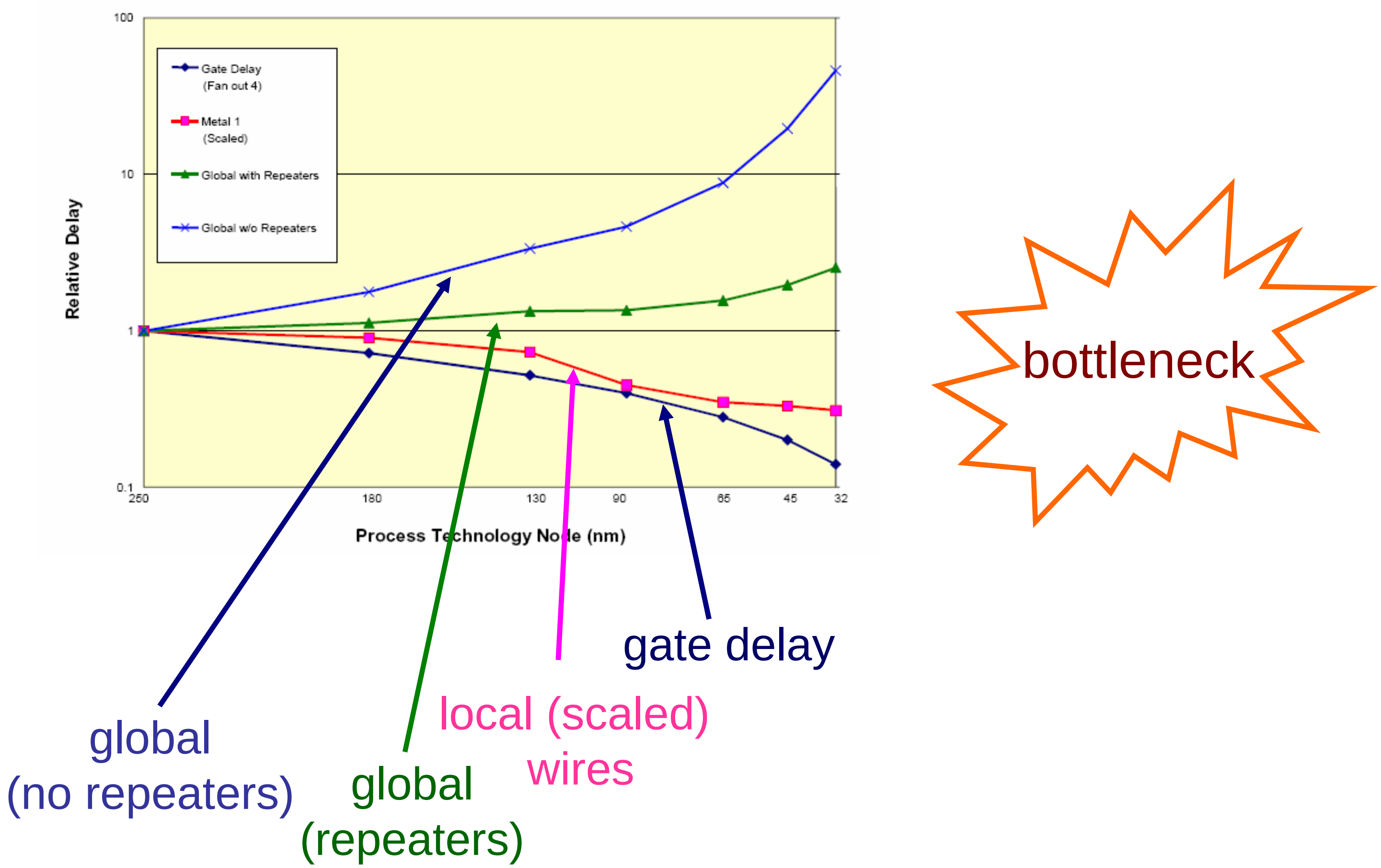
Interconnect delay is dominating gate delay



Al	*3.0 μΩ –cm*
Cu	*1.7 μΩ –cm*
SiO_2	*κ = 4.0*
Low κ	*κ = 2.0*
Al & Cu	*.8 μ Thick*
Al & Cu Line	*43 μ Long*

bottleneck

ITRS predictions imply global wires will most likely be buffered to reduce their delay



bottleneck

gate delay

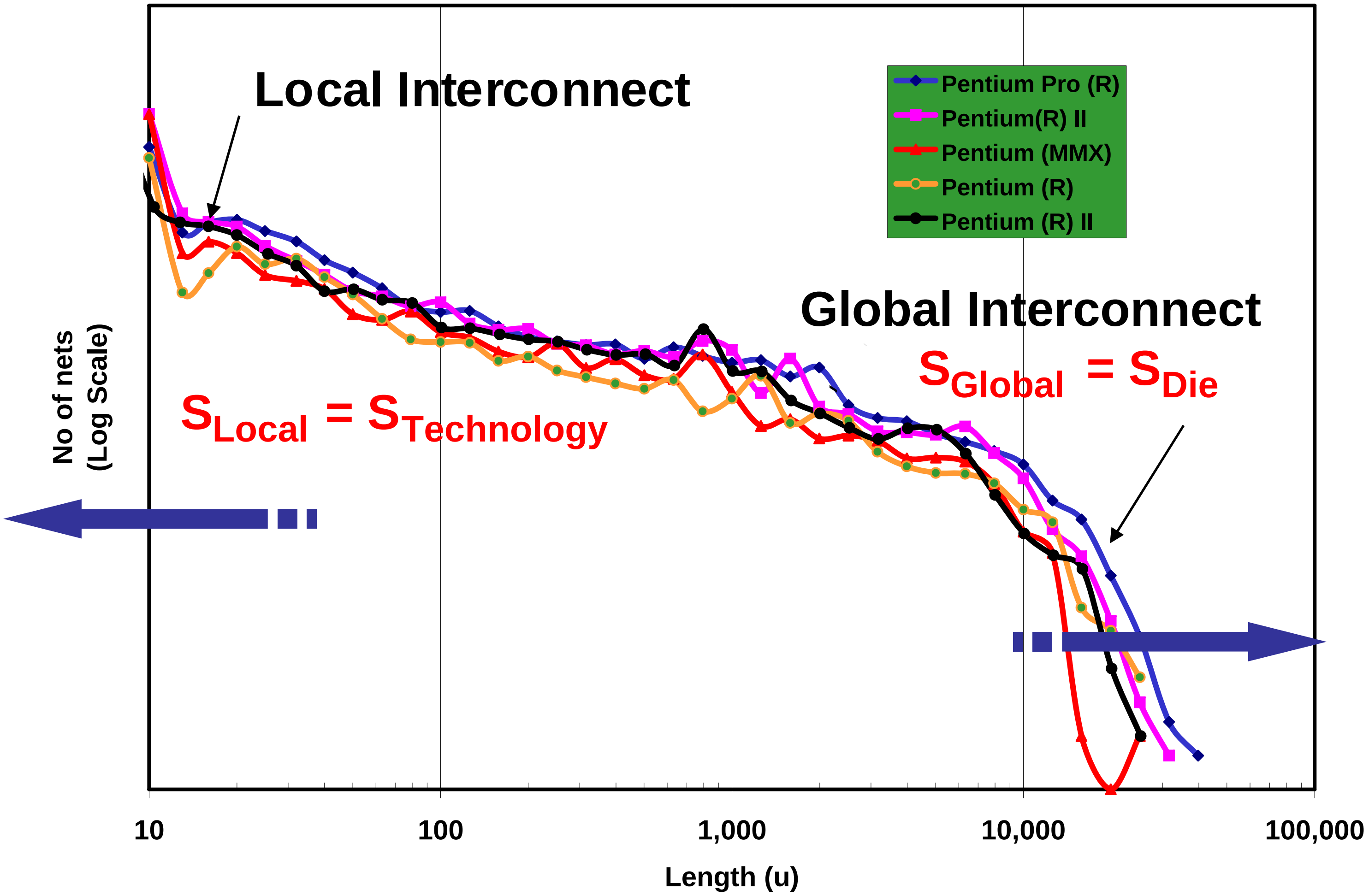
local (scaled) wires

global (no repeaters)

global (repeaters)

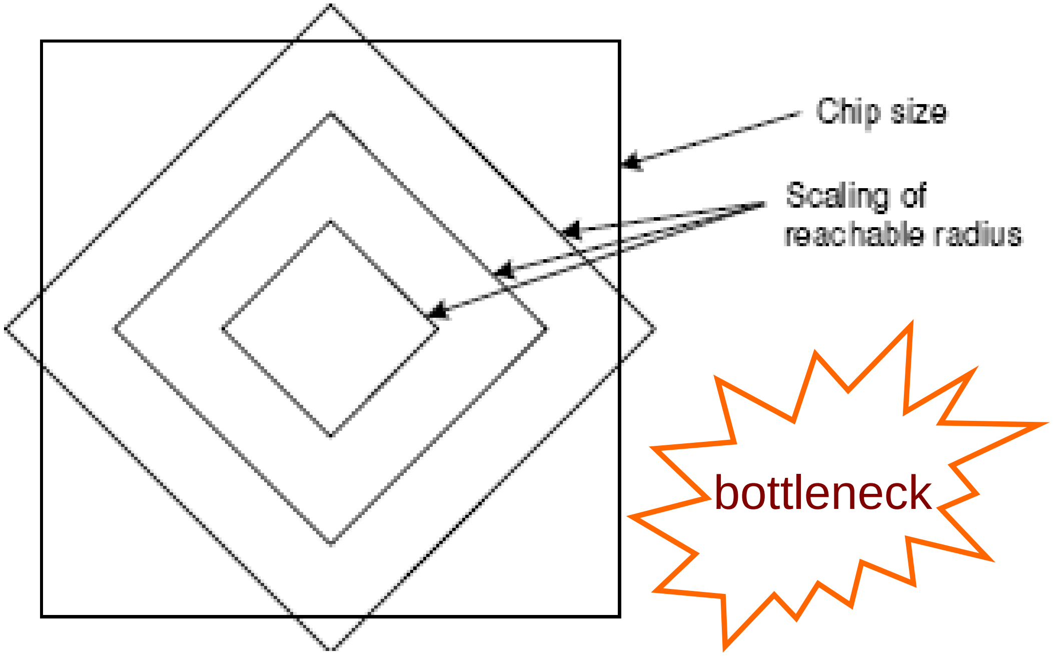
- Delay of local interconnects is relatively scaling well; global wires are a problem

Interconnect distribution is roughly the same; more local than global wires

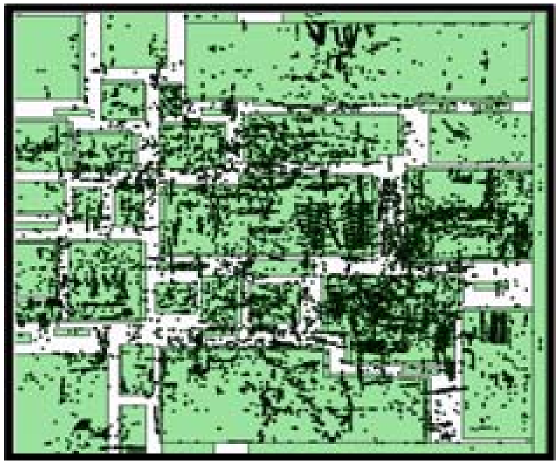


Source: Intel

With scaling the reachable radius of a buffer decreases → we need more and more buffers



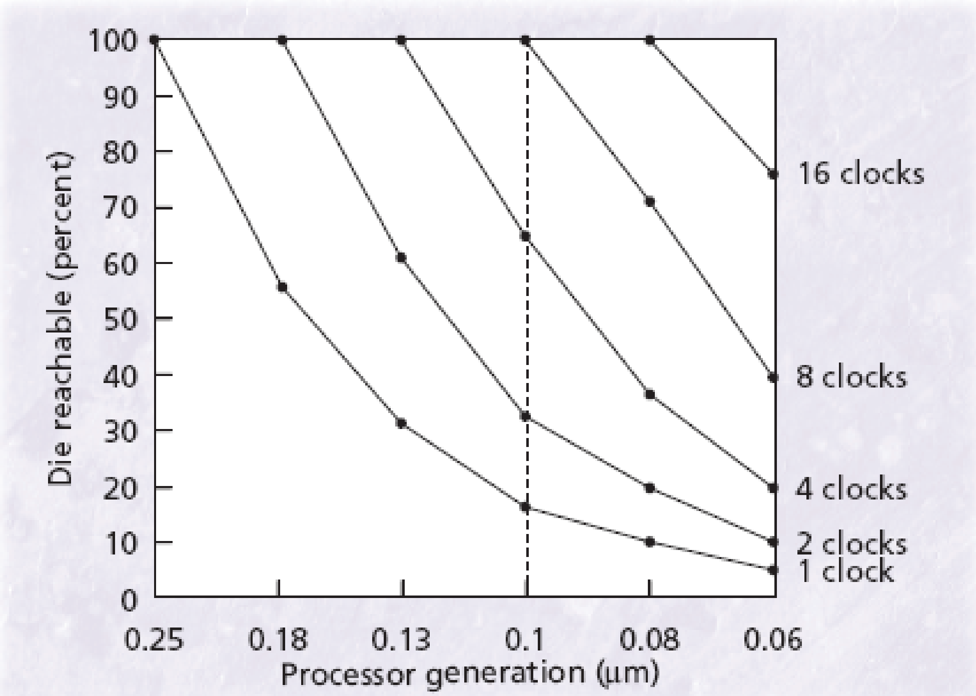
bottleneck



repeaters required to buffer Itanium global interconnects

- A corner-to-corner (BL-UR) wire in Itanium (180nm) requires 6 repeaters to span die
- Repeaters consume chip area; consume power; add vias

It takes an increasing number of clock cycles to span a die



[Matzke, TI'97]

⇒ Wires need to be pipelined (repeaters with states) to maintain synchronization in face of latency variations

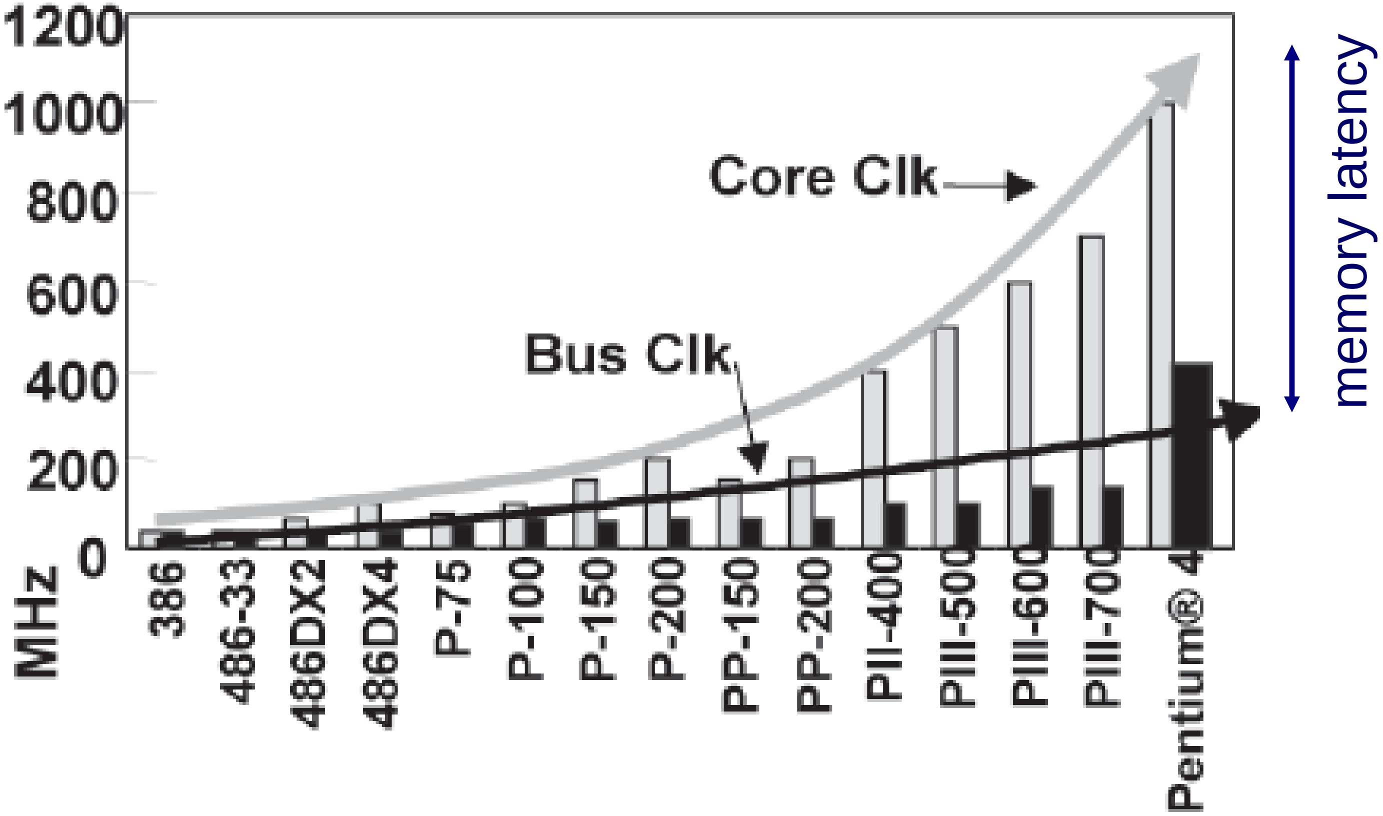
⇒ Use networks that route packets instead of global wires (network-on-a-chip NoC)

Lecture 02: CMOS scaling theory

- Device scaling
- Interconnect scaling
- More implications for design and architecture
- Readings and project assignments

Scaling of transistor delays (with ~constant power density) → scale frequency

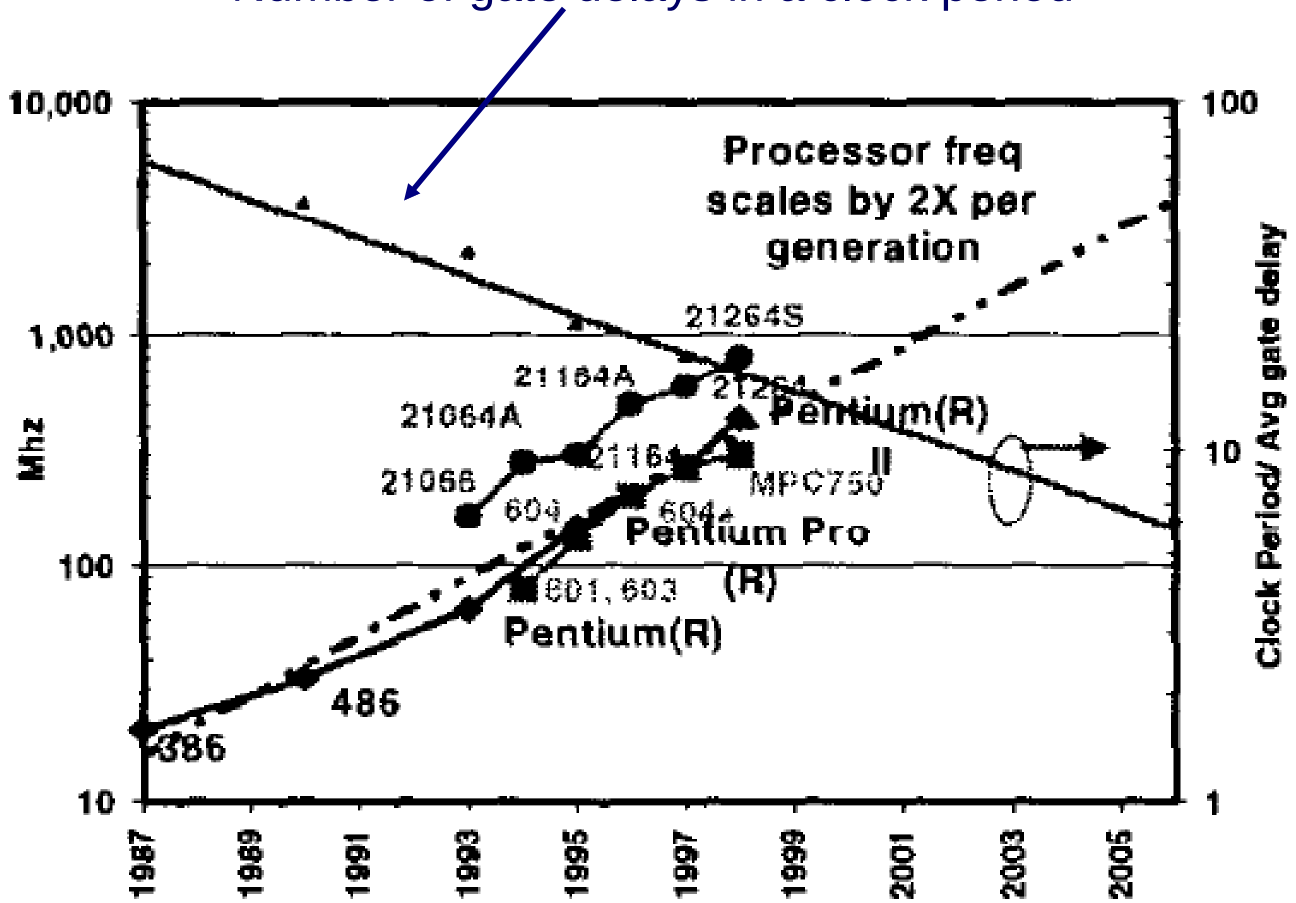
[Gelsinger'01]



memory latency

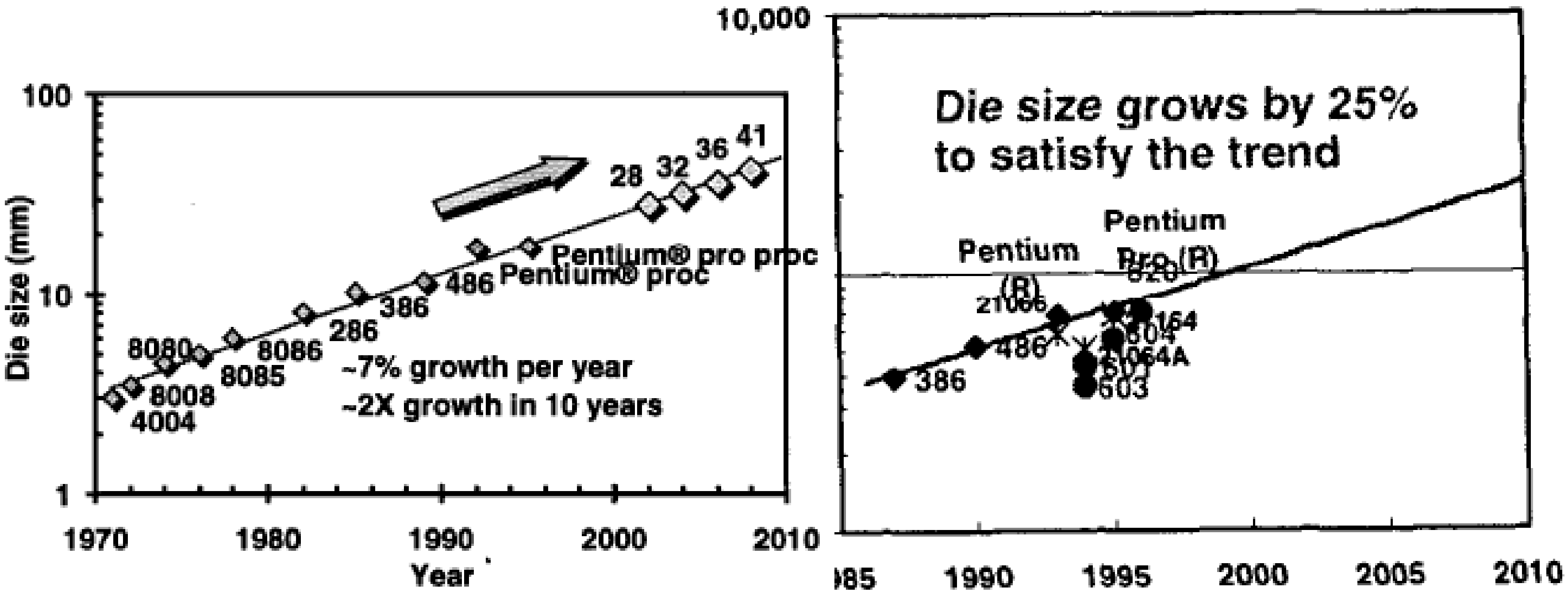
Pipelines depths were getting shorter → even larger frequency scaling

Number of gate delays in a clock period



Clock frequency was doubling every generation (not just increasing by 43%)

Die size trends

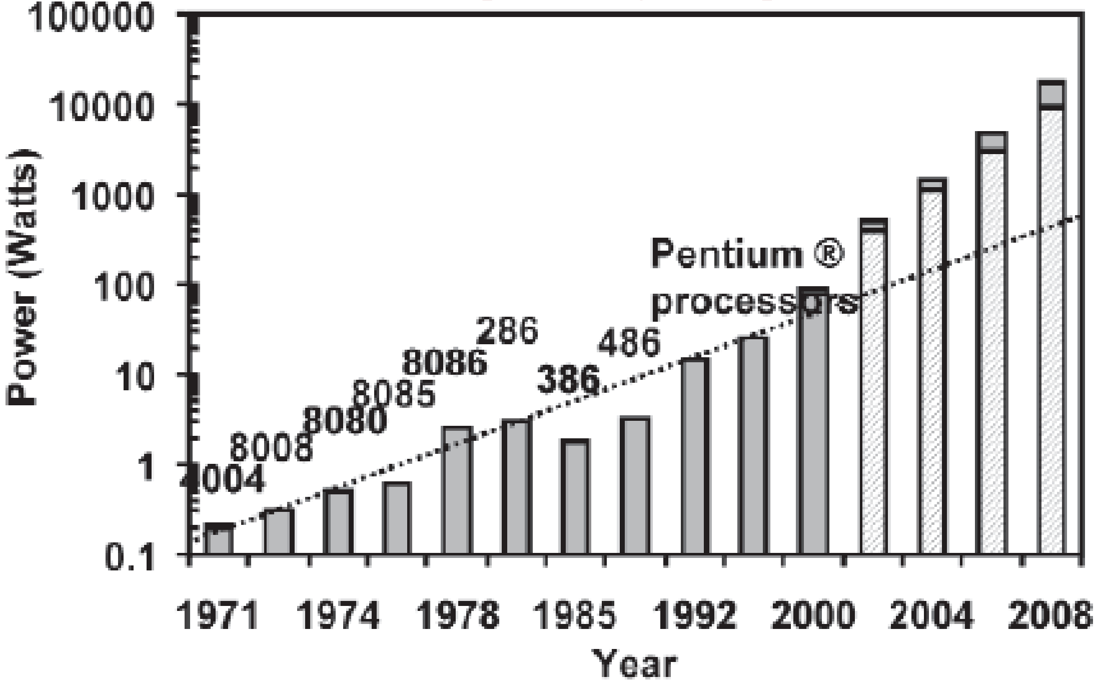


[source De/Brokar’99]

Borkar’99

Total power was increasing (mostly because of 2× frequency + die) until we hit a “wall”

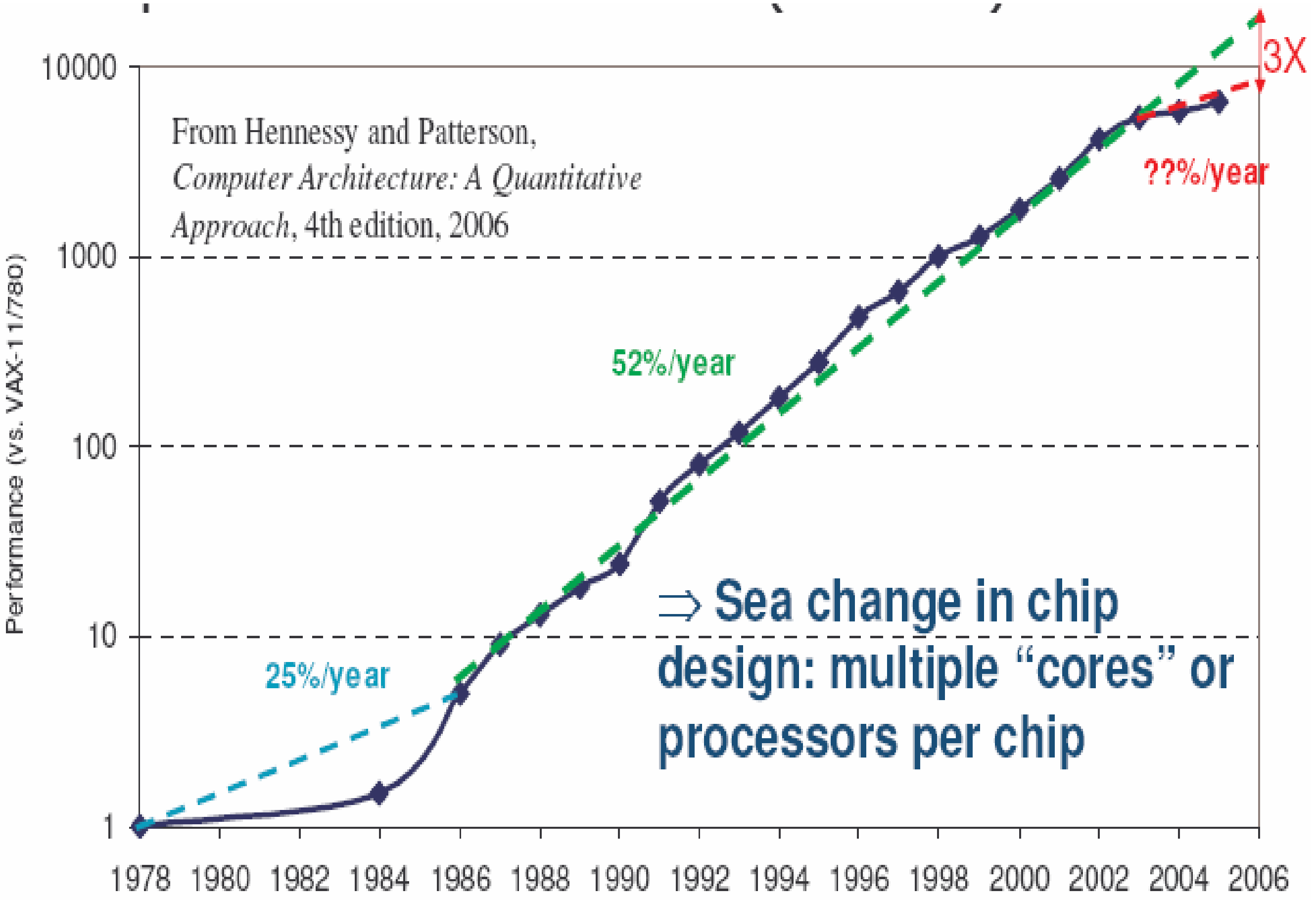
[Gelsinger’01]



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.”

Uniprocessor performance (SPECint)

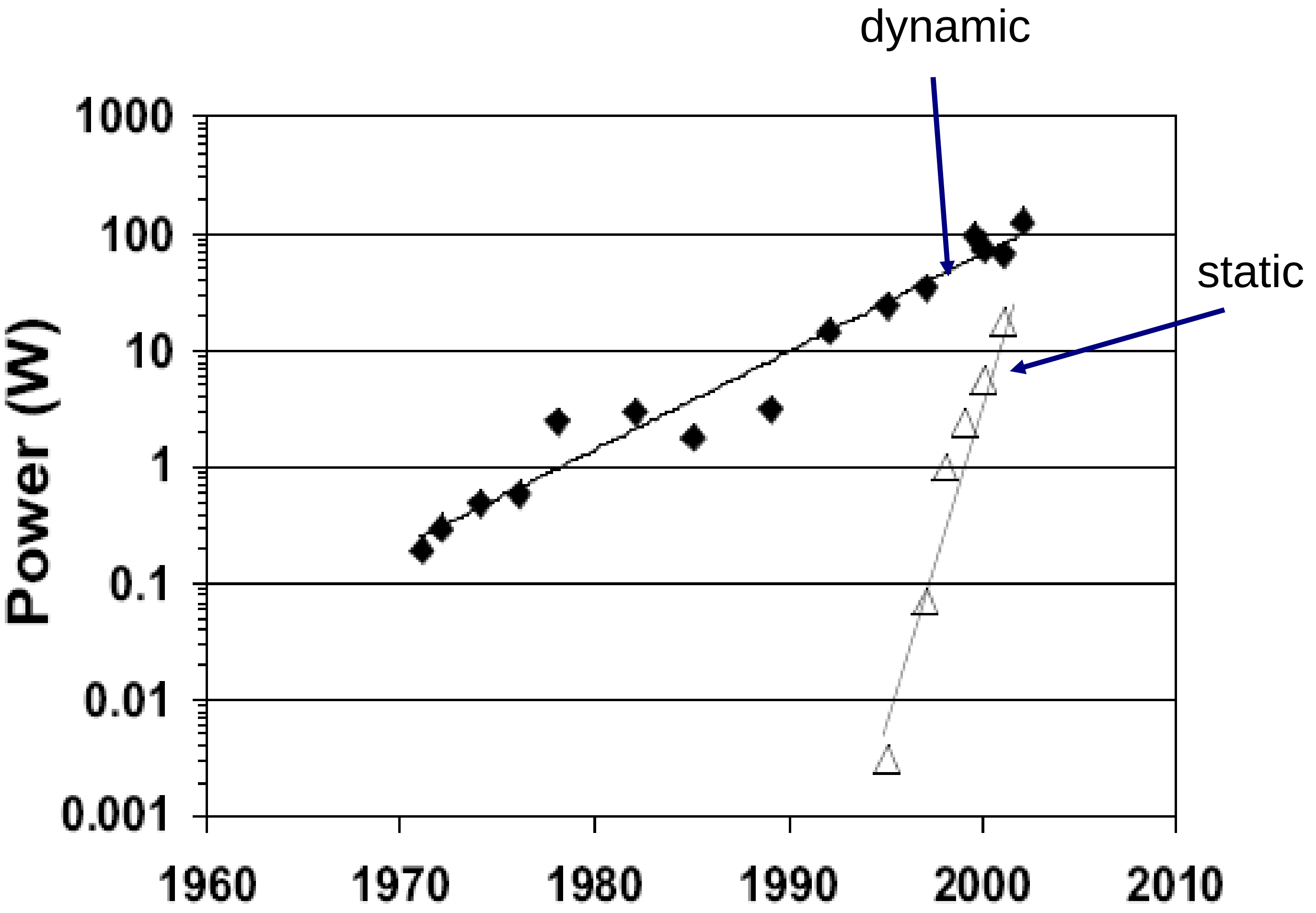


From Hennessy and Patterson, *Computer Architecture: A Quantitative Approach*, 4th edition, 2006

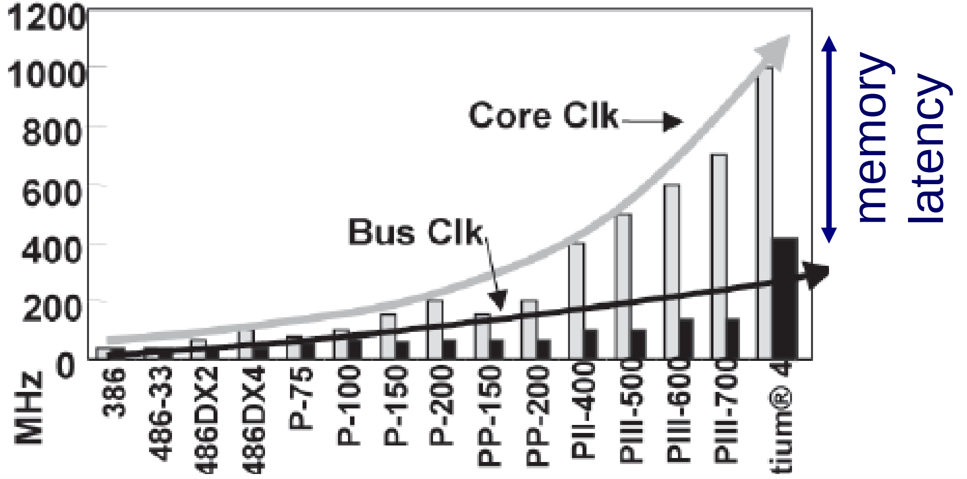
⇒ Sea change in chip design: multiple "cores" or processors per chip

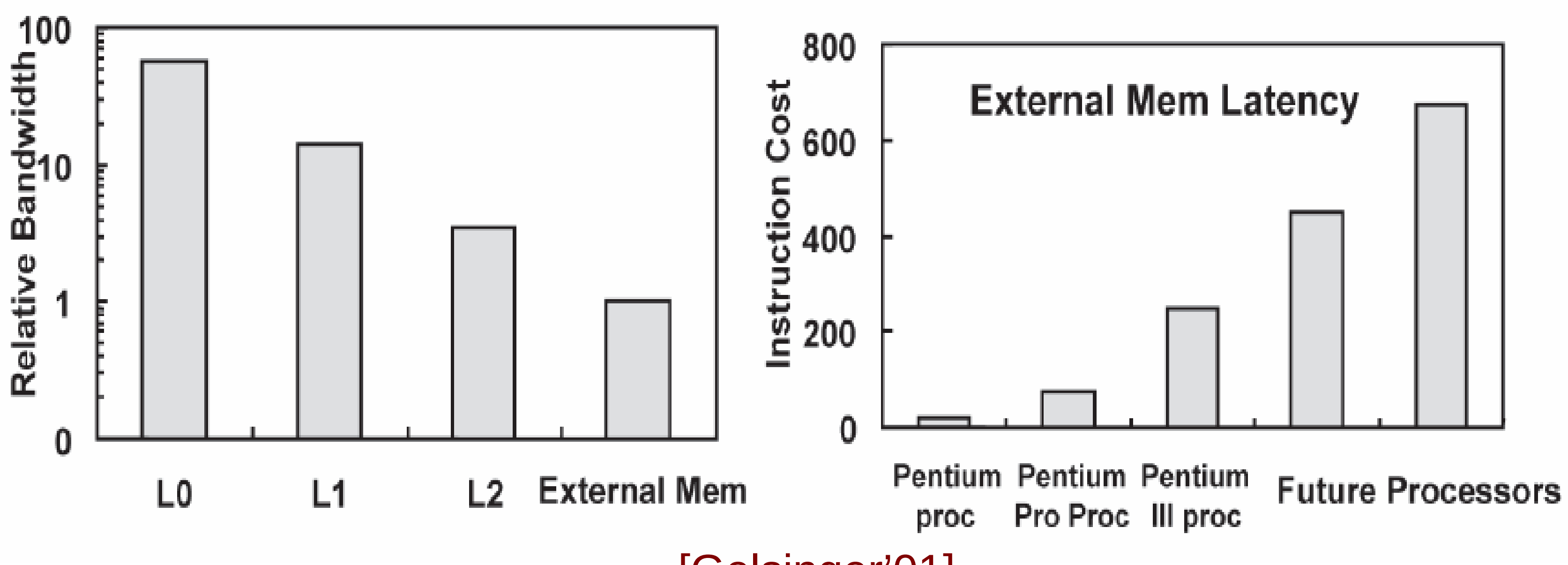
Uniprocessor performance improvement is slowing down (or even stopped)

Leakage power is becoming a bottleneck → increase in power density (w/o freq increase)



Another wall: external memory latency





[Gelsinger’01]

Lecture 02: CMOS scaling theory

- Device scaling
- Interconnect scaling
- More implications for design and architecture
- Readings and project assignments

Reading assignments for next lecture

- “Turning Silicon on Its Edge,” IEEE Circuits & Devices Magazine, Jan/Feb’04. ⇒ Yiwen

- “SOI technology for the GHz era,” IBM J. Res. & Dev., Vol. 46. ⇒ Cesare

- “Effect of increasing chip density on the evolution of computer architectures,” IBM J. Res & Dev, Vol. 46.

  ⇒ Brendan

- “Repeater scaling and its impact on CAD,” IEEE Trans. on CAD, Vol. 23(4) ⇒ Elif