Scaling of MOS Circuits

CONTENTS

Scaling of MOS Circuits

1.What is Scaling?

Proportional adjustment of the dimensions of an electronic device while maintaining the electrical properties of the device, results in a device either *larger* or *smaller* than the un-scaled device. Then *Which way do we scale the devices for VLSI? BIG and SLOW … or* ***SMALL*** *and* ***FAST****? What do we gain?*

2.Why Scaling?...

Scale the devices and wires down, Make the chips 'fatter' – functionality, intelligence, memory – and – faster, Make more chips per wafer – increased yield, Make the end user Happy by giving more for less and therefore, make MORE MONEY!!

3.FoM for Scaling

Impact of scaling is characterized in terms of several indicators:

- Minimum feature size
- Number of gates on one chip
- Power dissipation
- Maximum operational frequency
- Die size
- Production cost

Many of the FoMs can be improved by shrinking the dimensions of transistors and interconnections. Shrinking the separation between features – transistors and wires Adjusting doping levels and supply voltages.

3.1 Technology Scaling

Goals of scaling the dimensions by 30%:

Reduce gate delay by 30% (increase operating frequency by 43%)

Double transistor density

Reduce energy per transition by 65% (50% power savings @ 43% increase in frequency)

Die size used to increase by 14% per generation

Technology generation spans 2-3 years

Figure1 to Figure 5 illustrates the technology scaling in terms of minimum feature size, transistor count, prapogation delay, power dissipation and density and technology generations.

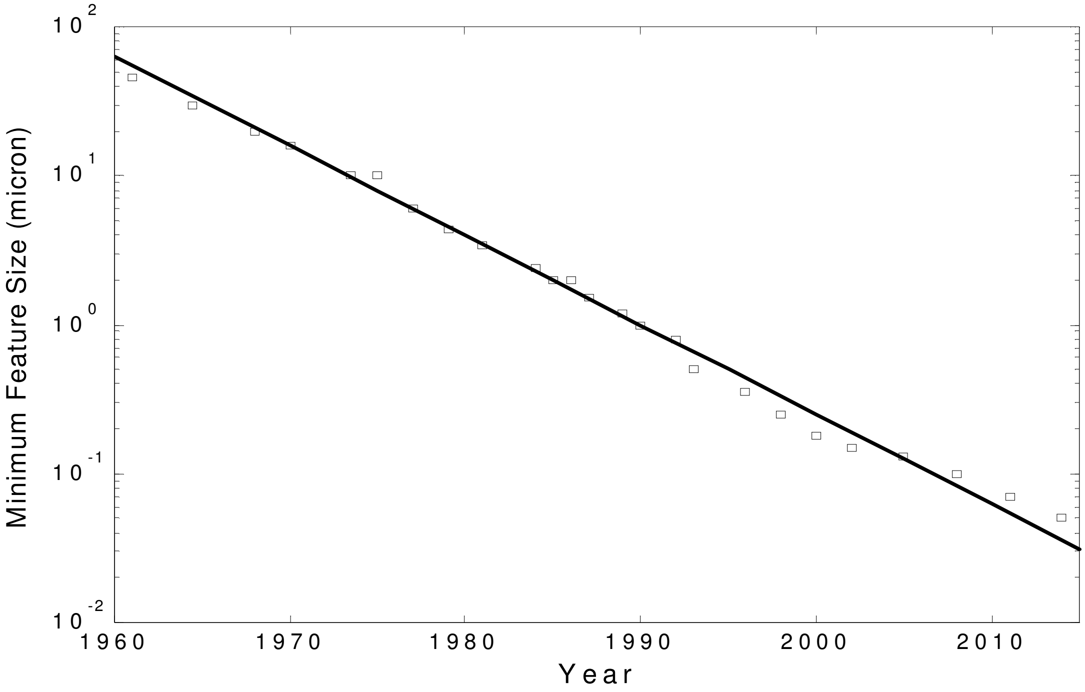


Figure-1:Technology Scaling (1)

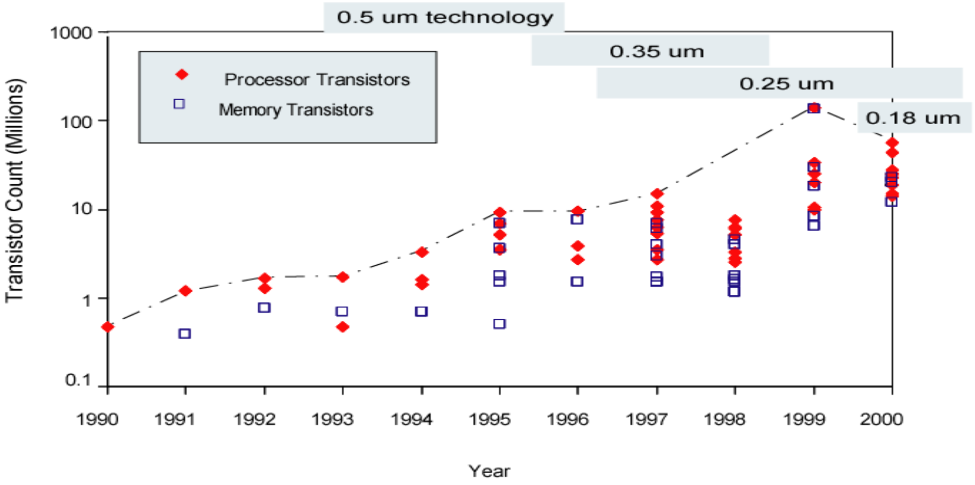
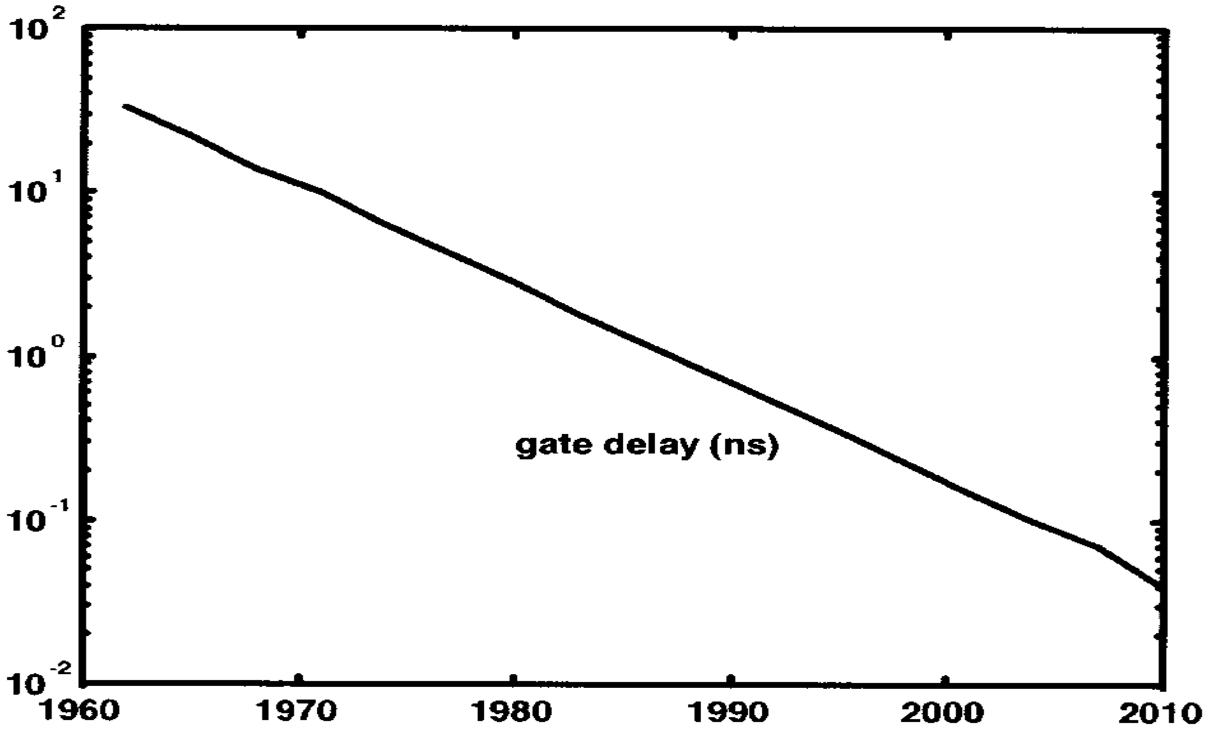
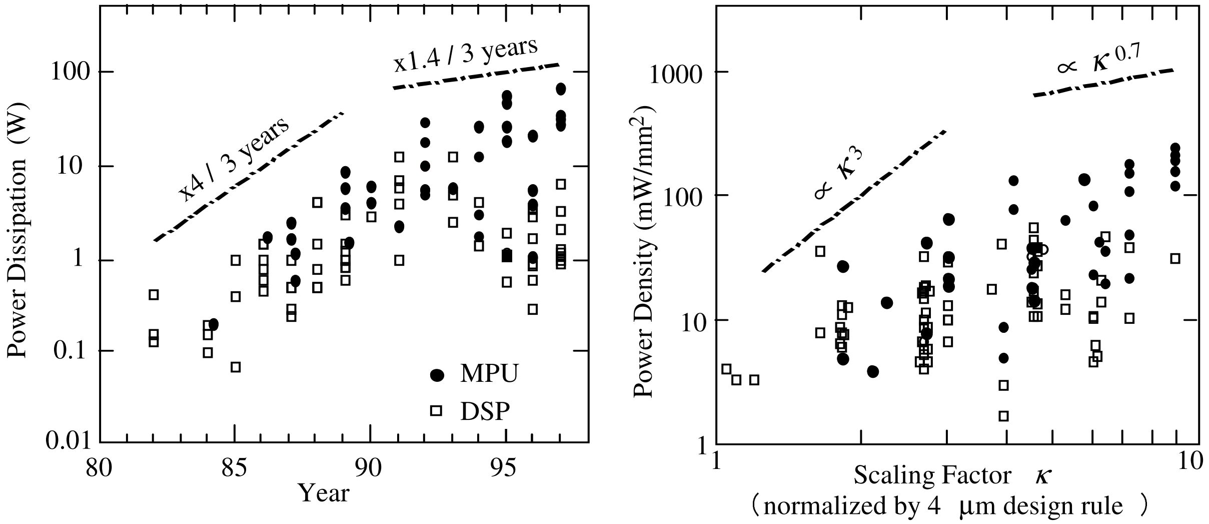


Figure-2:Technology Scaling (2)



Propagation Delay

Figure-3**:**Technology Scaling (3)



(a) Power dissipation vs. year.

(b) Power density vs. scaling factor.

Figure-4**:**Technology Scaling (4)

Technology Generations

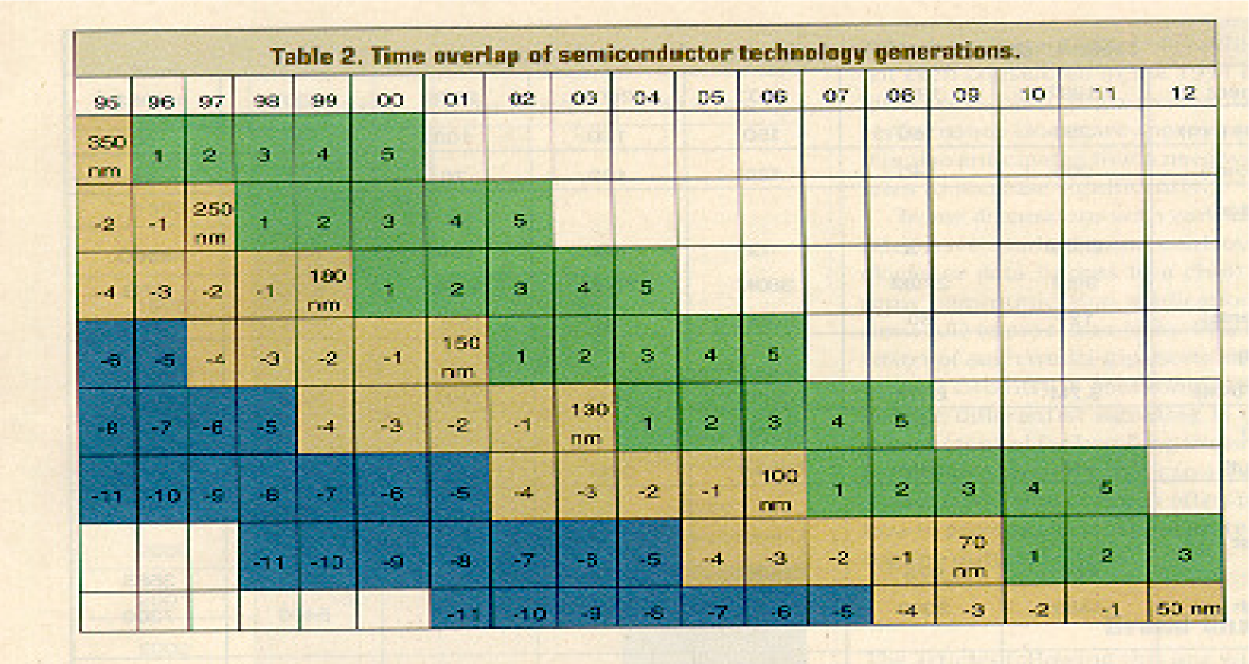


Table 2. Time overlap of semiconductor technology generations.

95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12
350 nm	1	2	3	4	5												
-2	-1	250 nm	1	2	3	4	5										
-4	-3	-2	-1	180 nm	1	2	3	4	5								
-6	-5	-4	-3	-2	-1	150 nm	1	2	3	4	5						
-8	-7	-6	-5	-4	-3	-2	-1	130 nm	1	2	3	4	5				
-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	100 nm	1	2	3	4	5	
			-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	70 nm	1	2	3
						-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	50 nm

Figure-5:Technology generation

4. International Technology Roadmap for Semiconductors (ITRS)

Table 1 lists the parameters for various technologies as per ITRS.

Year of Introduction	1999	2000	2001	2004	2008	2011	2014
Technology node [nm]	180		130	90	60	40	30
Supply [V]	1.5-1.8	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6	0.3-0.6
Wiring levels	6-7	6-7	7	8	9	9-10	10
Max frequency [GHz],Local-Global	1.2	1.6-1.4	2.1-1.6	3.5-2	7.1-2.5	11-3	14.9 -3.6
Max µP power [W]	90	106	130	160	171	177	186
Bat. power [W]	1.4	1.7	2.0	2.4	2.1	2.3	2.5

Node years: 2007/65nm, 2010/45nm, 2013/33nm, 2016/23nm

Table 1: ITRS

5.Scaling Models

- ❑ Full Scaling (Constant Electrical Field)

Ideal model – dimensions and voltage scale together by the same scale factor

- ❑ Fixed Voltage Scaling

Most common model until recently – only the dimensions scale, voltages remain constant

- ❑ General Scaling

Most realistic for today’s situation – voltages and dimensions scale with different factors

6.Scaling Factors for Device Parameters

Device scaling modeled in terms of generic scaling factors:
1/α and 1/β

- 1/β: scaling factor for supply voltage V_{DD} and gate oxide thickness D
- 1/α: linear dimensions both horizontal and vertical dimensions

Why is the scaling factor for gate oxide thickness different from other linear horizontal and vertical dimensions? Consider the cross section of the device as in Figure 6,various parameters derived are as follows.

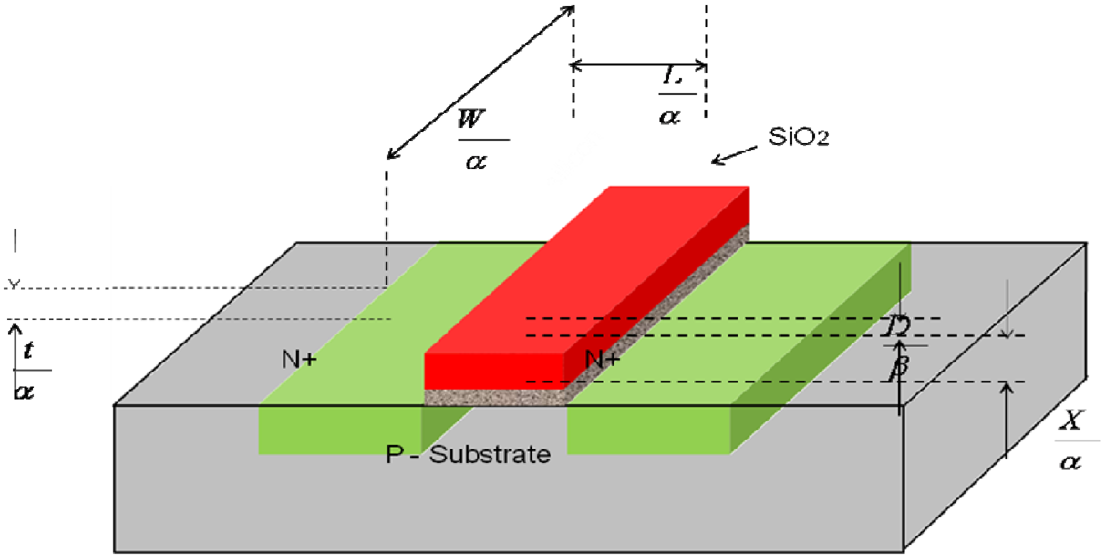


Figure-6:Technology generation

- Gate area A_g

$$A_g = L * W$$

Where L: Channel length and W: Channel width and both are scaled by $1/\alpha$

Thus A_g is scaled up by $1/\alpha^2$

- Gate capacitance per unit area C_o or C_{ox}

$C_{ox} = \varepsilon_{ox}/D$

Where ε_{ox} is permittivity of gate oxide(thin-ox)= $\varepsilon_{ins}\varepsilon_o$ and D is the gate oxide thickness scaled by $1/\beta$

Thus C_{ox} is scaled up by $\dfrac{1}{\left(\dfrac{1}{\beta}\right)} = \beta$

- Gate capacitance C_g $C_g = C_o * L * W$

Thus C_g is scaled up by $\beta * 1/\alpha^2 = \beta/\alpha^2$

- Parasitic capacitance C_x

C_x is proportional to A_x/d

where d is the depletion width around source or drain and scaled by $1/\alpha$

A_x is the area of the depletion region around source or drain, scaled by $(1/\alpha^2)$.

Thus C_x is scaled up by $\{1/(1/\alpha)\} * (1/\alpha^2) = 1/\alpha$

- Carrier density in channel Q_{on}

$Q_{on} = C_o * V_{gs}$

where Q_{on} is the average charge per unit area in the 'on' state.

C_o is scaled by β and V_{gs} is scaled by $1/\beta$

Thus Q_{on} is scaled by 1

- Channel Resistance R_{on}

$$R_{on} = \frac{L}{W} * \frac{1}{Q_{on} * \mu}$$

Where μ = channel carrier mobility and assumed constant

Thus R_{on} is scaled by 1.

- Gate delay T_d

T_d is proportional to $R_{on}*C_g$

T_d is scaled by $\frac{1}{\alpha^2}*\beta=\frac{\beta}{\alpha^2}$

- Maximum operating frequency f_o

$$f_o=\frac{W}{L}*\frac{\mu C_o V_{DD}}{C_g}$$

f_o is inversely proportional to delay T_d and is scaled by

$$\beta*\left(\frac{1}{\beta^2}\right)=\frac{1}{\beta}$$

- Saturation current I_{dss}

$$I_{dss}=\frac{C_o\mu}{2}*\frac{W}{L}*(V_{gs}-V_t)^2$$

Both V_{gs} and V_t are scaled by $(1/\ \beta)$. Therefore, I_{dss} is scaled by $\frac{1}{\left(\beta/\alpha^2\right)}=\frac{\alpha^2}{\beta}$

- Current density J

Current density, $J=\frac{I_{dss}}{A}$ where A is cross sectional area of the Channel in the “on” state which is scaled by $(1/\ \alpha^2)$.

So, J is scaled by

$$\frac{1/\beta}{1/\alpha^2}=\frac{\alpha^2}{\beta}$$

- Switching energy per gate E_g

$$E_g=\frac{1}{2}C_g V_{DD}^2$$

So E_g is scaled by

$$\frac{\beta}{\alpha^2}*\left(\frac{1}{\beta^2}\right)=\frac{1}{\alpha^2\beta}$$

- Power dissipation per gate P_g

$$P_g = P_{gs} + P_{gd}$$

P_g comprises of two components: static component P_{gs} and dynamic component P_{gd}:

Where, the static power component is given by: $P_{gs} = \frac{V_{DD}^2}{R_{on}}$

And the dynamic component by: $P_{gd} = E_g f_o$

Since V_{DD} scales by (1/β) and R_{on} scales by 1, P_{gs} scales by $(1/\beta^2)$.

Since Eg scales by $(1/\alpha^2\ \beta)$ and f_o by $(\alpha_2\ /\beta)$, P_{gd} also scales by $(1/\beta^2)$. Therefore, P_g scales by $(1/\beta^2)$.

- Power dissipation per unit area P_a

$$P_a = \frac{P_g}{A_g} = \frac{\left(\frac{1}{\beta^2}\right)}{\left(\frac{1}{\alpha^2}\right)} = \frac{\alpha^2}{\beta^2}$$

- Power – speed product P_T $\quad P_T = P_g * T_d = \frac{1}{\beta^2}\left(\frac{\beta}{\alpha^2}\right) = \frac{1}{\alpha^2\beta}$

6.1 Scaling Factors …Summary

Various device parameters for different scaling models are listed in Table 2 below.

Table 2: Device parameters for scaling models

NOTE: for Constant E: β=α; for Constant V: β=1

Parameters	Description	General (Combined V and Dimension)	Constant E	Constant V
V_{DD}	Supply voltage	1/β	1/α	1
L	Channel length	1/α	1/α	1/α
W	Channel width	1/α	1/α	1/α
D	Gate oxide thickness	1/β	1/α	1
A_g	Gate area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha^2$
C_o (or C_{ox})	Gate capacitance per unit area	β	α	1
C_g	Gate capacitance	β/α^2	1/α	$1/\alpha^2$
C_x	Parsitic capacitance	1/α	1/α	1/α
Q_{on}	Carrier density	1	1	1
R_{on}	Channel resistance	1	1	1
I_{dss}	Saturation current	1/β	1/α	1

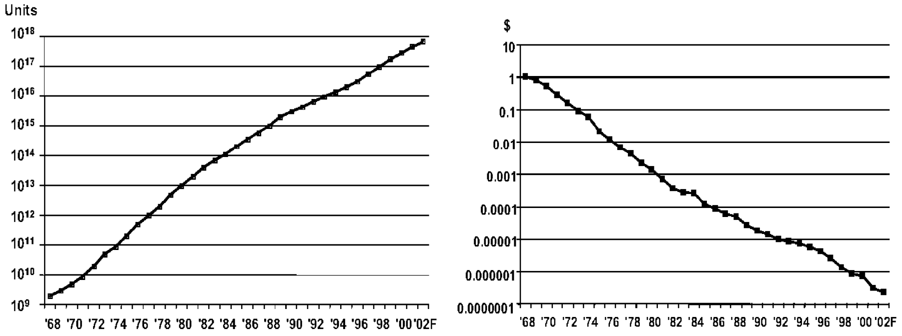
Parameters	**Description**	**General (Combined V and Dimension)**	**Constant E**	**Constant V**
A_c	**Conductor cross section area**	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha^2$
J	**Current density**	α^2 / β	α	α^2
V_g	**Logic 1 level**	$1 / \beta$	$1 / \alpha$	1
E_g	**Switching energy**	$1 / \alpha^2 \beta$	$1 / \alpha^3$	$1/\alpha^2$
P_g	**Power dissipation per gate**	$1 / \beta^2$	$1/\alpha^2$	1
N	**Gates per unit area**	α^2	α^2	α^2
P_a	**Power dissipation per unit area**	α^2 / β^2	1	α^2
T_d	**Gate delay**	β / α^2	$1 / \alpha$	$1/\alpha^2$
f_o	**Max. operating frequency**	α^2 / β	α	α^2
P_T	**Power speed product**	$1 / \alpha^2 \beta$	$1 / \alpha^3$	$1/\alpha^2$

7.Implications of Scaling

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

7.1Cost Improvement

– Moore’s Law is still going strong as illustrated in Figure 7.



Source: Dataquest/Intel

Source: Dataquest/Intel

Figure-7:Technology generation

7.2:Interconnect Woes

- Scaled transistors are steadily improving in delay, but scaled wires are holding constant or getting worse.
- SIA made a gloomy forecast in 1997
  – Delay would reach minimum at 250 – 180 nm, then get worse because of wires
- But…
- For short wires, such as those inside a logic gate, the wire RC delay is negligible.
- However, the long wires present a considerable challenge.

- Scaled transistors are steadily improving in delay, but scaled wires are holding constant or getting worse.
- SIA made a gloomy forecast in 1997
  – Delay would reach minimum at 250 – 180 nm, then get worse because of wires
- But…
- For short wires, such as those inside a logic gate, the wire RC delay is negligible.
- However, the long wires present a considerable challenge.

Figure 8 illustrates delay Vs. generation in nm for different materials.

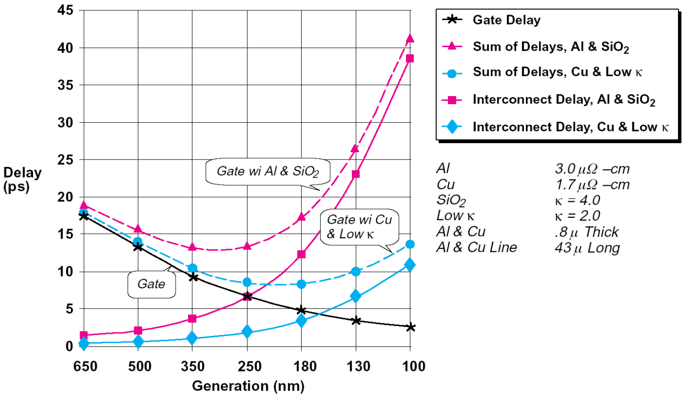


Figure-8:Technology generation

7.3 Reachable Radius

- We can’t send a signal across a large fast chip in one cycle anymore
- But the microarchitect can plan around this as shown in Figure 9.
  - Just as off-chip memory latencies were tolerated

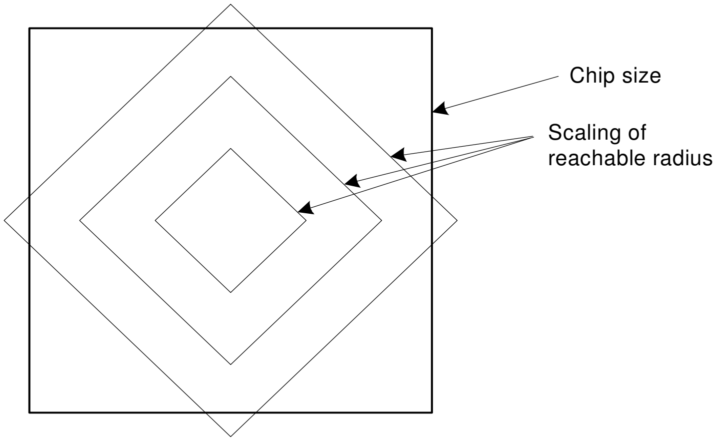


Figure-9:Technology generation

7.4 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(Figure 10)

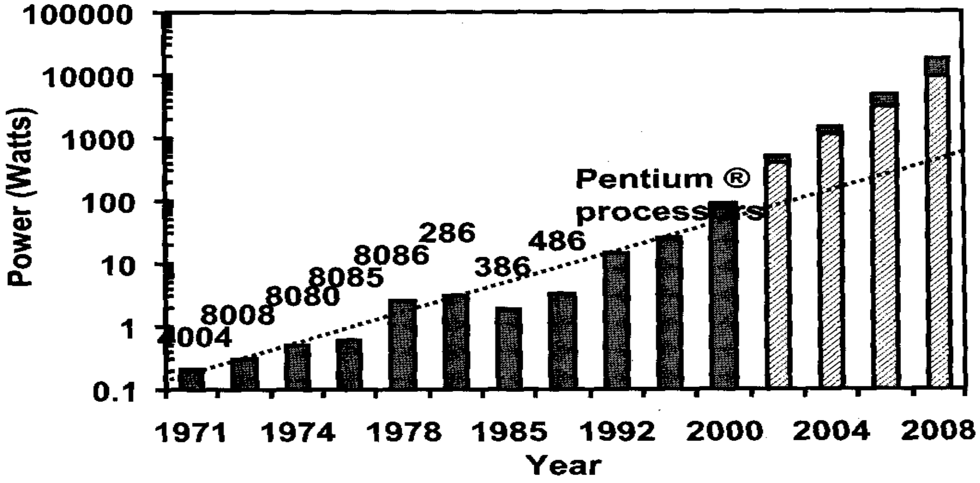
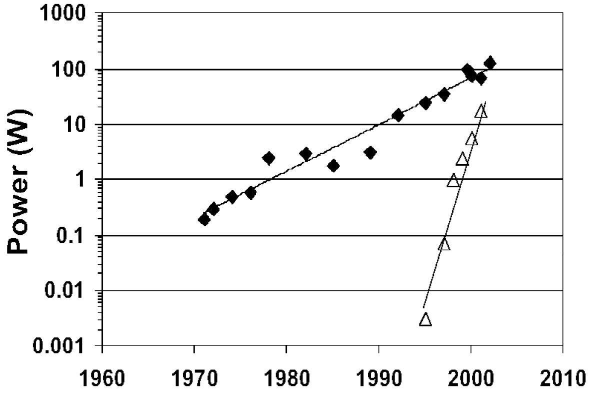


Figure-10:Technology generation

7.5 Static Power

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

A Major future challenge(Figure 11)



Moore(03)

Figure-11:Technology generation

7.6 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

7.7 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
  - Sub-threshold leakage, tunneling
  - Short channel effects
  - Fabrication costs
  - Electro-migration
  - Interconnect delay
- Rumors of demise have been exaggerated

8. Limitations of Scaling

Effects, as a result of scaling down- which eventually become severe enough to prevent further miniaturization.

- Substrate doping
- Depletion width
- Limits of miniaturization

- Limits of interconnect and contact resistance
- Limits due to sub threshold currents
- Limits on logic levels and supply voltage due to noise
- Limits due to current density

8.1 Substrate doping

- Substrate doping
- Built-in(junction) potential V_B depends on substrate doping level – can be neglected as long as V_B is small compared to $V_{DD.}$
- As length of a MOS transistor is reduced, the depletion region width –scaled down to prevent source and drain depletion region from meeting.
- the depletion region width d for the junctions is $d = \sqrt{\frac{2}{q}\frac{\xi_{si}}{N_B}\frac{\xi_0 V}{1}}$
- ε_{si} relative permittivity of silicon
- ε_0 permittivity of free space($8.85*10^{-14}$ F/cm)
- V effective voltage across the junction $V_a + V_b$
- q electron charge
- N_B doping level of substrate
- V_a maximum value Vdd-applied voltage
- V_b built in potential and $V_B = \frac{KT}{q}\ln\left[\frac{N_B}{n_i}\frac{N_D}{n_i}\right]$

8.2 Depletion width

- N_B is increased to reduce d , but this increases threshold voltage V_t -against trends for scaling down.
- Maximum value of N_B ($1.3*10^{19}$ cm^{-3} , at higher values, maximum electric field applied to gate is insufficient and no channel is formed.
- N_B maintained at satisfactory level in the channel region to reduce the above problem.
- E_{max} maximum electric field induced in the junction. $E_{max} = \frac{2V}{d}$

$$\sqrt{\frac{\ln\alpha}{\alpha}}$$

If N $_B$ is inreased by α Va =0 Vb increased by ln α and d is decreased by

- Electric field across the depletion region is increased by

$$1/\sqrt{\frac{\ln\alpha}{\alpha}}$$

- Reach a critical level E$_{crit}$ with increasing N $_B$

$$d=\sqrt{\frac{2}{q}\frac{\xi_{si}}{N_B}\xi\left\{\frac{E_{crit.}d}{2}\right\}}$$

Where $$d=\frac{\xi_{si}}{q}\frac{\xi_0}{N_B}(E_{crit})$$

Figure 12 , Figure 13 and Figure 14 shows the relation between substrate concentration Vs depletion width , Electric field and transit time.
Figure 15 demonstrates the interconnect length Vs. propagation delay and Figure 16 oxide thickness Vs. thermal noise.

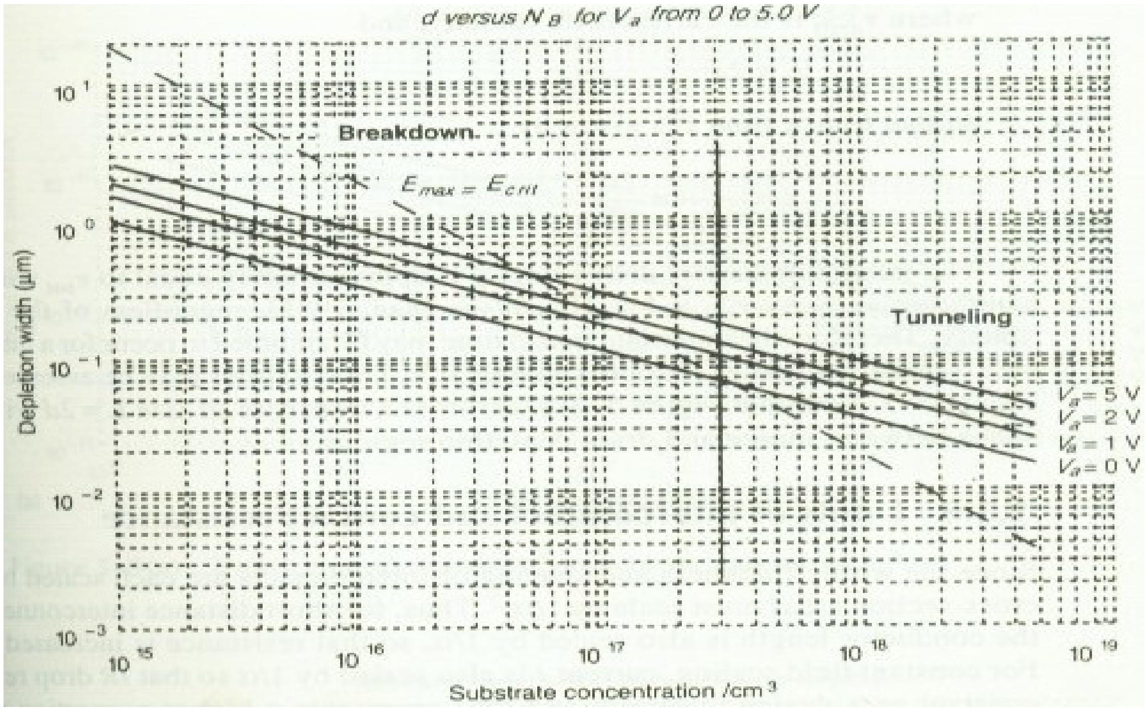


Figure-12:Technology generation

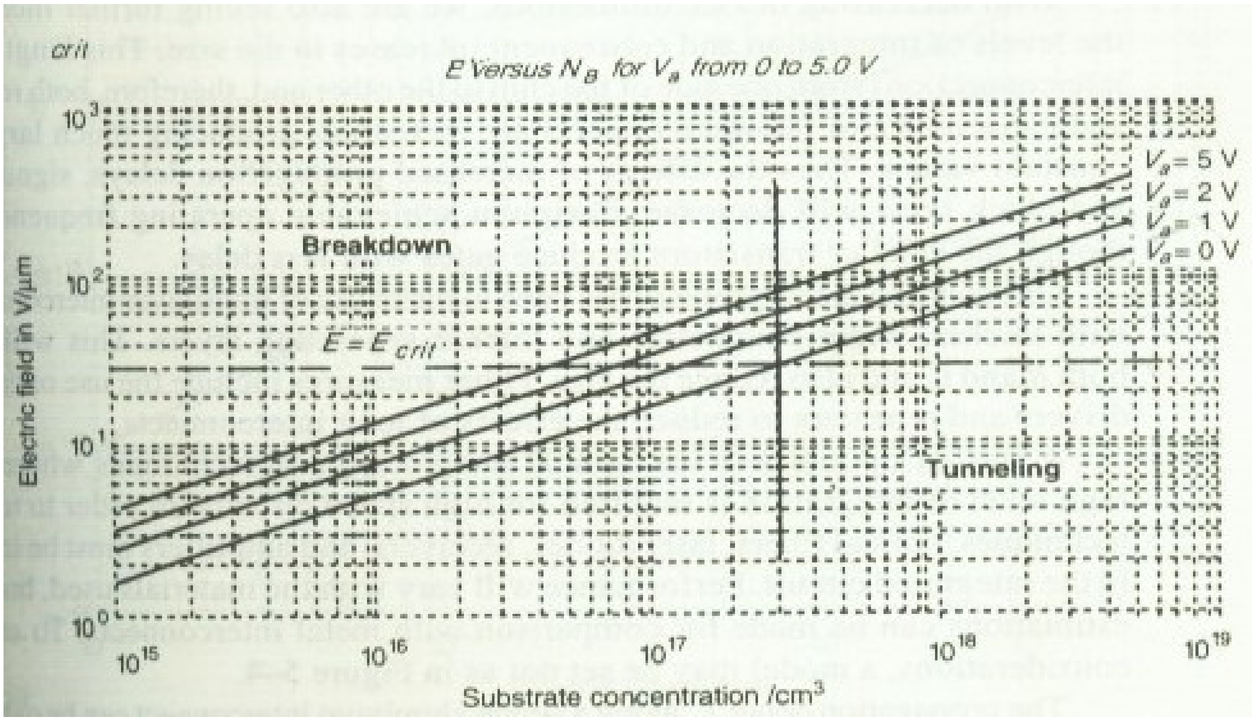


Figure-13:Technology generation

8.3 Limits of miniaturization

- minimum size of transistor; process tech and physics of the device
- Reduction of geometry; alignment accuracy and resolution
- Size of transistor measured in terms of channel length L

  L=2d (to prevent push through)
- L determined by N_B and Vdd
- Minimum transit time for an electron to travel from source to drain is

$$v_{drift} = \mu E$$

$$t = \frac{L}{V_{drift}} = \frac{2d}{\mu E}$$

- $_s$maximum carrier drift velocity is approx. Vsat,regardless of supply voltage

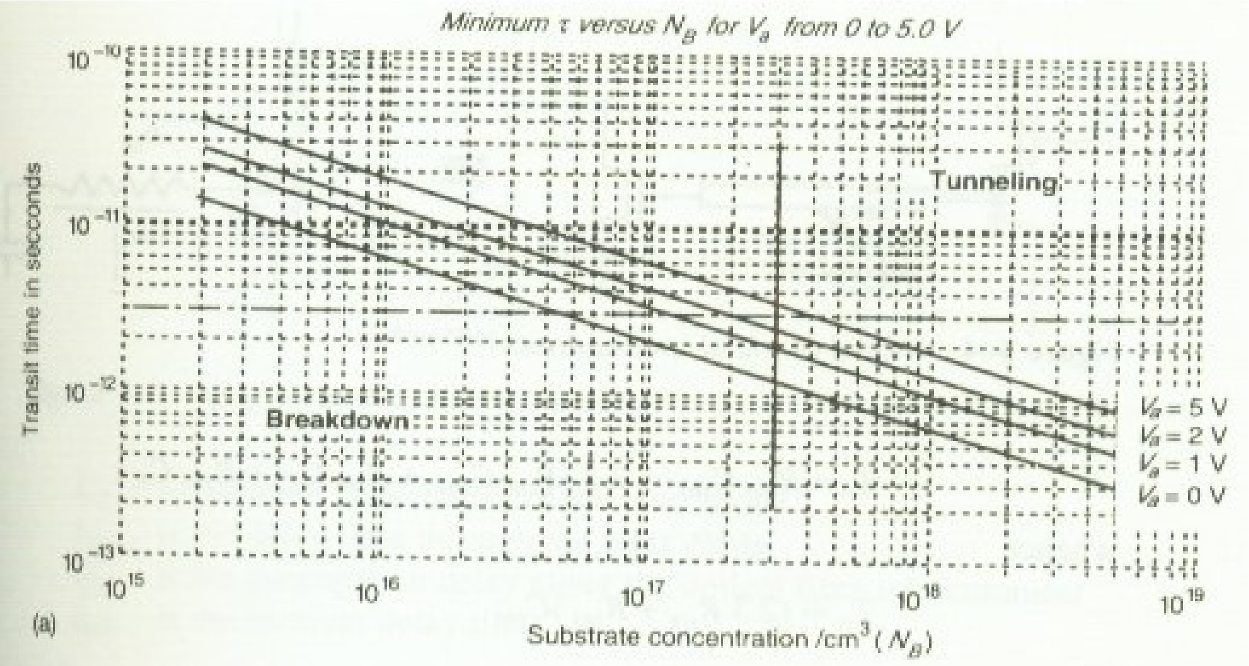


Figure-14**:**Technology generation

8.4 Limits of interconnect and contact resistance

- Short distance interconnect- conductor length is scaled by 1/α and resistance is increased by α
- For constant field scaling, I is scaled by 1/ α so that IR drop remains constant as a result of scaling.-driving capability/noise margin.

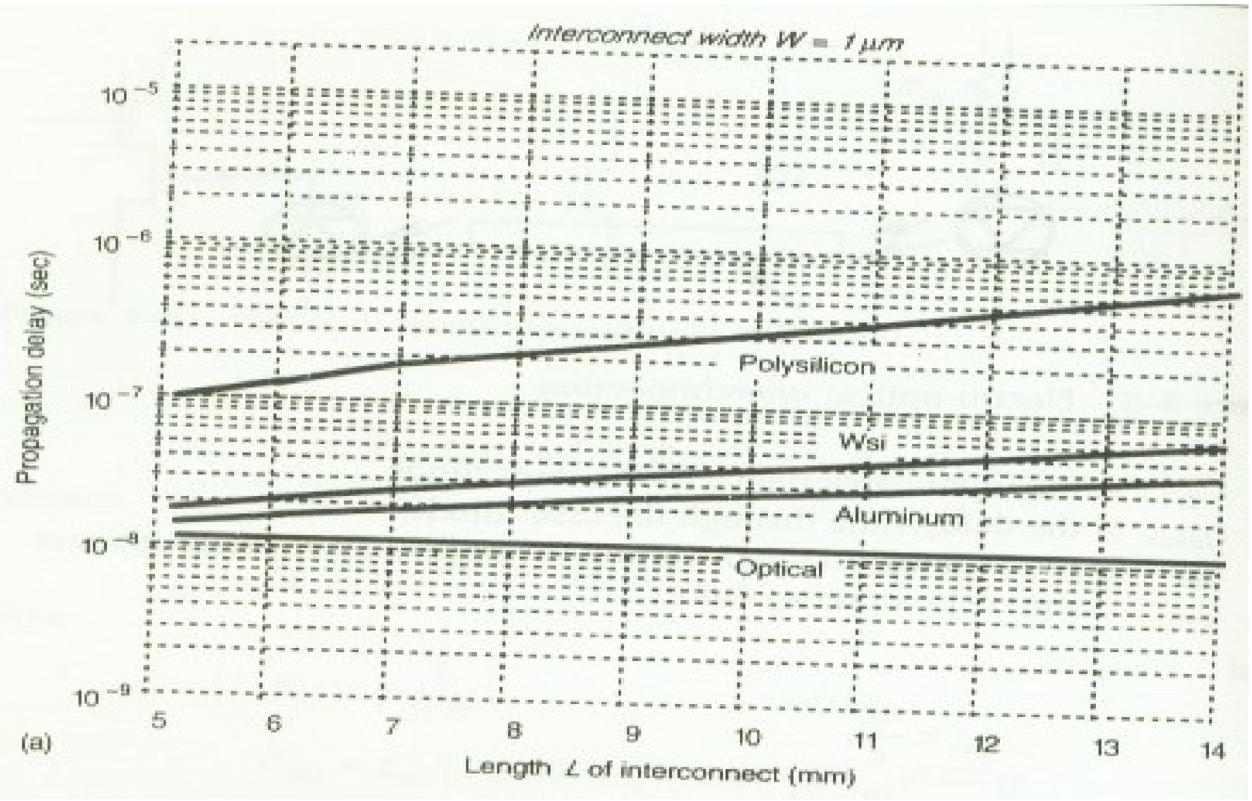


Figure-15:Technology generation

8.5 Limits due to subthreshold currents

- Major concern in scaling devices.
- I $_{sub}$ is directly praportinal exp (Vgs – Vt) q/KT
- As voltages are scaled down, ratio of Vgs-Vt to KT will reduce-so that threshold current increases.
- Therefore scaling Vgs and Vt together with Vdd .
- Maximum electric field across a depletion region is

$$E_{\max} = 2\{V_a + V_b\} / d$$

8.6 Limits on supply voltage due to noise

Decreased inter-feature spacing and greater switching speed –result in noise problems

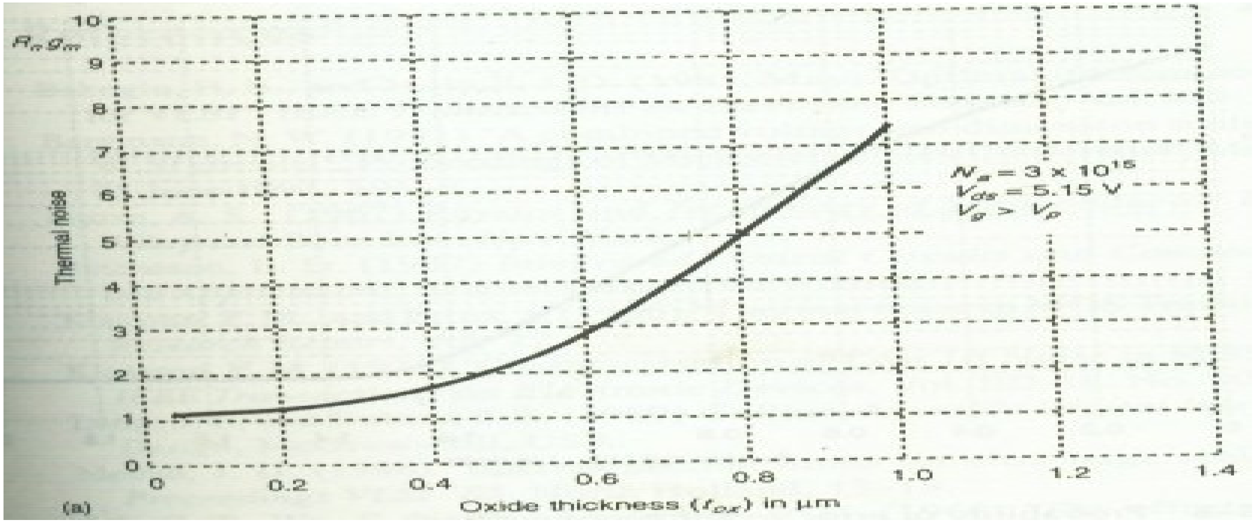


Figure-16:Technology generation

9. Observations – Device scaling

- ○ 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

9.1 Observations – Interconnect scaling

- ○ Capacitance per micron is remaining constant
  - ○ About 0.2 fF/mm
  - ○ 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

***10. Sum*mary**

- Scaling allows people to build more complex machines

  – That run faster too
- It does not to first order change the difficulty of module design

– Module wires will get worse, but only slowly

– You don’t think to rethink your wires in your adder, memory

Or even your super-scalar processor core

- It does let you design more modules
- Continued scaling of uniprocessor performance is getting hard

-Machines using global resources run into wire limitations

-Machines will have to become more explicitly parallel