CS184a: Computer Architecture (Structure and Organization)

Day 6: January 22, 2003
VLSI Scaling

Today

- VLSI Scaling Rules
- Effects
- Historical/predicted scaling
- Variations (cheating)
- Limits

Why Care?

- In this game, we must be able to predict the future
- Rapid technology advance
- Reason about changes and trends
- re-evaluate prior solutions given technology at time X.

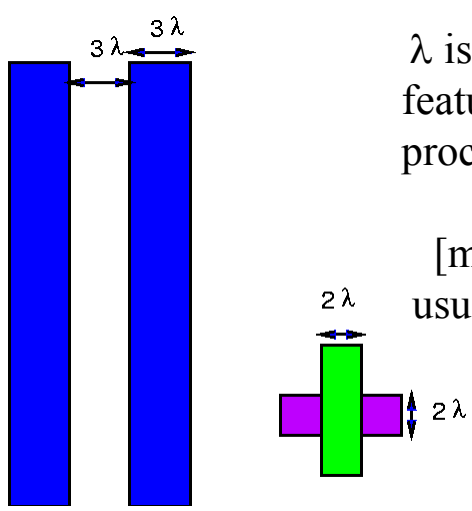
Why Care

- Cannot compare against what competitor does today
  - but what they can do at time you can ship
- Careful not to fall off curve
  - lose out to someone who can stay on curve

Scaling

- **Premise:** features scale "uniformly"
  - everything gets better in a predictable manner
- **Parameters:**
  - λ (lambda) -- Mead and Conway (class)
  - S -- Bohr
  - $1/\kappa$ -- Dennard

Feature Size

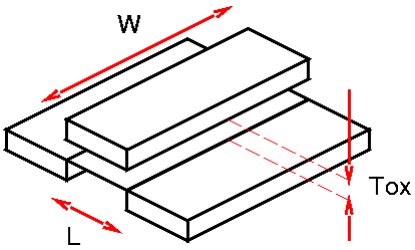


λ is half the minimum feature size in a VLSI process

[minimum feature usually channel width]

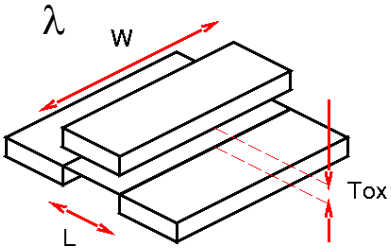
Scaling

- Channel Length (L)
- Channel Width (W)
- Oxide Thickness (T_{ox})
- Doping (N_a)
- Voltage (V)



Scaling

- Channel Length (L) λ
- Channel Width (W) λ
- Oxide Thickness (T_{ox}) λ
- Doping (N_a) $1/\lambda$
- Voltage (V) λ



Effects?

- Area
- Capacitance
- Resistance
- Threshold (V_{th})
- Current (I_d)
- Gate Delay (τ_{gd})
- Wire Delay (τ_{wire})
- Power

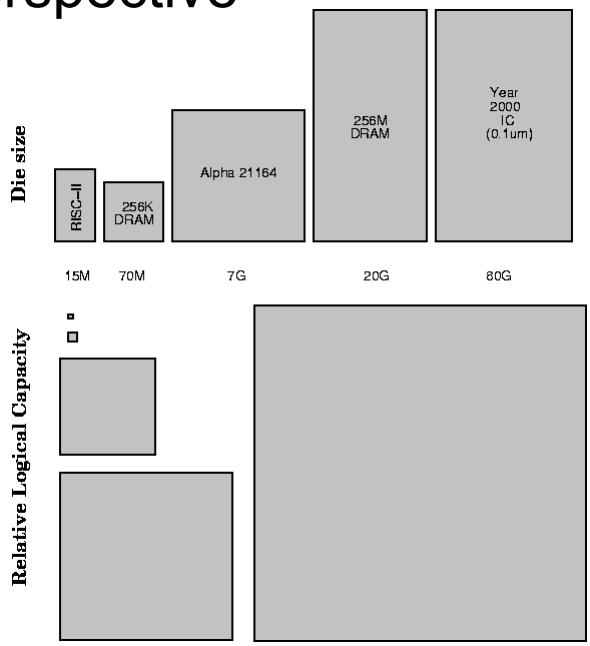
Area

- $\lambda \rightarrow \lambda/\kappa$
- $A = L * W$
- $A \rightarrow A/\kappa^2$

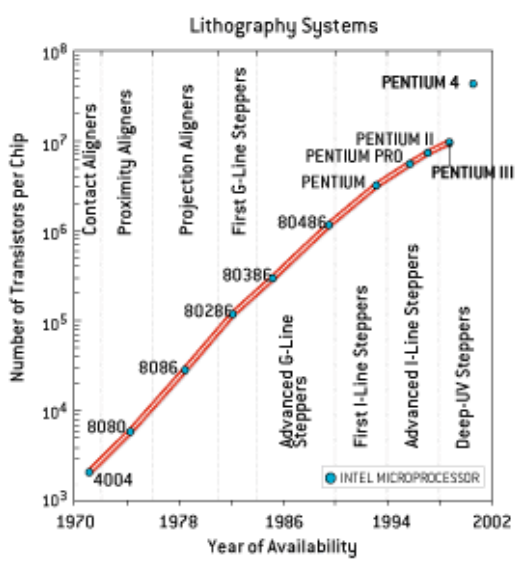
- $0.35\mu m \rightarrow 0.25\mu m$
- 50% area
- 2x capacity same area

Area Perspective

[2000 tech.]
18mm×18mm
0.18μm
60G λ^2

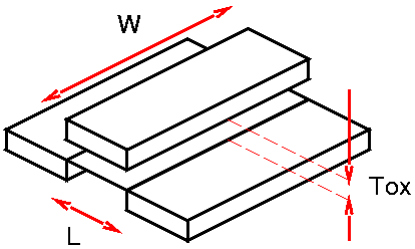


Capacity Scaling from Intel



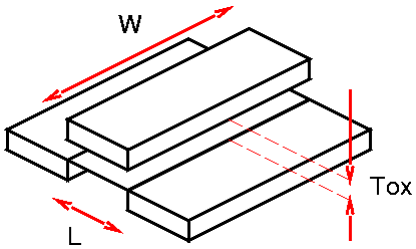
Capacitance

- Capacitance per unit area
  - $C_{ox} = \varepsilon_{SiO_2}/T_{ox}$
  - $T_{ox} \rightarrow T_{ox}/\kappa$
  - $C_{ox} \rightarrow \kappa\, C_{ox}$



Capacitance

- Gate Capacitance
  - $C_{gate} = A * C_{ox}$
  - $A \rightarrow A/\kappa^2$
  - $C_{ox} \rightarrow \kappa\, C_{ox}$
  - $C_{gate} \rightarrow C_{gate}/\kappa$



Threshold Voltage

Before:

$$V_{th} = \frac{1}{C_{OX}}\left(-Q_{eff} + \left(2\epsilon_{Si} q N_a \left(\phi_s + V_{s\text{-}sub}\right)\right)^{1/2}\right) + \left(W_f + \phi_s\right)$$

$$\left(W_f + \phi_s\right) \approx 0$$

adjust $V_{s\text{-}sub}$ **so** $\left(\phi_s + V_{s\text{-}sub}\right) \rightarrow \frac{\left(\phi_s + V_{s\text{-}sub}\right)}{\kappa}$

After:

$$V'_{th} = \frac{1}{\kappa C_{OX}}\left(-Q_{eff} + \left(2\epsilon_{Si} q \kappa N_a \frac{\left(\phi_s + V_{s\text{-}sub}\right)}{\kappa}\right)^{1/2}\right)$$

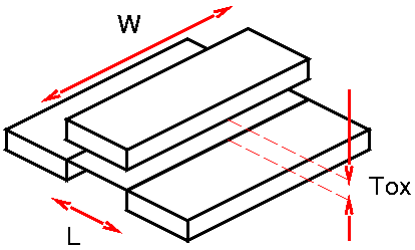
$$V'_{th} \approx \frac{V_{th}}{\kappa}$$

Threshold Voltage

- $V_{TH} \rightarrow V_{TH}/\kappa$

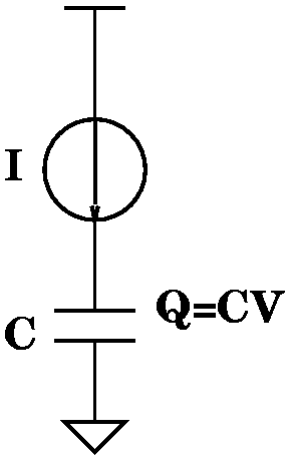
Current

- Saturation Current
  - $I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$
  - $V_{gs=}V \rightarrow V/\kappa$
  - $V_{TH} \rightarrow V_{TH}/\kappa$
  - $W \rightarrow W/\kappa$
  - $C_{ox} \rightarrow \kappa C_{ox}$
  - $I_d \rightarrow I_d/\kappa$



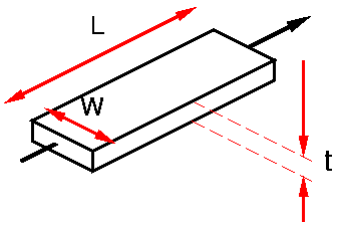
Gate Delay

- $\tau_{gd}=Q/I=(CV)/I$
- $V \rightarrow V/\kappa$
- $I_d \rightarrow I_d/\kappa$
- $C \rightarrow C/\kappa$
- $\tau_{gd} \rightarrow \tau_{gd}/\kappa$



Resistance

- $R=\rho L/(W*t)$
- $W \rightarrow W/\kappa$
- L, t similar
- $R \rightarrow \kappa R$



Wire Delay

- $\tau_{wire}=R\times C$
- $R \rightarrow \kappa R$
- $C \rightarrow C/\kappa$
- $\tau_{wire} \rightarrow \tau_{wire}$

- …assuming (logical) wire lengths remain constant...
- Assume short wire or buffered wire
- (unbuffered wire ultimately scales as length squared)

Power Dissipation (Static)

- Resistive Power
  - $P=V*I$
  - $V \rightarrow V/\kappa$
  - $I_d \rightarrow I_d/\kappa$
  - $P \rightarrow P/\kappa^2$

Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - $P=(1/2)CV^2f$
  - $V \rightarrow V/\kappa$
  - $C \rightarrow C/\kappa$
  - $P \rightarrow P/\kappa^3$
- Increase Frequency?
  - $f \rightarrow \kappa f$?
  - $P \rightarrow P/\kappa^2$

Effects?

- Area $1/\kappa^2$
- Capacitance $1/\kappa$
- Resistance κ
- Threshold (V_{th}) $1/\kappa$
- Current (I_d) $1/\kappa$
- Gate Delay (τ_{gd}) $1/\kappa$
- Wire Delay (τ_{wire}) 1
- Power $1/\kappa^2 \rightarrow 1/\kappa^3$

ITRS Roadmap

- Semiconductor Industry rides this scaling curve
- Try to predict where industry going
  - (requirements…self fulfilling prophecy)
- http://public.itrs.net

MOS Transistor Scaling (1974 to present)

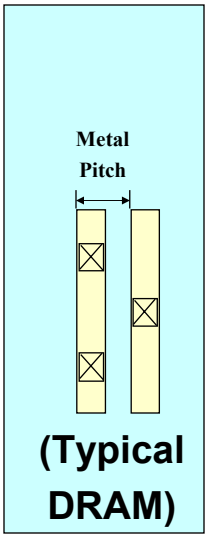
S=0.7

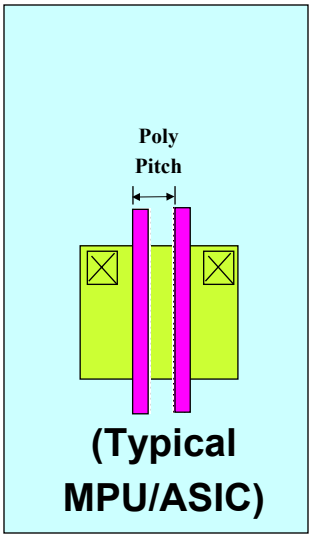
[0.5x per 2 nodes]

Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

Half Pitch (= Pitch/2) Definition

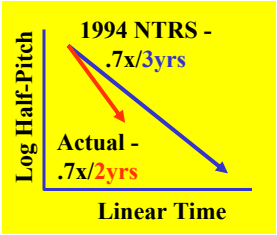




Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

Scaling Calculator + Node Cycle Time:



0.7x 0.7x

250 -> 180 -> 130 -> 90 -> 65 -> 45 -> 32 -> 22 -> 16

0.5x

N N+1 N+2

Node Cycle Time
(T yrs):

*CARR(T) =

[(0.5)^(1/2T yrs)] - 1

CARR(3 yrs) = -10.9%

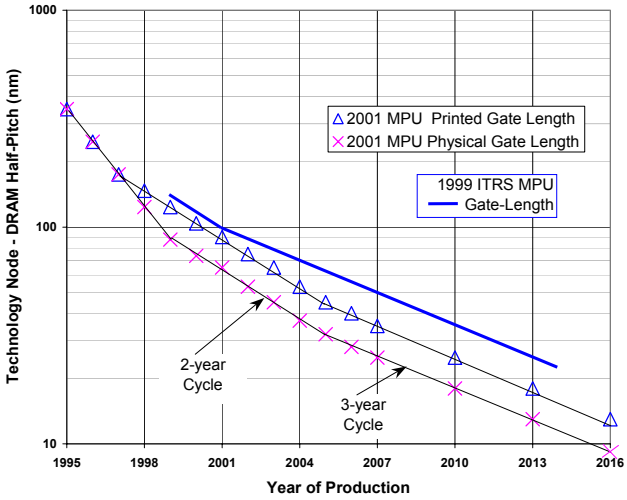
CARR(2 yrs) = -15.9%

* CARR(T) = Compound Annual Reduction Rate
(@ cycle time period, T)

Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

ITRS Roadmap Acceleration Continues…Gate Length



Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

Delays?

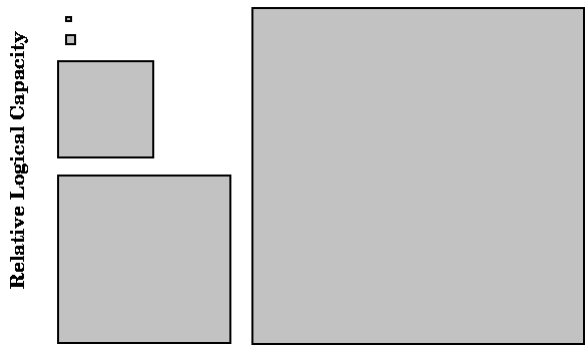
- If delays in gates/switching?
- If delays in interconnect?
- Logical interconnect lengths?

Delays?

- If delays in gates/switching?
  - Delay reduce with $1/\kappa$ [λ]

Delays

- Logical capacities growing
- Wirelengths?
  - No locallity→κ (slower!)
  - Rent's Rule
    - $L \rightarrow n^{(p-0.5)}$
    - [p>0.5]



Compute Density

- Density = compute / (Area * Time)
- κ^3>compute density scaling>κ
- κ^3: gates dominate, p<0.5
- κ^2: moderate p, good fraction of gate delay
  - [p from Rent's Rule again – more on Day12]
- κ : large p (wires dominate area and delay)

Power Density

- P–> P /κ^2 (static, or increase frequency)
- P–> P/κ^3 (dynamic, same freq.)

□ A –> A/κ^2

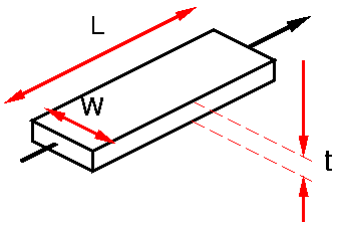
- P/A → P/A … or … P/κA

Cheating…

- Don't like some of the implications
  - High resistance wires
  - Higher capacitance
  - Need for more wiring
  - Not scale speed fast enough

Improving Resistance

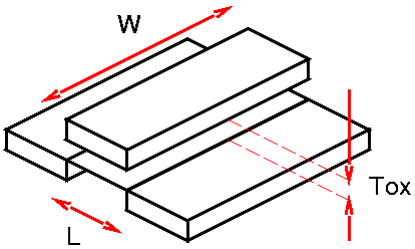
- $R=\rho L/(W*t)$
- $W \rightarrow W/\kappa$
- L, t similar
- $R \rightarrow \kappa R$



- Don't scale t quite as fast.
- Decrease ρ (copper)

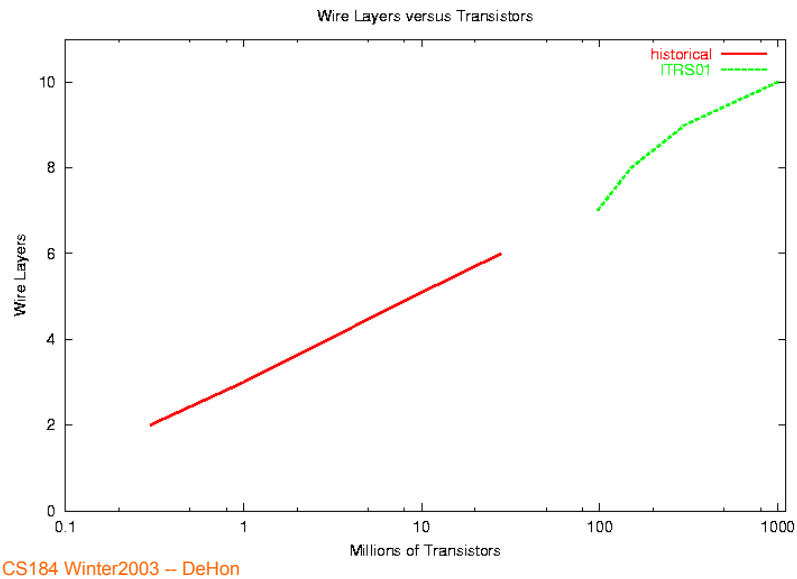
Improving Capacitance

- Capacitance per unit area
  - $C_{ox} = \varepsilon_{SiO_2}/T_{ox}$
  - $T_{ox} \rightarrow T_{ox}/\kappa$
  - $C_{ox} \rightarrow \kappa C_{ox}$

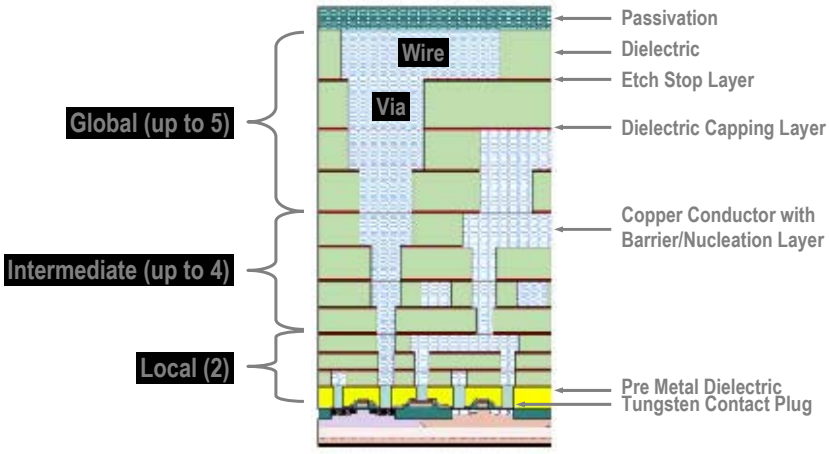


Reduce Dielectric Constant ε

Wire Layers = More Wiring



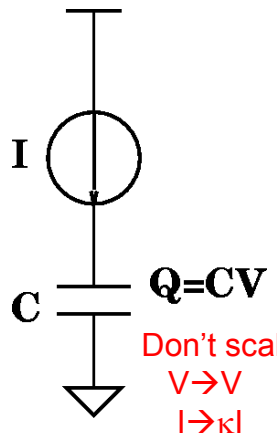
Typical chip cross-section illustrating hierarchical scaling methodology



[from Andrew Kahng]

Improving Gate Delay

- $\tau_{gd}=Q/I=(CV)/I$
- $V \rightarrow V/\kappa$
- $I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$
- $I_d \rightarrow I_d/\kappa$
- $C \rightarrow C/\kappa$
- $\tau_{gd} \rightarrow \tau_{gd}/\kappa$



- Lower C.
- Don't scale V.

Don't scale V:
$V \rightarrow V$
$I \rightarrow \kappa I$
$\tau_{gd} \rightarrow \tau_{gd}/\kappa^2$

…But Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - $P=(1/2)CV^2f$
  - $V \rightarrow V/\kappa$
  - $C \rightarrow C/\kappa$
  - $P \rightarrow P/\kappa^3$
- Increase Frequency?
  - $f \rightarrow \kappa f$?
  - $P \rightarrow P/\kappa^2$

If not scale V, power dissipation not scale.

…And Power Density

- P→ P (increase frequency)
- P→> P/κ (dynamic, same freq.)

□ A → A/κ^2

- P/A → κP/A … or … κ^2P/A

- Power Density Increases

Physical Limits

- Doping?
- Features?

Physical Limits

- Depended on
  - bulk effects
    - doping
    - current (many electrons)
    - mean free path in conductor
  - localized to conductors
- Eventually
  - single electrons, atoms
  - distances close enough to allow tunneling

What Is A "Red Brick" ?

- Red Brick = ITRS Technology Requirement with <u>no known solution</u>
- Alternate definition: Red Brick = something that REQUIRES billions of dollars in R&D investment

[from Andrew Kahng]

The "Red Brick Wall" - 2001 ITRS vs 1999

Table 1. 2001 Status of Red Brick Wall

Year of production	2001	2003	2005	2007	2010	2016
DRAM half-pitch (nm)	130	100	80	65	45	22
Overlay accuracy (nm)	46	35	28	23	18	9
MPU gate length (nm)	90	65	45	35	25	13
CD control (nm)	8	5.5	3.9	3.1	2.2	1.1
T_{ox} (equivalent) (nm)	1.3-1.6	1.1-1.6	0.8-1.3	0.6-1.1	0.5-0.8	0.4-0.5
Junction depth (nm)	48-95	33-66	24-47	18-37	13-26	7-13
Metal cladding thickness (nm)	16	12	9	7	5	2.5
Intermetal dielectric constant, k	3.0-3.6	3.0-3.6	2.6-3.1	2.3-2.7	2.1	1.8

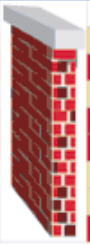


Table 2. 1999 Status of Red Brick Wall

Year of production	1999	2002	2005	2008	2011	2014
DRAM half-pitch (nm)	180	130	100	70	50	35
Overlay accuracy (nm)	65	45	35	25	20	15
MPU gate length (nm)	140	85-90	65	45	30-32	20-22
CD control (nm)	14	9	6	4	3	2
T_{ox} (equivalent) (nm)	1.9-2.5	1.5-1.9	1.0-1.5	0.8-1.2	0.6-0.8	0.5-0.6
Junction depth (nm)	42-70	25-43	20-33	16-26	11-19	8-13
Metal cladding thickness (nm)	17	13	10	0	0	0
Intermetal dielectric constant, k	3.5-4.0	2.7-3.56	1.6-2.2	1.5	<1.5	<1.5

Source: Semiconductor International - http://www.e-insite.net/semiconductor/index.asp?layout=article&articleId=CA187876

 [from Andrew Kahng]

Finishing Up...

Big Ideas [MSB Ideas]

- Moderately predictable VLSI Scaling
  - unprecedented capacities/capability growth for engineered systems
  - **change**
  - be prepared to exploit
  - account for in comparing across time

Big Ideas [MSB-1 Ideas]

- Uniform scaling reasonably accurate for past couple of decades
- Area increase κ^2
  - Real capacity maybe a little less?
- Gate delay decreases ($1/\kappa$)
- Wire delay not decrease, maybe increase
- Overall delay decrease less than ($1/\kappa$)