CS184a: Computer Architecture (Structure and Organization)

Day 6: January 19, 2005

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.

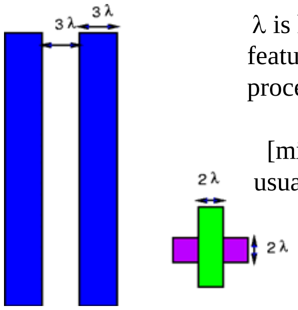
Why Care

- 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

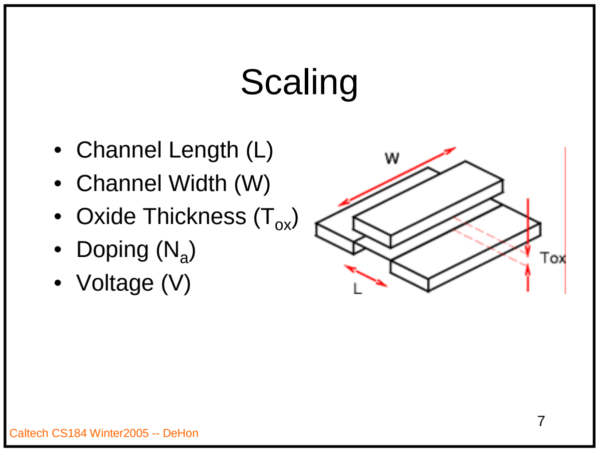
- **Premise:** features scale “uniformly”
  - everything gets better in a predictable manner
- **Parameters:**
  - λ (lambda) -- Mead and Conway (class)
  - S -- Bohr
  - 1/κ -- Dennard

Feature Size



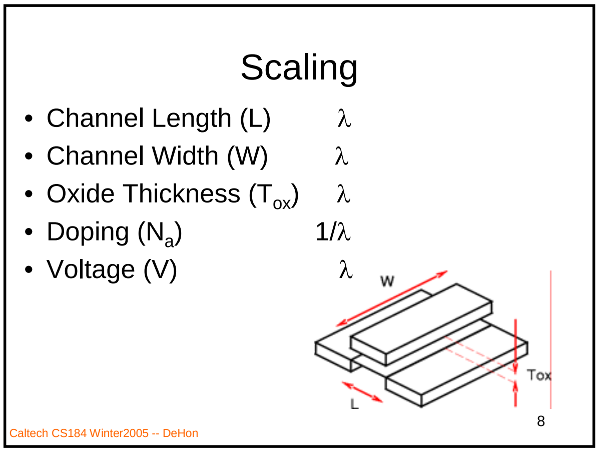
λ is half the minimum feature size in a VLSI process

[minimum feature usually channel width]



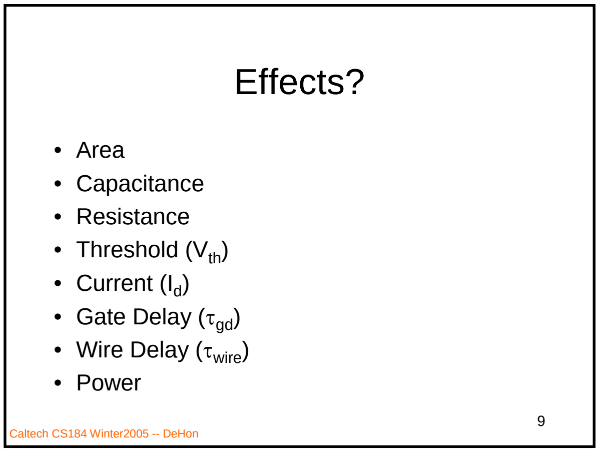
Scaling

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



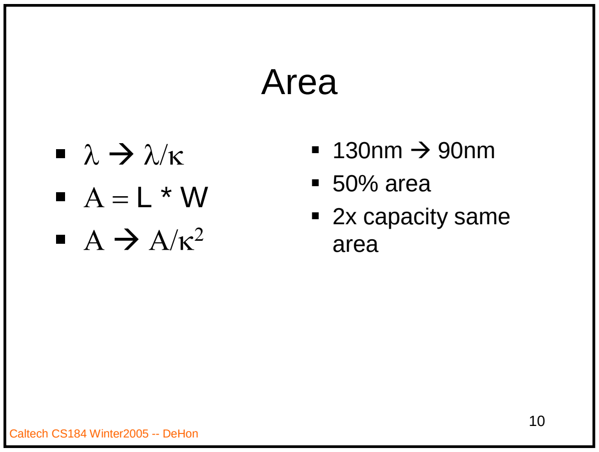
Scaling

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



Effects?

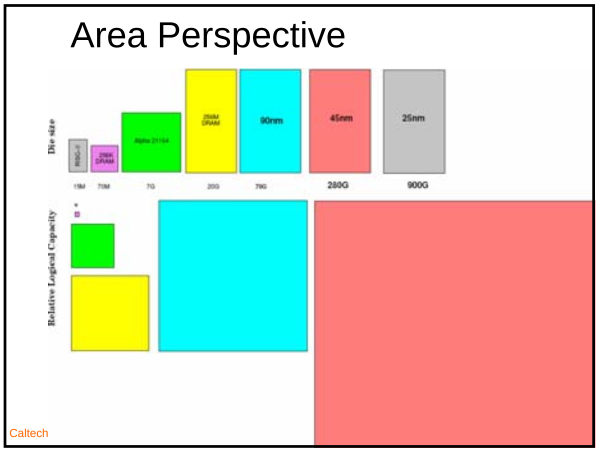
- Area
- Capacitance
- Resistance
- Threshold (V_{th})
- Current (I_d)
- Gate Delay (τ_{gd})
- Wire Delay (τ_{wire})
- Power



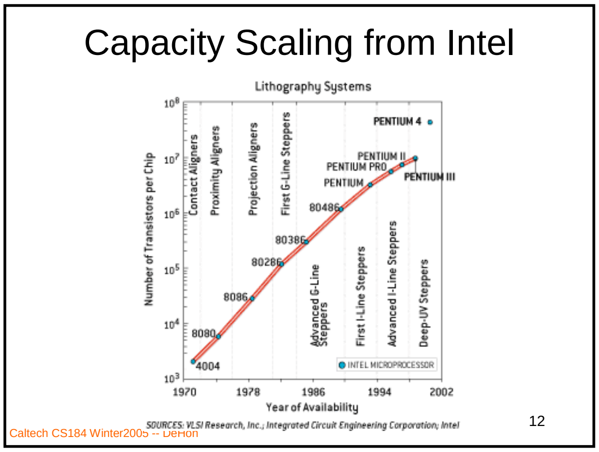
Area

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

- 130nm → 90nm
- 50% area
- 2x capacity same area



Area Perspective

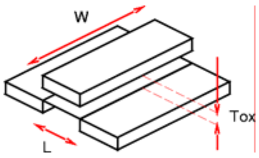


Capacity Scaling from Intel

SOURCES: VLSI Research, Inc.; Integrated Circuit Engineering Corporation; Intel

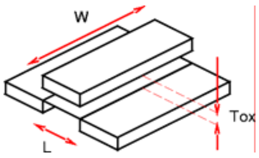
Capacitance

- 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:

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

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

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

After:

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

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

Threshold Voltage

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

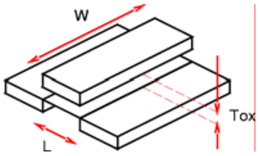
Current

- Saturation Current

  $I_d = (\mu C_{OX}/2)(W/L)(V_{gs} - V_{TH})^2$

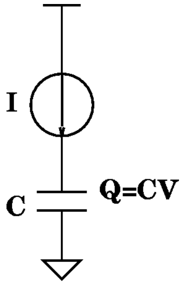
  $V_{gs=}V \rightarrow V/\kappa$
  $V_{TH} \rightarrow V_{TH}/\kappa$
  $W \rightarrow W/\kappa$
  $L \rightarrow L/\kappa$
  $C_{ox} \rightarrow \kappa\ C_{ox}$

  $I_d \rightarrow I_d/\kappa$



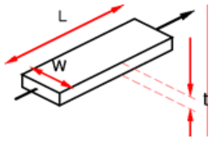
Gate Delay

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



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 Load)

- 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?

  - $\tau_{gd} \rightarrow \tau_{gd}/\kappa$

  - So: $f \rightarrow \kappa f$?

  - $P \rightarrow P/\kappa^2$

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

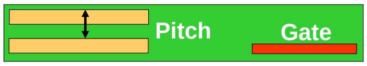
- Semiconductor Industry rides this scaling curve
- Try to predict where industry going
  - (requirements…self fulfilling prophecy)

- http://public.itrs.net

MOS Transistor *Scaling* (1974 to present)

S=0.7

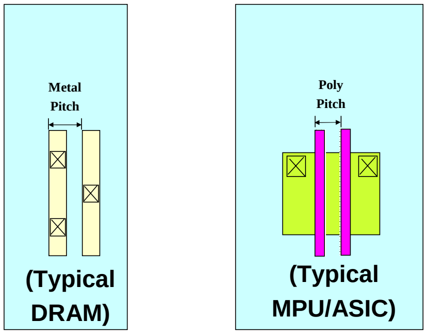
[0.5x per 2 nodes]



Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

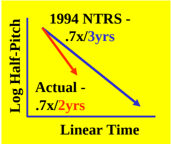
Half Pitch (= Pitch/2) Definition



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

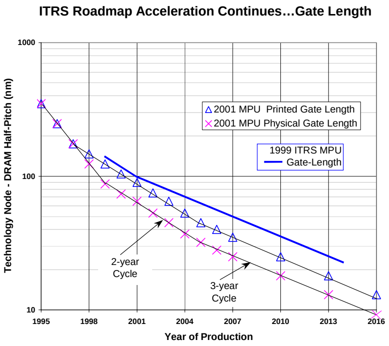
CARR(3 yrs) = -10.9%

CARR(2 yrs) = -15.9%

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

Source: 2001 ITRS - Exec. Summary, ORTC Figure

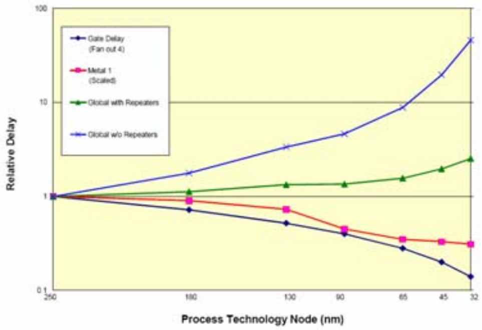
[from Andrew Kahng]



Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

ITRS 2003 Gate/Wire Scaling



What happens to delays?

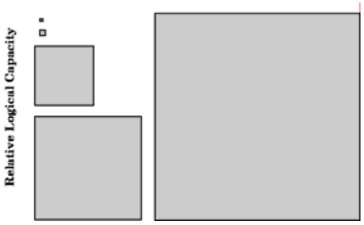
- If delays in gates/switching?
- If delays in interconnect?
- Logical interconnect lengths?

Delays?

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

Delays

- Logical capacities growing
- Wirelengths?
  - No locallity: $L \rightarrow \kappa$ (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

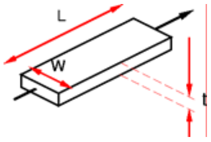
- P–> P $/\kappa^2$ (static, or increase frequency)
- P–> P$/\kappa^3$ (dynamic, same freq.)
- A –> A$/\kappa^2$
- P/A → P/A … or … P/κA

Cheating…

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

Improving Resistance

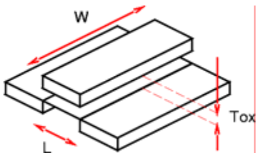
- $R=\rho L/(W*t)$
- W→ W/κ
- L, t similar
- R → κ R



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

Capacitance and Leakage

- 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 (interconnect)

or Substitute for scaling T_{ox} (gate quantum tunneling)

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:

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

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

ITRS 2003

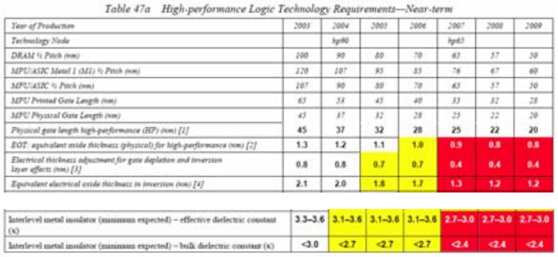


Table 47a High-performance Logic Technology Requirements—Near-term

Year of Production	2003	2004	2005	2006	2007	2008	2009
Technology Node		hp90			hp65		
DRAM ½ Pitch (nm)	100	90	80	70	65	57	50
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)	120	107	95	85	76	67	60
MPU/ASIC ½ Pitch (nm)	107	90	80	70	65	57	50
MPU Printed Gate Length (nm)	65	53	45	40	35	32	28
MPU Physical Gate Length (nm)	45	37	32	28	25	22	20
Physical gate length high-performance (HP) (nm) [1]	45	37	32	28	25	22	20
EOT: equivalent oxide thickness (physical) for high-performance (nm) [2]	1.3	1.2	1.1	1.0	0.9	0.8	0.8
Electrical thickness adjustment for gate depletion and inversion layer effects (nm) [3]	0.8	0.8	0.7	0.7	0.4	0.4	0.4
Equivalent electrical oxide thickness in inversion (nm) [4]	2.1	2.0	1.8	1.7	1.3	1.2	1.2

Interlevel metal insulator (minimum expected) – effective dielectric constant (κ)	3.3–3.6	3.1–3.6	3.1–3.6	3.1–3.6	2.7–3.0	2.7–3.0	2.7–3.0
Interlevel metal insulator (minimum expected) – bulk dielectric constant (κ)	<3.0	<2.7	<2.7	<2.7	<2.4	<2.4	<2.4

Table 81a

High-K dielectric Survey

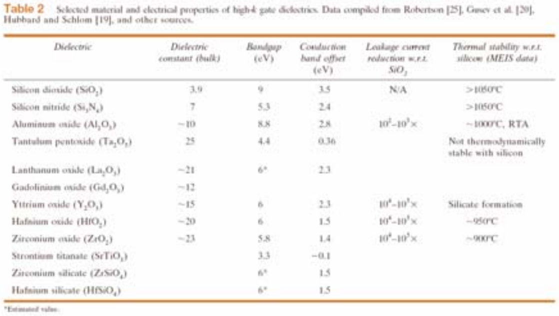


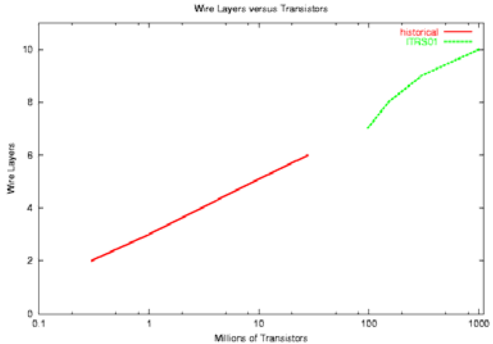
Table 2 Selected material and electrical properties of high-k gate dielectrics. Data compiled from Robertson [25], Gusev et al. [20], Hubbard and Schlom [19], and other sources.

Dielectric	*Dielectric constant (bulk)*	*Bandgap (eV)*	*Conduction band offset (eV)*	*Leakage current reduction w.r.t. SiO₂*	*Thermal stability w.r.t. silicon (MEIS data)*
Silicon dioxide (SiO_2)	3.9	9	3.5	N/A	>1050°C
Silicon nitride (Si_3N_4)	7	5.3	2.4		>1050°C
Aluminum oxide (Al_2O_3)	~10	8.8	2.8	10^2–10^3×	~1000°C, RTA
Tantulum pentoxide (Ta_2O_5)	25	4.4	0.36		Not thermodynamically stable with silicon
Lanthanum oxide (La_2O_3)	~21	6*	2.3		
Gadolinium oxide (Gd_2O_3)	~12				
Yttrium oxide (Y_2O_3)	~15	6	2.3	10^4–10^5×	Silicate formation
Hafnium oxide (HfO_2)	~20	6	1.5	10^4–10^5×	~950°C
Zirconium oxide (ZrO_2)	~23	5.8	1.4	10^4–10^5×	~900°C
Strontium titanate ($SrTiO_3$)		3.3	~0.1		
Zirconium silicate ($ZrSiO_4$)		6*	1.5		
Hafnium silicate ($HfSiO_4$)		6*	1.5		

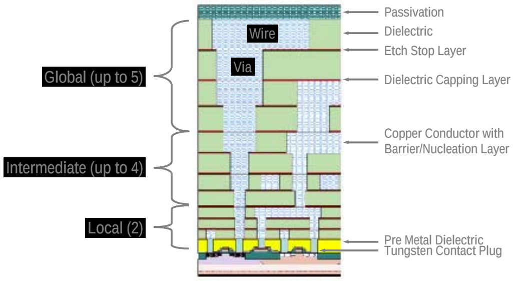
*Estimated value.

Wong/IBM J. of R&D, V46N2/3P133--168

Wire Layers = More Wiring



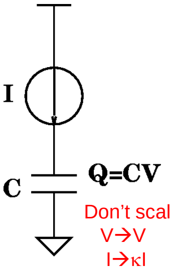
Typical chip cross-section illustrating hierarchical scaling methodology



[from Andrew Kahng]

Improving Gate Delay

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



Don’t scale V:
V→V
I→κI
τ_{gd} → τ_{gd} /κ^2

- Lower C.
- Don’t scale V.

…But
Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - P=(1/2)CV^{2}f
  - V→ V /κ
  - C → C /κ
  - P→ P/κ^3
- Increase Frequency?
  - f → κf ?
  - P → P/κ^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^{2}P/A

- Power Density Increases

…this is where some companies have gotten into trouble…

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 no known solution
- 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

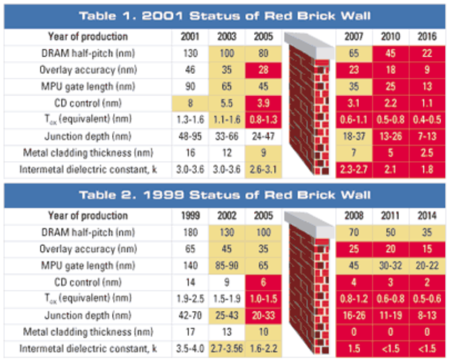


Table 1. 2001 Status of Red Brick Wall

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]

Conventional Scaling

- Ends in your lifetime
- …perhaps in your first few years after grad school…

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
  - …but not for much longer

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 $(1/\kappa)$