ESE680-002 (ESE534): Computer Organization

Day 6: January 29, 2007
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

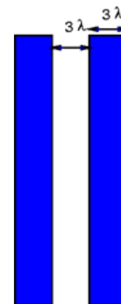
Scaling

- **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

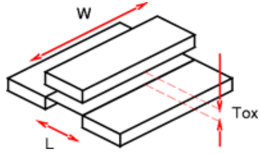
λ is half the minimum feature size in a VLSI process

[minimum feature usually channel width]



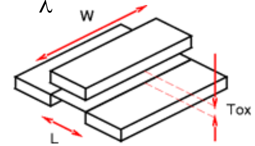
Scaling

- 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) $1/\lambda$
- Voltage (V) λ



Effects?

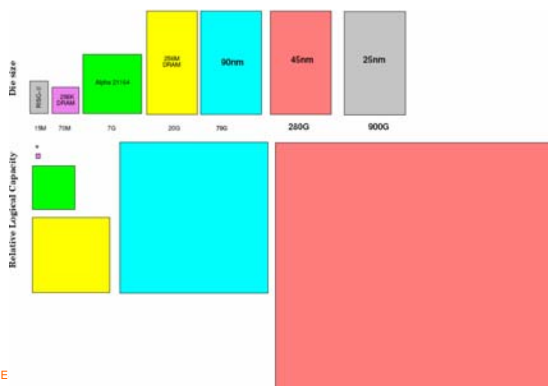
- Area
- Capacitance
- Resistance
- Threshold (V_{th})
- Current (I_d)
- Gate Delay (τ_{gd})
- Wire Delay (τ_{wire})
- Power

Area

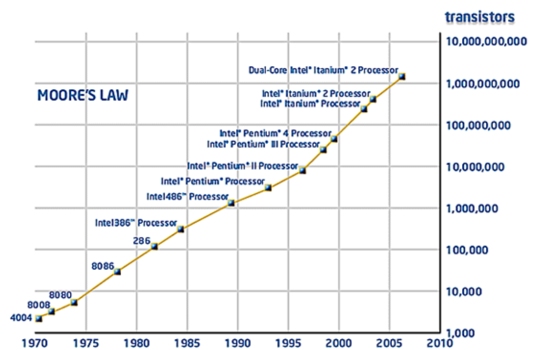
- $\lambda \rightarrow \lambda/\kappa$
- $A = L * W$
- $A \rightarrow A/\kappa^2$

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

Area Perspective



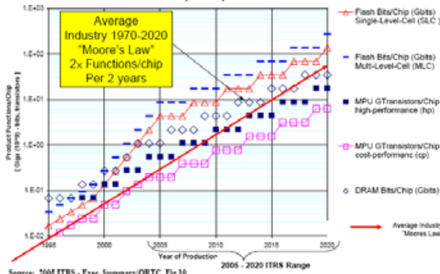
Capacity Scaling from Intel



ITRS 2005 Moore's Law

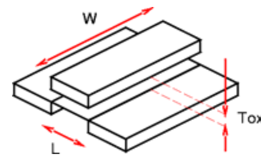
Figure 10 ITRS Product Functions per Chip

2005 ITRS Product Technology Trends - Functions per Chip



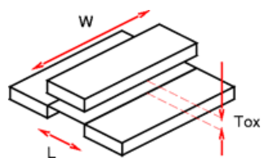
Capacitance

- 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:

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

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

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:

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

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

Threshold Voltage

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

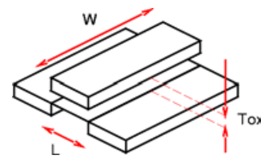
Current

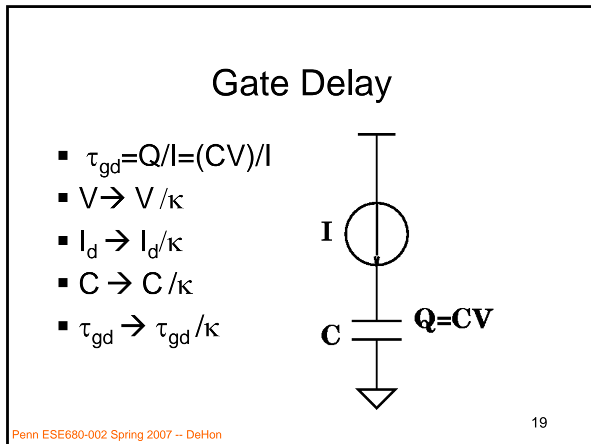
- Saturation Current

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

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

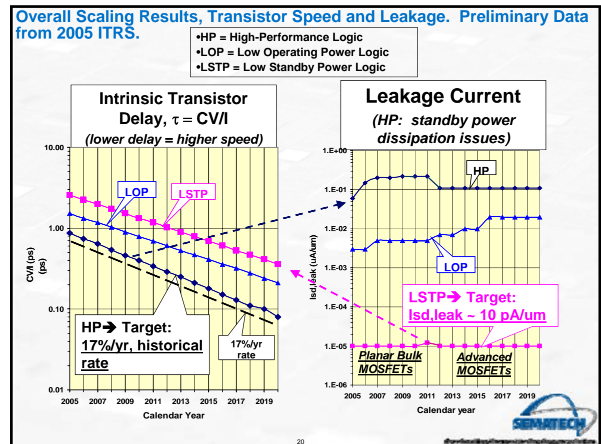
  $I_d \rightarrow I_d/\kappa$

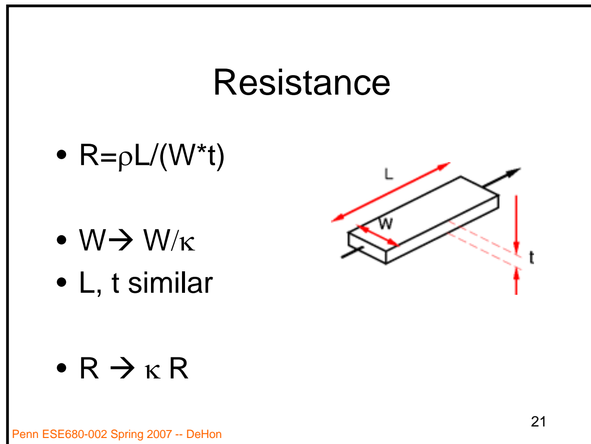




Gate Delay

- τ_{gd}=Q/I=(CV)/I
- V→ V /κ
- I_d → I_d/κ
- C → C /κ
- τ_{gd} → τ_{gd} /κ





Resistance

- R=ρL/(W*t)
- W→ W/κ
- L, t similar
- R → κ R

Wire Delay

- τ_{wire}=R×C
- R –> κ R
- C –> C /κ
- τ_{wire} –> τ_{wire}

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

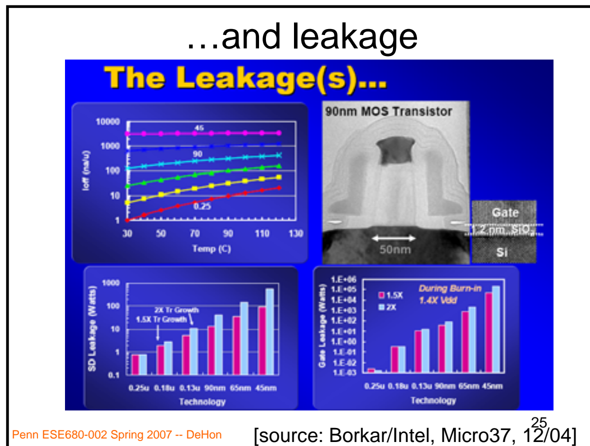
Power Dissipation (Static Load)

- Resistive Power
  - P=V*I
  - V→ V /κ
  - I_d→ I_d/κ
  - P→ P /κ²

Power Dissipation (Dynamic)

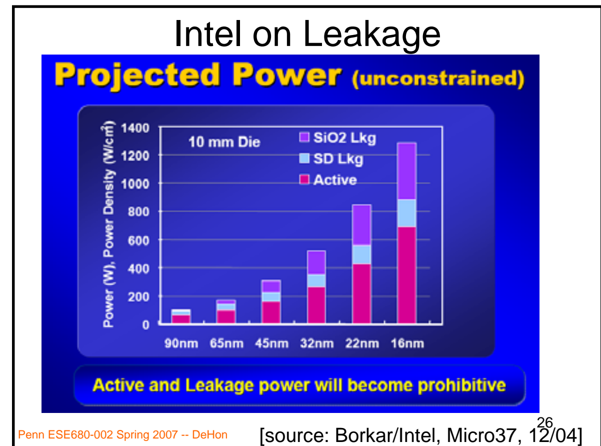
- Capacitive (Dis)charging
  - P=(1/2)CV²f
  - V→ V /κ
  - C → C /κ
  - P→ P/κ³
- Increase Frequency?
  - τ_{gd} → τ_{gd} /κ
  - So: f → κf ?
  - P → P/κ²

…and leakage



Penn ESE680-002 Spring 2007 -- DeHon [source: Borkar/Intel, Micro37, 12/04] 25

Intel on Leakage



Penn ESE680-002 Spring 2007 -- DeHon [source: Borkar/Intel, Micro37, 12/04] 26

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 $1/\kappa^2 \rightarrow 1/\kappa^3$

Penn ESE680-002 Spring 2007 -- DeHon 27

ITRS Roadmap

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

Penn ESE680-002 Spring 2007 -- DeHon 28

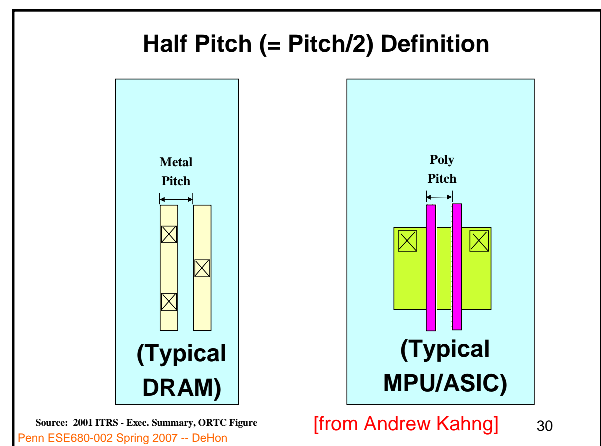
MOS Transistor *Scaling* (1974 to present)

S=0.7

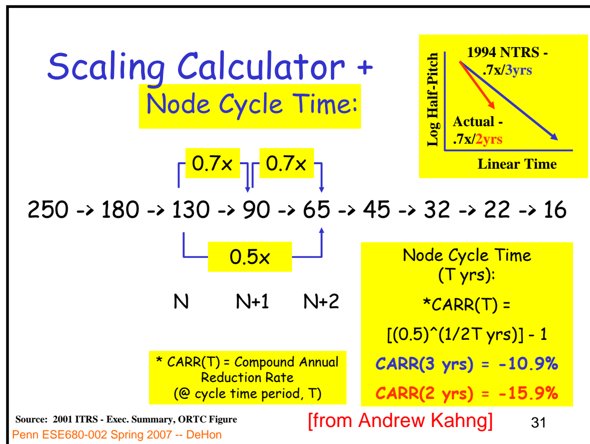
[0.5x per 2 nodes]

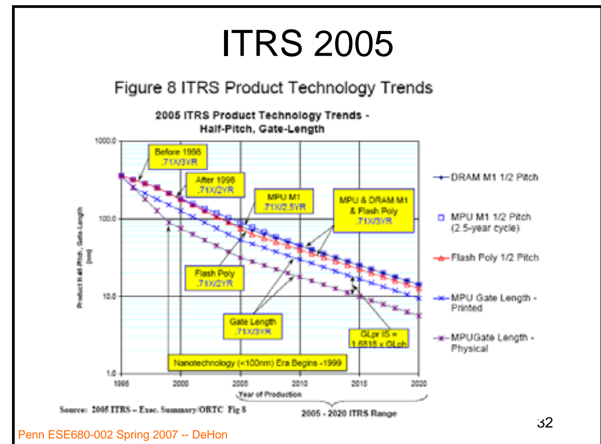
Pitch Gate

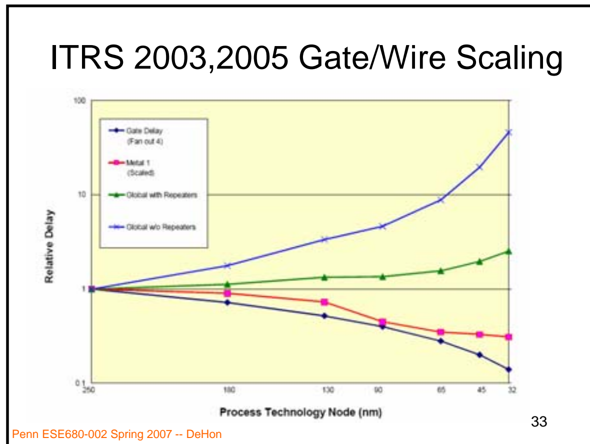
Source: 2001 ITRS - Exec. Summary, ORTC Figure [from Andrew Kahng] 29
Penn ESE680-002 Spring 2007 -- DeHon



Source: 2001 ITRS - Exec. Summary, ORTC Figure [from Andrew Kahng] 30
Penn ESE680-002 Spring 2007 -- DeHon







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 locality: $L \rightarrow \kappa$ (slower!)
  - Rent's Rule
    - $L \rightarrow n^{(p-0.5)}$
    - [p>0.5]

Relative Logical Capacity

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 Day14]
- □ κ : large p (wires dominate area and delay)

Power Density

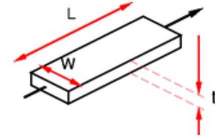
- P–> P /κ^2 (static, or increase frequency)
- P–> P/κ^3 (dynamic, same freq.)
- A –> A/κ^2
- P/A → P/A … or … P/κA

Cheating…

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

Improving Resistance

- R=ρL/(W*t)
- W→ W/κ
- L, t similar
- R → κ R



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

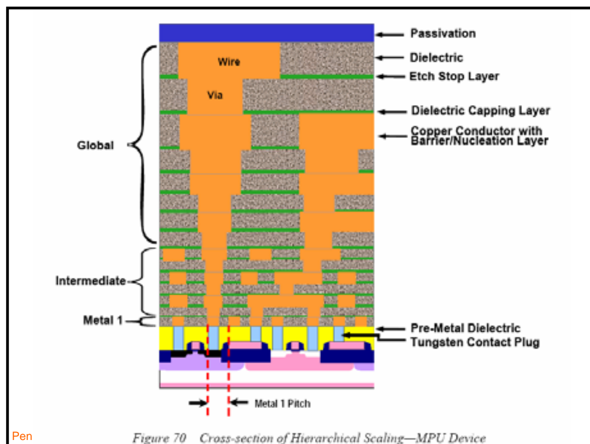
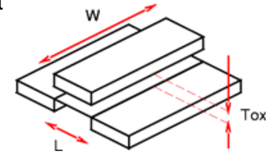


Figure 70 Cross-section of Hierarchical Scaling—MPU Device

Capacitance and Leakage

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



Reduce Dielectric Constant ε (interconnect)

and Increase Dielectric to substitute for scaling T_{ox} (gate quantum tunneling)

Threshold Voltage

Before:

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

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

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:

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

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

ITRS 2005

Table 80a MPU Interconnect Technology Requirements—Near-term Years (continued)

Year of Production	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*
DRAM ½ Pitch (nm) (contacted)	*80*	*70*	*65*	*57*	*50*	*45*	*40*	*36*	*32*
MPU/ASIC Metal 1 ½ Pitch (nm)(contacted)	*90*	*78*	*68*	*59*	*52*	*45*	*40*	*36*	*32*
MPU Physical Gate Length (nm)	*32*	*28*	*25*	*22*	*20*	*18*	*16*	*14*	*13*
Line length (μm) where 25% of switching voltage is induced on victim minimum global wire by crosstalk [4]	170	147	137	130	128	124	120	115	97
Cu thinning of maximum width global wiring due to dishing and erosion (nm), 10% × height, 80% areal density	220	220	230	230	240	240	240	250	250
Cu thinning global wiring due to dishing (nm), 100 μm wide feature	24	21	19	17	15	14	13	13	10
Conductor effective resistivity (μΩ–cm) Cu wiring, assumes no scattering	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2
Interlevel metal insulator – effective dielectric constant (κ)	3.1–3.4	3.1–3.4	2.7–3.0	2.7–3.0	2.5–2.8	2.5–2.8	2.5–2.8	2.1–2.4	2.1–2.4
Interlevel metal insulator (minimum expected) – bulk dielectric constant (κ)	≤ 2.7	≤ 2.7	≤ 2.4	≤ 2.4	≤ 2.2	≤ 2.2	≤ 2.2	≤ 2.0	≤ 2.0

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. SiO_2*	*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
Tantalum 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		

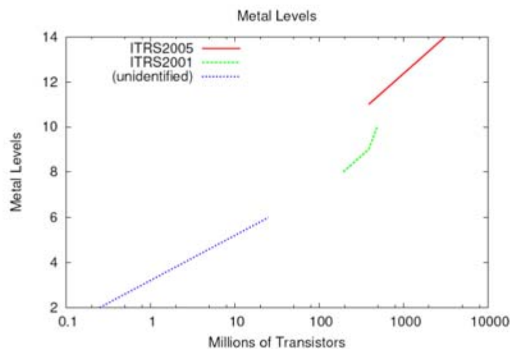
*Estimated value.

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

Intel Saturday NYT Announcement

- **Intel Says Chips Will Run Faster, Using Less Power**
  - NYT 1/27/07, John Markov
  - Claim: "most significant change in the materials used to manufacture silicon chips since Intel pioneered the modern integrated-circuit transistor more than four decades ago"
  - "Intel's advance was in part in finding a new insulator composed of an alloy of hafnium…will replace the use of silicon dioxide."

Wire Layers = More Wiring



Typical chip cross-section illustrating hierarchical scaling methodology

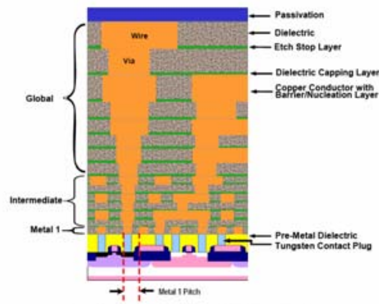
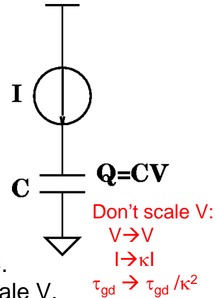


Figure 70 Cross-section of Hierarchical Scaling—MPU Device

[ITRS2005 Interconnect Chapter]

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
  - Lower C.
  - Don't scale V.



Don't scale V:
V→V
I→κI
τ_{gd} → τ_{gd} /κ²

…But Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - P=(1/2)CV²f
  - V→ V /κ
  - C → C /κ
  - P→ P/κ³
- Increase Frequency?
  - f → κf ?
  - P → P/κ²

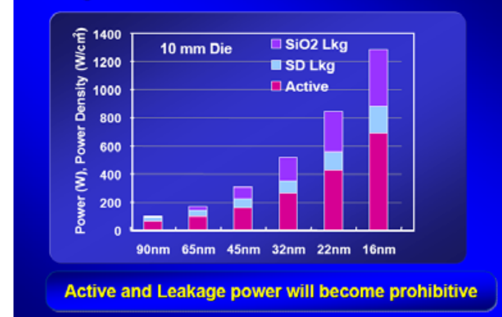
If not scale V, power dissipation not scale.

…And Power Density

- P→ P (increase frequency)
- P→> P/κ (dynamic, same freq.)
- A → A/κ²
- P/A → κP/A … or … κ²P/A
- Power Density Increases

…this is where some companies have gotten into trouble…

Intel on Leakage



[source: Borkar/Intel, Micro37, 12/04]

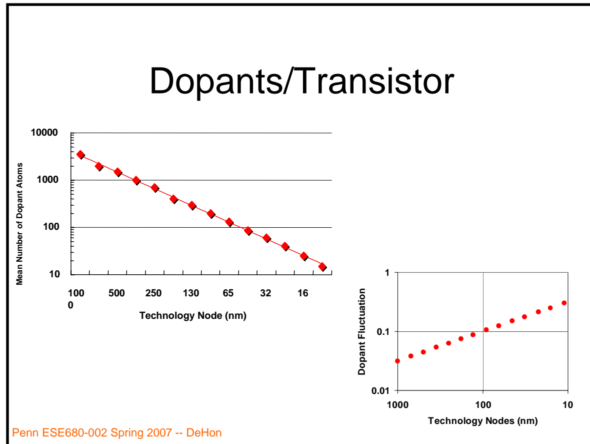
Physical Limits

- 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

Dopants/Transistor



Electrons

Table 40b High-Performance Logic Technology Requirements—Long-term (continued)

Grey cells delineate one of two time periods: either before initial production ramp has started for ultra-thin body fully depleted (UTB FD) SOI or double-gate (DG) MOSFETs, or beyond when planar bulk or UTB FD MOSFETs have reached the limits of practical scaling (see the text and the table notes for further discussion).

Year of Production	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
DRAM ½ Pitch (nm) (contacted)	28	25	22	20	18	16	14
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted)	28	25	22	20	18	16	14
MPU Physical Gate Length (nm)	11	10	9	8	7	6	6
R_{sd}: Effective Parasitic series source/drain resistance [12]							
Planar Bulk (Ω-µm)							
UTB FD (Ω-µm)	75	75					
DG (Ω-µm)	85	80	75	70	65	60	55
Cg ideal: Ideal NMOS Device Gate Capacitance [13]							
Extended Planar Bulk (F/µm)							
UTB FD (F/µm)	4.22E-16	3.83E-16					
DG (F/µm)	3.80E-16	3.45E-16	3.45E-16	3.07E-16	2.68E-16	2.30E-16	1.92E-16
$C_{g,total}$: Total gate capacitance for calculation of CV/I [14]							
Extended Planar Bulk (F/µm)							
UTB FD (F/µm)	5.42E-16	5.03E-16					
DG (F/µm)	5.59E-16	5.25E-16	5.25E-16	4.87E-16	4.48E-16	4.10E-16	3.62E-16

$e = 1.6 \times 10^{-19}$ C

How many electrons?

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]

ITRS 2005 …

Table 40a High-Performance Logic Technology Requirements—Near-term (continued)

Grey cells delineate one of two time periods: either before initial production ramp has started for ultra-thin body fully depleted (UTB FD) SOI or double-gate (DG) MOSFETs, or beyond when planar bulk or UTB FD MOSFETs have reached the limits of practical scaling (see the text and the table notes for further discussion).

Year of Production	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*
DRAM ½ Pitch (nm) (contacted)	80	70	65	57	50	45	40	36	32
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted)	90	78	68	59	52	45	40	36	32
MPU Physical Gate Length (nm)	32	28	25	22	20	18	16	14	13
R_{sd}: Effective Parasitic series source/drain resistance [12]									
Planar Bulk (Ω-µm)	180	170	140	140	120	105	80	70	
UTB FD (Ω-µm)				155	140	125	110	90	75
DG (Ω-µm)							110	100	90
$C_{g,ideal}$: Ideal NMOS Device Gate Capacitance [13]									
Extended Planar Bulk (F/µm)	5.73E-16	5.25E-16	4.69E-16	6.37E-16	6.72E-16	6.78E-16	7.39E-16	6.41E-16	
UTB FD (F/µm)				5.84E-16	5.75E-16	5.65E-16	5.52E-16	5.37E-16	4.98E-16
DG (F/µm)							4.60E-16	4.39E-16	4.48E-16
$C_{g,total}$: Total gate capacitance for calculation of CV/I [14]									
Extended Planar Bulk (F/µm)	8.13E-16	7.65E-16	6.99E-16	8.47E-16	8.42E-16	8.28E-16	8.59E-16	7.51E-16	
UTB FD (F/µm)				8.04E-16	7.55E-16	7.35E-16	6.92E-16	6.67E-16	6.18E-16
DG (F/µm)							6.50E-16	6.29E-16	6.28E-16
τ = CV/I: NMOSFET intrinsic delay (ps) [15]	0.870	0.740	0.640	0.540	0.460	0.400	0.340	0.290	0.250
1/τ: NMOSFET intrinsic switching speed (GHz) [16]	1149	1351	1563	1852	2174	2500	2941	3448	4000

Conventional Scaling

- Ends in your lifetime
- …perhaps in your first few years out of school…
- Perhaps already:
  - "Basically, this is the end of scaling."
    - May 2005, Bernard Meyerson, V.P. and chief technologist for IBM's systems and technology group

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 $(1/\kappa)$