ESE534: Computer Organization

Day 5: January 30, 2012
VLSI Scaling

Today

- VLSI Scaling Rules
- Effects
- Historical/predicted scaling
- Variations (cheating)
- Limits
- **Note:** Most gory MOSFET equations will see → goal is to understand trends
  - Give equations … then push through scaling implications together

Why Care?

- In this game, we must be able to predict the future
- Technology advances rapidly
- Reason about changes and trends
- Re-evaluate prior solutions given technology at time X.
- Make direct comparison across technologies
  - *E.g.* to understand older designs
    - What comes from process vs. architecture

Why Care

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

Preclass

- When will we have 100-core processors
- Bits/DRAM chip in 2024?

Scaling

- **Premise:** features scale "uniformly"
  - everything gets better in a predictable manner
- **Parameters:**
  - λ (lambda) -- Mead and Conway
  - F -- Half pitch – ITRS (F=2λ)
  - S – scale factor – Rabaey
    - $F' = S \times F$

ITRS Roadmap

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

Preclass

- Scale factor S from 45nm → 32nm?

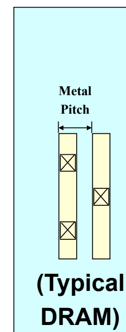
MOS Transistor Scaling (1974 to present)

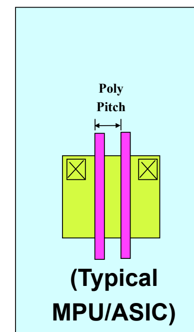
S=0.7

[0.5x per 2 nodes]

Source: 2001 ITRS - Exec. Summary, ORTC Figure [from Andrew Kahng]

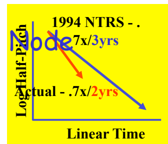
Half Pitch (= Pitch/2) Definition





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%

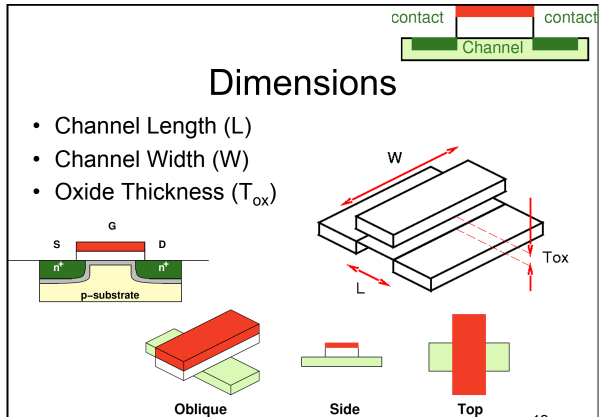
CARR(2 yrs) = -15.9%

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

Source: 2001 ITRS - Exec. Summary, ORTC Figure [from Andrew Kahng]

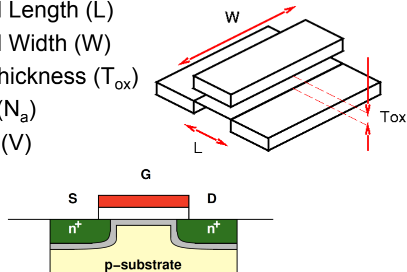
Dimensions

- Channel Length (L)
- Channel Width (W)
- Oxide Thickness (T_{ox})



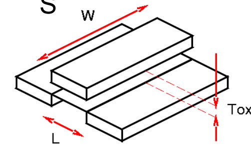
Scaling

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



Full Scaling

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



Effects on Physical Properties?

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

- Go through full (ideal)
- …then come back and ask what still makes sense today.

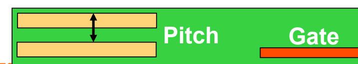
Area

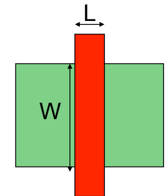
- $\lambda \rightarrow \lambda S$
- Area impact?
- $A = L \times W$
- $A \rightarrow AS^2$

- 45nm → 32nm
- 50% area
- 2× capacity same area

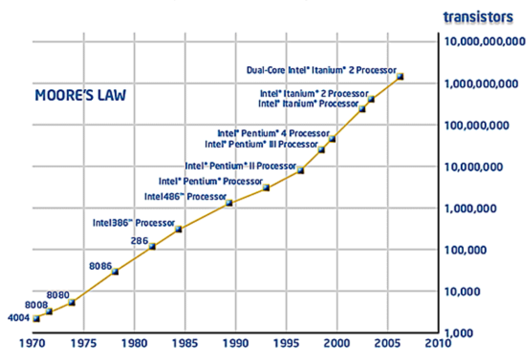
S=0.7

[0.5x per 2 nodes]



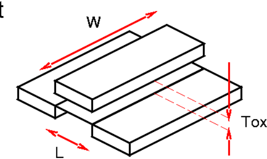


Capacity Scaling from Intel



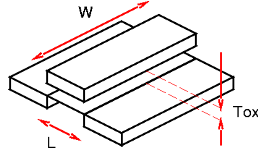
Capacitance

- Capacitance per unit area scaling:
  - $C_{ox} = \varepsilon_{SiO_2}/T_{ox}$
  - $T_{ox} \rightarrow S \times T_{ox}$
  - $C_{ox} \rightarrow C_{ox}/S$



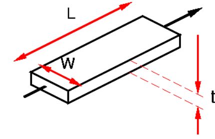
Capacitance

- Gate Capacitance scaling?
  - $C_{gate} = A \times C_{ox}$
  - $A \rightarrow A \times S^2$
  - $C_{ox} \rightarrow C_{ox}/S$
  - $C_{gate} \rightarrow S \times C_{gate}$



Resistance

- Resistance scaling?
- $R = \rho L/(W^*t)$
- $W \rightarrow S \times W$
- L, t similar
- $R \rightarrow R/S$



Threshold Voltage

- $V_{TH} \rightarrow S \times V_{TH}$

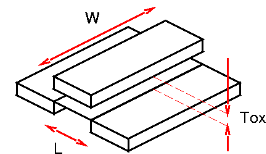
Current

- Saturation Current scaling?

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

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

$I_d \rightarrow S \times I_d$



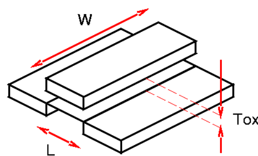
Current

$$V_{DSAT} \approx \frac{L v_{sat}}{\mu_n}$$

- Velocity Saturation Current scaling:

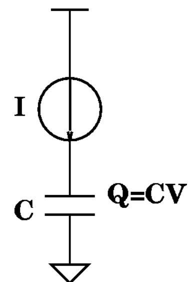
$$I_{DS} \approx v_{sat} C_{OX} W \left(V_{GS} - V_T - \frac{V_{DSAT}}{2} \right)$$

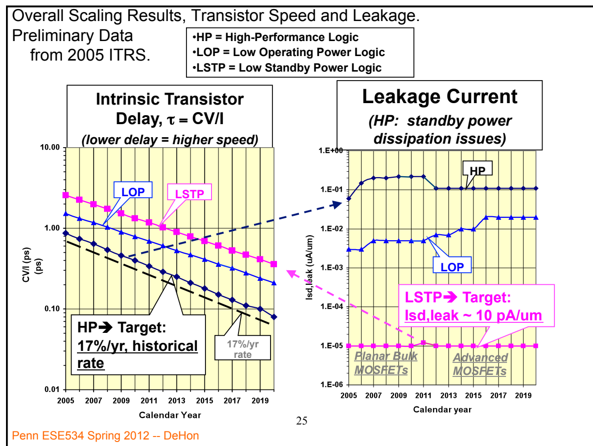
$V_{gs} = V \rightarrow S \times V$
$V_{TH} \rightarrow S \times V_{TH}$
$L \rightarrow S \times L$
$V_{DSAT} \rightarrow S \times V_{DSAT}$
$W \rightarrow S \times W$
$C_{ox} \rightarrow C_{ox}/S$
$I_d \rightarrow S \times I_d$

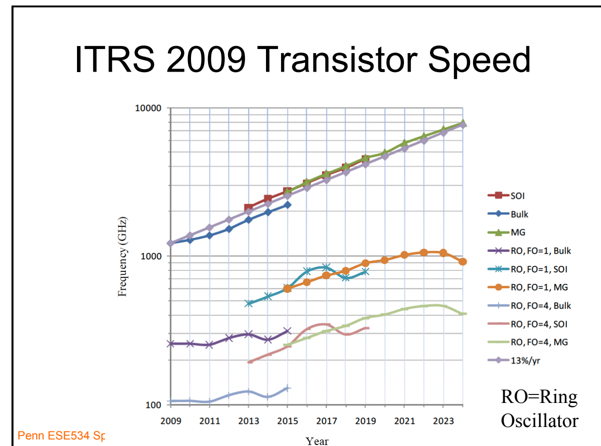


Gate Delay

- Gate Delay scaling?
- $\tau_{gd} = Q/I = (CV)/I$
- $V \rightarrow S \times V$
- $I_d \rightarrow S \times I_d$
- $C \rightarrow S \times C$
- $\tau_{gd} \rightarrow S \times \tau_{gd}$







Wire Delay

- Wire delay scaling?
- τ_{wire}=R×C
- R → R/S
- C → S×C
- τ_{wire} → τ_{wire}

- …assuming (logical) wire lengths remain constant...

27

Impact of Wire and Gate Delay Scaling

- If gate delay scales down but wire delay does not scale, what does that suggest about the relative contribution of gate and wire delays to overall delay as we scale?

28

Energy

- Switching Energy per operation scaling?
- $E=1/2\ CV^2$
- V→ S×V
- C → S×C
- E → S^3×E

29

Power Dissipation (Dynamic)

- Capacitive (Dis) charging scaling?
  - $P=(1/2)CV^2f$
  - V→ S×V
  - C → S×C
  - P→ S^3×P
- Increase Frequency?
  - τ_{gd} → S×τ_{gd}
  - So: f → f/S ?
  - P → S^2×P

30

Effects?

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

Power Density

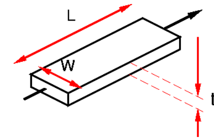
- P → S^2P (increase frequency)
- P → S^3P (dynamic, same freq.)
- A → S^2A
- Power Density: P/A two cases?
  - P/A → P/A increase freq.
  - P/A → S×P/A same freq.

Cheating…

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

Improving Resistance

- $R=\rho L/(W\times t)$
- W → S×W
- L, t similar
- R → R/S



- What might we do?
- Don't scale t quite as fast → now taller than wide.
- Decrease ρ (copper) – introduced 1997

http://www.ibm.com/ibm100/us/en/icons/copperchip/

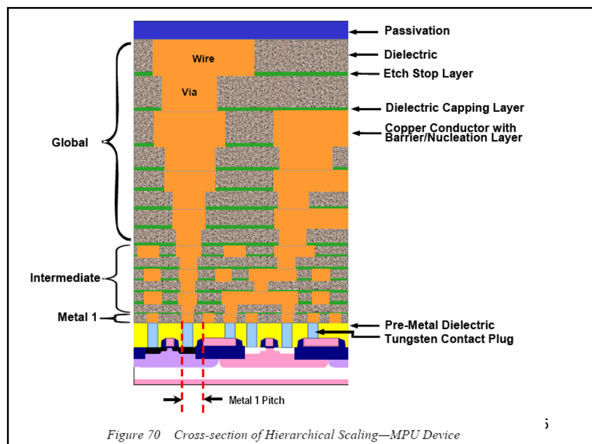
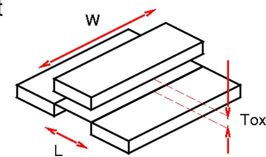
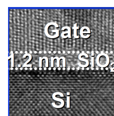


Figure 70 Cross-section of Hierarchical Scaling—MPU Device

Capacitance and Leakage

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



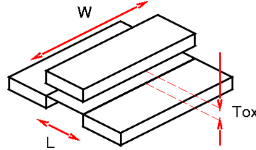


What's wrong with Tox = 1.2nm?

source: Borkar/Micro 2004

Capacitance and Leakage

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



What might we do?

Reduce Dielectric Constant ε (interconnect)

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

ITRS 2009

Table PIDS3B Low Operating Power Technology Requirements

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 multi-gate (MG) 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	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
MPU/ASIC Metal 1 (M1) ½ Pitch (nm) (contacted)	54	45	38	32	27	24	21	18.9	16.9	15	13.4	11.9	10.6	9.5	8.4	7.5
Lg: Physical Lgate for High Performance logic (nm)	29	27	24	22	20	18	17	15.3	14	12.8	11.7	10.7	9.7	8.9	8.1	7.4
L_g: Physical Lgate for Low OperatingPower (LOP) logic (nm) [1]	32	29	27	24	22	18	17	15.3	14	12.8	11.7	10.7	9.7	8.9	8.1	7.4
EOT: Equivalent Oxide Thickness (nm) [2]																
Extended planar bulk	1	0.9	0.9	0.85	0.8											
UTB FD			-	-	0.9	0.85	0.8	0.75	0.7							
MG			-	-	-	-	0.8	0.8	0.75	0.73	0.7	0.7	0.65	0.65	0.6	0.6
Gate poly depletion (nm) [3]																
Bulk	0.27	0.27	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Channel doping (E18 /cm3) [4]																
Extended Planar Bulk	3	3.7	4.5	5	5.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Junction depth or body Thickness (nm) [5]																
Extended Planar Bulk (junction)	14	13	11.5	10	9											
UTB FD (body)					7	6.2	6	5.1	4.7							
MG (body)							8	7.6	7	6.4	5.8	5.4	4.8	4.4	4.2	4
EOT_{elec}: Electrical Equivalent Oxide Thickness (nm) [6]																
Extended Planar Bulk	1.64	1.53	1.23	1.18	1.14											
UTB FD					1.3	1.25	1.2	1.15	1.1							
MG							1.2	1.2	1.15	1.13	1.1	1.1	1.05	1.05	1	1

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–$10^3\times$	~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–$10^5\times$	Silicate formation
Hafnium oxide (HfO_2)	~20	6	1.5	10^4–$10^5\times$	~950°C
Zirconium oxide (ZrO_2)	~23	5.8	1.4	10^4–$10^5\times$	~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.

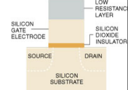
Wong/IBM J. of R&D, V46N2/3P133—168, 2002

Intel NYT Announcement

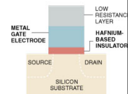
- **Intel Says Chips Will Run Faster, Using Less Power**
  - NYT 1/27/07, John Markov
  - Claim: "most significant change in the materials used to manufacture silicon chips since Intel pioneered the modern integrated-circuit transistor more than four decades ago"
  - "Intel's advance was in part in finding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

As microprocessor transistors become smaller, stopping undesired current leakage becomes more difficult. This leakage leads to shortened battery life. Intel's coming chips use a new insulation material to prevent this, reducing power consumption.

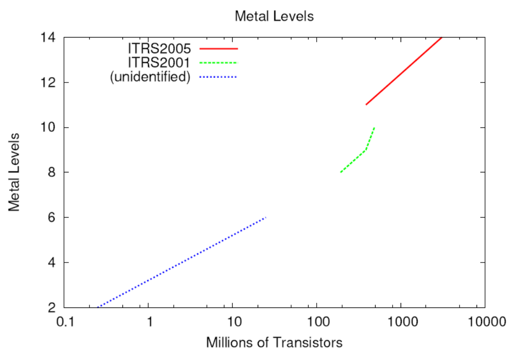
Current transistors use extremely thin silicon dioxide insulators, which lead to current leakage. Thickening them decreases this leakage but reduces the electric charge passing through, impeding performance.



New transistors use a hafnium-based insulator and a metal gate electrode. Hafnium provides stronger electrical coupling, so the insulator can be made thicker to reduce leakage without degrading the performance of the transistor.



Wire Layers = More Wiring



Typical chip cross-section illustrating hierarchical scaling methodology

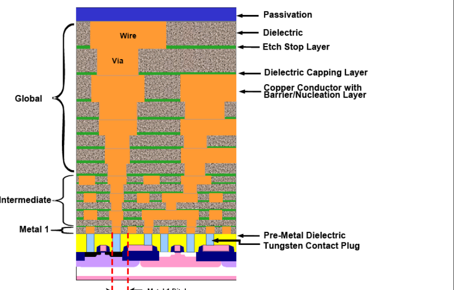


Figure 70 Cross-section of Hierarchical Scaling—MPU Device

[ITRS2005 Interconnect Chapter]

Improving Gate Delay

- τ_{gd}=Q/I=(CV)/I
- V→ S×V
- $I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$
- I_d → $S \times I_d$
- C → S×C
- τ_{gd} → $S \times \tau_{gd}$
  - Lower C.
  - Don't scale V.

How might we accelerate?

Don't scale V:
V→V
I→I/S
τ_{gd} → $S^2 \times \tau_{gd}$

…But Power Dissipation (Dynamic)

- Capacitive (Dis) charging
  - P=(1/2)CV^2f
  - V→ V
  - C → S×C
- Increase Frequency?
  - f → f/S^2 ?
  - P → P/S

If not scale V, power dissipation not scale down.

…And Power Density

- P→ P/S (increase frequency)
- But… A → S^2×A
- What happens to power density?
- P/A → $(1/S^3)$P
- Power Density Increases

…this is where some companies have gotten into trouble…

Historical Voltage Scaling

http://software.intel.com/en-us/articles/gigascale-integration-challenges-and-opportunities/

- Frequency impact?
- Power Density impact?

uProc Clock Frequency

The Future of Computing Performance: Game Over or Next Level?
National Academy Press, 2011

 http://www.nap.edu/catalog.php?record_id=12980

uP Power Density

The Future of Computing Performance: Game Over or Next Level?
National Academy Press, 2011

 http://www.nap.edu/catalog.php?record_id=12980

Power Density Impact

- What is impact on power density growing to hit our practical limit?

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]

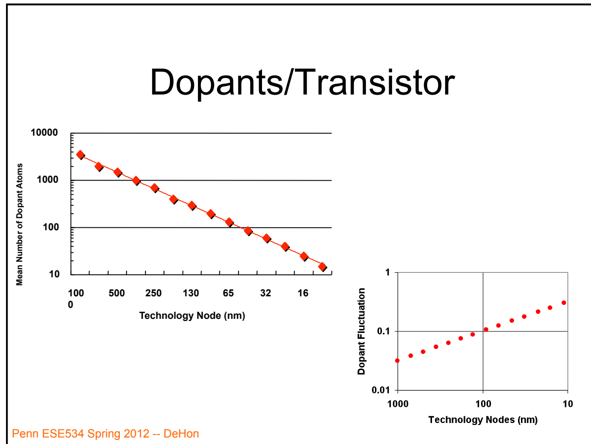
ITRS 2009

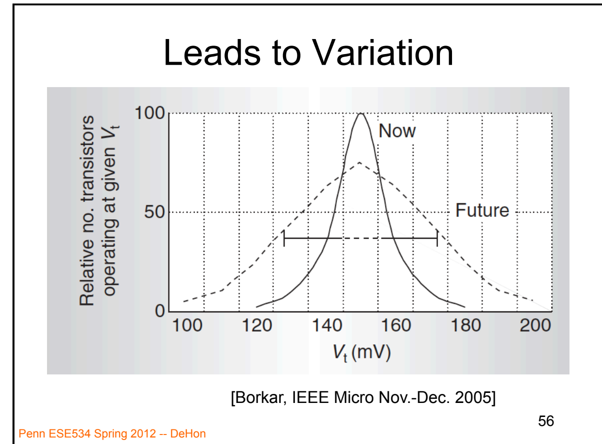
Physical Limits

- 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





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

How many electrons?

$e=1.6\times10^{-19}$ C

F=14nm C_{min}=1.4E-2× 3.6E-16=5E-18=30e

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...

Admin

- HW3 out
- Not graded HW1, yet ☹
- Reading for Wednesday on Web
- No Office Hours this Tuesday
- Guest lecturer on Wednesday
  - Topic likely important to project

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 $1/S^2$
  - Real capacity maybe a little less?
- Gate delay decreases (S)
  - …maybe a little less
- Wire delay not decrease, maybe increase
- Overall delay decrease less than (S)
- Lack of V scale → Power density limit