ESE370: Circuit-Level Modeling, Design, and Optimization for Digital Systems

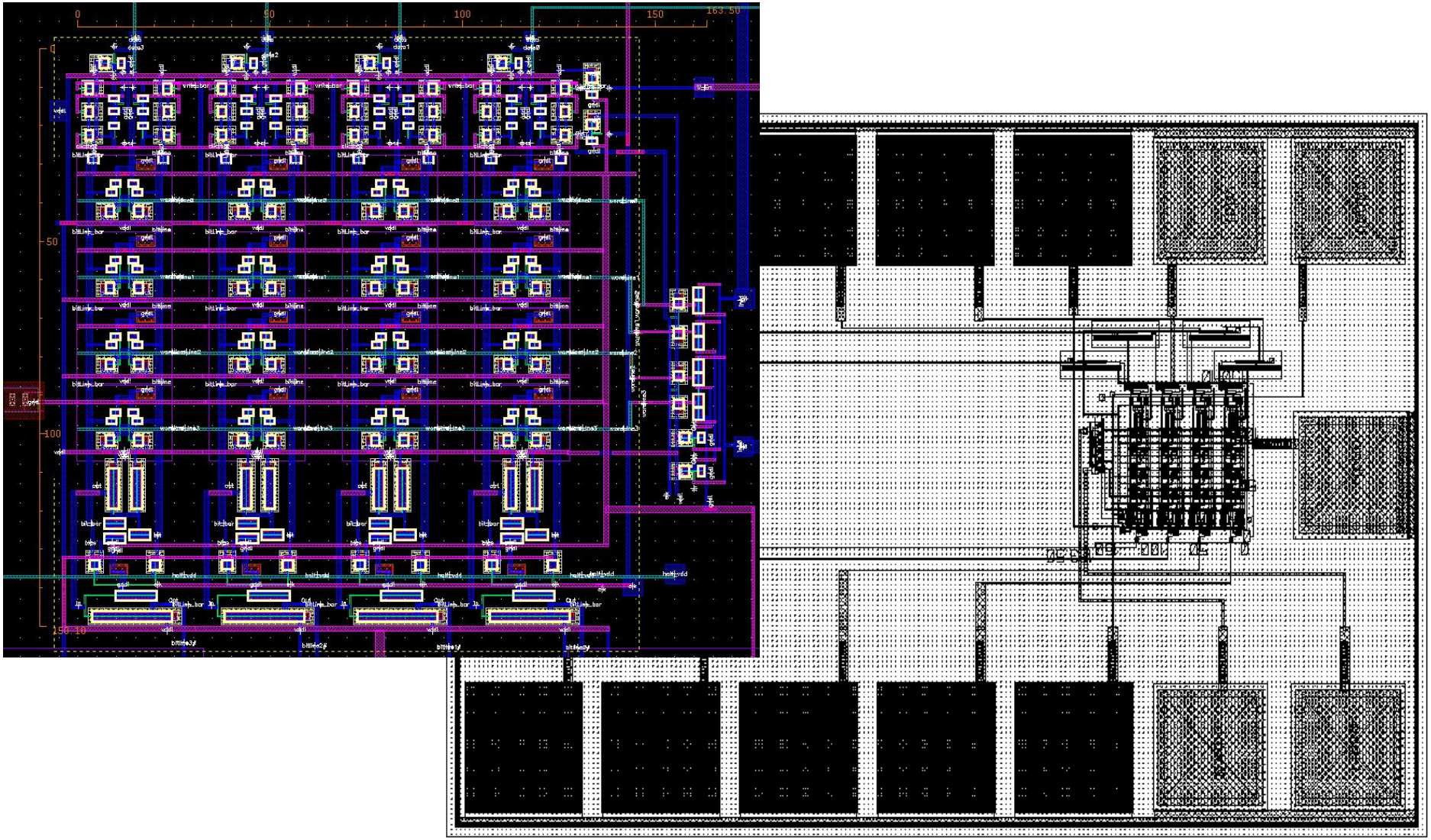
Lec 12: October 4, 2021

Scaling

Today

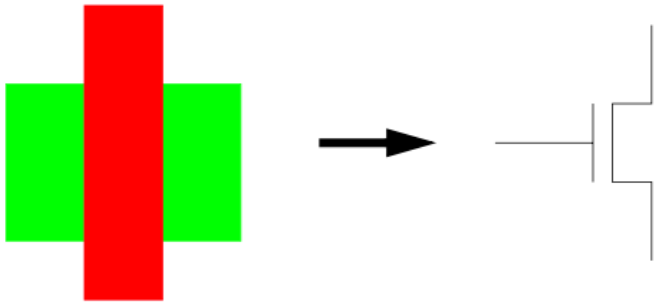
- Layout Roundup
  - Extraction
- VLSI Scaling Trends/Disciplines
- Effects
- Alternatives (cheating)

4x4 6T SRAM Memory



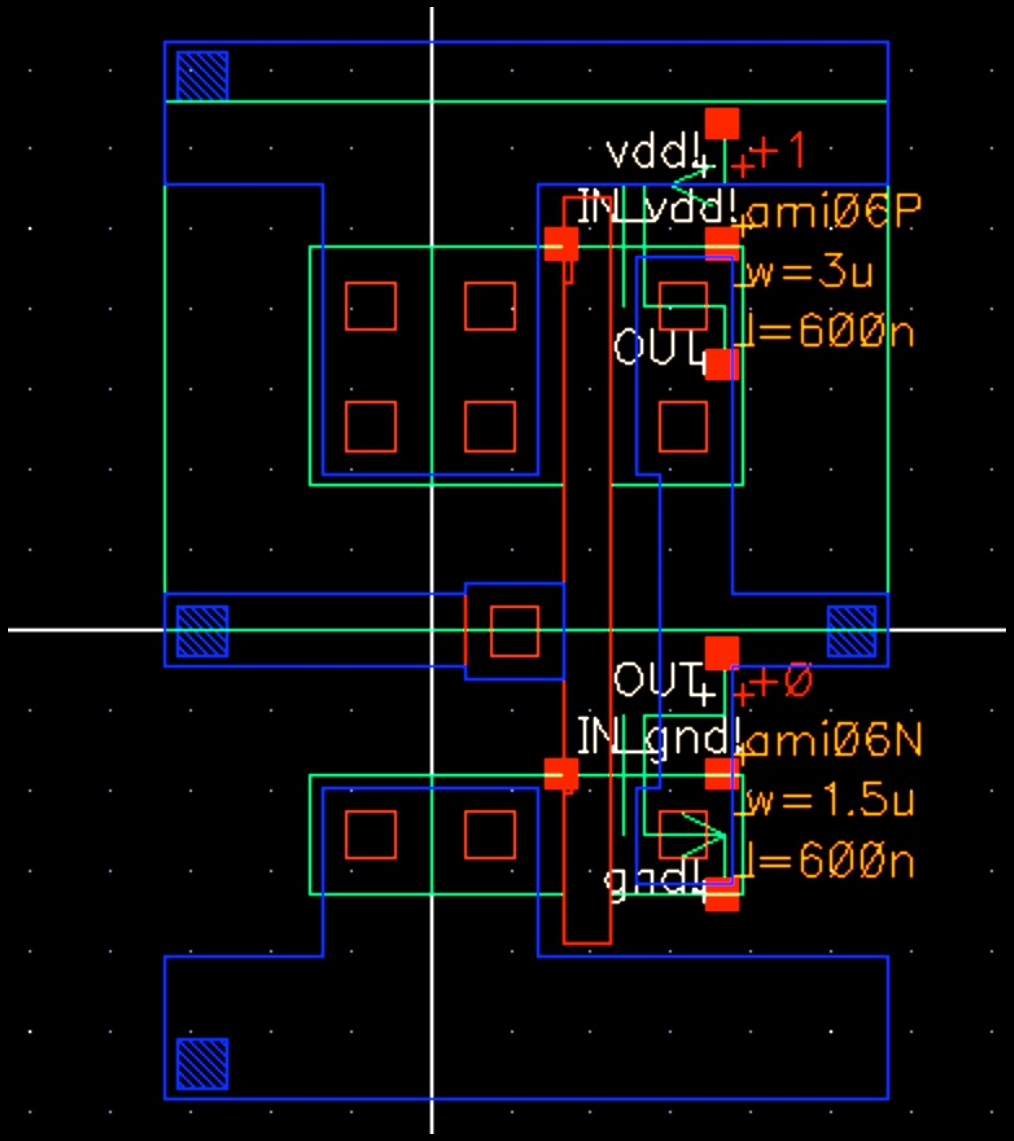
Circuit Extraction

- Circuit extraction extracts a schematic representation of a layout, including transistors, wires, and possibly wire and device resistance and capacitance.

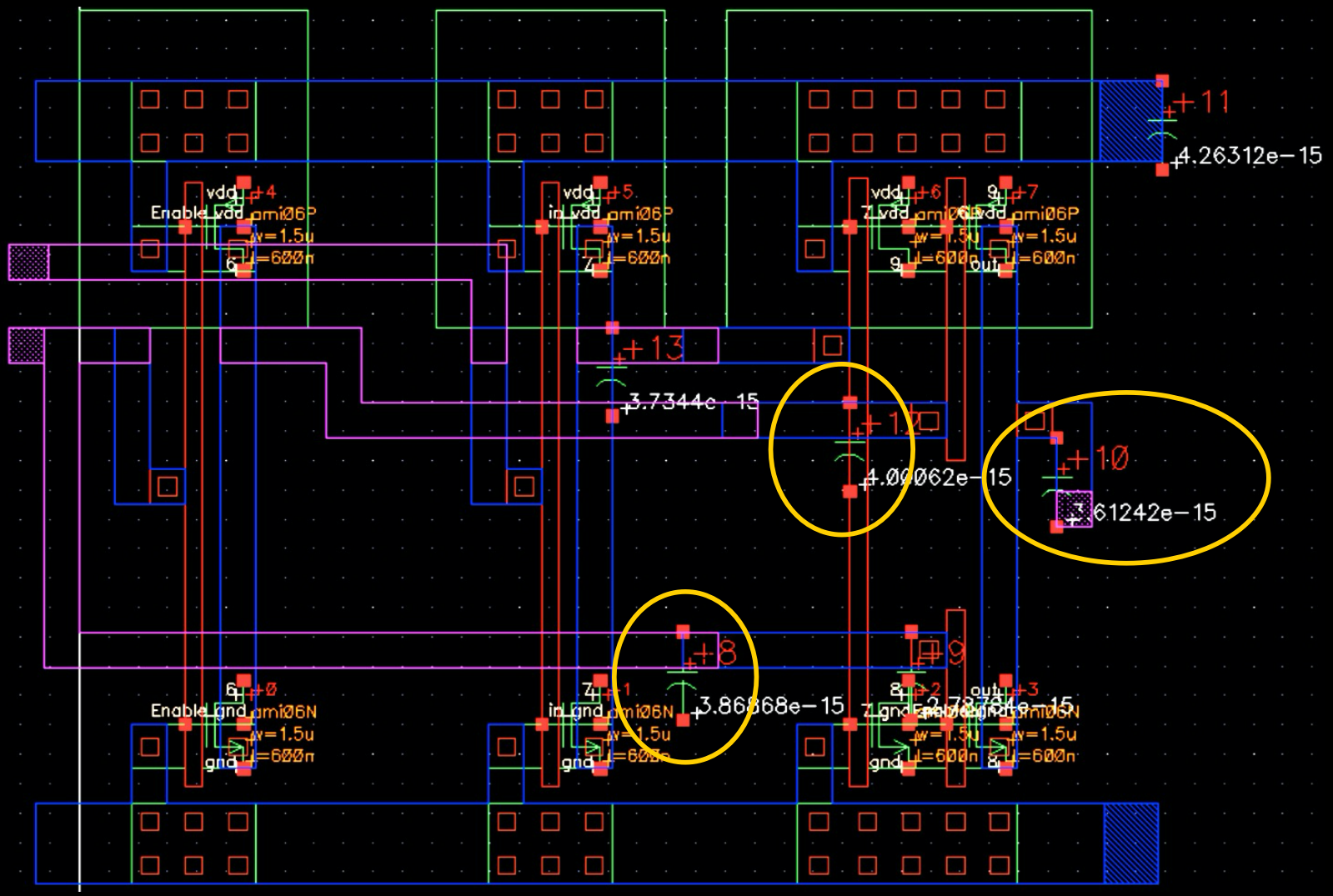


- Circuit extraction is used for LVS, and for spice simulation of layouts

Circuit Extraction

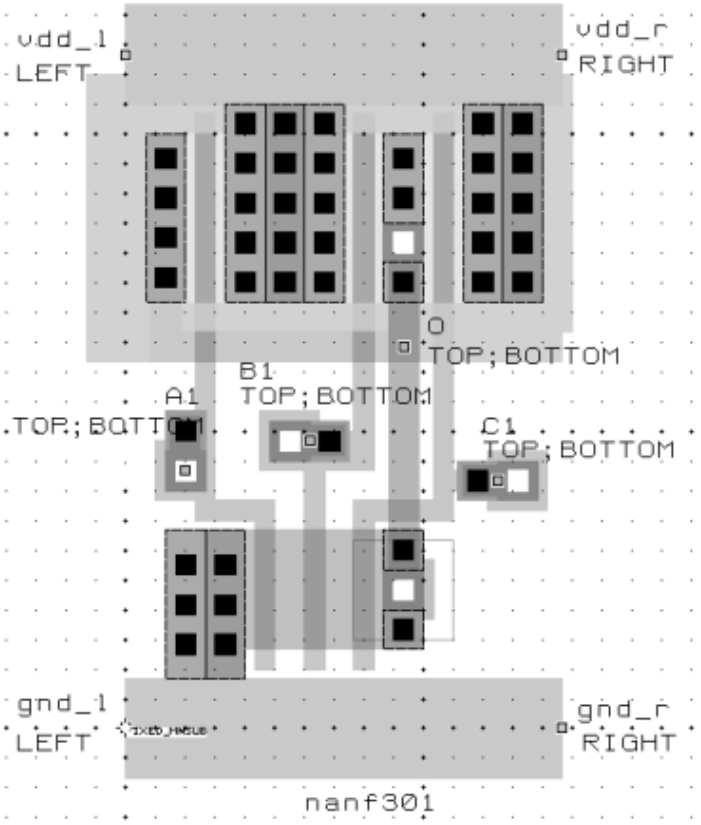


Circuit Extraction



Scaling

Standard Cells



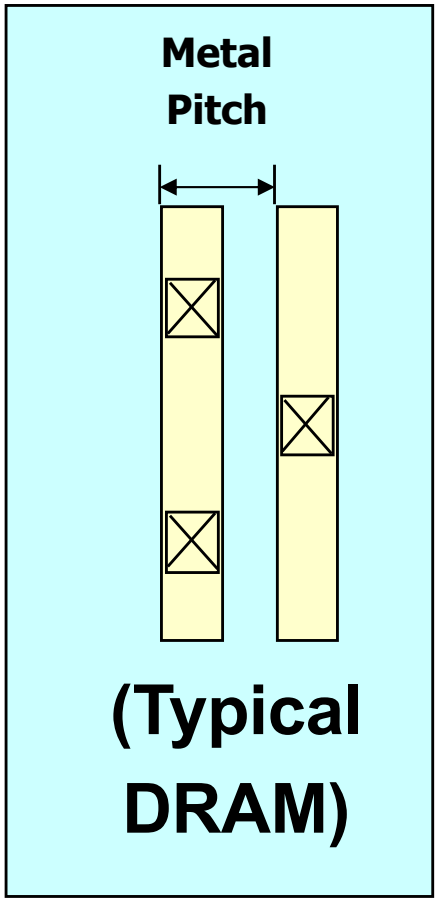
Fanout 4x	0.5 μm	1.0 μm	2.0 μm
A1_tphl	0.595	0.711	0.919
A1_tplh	0.692	0.933	1.360
B1_tphl	0.591	0.739	1.006
B1_tplh	0.620	0.825	1.1.81
C1_tphl	0.574	0.740	1.029
C1_tplh	0.554	0.728	1.026

3-input NAND cell
(from Mississippi State Library)
characterized for fanout of 4 and
for three different technologies

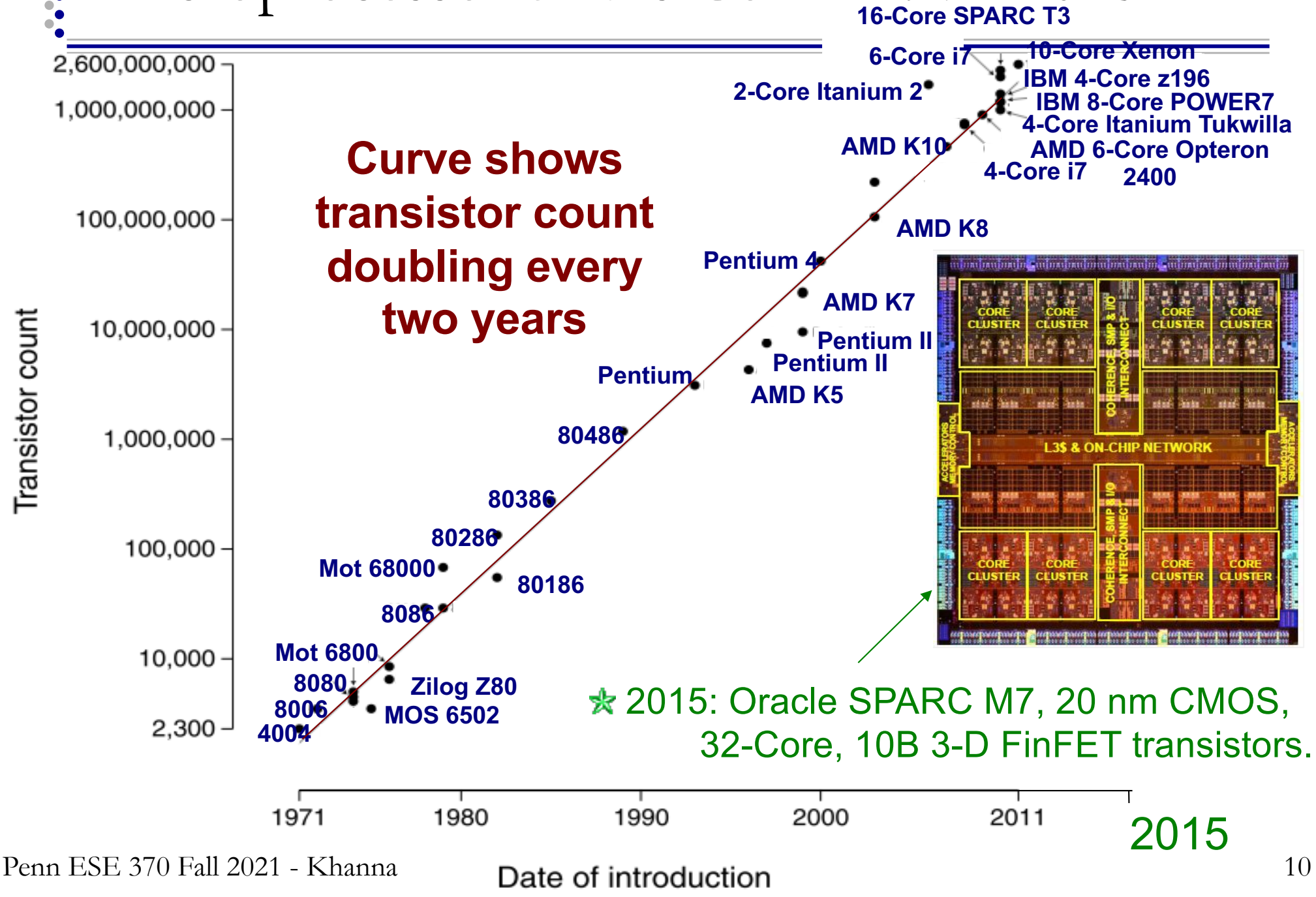
Scaling Technology

- **Premise:** features scale "uniformly"
  - everything gets better in a predictable manner

- **Parameters:**
  - λ (lambda) -- Mead and Conway (L=2λ)
  - F -- Half pitch – ITRS (F=2λ=L)
  - S – scale factor – Rabaey
    - F'=F/S
      - S>1

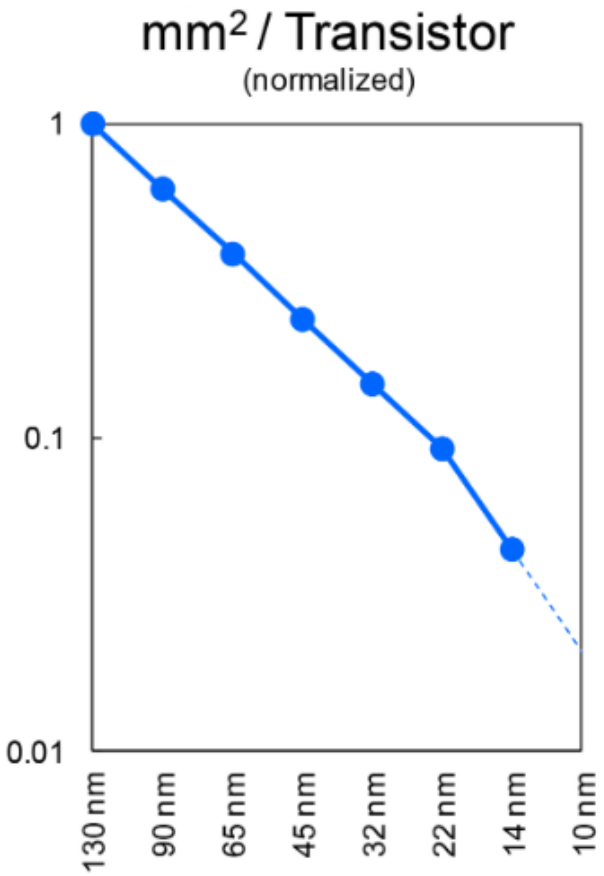


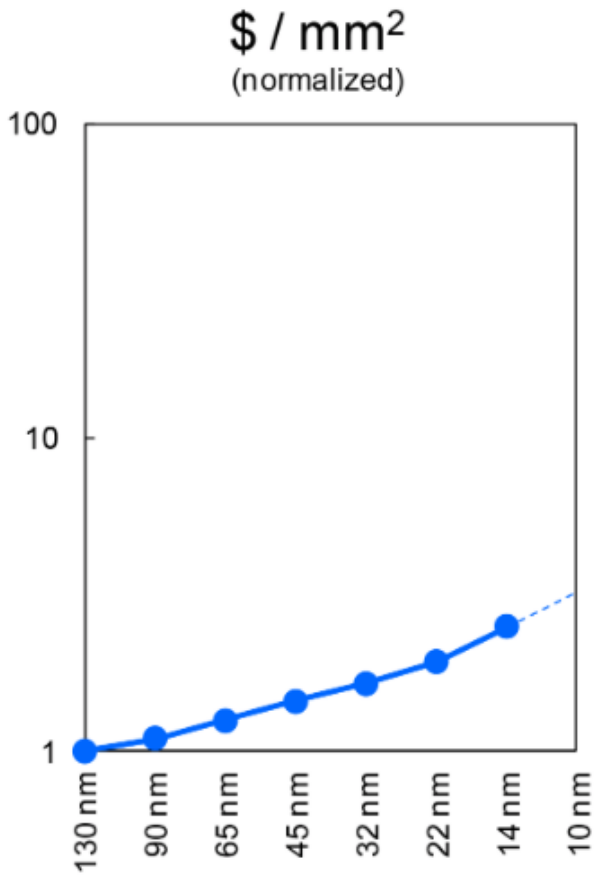
Microprocessor Trans Count 1971-2015

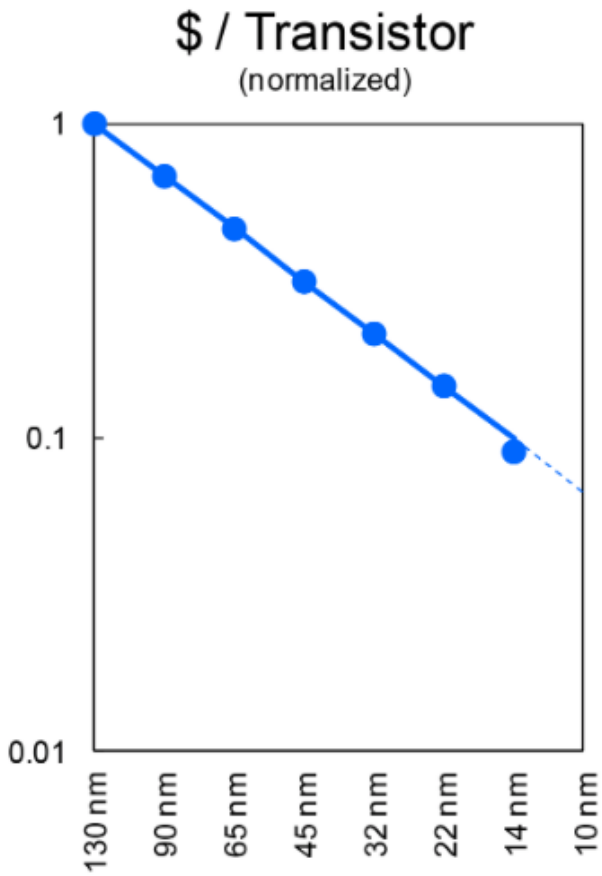


★ 2015: Oracle SPARC M7, 20 nm CMOS, 32-Core, 10B 3-D FinFET transistors.

Intel Cost Scaling





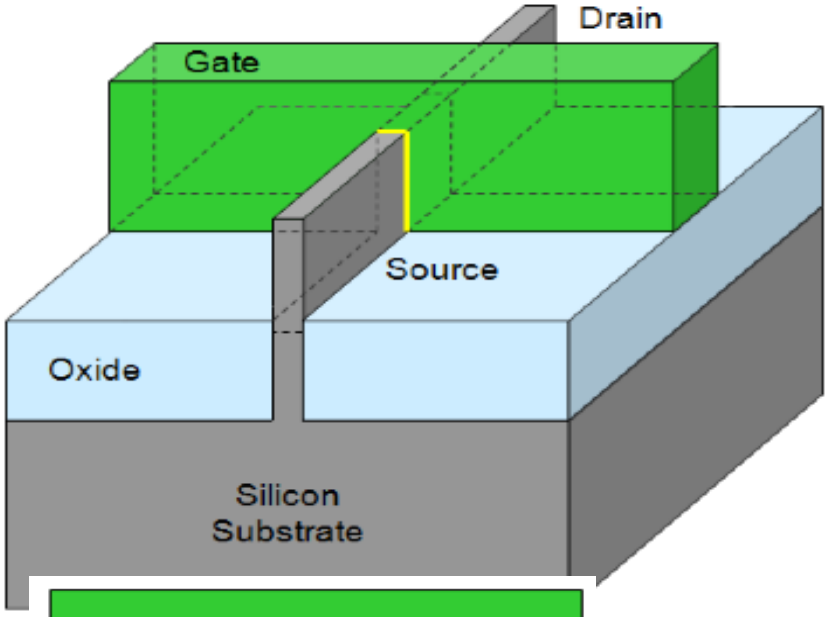


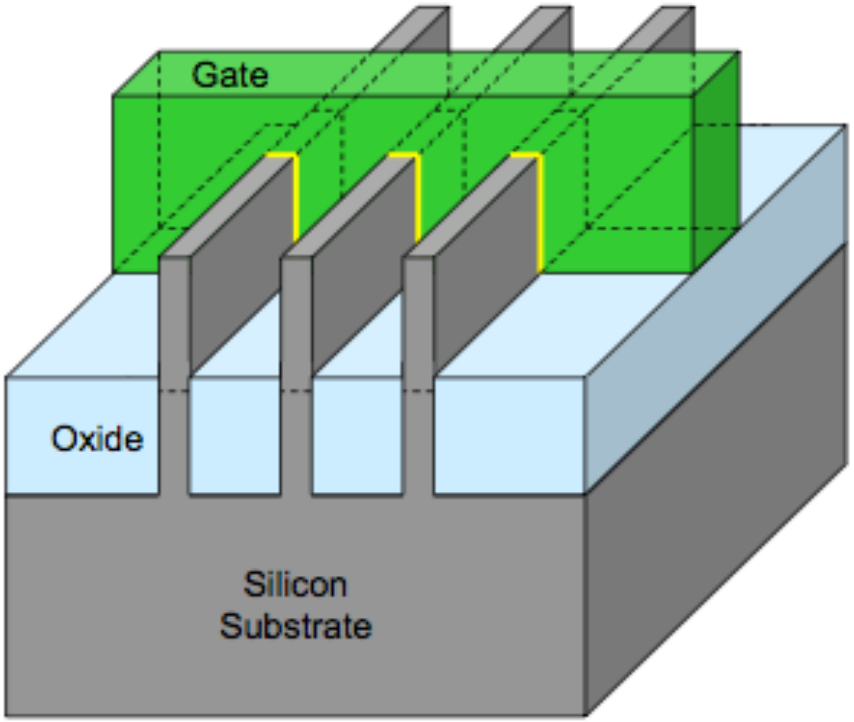
http://www.anandtech.com/show/8367/intels-14nm-technology-in-detail

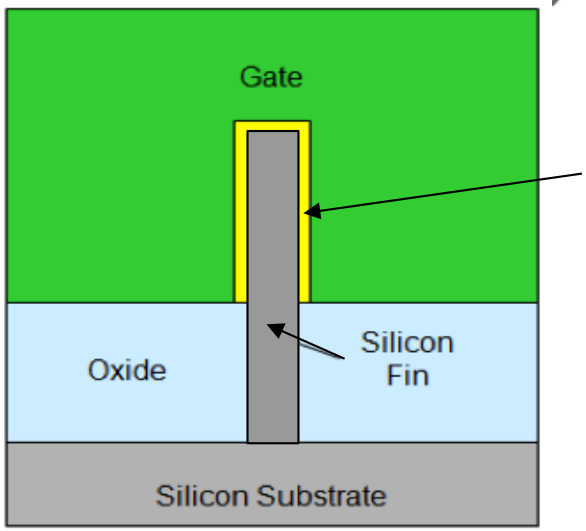
More Moore → Scaling

- Geometrical Scaling
  - continued shrinking of horizontal and vertical physical feature sizes
- Design Equivalent Scaling
  - design technologies that enable high performance, low power, high reliability, low cost, and high design productivity even if neither geometrical nor equivalent scaling can be used

22nm 3D FinFET Transistor







High-k gate dielectric

Tri-Gate transistors with multiple fins connected together increases total drive strength for higher performance

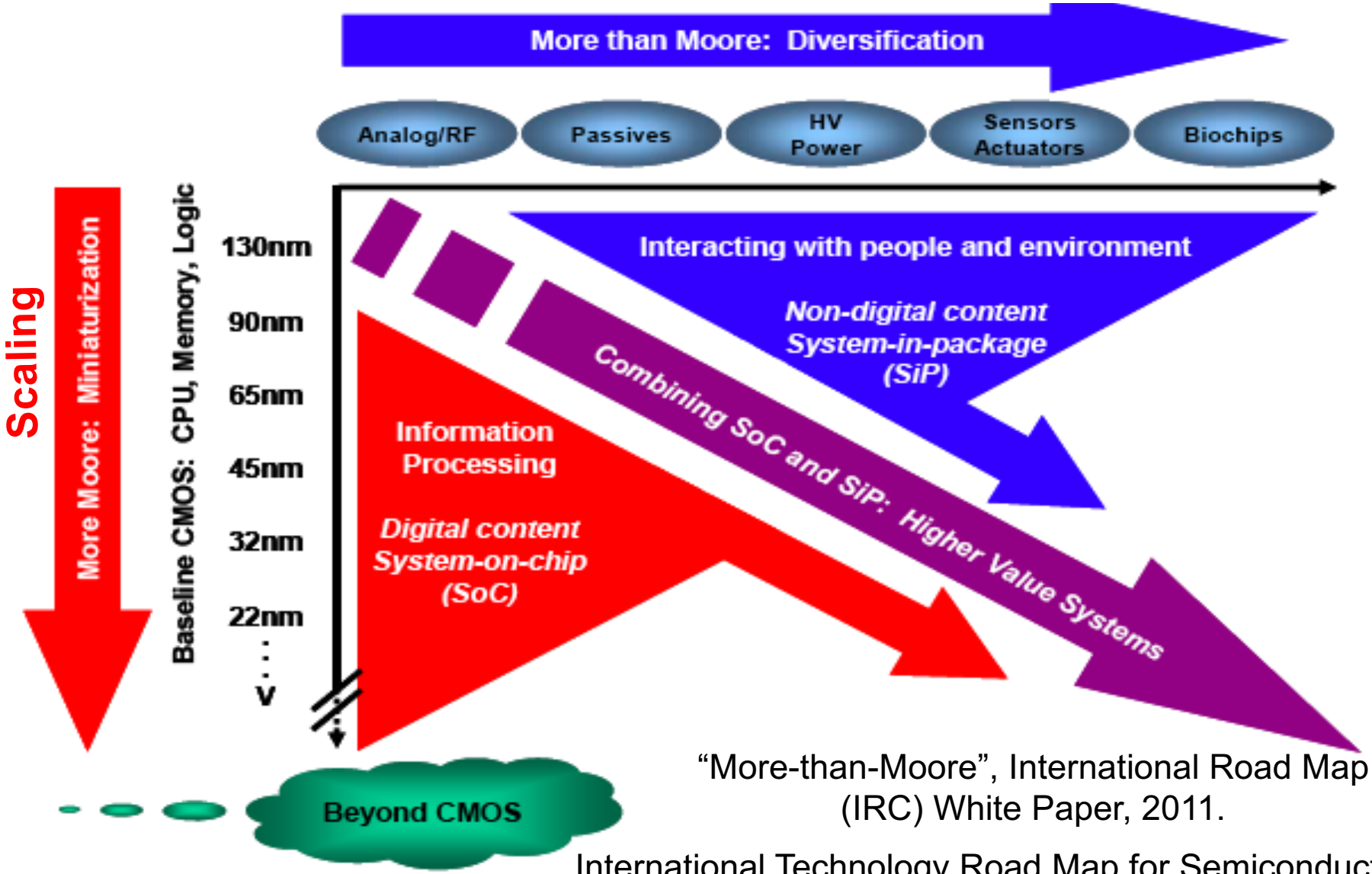
http://download.intel.com/newsroom/kits/22nm/pdfs/22nm-Details_Presentation.pdf

ITRS Roadmap

- International Technology Roadmap for Semiconductors
  - Try to predict where industry going
- ITRS 2.0 started in 2015 with new focus
  - System Integration, Heterogeneous Integration, Heterogeneous Components, Outside System Connectiviy, More Moore, Beyond CMOS and Factory Integartion.

- http://www.itrs2.net/

More-than-Moore



"More-than-Moore", International Road Map (IRC) White Paper, 2011.

International Technology Road Map for Semiconductors

Preclass 1

- Scaling from 32nm → 22nm, what is 1/S?
  - Scaling minimum gate length
  - And pitch distance

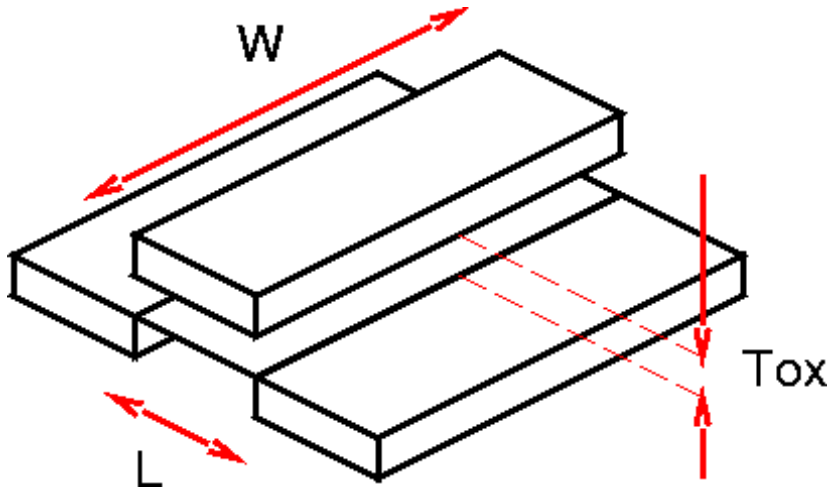
MOS Transistor Scaling - (1974 to present)

1/S=0.7
per technology node
[0.5x per 2 nodes]

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

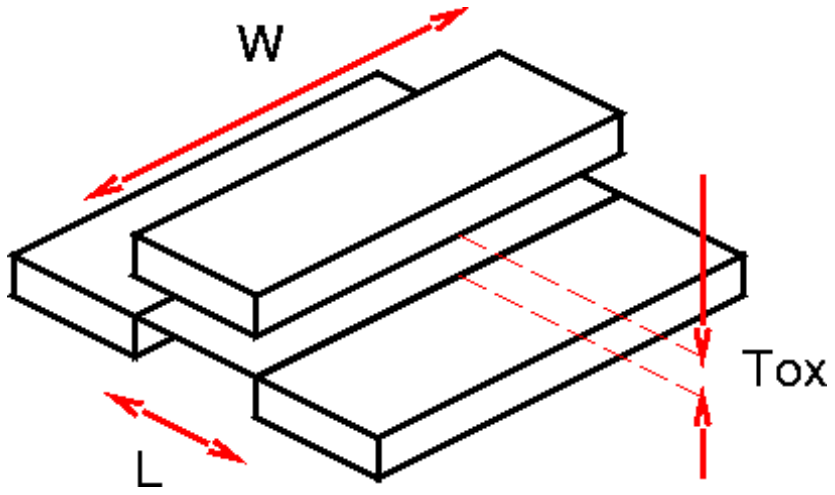
Scaling

- Channel Length (L)
- Channel Width (W)
- Oxide Thickness (t_{ox})
- Doping (N_a)
- Voltage (V_{DD},V_t,)



Full Scaling (Ideal Scaling)

- Channel Length (L) 1/S
- Channel Width (W) 1/S
- Oxide Thickness (t_{ox}) 1/S
- Doping (N_a) S
- Voltage (V_{DD},V_t,) 1/S



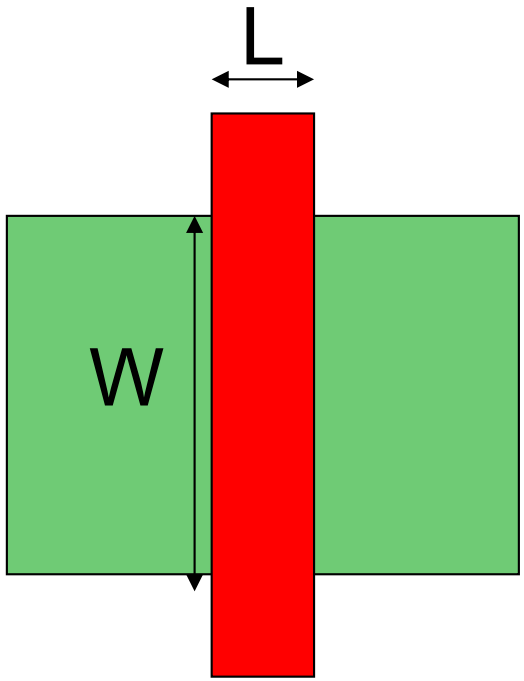
Effects on Physical Properties and Specs?

- Area
- Capacitance
  - C_{ox} and C_{gate}
- Resistance
- Current (I_d)
- Gate Delay (τ_{gd})
- Wire Delay (τ_{wire})
- Power
  - Same frequency
  - Scaled frequency
- Power Density
  - Same frequency
  - Scaled frequency

Preclass 2 - table

Area

- λ' ➔ λ/S
- Area impact?
- $A = L \times W$
-

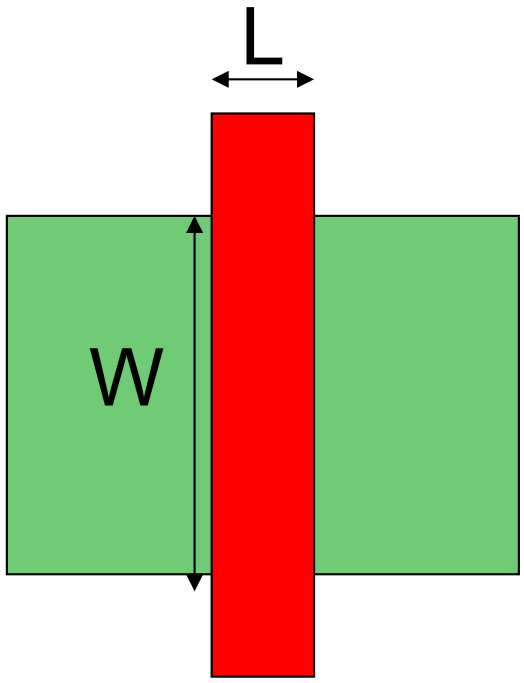


Area

- λ' ➔ λ/S
- Area impact?
- A = L × W
- A' ➔ A/S^2

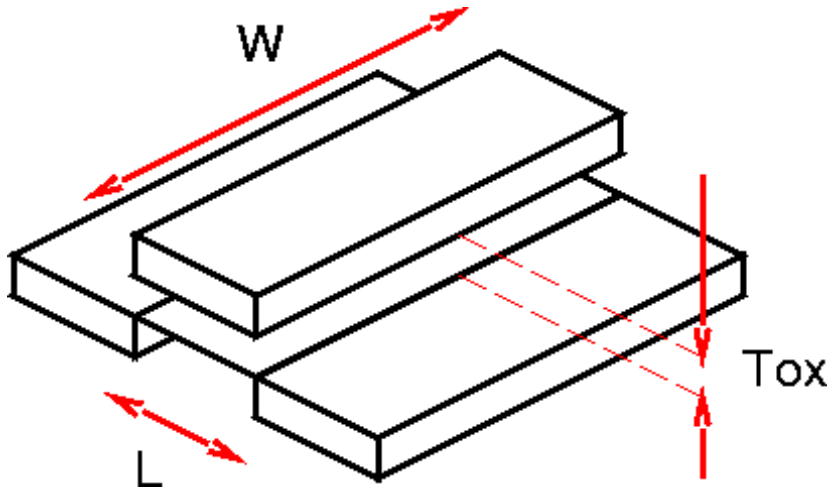
- 32nm ➔ 22nm
- 50% area
- 2 × transistor capacity for same area

1/S=0.7



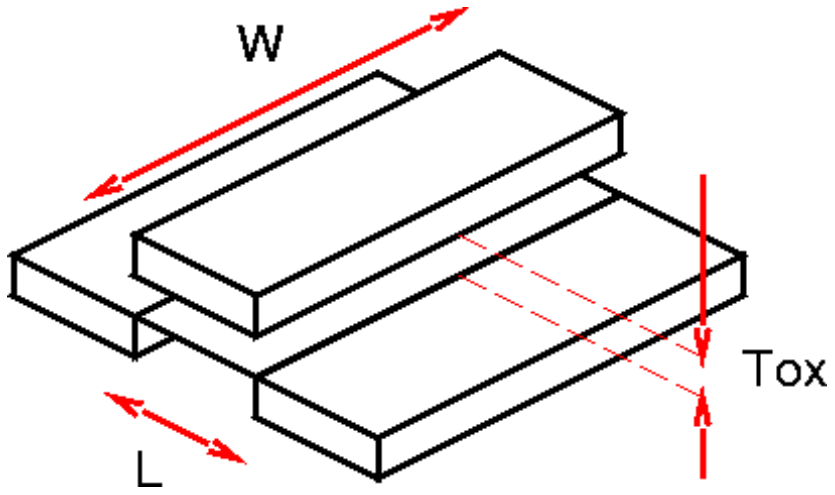
Capacitance

- Capacitance per unit area scaling?
  - $C_{ox} = \varepsilon_{SiO_2}/t_{ox}$
  - $t'_{ox} \rightarrow t_{ox}/S$



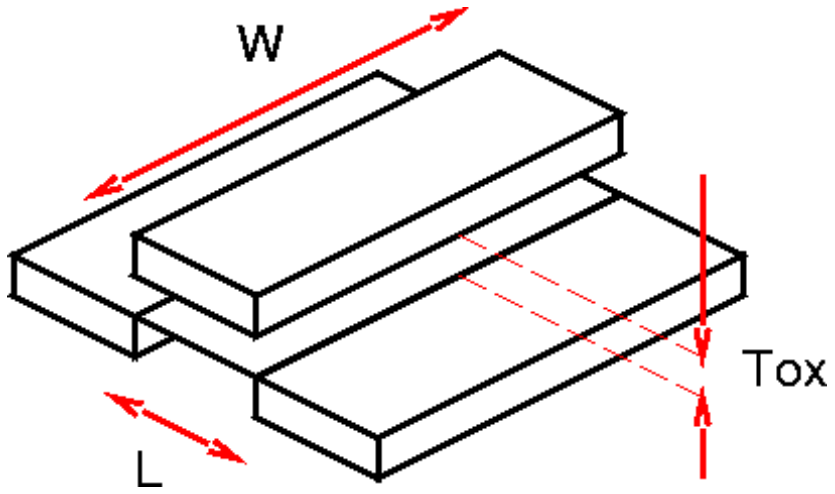
Capacitance

- Capacitance per unit area scaling?
  - $C_{ox} = \varepsilon_{SiO_2}/t_{ox}$
  - $t'_{ox} \rightarrow t_{ox}/S$
  - $C'_{ox} \rightarrow C_{ox} \times S$



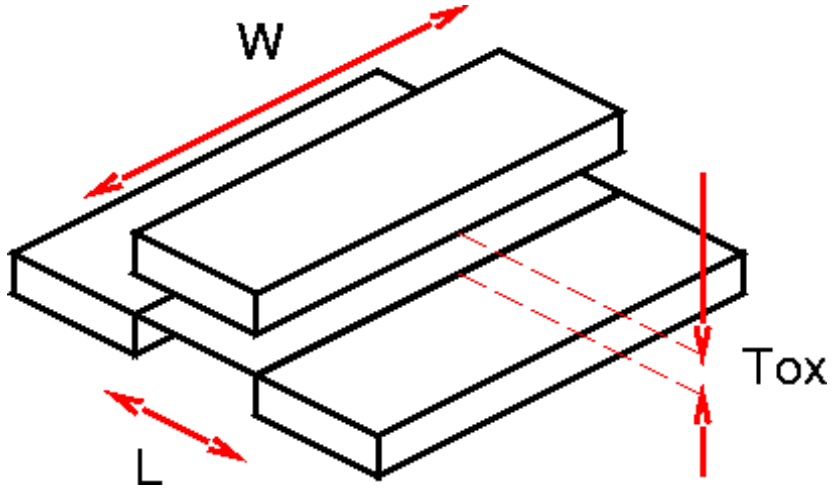
Capacitance

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



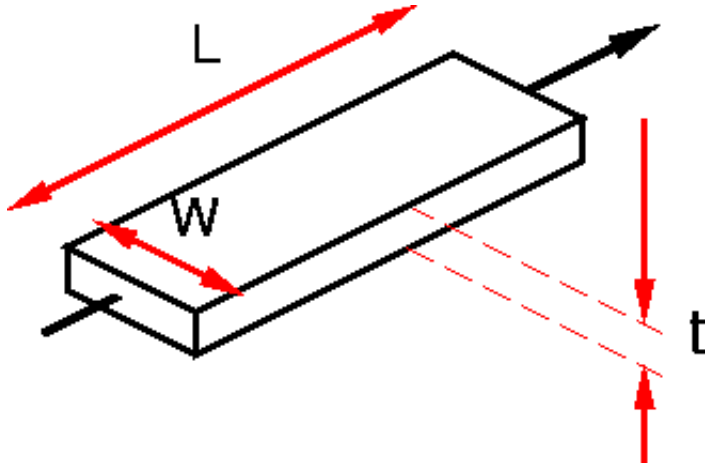
Capacitance

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



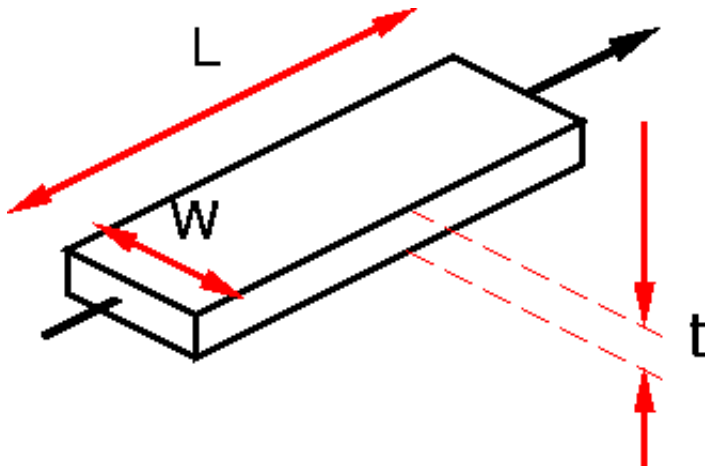
Wire Resistance

- Resistance scaling?
- $R=\rho L/(W*t)$
  - L, t remain similar (not scaled)
- W→ W/S



Wire Resistance

- Resistance scaling?
- $R=\rho L/(W*t)$
  - L, t remain similar (not scaled)
- W → W/S
- R' → R×S



Current

- Which Voltages matters here? (V_{gs},V_{ds},V_{th}…)
- Transistor charging looks like voltage-controlled current source
- Saturation Current scaling?

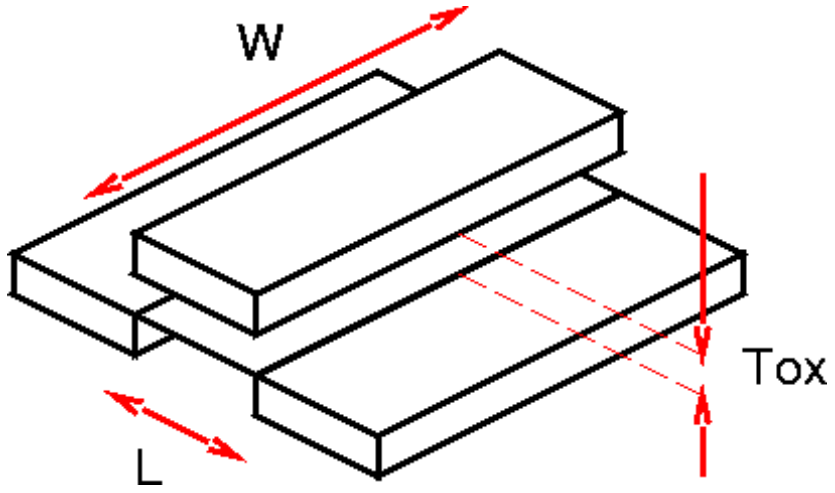
$$I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$$

V_{gs},V_{TH}: V'→ V/S

W'→ W/S

L'→ L/S

C'_{ox} → C_{ox}×S



Current

- Which Voltages matters here? (V_{gs},V_{ds},V_{th}…)
- Transistor charging looks like voltage-controlled current source
- Saturation Current scaling?

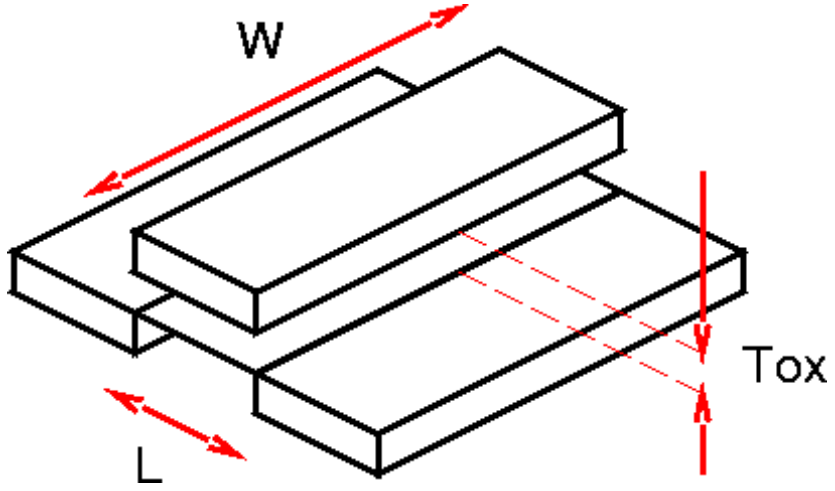
$$I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$$

V_{gs},V_{TH}: V'→ V/S

W'→ W/S

L'→ L/S

C'_{ox} → C_{ox}×S



$$I'_d=(\mu C_{OX}S/2)((W/S)/(L/S))(V_{gs}/S-V_{TH}/S)^2$$

Current

- Which Voltages matters here? (V_{gs},V_{ds},V_{th}…)
- Transistor charging looks like voltage-controlled current source
- Saturation Current scaling?

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

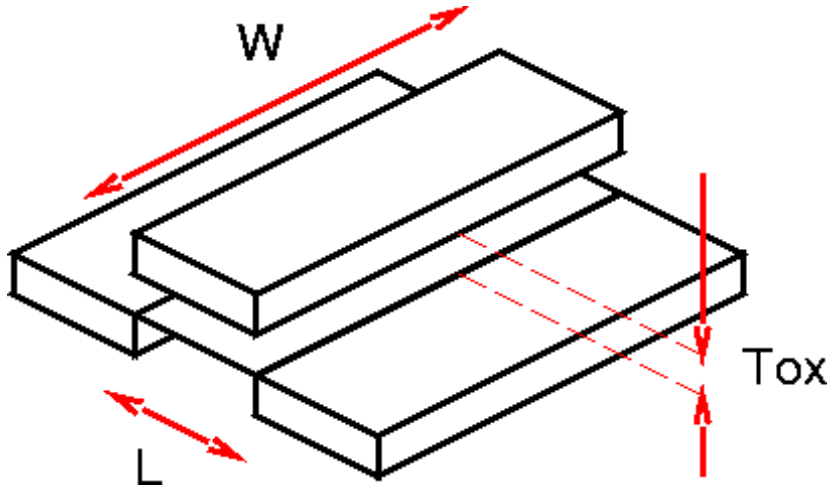
V_{gs},V_{TH}: V'→ V/S

W'→ W/S

L'→ L/S

C'_{ox} → C_{ox}×S

I'_d→I_d/S



Current

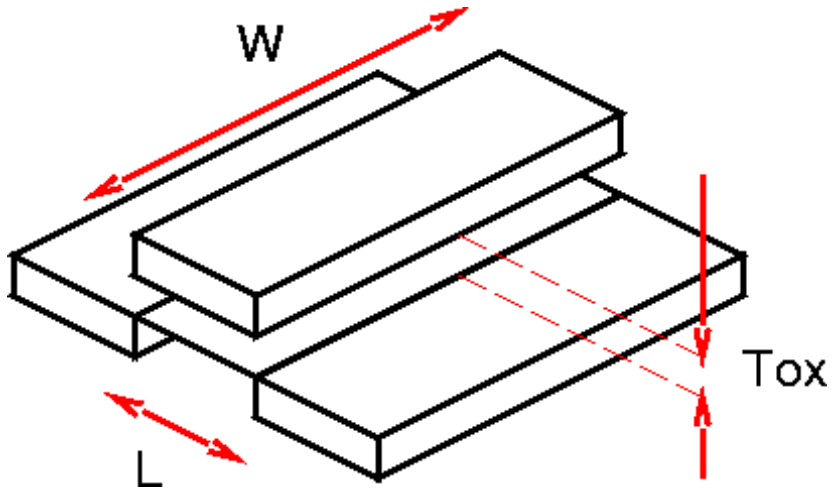
- Velocity Saturation Current scaling?

$V_{gs}, V_{TH}: V' \rightarrow V/S$

$L' \rightarrow L/S$

$W' \rightarrow W/S$

$C'_{ox} \rightarrow C_{ox}S$



Current

- Velocity Saturation Current scaling?

$V_{gs}, V_{TH}: V' \rightarrow V/S$

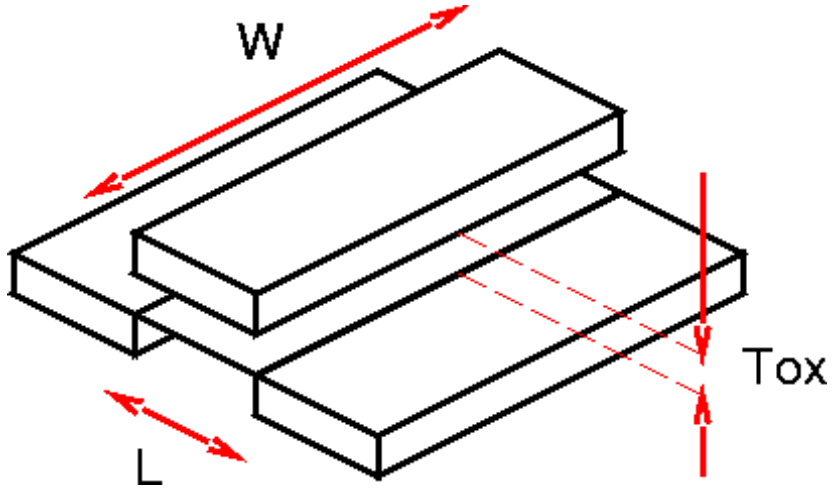
$L' \rightarrow L/S$

$W' \rightarrow W/S$

$C'_{ox} \rightarrow C_{ox}S$

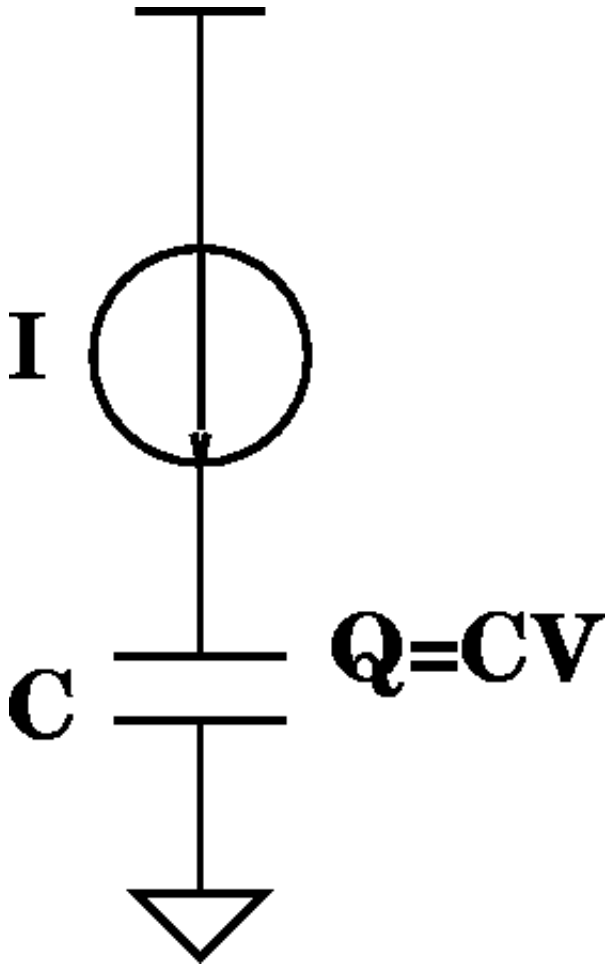
$I'_d \rightarrow I_d/S$

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



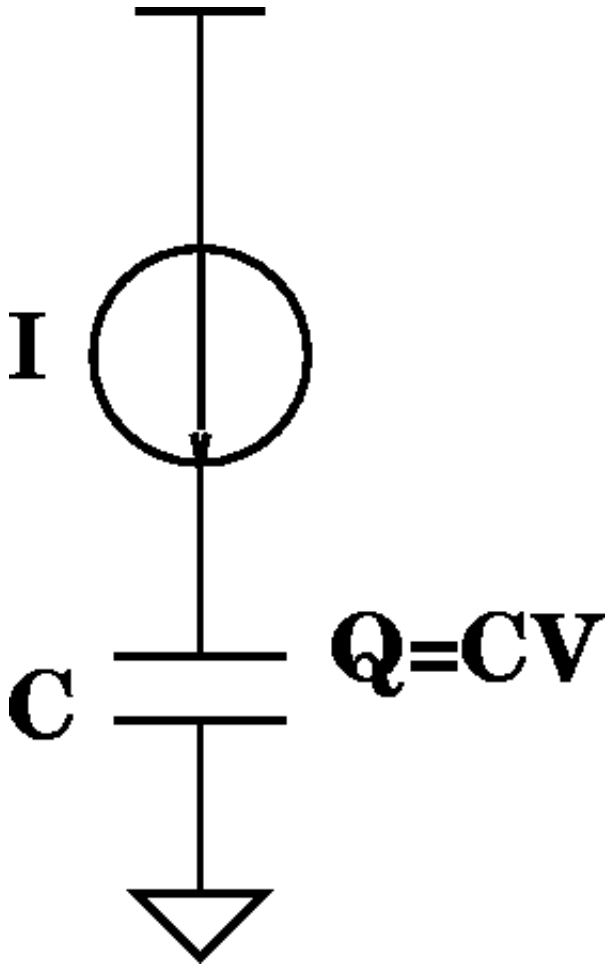
Gate Delay

- Gate Delay scaling?
- $\tau_{gd}=Q/I=(CV)/I$
- V'→ V/S
- $I'_d \rightarrow I_d/S$
- $C_g' \rightarrow C_g/S$



Gate Delay

- Gate Delay scaling?
- $\tau_{gd}=Q/I=(CV)/I$
- V'→ V/S
- $I'_d \rightarrow I_d/S$
- $C_g' \rightarrow C_g/S$
- $\tau'_{gd} \rightarrow \tau_{gd}/S$



Wire Delay

- Wire delay scaling?
- $\tau_{wire} = R \times C$
- R' → R×S
- C' → C/S

  - Again assuming (logical) wire lengths remain constant

Wire Delay

- Wire delay scaling?
- $\tau_{wire} = R \times C$
- R' → R×S
- C' → C/S
- $\tau'_{wire} \rightarrow \tau_{wire}$

  - Again assuming (logical) wire lengths remain constant

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - V' → V/S
  - C' → C/S

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - V'→ V/S
  - C' → C/S
  - P'→ P/S^3

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - $V' \rightarrow V/S$
  - $C' \rightarrow C/S$
  - $P' \rightarrow P/S^3$
- Increase Frequency?
  - $\tau_{gd} \rightarrow \tau_{gd}/S$
  - So: $f \rightarrow f \times S$

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - $V' \rightarrow V/S$
  - $C' \rightarrow C/S$
  - $P' \rightarrow P/S^3$
- Increase Frequency?
  - $\tau_{gd} \rightarrow \tau_{gd}/S$
  - So: $f \rightarrow f \times S$
  - $P \rightarrow P/S^2$

Effects? (Preclass 2)

Area	$1/S^2$
Capacitance (C_{ox}, C_g)	S, 1/S
Resistance	S
Threshold (V_{th})	1/S
Current (I_d)	1/S
Gate Delay (τ_{gd})	1/S
Wire Delay (τ_{wire})	1
Power	$1/S^3$, $1/S^2$

1/S=0.7

Power Density

- P' ➔ P/S^2 (increased frequency)
- P' ➔ P/S^3 (same frequency)
- A' ➔ A/S^2

- Power Density: P/A two cases?

Power Density

- P' ➔ P/S^2 (increased frequency)
- P' ➔ P/S^3 (same frequency)
- A' ➔ A/S^2

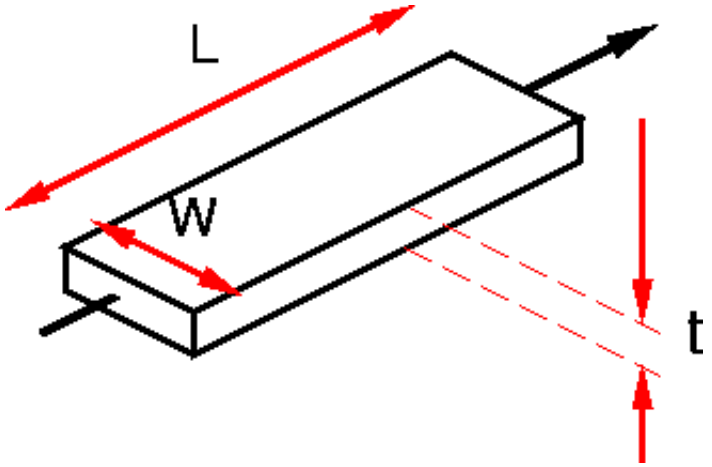
- Power Density: P/A two cases?
  - P/A ➔ P/A increase freq.
  - P/A ➔ (P/A)/S same freq.

But…

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

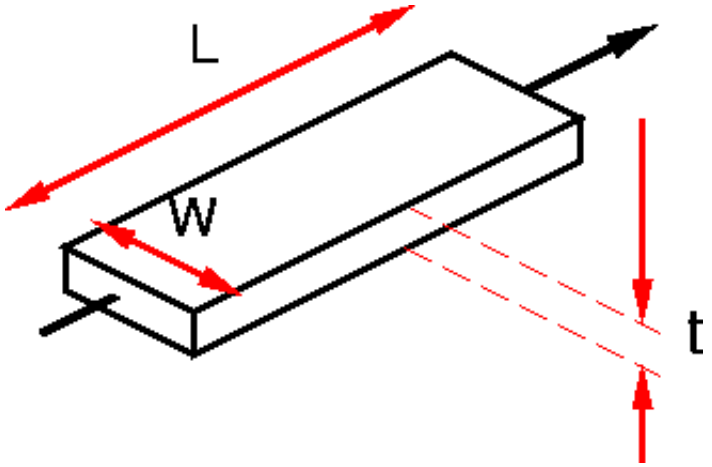
Improving Resistance

- $R=\rho L/(W\times t)$
- W'→ W/S
  - L, t similar
- R' → R×S



Improving Resistance

- $R=\rho L/(W\times t)$
- W'→ W/S
  - L, t similar
- R' → R×S



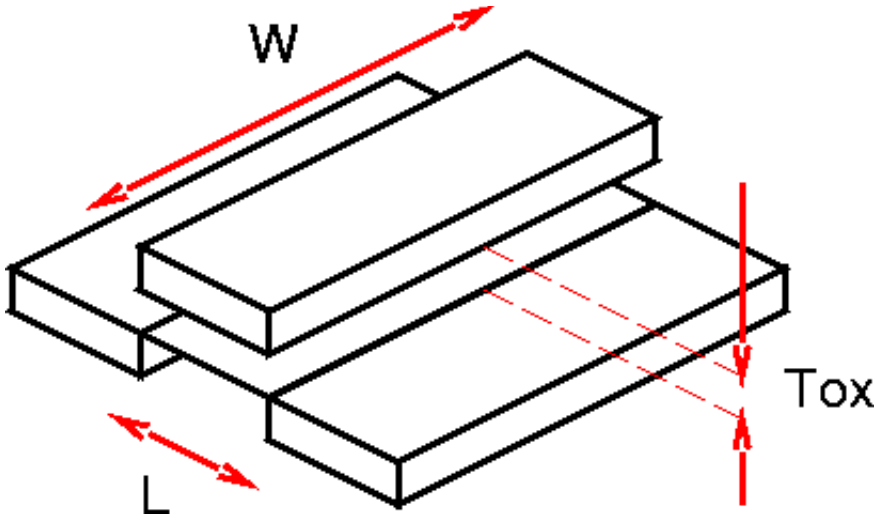
What might we do?

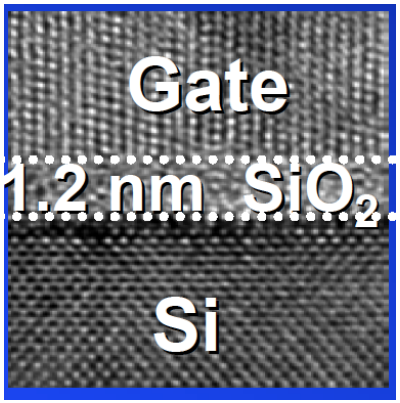
Decrease ρ (copper) – introduced 1997

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

Capacitance and Leakage

- Capacitance per unit area
  - $C_{ox} = \varepsilon_{SiO_2}/t_{ox}$
  - $t'_{ox} \rightarrow t_{ox}/S$
  - $C'_{ox} \rightarrow C_{ox} \times S$



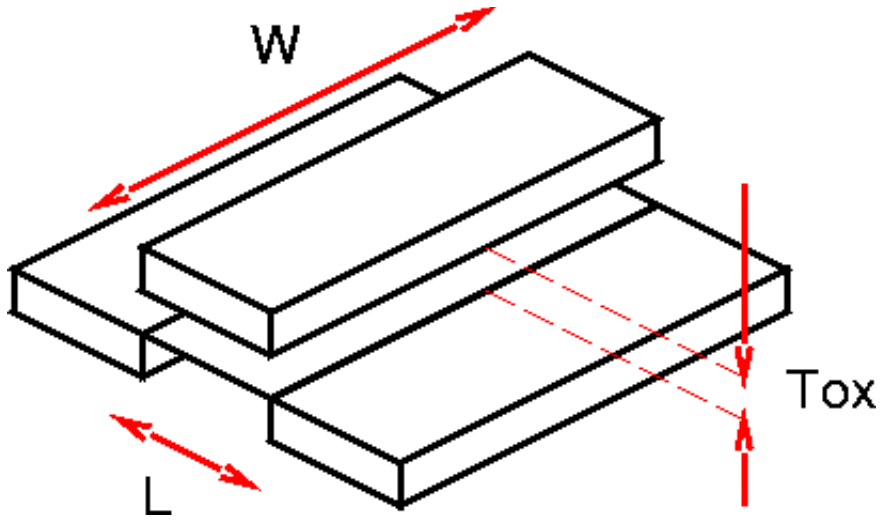


What's wrong with t_{ox} = 1.2nm?

source: Borkar/Micro 2004

Capacitance and Leakage

- Capacitance per unit area
  - $C_{ox} = \varepsilon_{SiO_2}/t_{ox}$
  - $t'_{ox} \rightarrow t_{ox}/S$
  - $C'_{ox} \rightarrow C_{ox} \times S$



What might we do?

Reduce dielectric constant, ε, and not scale thickness to mimic t_{ox} scaling.

High-K dielectric Survey

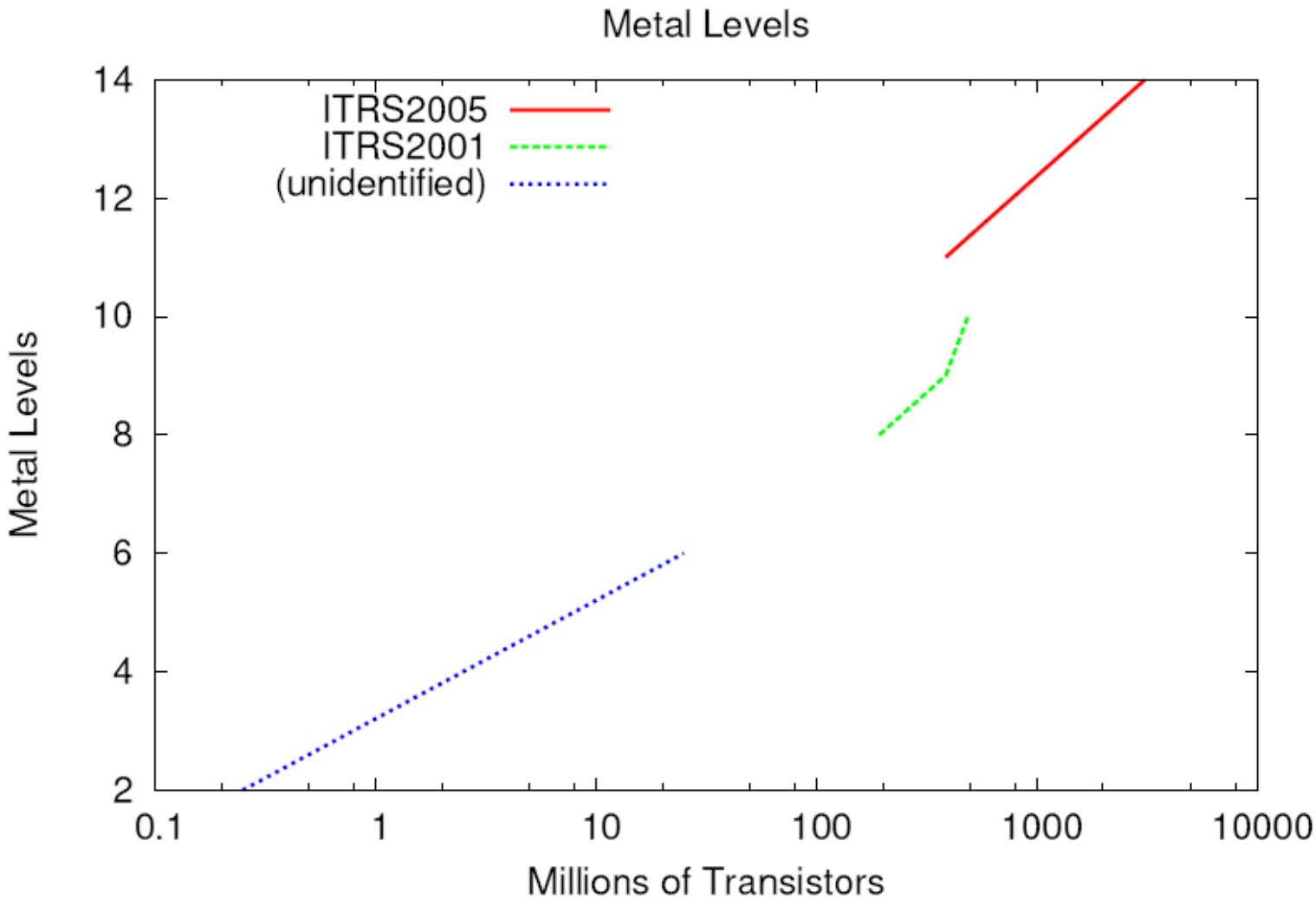
Table 2 Selected material and electrical properties of high-*k* gate dielectrics. Data compiled from Robertson [25], Gusev et al. [20], Hubbard and Schlom [19], and other sources.

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
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, 2002

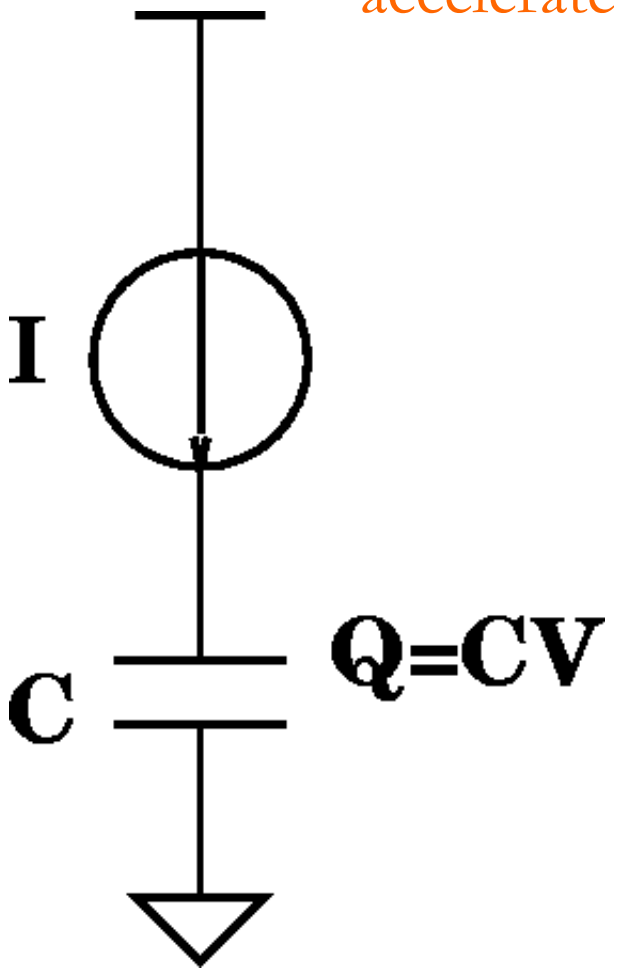
Wire Layers = More Wiring



Gate Delay

- $\tau_{gd}=Q/I=(CV)/I$
- $V' \rightarrow V/S$
- $I'_d \rightarrow I_d/S$
- $C_g' \rightarrow C_g/S$
- $\tau'_{gd} \rightarrow \tau_{gd}/S$

How might we accelerate speed up?

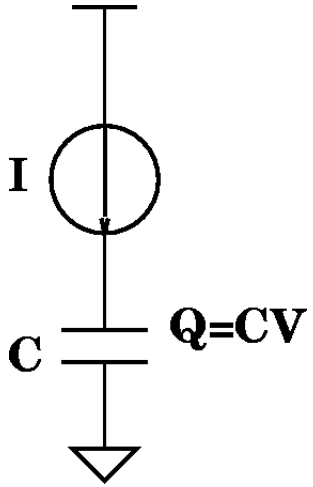


Gate Delay

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

How might we accelerate speed up?

Don't scale V!

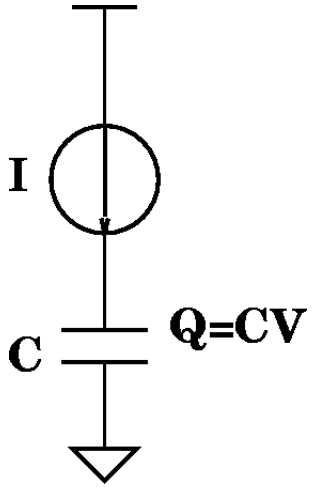


Gate Delay

- $\tau_{gd}=Q/I=(CV)/I$
- $V' \rightarrow V$
  - $I'_d=(\mu C_{OX} S/2)((W/S)/(L/S))(V_{gs}-V_{TH})^2$
- $I'_d \rightarrow I_d \times S$
- $C_g' \rightarrow C_g/S$
- $\tau'_{gd} \rightarrow \tau_{gd}/S^2$

How might we accelerate speed up?

Don't scale V!



But… Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - $P=(1/2)CV^2f$
  - V'→ V
  - C' → C/S
  - P'→ P/S

But… Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - $P=(1/2)CV^2f$
  - $V' \rightarrow V$
  - $C' \rightarrow C/S$
  - $P' \rightarrow P/S$
- Increase Frequency?
  - $\tau'_{gd} \rightarrow \tau_{gd}/S^2$
  - $f' \rightarrow f \times S^2$
  - $P' \rightarrow P \times S$

If don't scale V, power dissipation doesn't scale down!

…And Power Density

- P'→ P×S (increase frequency)
- A' → A/S^2
- What happens to power density?

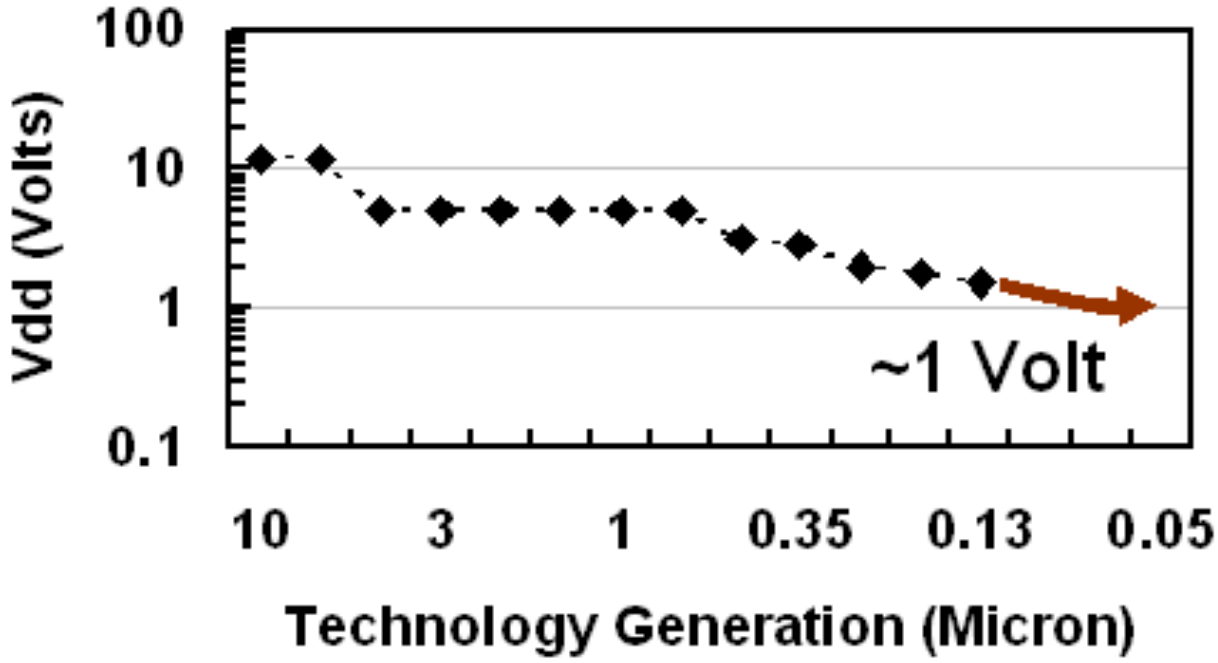
…And Power Density

- P'→ P×S (increase frequency)
- A' → A/S^2
- What happens to power density?

- $P/A \rightarrow S^3 \times P$

- Power Density Increases!

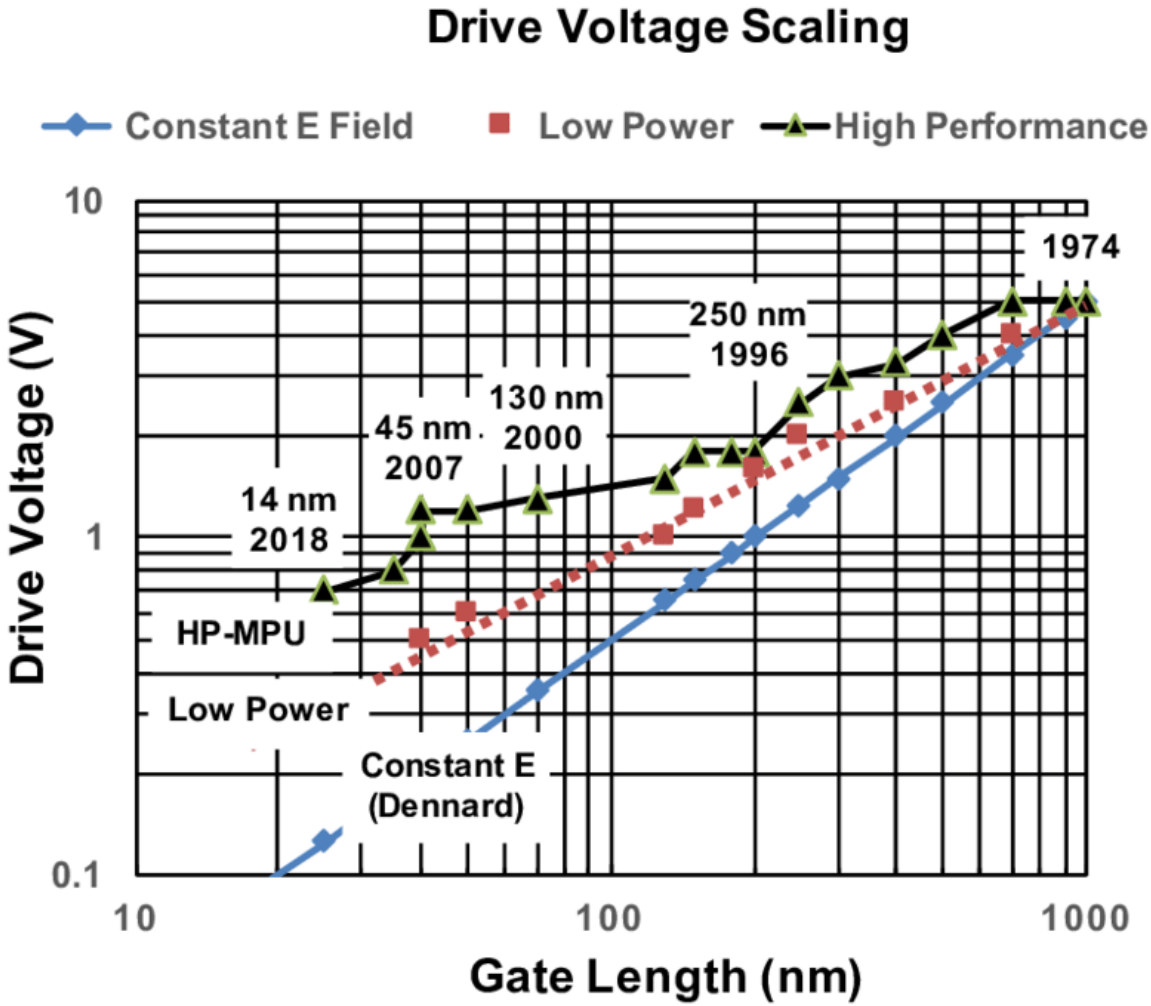
Historical Voltage Scaling



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

- Frequency impact?
- Power Density impact?

Historical Voltage Scaling

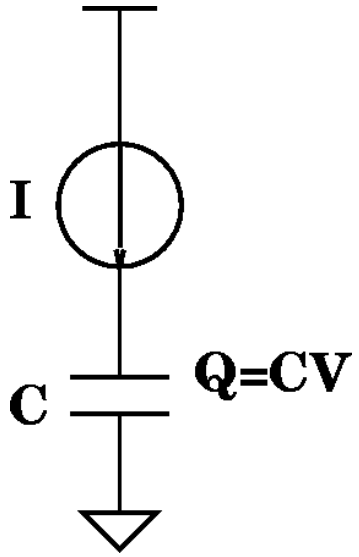


http://software opportunities/

- Freq
- Pow

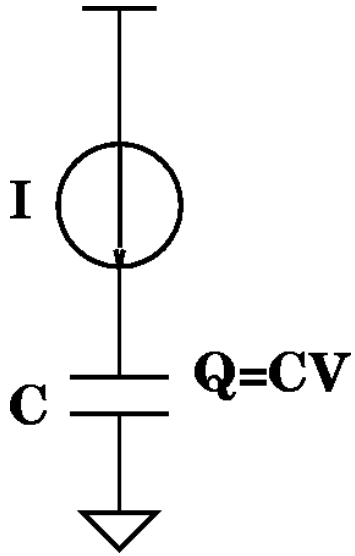
Scale V separately with Factor U, (U<S)

- $\tau_{gd}=Q/I=(CV)/I$
- V'→V/U



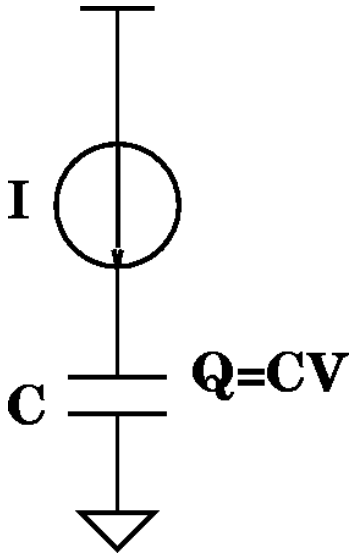
Scale V separately with Factor U

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



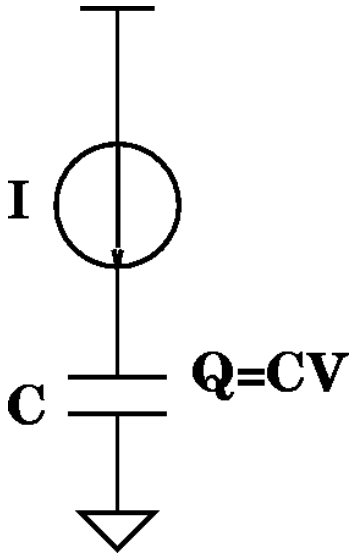
Scale V separately with Factor U

- $\tau_{gd}=Q/I=(CV)/I$
- $V' \rightarrow V/U$
  - $I'_d=(\mu C_{OX} S/2)((W/S)/(L/S)(V_{gs}/U-V_{TH}/U)^2$
- $I'_d \rightarrow S/U^2 \times I_d$
- $C' \rightarrow C/S$
- $\tau'_{gd} \rightarrow ((1/(SU)) / (S/U^2)) \times \tau_{gd}$
- $\tau'_{gd} \rightarrow (U/S^2) \times \tau_{gd}$



Scale V separately with Factor U

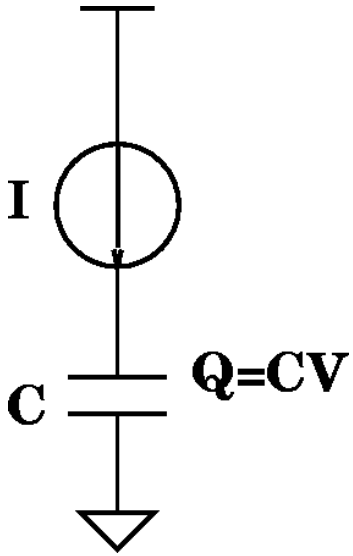
- $\tau_{gd}=Q/I=(CV)/I$
- $V' \rightarrow V/U$
  - $I'_d=(\mu C_{OX} S/2)((W/S)/(L/S)(V_{gs}/U-V_{TH}/U)^2$
- $I'_d \rightarrow S/U^2 \times I_d$
- $C' \rightarrow C/S$
- $\tau'_{gd} \rightarrow ((1/(SU)) / (S/U^2)) \times \tau_{gd}$
- $\tau'_{gd} \rightarrow (U/S^2) \times \tau_{gd}$
- $f' \rightarrow (S^2/U) \times f$



Scale V separately with Factor U

Ideal scale factors:
S=100
U=100
τ=1/100
f_{ideal}=100

- $\tau_{gd}=Q/I=(CV)/I$
- V'→V/U
  - $I'_d=(\mu C_{OX}S/2)((W/S)/(L/S)(V_{gs}/U-V_{TH}/U)^2$
- $I'_d \rightarrow S/U^2 \times I_d$
- $C' \rightarrow C/S$
- $\tau'_{gd} \rightarrow ((1/(SU)) / (S/U^2)) \times \tau_{gd}$
- $\tau'_{gd} \rightarrow (U/S^2) \times \tau_{gd}$
- $f' \rightarrow (S^2/U) \times f$



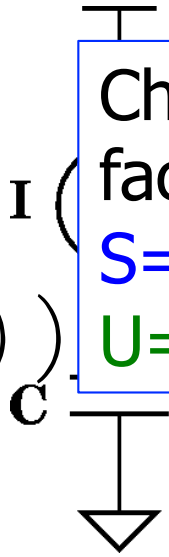
Preclass 3

- Assuming V_{dd}=10V in a 10μm process and V_{dd}=1V in a 100nm process, what are S and U? (assume everything else scales according to ideal scaling.)
  - S =
  - U =

Scale V separately with Factor U

- $\tau_{gd}=Q/I=(CV)/I$
- $V' \rightarrow V/U$
  - $I'_d=(\mu C_{OX} S/2)((W/S)/(L/S)(V_{gs}/U-V_{TH}/U)^2$
- $I'_d \rightarrow S/U^2 \times I_d$
- $C' \rightarrow C/S$
- $\tau'_{gd} \rightarrow \tau'_{gd} \rightarrow ((1/(SU)) / (S/U^2))$
- $\tau'_{gd} \rightarrow (U/S^2) \times \tau_{gd}$
- $f' \rightarrow (S^2/U) \times f$

Ideal scale factors:
S=100
U=100
τ=1/100
f_{ideal}=100



Cheating factors:
S=100
U=10

How much faster are gates?

Scale V separately with Factor U

- $\tau_{gd}=Q/I=(CV)/I$
- $V' \rightarrow V/U$
  - $I'_d=(\mu C_{OX} S/2)((W/S)/(L/S)(V_{gs}/U-V_{TH}/U)^2$
- $I'_d \rightarrow S/U^2 \times I_d$
- $C' \rightarrow C/S$
- $\tau'_{gd} \rightarrow \tau'_{gd} \rightarrow ((1/(SU)) / (S/U^2))$
- $\tau'_{gd} \rightarrow (U/S^2) \times \tau_{gd}$
- $f' \rightarrow (S^2/U) \times f$

Ideal scale factors:
S=100
U=100
τ=1/100
f_{ideal}=100

Cheating factors:
S=100
U=10
τ=1/1000
f_{new}=1000

How much faster are gates?

Scale V separately with Factor U

- $\tau_{gd}=Q/I=(CV)/I$
- $V' \rightarrow V/U$
  - $I'_d=(\mu C_{OX} S/2)((W/S)/(L/S)(V_{gs}/U-V_{TH}/U)^2$
- $I'_d \rightarrow S/U^2 \times I_d$
- $C' \rightarrow C/S$
- $\tau'_{gd} \rightarrow ((1/(SU)) / (S/U^2)) \times \tau_{gd}$
- $\tau'_{gd} \rightarrow (U/S^2) \times \tau_{gd}$
- $f' \rightarrow (S^2/U) \times f$

Ideal scale factors:
S=100
U=100
τ=1/100
f_{ideal}=100

Cheating factors:
S=100
U=10
τ=1/1000
f_{new}=1000

$f_{new}/f_{ideal}=10$

Power Density Impact

- $P = (1/2)CV^2 f$
- $P \rightarrow (1/S)\ (1/U^2)\ (S^2/U) = S/U^3$
- $P/A \rightarrow (S/U^3)\ /\ (1/S^2) = S^3/U^3$

Power Density Impact

- $P = (1/2)CV^2 f$
- $P \rightarrow (1/S)\ (1/U^2)\ (S^2/U) = S/U^3$
- $P/A \rightarrow (S/U^3)\ /\ (1/S^2) = S^3/U^3$

- $U=10 \quad S=100$
- $P/A \rightarrow 1000\ (P/A)$

Power Density Impact

- $P = (1/2)CV^2 f$
- $P \rightarrow (1/S)\ (1/U^2)\ (S^2/U) = S/U^3$
- $P/A \rightarrow (S/U^3)\ /\ (1/S^2) = S^3/U^3$

- $U=10 \quad S=100$
- $P/A \rightarrow 1000\ (P/A)$

- Compare with ideal scaling:
- $P/A \rightarrow S^3 \times P$ (ideal scaling)
- $P/A \rightarrow 1{,}000{,}000\ (P/A)$ (ideal scaling)

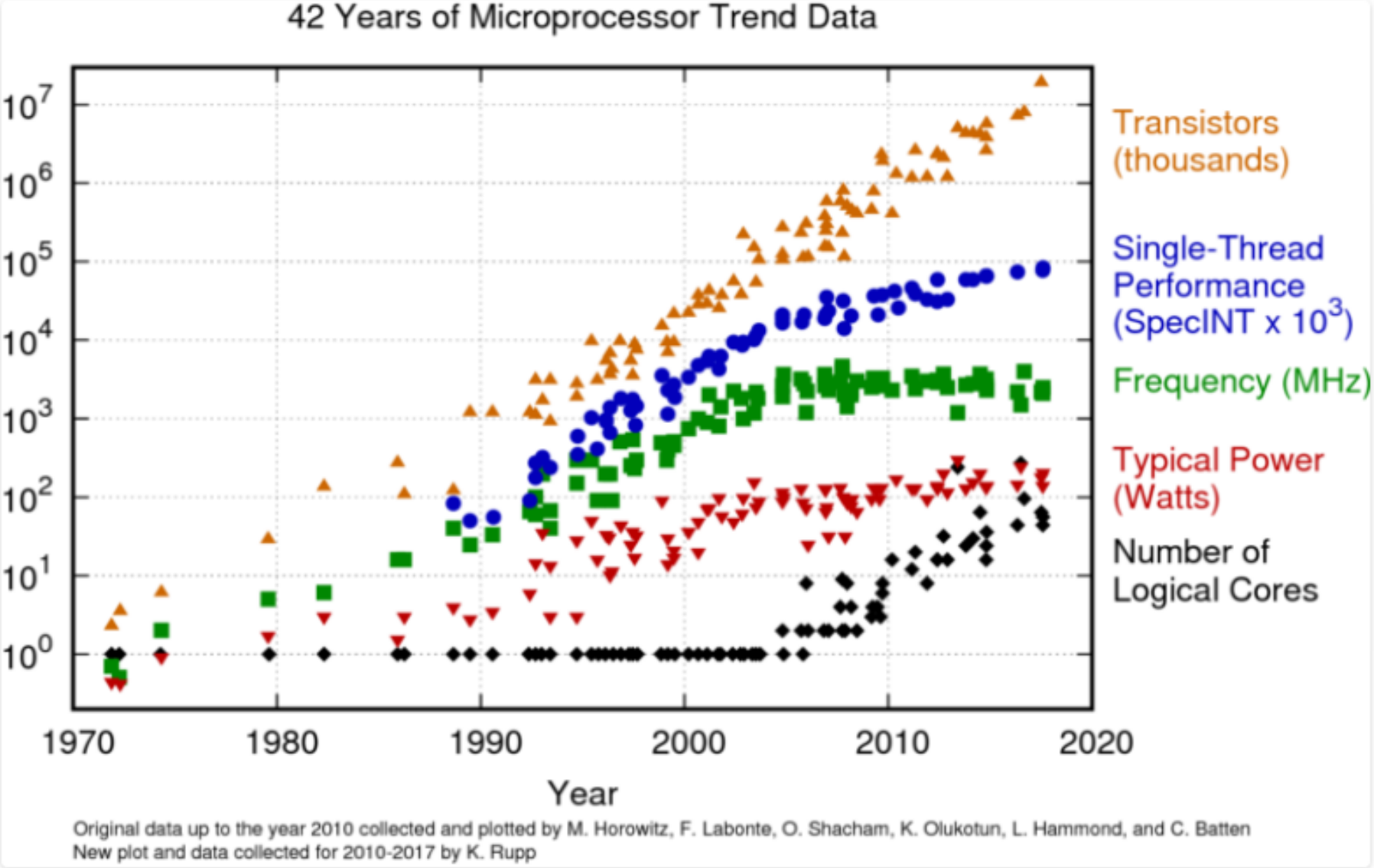
Scaling Methods

Table 3.8 Scaling scenarios for short-channel devices.

Parameter	Relation	Constant Field	General Scaling	Constant Voltage
W, L, t_{ox}		$1/S$	$1/S$	$1/S$
V_{DD}, V_T		$1/S$	$1/U$	1
N_{SUB}	V/W_{depl}^2	S	S^2/U	S^2
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
C_{ox}	$1/t_{ox}$	S	S	S
C_{gate}	$C_{ox}WL$	$1/S$	$1/S$	$1/S$
k_n, k_p	$C_{ox}W/L$	S	S	S

U<S

42 Years of uP Trend Data

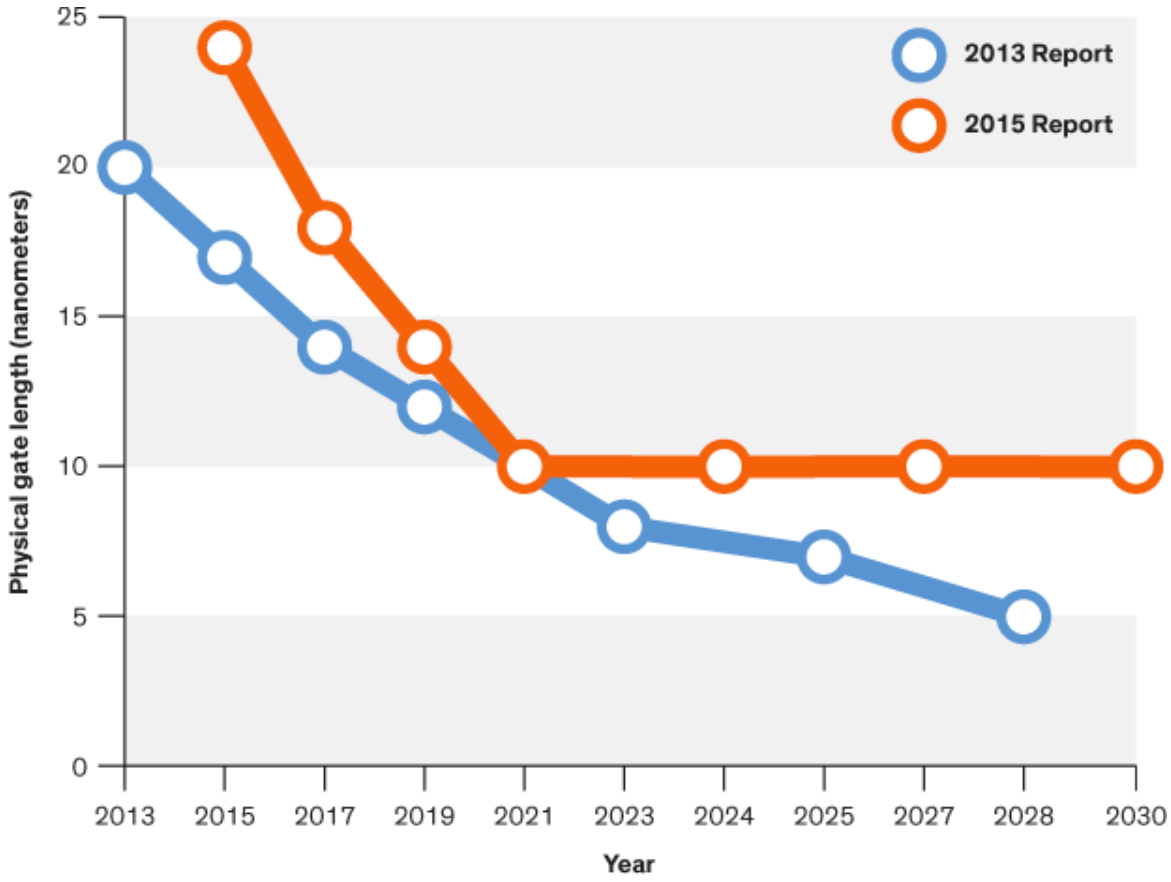


Conventional Scaling

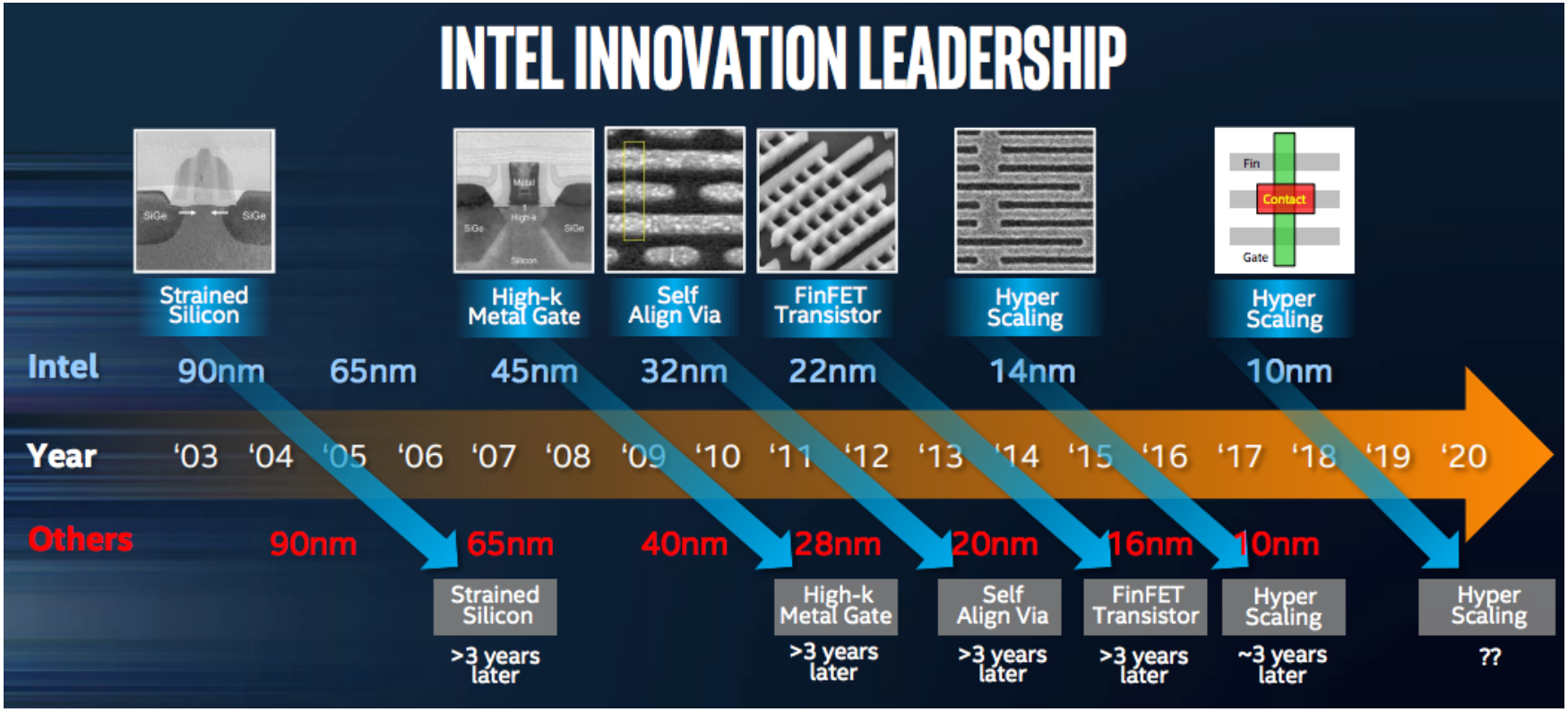
- Ends in your lifetime
- Perhaps already:
  - "Basically, this is the end of scaling."
    - May 2005, Bernard Meyerson, V.P. and chief technologist for IBM's systems and technology group

ITRS 2.0 Report 2015

- "After 2021, the report forecasts, it will no longer be economically desirable for companies to continue traditional transistor miniaturization in microprocessors."

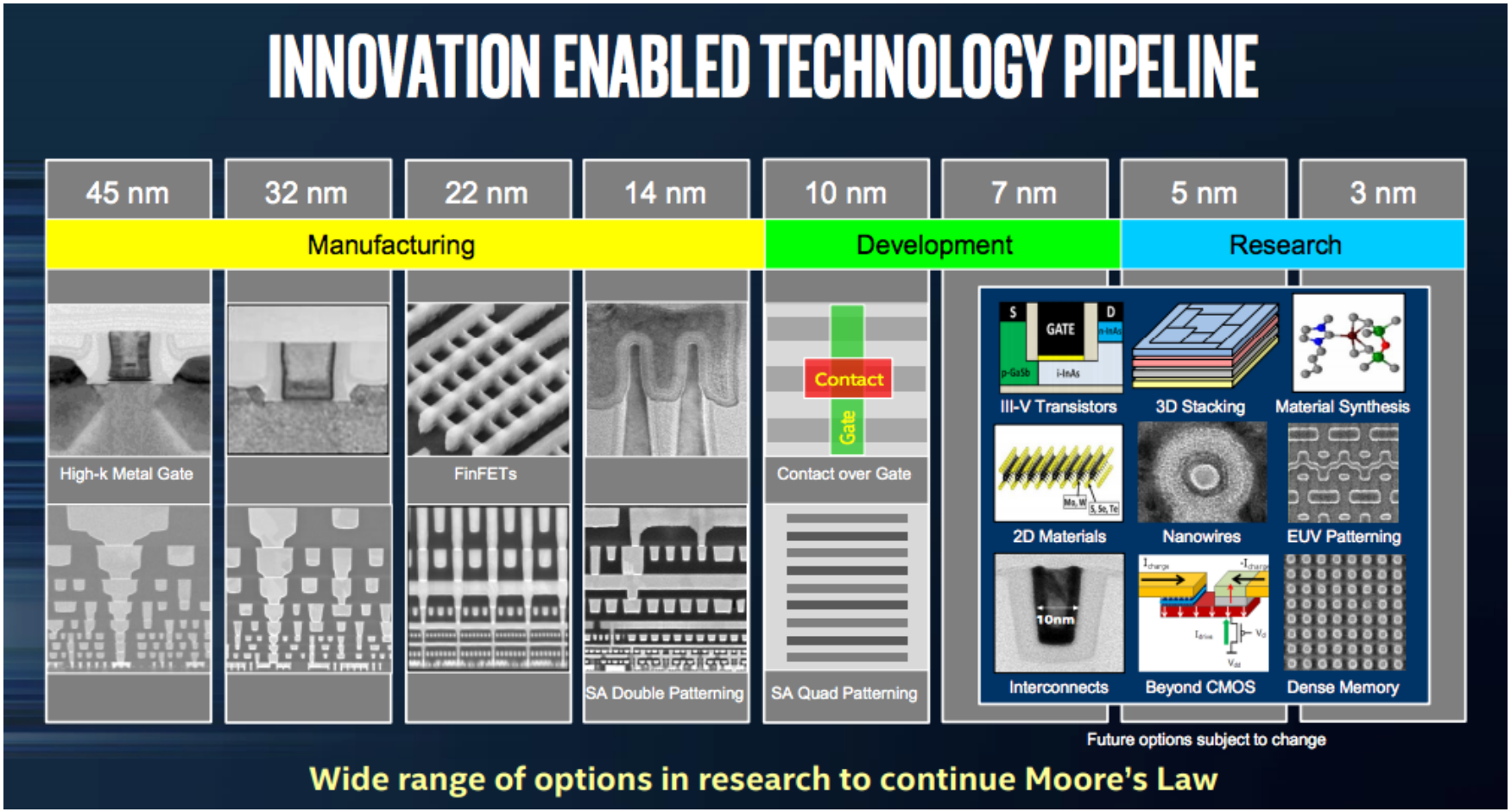


BUT…



Source:https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/09/mark-bohr-on-continuing-moores-law.pdf

BUT…



Source:https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/09/mark-bohr-on-continuing-moores-law.pdf

Big Ideas

- Moderately predictable VLSI Scaling
  - unprecedented capacities/capability growth for engineered systems
  - …but not for much longer

Admin

- HW 4 due Friday 10/8