ESE532: System-on-a-Chip Architecture

Day 21: April 5, 2017
VLSI Scaling

Today

- VLSI Scaling Rules
- Effects
- Historical/predicted scaling
- Variations (cheating)
- Limits
- **Note:** gory equations

→ goal is to understand trends

– Give equations … then push through scaling implications together

Message

- Technology advances rapidly
- Must account for in understanding
  - …platform will be available
  - …platforms will be inexpensive
  - …what our competitors can build
  - …new challenges and opportunities

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: Custom SoC

- 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

Scaling

- **Old 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×F

ITRS Roadmap

- Semiconductor Industry rides this scaling curve
- Try to predict where industry going
  - (requirements…self fulfilling prophecy)
- http://public.itrs.net
- http://www.semiconductors.org/main/2015_international_technology_roadmap_for_semiconductors_itrs/

Preclass

- Scale factor S from 28nm → 20nm?

MOS Transistor *Scaling* (1974 to present)

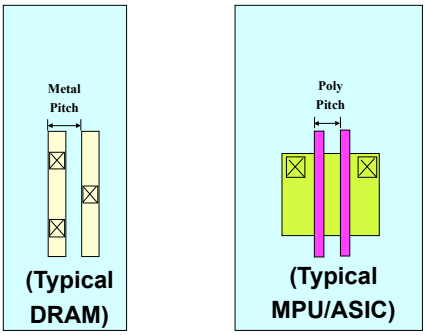
S=0.7

[0.5x per 2 nodes]

Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

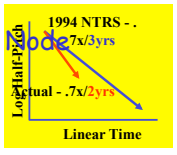
Half Pitch (= Pitch/2) Definition



Source: 2001 ITRS - Exec. Summary, ORTC Figure

[from Andrew Kahng]

Scaling Calculator + 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

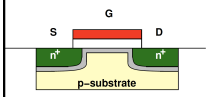
[from Andrew Kahng]

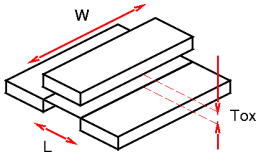
Warning

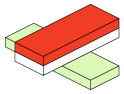
- Dive into detail equations
  - Not expect you necessarily know before
    - Unless took 370, 570, 534…
  - Won't expect you use later
- …but, you want to have an idea of the implications (area, performance, energy)
- If I just showed you results
  - I think would be hard to follow
  - Not engaged
- So, we will do calculations together

Dimensions

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







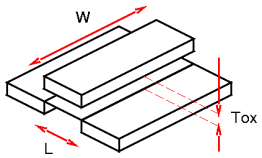
Oblique

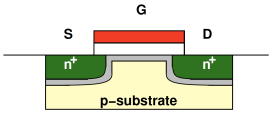
Side

Top

Scaling

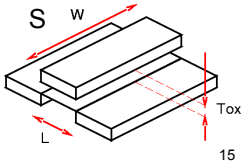
- Channel Length (L)
- Channel Width (W)
- Oxide Thickness (T_{ox})
- Doping (N_a)
- Voltage (V)





Full Scaling (Ideal Scaling, Dennard 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
- These are more the take-aways

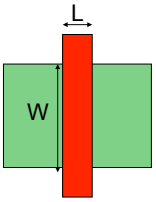
Area

- F➔ FS
- Area impact?
- $A = L \times W$
- A ➔ AS^2

- 28nm ➔ 20nm
- 50% area
- 2× capacity same area

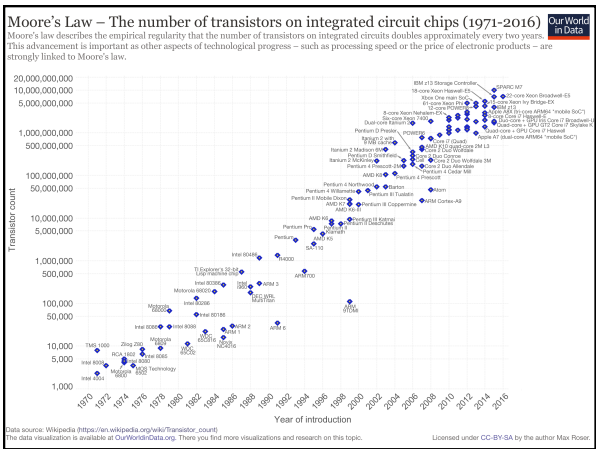
$S=0.7$

[0.5x per 2 nodes]



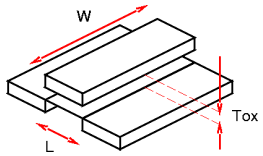
Preclass

- When will have 100-core processor
  - What feature size?
  - What time frame?



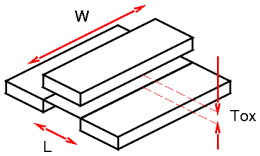
Capacitance

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



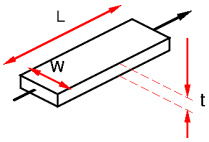
Capacitance

- 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

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

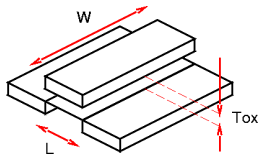
Current

- Saturation Current scaling?

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

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

$I_d \rightarrow S \times I_d$



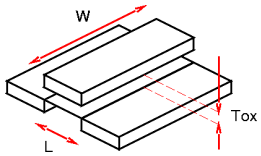
Current

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

- Velocity Saturation Current scaling:

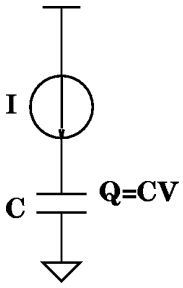
$$I_{DS} \approx v_{sat} C_{OX} W \left(V_{GS} - V_T - \frac{V_{DSAT}}{2} \right)$$

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



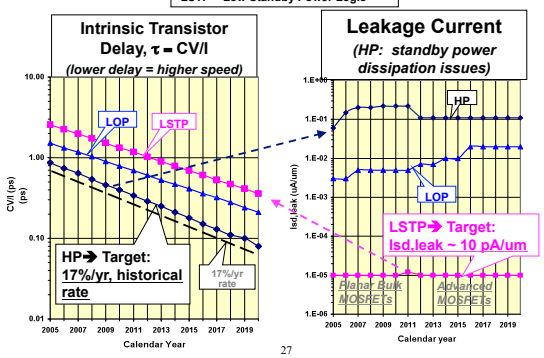
Gate Delay

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

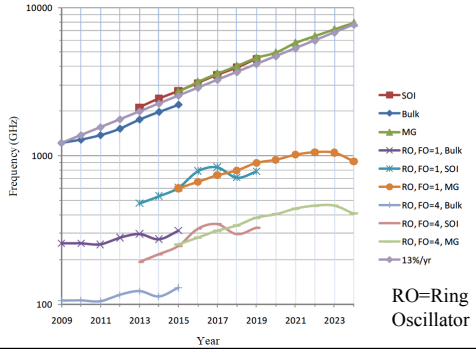


Overall Scaling Results, Transistor Speed and Leakage.
Preliminary Data from 2005 ITRS.

•HP = High-Performance Logic
•LOP = Low Operating Power Logic
•LSTP = Low Standby Power Logic



ITRS 2009 Transistor Speed



RO=Ring Oscillator

Wire Delay

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

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

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?

Energy

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

Power Dissipation (Dynamic)

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

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

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

Power Density

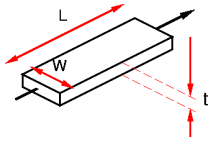
- P/A → P/A very important
- Says, it doesn't get any harder to cool as we scale exponentially to
  - More gates switching
  - Higher clock rates
- Don't create an cooling bottleneck

Cheating…

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

Improving Resistance

- R=ρL/(W×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/

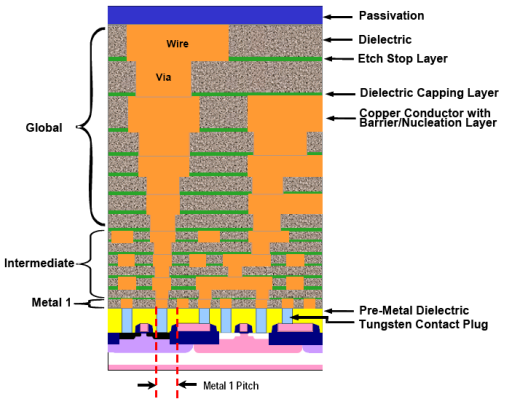
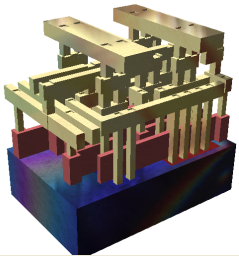


Figure 70 Cross-section of Hierarchical Scaling—MPU Device

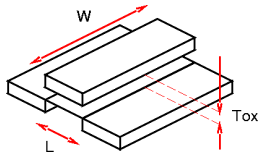
3D View

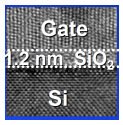


Source: https://en.wikipedia.org/wiki/File:Silicon_chip_3d.PNG

Capacitance and Leakage

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



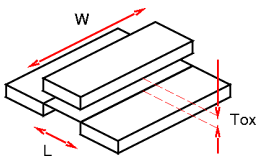


What's wrong with Tox = 1.2nm?

source: Borkar/Micro 2004

Capacitance and Leakage

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



What might we do?

Reduce Dielectric Constant ε (interconnect)

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

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		

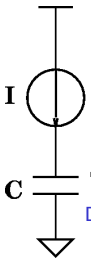
*Estimated value.

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

Improving Gate Delay

Normal scale

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



What happens if don't scale V? (note V impacts I)

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

…But Power Dissipation (Dynamic)

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

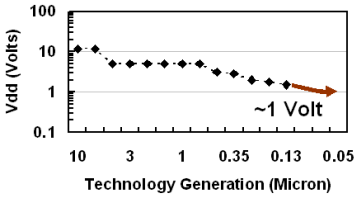
If not scale V, power dissipation not scale down.

…And Power Density

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

…this is where some companies have gotten into trouble…

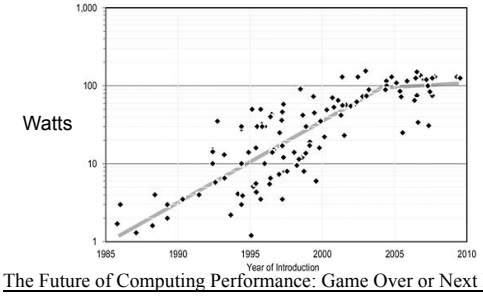
Historical Voltage Scaling



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

- …and, we're running into limits prevent voltage scaling going forward.

uP Power Density

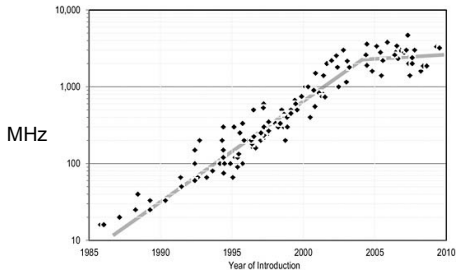


The Future of Computing Performance: Game Over or Next Level?
National Academy Press, 2011
http://www.nap.edu/catalog.php?record_id=12980

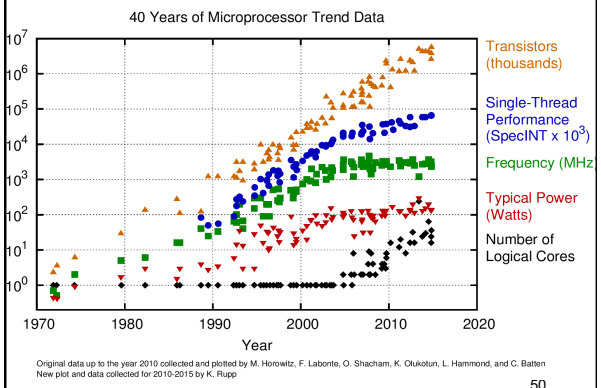
Power Density

- Has become the new bottleneck
- Can put more transistors on a chip than we can afford to turn on
  - Can afford to operate at the frequency at which they are capable of switching
- Energy and Power, not capacity, limits performance

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



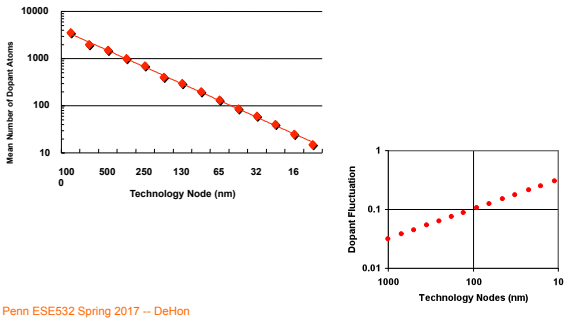
Physical Limits

- Doping?
- Features?

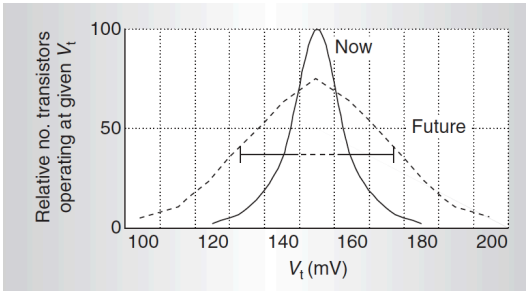
Physical Limits

- 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



Leads to Variation



[Borkar, IEEE Micro Nov.-Dec. 2005]

Conventional Scaling

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

It Was a Great Ride, But It's Ending

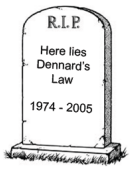
Moore's Law put lots more transistors on a chip…

…but it's Dennard's Law that made them useful

Dennard's Law has been repealed.

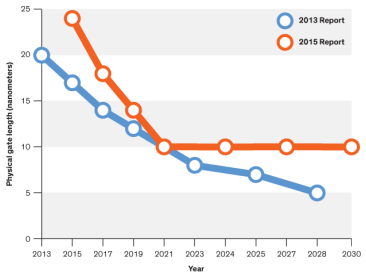
Moore's Law: 2x transistors every 2 years

Dennard's Law: transistors will be faster and lower energy



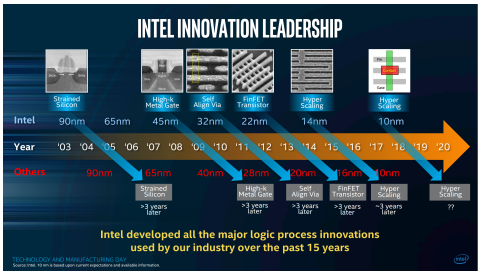
Bob Colwell
March 2013

Feature Size Scaling ITRS



IEEE Spectrum, Sep. 2016

Intel says they've got 10nm covered



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
- Dennard Scaling ended
- Feature scaling end at 10nm?
- Transistor count integration may continue…
  - E.g. 3D

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 $1/S^2$
  - 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

Admin

- Project 4x and area Milestone
  - Due Friday