

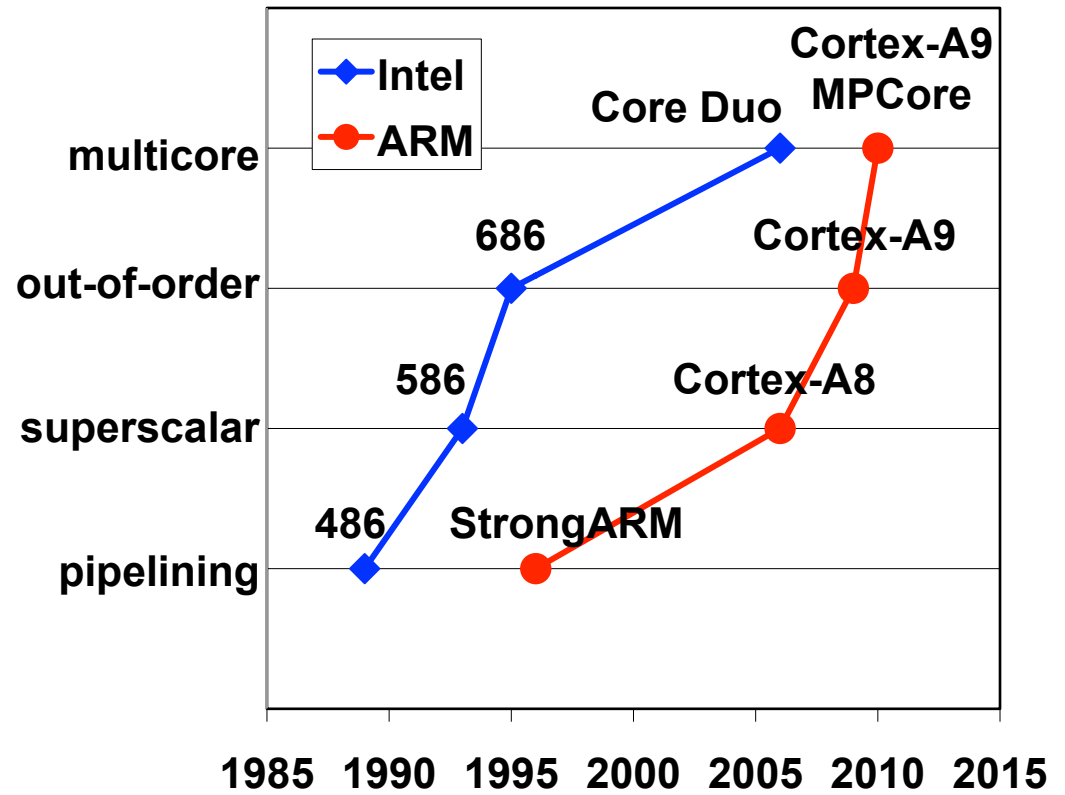
CSE 291:

Mobile Application

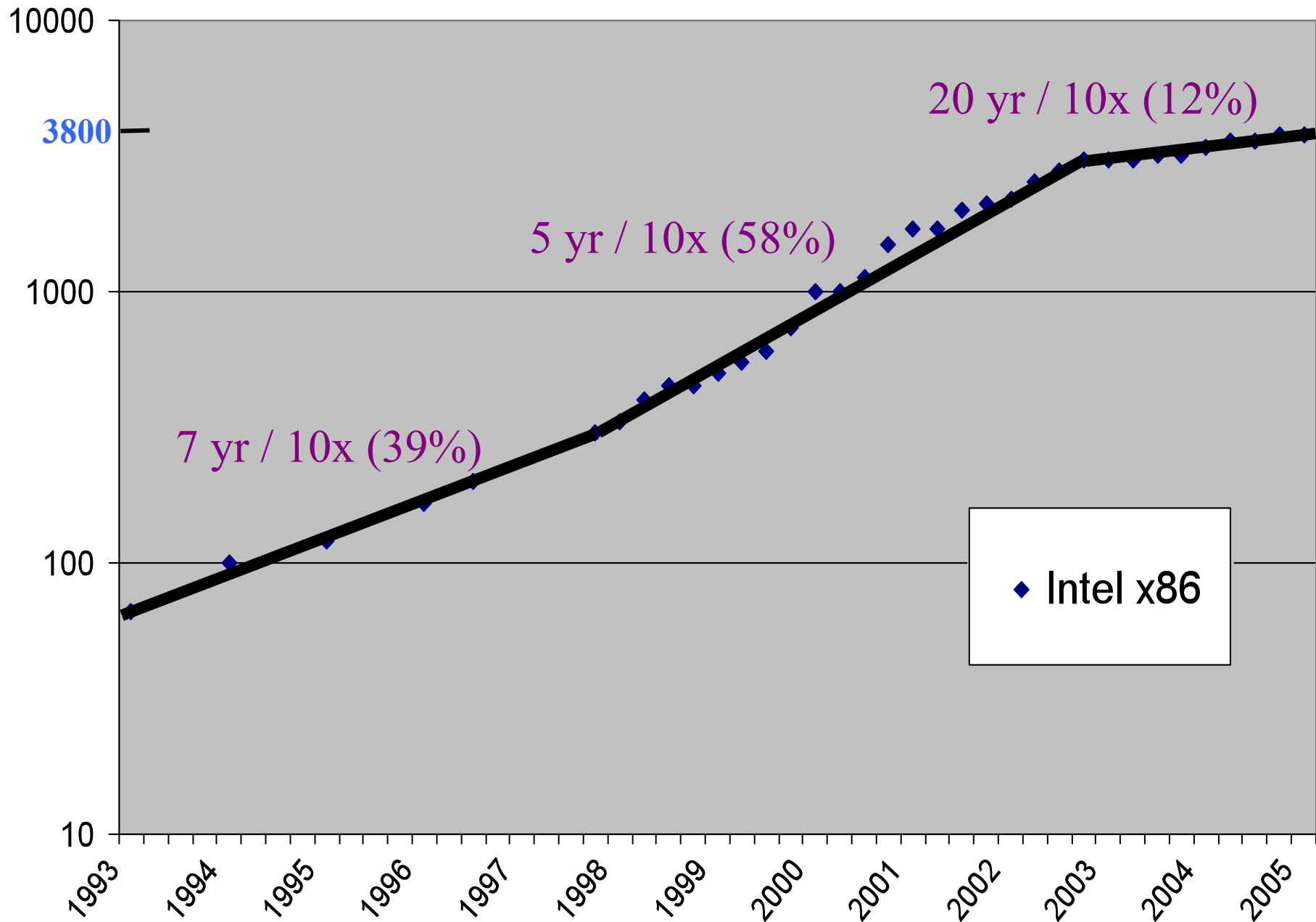
Processor Design

Mobile Application Processors are where the action are

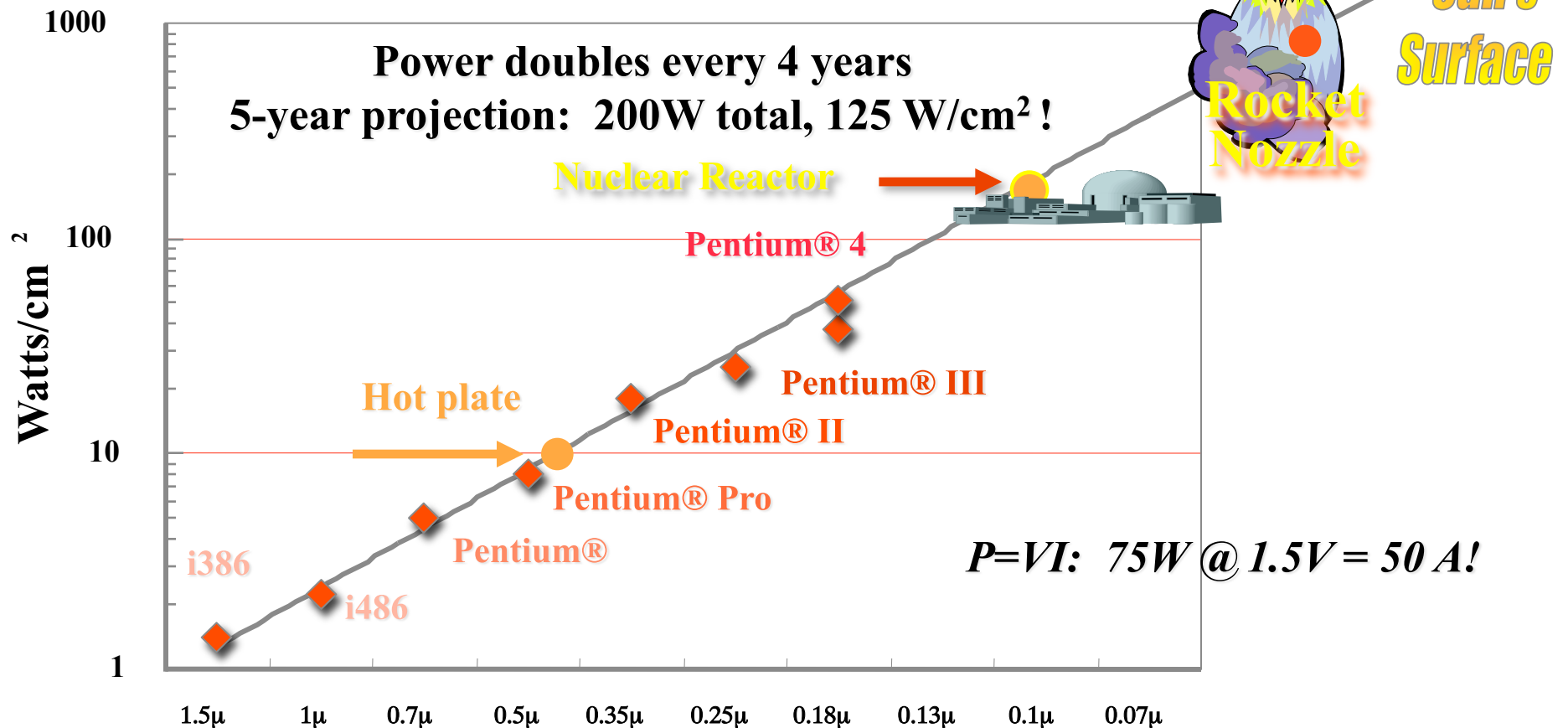
- The evolution of mobile application processors mirrors that of microprocessors mirrors that of mainframes ..
- Not much left to borrow!
- Faster iterative cycle than desktops:
obsolete vs. worn out
- Mobile Application Processors are the new centroid of microprocessor evolution!



Desktop micros: frequency versus time



Power Density



From "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies"
 – Fred Pollack, Intel Corp. Micro32 conference key note - 1999.

Course Staff



Prof. Michael Taylor
CSE 3202

Office Hours: Right After Class



Dr. Jack Sampson
CSE 3204

Office Hours: TBD

Where to find information:

Wherever you can get it!

IEEE Explore website, free from UCSD network
ISSCC (Chip Tapeout Papers)

ACM Portal website, free from UCSD network

google.com – I'm serious!

Powerpoint presentations by Qualcomm, Nvidia, Etc.

Commentary articles

Clever search terms

Reverse engineering sites (chipworks)

Microprocessor Report leaked articles

anandtech.com – Technical fanboy site

fandroid.com etc – I'm serious!

qualcomm, ti, nvidia, google, arm – developer sites and blogs

Project Structure

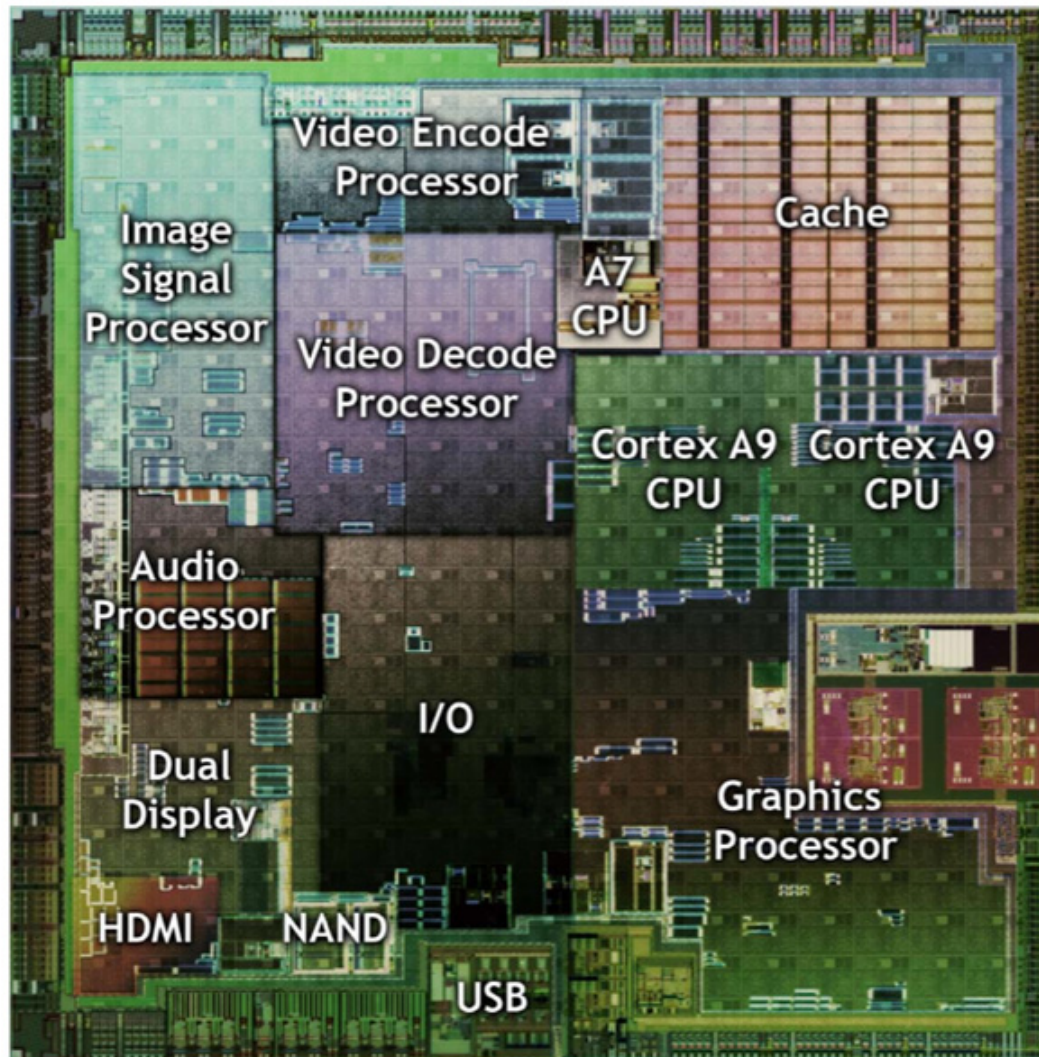
■ First Two Weeks:

- Literature Review
- Related Work
- Project Proposal

■ Last 6 Weeks

- Project
- Midpoint Review
- Final Project Report
 - 6 page DAC submission
 - DAC submission deadline shortly after class ends
- Final Project Presentation
 - Last Day of Class

NVidia Tegra 2: Example MAP

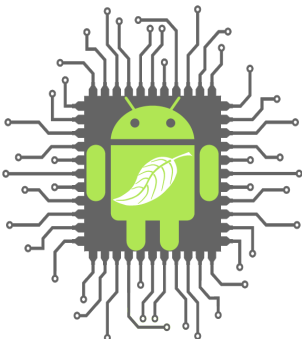


Discussion

- RISC vs CISC Battle, Round II
- Vertical Integration (Apple) versus Competitive Bazaar (Wintel/AndroidQualTISamidia)
- AMD vs. Intel
- Nvidia vs. Intel
- Nvidia vs. AMD/ATI
- Mobile Monopoly Theory
 - Qualcomm
 - (with AMD Adreno)
 - Nvidia
 - (with Icera)
 - TI
 - (historical, with OMAP)
 - Apple
 - Soon enough
- Mobile Snafus
 - Intel sells Xscale to Marvell
 - AMD sold Adreno Mobile GPU (\$65m) to Qcom

GreenDroid: An Architecture for the Dark Silicon Era

*UCSD Center for Dark Silicon
Department of Computer Science and Engineering,
University of California, San Diego*



This Talk

The Dark Silicon Problem

*How to use Dark Silicon to improve energy efficiency
(Conservation Cores)*

The GreenDroid Mobile Application Processor

GreenDroid Highlights

Where does dark silicon come from? And how dark is it going to be?

The Utilization Wall:

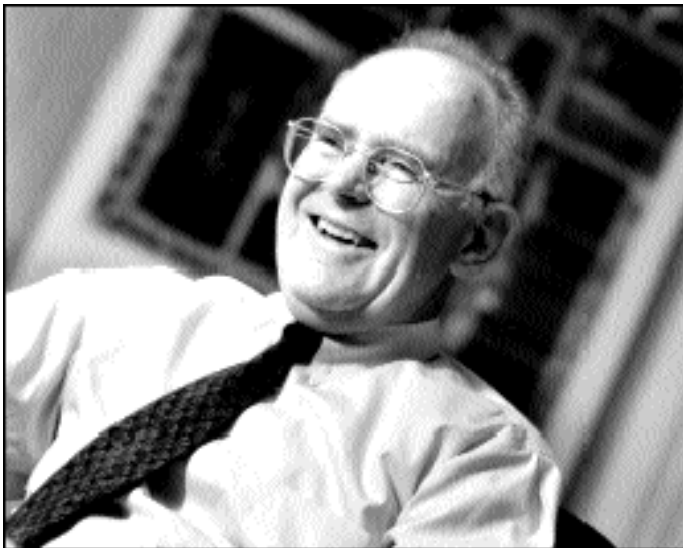
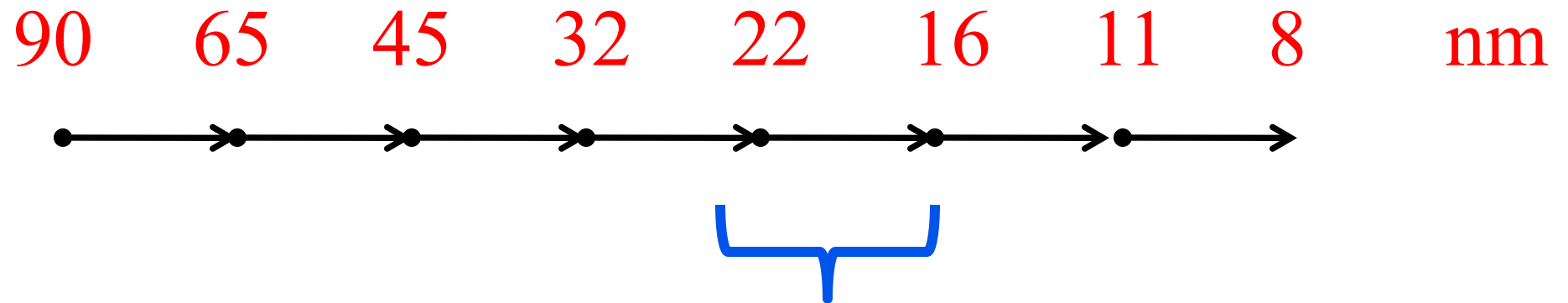
With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- Scaling theory
 - Transistor and power budgets are no longer balanced
 - Exponentially increasing problem!
- Experimental results
 - Replicated a small datapath
 - More "dark silicon" than active
- Observations in the wild
 - Flat frequency curve
 - "Turbo Mode"
 - Increasing cache/processor ratio

Scaling 101: Moore's Law



$$S = \frac{22}{16} = \sim 1.4x$$

Scaling 101:

Transistors scale as S^2

180 nm

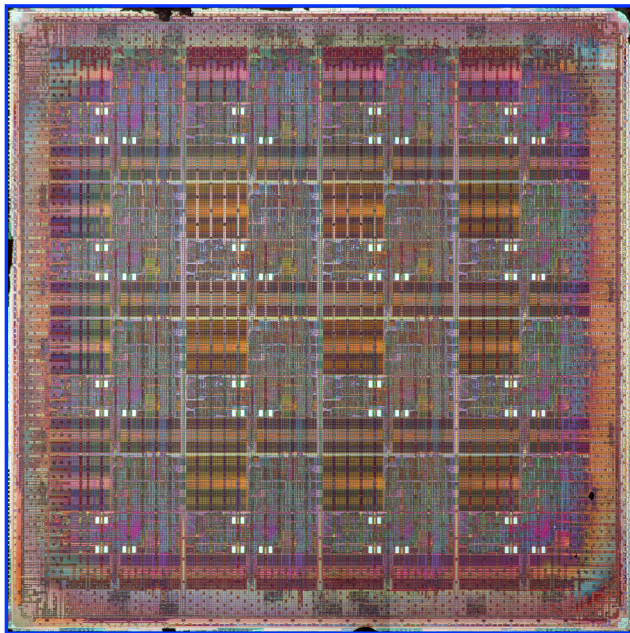
16 cores

$S = 2\times$

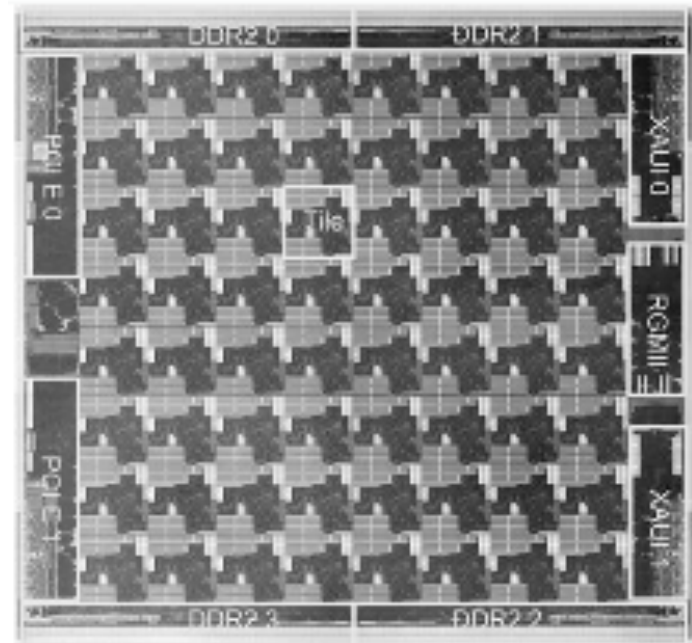
Transistors = $4\times$

90 nm

64 cores



MIT Raw



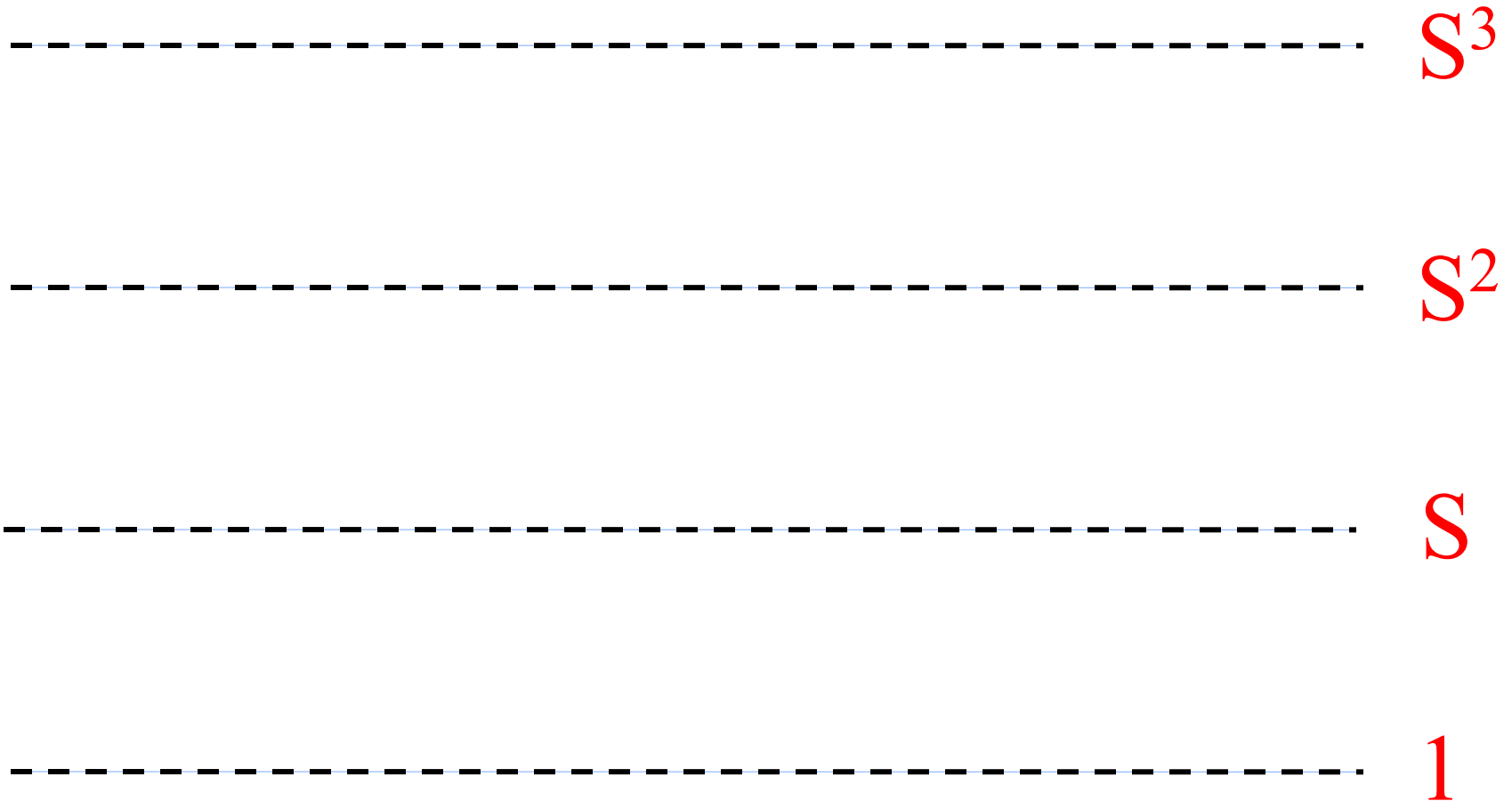
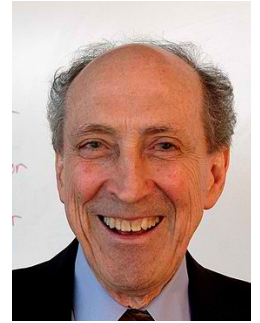
Tiler TILE64

Advanced Scaling:

Dennard: “Computing Capabilities

If $S=1.4\times \dots$

Scale by $S^3 = 2.8\times$ ”



Design of Ion-Implanted MOSFETs with Very Small Dimensions

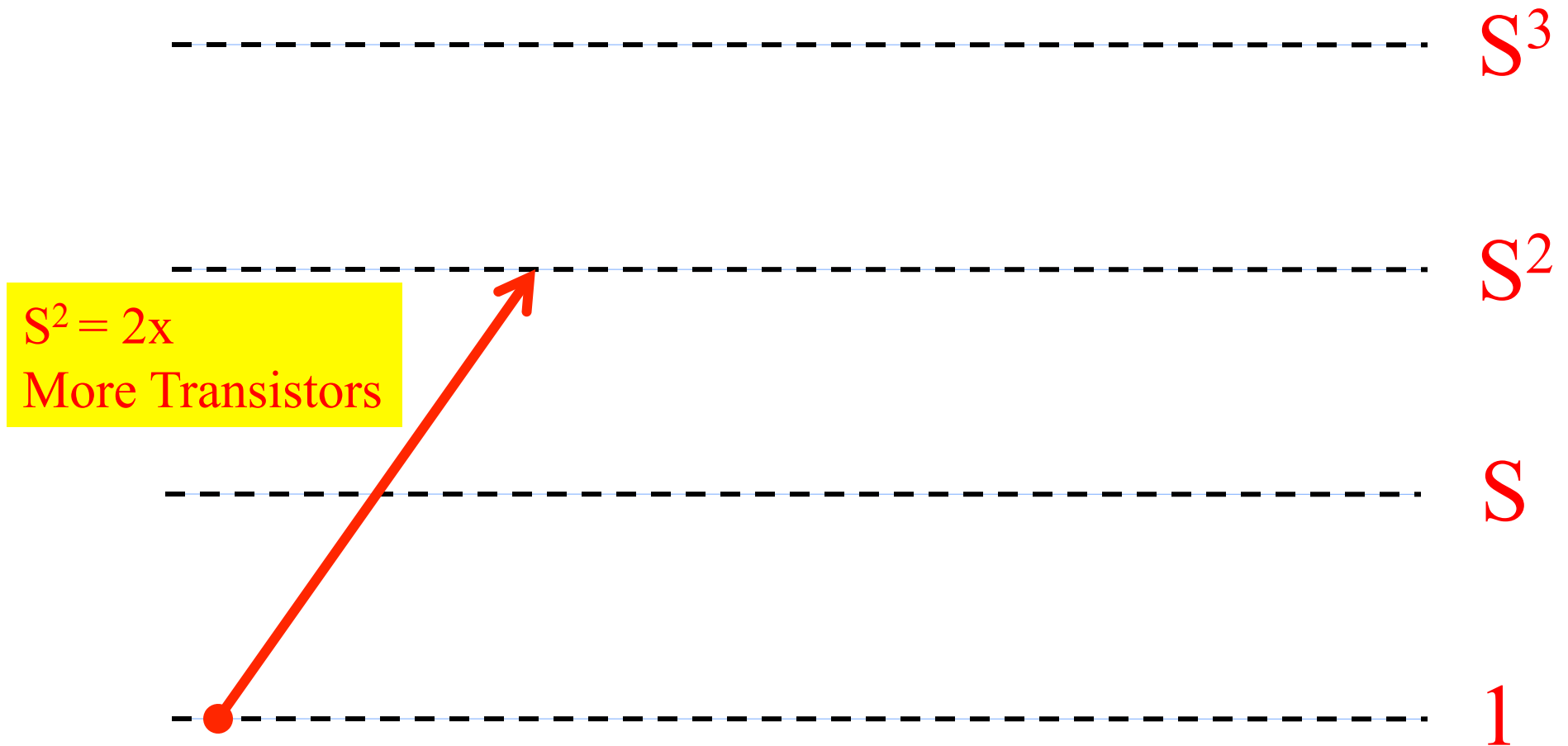
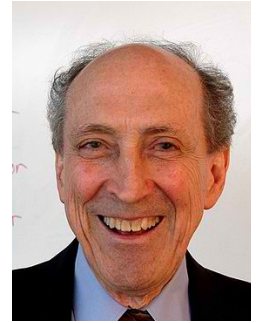
Dennard et al, 1974

Advanced Scaling:

Dennard: “Computing Capabilities

If $S=1.4\times \dots$

Scale by $S^3 = 2.8\times$ ”

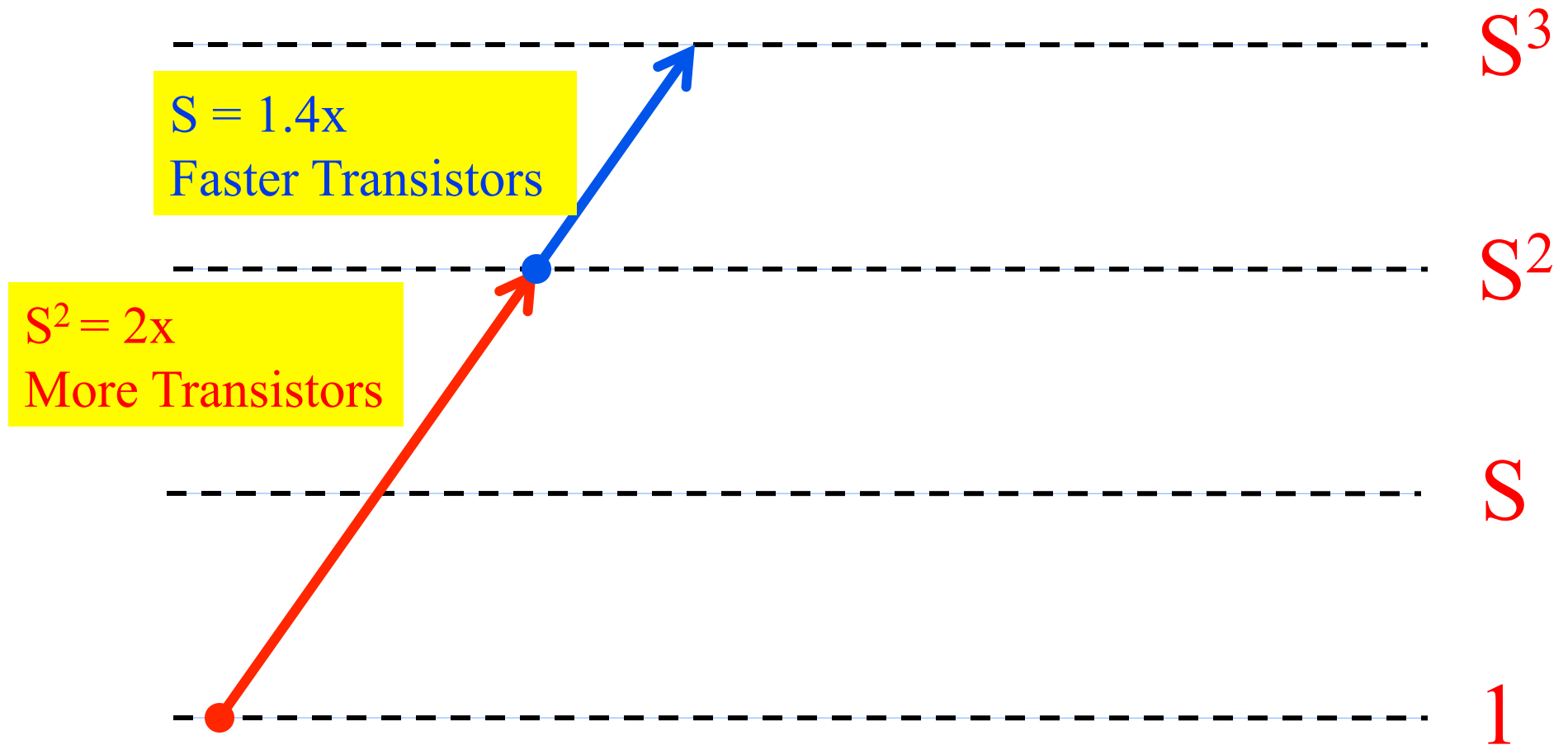
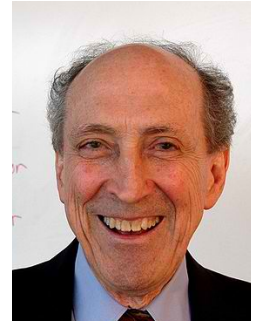


Advanced Scaling:

Dennard: “Computing Capabilities

If $S=1.4x$...

Scale by $S^3 = 2.8x$ ”

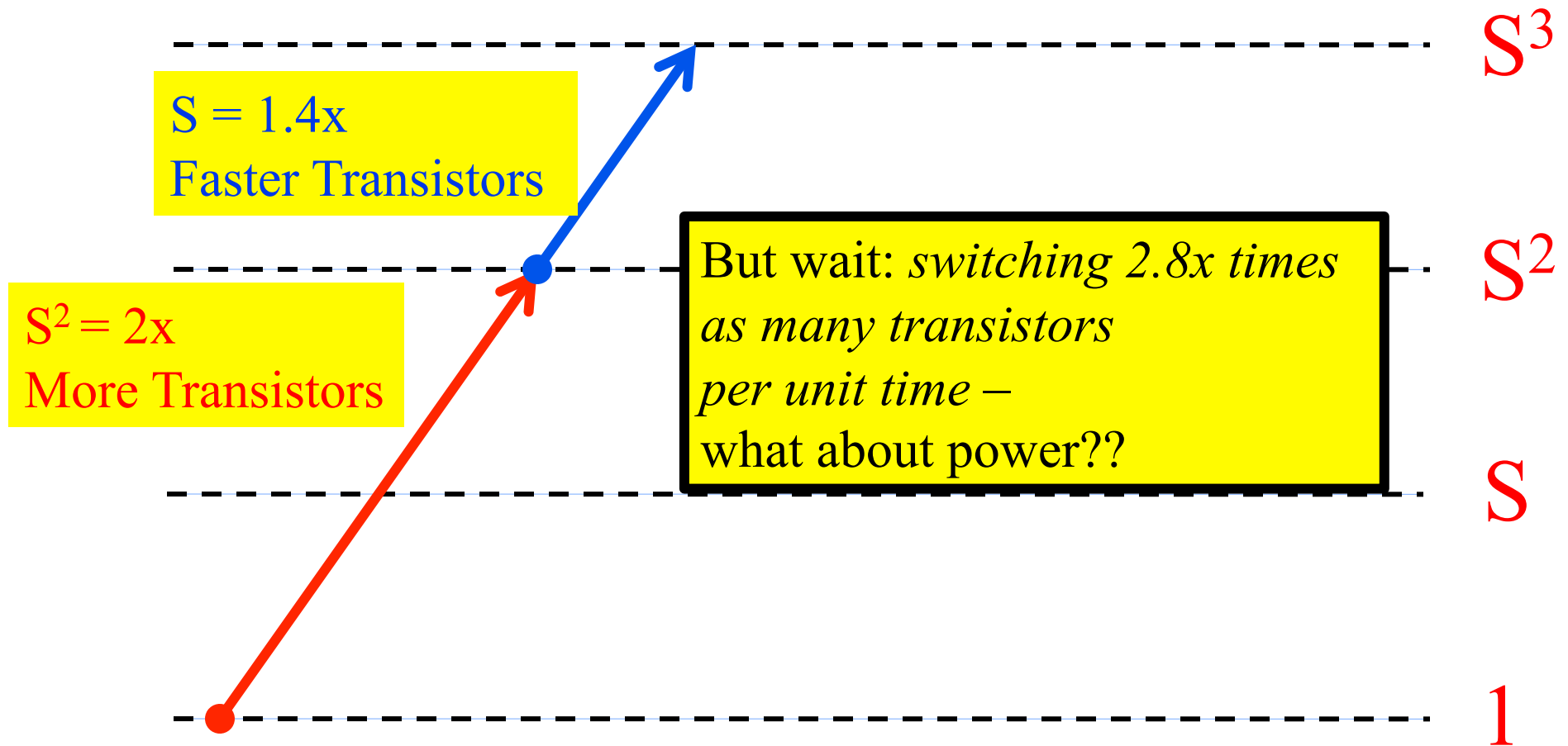
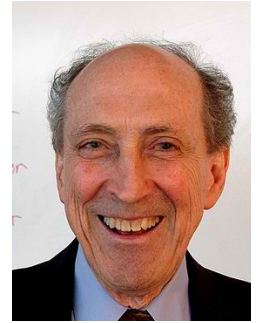


Advanced Scaling:

Dennard: “Computing Capabilities

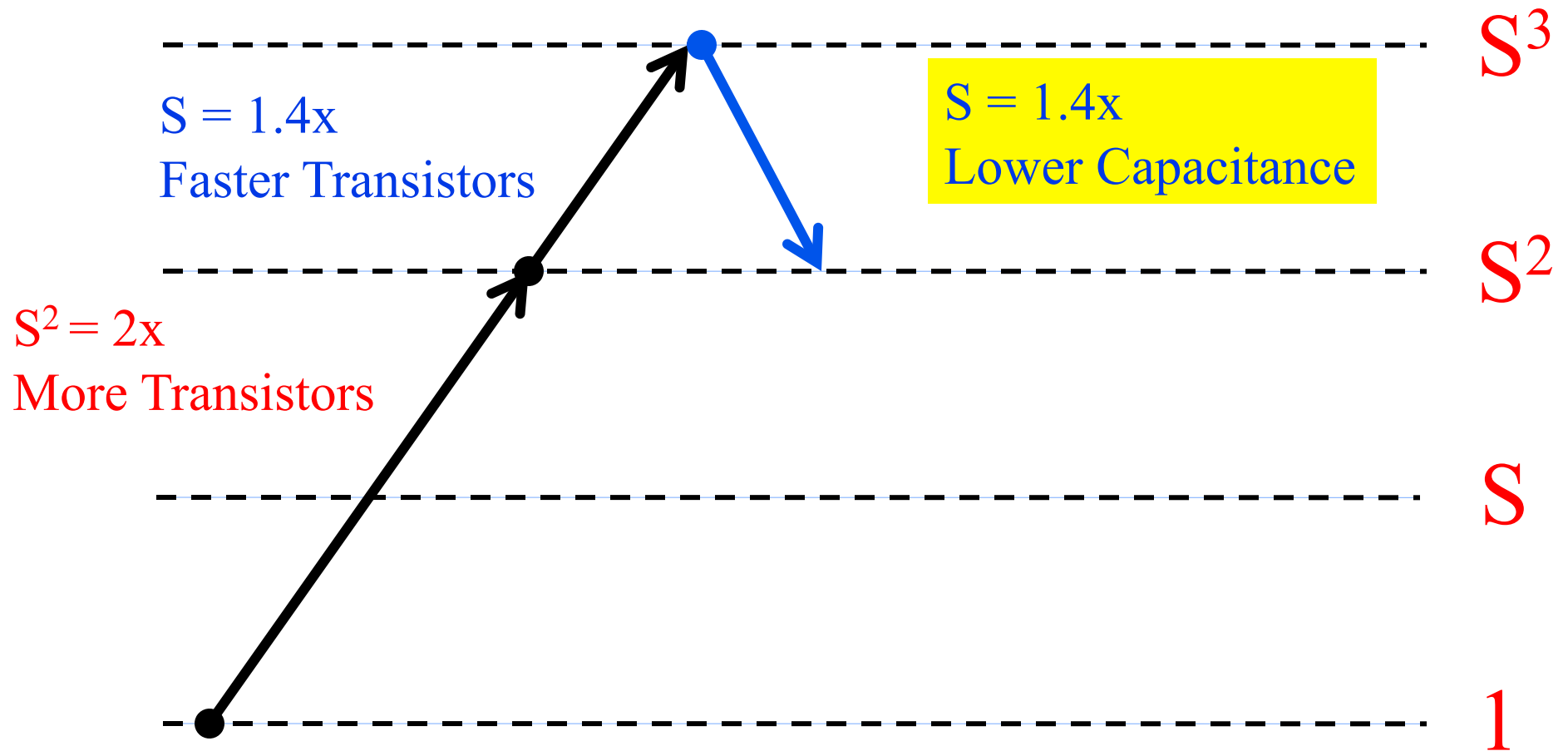
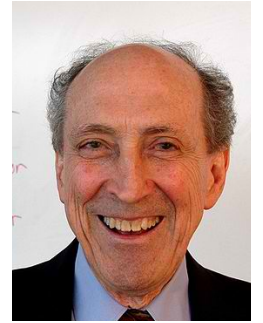
If $S=1.4x$...

Scale by $S^3 = 2.8x$ ”



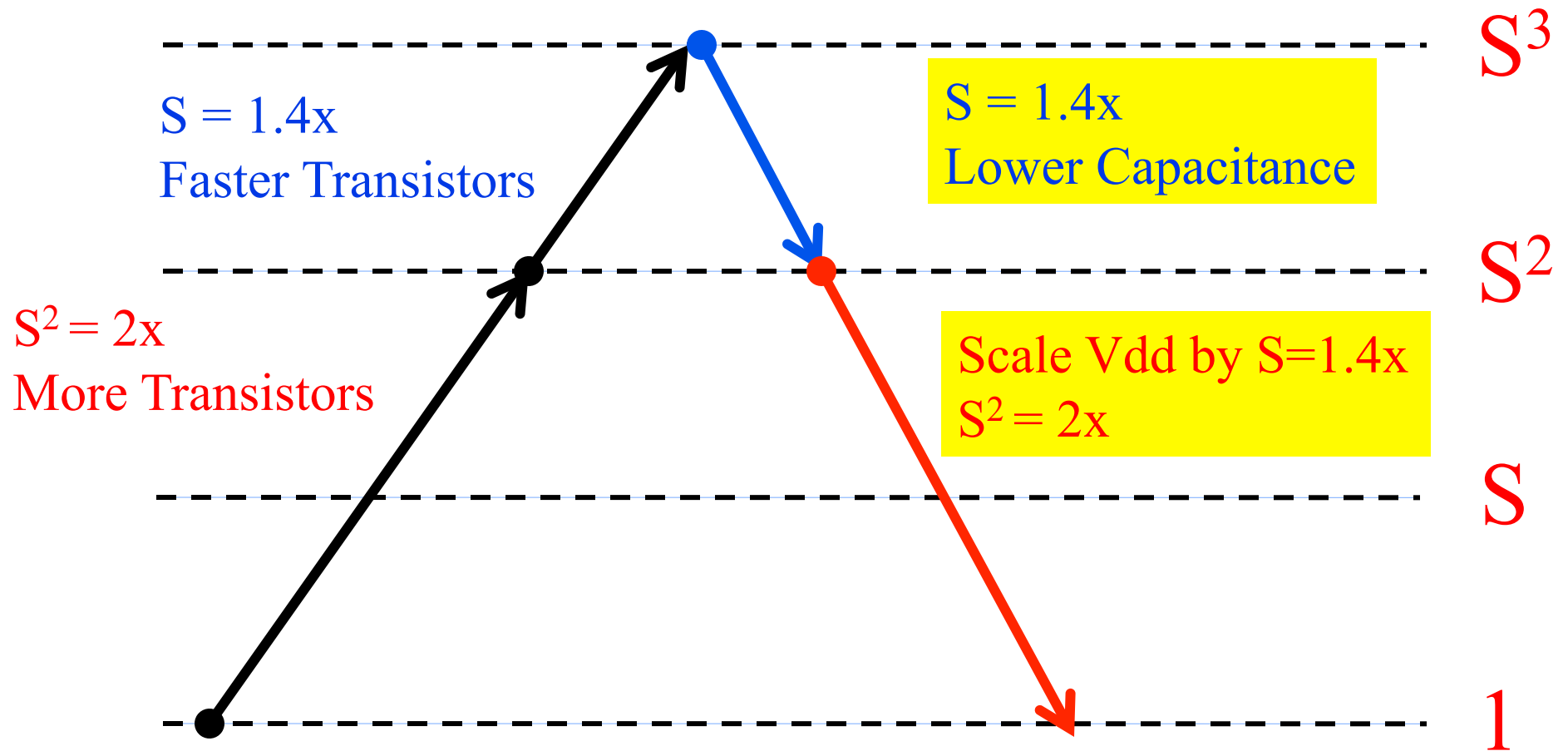
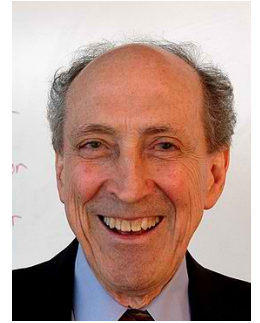
Dennard:

“We can keep power consumption constant”



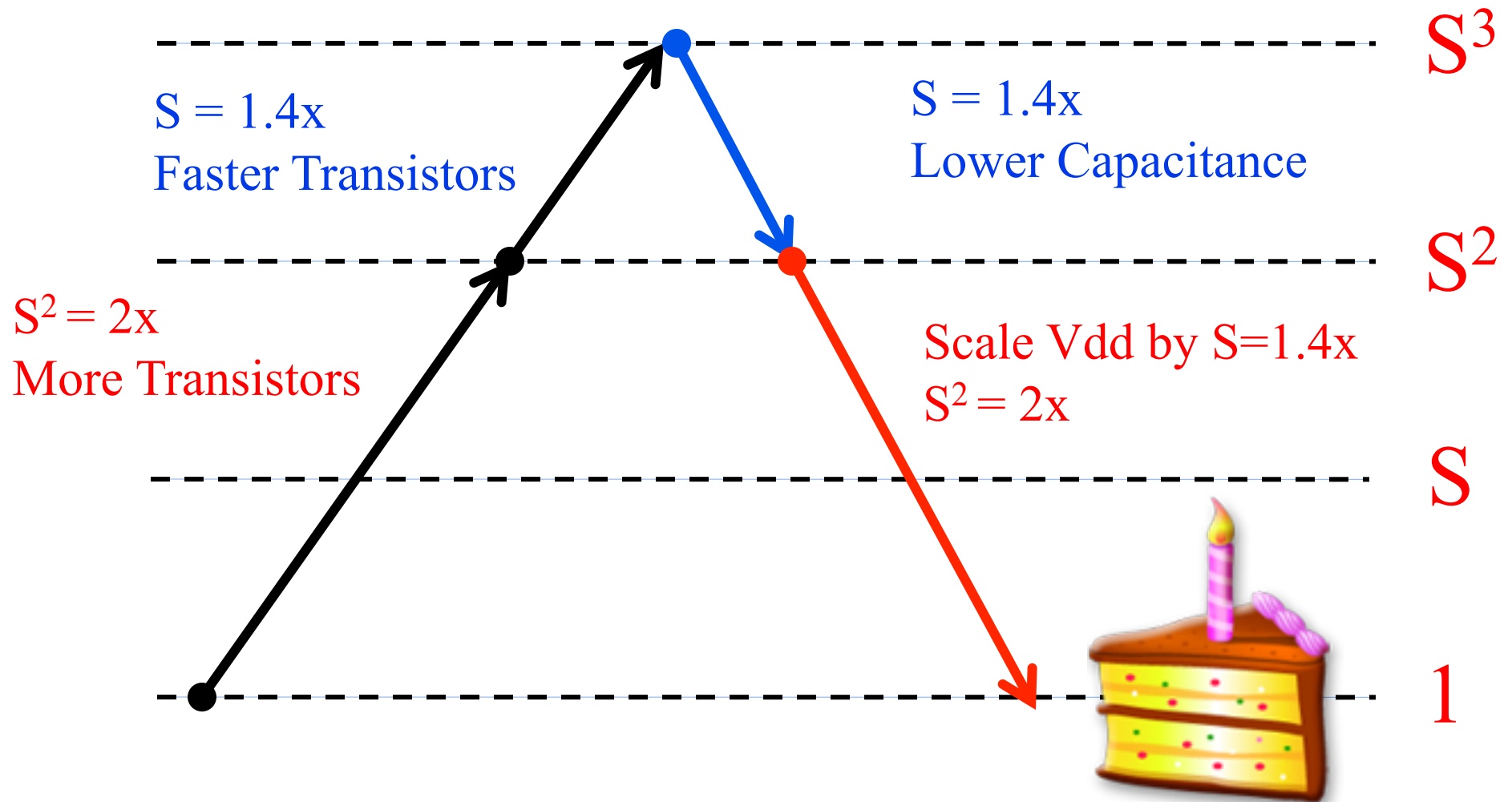
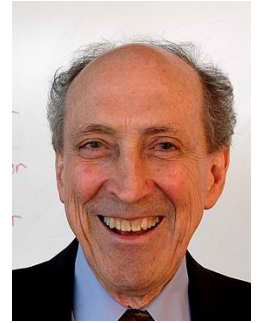
Dennard:

“We can keep power consumption constant”



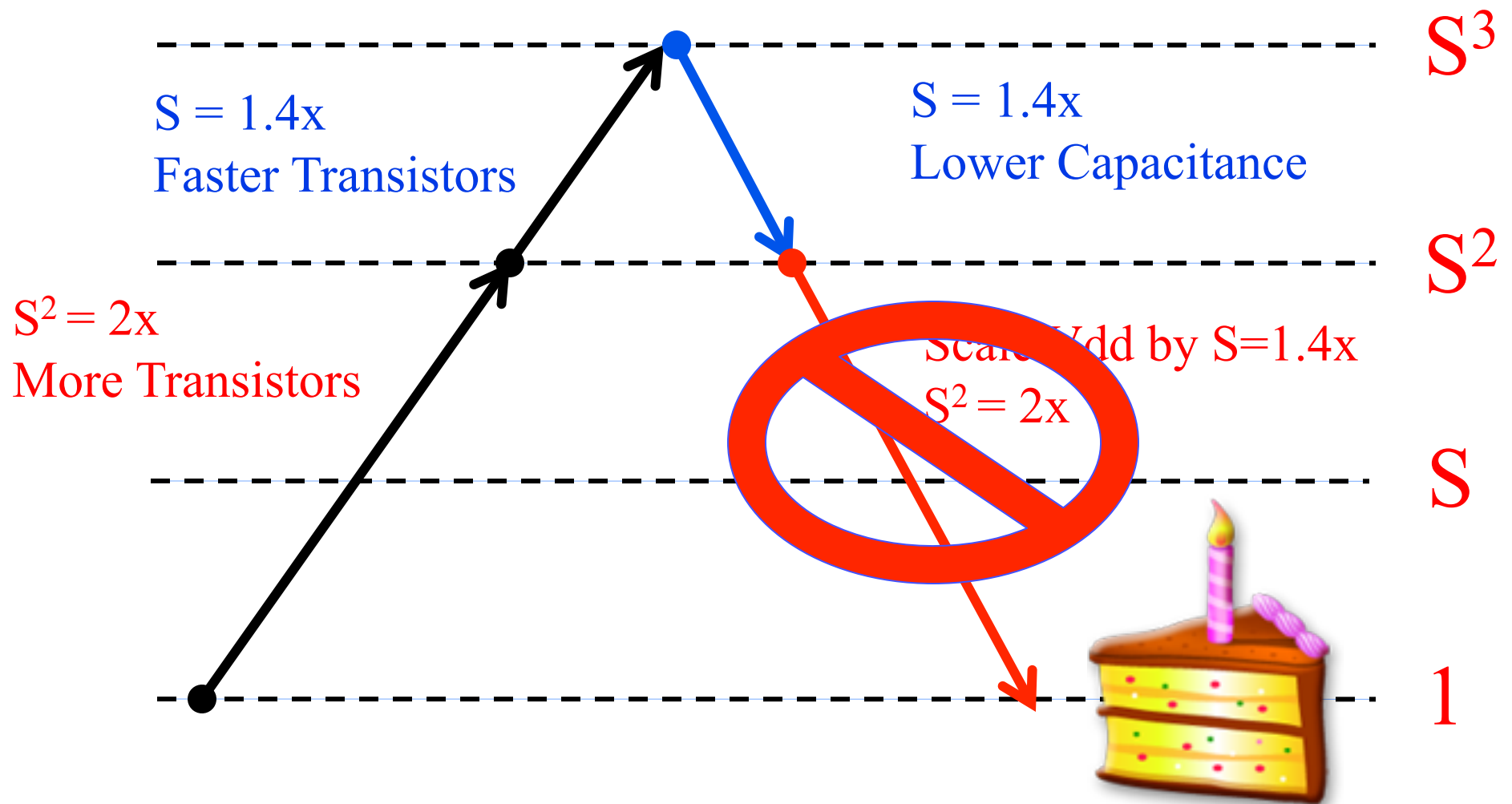
Dennard:

“We can keep power consumption constant”



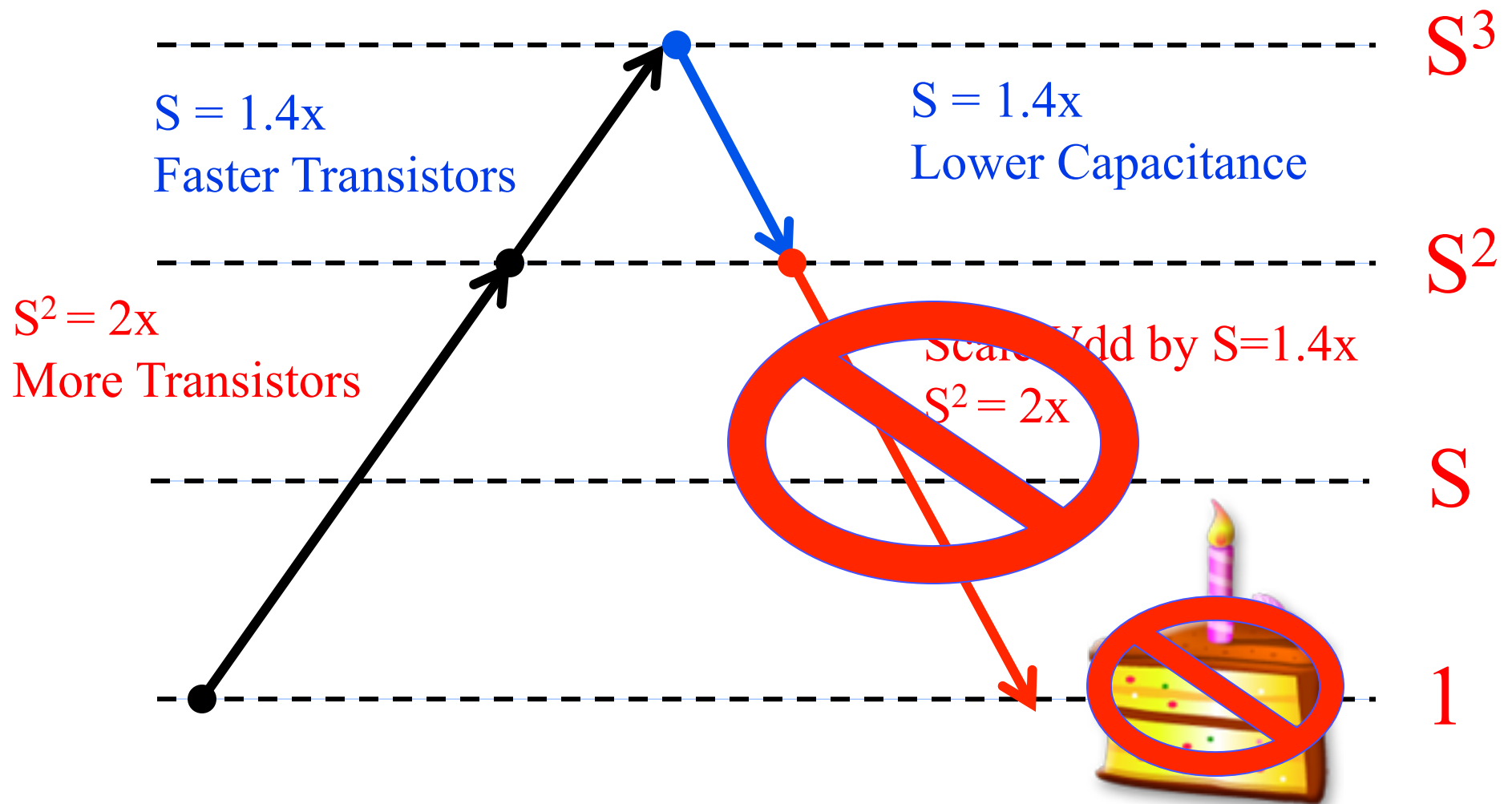
Fast forward to 2005:

Threshold Scaling Problems due to Leakage Prevents Us From Scaling Voltage

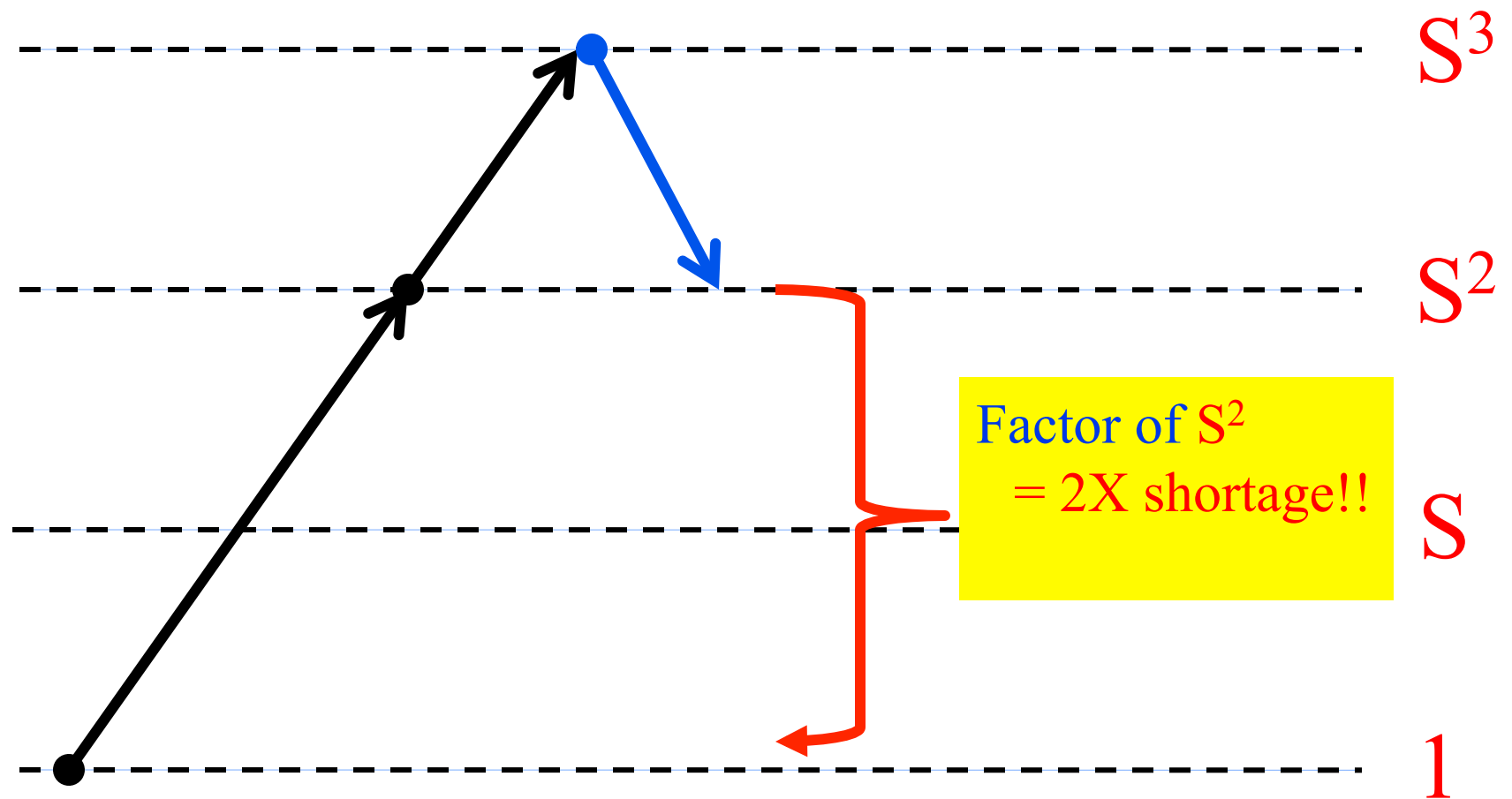


Fast forward to 2005:

Threshold Scaling Problems due to Leakage Prevents Us From Scaling Voltage



***Full Chip, Full Frequency Power Dissipation
Is increasing exponentially by 2x with
every process generation***



We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

■ Scaling theory

- Transistor and power budgets are no longer balanced
- Exponentially increasing problem!

■ Experimental results

- Replicated a small datapath
- More "dark silicon" than active

■ Observations in the wild

- Flat frequency curve
- "Turbo Mode"
- Increasing cache/processor ratio

Classical scaling

Device count	S^2	} 1
Device frequency	S	
Device cap (power)	$1/S$	
Device V_{dd} (power)	$1/S^2$	
Utilization ?		

Leakage-limited scaling

Device count	S^2	} S^2
Device frequency	S	
Device cap $\rightarrow$ power	$1/S$	
Device V_{dd} (power)	~ 1	
Utilization ?		

We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

■ Scaling theory

- Transistor and power budgets are no longer balanced
- Exponentially increasing problem!

■ Experimental results

- Replicated a small datapath
- More "dark silicon" than active

■ Observations in the wild

- Flat frequency curve
- "Turbo Mode"
- Increasing cache/processor ratio

Classical scaling

Device count	S^2	} 1
Device frequency	S	
Device cap (power)	$1/S$	
Device V_{dd} (power)	$1/S^2$	
Utilization ?	1	

Leakage-limited scaling

Device count	S^2	} S^2
Device frequency	S	
Device cap $\rightarrow$ power	$1/S$	
Device V_{dd} (power)	~ 1	
Utilization ?	$1/S^2$	

We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

■ Scaling theory

- Transistor and power budgets are no longer balanced
- Exponentially increasing problem!

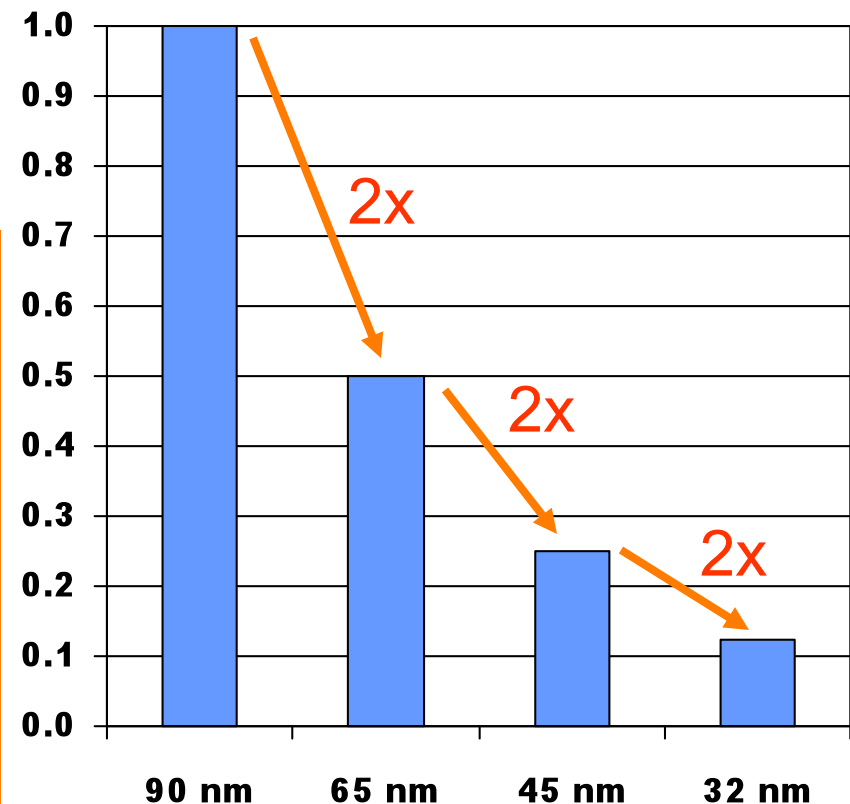
■ Experimental results

- Replicated a small datapath
- More "dark silicon" than active

■ Observations in the wild

- Flat frequency curve
- "Turbo Mode"
- Increasing cache/processor ratio

Expected utilization for fixed area and power budget



We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

■ Scaling theory

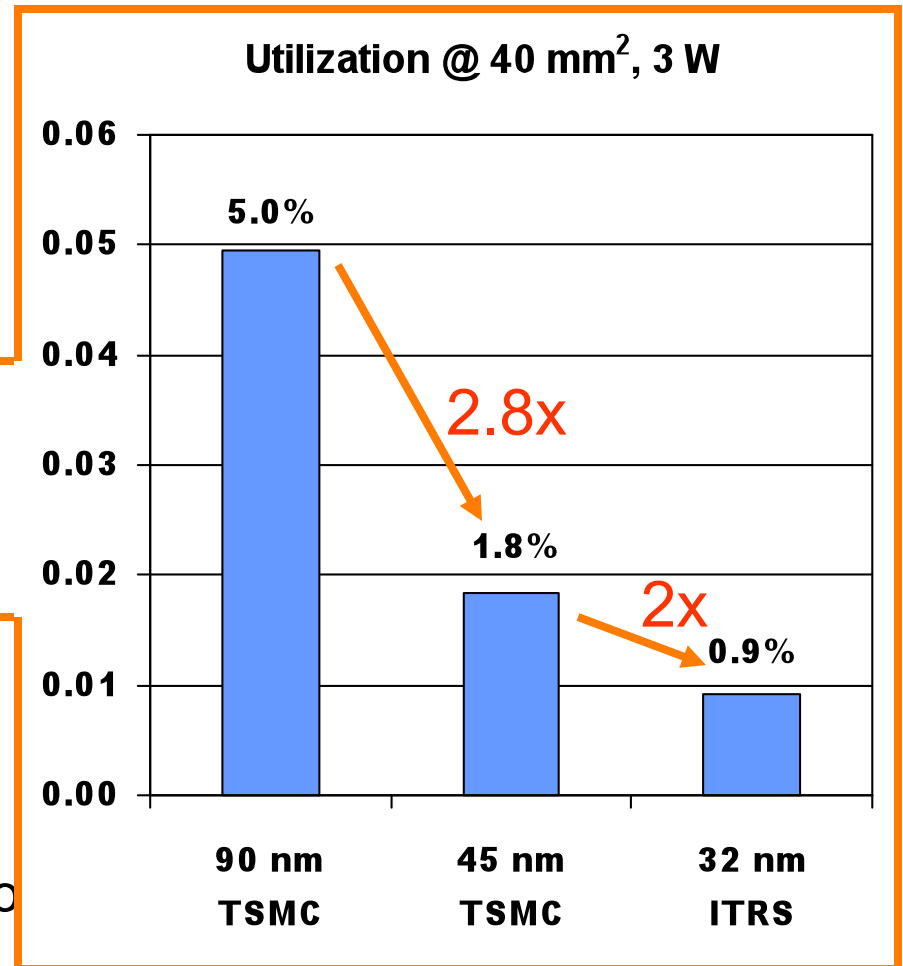
- Transistor and power budgets are no longer balanced
- Exponentially increasing problem!

■ Experimental results

- Replicated a small datapath
- More "dark silicon" than active

■ Observations in the wild

- Flat frequency curve
- "Turbo Mode"
- Increasing cache/processor ratio



We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

■ Scaling theory

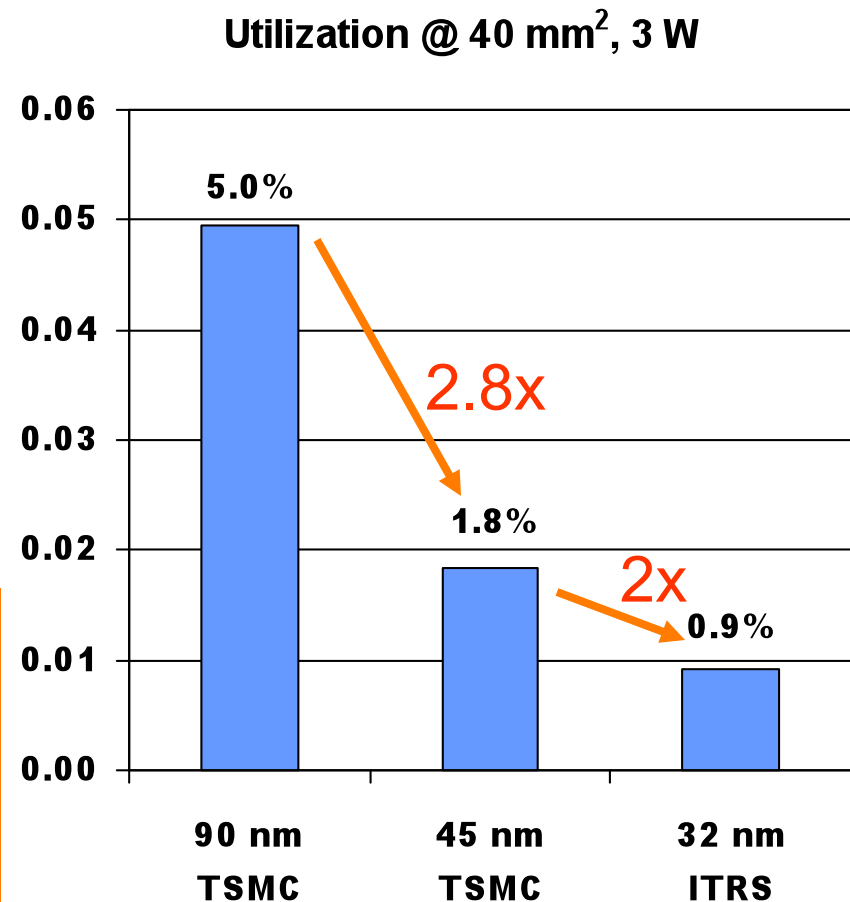
- Transistor and power budgets are no longer balanced
- Exponentially increasing problem!

■ Experimental results

- Replicated a small datapath
- More "dark silicon" than active

■ Observations in the wild

- Flat frequency curve
- "Turbo Mode"
- Increasing cache/processor ratio



We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

■ Scaling theory

Utilization @ 40 mm², 3 W

The utilization wall will change the way everyone builds processors.

■ Ex

– Repeated a small datapath

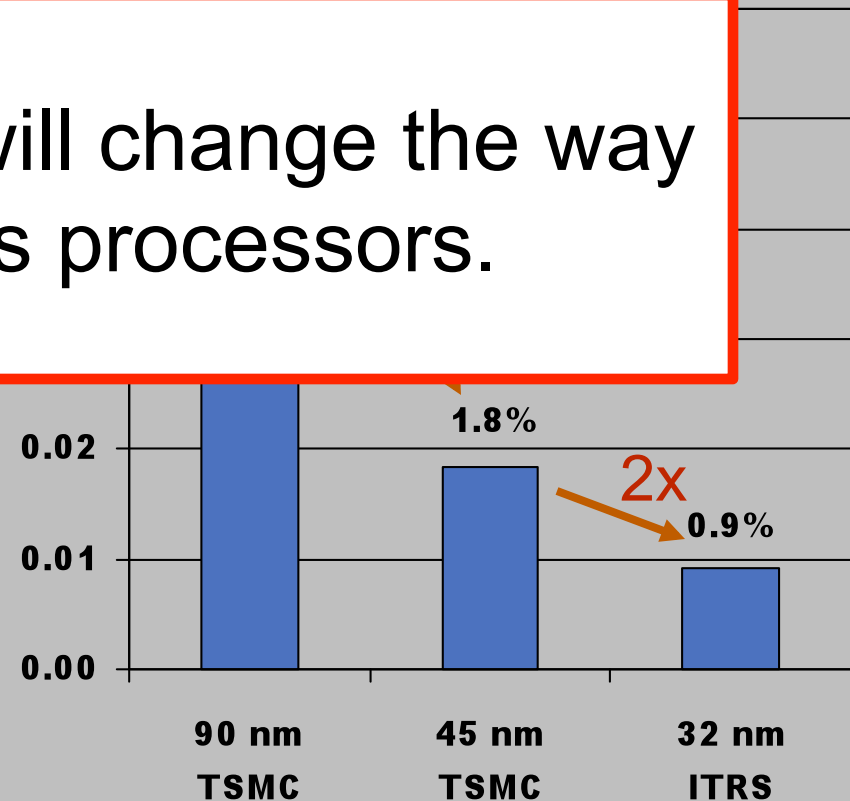
– More "dark silicon" than active

■ Observations in the wild

– Flat frequency curve

– "Turbo Mode"

– Increasing cache/processor ratio



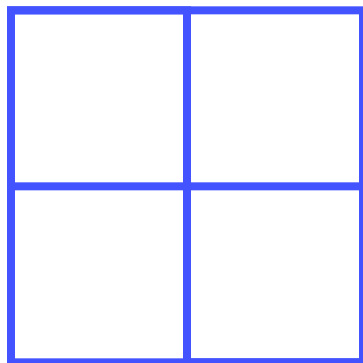
Utilization Wall: Dark Silicon Leads to Dark Implications for Multicore

Spectrum of tradeoffs
between # of cores and
frequency

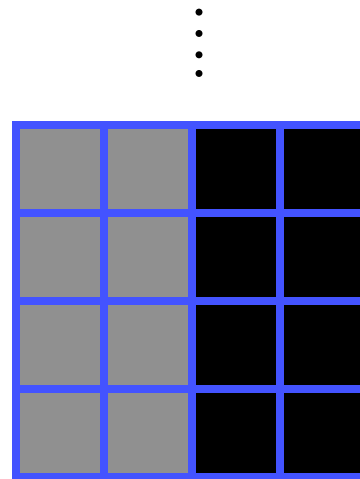
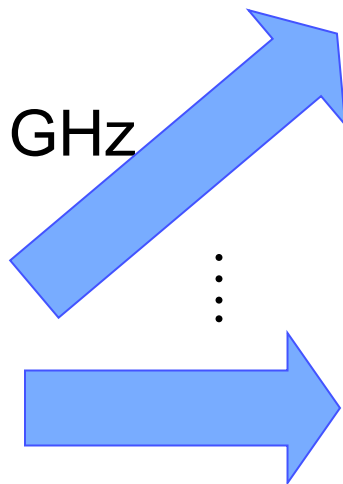
Example:

65 nm $\rightarrow$ 32 nm ($S = 2$)

4 cores @ 1.8 GHz

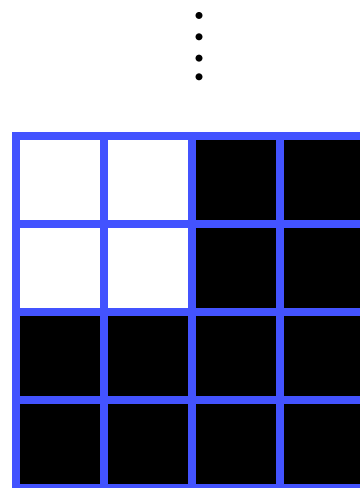


65 nm



2x4 cores @ 1.8 GHz
(8 cores dark, 8 dim)

*(Industry's Choice,
next slide)*



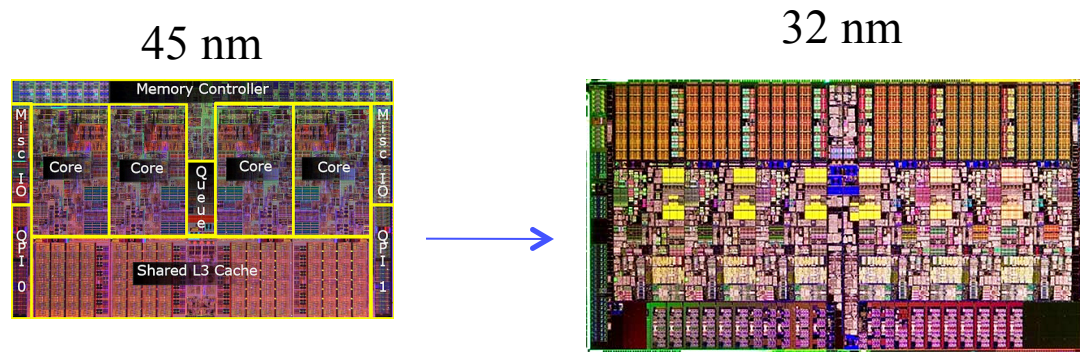
4 cores @ 2x1.8 GHz
(12 cores dark)

32 nm

[Hotchips 2010]

Utilization Wall in “Real Life”

45 nm $\rightarrow$ 32 nm ($S=1.4x$)



Nehalem
3.2 GHz
4 cores
120 W

Gulftown
3.3 GHz
6 cores
120 W

$S=1.4x$
 $\Delta \text{cores} = 1.5x$
 $\Delta \text{freq} = 1.03x$

As predicted by the utilization wall, compute capability scaled as S and not as S^3

Improvements in performance are gated by improvements in energy efficiency, which primarily result from decreases in capacitance, not improvements in frequency or area.