High Performance Computing

1. The Multicore revolution

Andrea Marongiu
(andrea.marongiu@unimore.it)
AA 2018-2019

Credits

- Michael B. Taylor, Is Dark Silicon Useful?, DAC 2012
- Moreno Marzolla, High Performance Computing, Università di Bologna, 2018
- Luca Benini, Metodologie di Progettazione HW/SW, Università di Bologna, 2018
- Marwedel, Embedded System Design, Springer 2018,
- Wolf, Computers as Components 4th Ed. Morgan Kaufmann 2016
- Wolf, High-Performance Embedded Computing 2nd Ed. Morgan Kaufmann 2014
- Lee, Seshia: Introduction to Embedded Systems, A Cyber-Physical Systems Approach, 2nd Ed., MIT Press, 2017

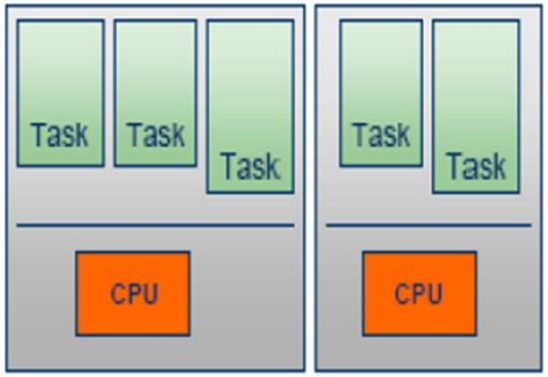
How are HP computers architected?

- Independent of the target compute domain, the architectural paradigms are similar
  - Convergence between traditional HPC systems and high-performance embedded systems

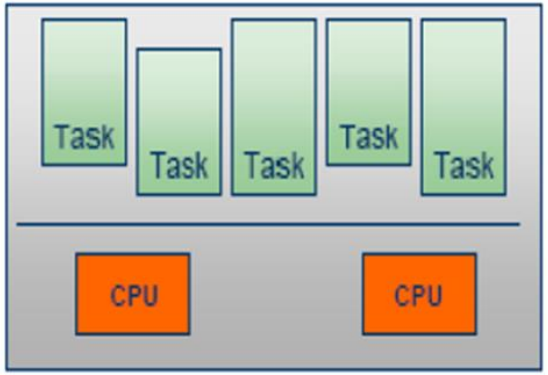
- We have already stressed a few keywords
  - *Architectural Heterogeneity*
  - *Massive Parallelism*

Vocabulary in the multi-era

- AMP (Asymmetric MP)
  - Each processor has local memory
  - Tasks statically allocated to one processor

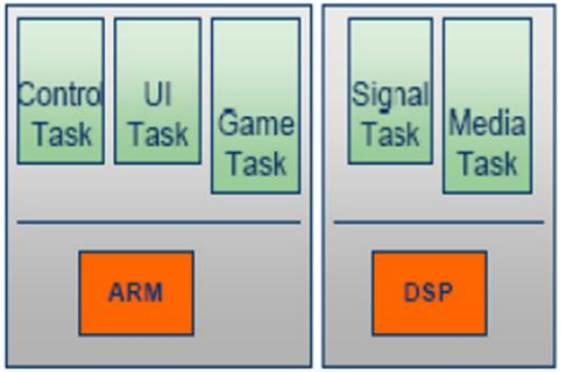


- SMP (Symmetric MP)
  - Processors share memory
  - Tasks dynamically scheduled to any processor

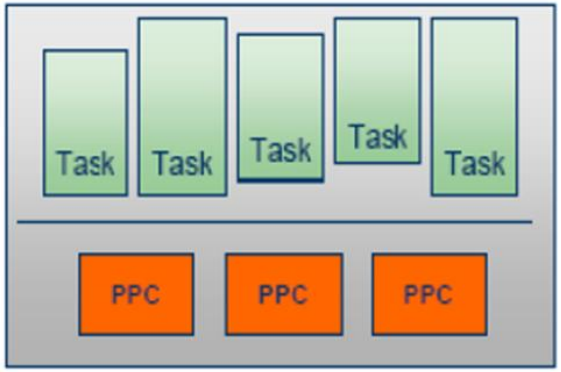


Vocabulary in the multi-era

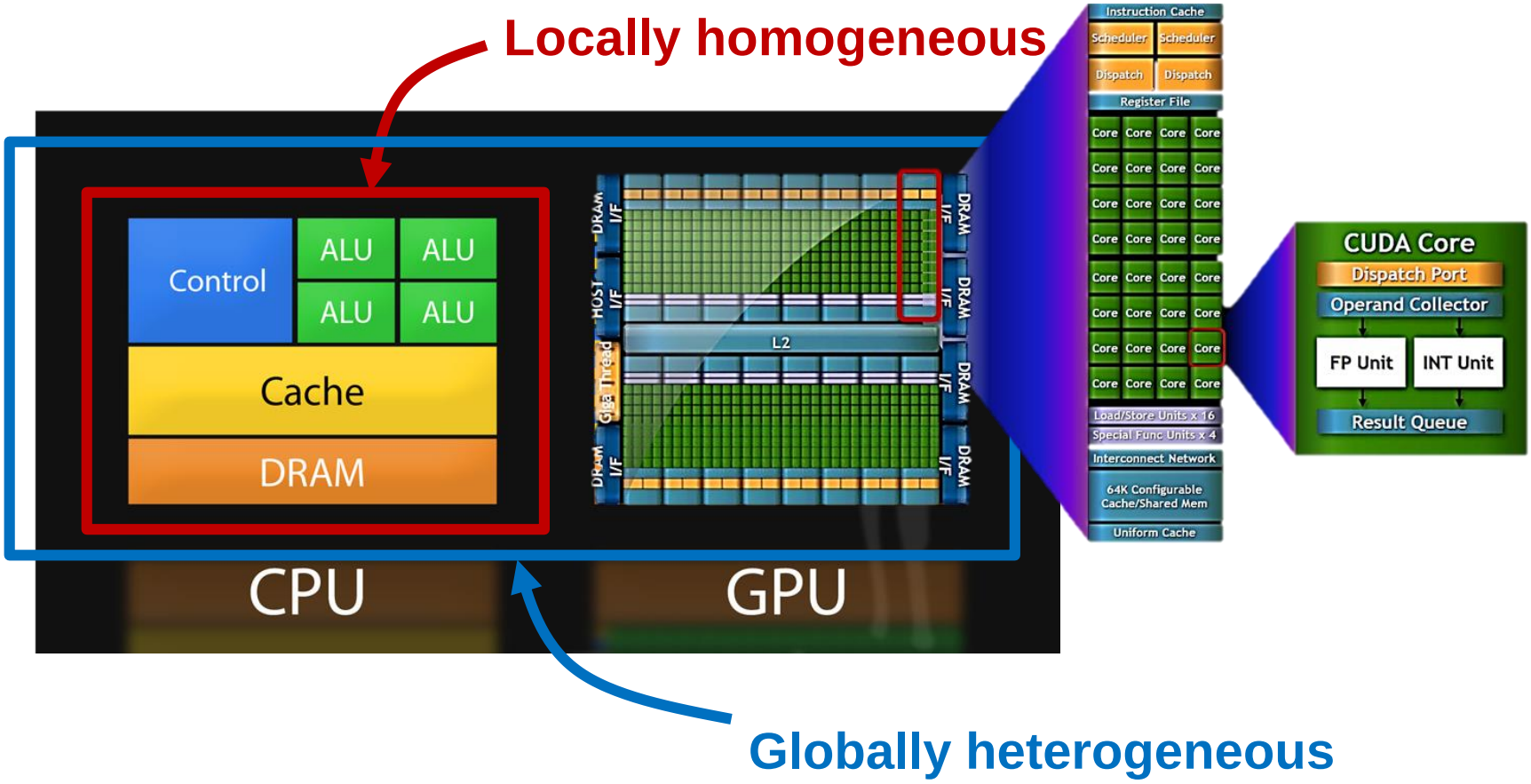
- Heterogeneous:
  - Specialization among processors
  - Often different instruction sets
  - Usually AMP design



- Homogeneous:
  - All processors have the same instruction set
  - Processors can run any task
  - Usually SMP design

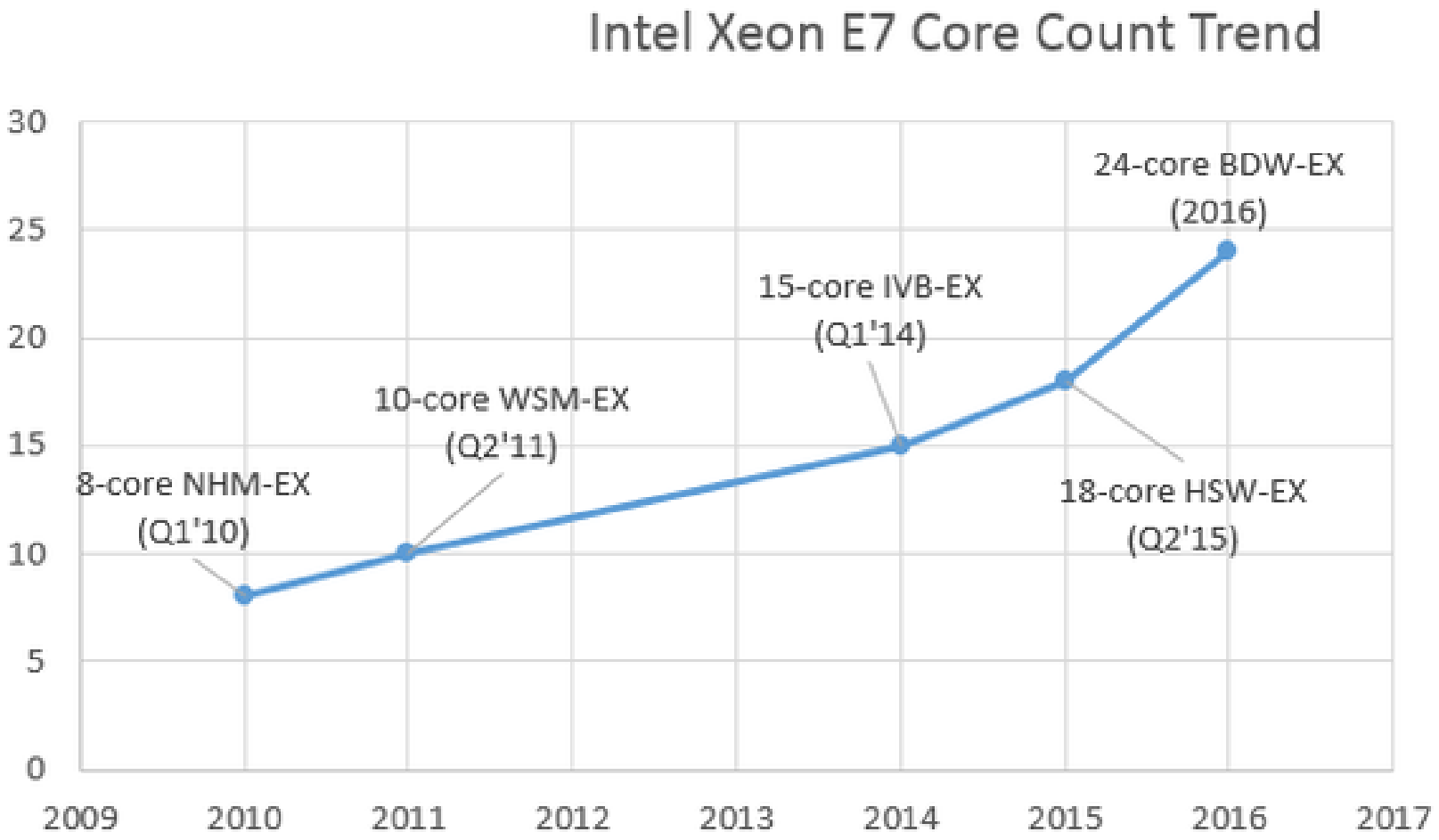


Modern HP systems



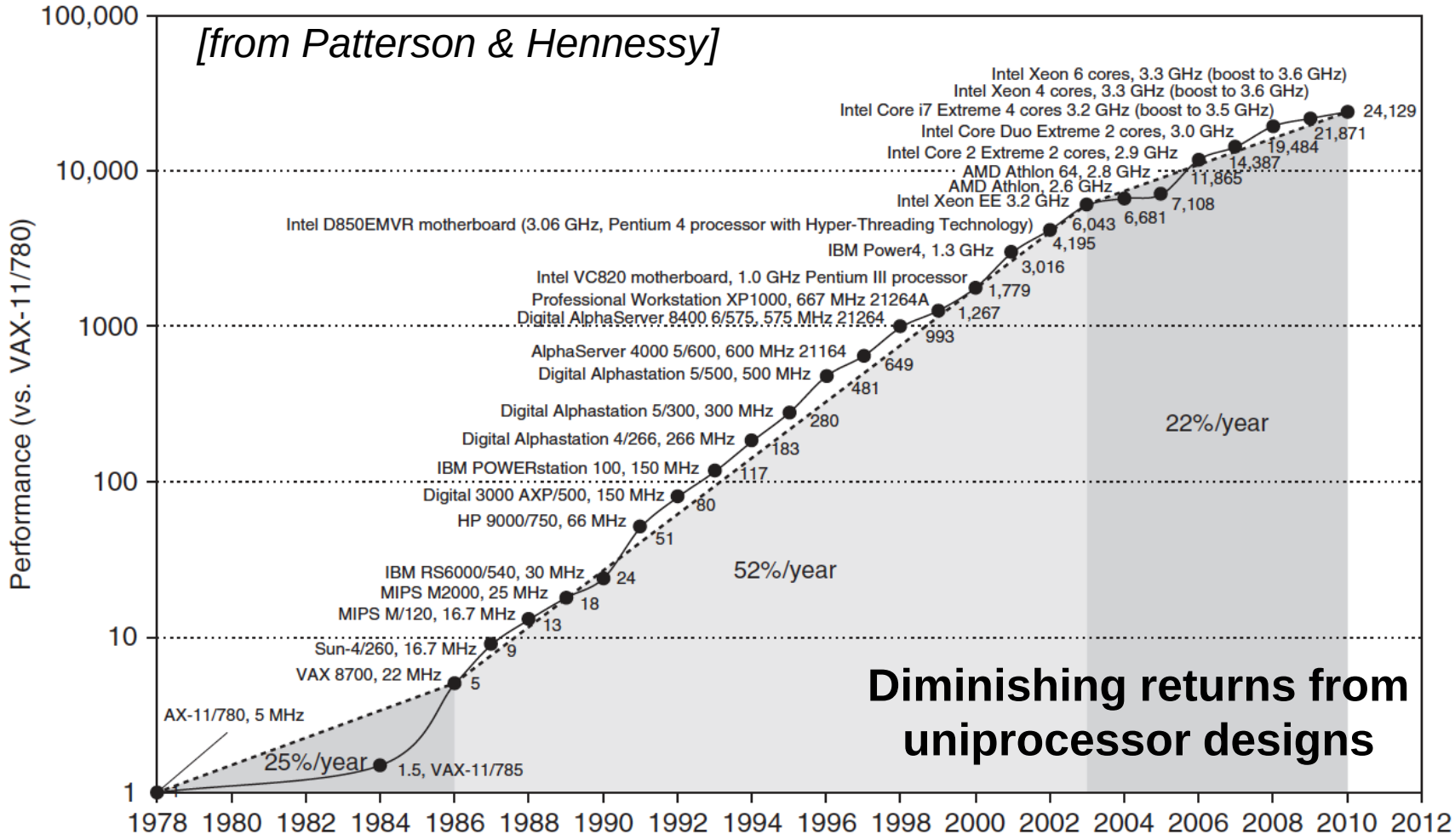
But how did we get there?

Multicore “Revolution”



Why Multicores?

The SPECint performance of the hottest chip grew by 52% per year from 1986 to 2002, and then grew only 20% in the next three years (about 6% per year).



Moore's Law

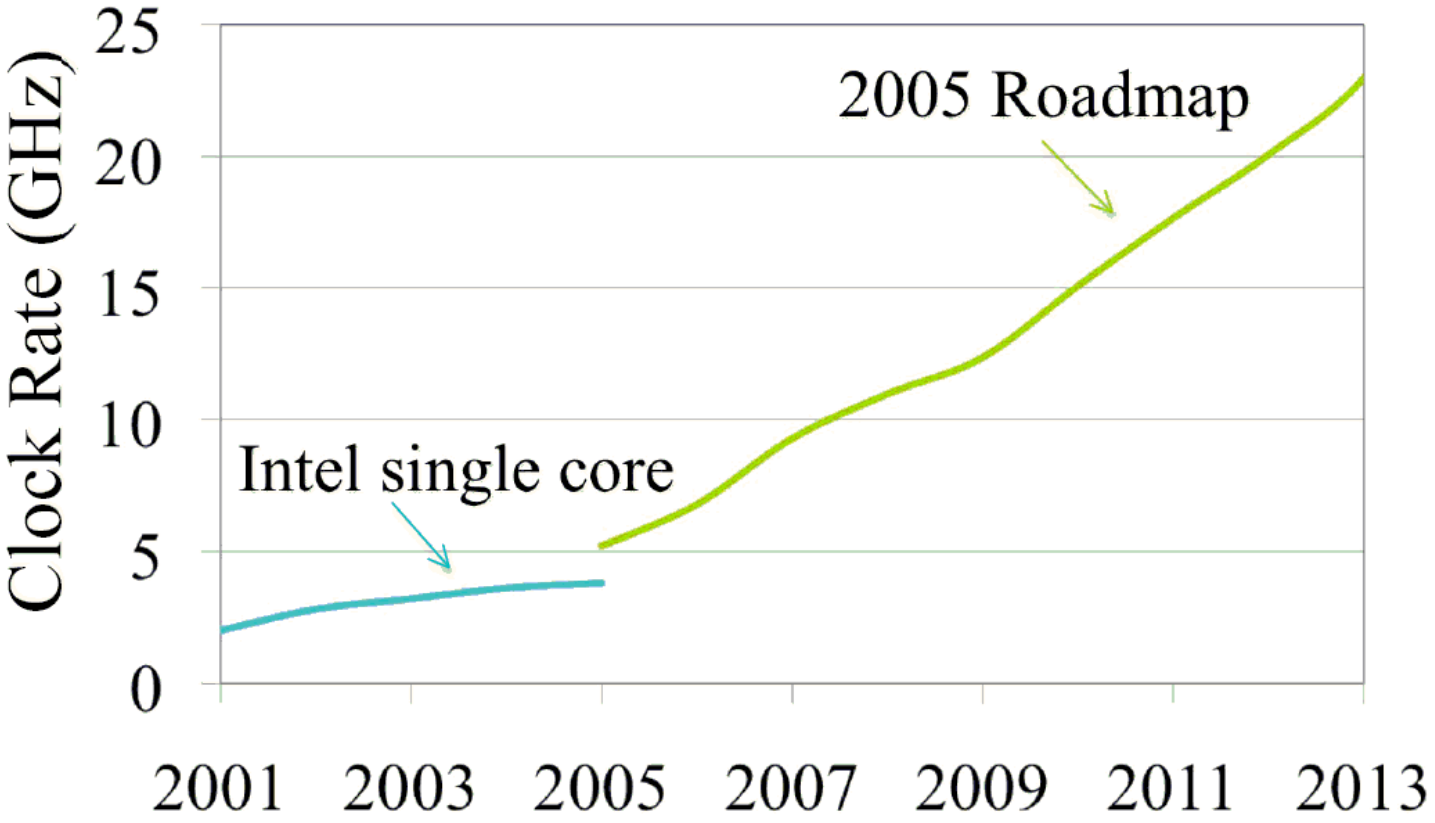
- *The number of transistors in an integrated circuit (IC) doubles every two years*
  - Until the early 2000's this meant an ever-increasing number of smaller transistors for each new processor generration

Moore, G.E., *Cramming more components onto integrated circuits*. Electronics, 38(8), April 1965

Gordon E. Moore (1929–)

Roadmap for CPU Clock Speed: Circa 2005

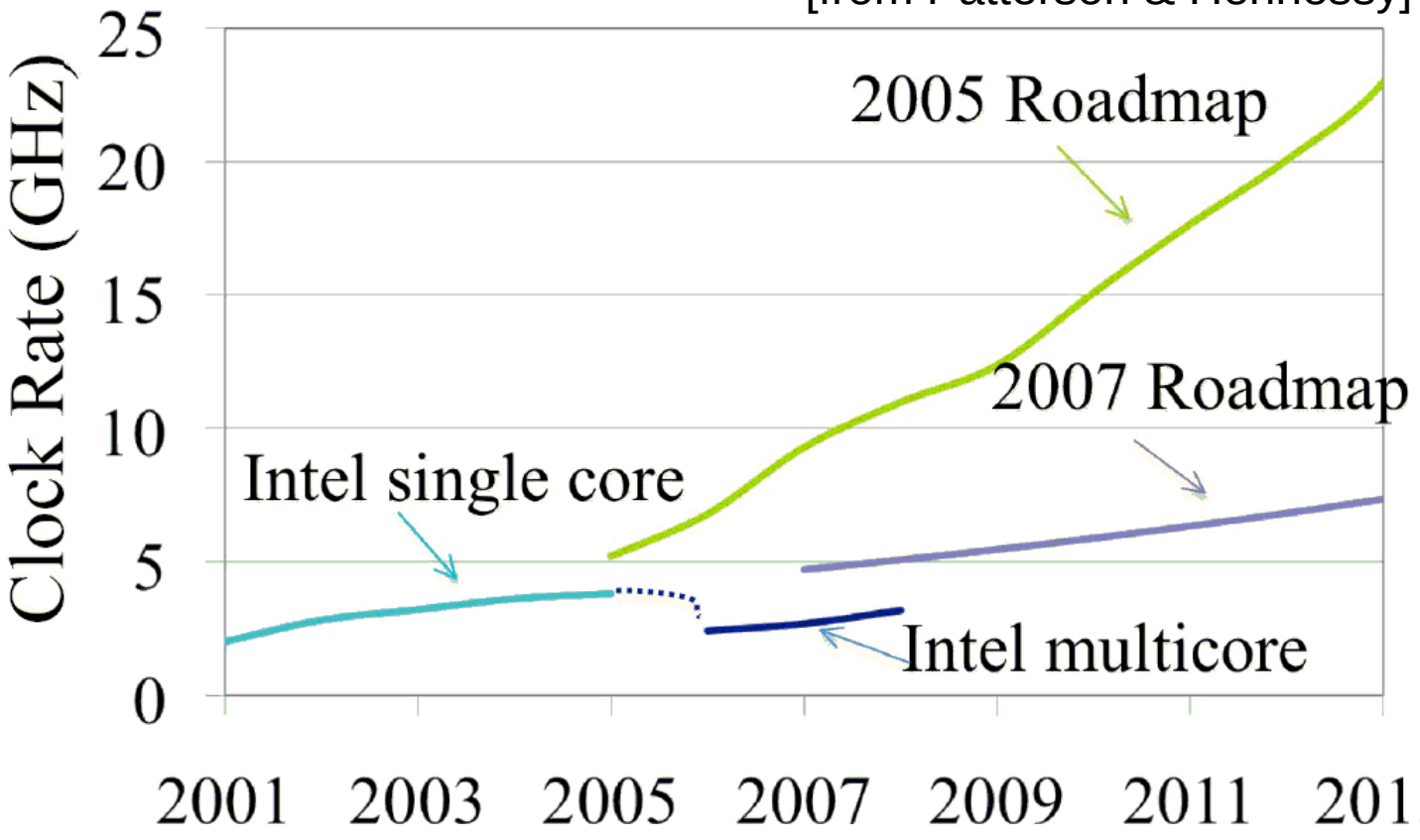
[from Patterson & Hennessy]



Here is the result of the best thought in 2005. By 2015, the clock speed of the top “hot chip” would be in the 20 – 25 GHz range.

The CPU Clock Speed Roadmap (A Few Revisions Later)

[from Patterson & Hennessy]



This reflects the practical experience gained with dense chips that were literally "hot"; they radiated considerable thermal power and were difficult to cool.
Law of Physics: All electrical power consumed is eventually radiated as heat.

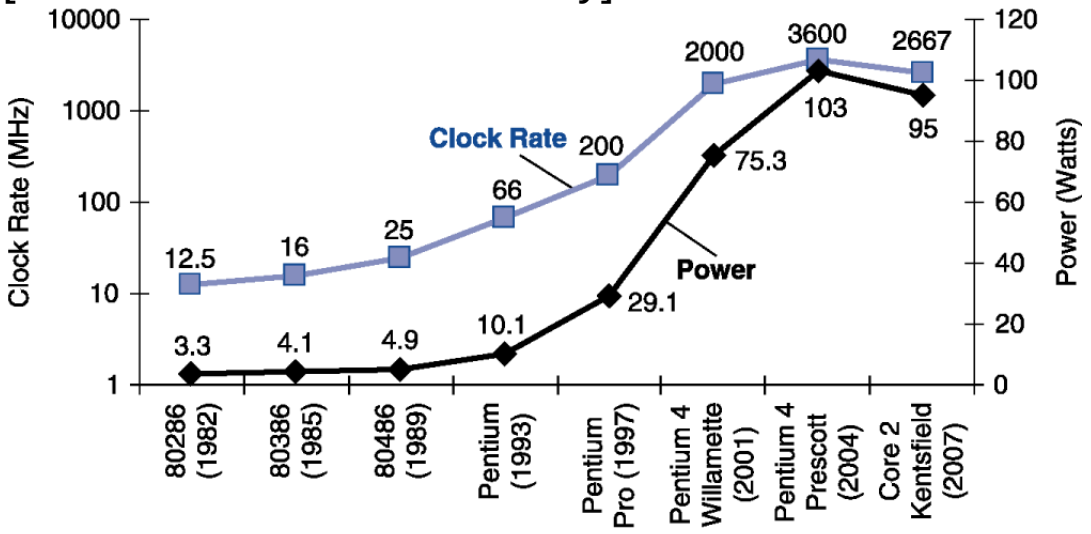
It's all about physics

- Smaller transistors → faster processor
- Faster processor → higher energy consumption
- Higher energy consumption → more heat
- More heat → unreliable processor

$$\text{Power} = \text{Capacitive load} \times \text{Voltage}^2 \times \text{Frequency}$$

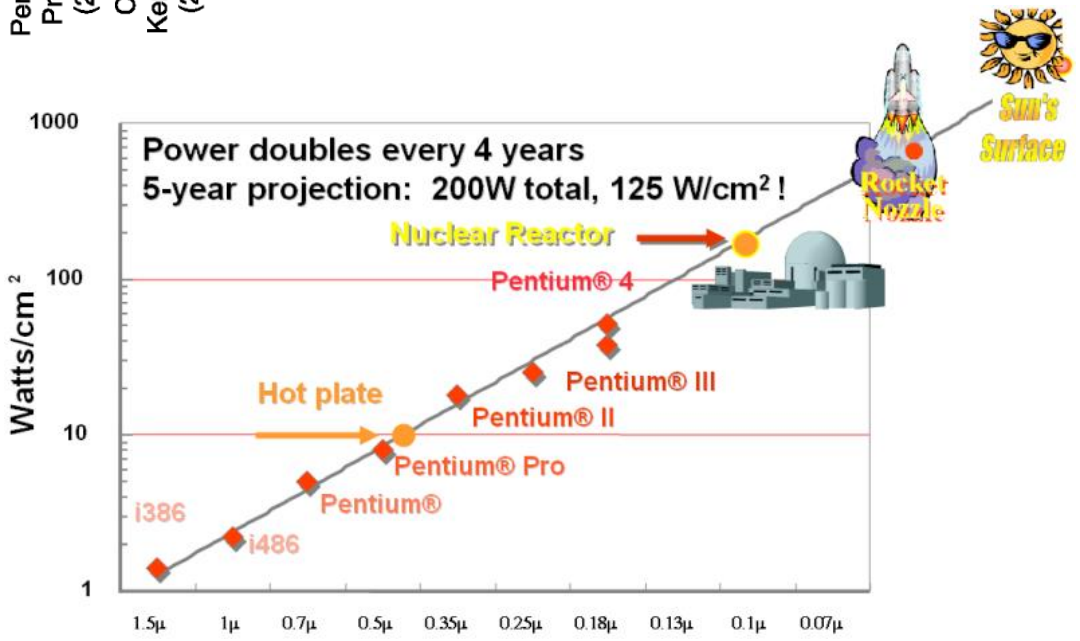
Power Wall

[from Patterson & Hennessy]



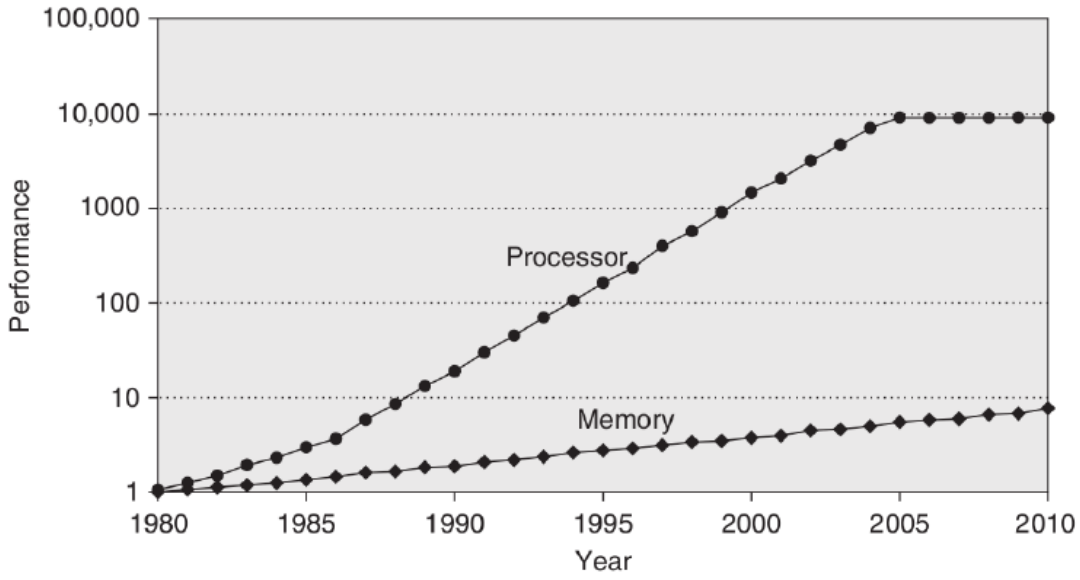
- The design goal for the late 1990's and early 2000's was to drive the clock rate up.
  - by adding more transistors to a smaller chip.

- Unfortunately, this increased the power dissipation of the CPU chip beyond the capacity of inexpensive cooling techniques



..and a few more "walls"

- The **memory wall**: access to data is a limiting factor.

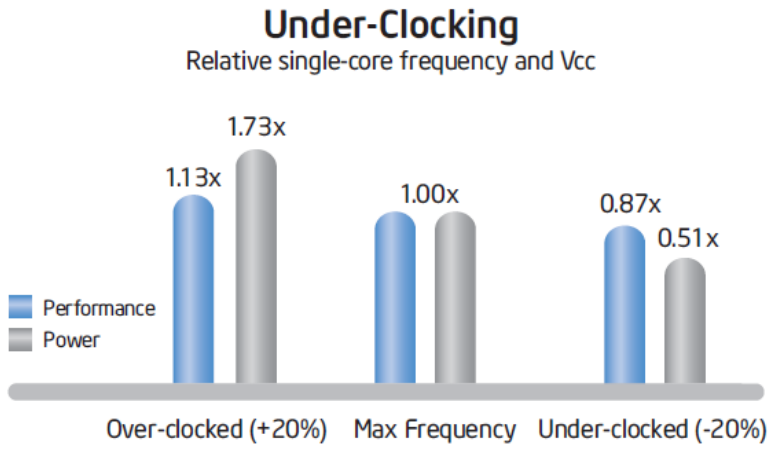


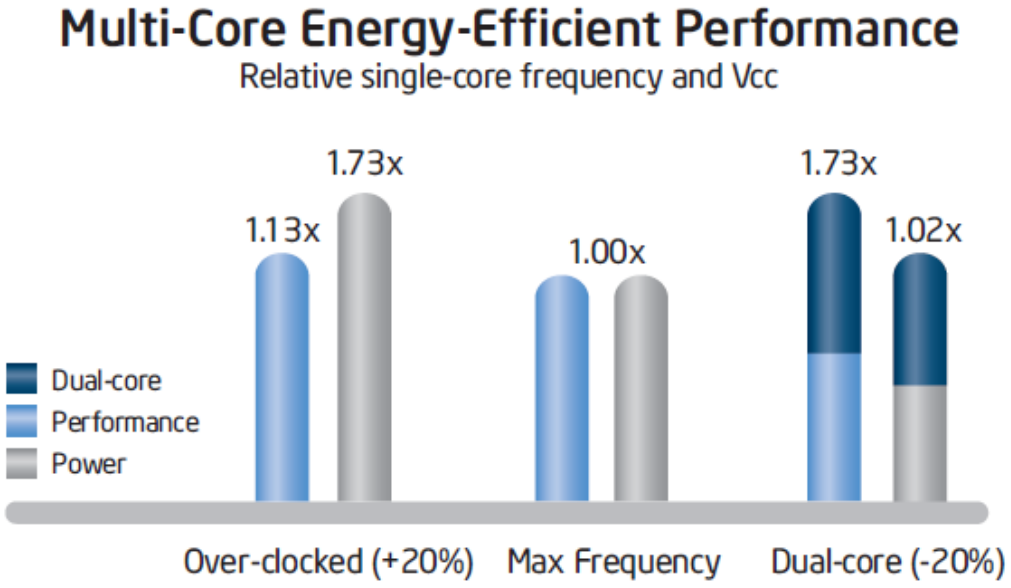
- The **ILP wall**: all the existing instruction-level parallelism (ILP) is already being used.
- **Conclusion**:
  - Explicit parallel mechanisms and explicit parallel programming are required for performance scaling

The Multicore Approach

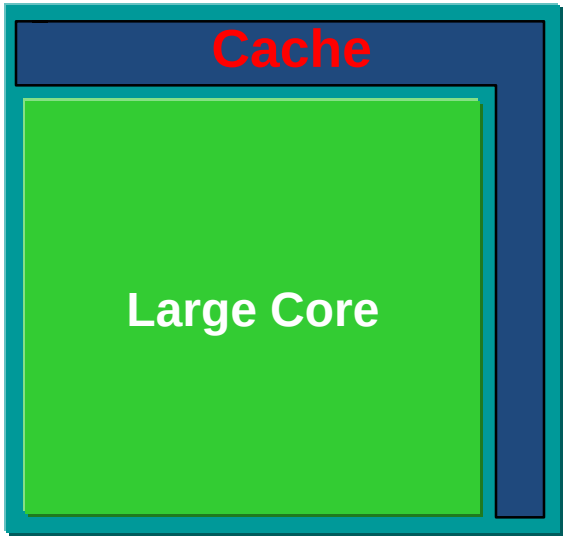
Multiple cores on the same chip

- Simpler
- Slower
- Less power demanding

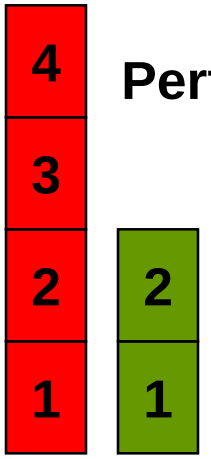




Multi-Core & Power



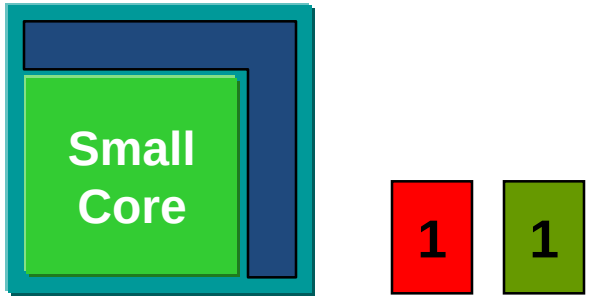
Power

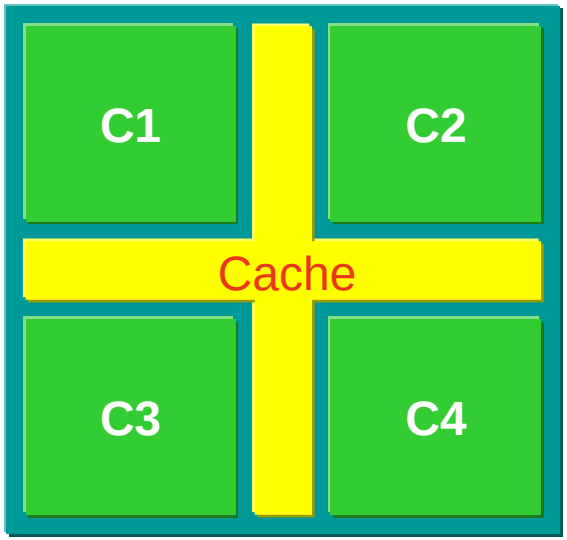


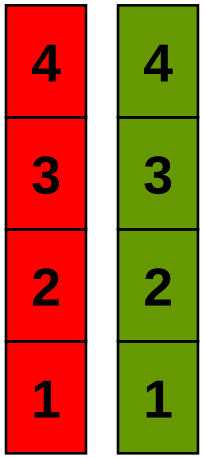
Performance

Power = 1/4

Performance = 1/2

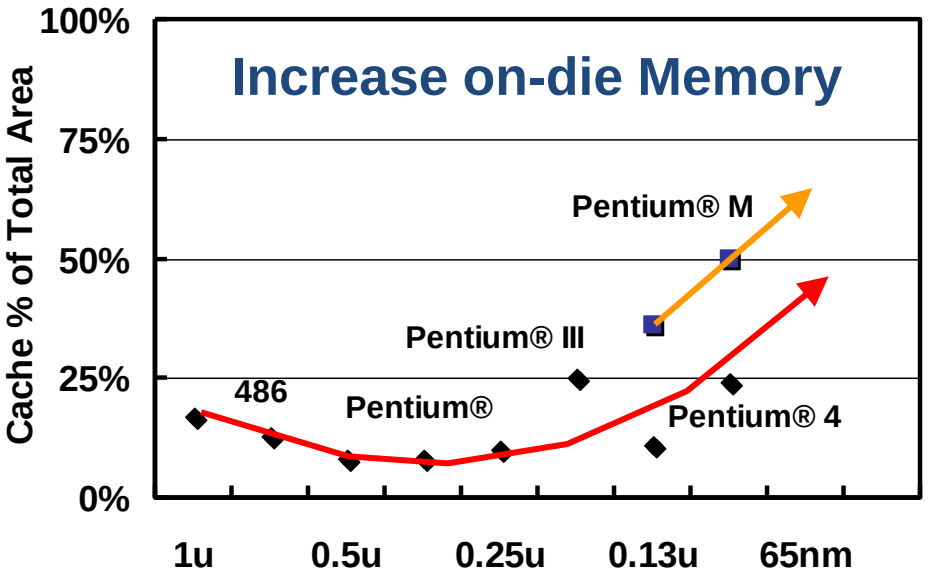




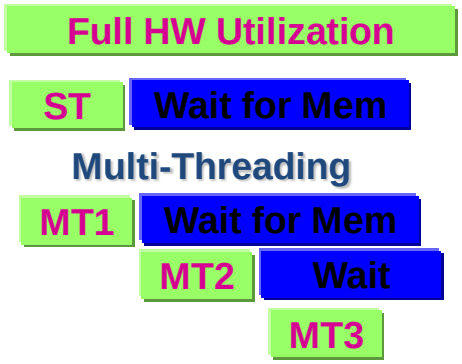


Multi-Core:
Power efficient
Better power and thermal management

µArchitecture Techniques

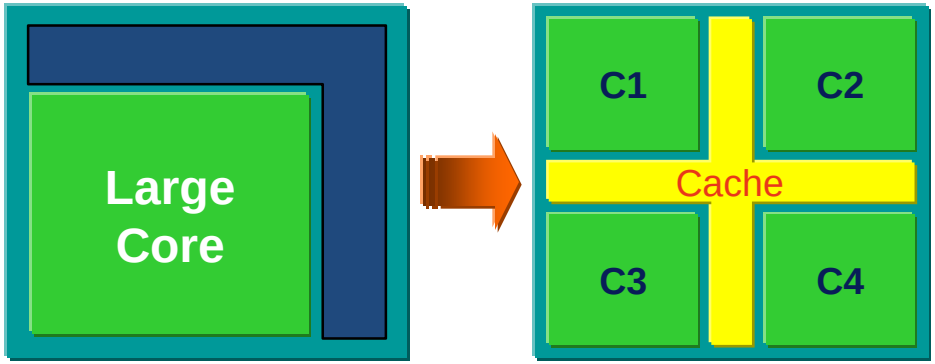


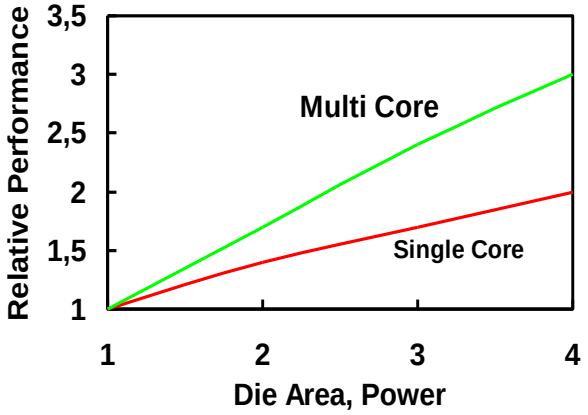
Multi-threading



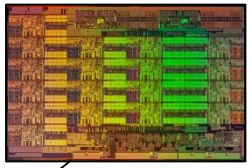
Improved performance, no impact on thermals & power delivery

Chip Multi-processing



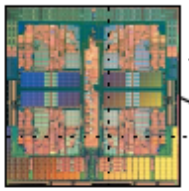


Transition to Multicore

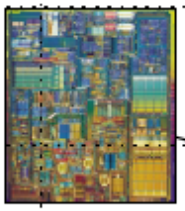


Intel Xeon (18cores)

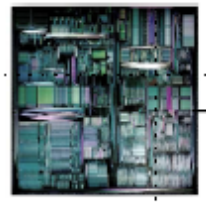
AMD Phenom (4 cores)



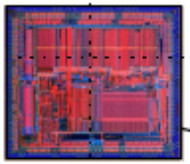
Intel Pentium 4



Intel Pentium Pro



MIPS R2000



Transistors (Thousands)

Parallel App Performance

Sequential App Performance

Frequency (MHz)

Typical Power (Watts)

Number of Cores

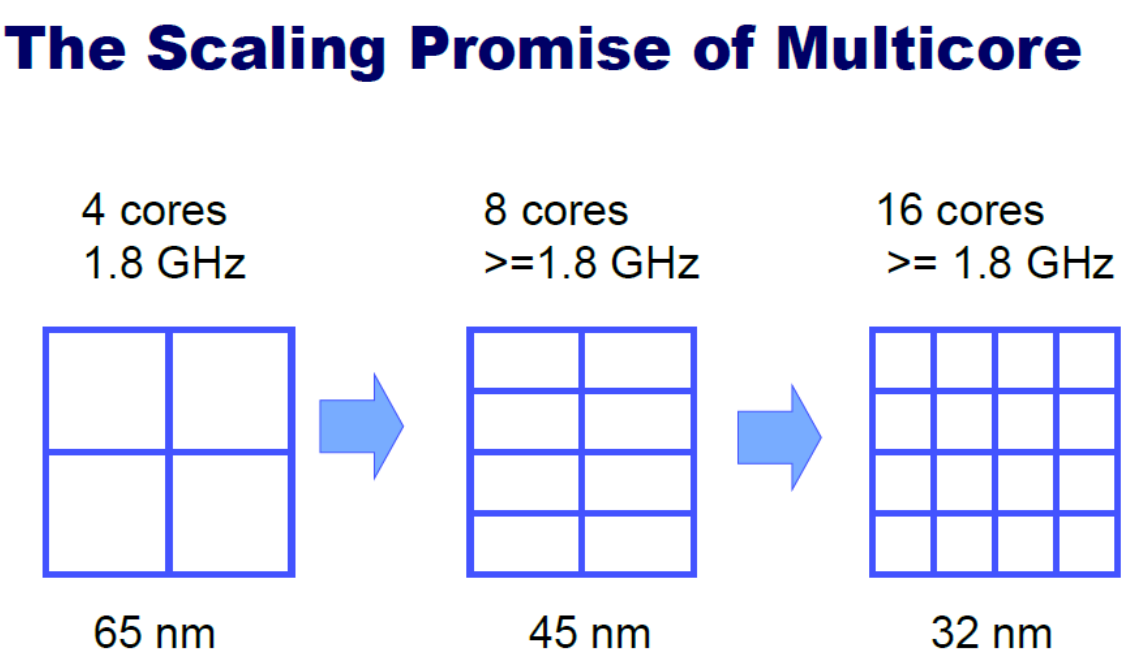
10^6 10^5 10^4 10^3 10^2 10^1 10^0

1975 1980 1985 1990 1995 2000 2005 2010 2015

Data partially collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond

And how about system heterogeneity?

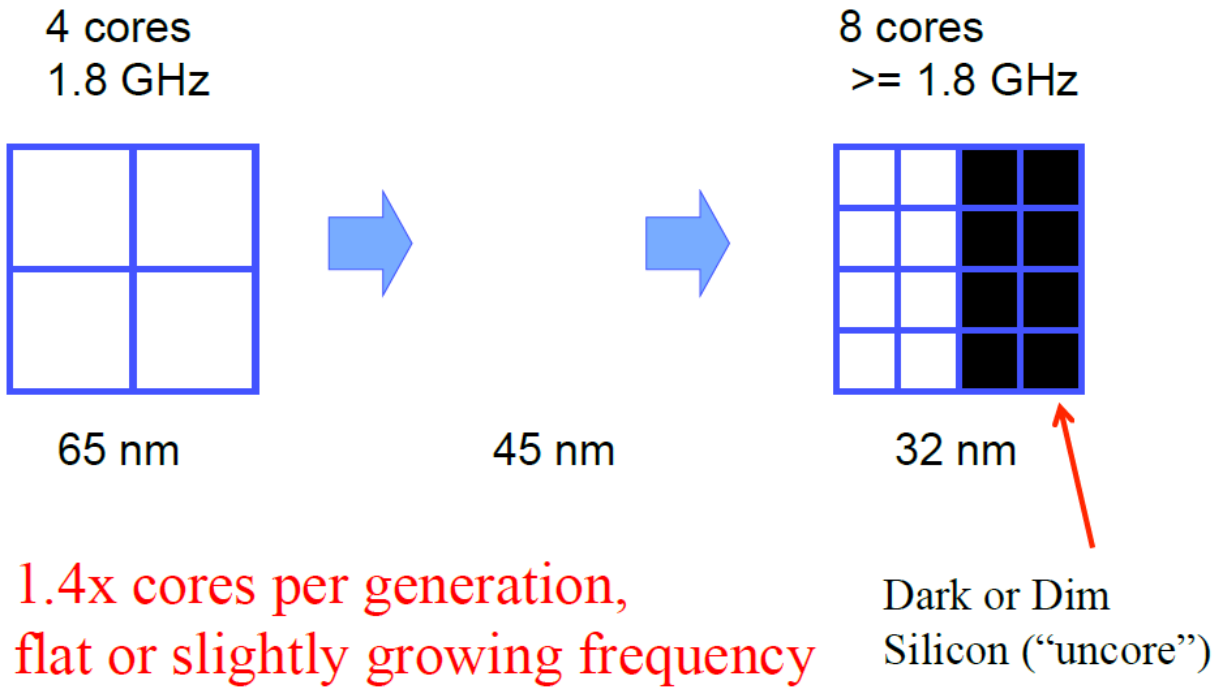
Dark silicon



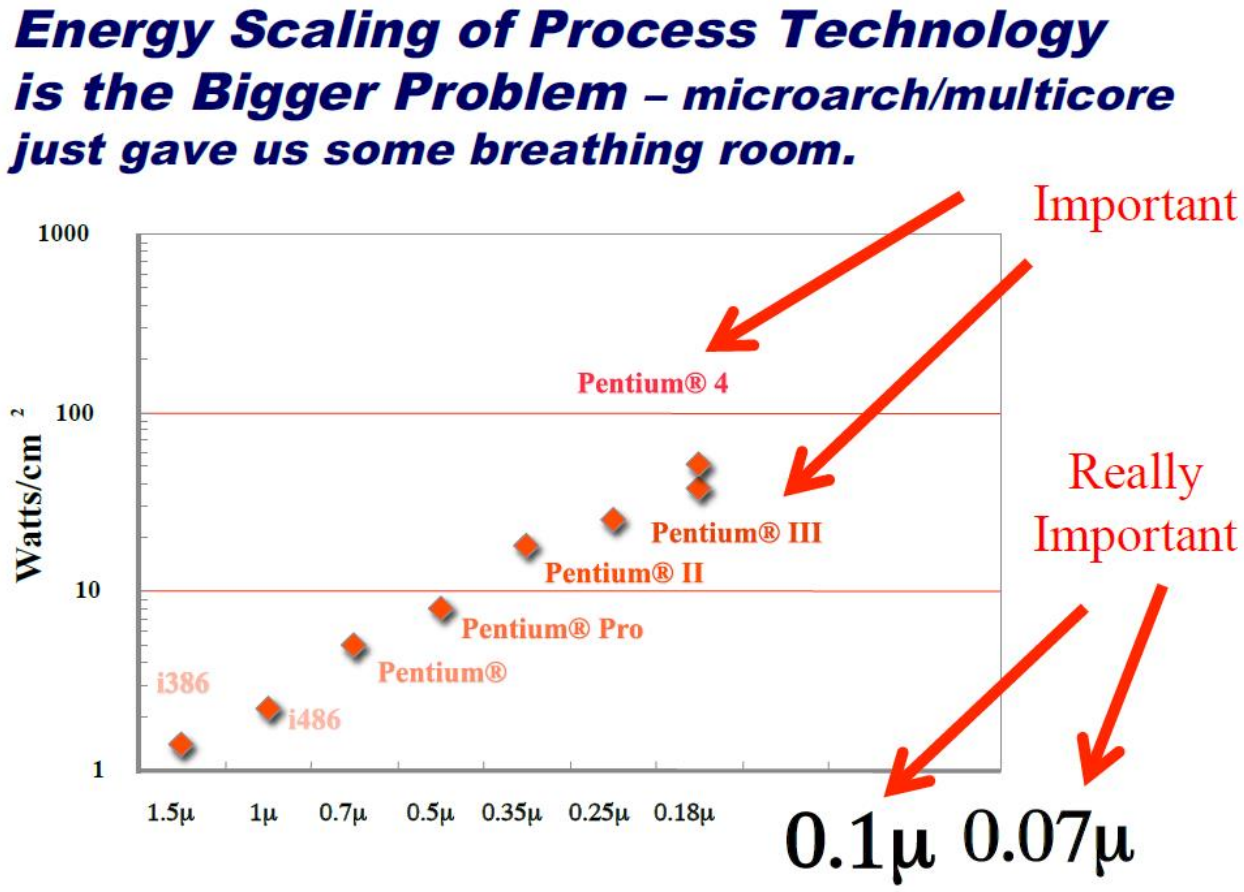
2x cores per generation,
flat or slightly growing frequency

Dark silicon

But actually,
that's not what's happening



Dark silicon



Dark silicon

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

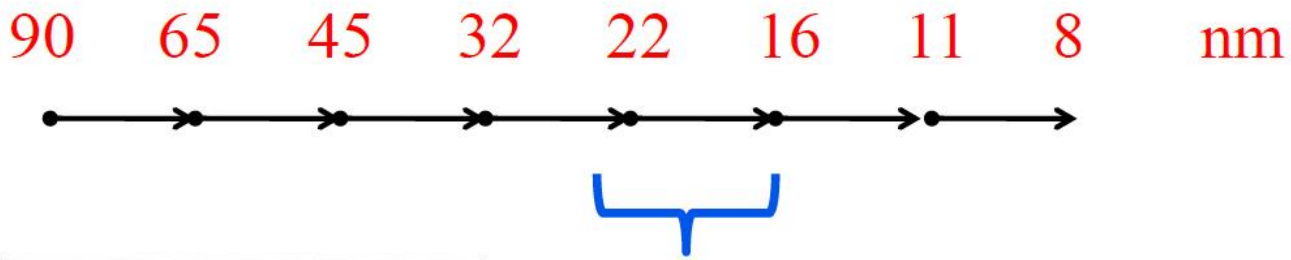
The Utilization Wall:

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

[Venkatesh, ASPLOS '10]

Dark silicon

Scaling 101: Moore's Law



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

Dark silicon

Scaling 101:

Transistors scale as S^2

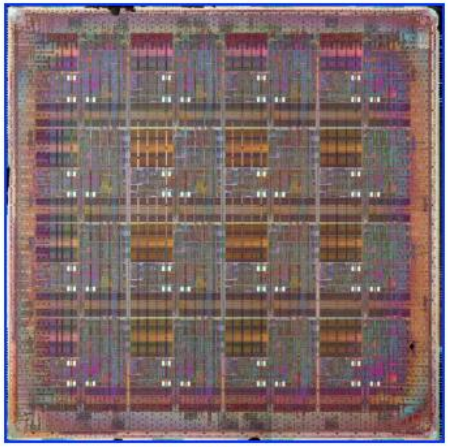
180 nm

16 cores

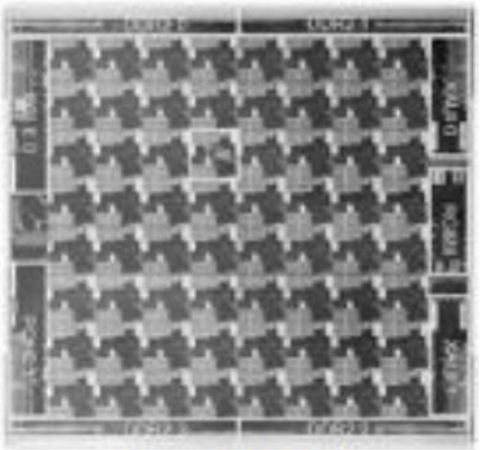
S = 2x
Transistors = 4x

90 nm

64 cores



MIT Raw



Tilera TILE64

Dark silicon

Advanced Scaling:

Dennard: "Computing Capabilities Scale by $S^3 = 2.8x$"

If S=1.4x …

S^3

S^2

S

1

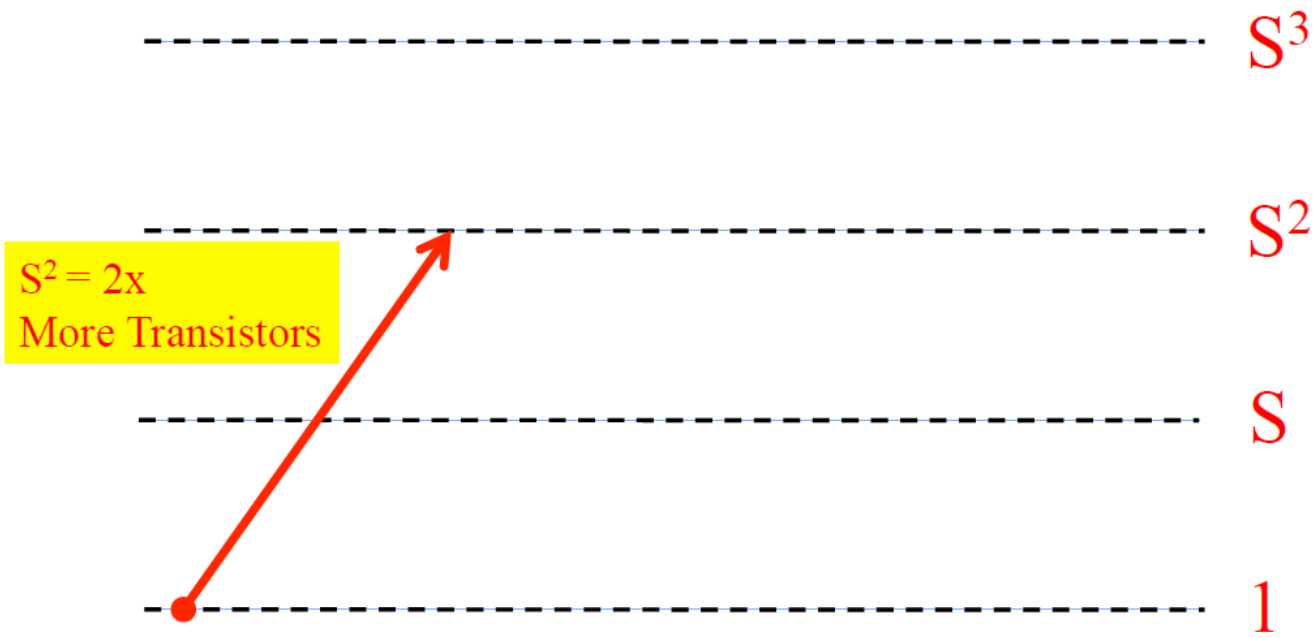
Design of Ion-Implanted MOSFETs with Very Small Dimensions
Dennard et al, 1974

Dark silicon

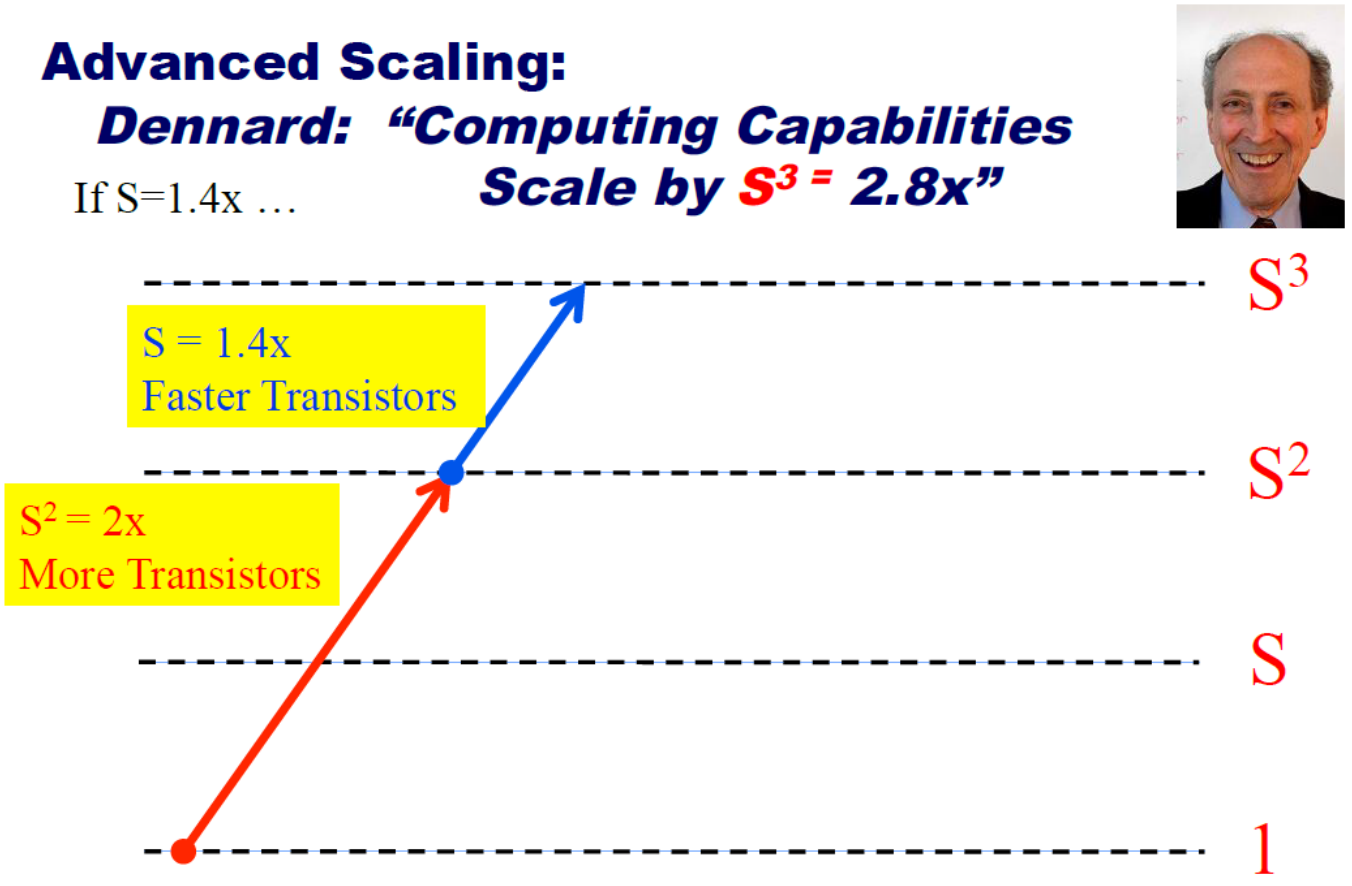
Advanced Scaling:

Dennard: "Computing Capabilities Scale by $S^3 = 2.8x$"

If S=1.4x …



Dark silicon

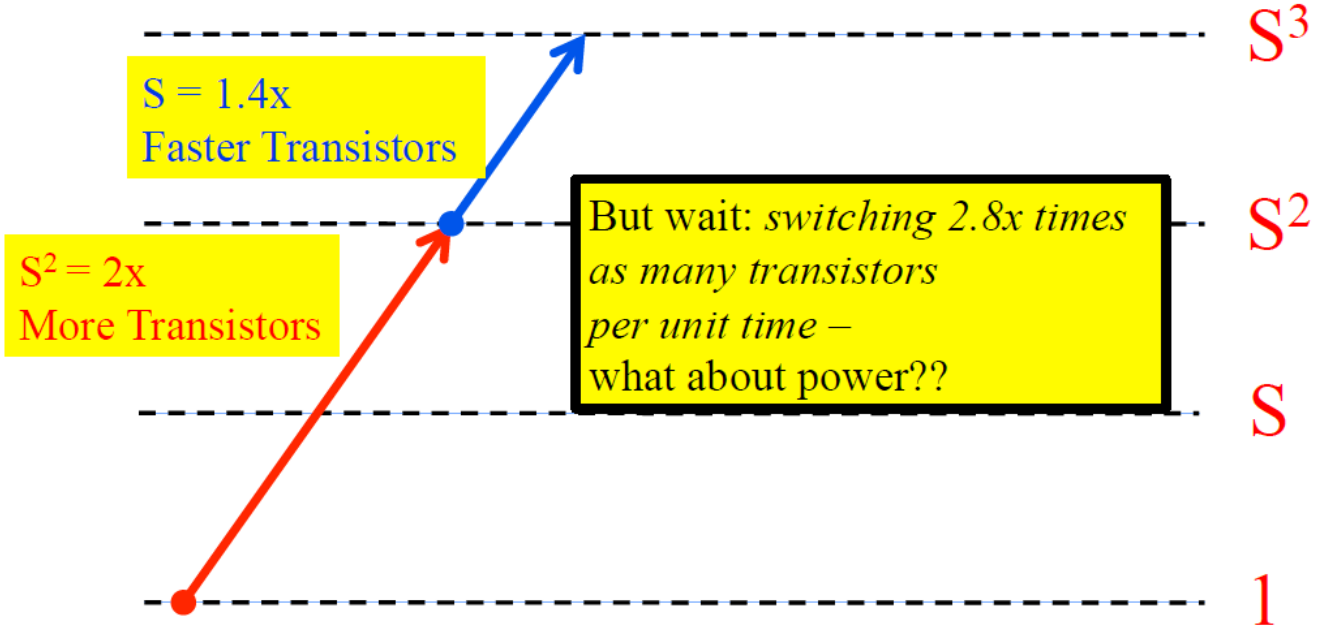


Dark silicon

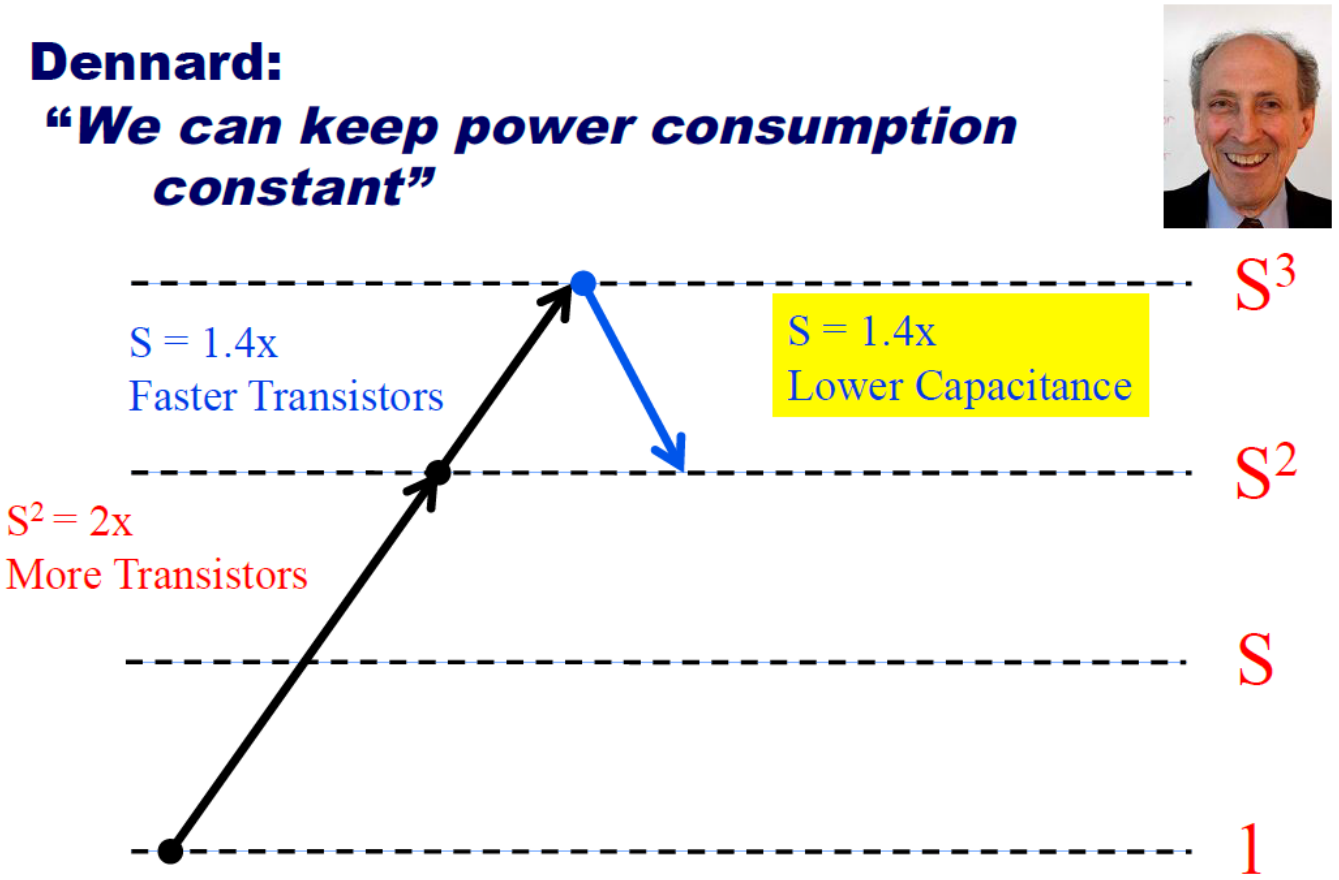
Advanced Scaling:

Dennard: "Computing Capabilities Scale by $S^3 = 2.8x$"

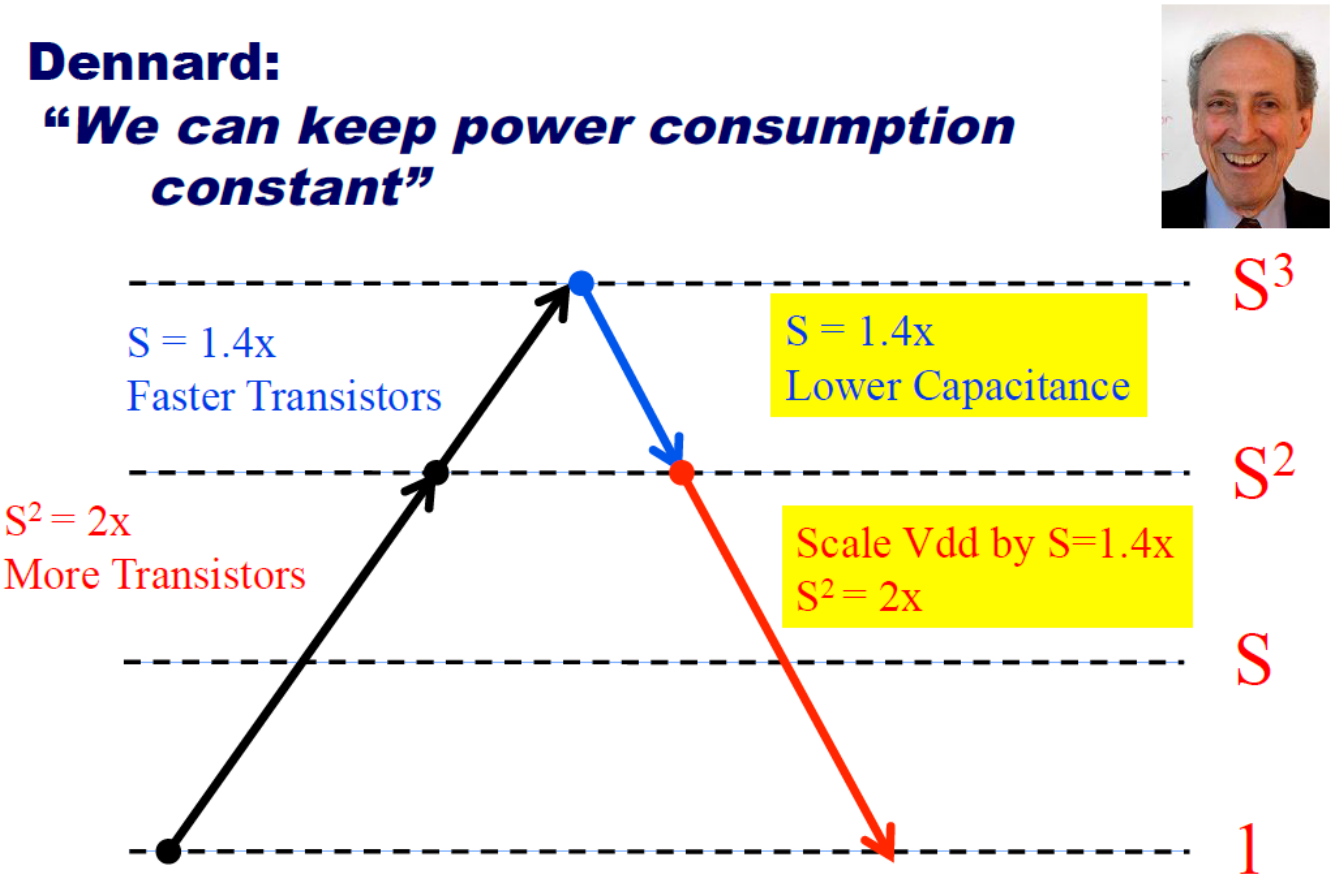
If S=1.4x …



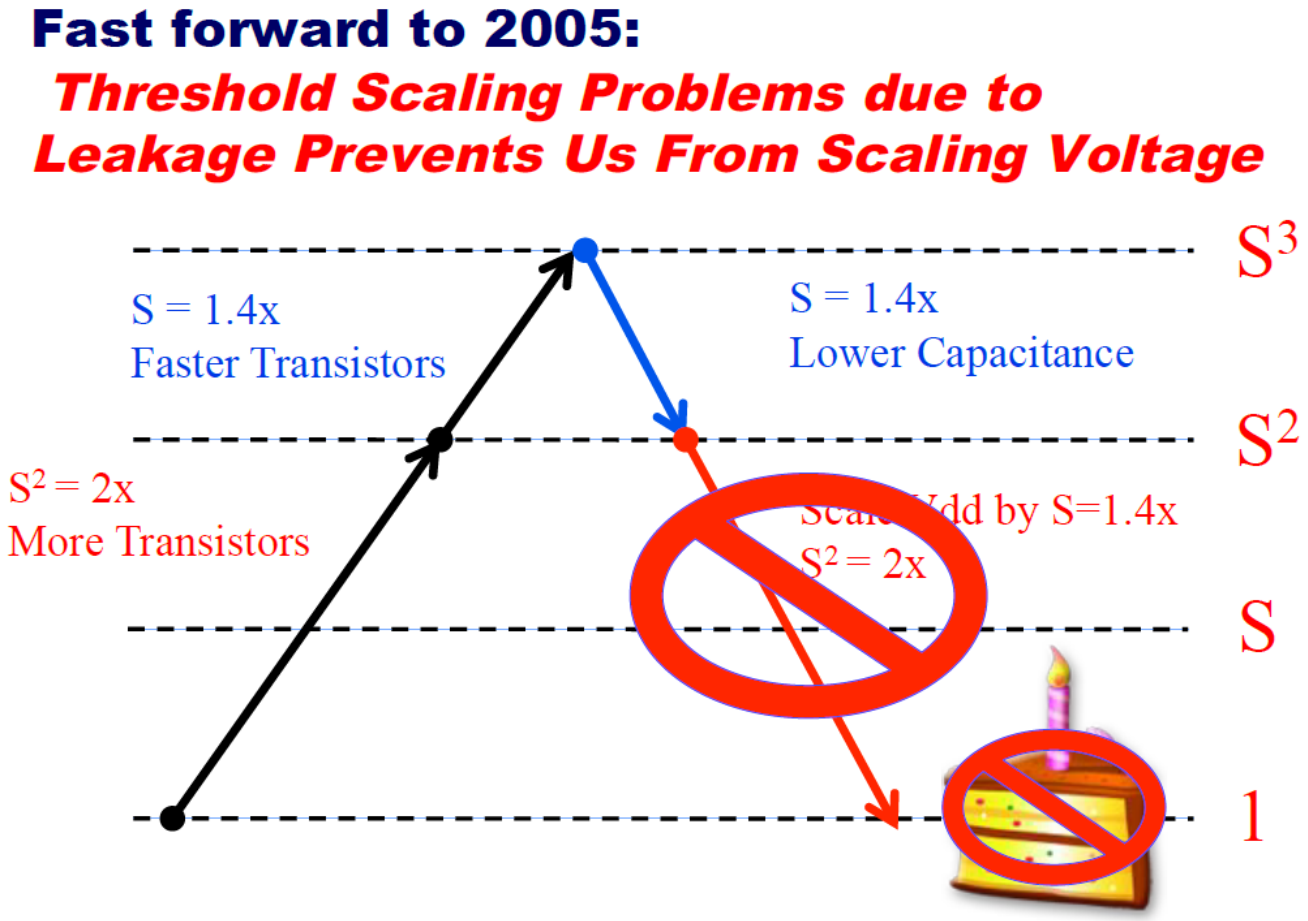
Dark silicon



Dark silicon

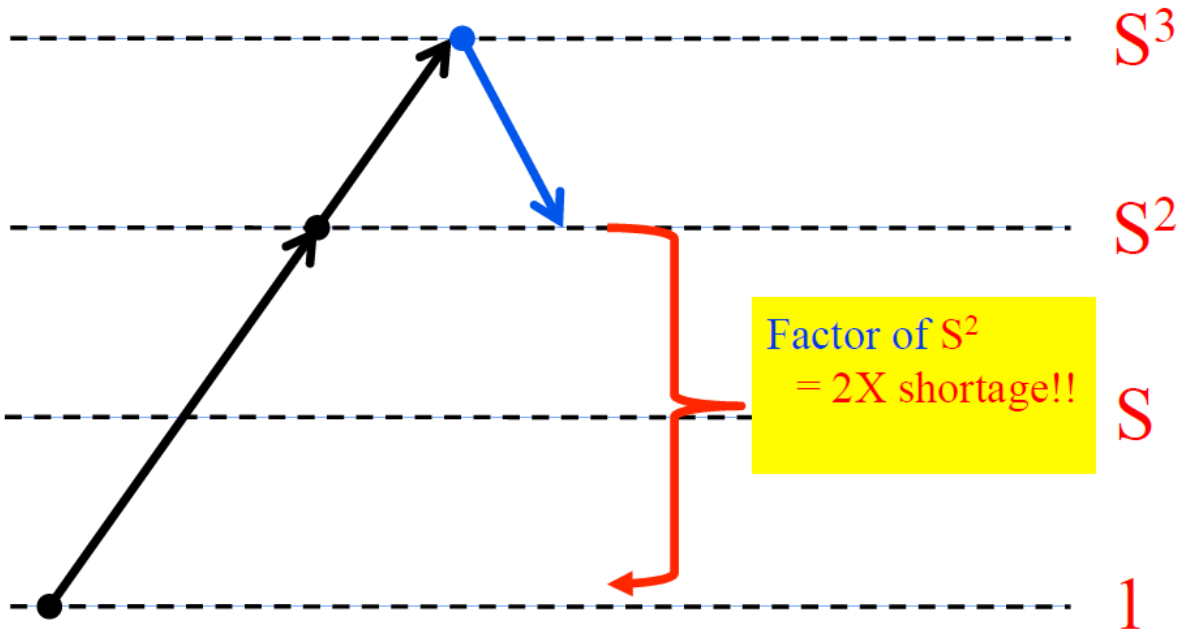


Dark silicon



Dark silicon

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

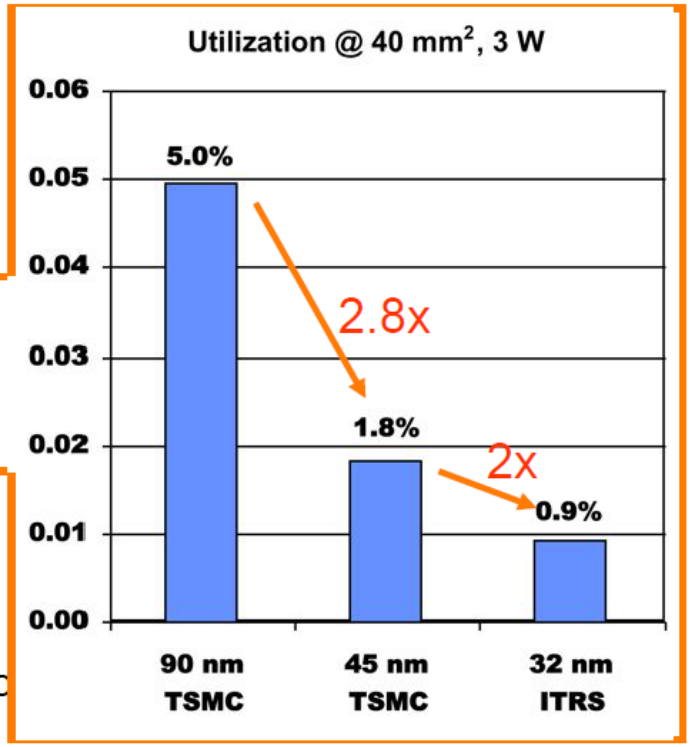


Dark silicon

We've Hit The Utilization Wall

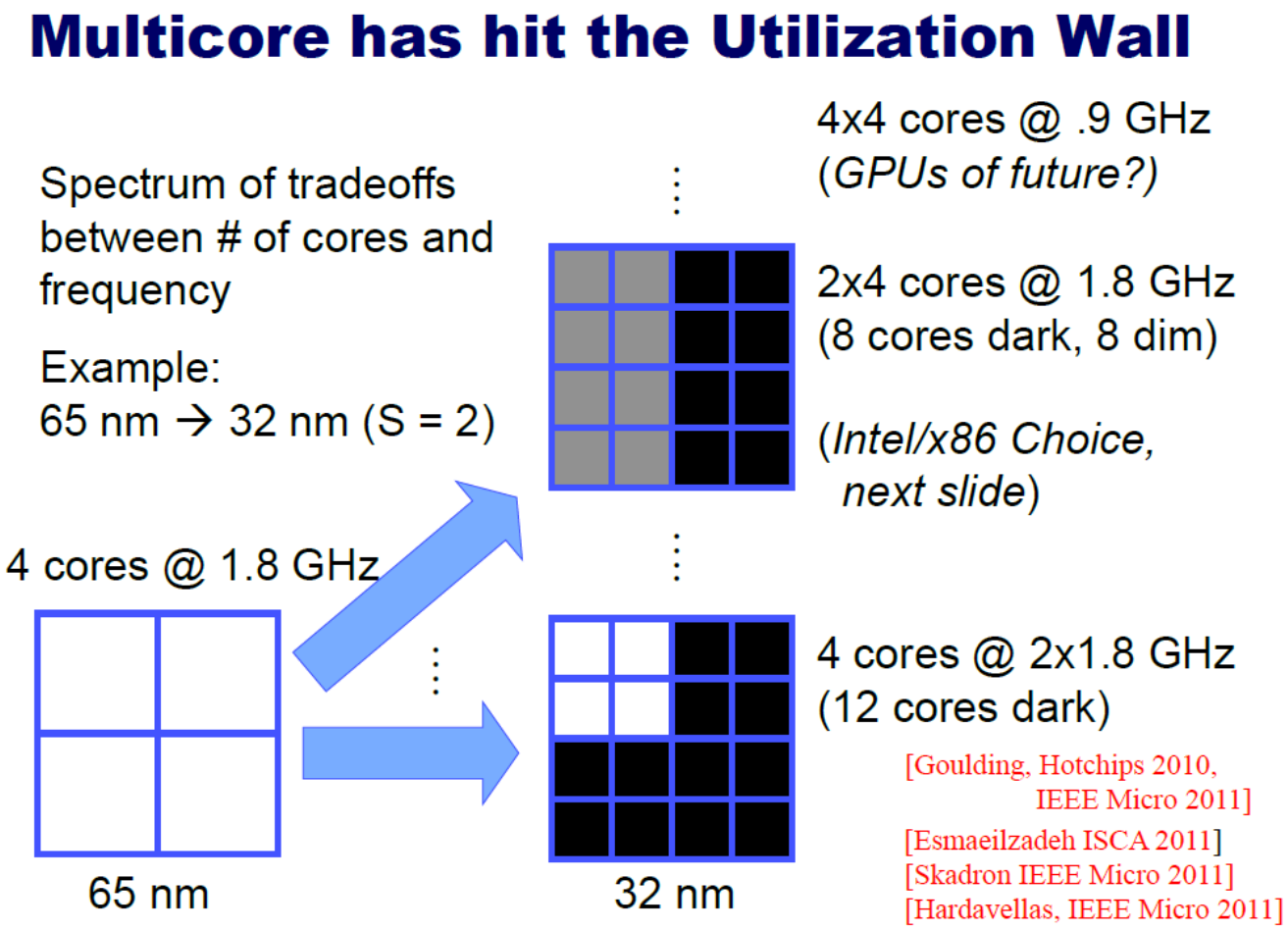
Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- 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



[Venkatesh, ASPLOS '10]

Dark silicon



Dark silicon

The Four Horsemen

What do we do with this dark silicon?

Four top contenders, each of which seemed like an unlikely candidate from the beginning, carrying unwelcome burdens in design, manufacturing and programming. None is ideal, but each has its benefit and the optimal solution probably incorporates all four of them…

I II III IV

Dark silicon

The Shrinking Horseman (#1)

"Area is expensive. Chip designers will just build smaller chips instead of having dark silicon in their designs!"

First, dark silicon doesn't mean *useless silicon,*
it just means it's under-clocked or not used all of the time.

There's lots of dark silicon in current chips:

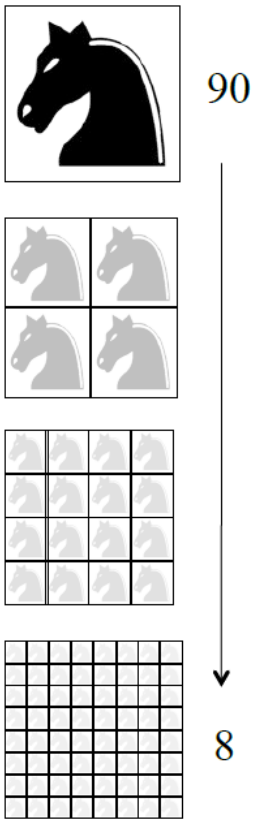
- On-chip GPU on AMD Fusion or Intel Sandybridge for GCC
- L3 cache is very dark for applications with small working sets
- SSE units for integer apps
- Many of the resources in FPGAs not used
  by many designs (DSP blocks, PCI-E, Gig-E etc)

Dark silicon

The Dim Horseman (#2)

"We will fill the chip with homogeneous cores that would exceed the power budget but we will underclock them (spatial dimming), or use them all only in bursts (temporal dimming)

… "dim silicon".



Spatial Dimming

- Gen1&2 multicores (higher core counts → lower freqs)
- Near-threshold voltage operation

Temporal Dimming

Thermally limited systems

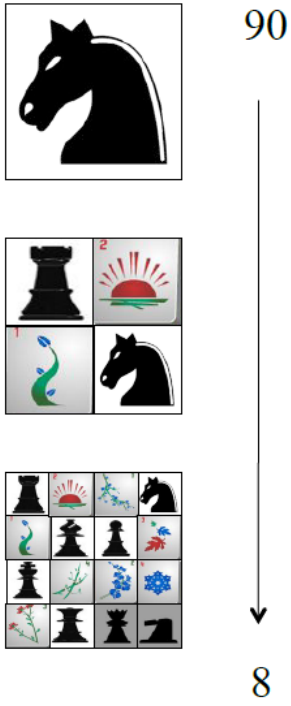
- **ARM Big-little** (A15 power usage way above sustainable for phone → 10sec burst at most)
- Battery-limited systems
- Quad-core mobile application processor

Dark silicon

The Specialized Horseman (#3)

"We will use all of that dark silicon area to build specialized cores, each of them tuned for the task at hand (10-100x more energy efficient), and only turn on the ones we need…"

[e.g., Venkatesh et al., ASPLOS 2010,
Lyons et al., CAL 2010,
Goulding et al., Hotchips 2010,
Hardavellas et al. IEEE Micro 2011]



Specialization is the goal behind architectural heterogeneity

Dark silicon

The Deus Ex Machina Horseman

"MOSFETs are the fundamental problem. "

We can switch to FinFets, Trigate, High-K, nanotubes, 3D, for one-time improvements, but none are sustainable solutions across process generations.

Device physics ("thermionic emission of carriers across a potential well") limit MOSFETS to 60 mV/decade subthreshold slope, which means the leakage problem is always there.."

- **Nano-electrical Mechanical Relays**

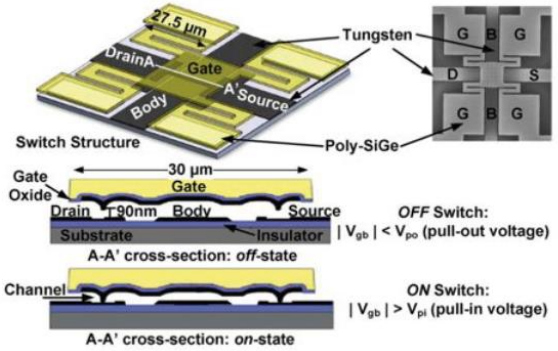


Fig. 1. SEM, diagram, and operating states of the MEM relay device.

[e.g, Spencer et al JSSC 2011]

- **Human Brain**
  - →100 trillion synapses @ 20 W!
  - →Very "dark" circuits