CS152: Computer Systems Architecture
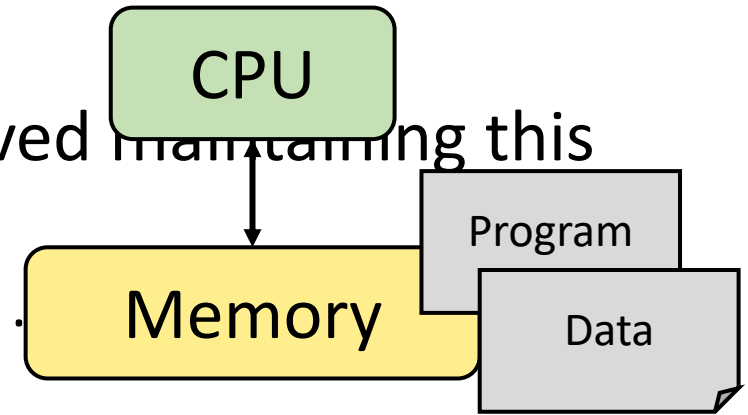# Moore's Law

Sang-Woo Jun

Winter 2021

UCI

Conventional performance scaling

❏ Traditional model of a computer is simple
  o Single, in-order flow of instructions on a processor
  o Simple, in-order memory model
❏ Large part of computer architecture research involved maintaining this abstraction while improving performance
  o Transparent caches, Transparent superscalar scheduling, ...
  o Same software runs faster tomorrow
  o (Slow software becomes acceptable tomorrow)
❏ Driven largely by continuing march of Moore's law

CPU

Memory

Program

Data

Moore's Law

❑ What exactly does it mean?

❑ What is it that is scaling?

Moore's Law

❑ Typically cast as:
  "Performance doubles every X months"

❑ Actually closer to:
  "Number of transistors per unit cost doubles every X months"

Moore's Law

The complexity for minimum component costs has increased at a rate of roughly a factor of two per year.

[…]

Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years.

-- Gordon Moore, Electronics, 1965

Why is Moore's Law conflated with processor performance?

Dennard Scaling:
From Moore's Law to performance

❑ "Power density stays constant as transistors get smaller"
   o Robert H. Dennard, 1974

❑ Intuitively:
   o Smaller transistors → shorter propagation delay → faster frequency
   o Smaller transistors → smaller capacitance → lower voltage

   o $Power \propto Capacitance \times Voltage^2 \times Frequency$

Moore's law → Faster performance @ Constant power!

Single-core performance scaling projection

What happened?
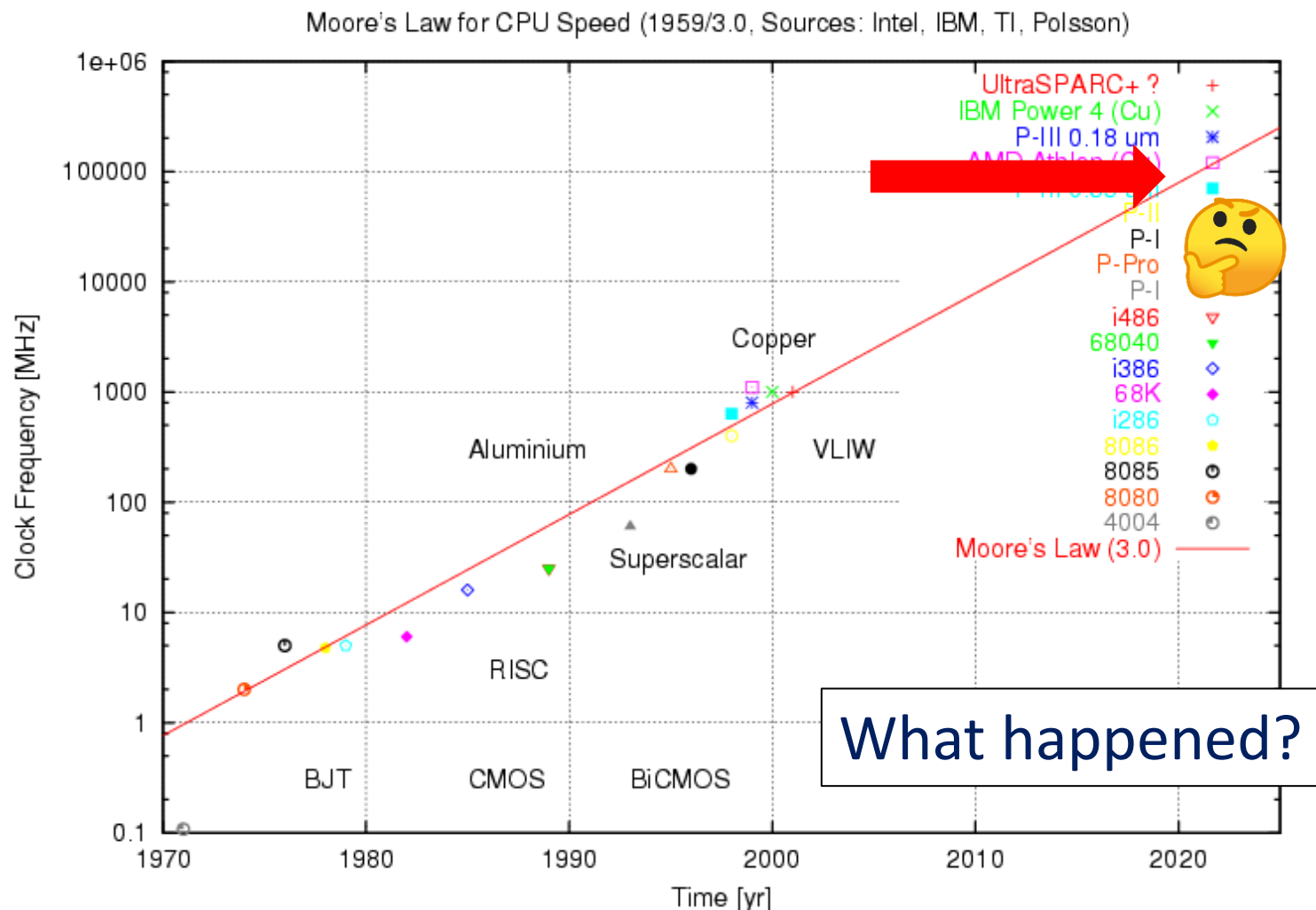
(Slightly) more accurate processor power consumption

Gate-oxide stopped scaling

Stopped scaling due to leakage
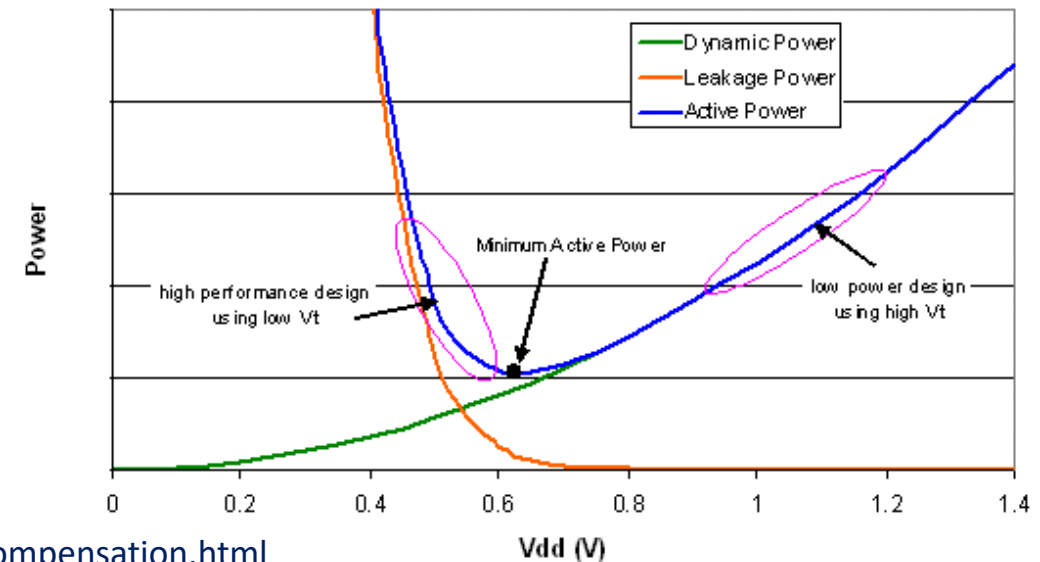
$$Power = (ActiveTransistors \times Capacitance \times Voltage^2 \times Frequency)$$

Dynamic power

$$+ (Voltage \times Leakage)$$

Static power

Total power consumption with constant frequency

https://www.design-reuse.com/articles/20296/power-management-leakage-control-process-compensation.html

End of Dennard Scaling

❑ Even with smaller transistors, we cannot continue reducing power
  o What do we do now?

❑ Option 1: Continue scaling frequency at increased power budget
  o Chip quickly become too hot to cool!
  o Thermal runaway:
           Hotter chip → increased resistance → hotter chip → …

Option 1: Continue scaling frequency at increased power budget
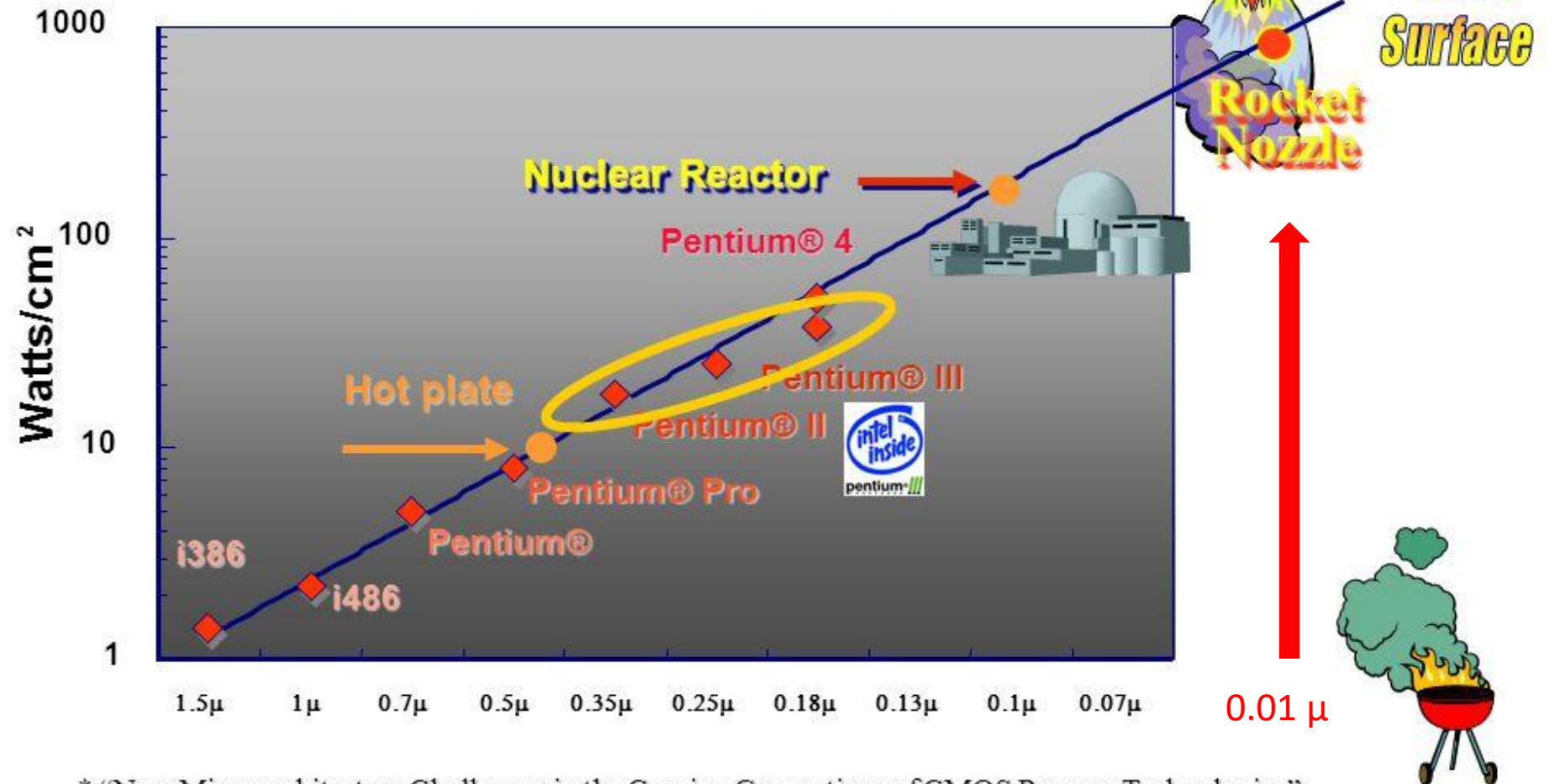
* "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies" – Fred Pollack, Intel Corp. Micro32 conference key note - 1999.

Option 2: Stop frequency scaling

Danowitz et.al., "CPU DB: Recording Microprocessor History," Communications of the ACM, 2012
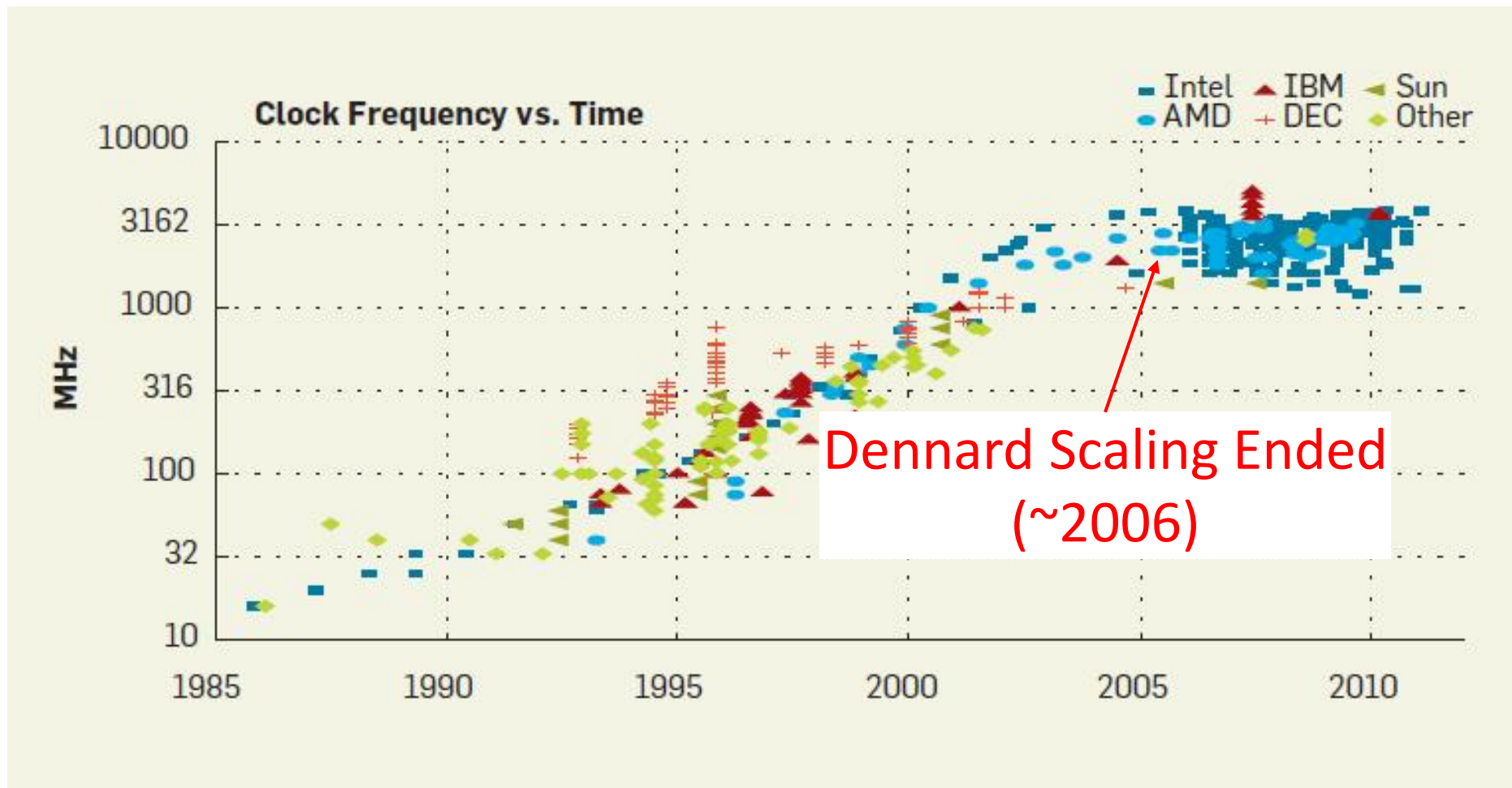
Looking back: change of predictions

Kogge et. al., "Yearly update : exascale projections for 2013,"Sandia National Laboratoris, 2013

But Moore's Law continues beyond 2006

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Our World in Data

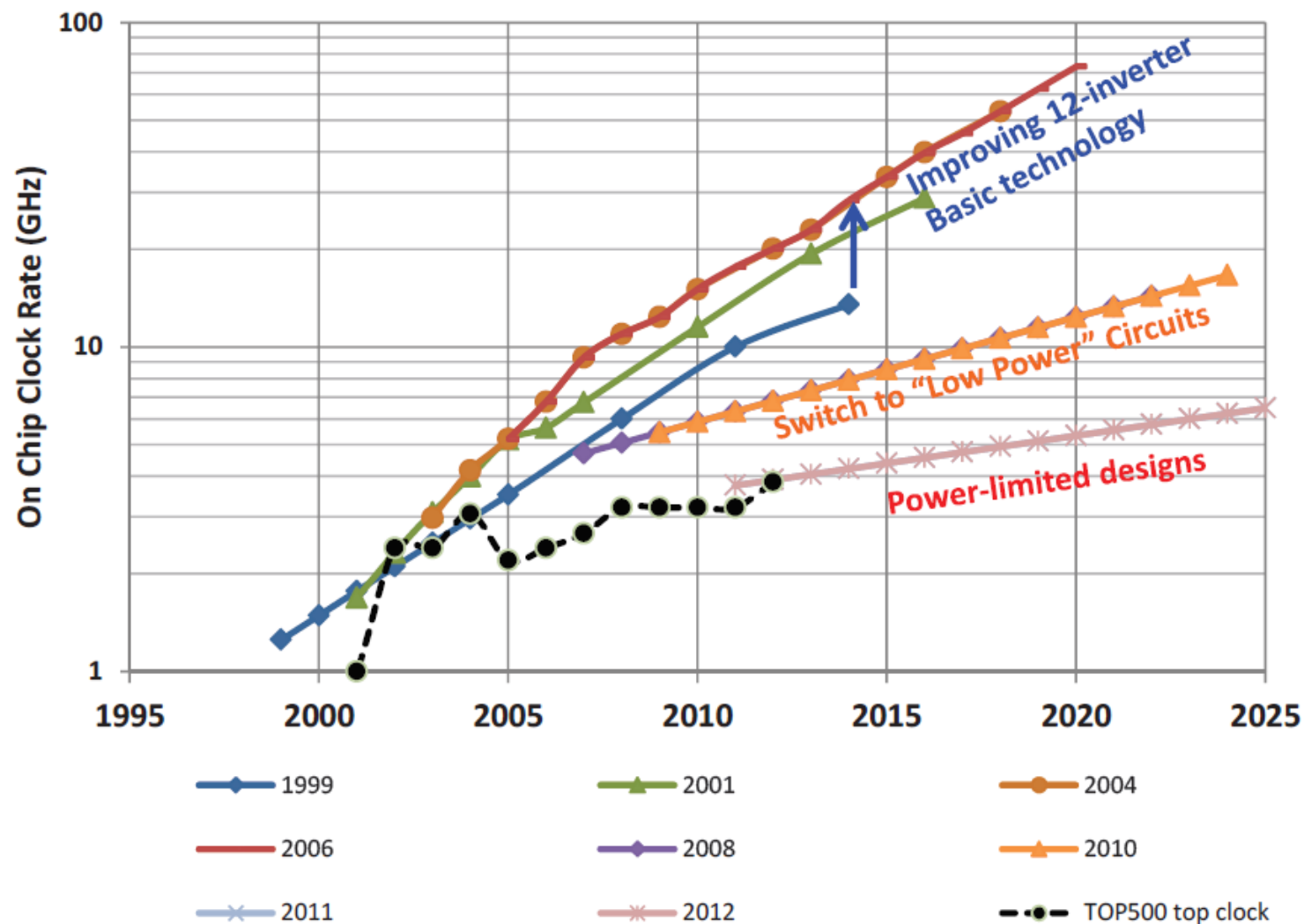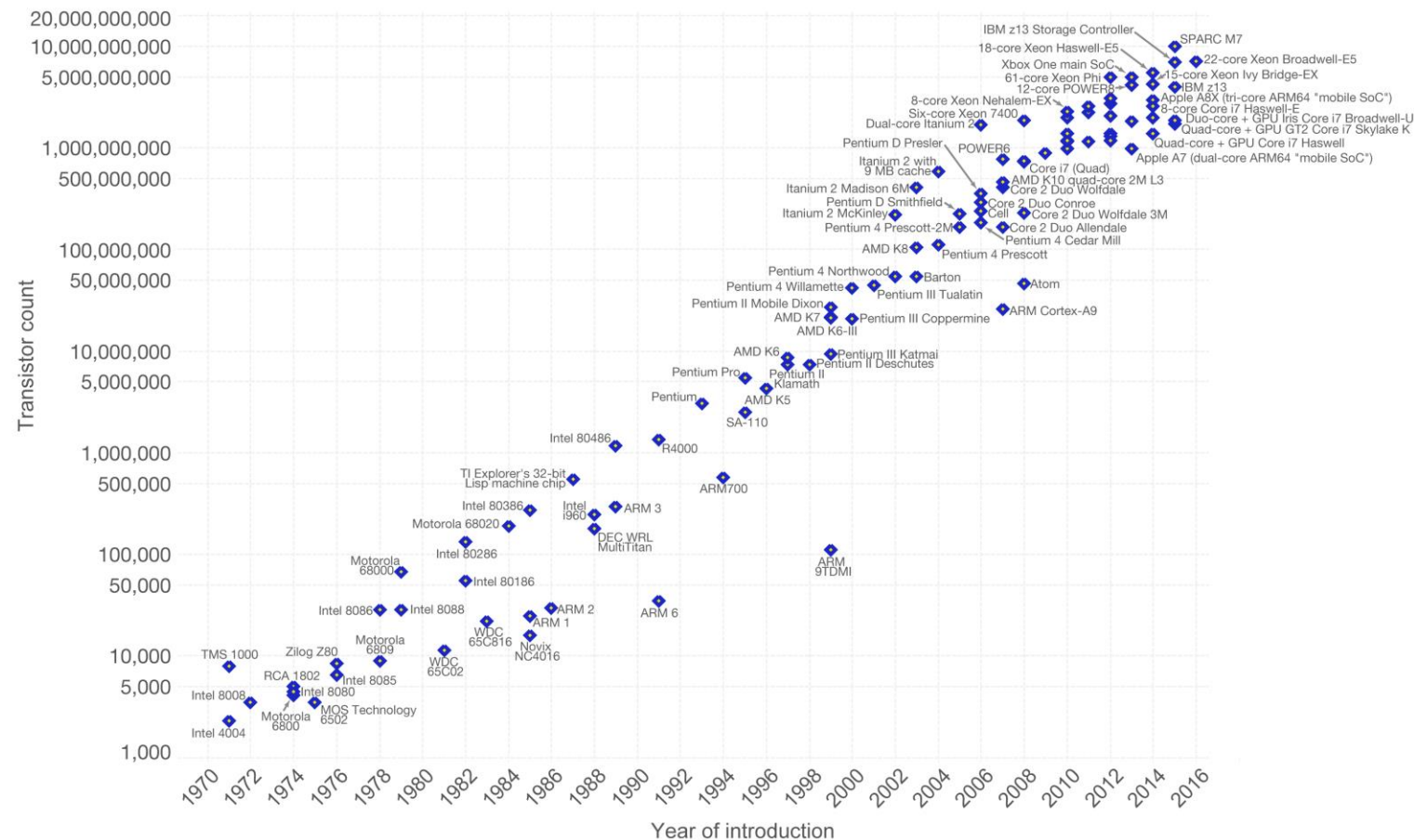Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.
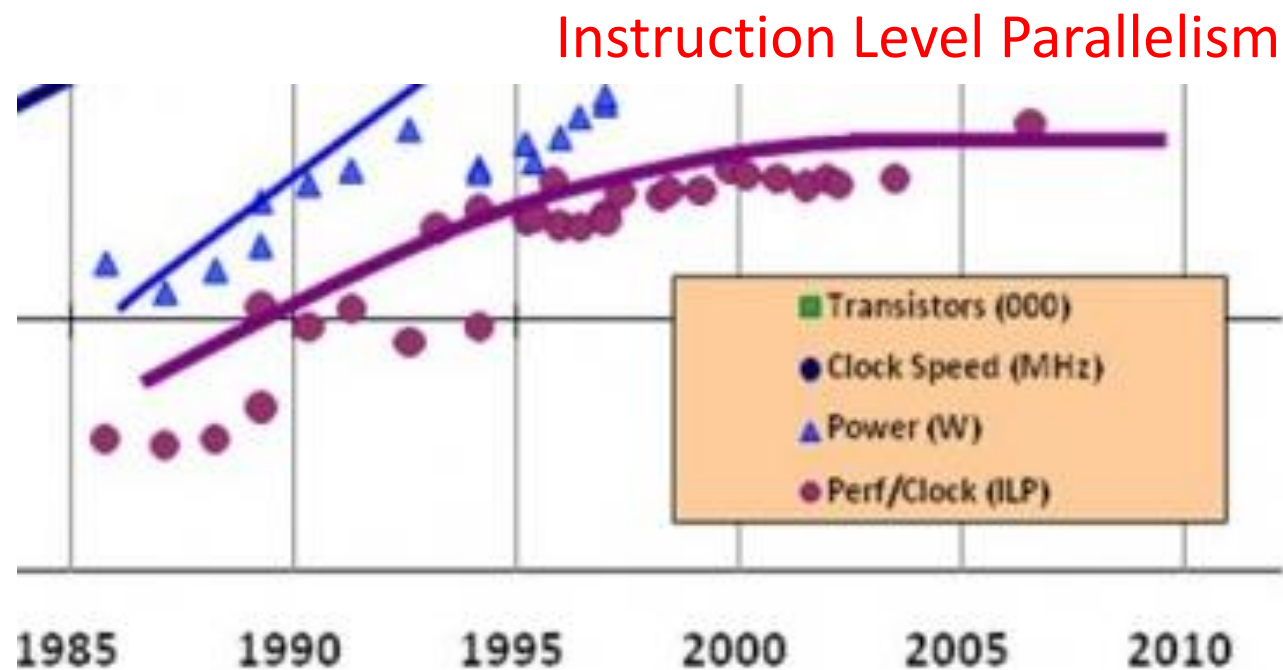
Transistor count (y-axis): 20,000,000,000; 10,000,000,000; 5,000,000,000; 1,000,000,000; 500,000,000; 100,000,000; 50,000,000; 10,000,000; 5,000,000; 1,000,000; 500,000; 100,000; 50,000; 10,000; 5,000; 1,000

Year of introduction (x-axis): 1970, 1972, 1974, 1976, 1978, 1980, 1982, 1984, 1986, 1988, 1990, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014, 2016

Data points labeled:
IBM z13 Storage Controller, SPARC M7, 18-core Xeon Haswell-E5, 22-core Xeon Broadwell-E5, Xbox One main SoC, 15-core Xeon Ivy Bridge-EX, 61-core Xeon Phi, IBM z13, 12-core POWER8, 8-core Xeon Nehalem-EX, Apple A8X (tri-core ARM64 "mobile SoC"), Six-core Xeon 7400, 8-core Core i7 Haswell-E, Dual-core Itanium 2, Duo-core + GPU Iris Core i7 Broadwell-U, Quad-core + GPU GT2 Core i7 Skylake K, Pentium D Presler, POWER6, Quad-core + GPU Core i7 Haswell, Itanium 2 with 9 MB cache, Core i7 (Quad), Apple A7 (dual-core ARM64 "mobile SoC"), AMD K10 quad-core 2M L3, Itanium 2 Madison 6M, Core 2 Duo Wolfdale, Pentium D Smithfield, Core 2 Duo Conroe, Itanium 2 McKinley, Cell, Core 2 Duo Wolfdale 3M, Pentium 4 Prescott-2M, Core 2 Duo Allendale, AMD K8, Pentium 4 Cedar Mill, Pentium 4 Prescott, Pentium 4 Northwood, Barton, Pentium 4 Willamette, Pentium III Tualatin, Atom, Pentium II Mobile Dixon, AMD K7, Pentium III Coppermine, ARM Cortex-A9, AMD K6-III, AMD K6, Pentium III Katmai, Pentium Pro, Pentium II Deschutes, Pentium II Klamath, Pentium, AMD K5, SA-110, Intel 80486, R4000, TI Explorer's 32-bit Lisp machine chip, ARM700, Intel 80386, Intel i960, ARM 3, Motorola 68020, DEC WRL MultiTitan, Intel 80286, ARM 9TDMI, Motorola 68000, Intel 80186, Intel 8086, Intel 8088, ARM 2, ARM 6, Motorola 6809, ARM 1, WDC 65C816, TMS 1000, Zilog Z80, Novix NC4016, WDC 65C02, RCA 1802, Intel 8085, Intel 8008, Intel 8080, MOS Technology 6502, Motorola 6800, Intel 4004

Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.
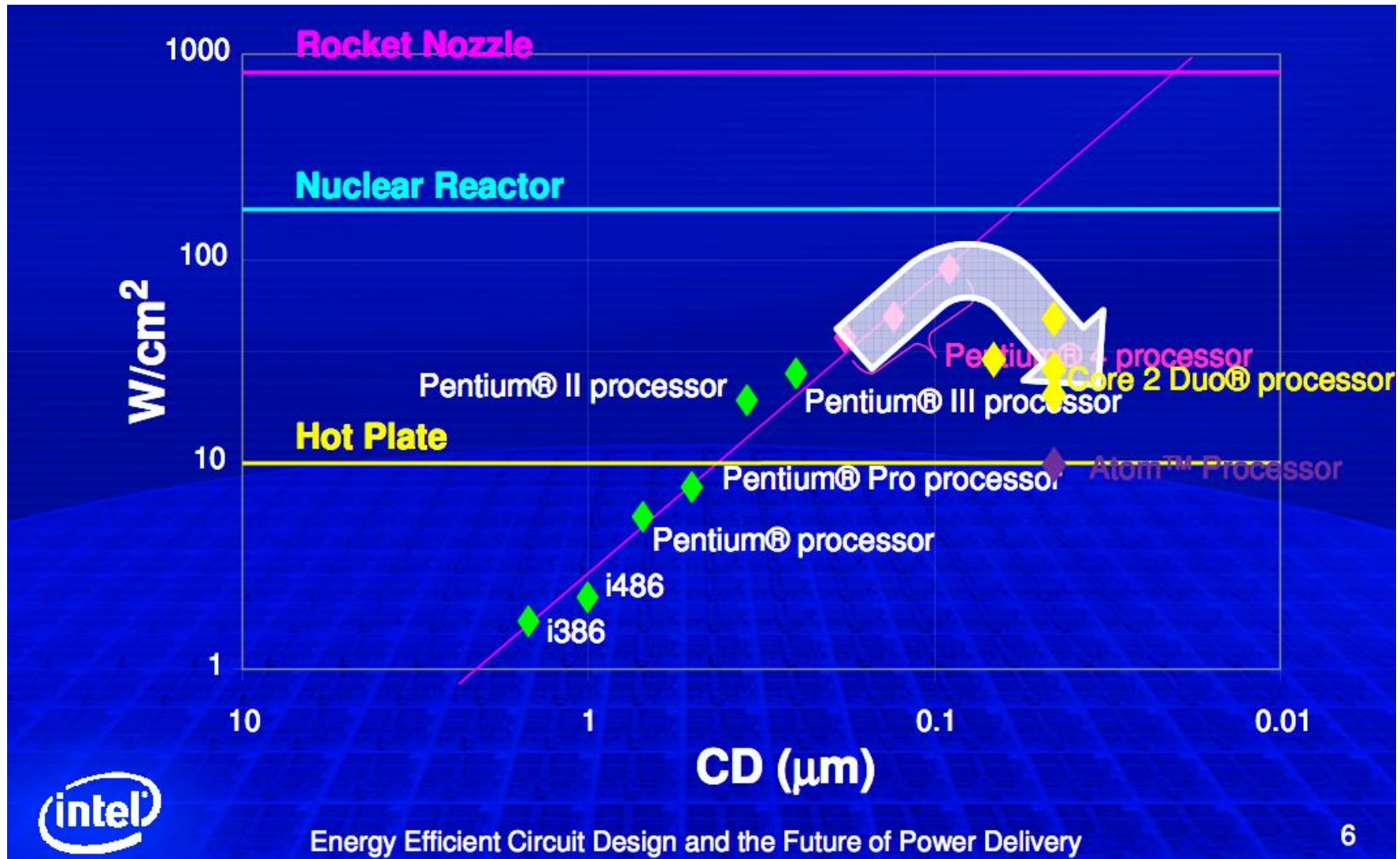Licensed under CC-BY-SA by the author Max Roser.

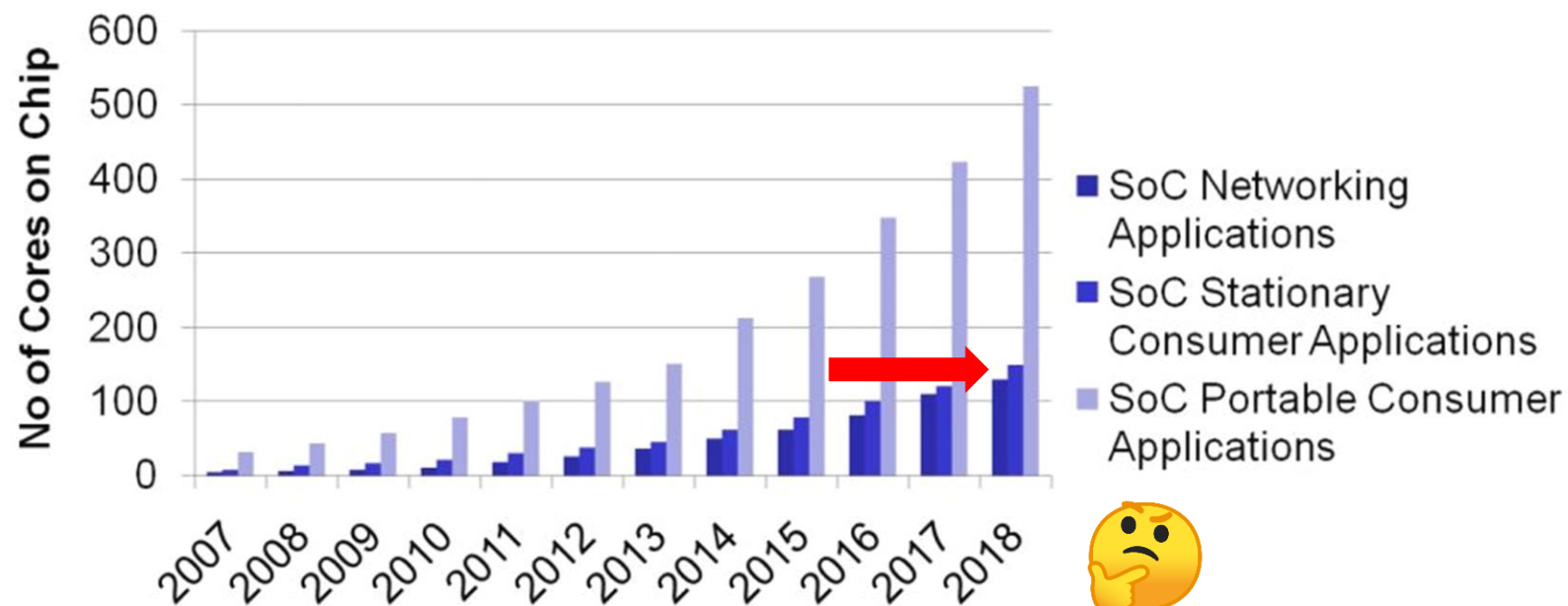# State of things at this point (2006)

❑ Single-thread performance scaling ended
- o Frequency scaling ended (Dennard Scaling)
- o Instruction-level parallelism scaling stalled ... also around 2005

❑ Moore's law continues
- o Double transistors every two years
- o What do we do with them?

Instruction Level Parallelism

- ■ Transistors (000)
- ● Clock Speed (MHz)
- ▲ Power (W)
- ● Perf/Clock (ILP)

K. Olukotun, "Intel CPU Trends"

Crisis averted with manycores?

Energy Efficient Circuit Design and the Future of Power Delivery

Crisis averted with manycores?

Source:
International Roadmap for Semiconductors 2007 edition (http://www.itrs.net/)

What happened?

Can't keep going up

Gate-oxide stopped scaling

Stopped scaling due to leakage

Stopped scaling due to thermal

$$Power =$$

$$(ActiveTransistors \times Capacitance \times Voltage^2 \times Frequency)$$

Dynamic power

Stopped scaling due to leakage

$$+ (Voltage \times LeakageCurrent)$$

Static power

"Utilization Wall"

Regardless of Moore's Law, a limited amount of gates can be active at a given time

Where To, From Here?

❑ The number of active transistors at a given time is limited

○ We won't get much performance improvements even if Moore's law continues

○ We need to make the best use of those active transistors!

Where To, From Here?

❑ Potential Solution 1: The software solution
- ○ Write efficient software to make the efficient use of hardware resources
- ○ No longer depend entirely on hardware performance scaling
- ○ "Performance engineering" software, using hardware knowledge

❑ Solution 2: The specialized architectural solution
- ○ Chip space is now cheap, but power is expensive
- ○ Stop depending on more complex general-purpose cores
- ○ Use space to build heterogeneous systems,
  with compute engines well-suited for each application

Taylor, Michael B. "Is dark silicon useful? Harnessing the four horsemen of the coming dark silicon apocalypse." DAC Design Automation Conference 2012. IEEE, 2012.

The Bottom Line:
Architecture is No Longer Transparent

❑ Optimized software requires architecture knowledge

❑ Special-purpose "accelerators" (GPU, FPGA, …) programmed explicitly

❑ Even general-purpose processors implement specialized instructions

  o Single-Instruction Multiple Data (SIMD) instructions such as AVX

  o Special-purpose instructions sets such as AES-NI