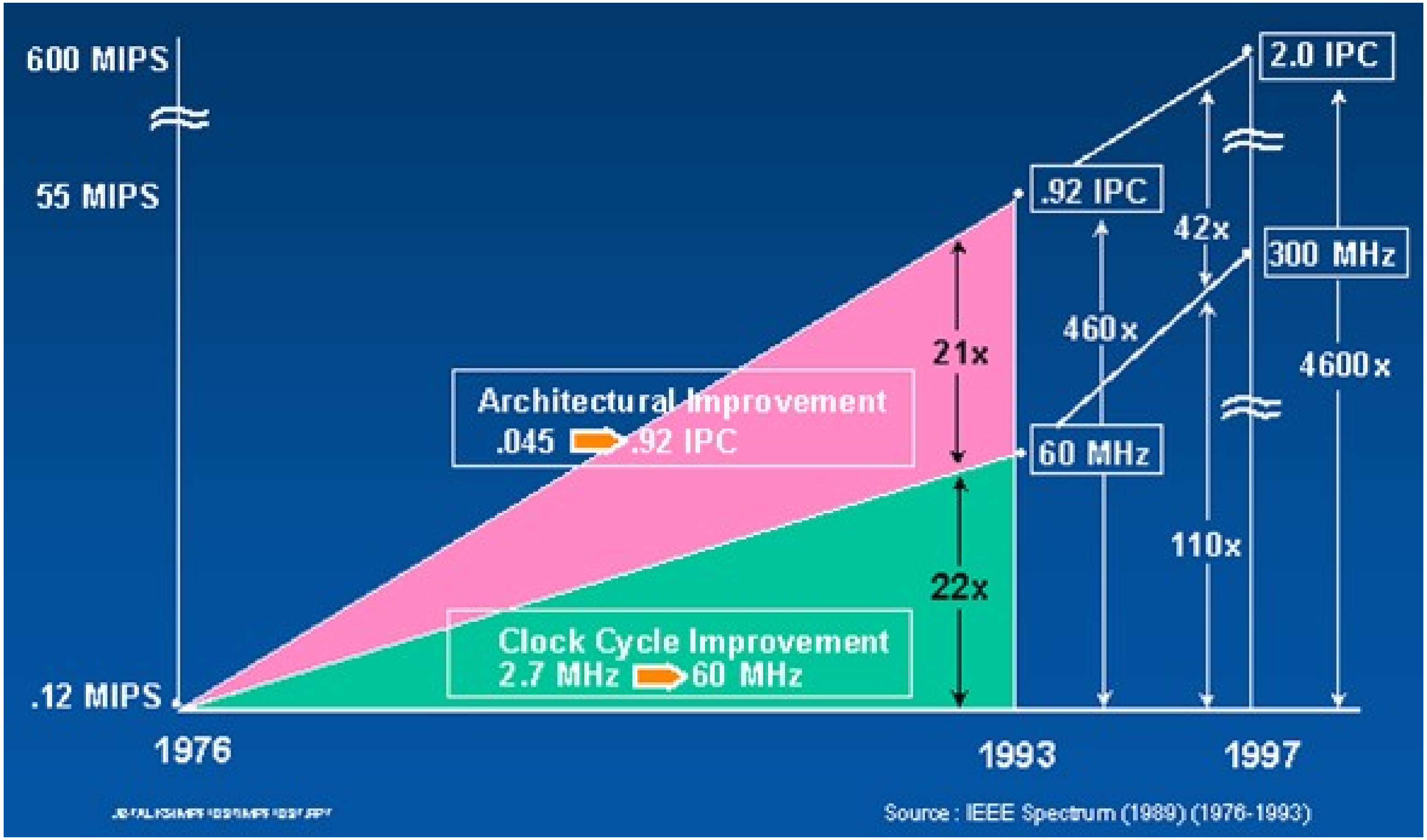
CUDA Programming

The Good Old Days for Software

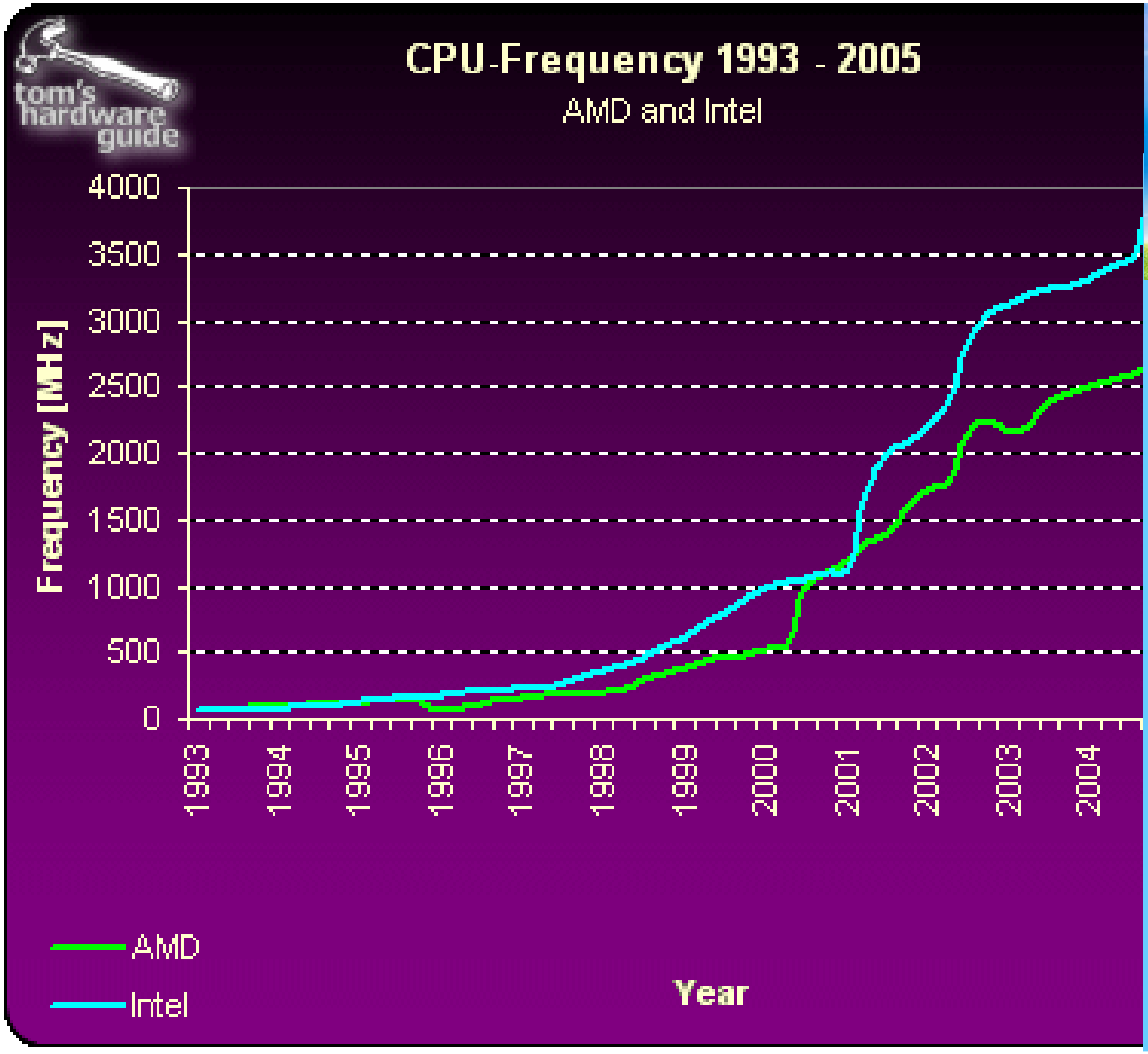
Source: J. Birnbaum



Source : IEEE Spectrum (1989) (1976-1993)

- Single-processor performance experienced dramatic improvements from **clock**, and **architectural** improvement (Pipelining, Instruction-Level-Parallelism).
- Applications experienced **automatic** performance improvement.

Hitting the Power Wall



toward

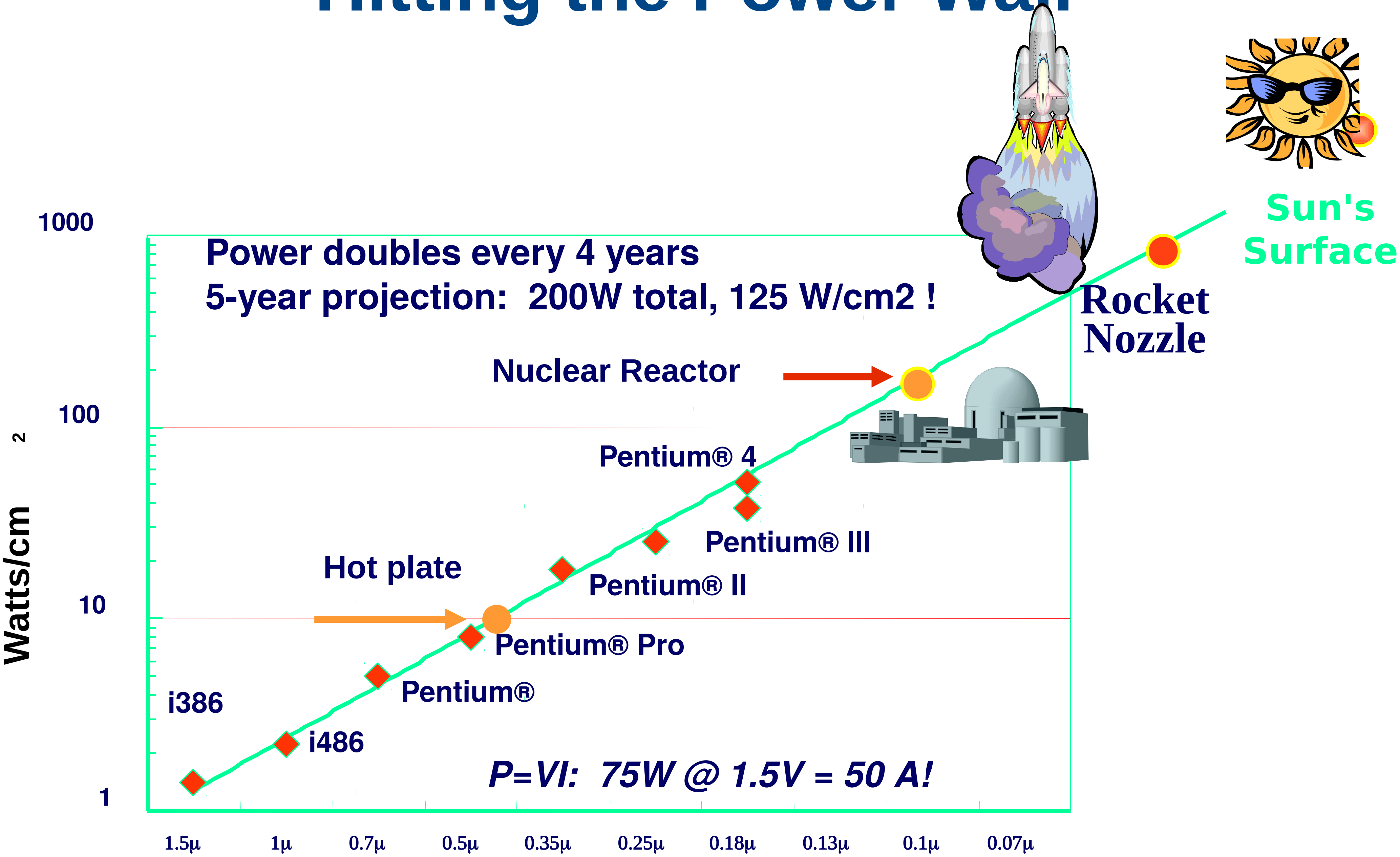
a

brighter

tomorrow

http://img.tomshardware.com/us/2005/11/21/the_mother_of_all_cpu_charts_2005/cpu_frequency.gif

Hitting the Power Wall



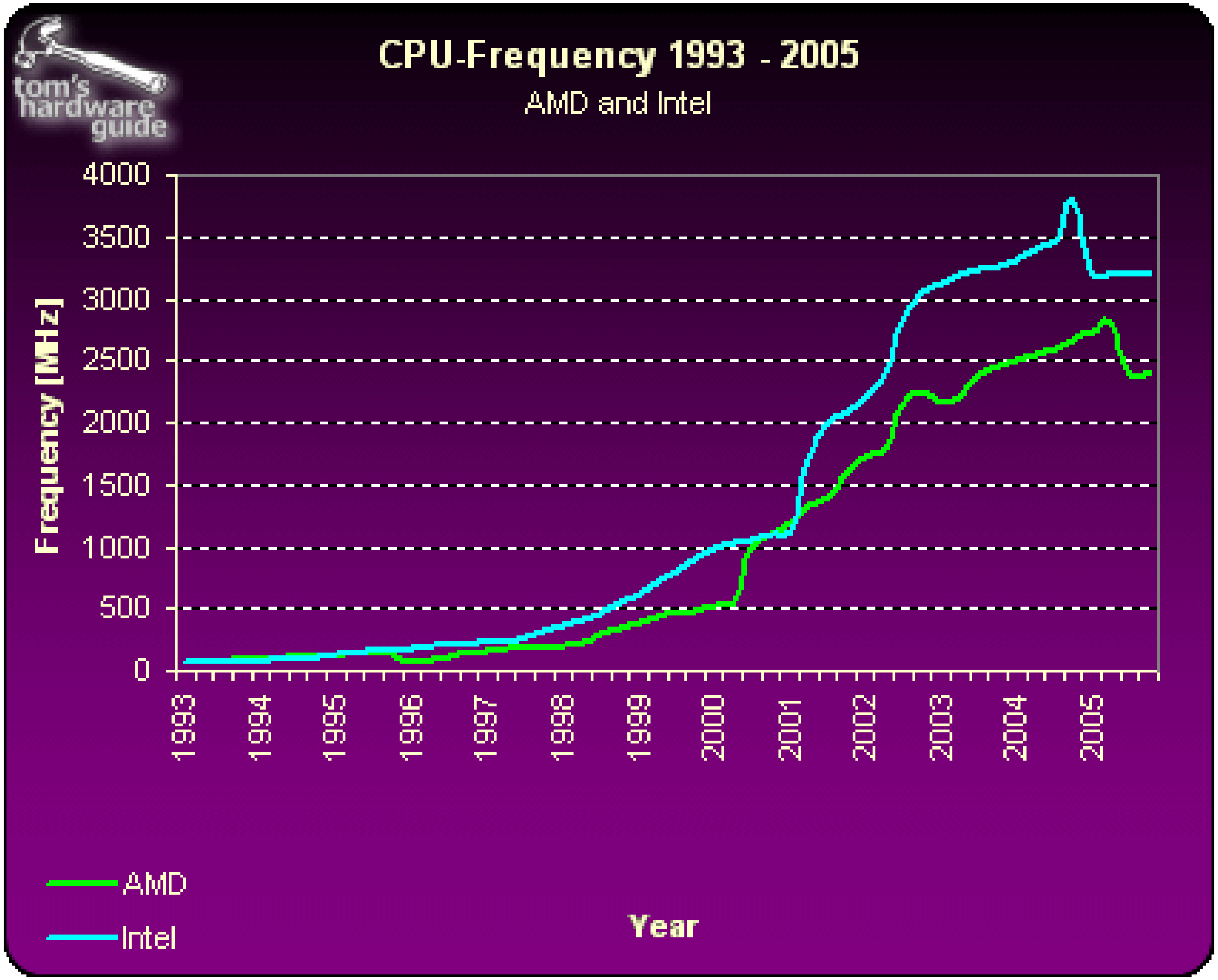
Power doubles every 4 years

5-year projection: 200W total, 125 W/cm2 !

P=VI: 75W @ 1.5V = 50 A!

“New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies” – Fred Pollack, Intel Corp. Micro32 conference key note - 1999. Courtesy Avi Mendelson, Intel.

Hitting the Power Wall



http://img.tomshardware.com/us/2005/11/21/the_mother_of_all_cpu_charts_2005/cpu_frequency.gif

2004 – Intel cancels Tejas and Jayhawk due to
heat problems due to the extreme power consumption of the core ...

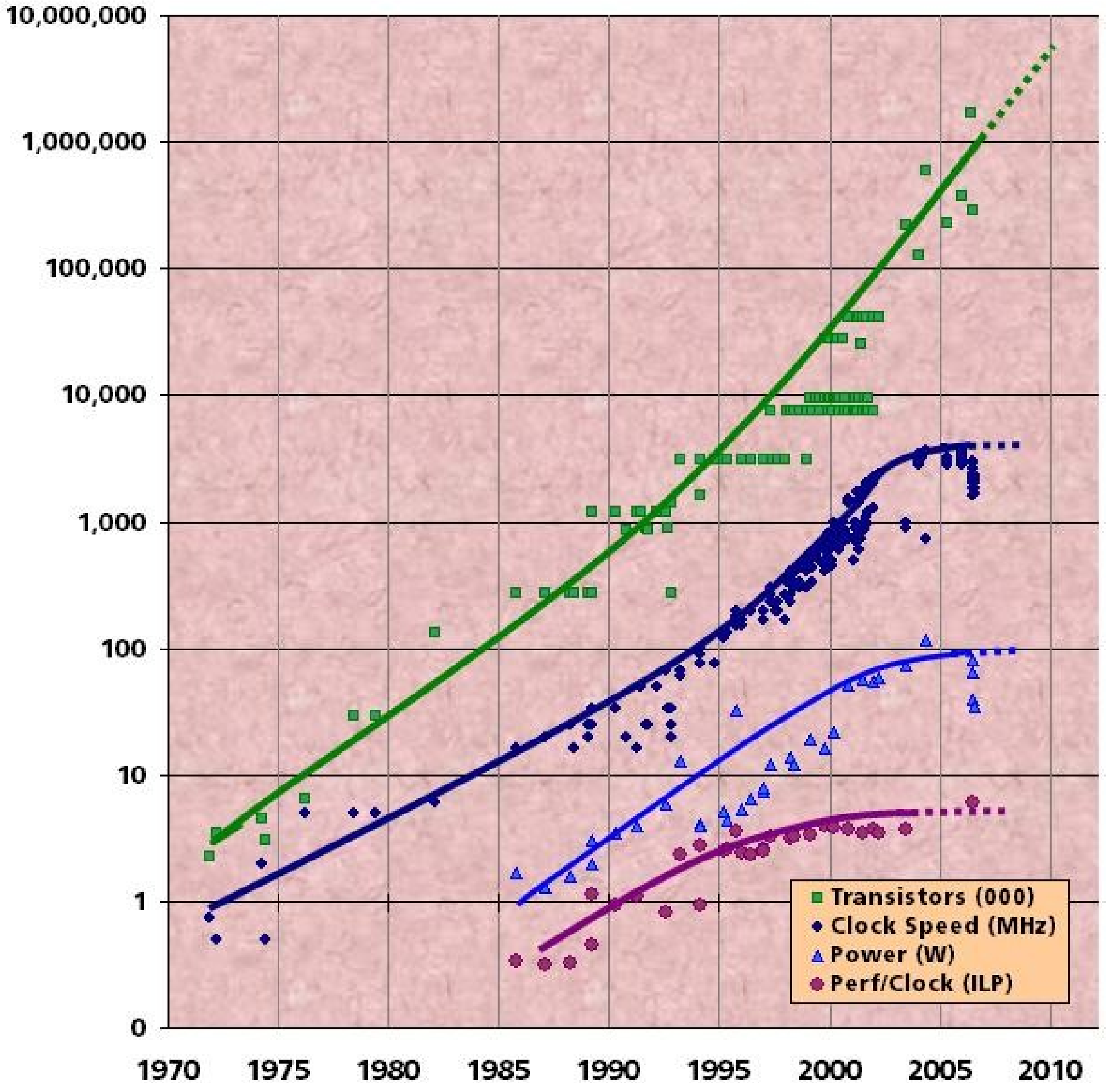
The Only Option: Use Many Cores

Chip density is increasing by ~2x every 2 years

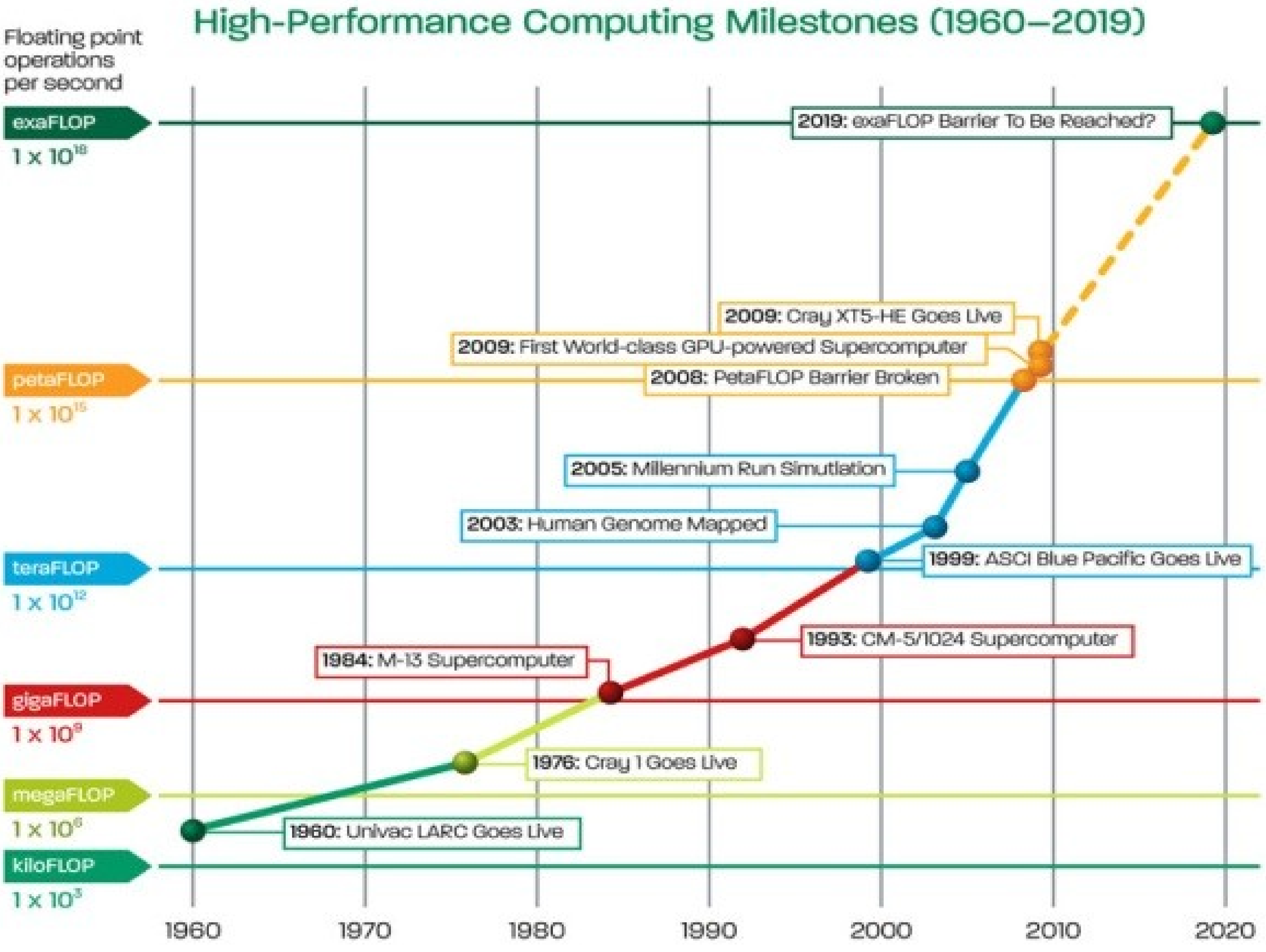
- Clock speed is not
- Number of processor cores may double

There is little or no more hidden parallelism (ILP) to be found

Parallelism must be exposed to and managed by software



Source: Intel, Microsoft (Sutter) and Stanford (Olukotun, Hammond)



High-Performance Computing Milestones (1960–2019)

2019: exaFLOP Barrier To Be Reached?

2009: Cray XT5-HE Goes Live

2009: First World-class GPU-powered Supercomputer

2008: PetaFLOP Barrier Broken

2005: Millennium Run Simutlation

2003: Human Genome Mapped

1999: ASCI Blue Pacific Goes Live

1993: CM-5/1024 Supercomputer

1984: M-13 Supercomputer

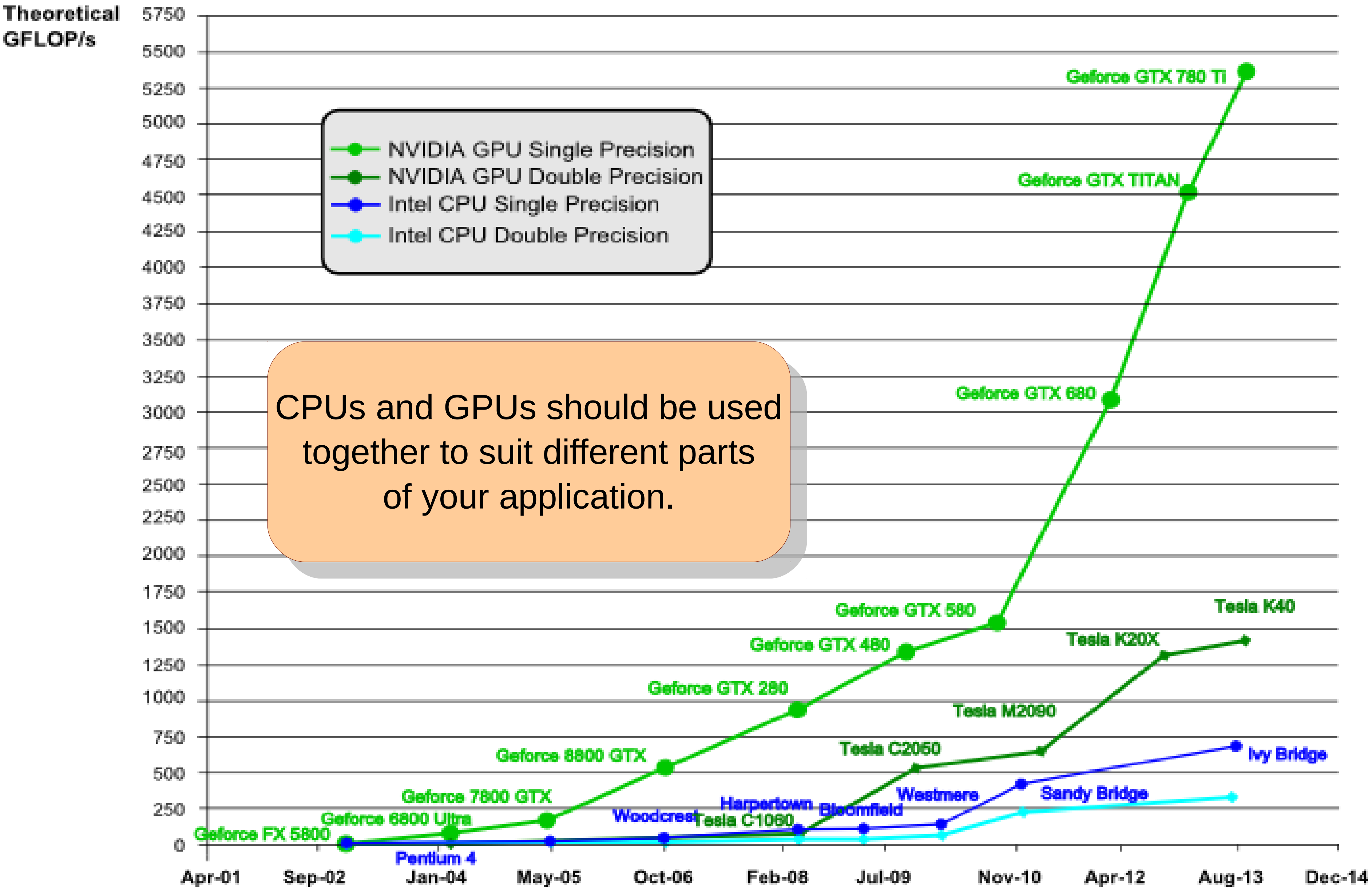
1976: Cray 1 Goes Live

1960: Univac LARC Goes Live

Parallel Platforms

- Shared memory systems (multi-core)
- Distributed systems (cluster)
- Graphics Processing Units (many-core)
- Field-Programmable Gate Arrays (configurable after manufacturing)
- Application-Specific Integrated Circuits
- Heterogeneous Systems

GPU-CPU Performance Comparison



CPUs and GPUs should be used together to suit different parts of your application.

Source: Thorsten Thormählen

In this course...

- Basic GPU Programming
  - Computation, Memory, Synchronization, Debugging
- Topics in GPU Programming
  - Unified virtual memory, multi-GPU, peer access

Logistics

- You need to arrange for your GPU.
  - Your laptop may have one.
  - With gmail account, you get some GPU time on Google colab.
  - You can use the central computing facilities at your institute.