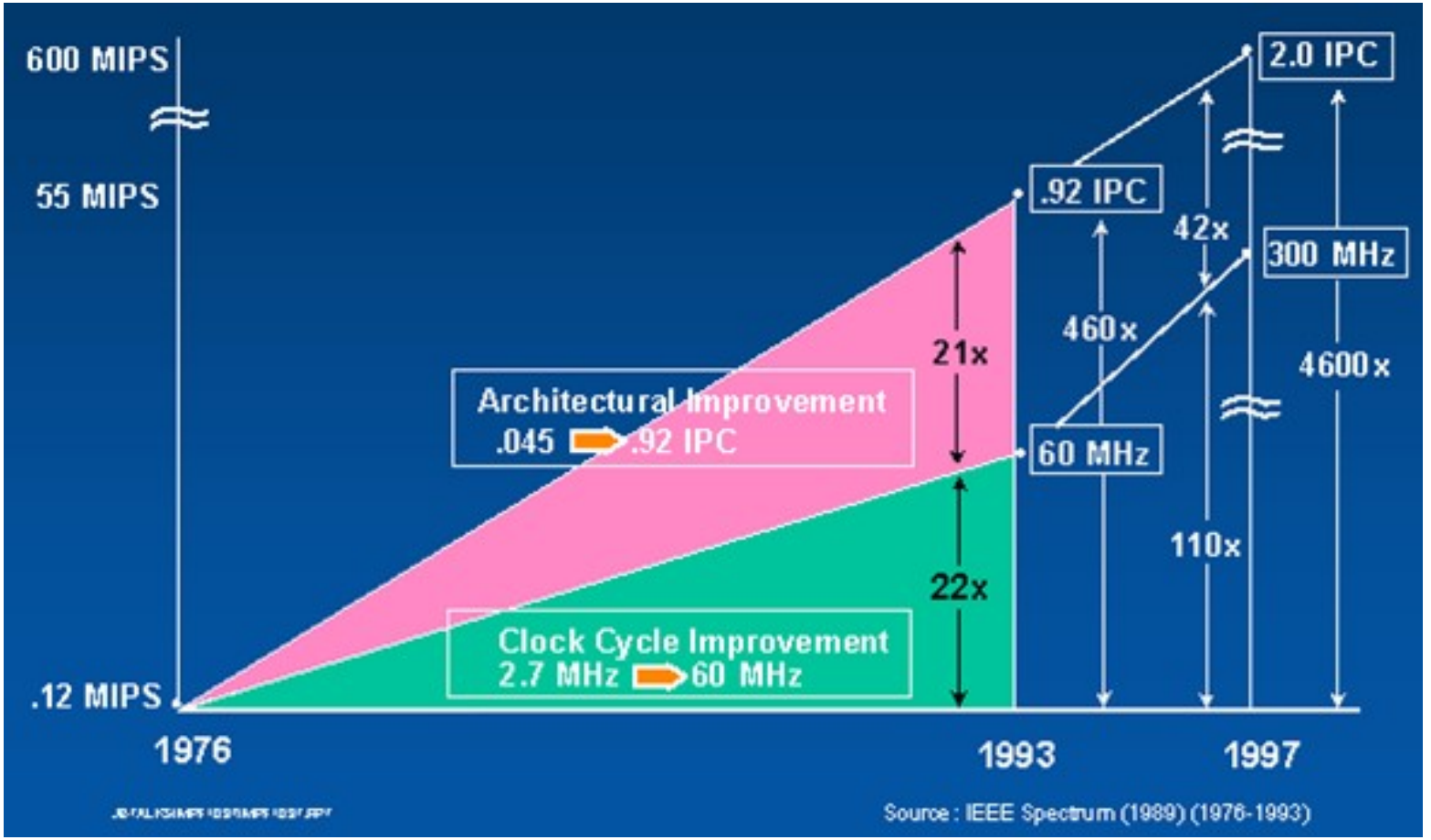
GPU Programming

Rupesh Nasre.
rupesh@cse.iitm.ac.in

IIT Madras
January 2022

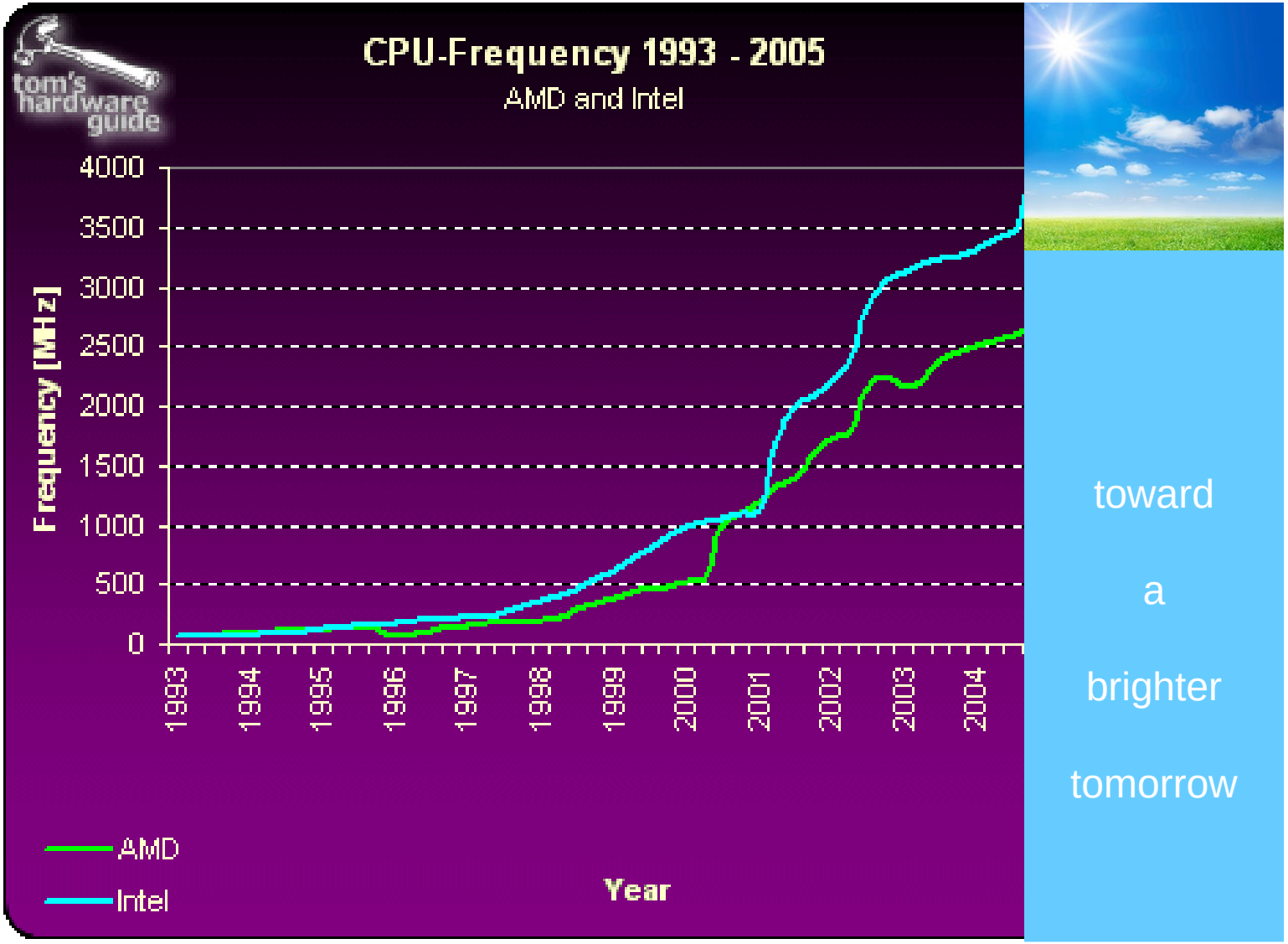
The Good Old Days for Software

Source: J. Birnbaum



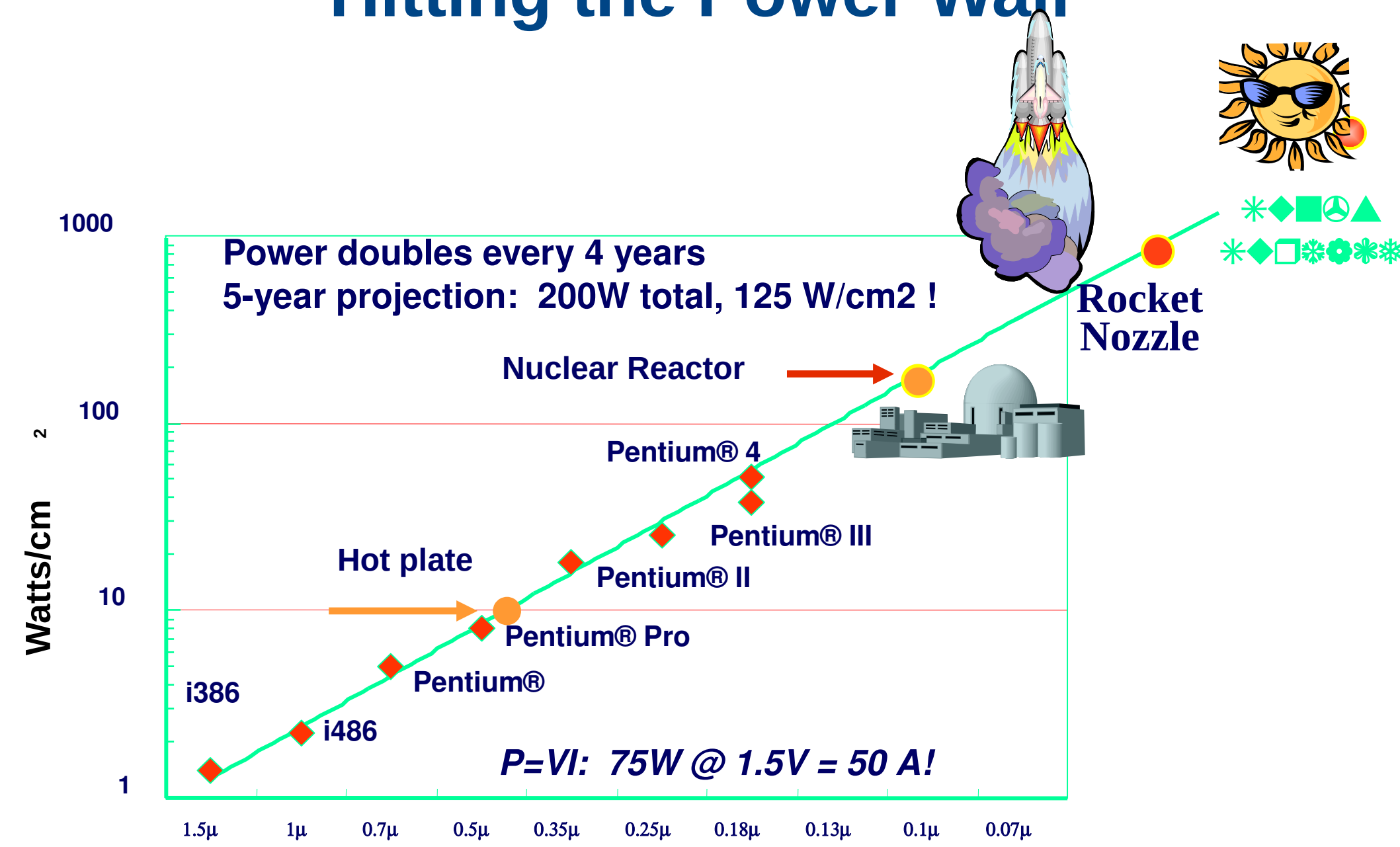
- Single-processor performance experienced dramatic improvements from **clock**, and **architectural** improvement (Pipelining, Instruction-Level-Parallelism).
- Applications experienced **automatic** performance improvement.

Hitting the Power Wall



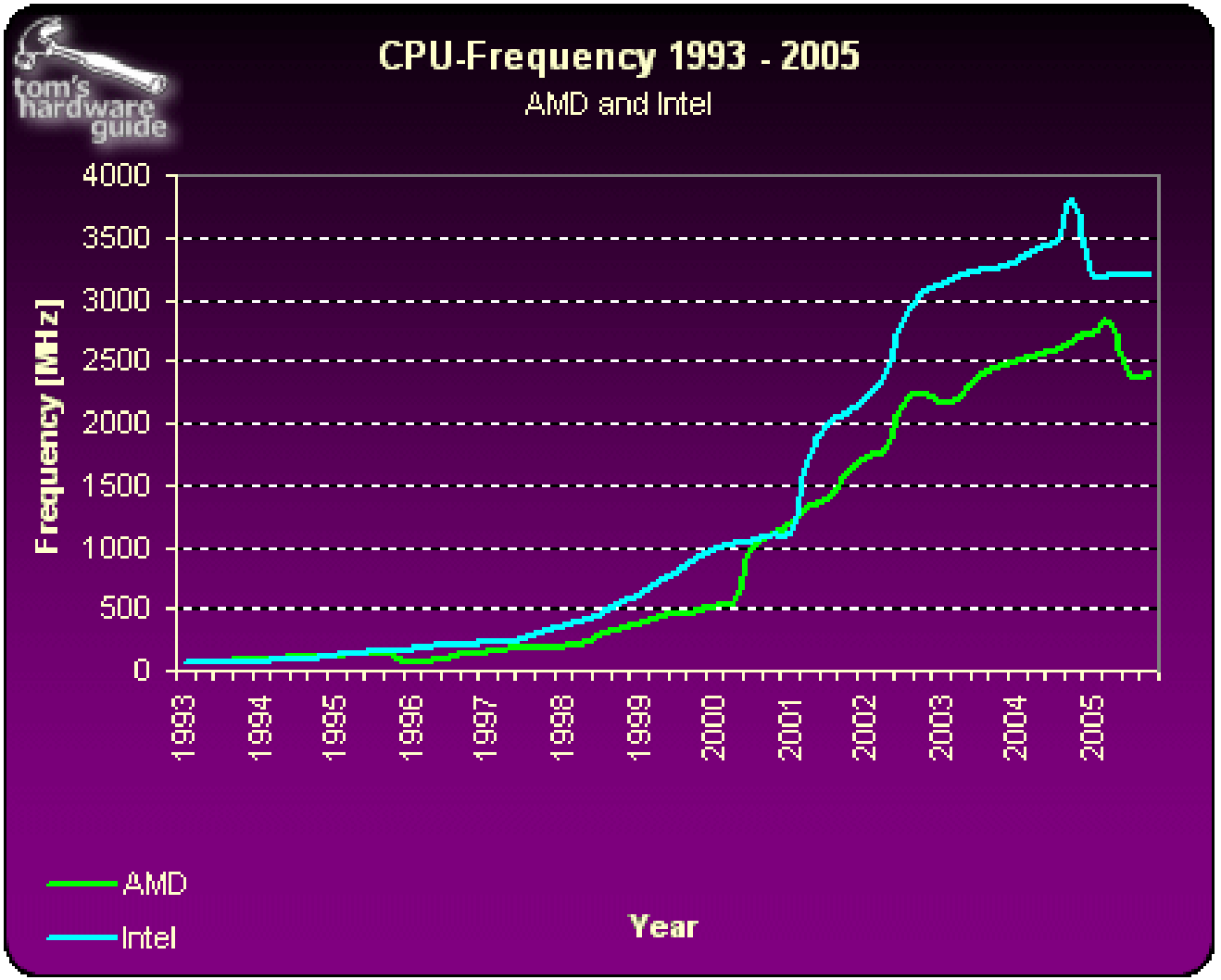
http://img.tomshardware.com/us/2005/11/21/the_mother_of_all_cpu_charts_2005/cpu_frequency.gif

Hitting the Power Wall



"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 ...

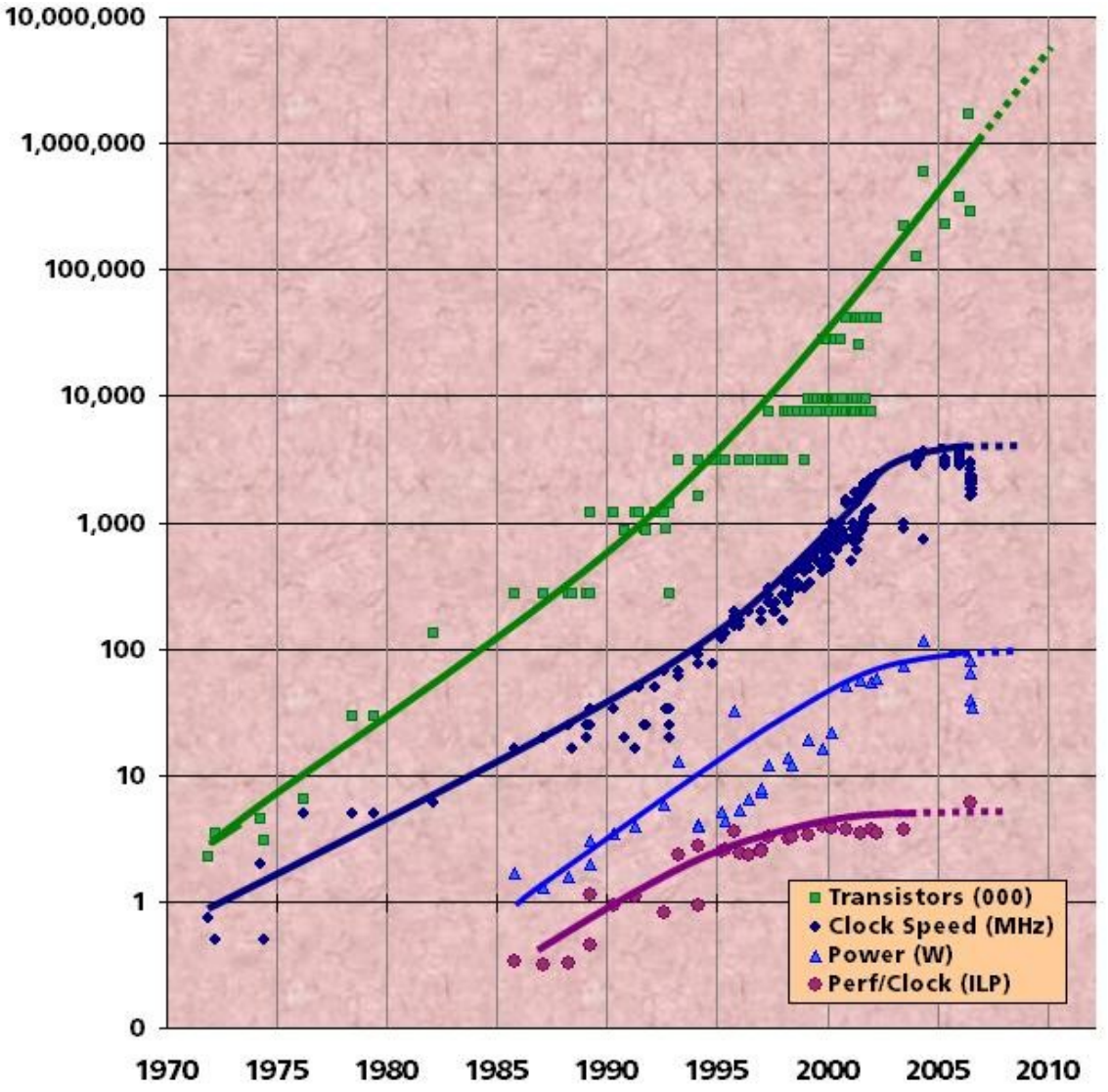
The Only Option: Use Many Cores

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

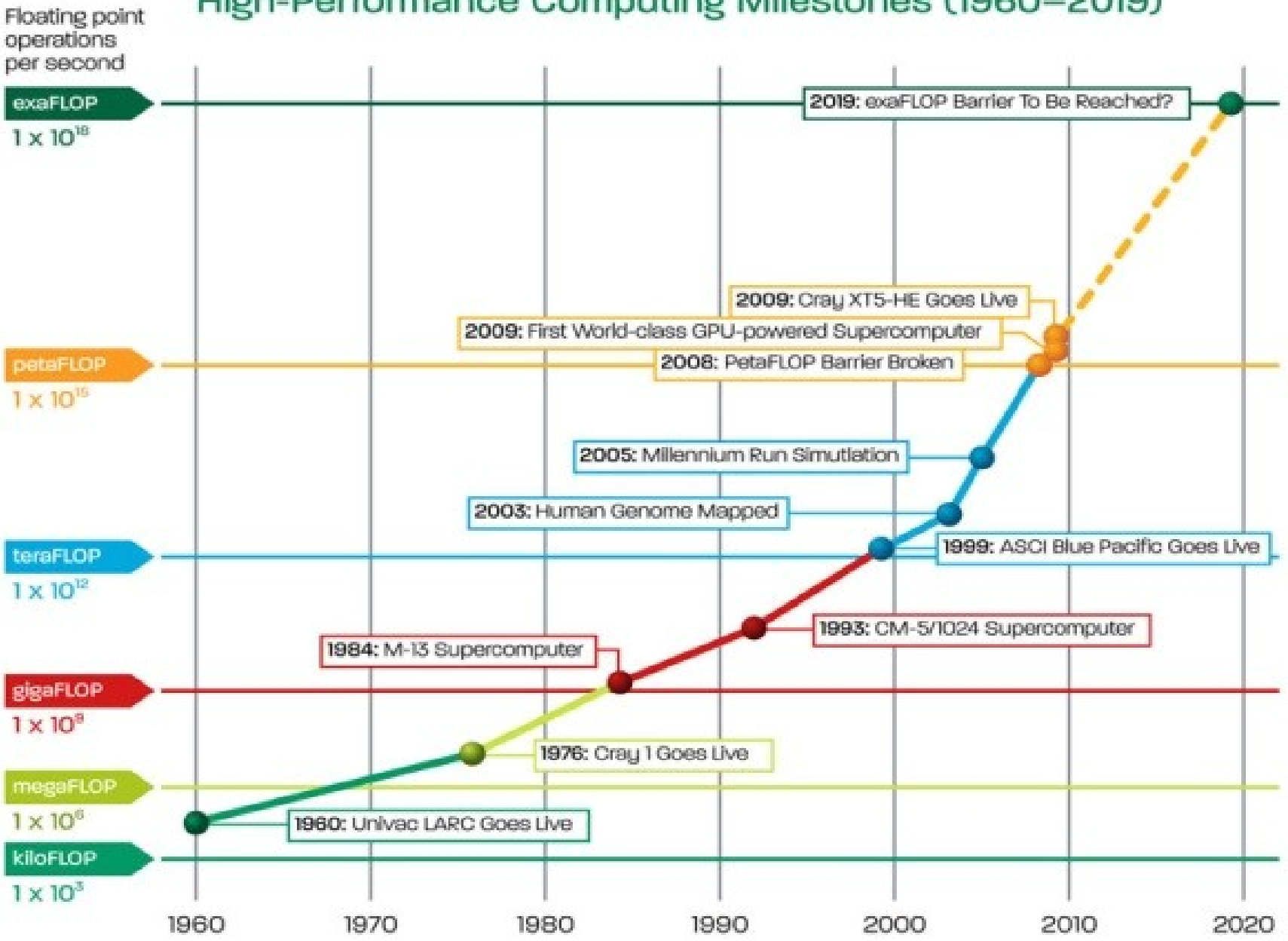
Parallelism must be exposed to and managed by software



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

AMD

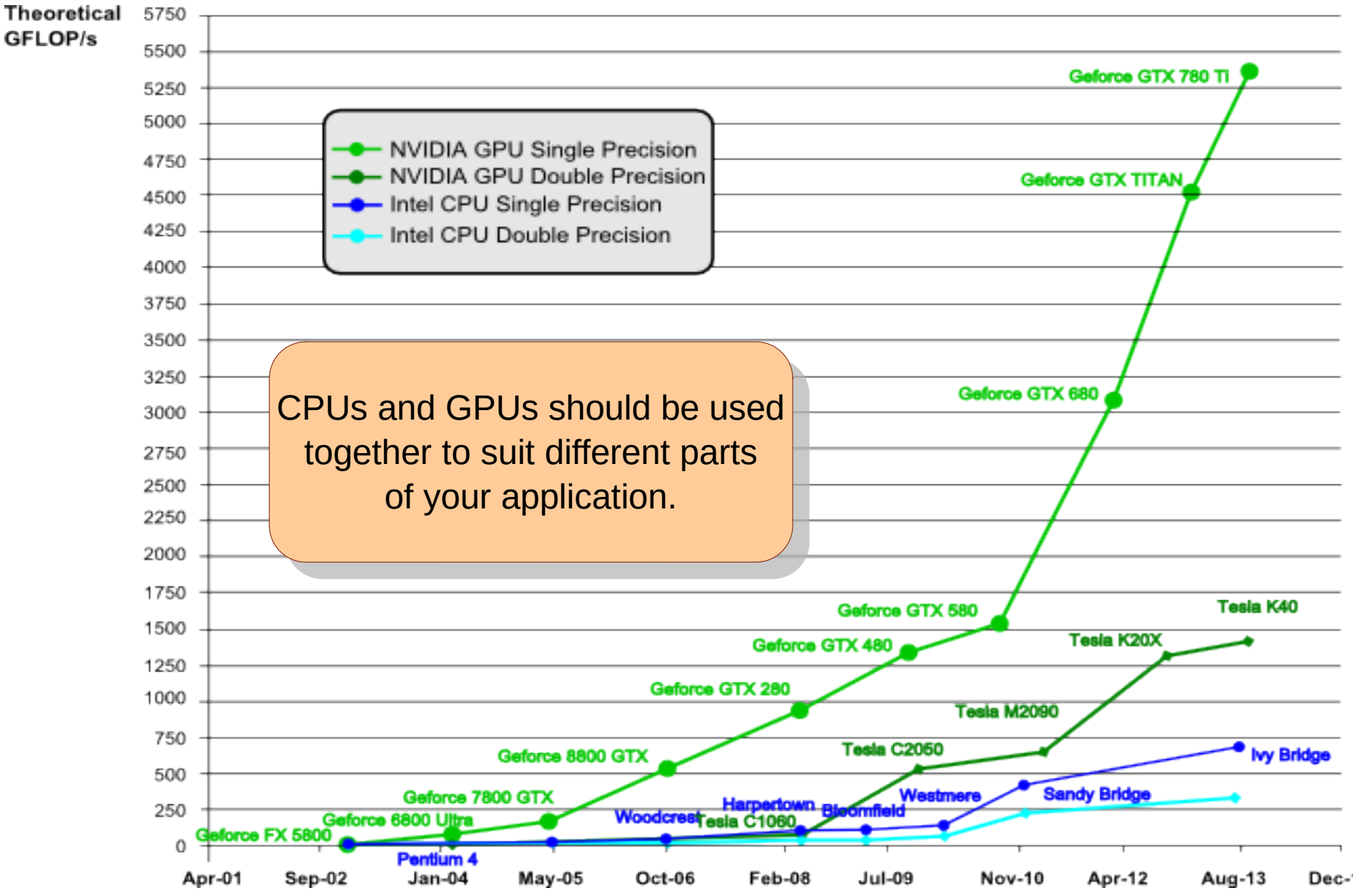
High-Performance Computing Milestones (1960–2019)



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



Source: Thorsten Thormählen

In this course...

- Basic GPU Programming
  - Computation, Memory, Synchronization, Debugging
- Advanced GPU Programming
  - Streams, Heterogeneous computing, Case studies
- Topics in GPU Programming
  - Unified virtual memory, multi-GPU, peer access

Logistics

- Tutorials and lectures would be intermixed.
  - And we can have separate doubt-sessions.
  - Video lectures are + would be available.
- You need to arrange for your GPU.
  - Your laptop may have one.
  - With gmail account, you get some GPU time on Google cloud (preferred by many in the past).
  - You can use the central computing facilities at the institute.

Logistics

- **Evaluation**
  - Four assignments (10 + 15 + 15 + **20**)
  - MidSem (20) + EndSem (20)
  - Dates are on the course webpage.
  - You have this week to suggest changes to dates.
- **Moodle**
  - Your responsibility to subscribe to it.
  - Exams would be on moodle.
  - Assignments are to be submitted on moodle.

To get the MOST out of this course

- Keep hands away from mouse, keyboard, and whatsapp.
- Solve questions during classwork.
  - Keep a copy with you. Take notes.
- Ask questions (others also haven't understood).
  - Do not let a few dominate the discussion.