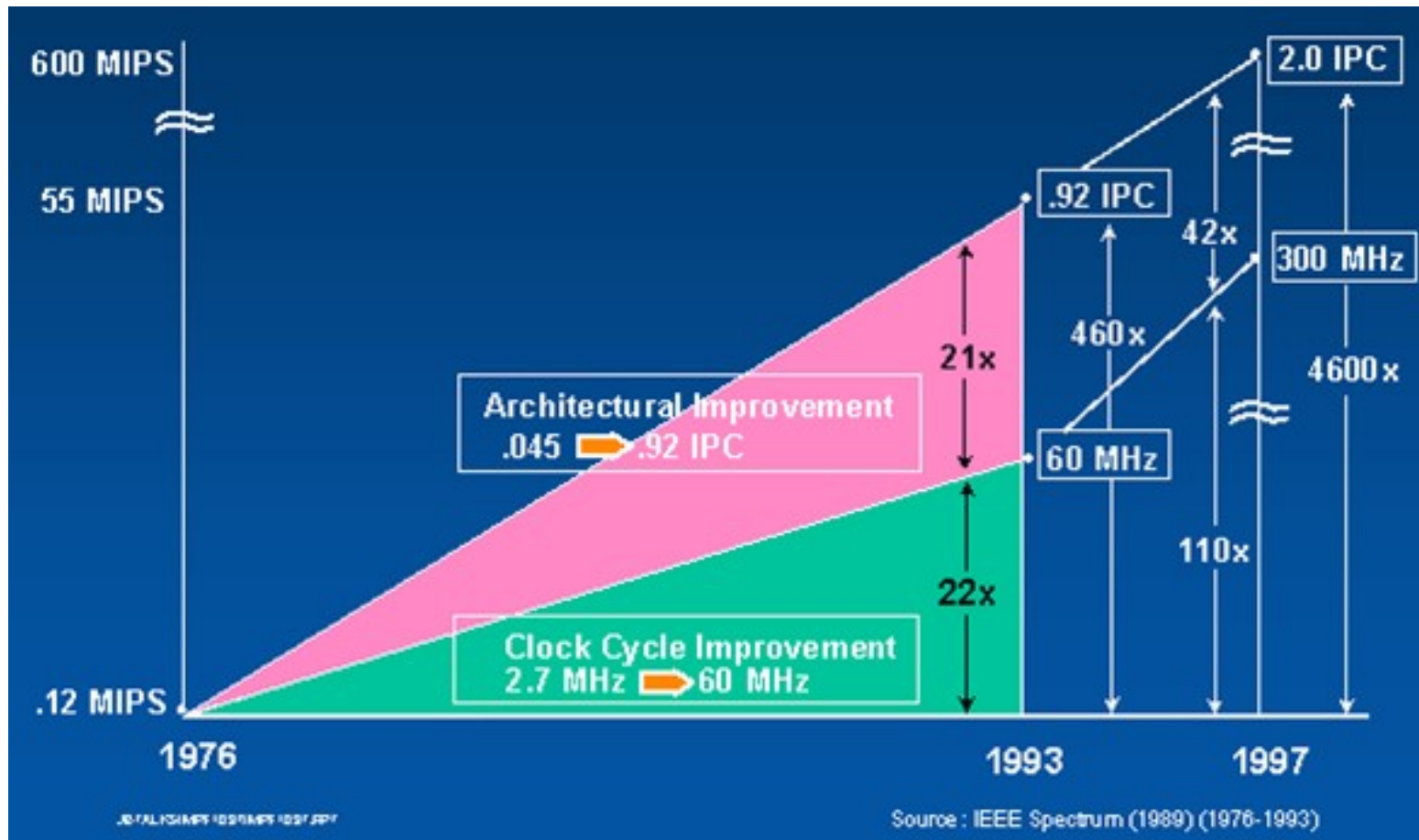# GPU Programming

Rupesh Nasre.

rupesh@cse.iitm.ac.in

IIT Madras
January 2023
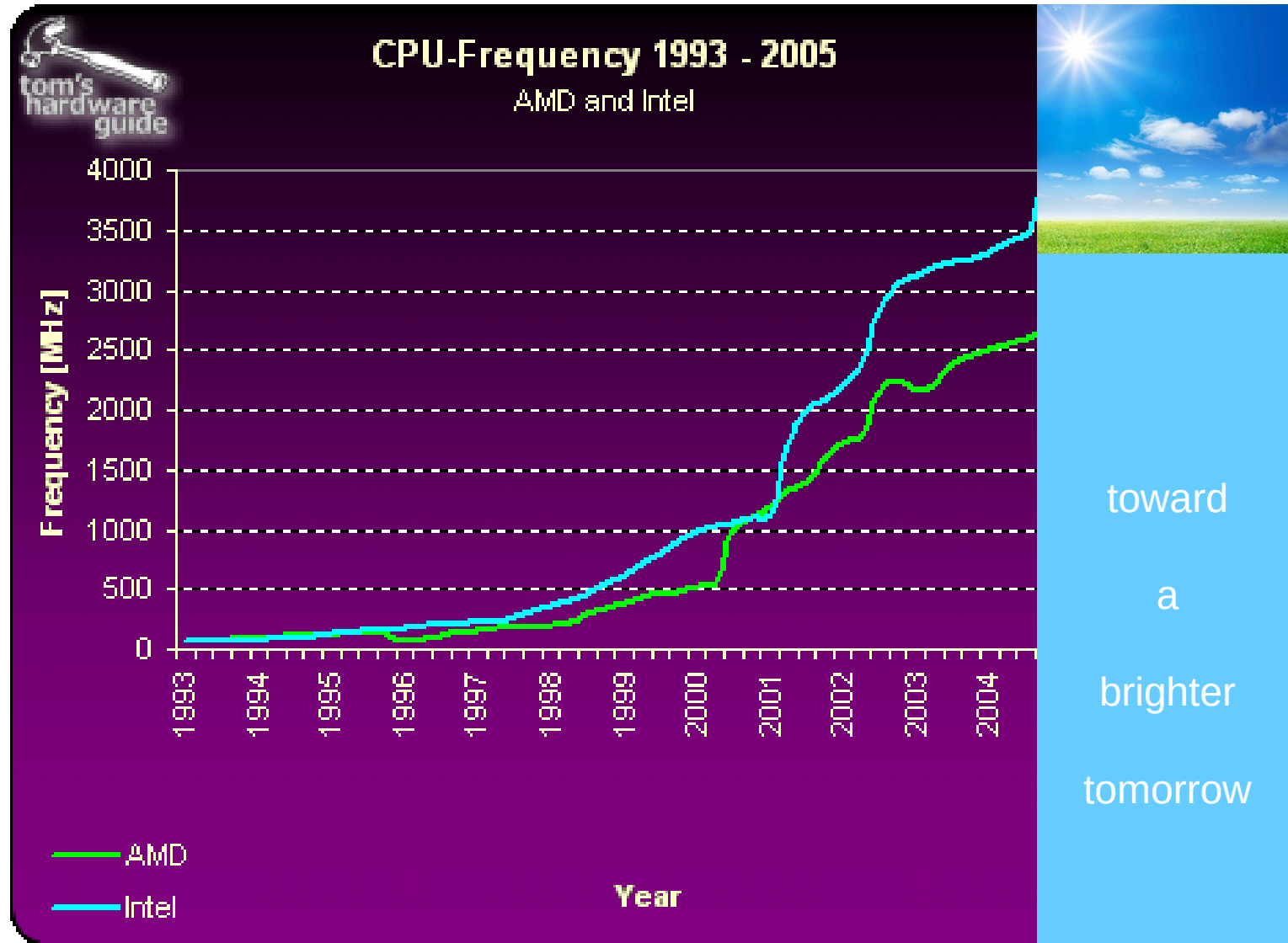
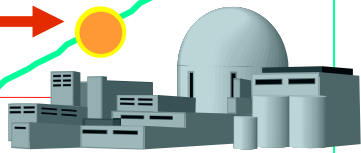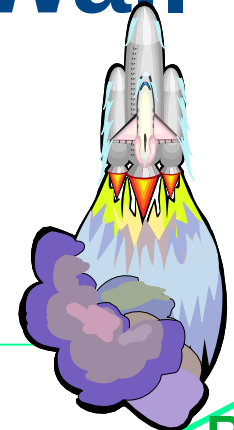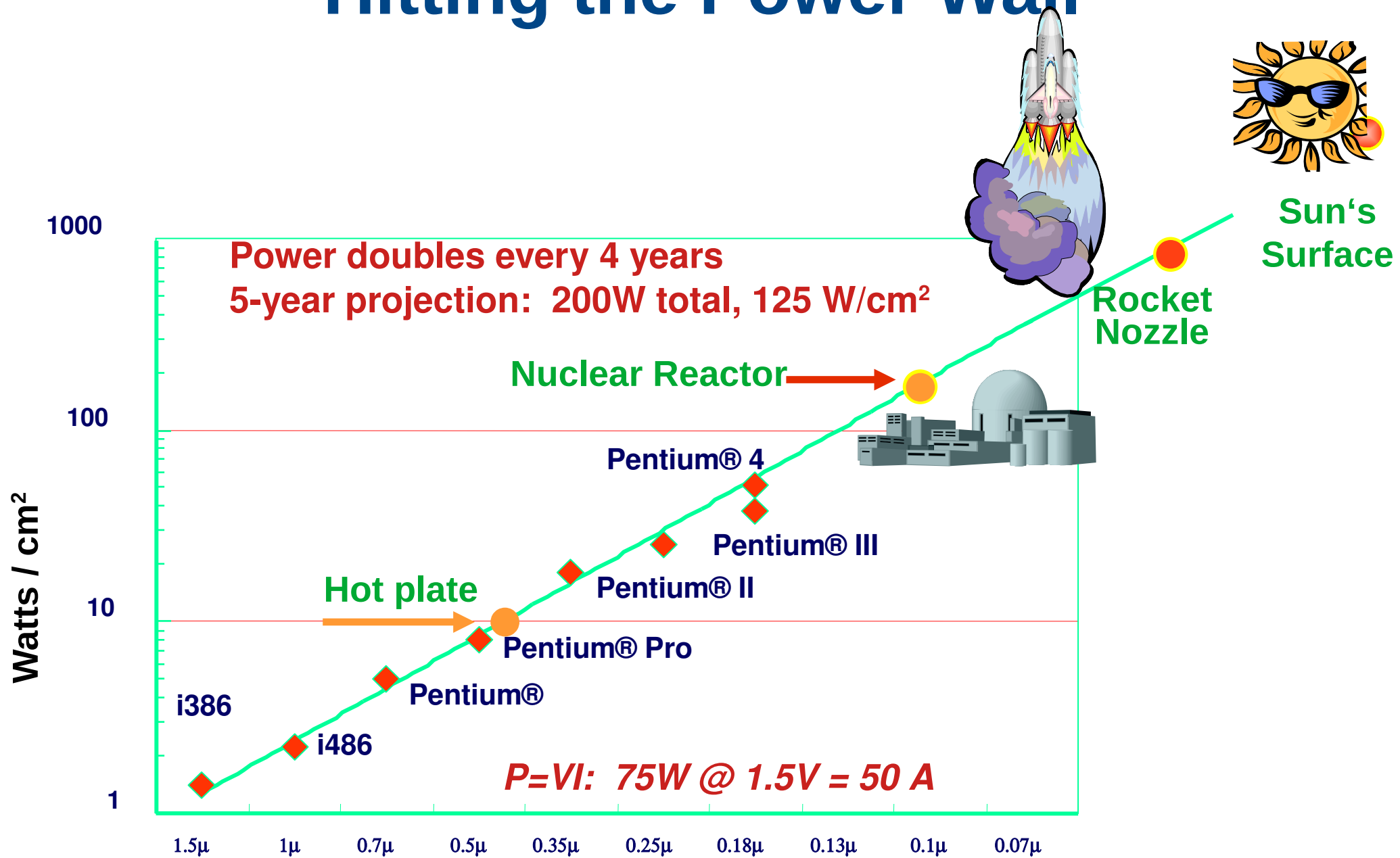# The Good Old Days for Software

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

Power doubles every 4 years
5-year projection: 200W total, 125 W/cm^2

Nuclear Reactor →

Sun's Surface

Rocket Nozzle

Watts / cm^2

1000

100

Pentium® 4

10

Pentium® III

Hot plate →

Pentium® II

Pentium® Pro

i386

Pentium®

i486

1

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

1.5µ   1µ   0.7µ   0.5µ   0.35µ   0.25µ   0.18µ   0.13µ   0.1µ   0.07µ

4

Hitting the Power Wall
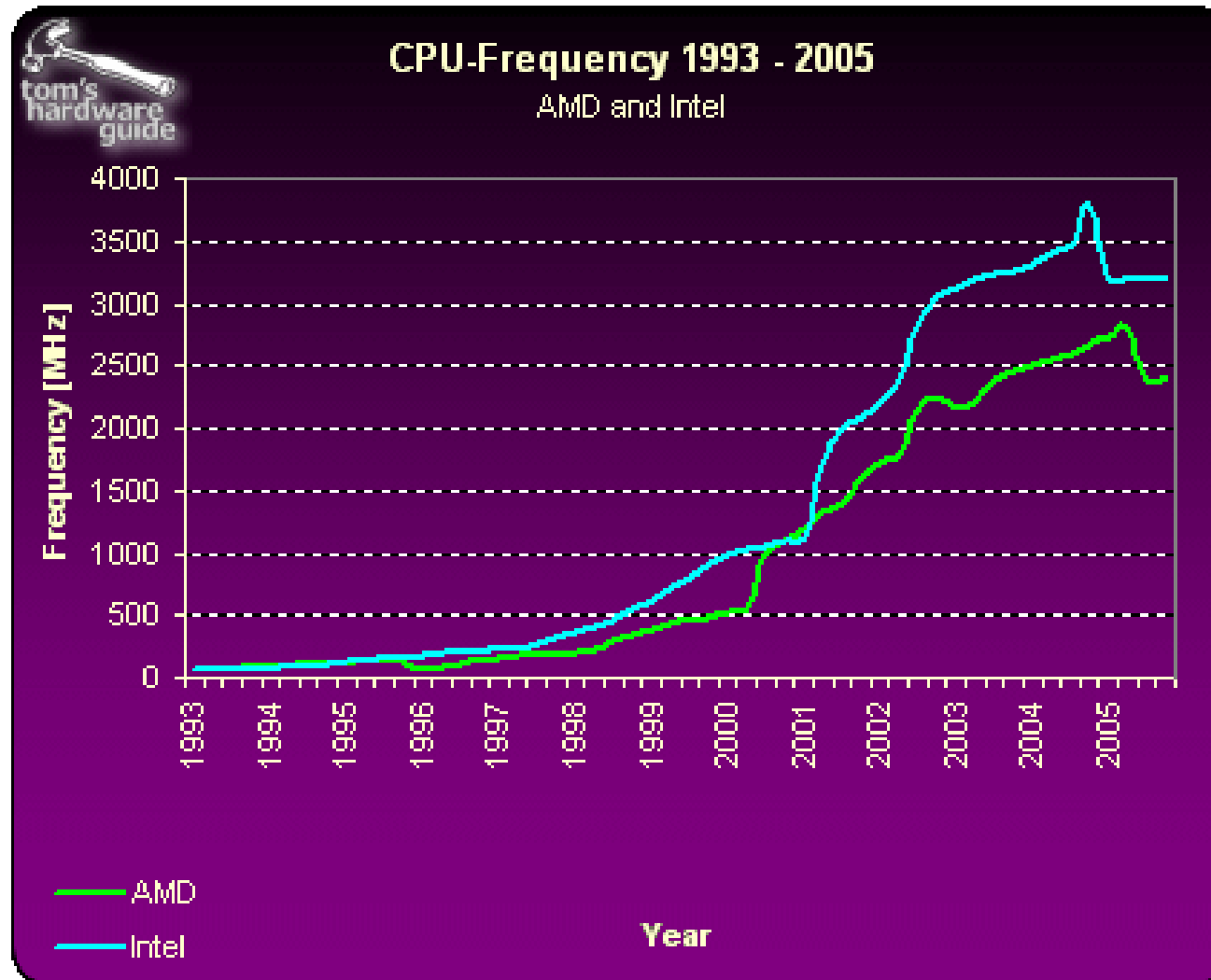
http://img.tomshardware.com/us/2005/11/21/the_mother_of_all_cpu_charts_2005/cpu_frequency.gif

2004 – Intel cancels Tejas and Jayhawk due to *heat problems due to the extreme power consumption of the core.*
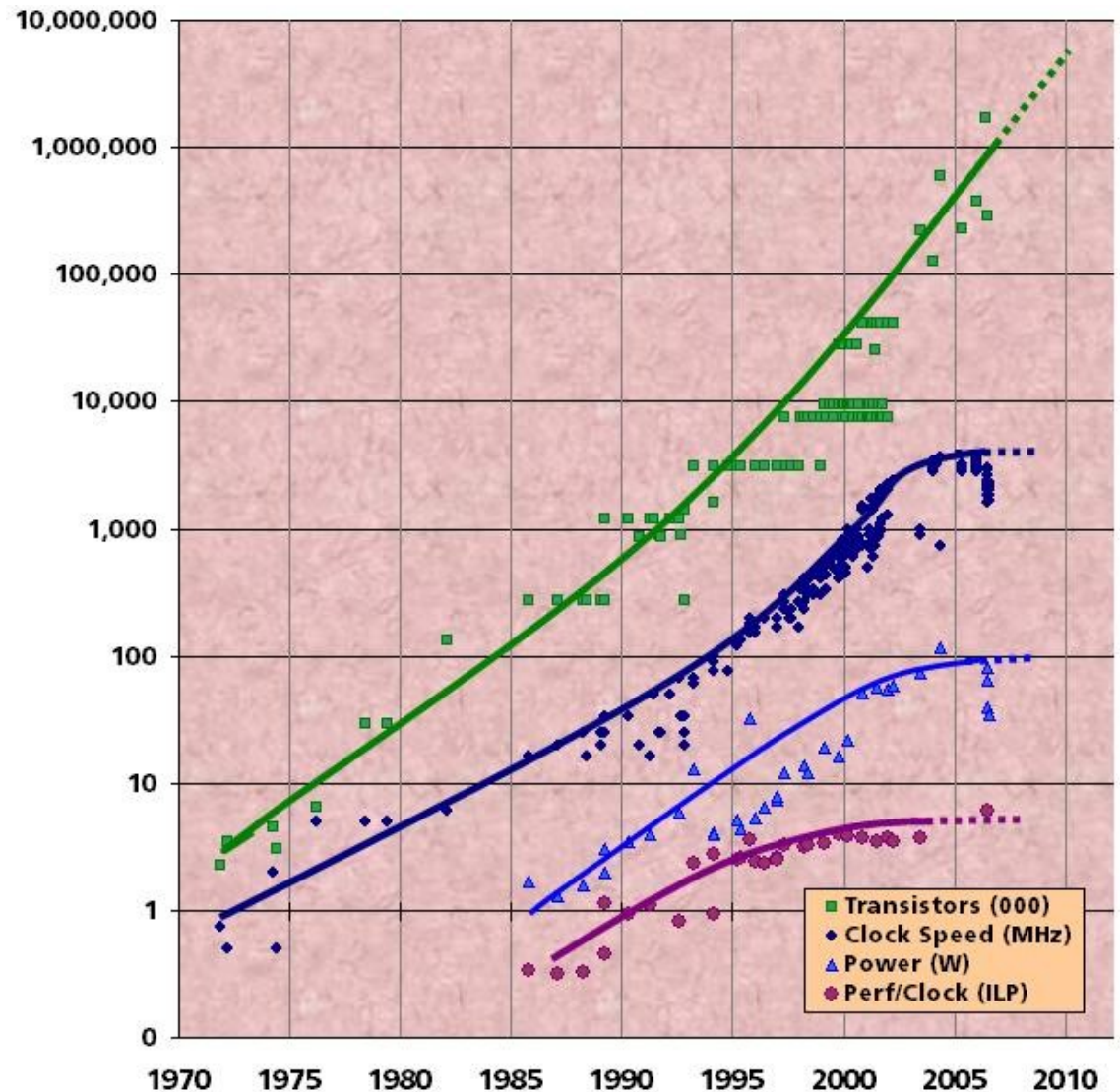
The Only Option: Use Many Cores

Chip density is increasing by ~2x every 2 years

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

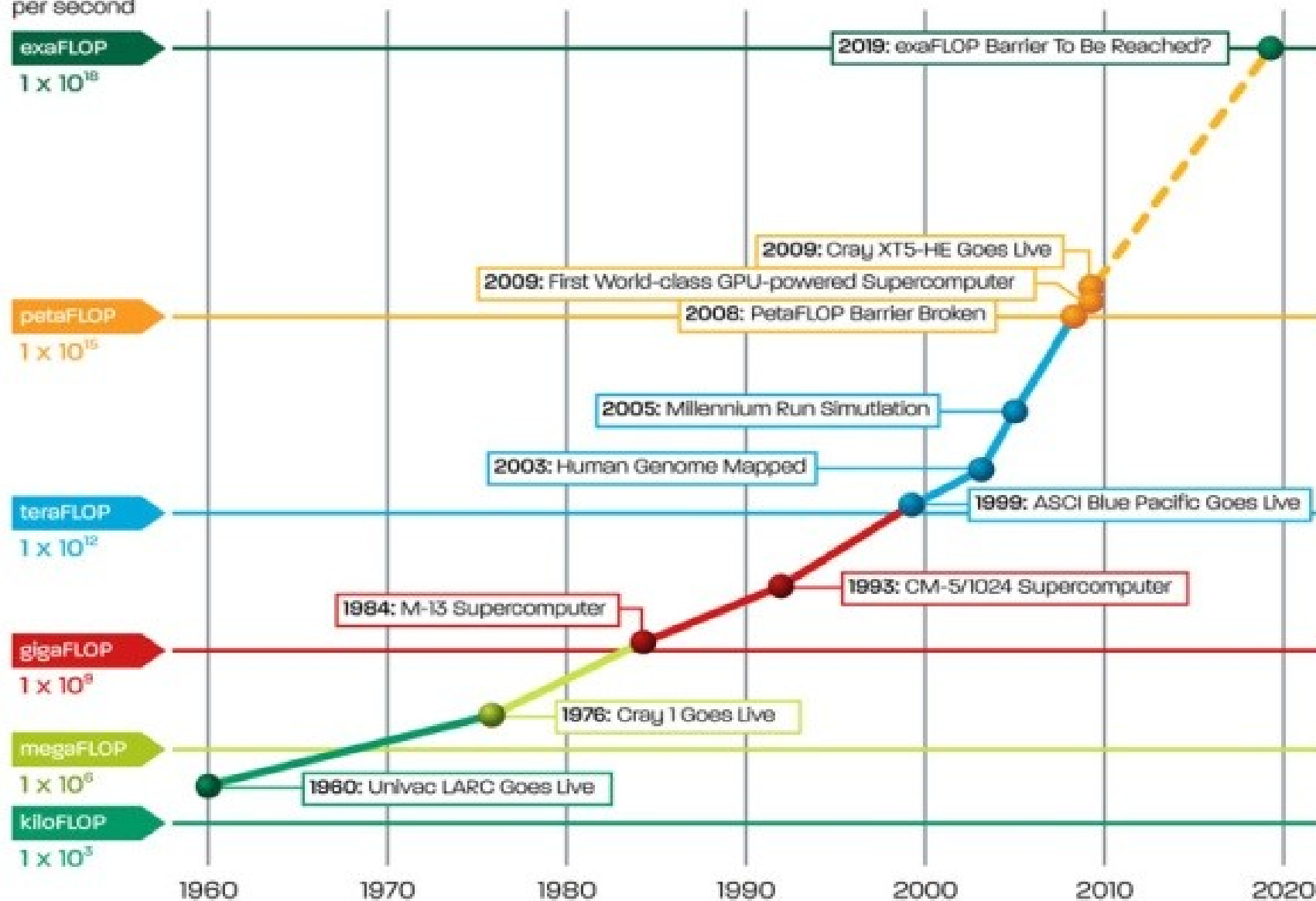There is little or no more hidden parallelism (ILP) to be found

Parallelism must be exposed to and managed by software

Legend:
- Transistors (000)
- Clock Speed (MHz)
- Power (W)
- Perf/Clock (ILP)

High-Performance Computing Milestones (1960–2019)

Floating point operations per second

exaFLOP 1 x 10^18
petaFLOP 1 x 10^15
teraFLOP 1 x 10^12
gigaFLOP 1 x 10^9
megaFLOP 1 x 10^6
kiloFLOP 1 x 10^3

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

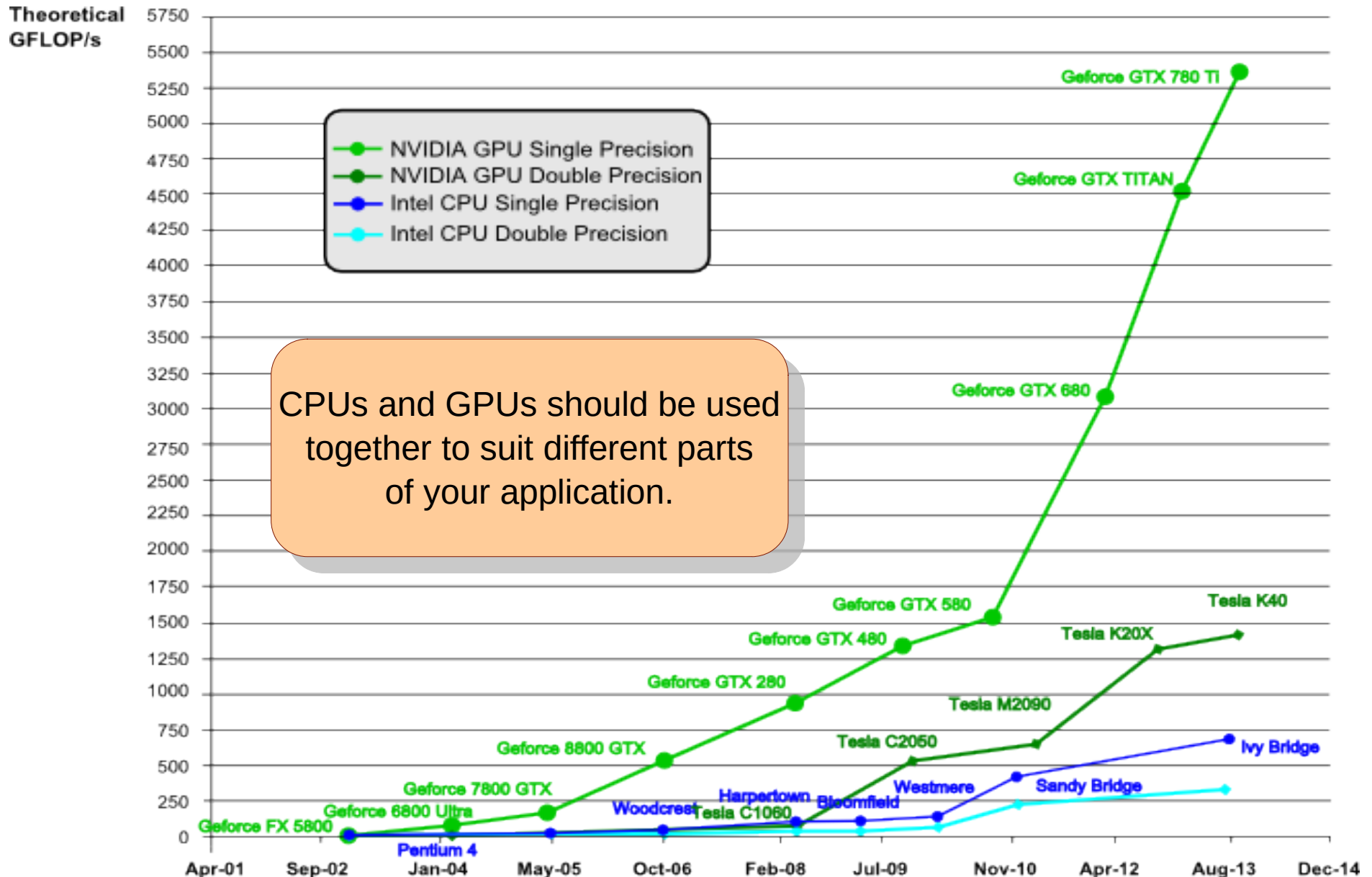1960   1970   1980   1990   2000   2010   2020

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.
  - In-class problem solving sessions
- 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 pen-paper based, open-book.
  - Assignments are to be submitted on moodle.