GPU Programming

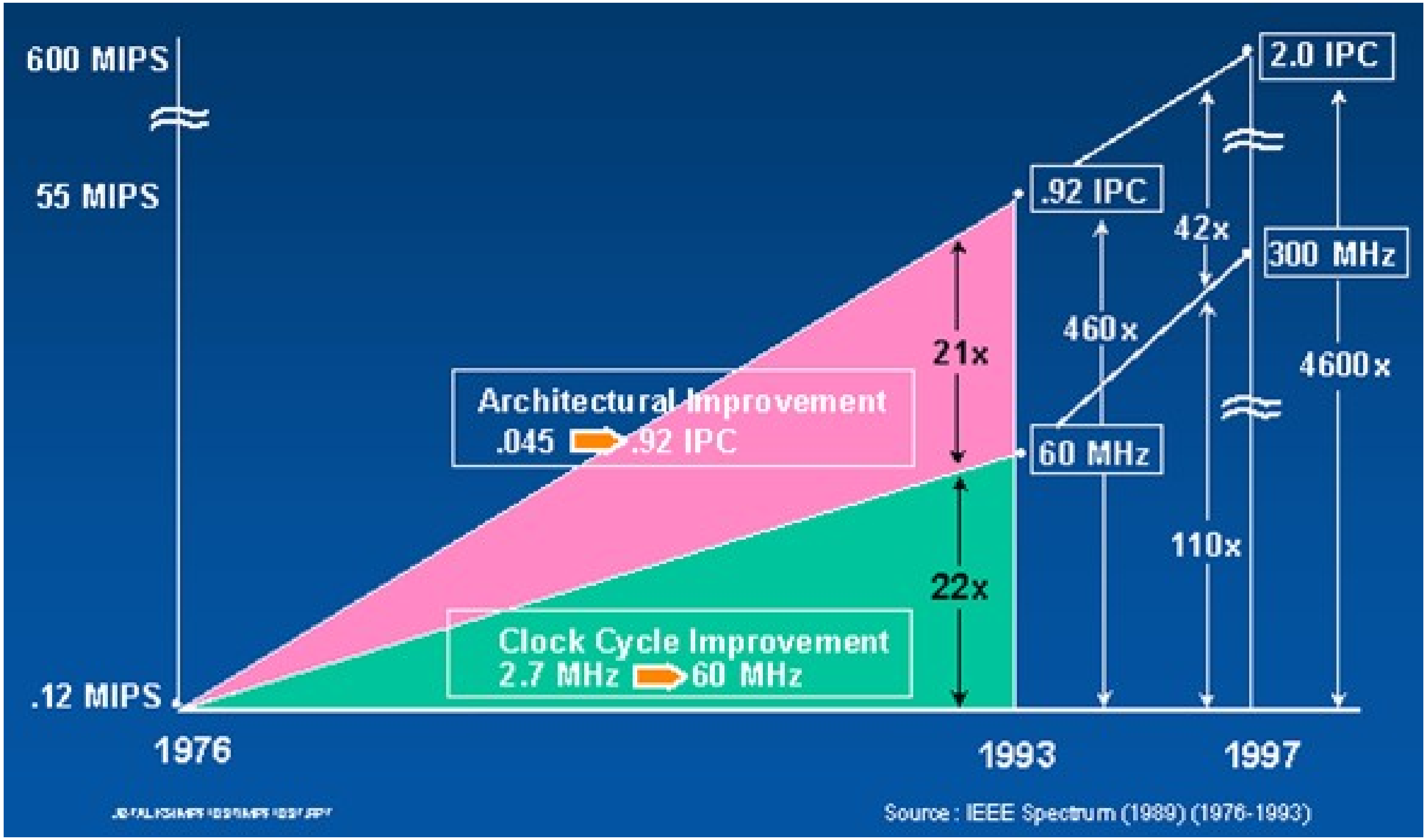
Rupesh Nasre.
rupesh@cse.iitm.ac.in

IIT Madras

January 2024

The Good Old Days for Software

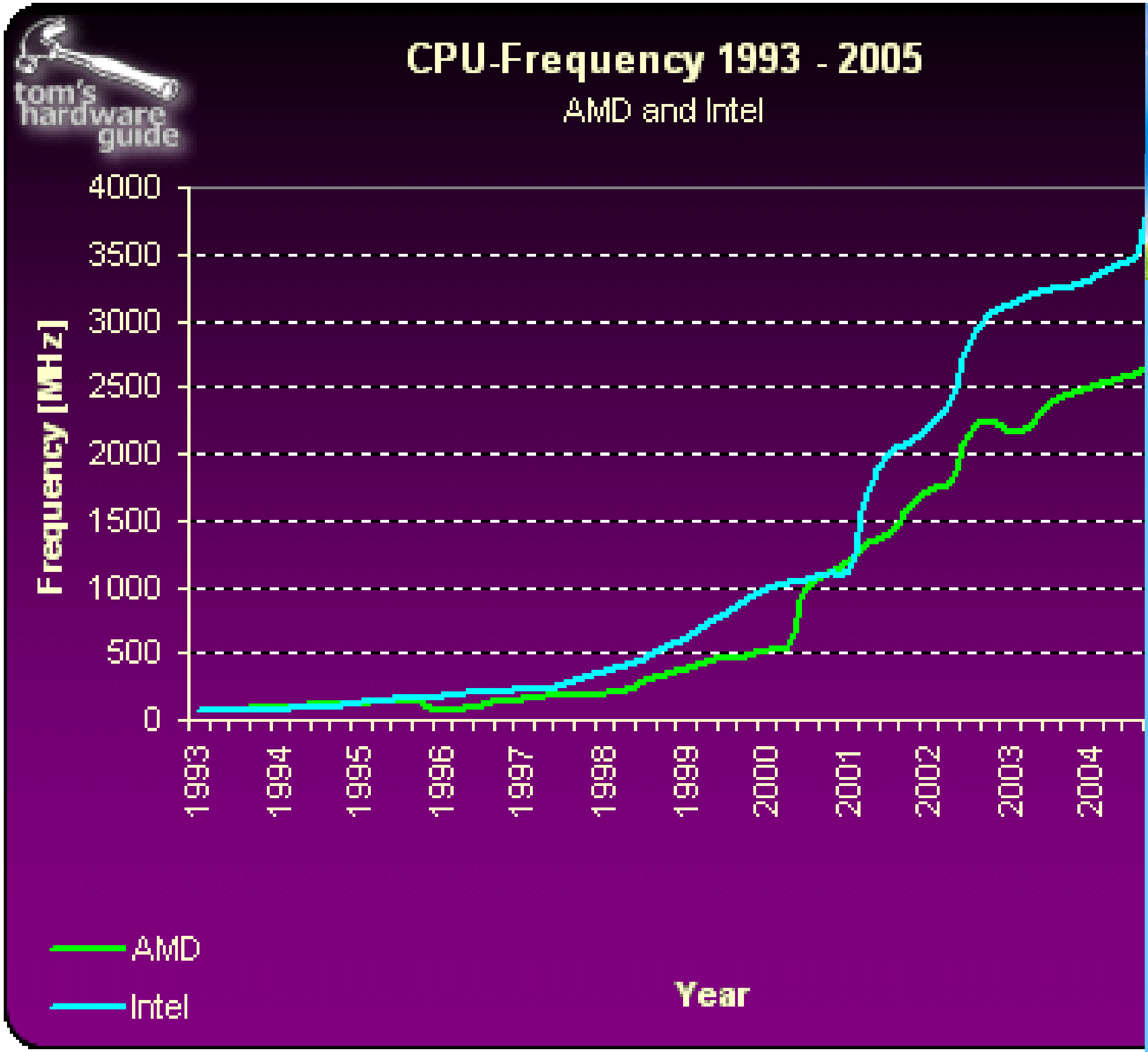
Source: J. Birnbaum



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

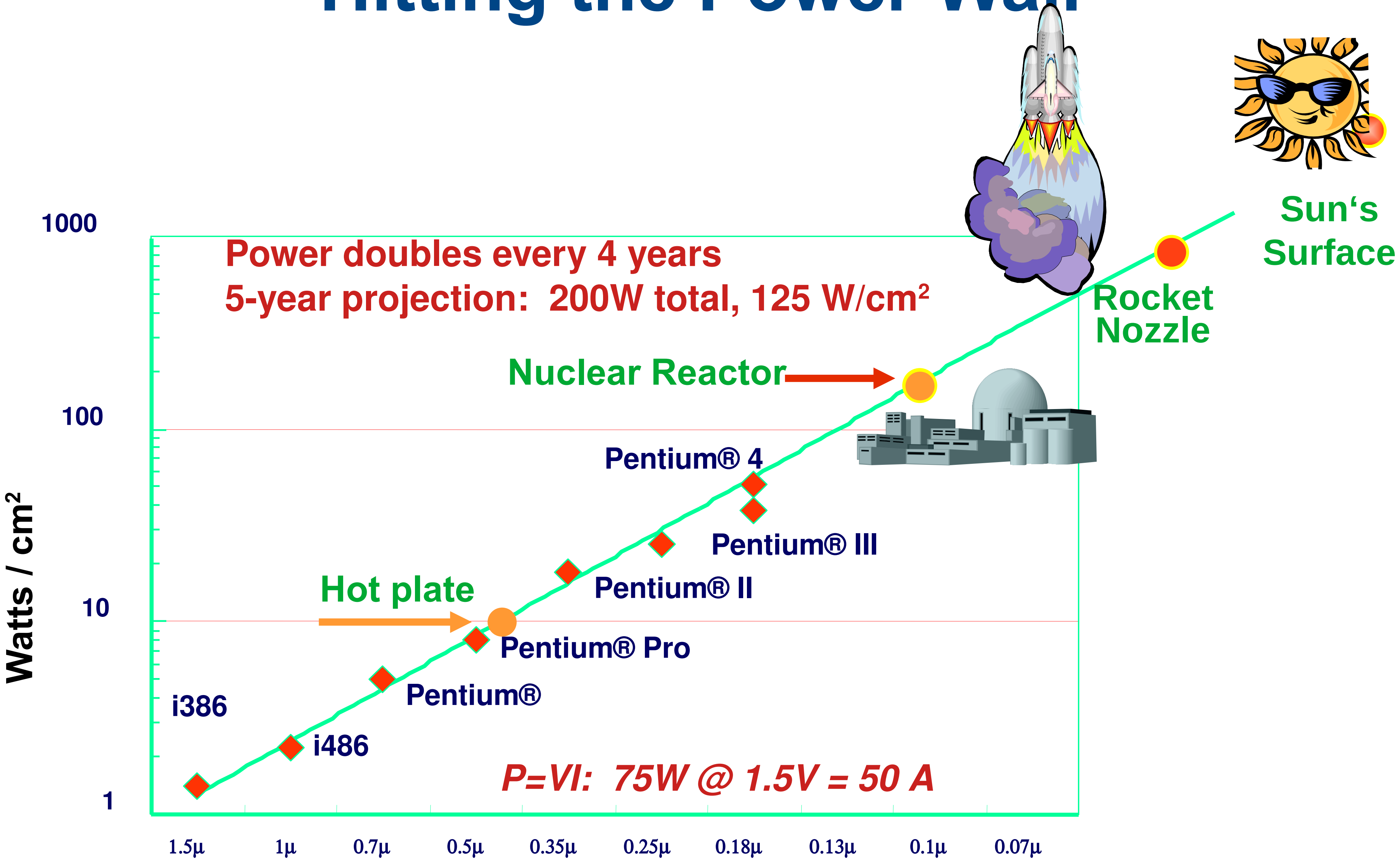
a

brighter

tomorrow

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

Hitting the Power Wall



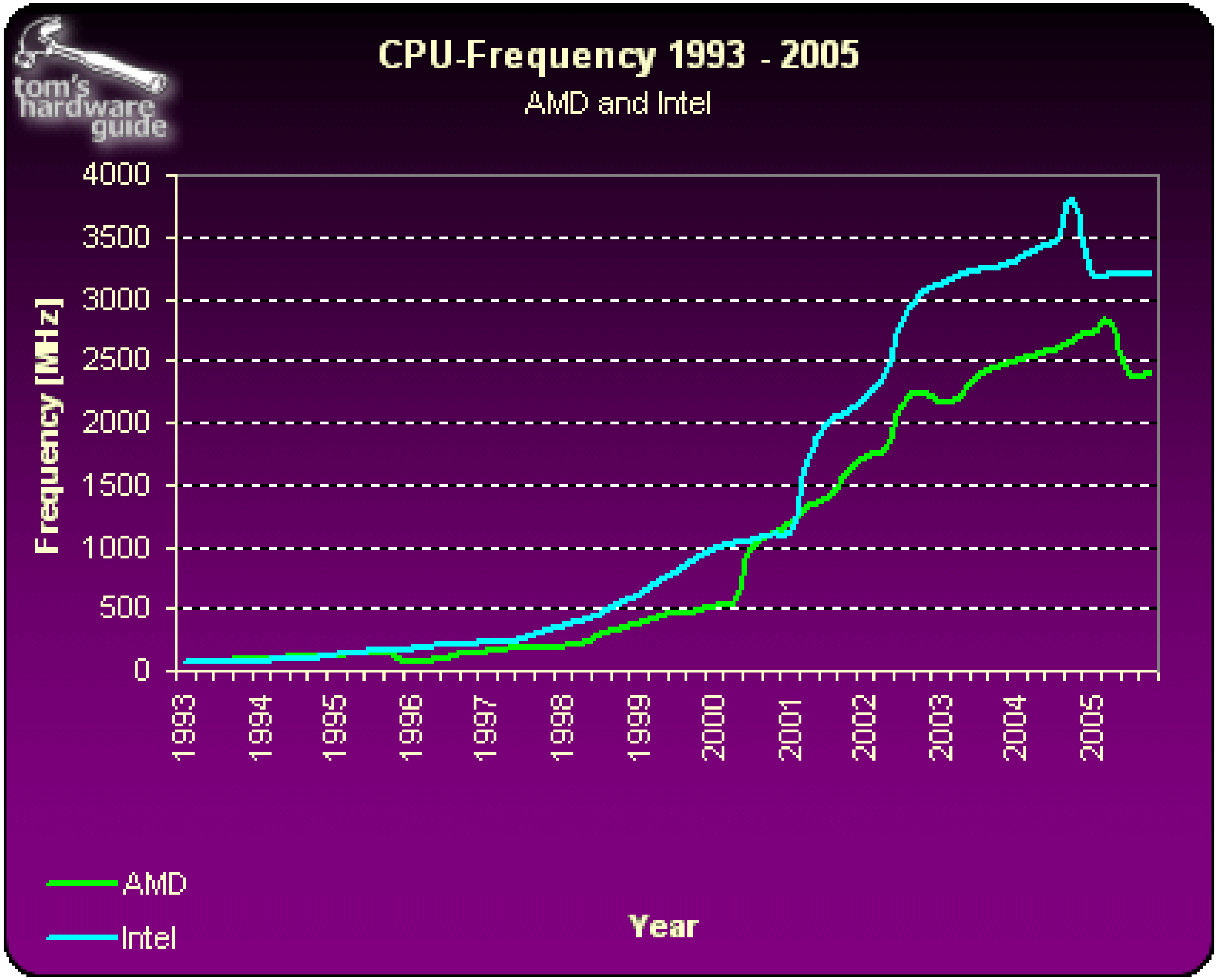
Power doubles every 4 years

5-year projection: 200W total, 125 W/cm^2

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.*

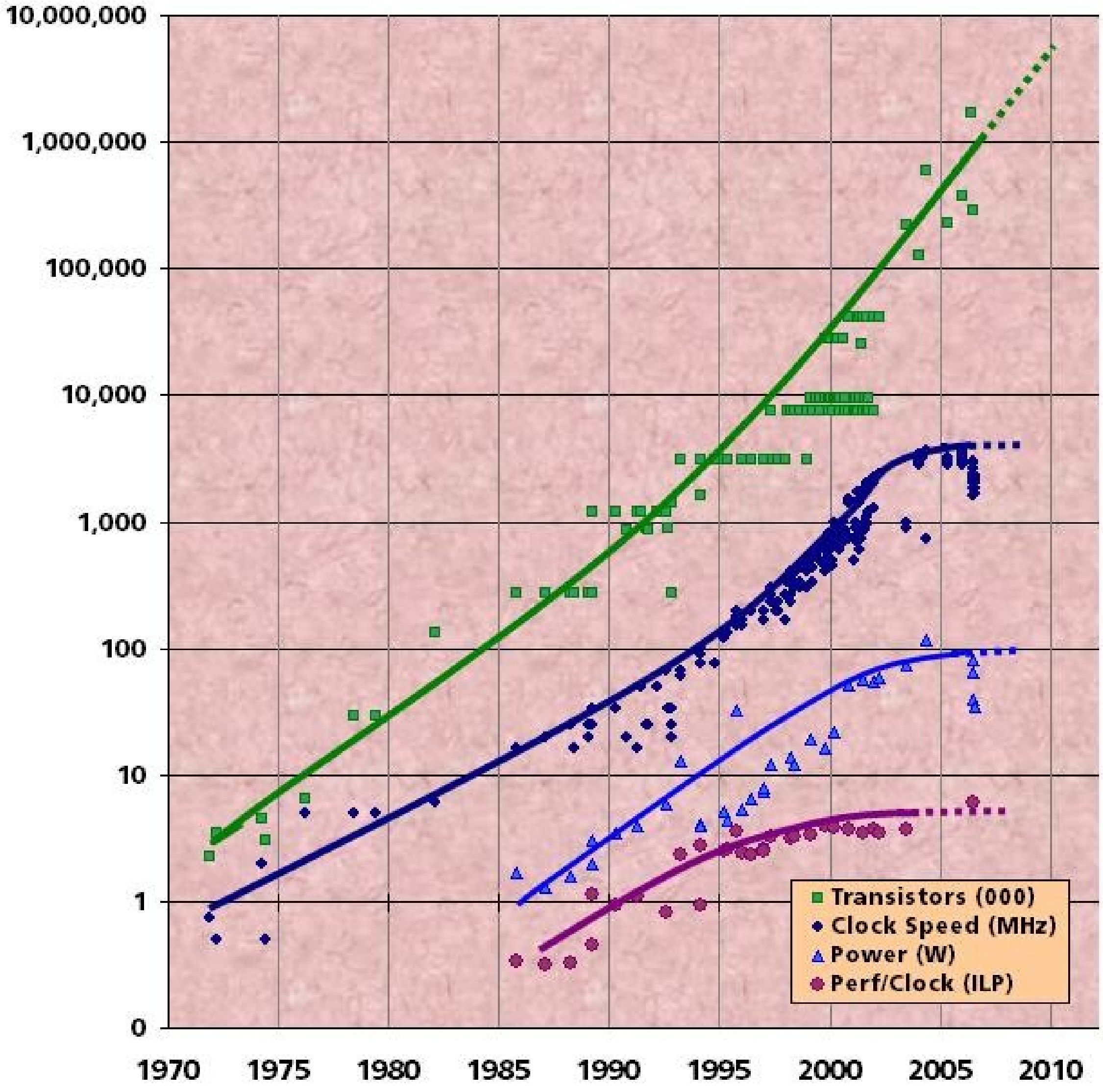
The Only Option: Use Many Cores

Chip density is increasing by ~2x every 2 years

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

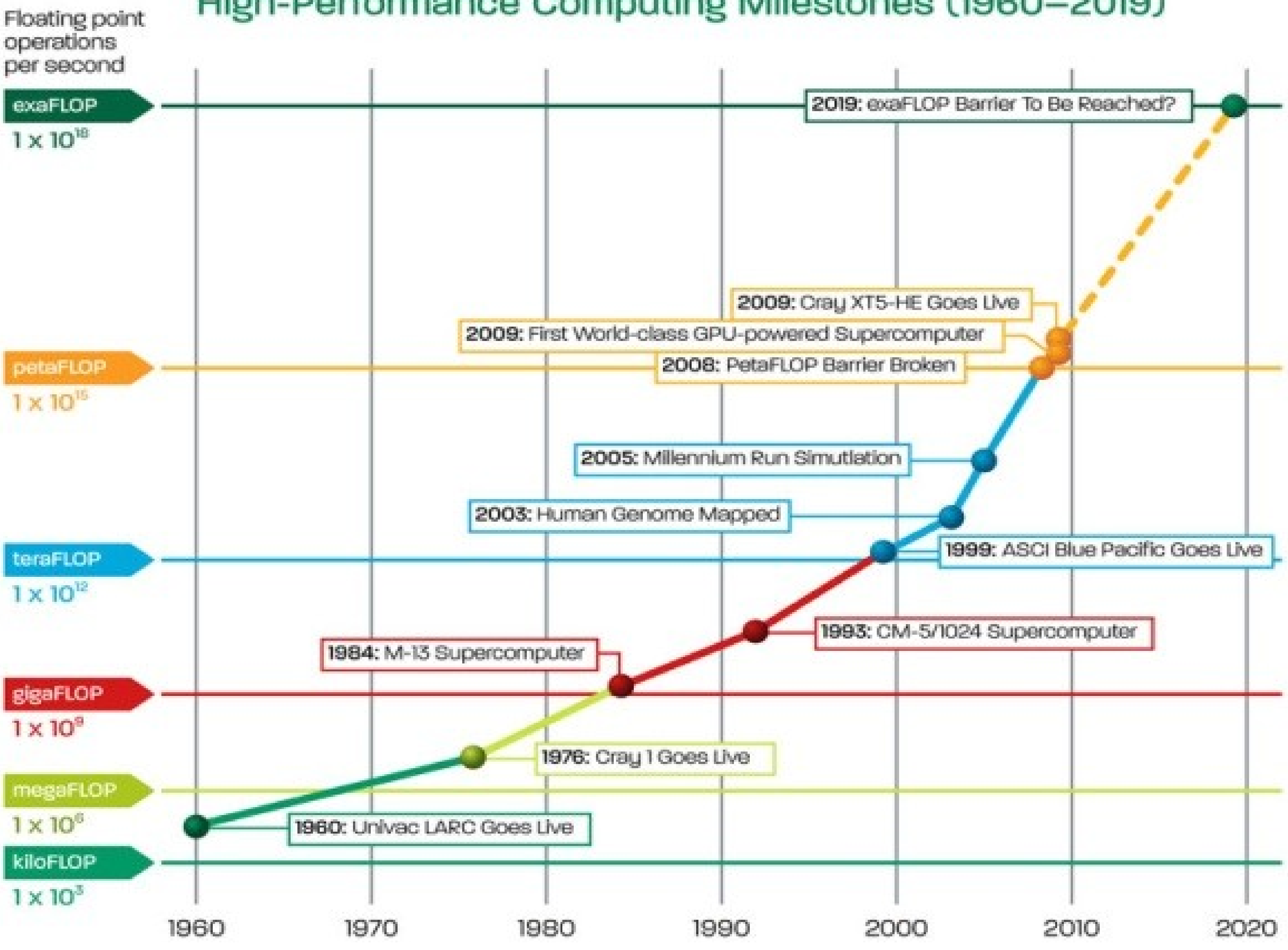
There is little or no more hidden parallelism (ILP) to be found

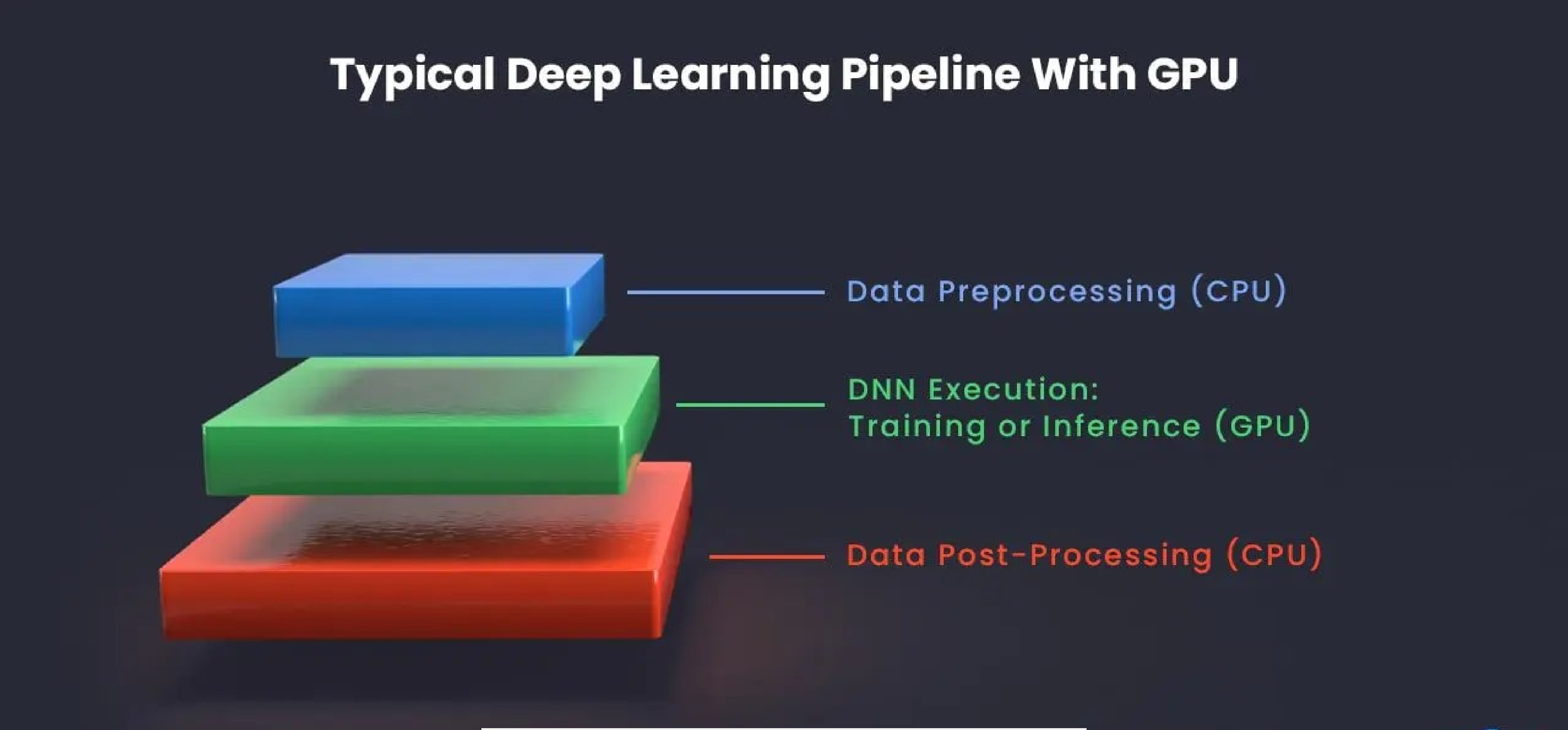
Parallelism must be exposed to and managed by software



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

High-Performance Computing Milestones (1960–2019)



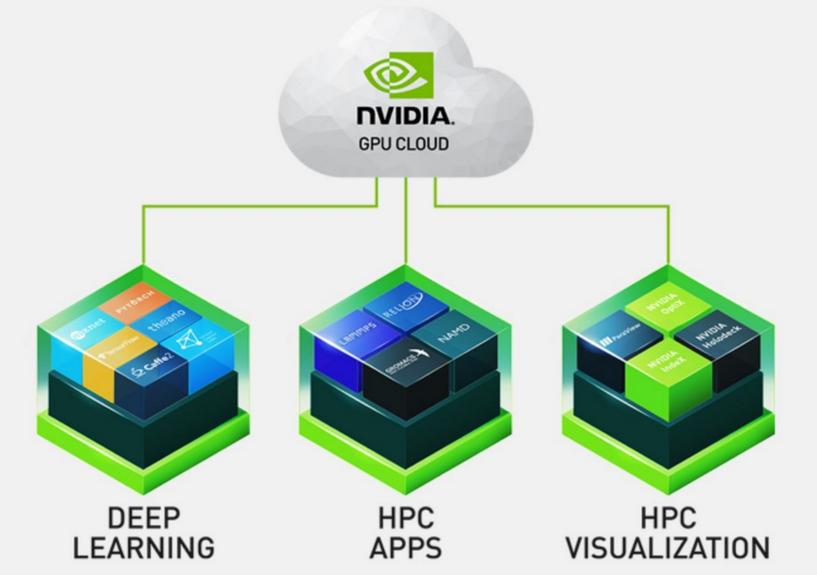


Typical Deep Learning Pipeline With GPU

Data Preprocessing (CPU)

DNN Execution:
Training or Inference (GPU)

Data Post-Processing (CPU)

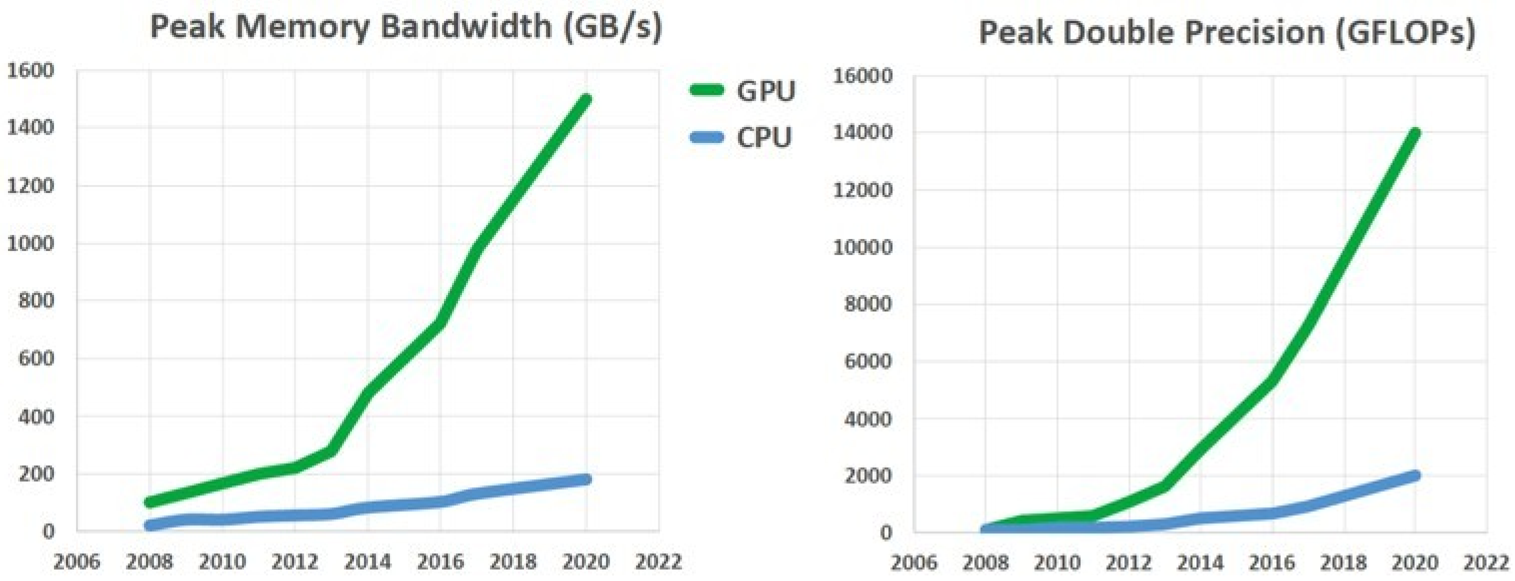


Creator: Patrick Kennedy

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: The Next Platform

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 or kaggle (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.

Reasons for Dropping the Course

- The instructor is strict about attendance. Does not shy giving W grades.
- The course-load is high compared to many other courses in the insti.
- There will be plagiarism checks on the submitted codes. Your grades will be reduced by two grades (S to B, D to U) for copying or for sharing your code or referred to DisCo.
- The assignment deadlines are not extended even when your real brother is getting married (except for specific certified health issues or if you represent IITM in an approved competition).
- The instructor does not cancel the 8 o'clock class.