

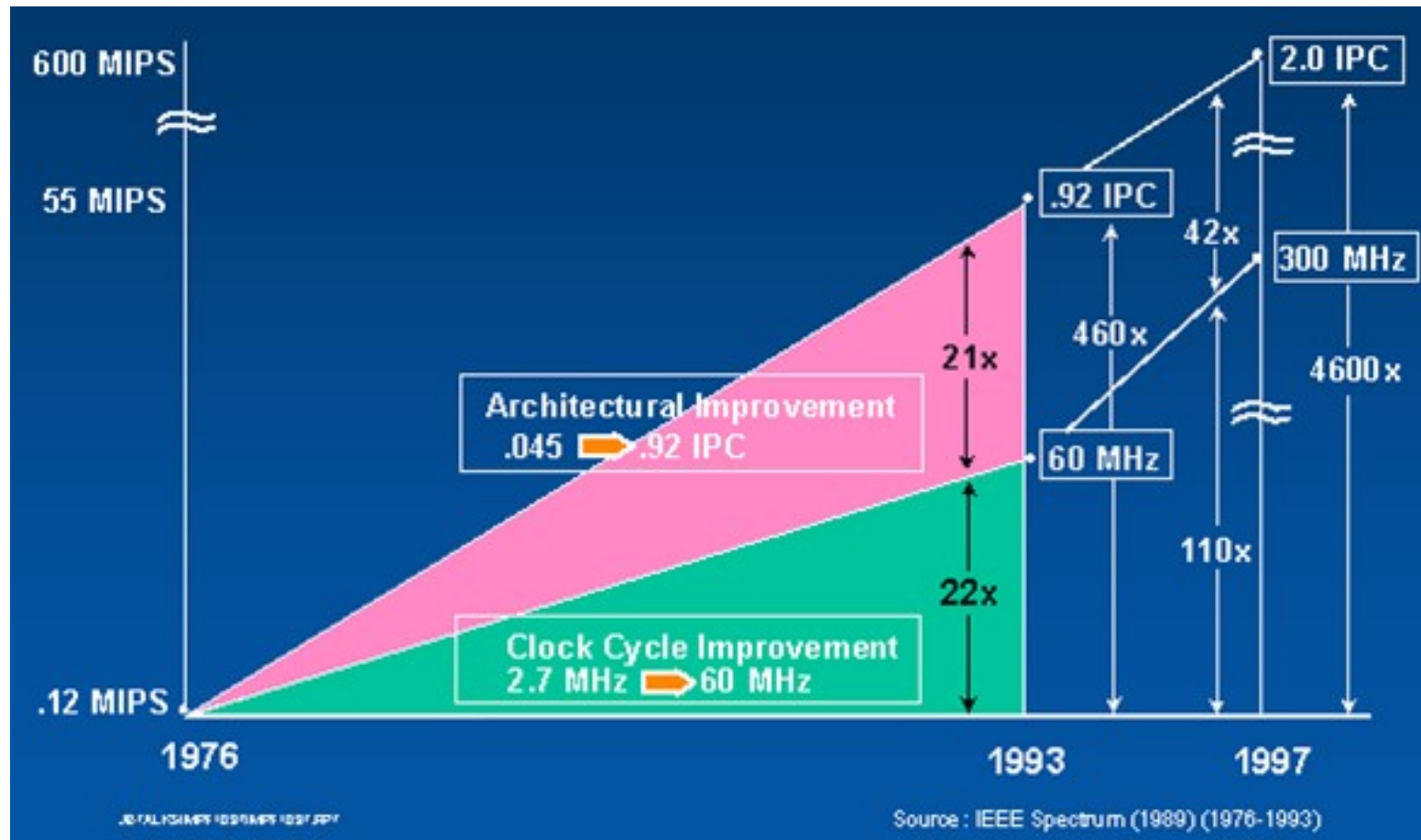
# GPU Programming

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

IIT Madras  
January 2024

# The Good Old Days for Software

Source: J. Birnbaum

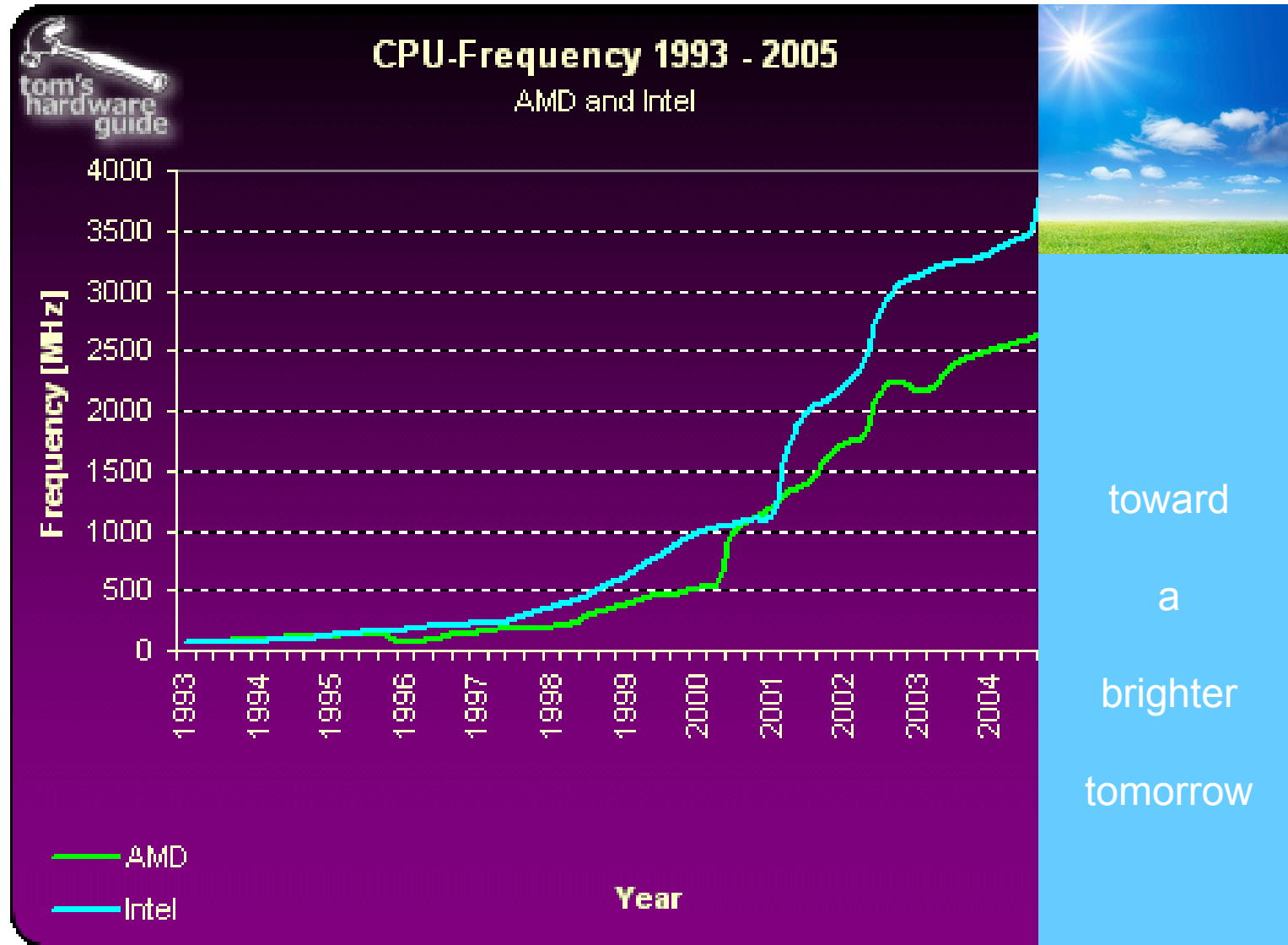


- Single-processor performance experienced dramatic improvements from **clock**, and **architectural improvement** (Pipelining, Instruction-Level-Parallelism).
- Applications experienced **automatic** performance improvement.

As the years pass by, the hardware improvements <sup>2</sup> made the programs run faster without the need of coding optimizations. Then we reached

# Hitting the Power Wall

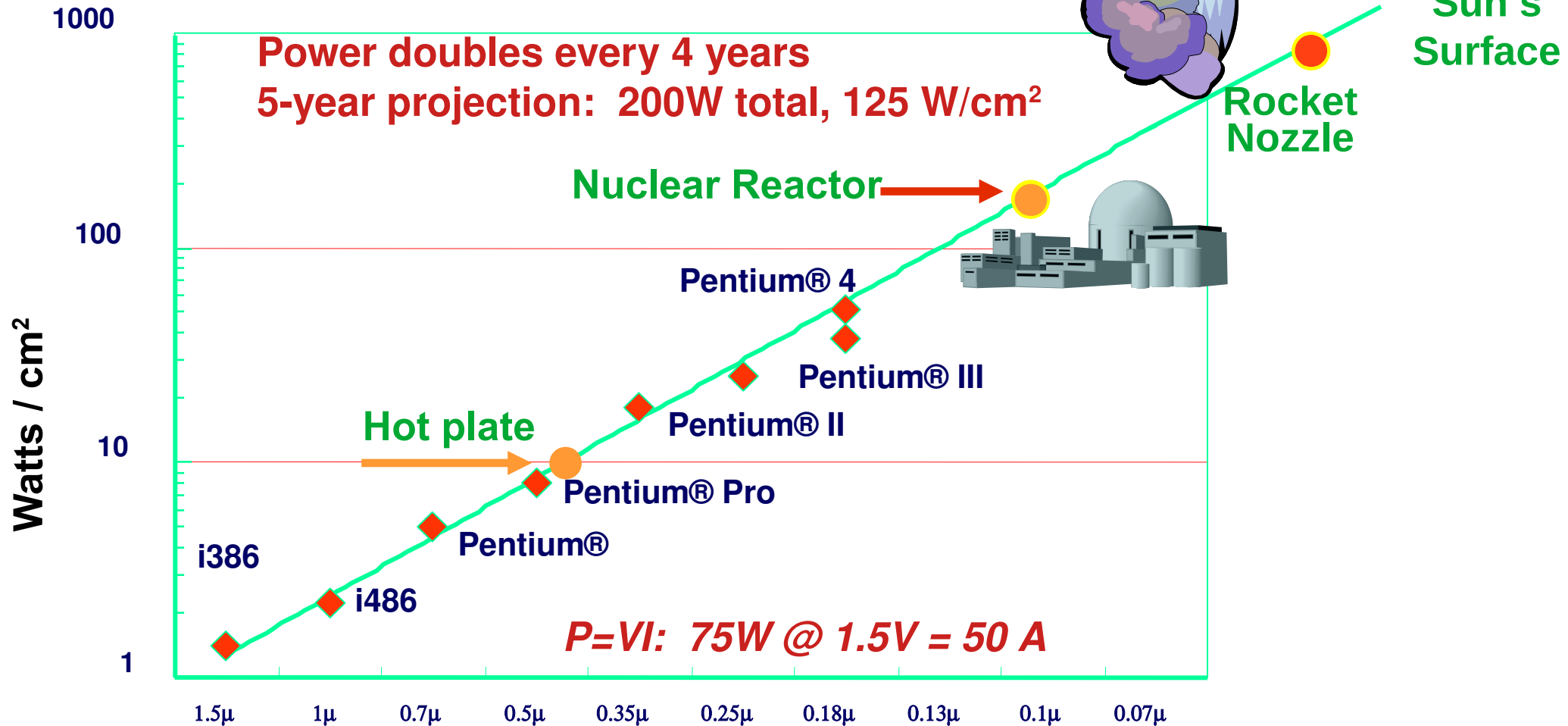
a point where we  
can no longer increase



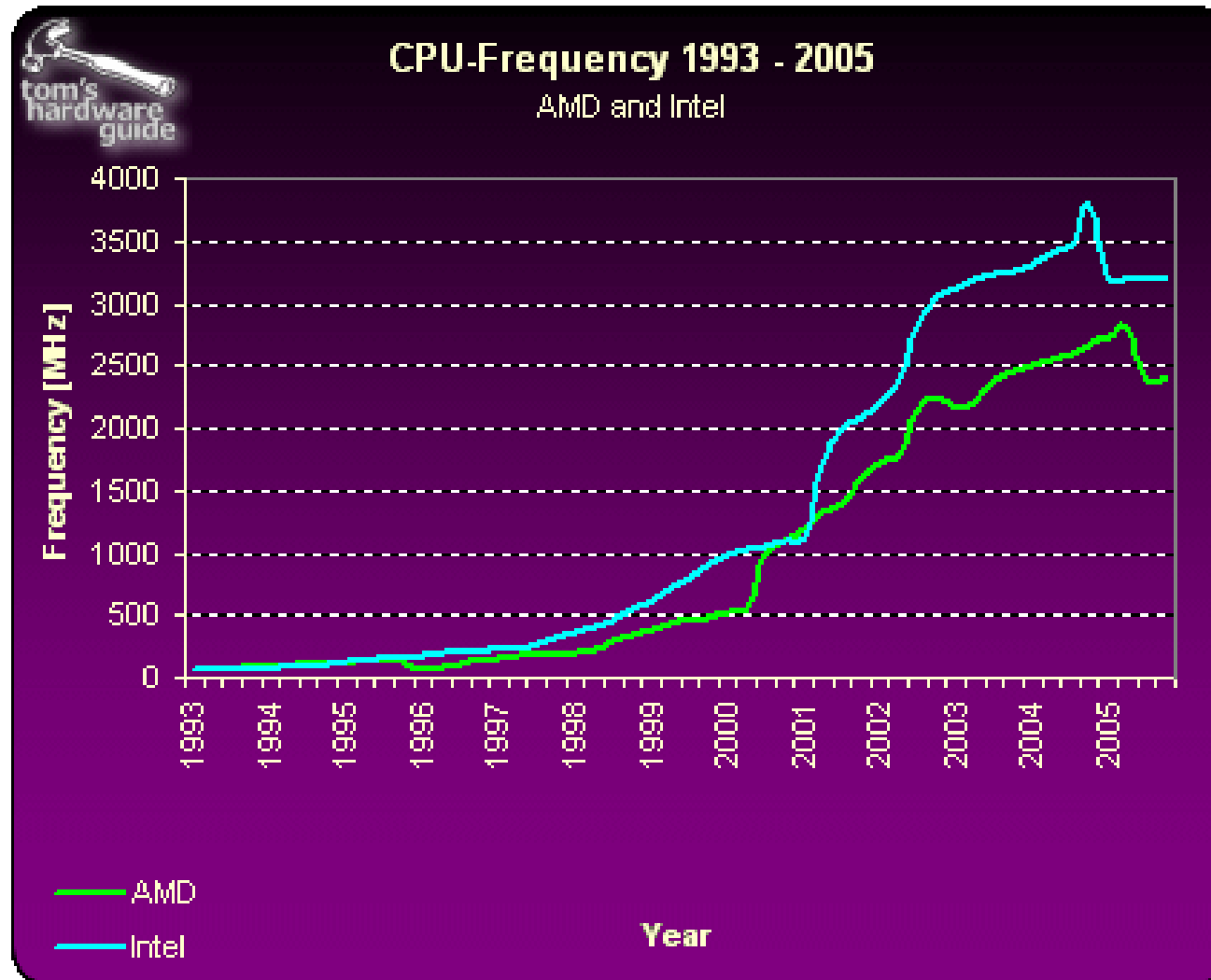
the  
hardware  
(cpu freq)  
now we  
are left  
with  
coding  
optimizations  
rather than  
improving  
hardware.

[http://img.tomshardware.com/us/2005/11/21/the\\_mother\\_of\\_all\\_cpu\\_charts\\_2005/cpu\\_frequency.gif](http://img.tomshardware.com/us/2005/11/21/the_mother_of_all_cpu_charts_2005/cpu_frequency.gif)

# Hitting the Power Wall



# Hitting the Power Wall



[http://img.tomshardware.com/us/2005/11/21/the\\_mother\\_of\\_all\\_cpu\\_charts\\_2005/cpu\\_frequency.gif](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.***

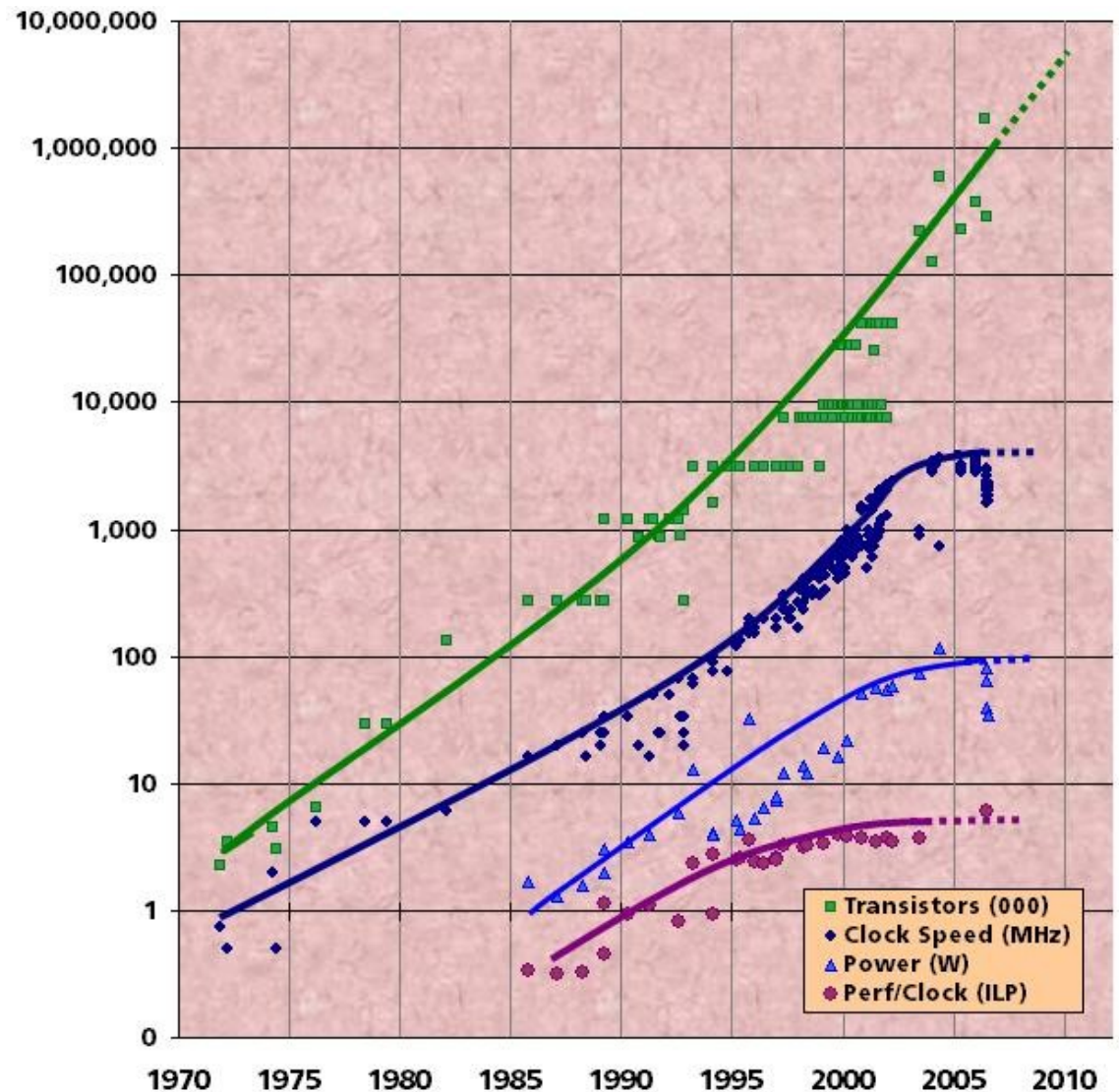
# The Only Option: Use Many Cores

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



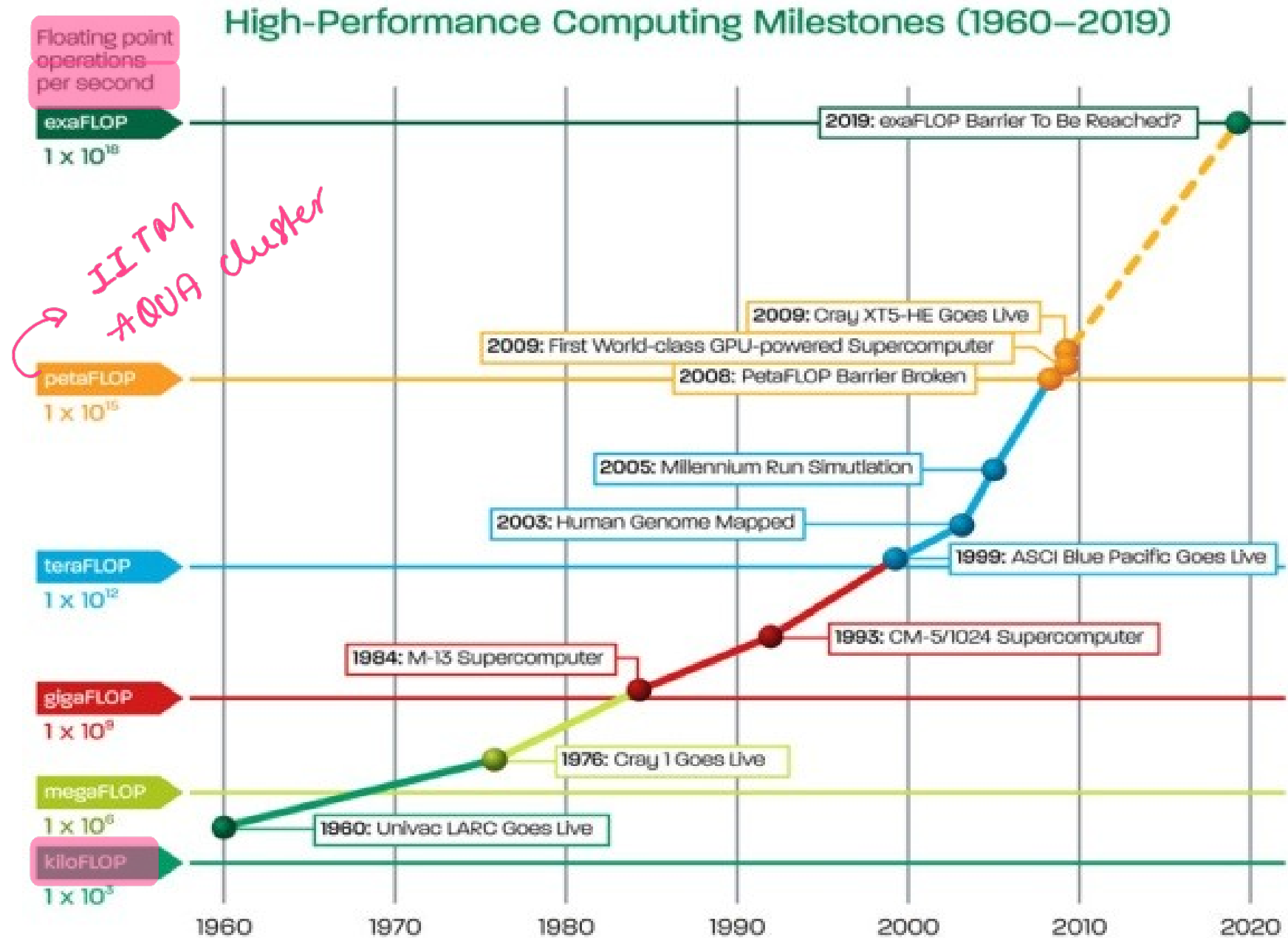
Source: Intel, Microsoft (Sutter) and Stanford (Olukotun, Hammond)

GPU's are also called accelerators.

→ Nvidia has 5120 WDA cores

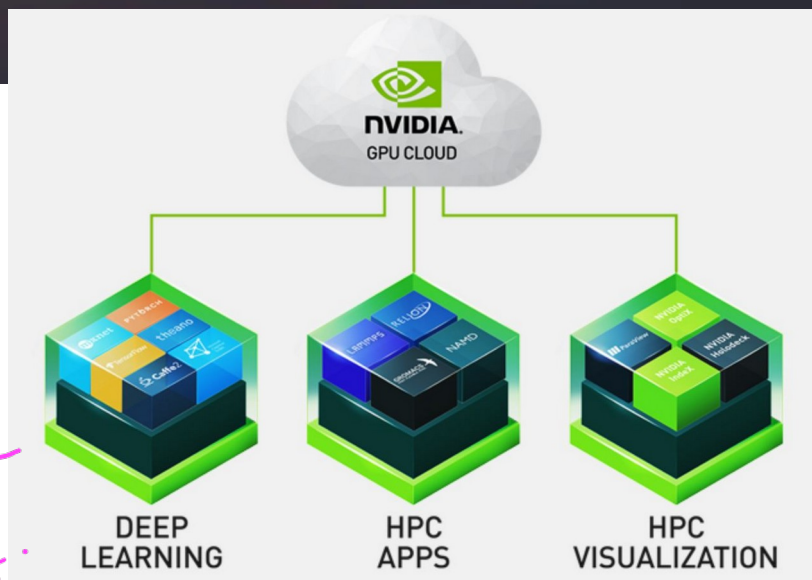
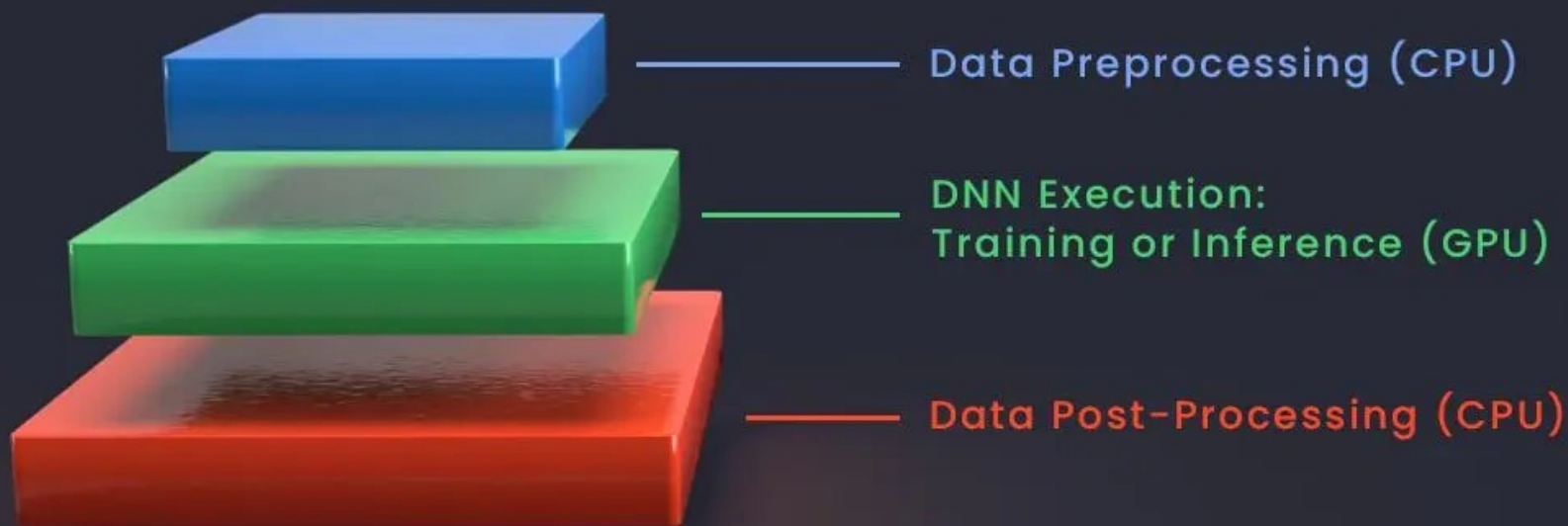
GPU's are also called accelerators.  
GPU's will have thousands of cores.

→ intel CPU has 48 cores.  
AMD





# Typical Deep Learning Pipeline With GPU



mobidev

*Super Computers are distributed systems.*

*are not the unit.*

*in it*



# Parallel Platforms

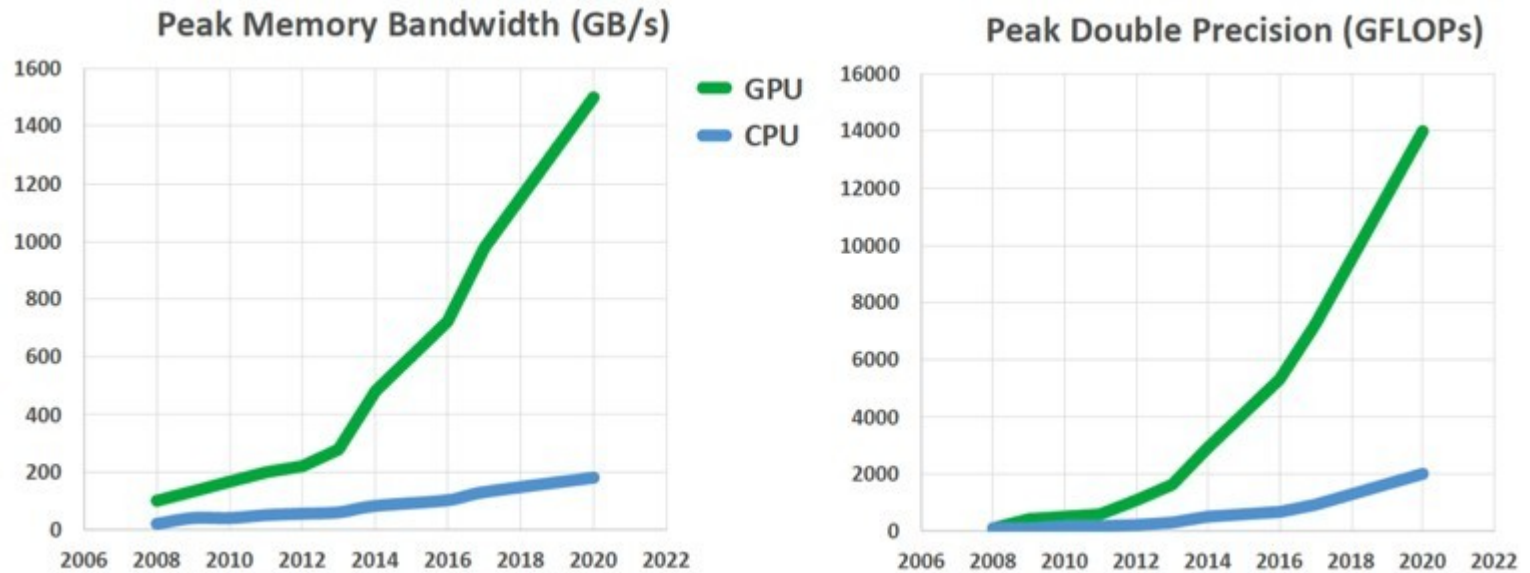
*note: GPUs  
Sole processing  
it acts as a co-processor.  
it works alongside the main  
processor unit  
i.e. CPU & intel + nvidia*

- Shared memory systems (multi-core) *→ our laptops*
- Distributed systems (cluster) *→ each node will have its own memory. if a node needs to access another nodes data, it makes a request*
- Graphics Processing Units (many-core)
- Field-Programmable Gate Arrays (configurable after manufacturing)
- Application-Specific Integrated Circuits
- Heterogeneous Systems

*A system with GPU and CPU is in a sense acts like a distributed system becoz CPU & GPU has its own memory*

and also a single GPU processing resembles shared memory systems.

# GPU-CPU Performance Comparison



CPU and GPU should be used together to suit different parts of your application.

# 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.