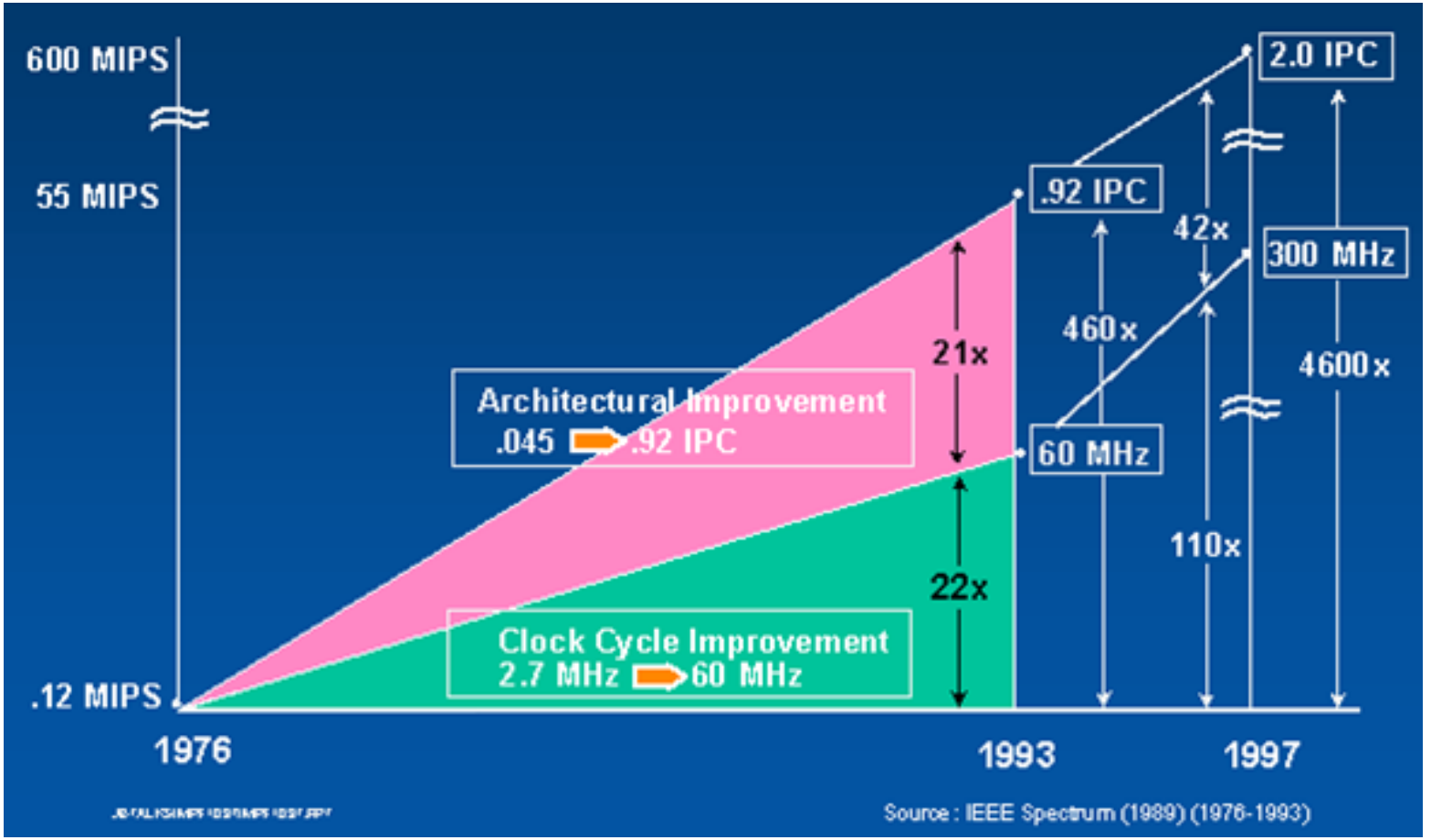
GPU Programming

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

IIT Madras
January 2025

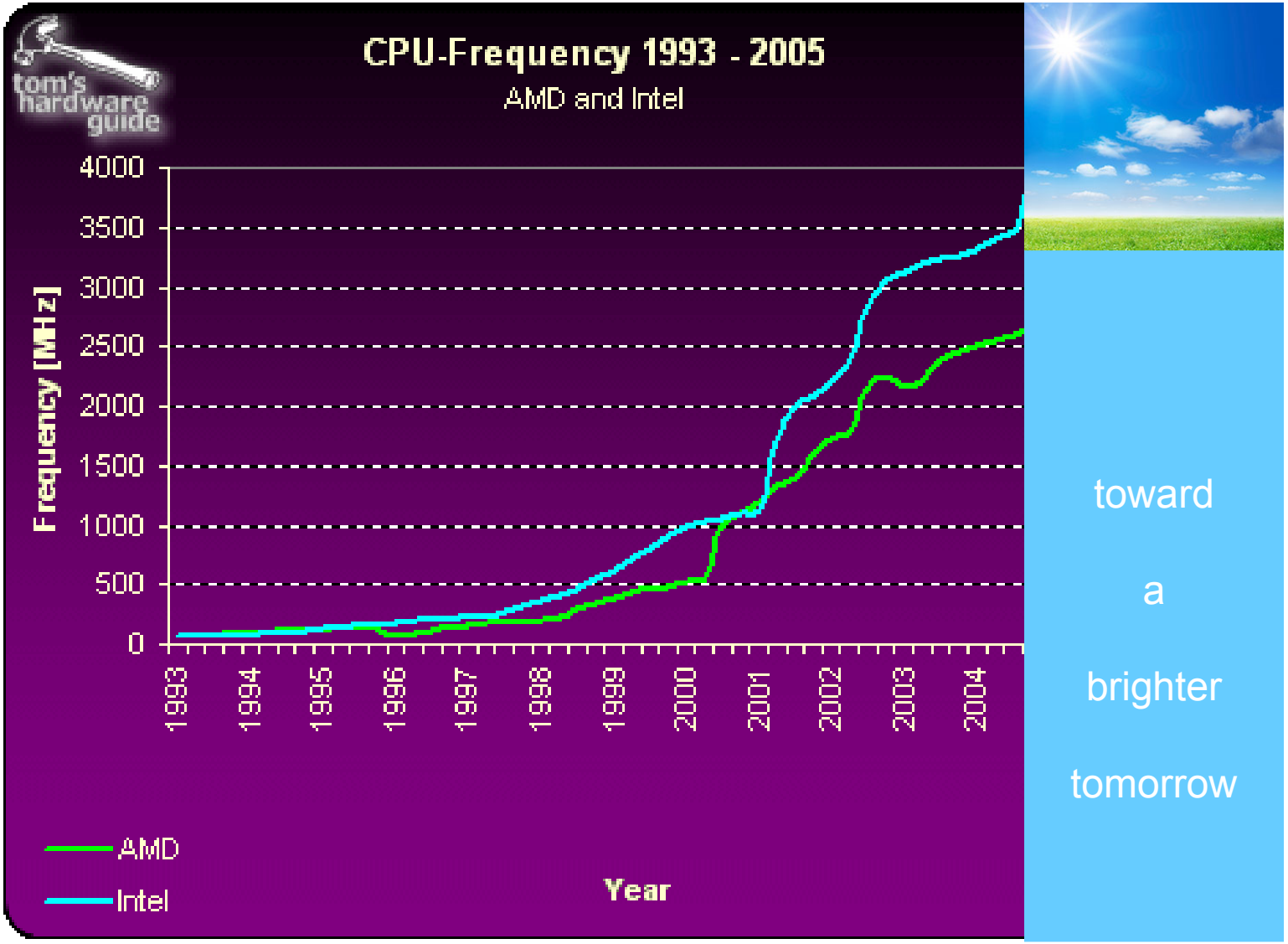
The Good Old Days for Software

Source: J. Birnbaum



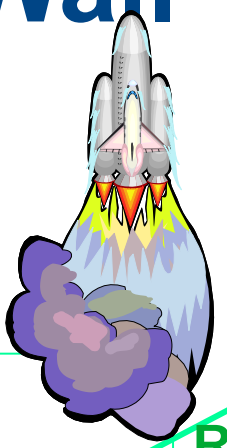
- Single-processor performance experienced dramatic improvements from **clock**, and **architectural** improvement (Pipelining, Instruction-Level-Parallelism).
- Applications experienced **automatic** performance improvement.

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



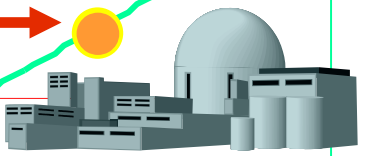
Sun's Surface

Rocket Nozzle

Power doubles every 4 years

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

Nuclear Reactor



Hot plate

Watts / cm²

1000

100

10

1

Pentium® 4

Pentium® III

Pentium® II

Pentium® Pro

Pentium®

i486

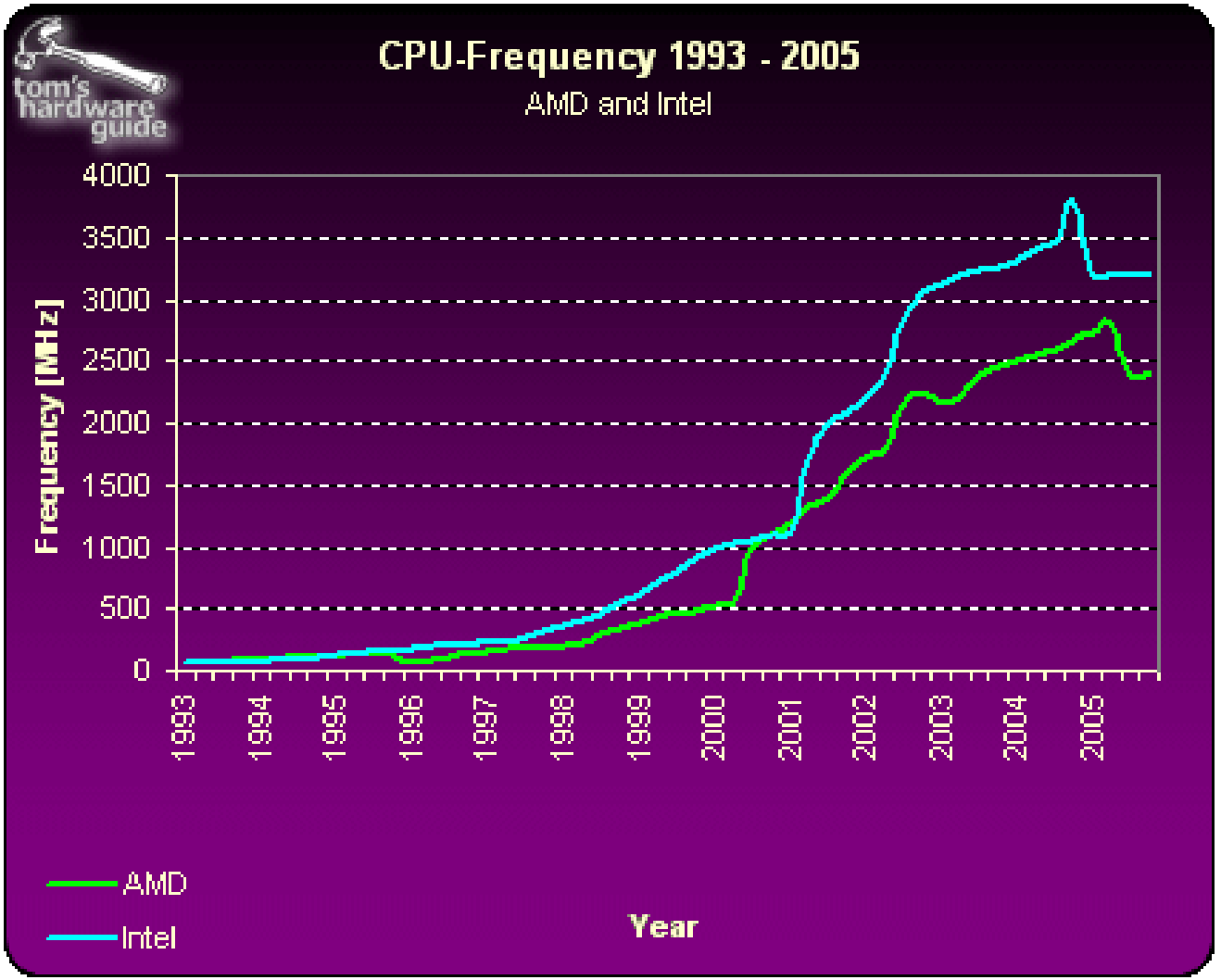
i386

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

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

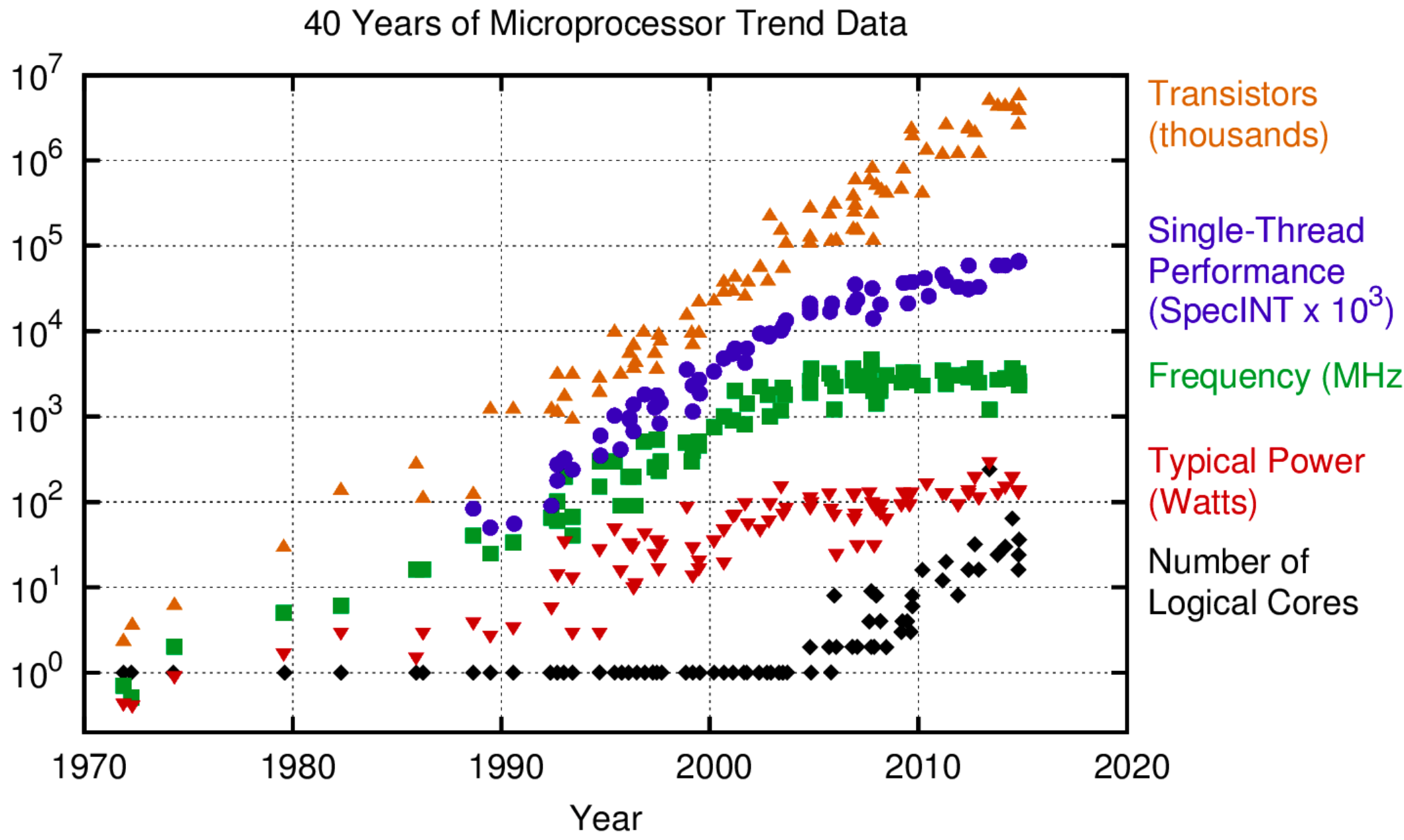
Hitting the Power Wall



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



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

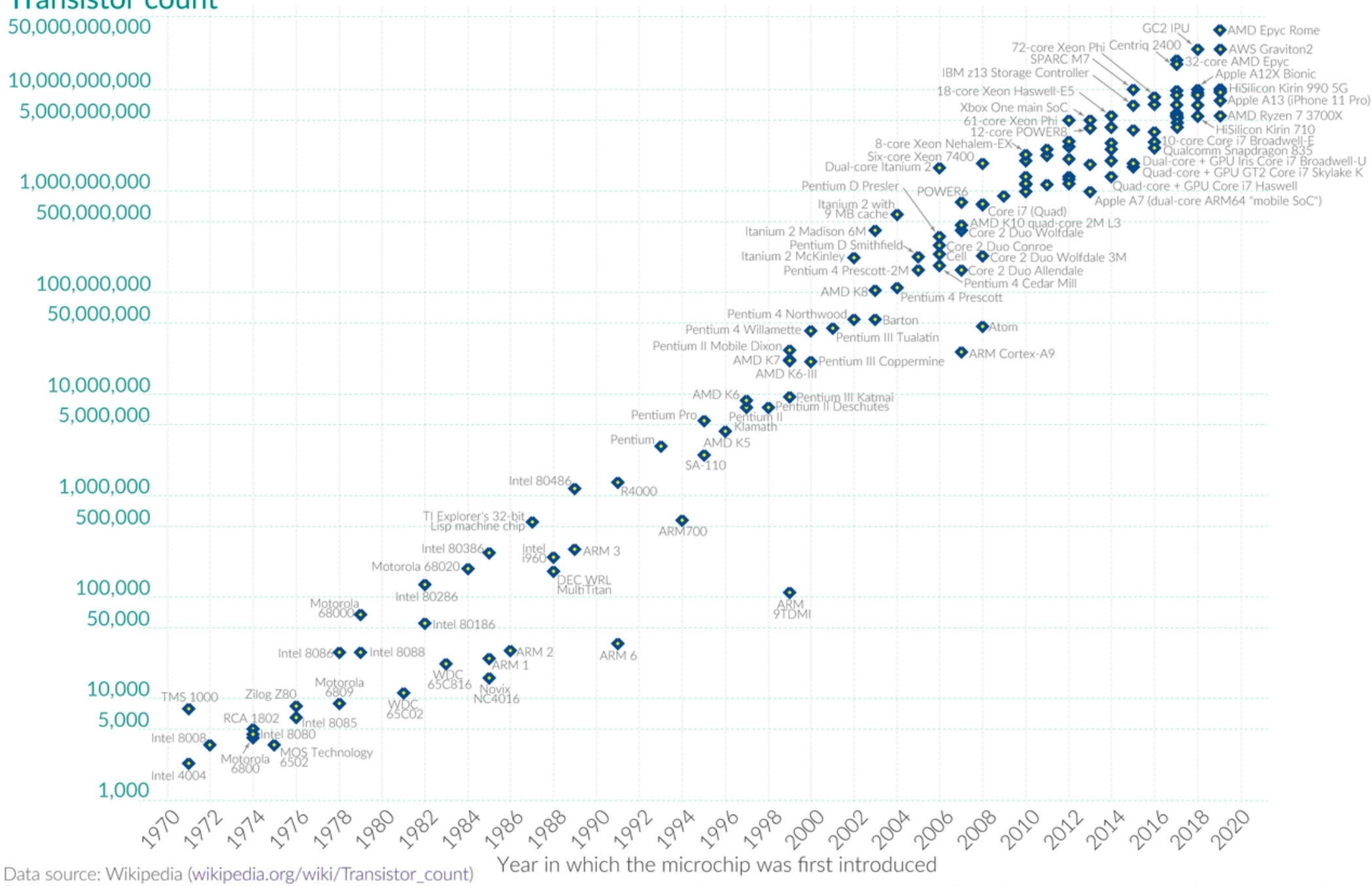
Moore's Law

- Number of transistors in an IC doubles about every two years.
  - It is an observation based on empirical evidence.
  - Helped companies predict, plan, prepare, and pursue.
  - Has been seen as a self-fulfilling prophecy.
- Jenson Huang (NVIDIA CEO) declared the law dead in 2022.

Moore's Law: The number of transistors on microchips doubles every two years

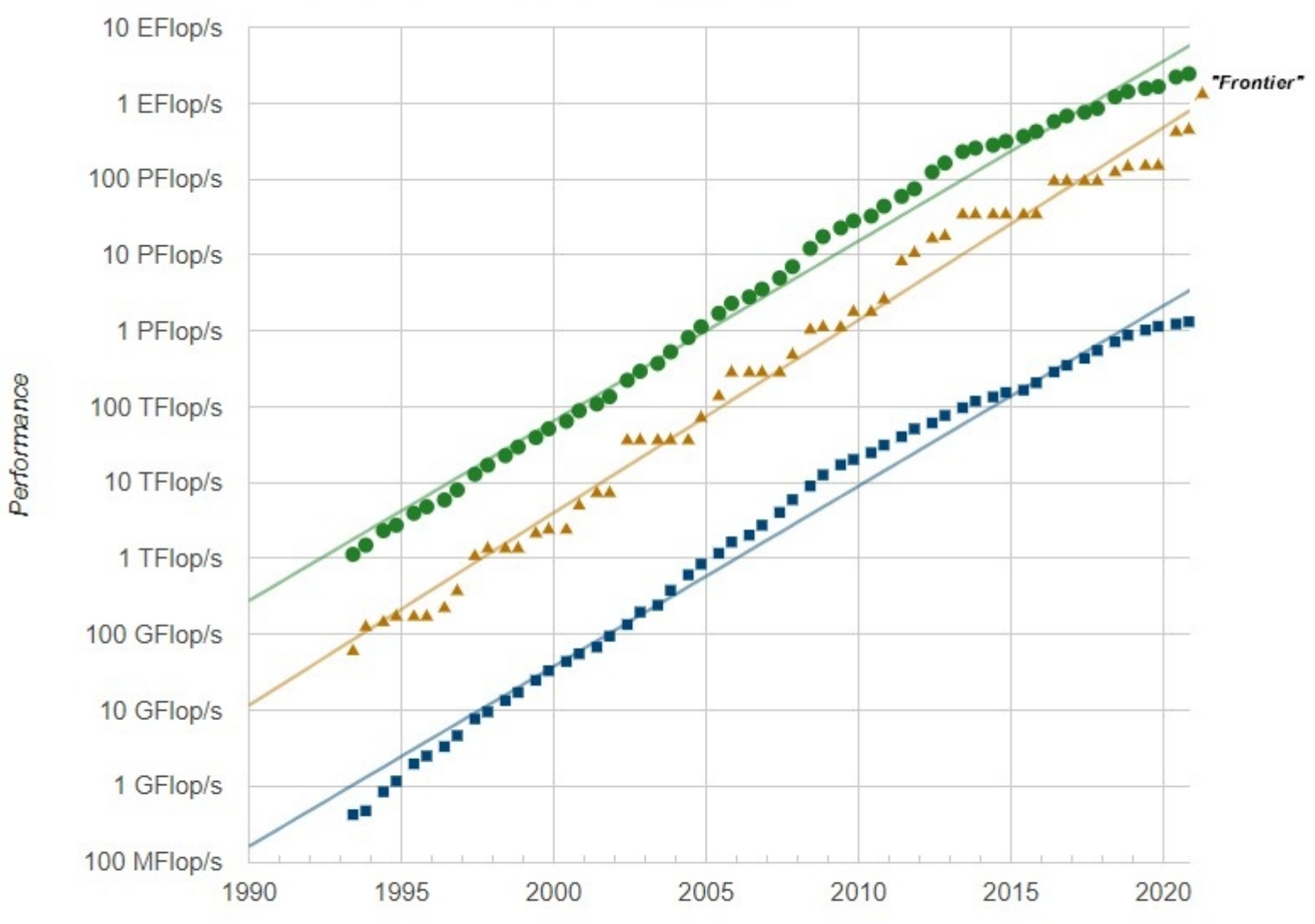
Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

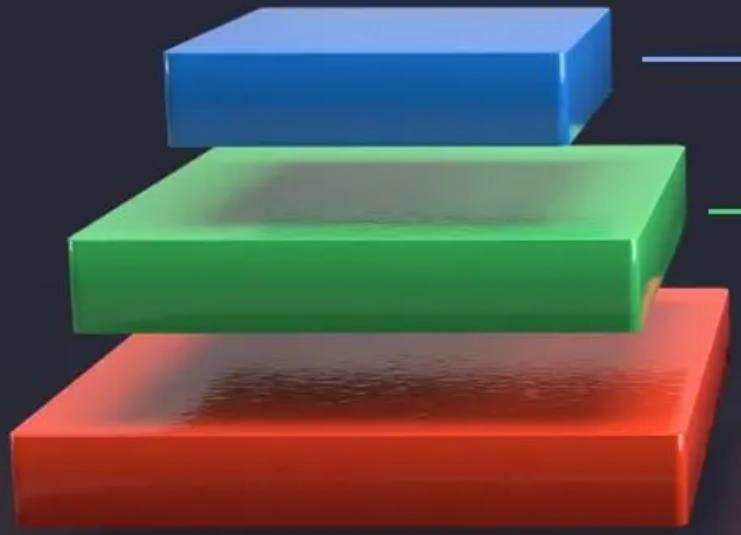


Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

HPC's Progress towards Exascale



Typical Deep Learning Pipeline With GPU

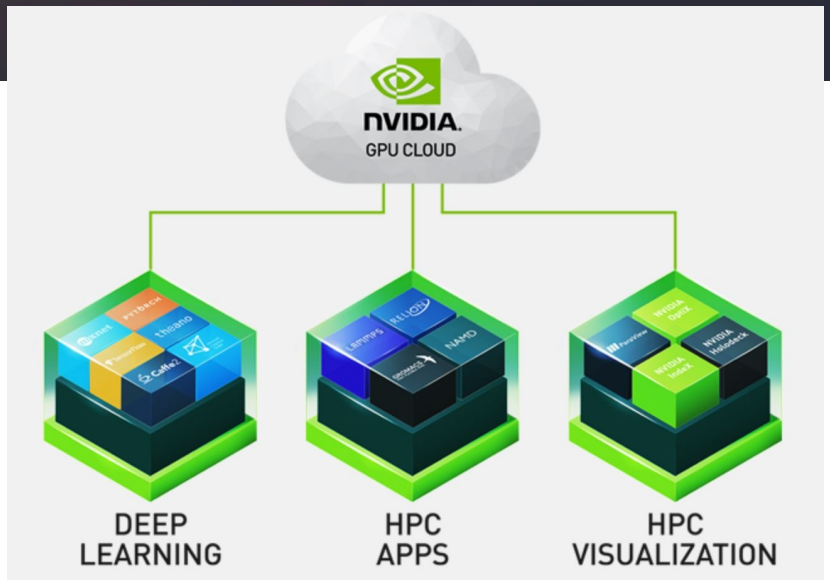


Data Preprocessing (CPU)

DNN Execution:
Training or Inference (GPU)

Data Post-Processing (CPU)

mobidev

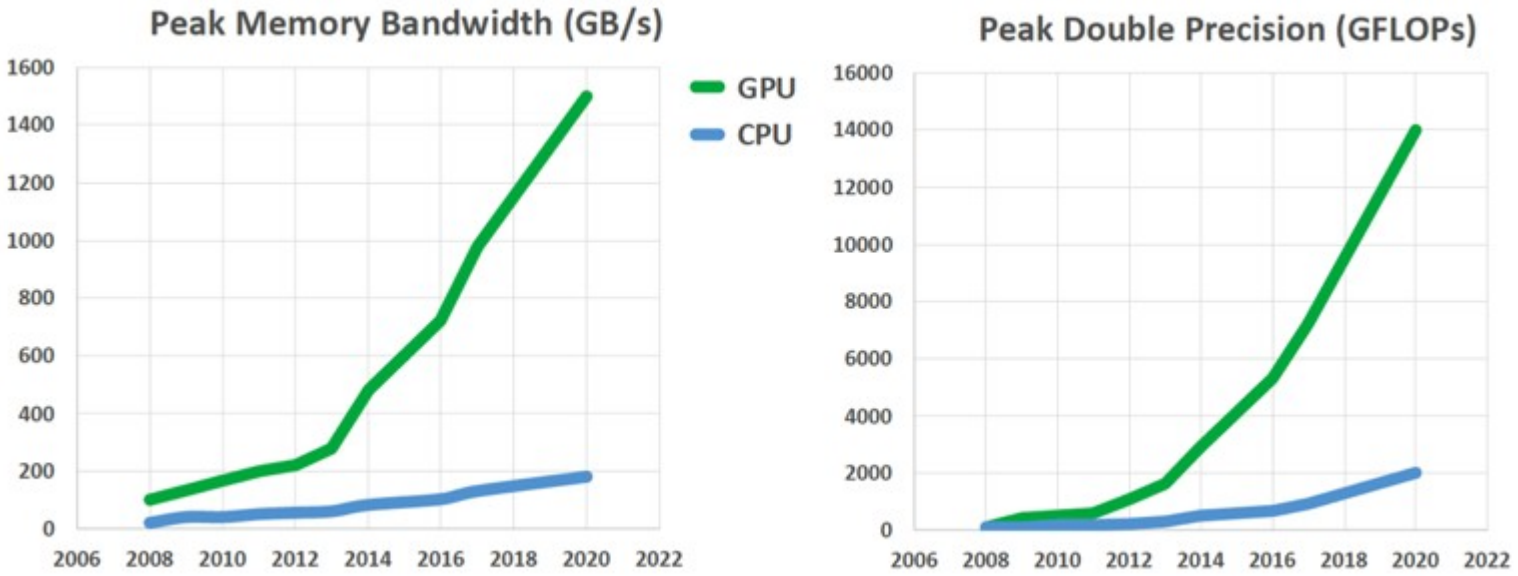


Creator: Patrick Kennedy

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 or olakrutrim (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 the next 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

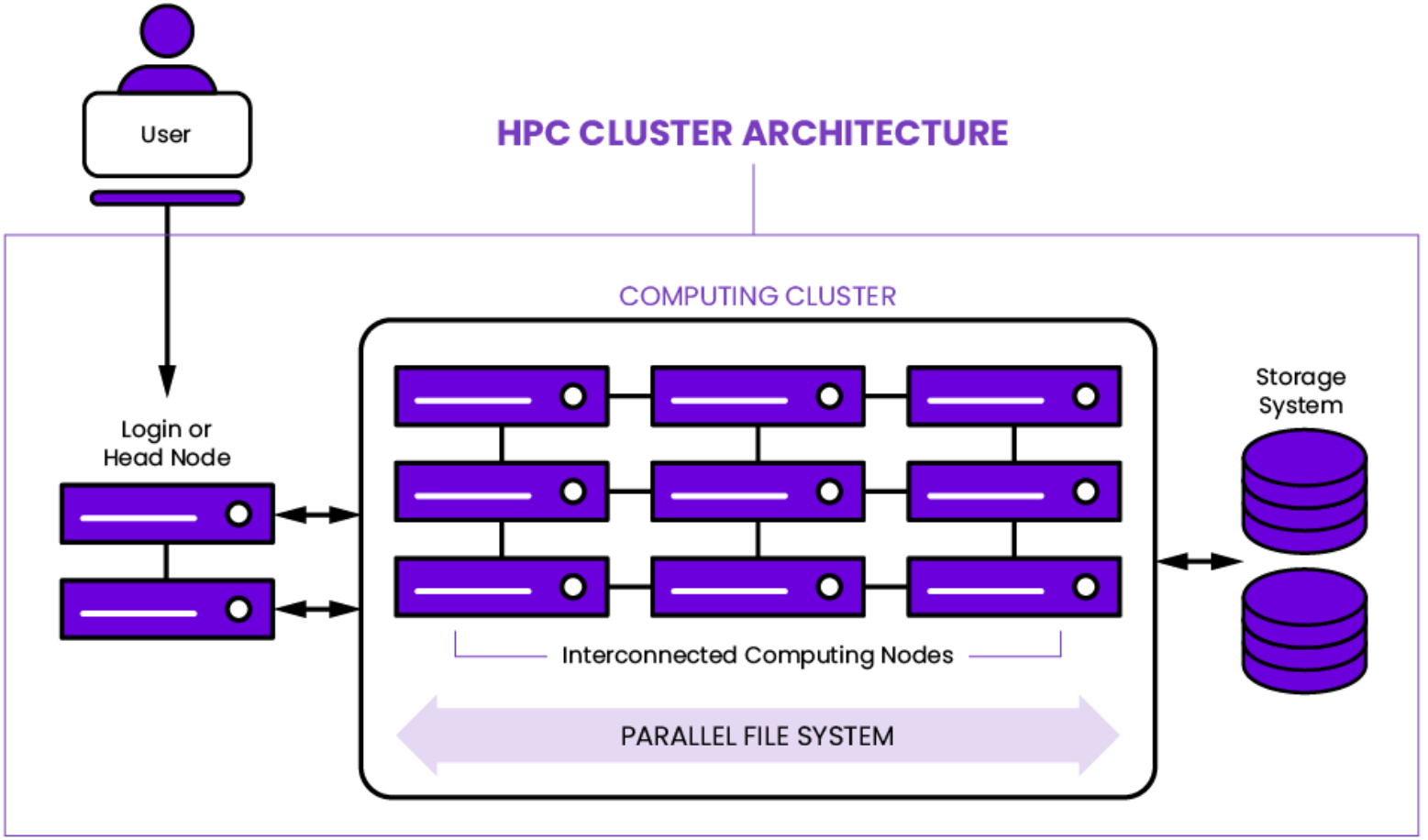
- The instructor takes attendance.
- The course-load is high compared to many other courses in the insti.
- There will be plagiarism checks on the submitted codes. Your grades will be reduced by two grades (S to B, D to U) for copying or for sharing your code or referred to DisCo.
- The assignment deadlines 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.~~

A few reactions on the course...

- Colleage from IBM: "I work on large scale distributed training and I just wanted to thank you for the amazing GPU programming course available on Youtube."
- PhD student at Illinois Institute of Technology: "I learned CUDA from your tutorial videos on YouTube and have found them to be very helpful."
- Graduate student from Bangadesh: "I am so charmed by your techniques of delivering lectures and noticed you indulge your students when they ask questions while you are still speaking."
- "Best lecture on Cuda programming Seen so far."
- "The best lecture on GPU in the entire YouTube."
- "lot of background noise.. in the video"
- Also taught under the National Supercomputing Mission (1500+ registrations), KLA, ...
- Our course material is used in AICTE, Gujarat TU, Bharath University, Lendi IET, Jagannath University, NITTR, ...

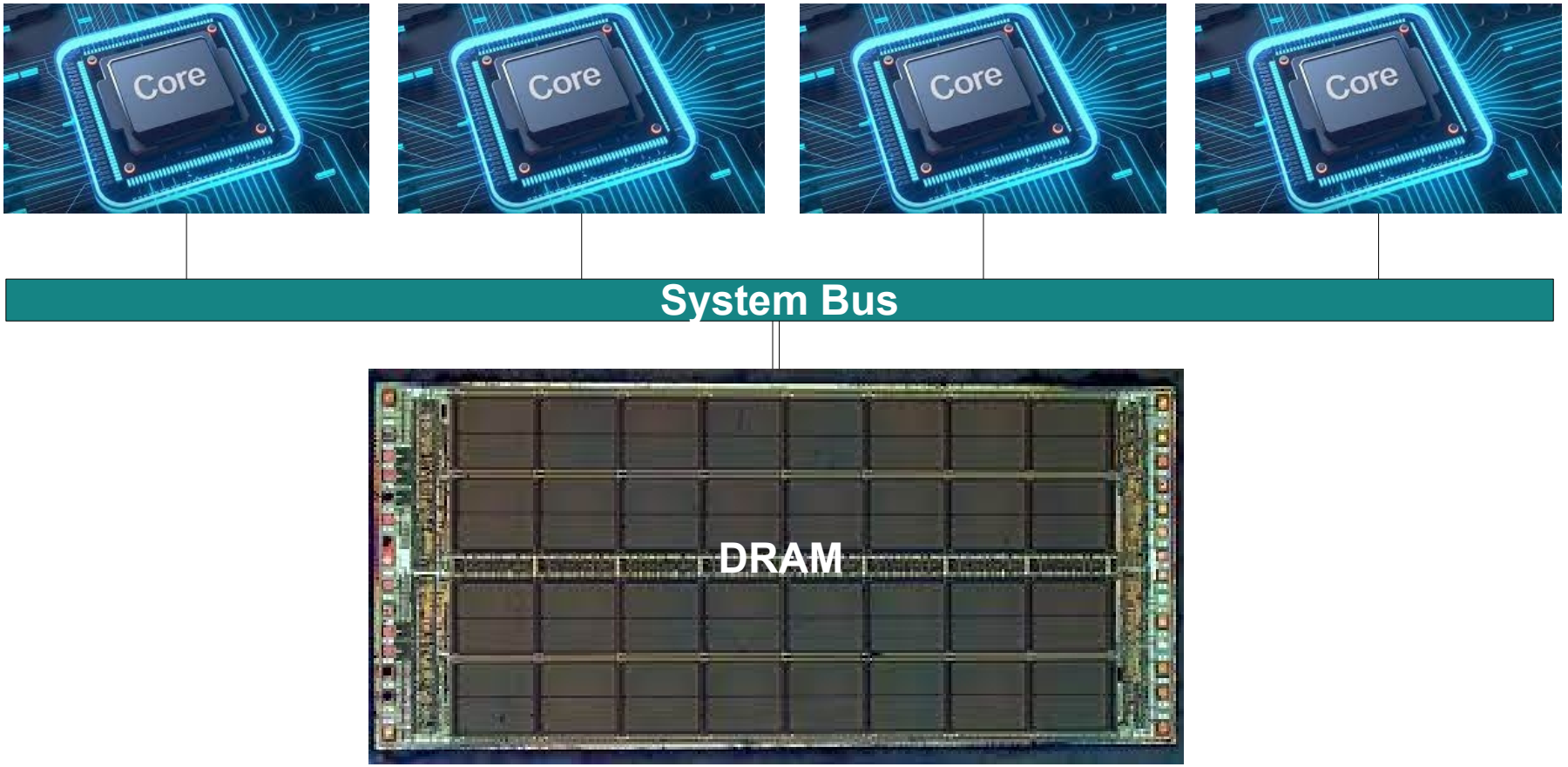
Additional Slides for Self Learning

HPC Platforms: Distributed



- No common memory; data and commands need to be explicitly communicated.
- Popular systems: cloud, map-reduce, Hadoop, WWW, …
- Programming with MPI (send, receive, barrier, …)
- Communication is the key!

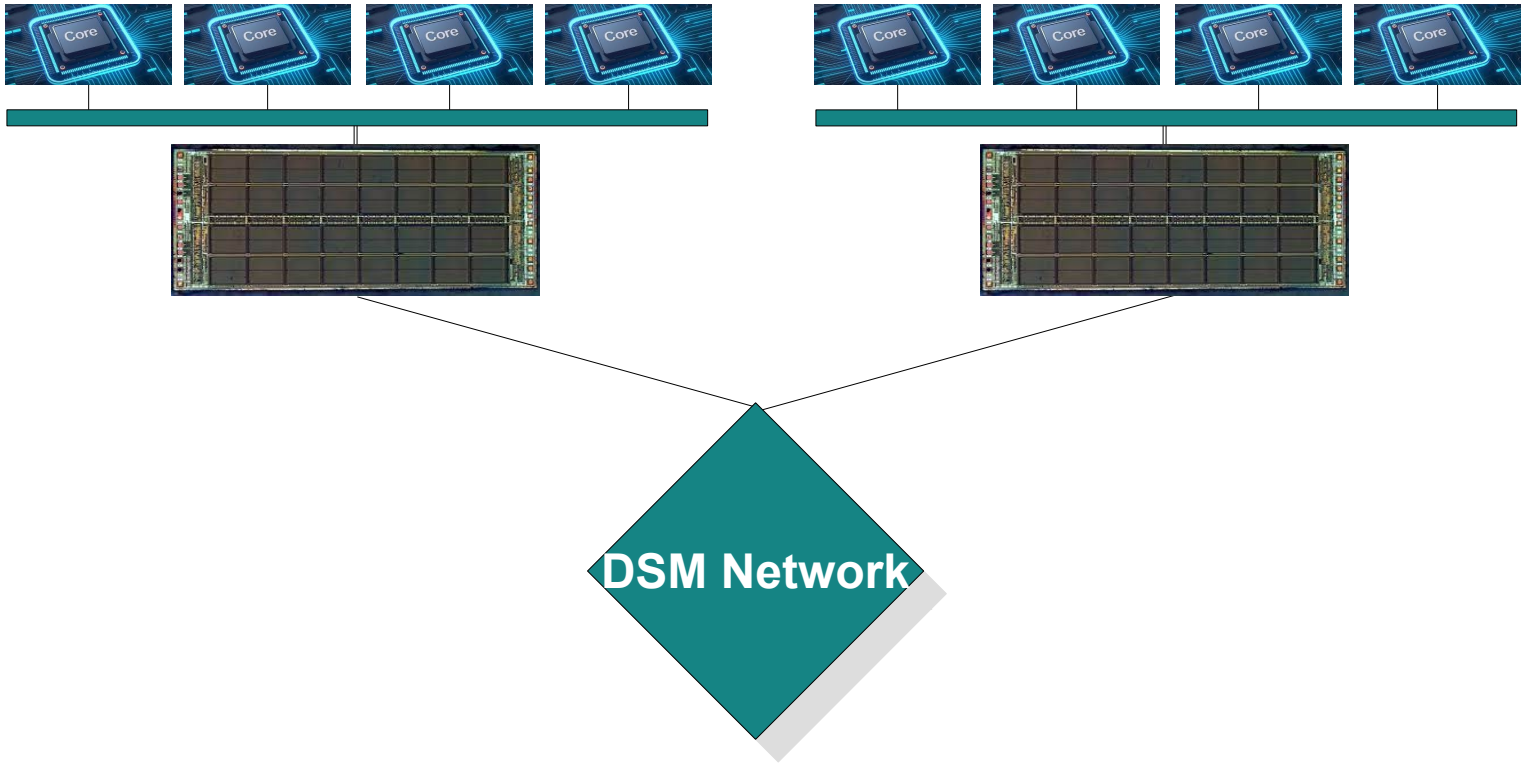
HPC Platforms: Shared Memory



- All the cores have access to the same memory.
- They can share data fast, and also need to synchronize.
- Popular systems: all our machines, including phones
- Programming with pthreads, OpenMP, ...

HPC Platforms: Shared Memory

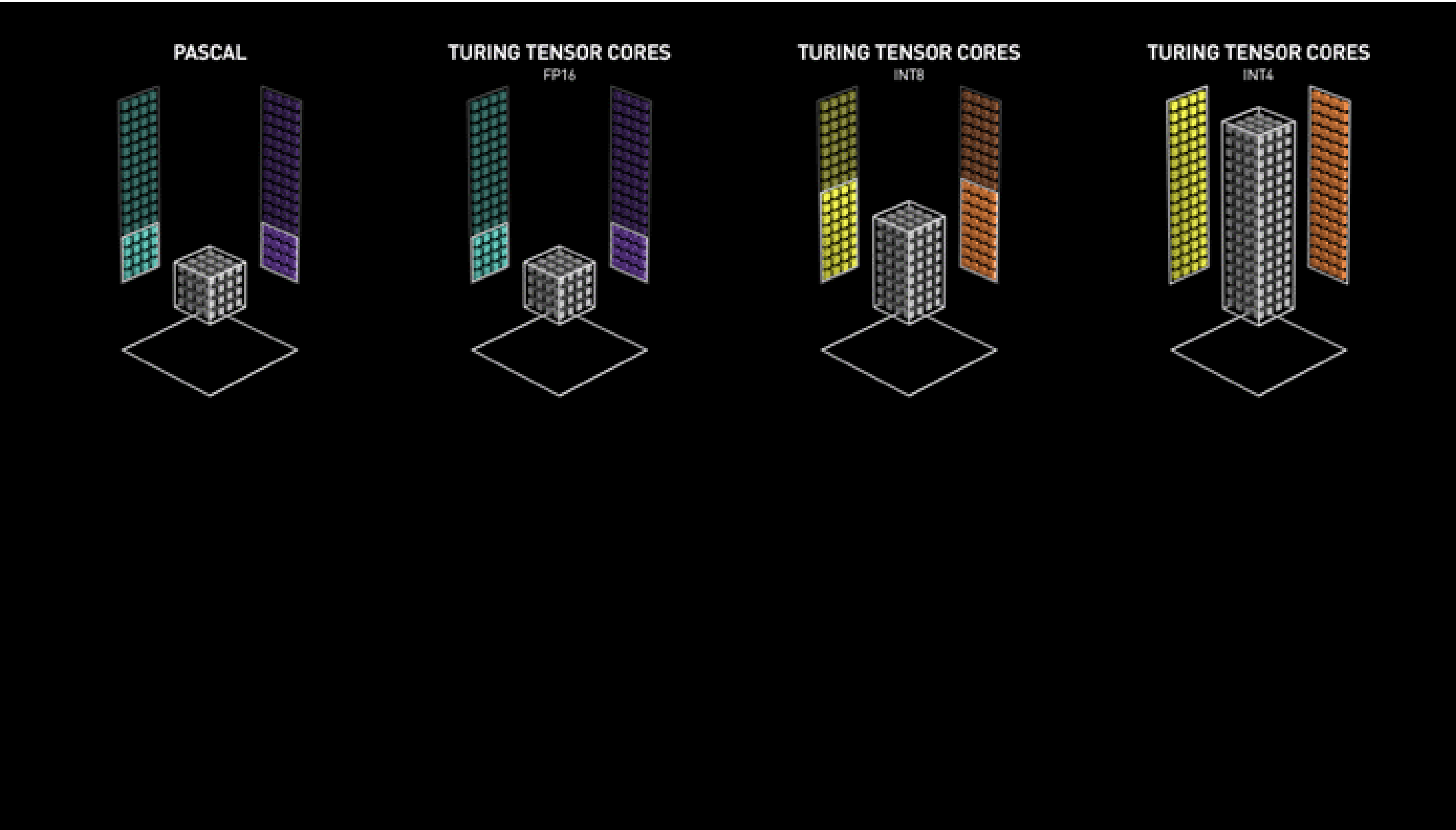
- Cores may get grouped into sockets.
  - e.g., dual-socket system
- Leads to Non-Uniform Memory Access (NUMA)
  - Data access needs to respect the memory architecture.



HPC Platforms: GPU

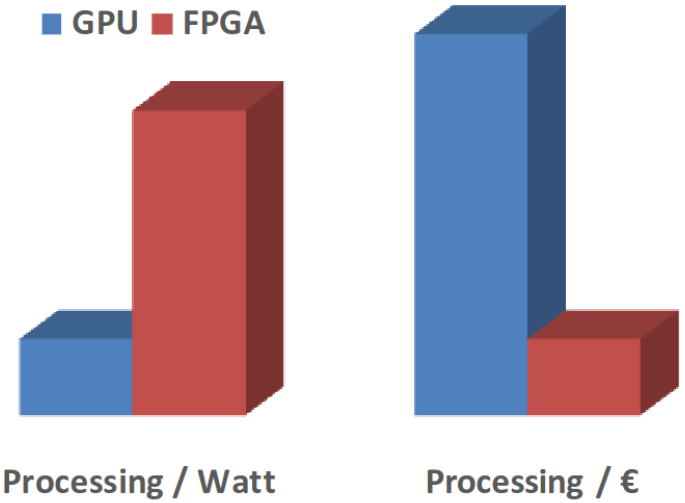
- Separate device, separate address space*
- Thousands of cores
- Massive multi-threading
- Connected via PCI-e, NVLink
- Vendors: NVIDIA, Intel, AMD, ...
- Example: NVIDIA GH100
  - 14,500 cores
  - offers 48 TF FP32 (recall CPU speed)
  - offers 400 TF TF32 (tensor cores)

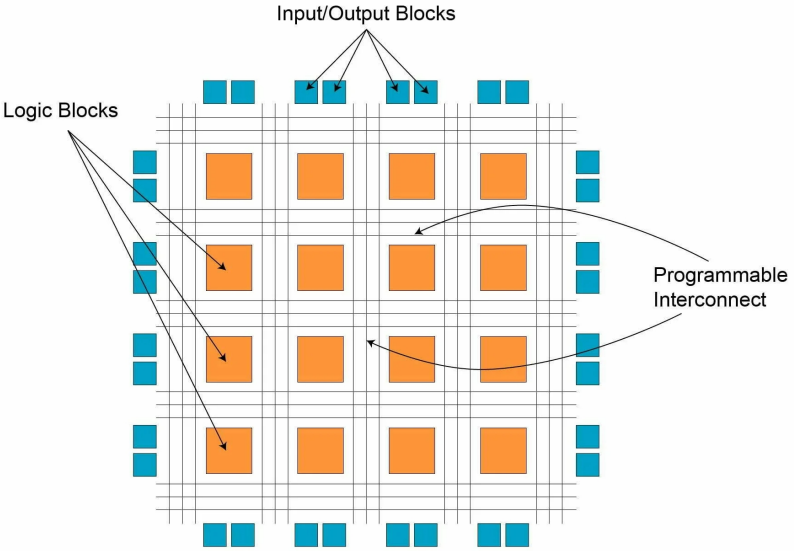
HPC Platforms: GPU



HPC Platforms: FPGA

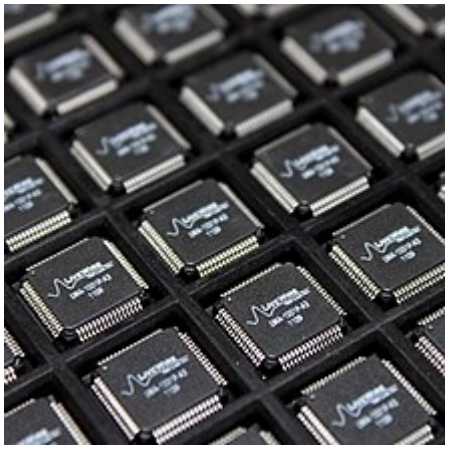
- Field Programmable Gate Arrays
- Reconfigurable interconnects
  - After manufacturing
- Useful for custom applications
- Vendors: Xilinx, Intel, Lattice, ...





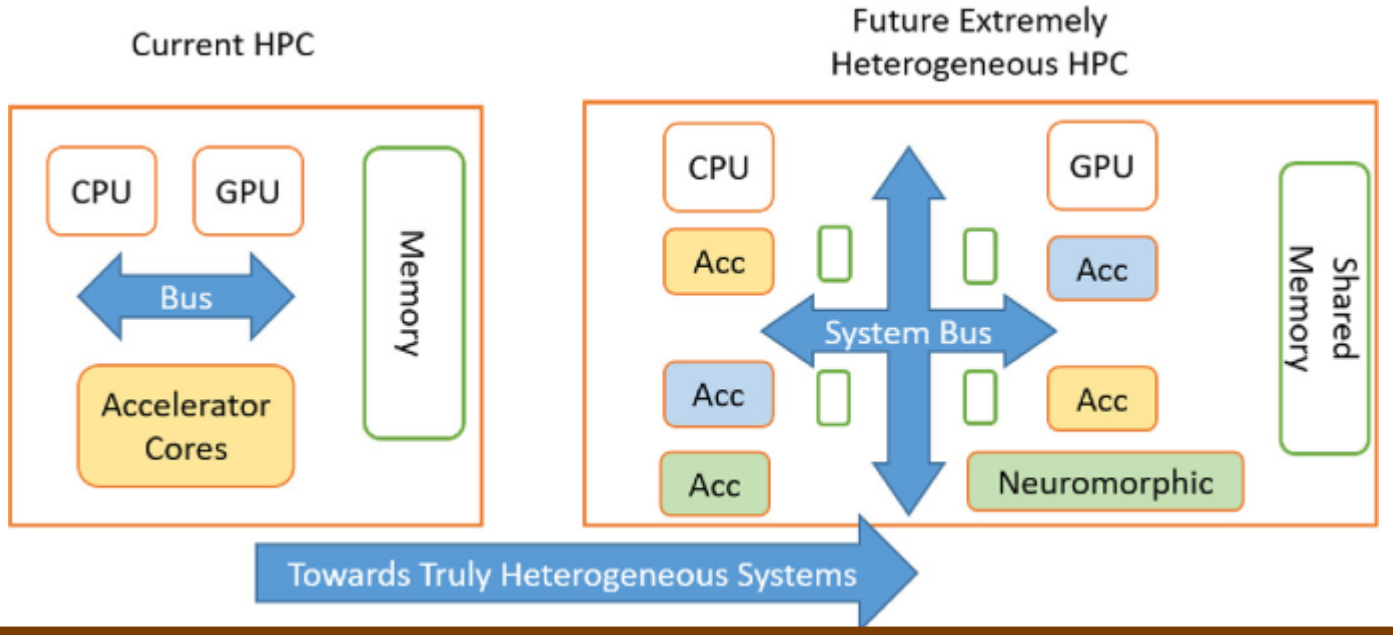
HPC Platforms: ASIC

- Application Specific Integrated Circuit
- Offers better performance
- Works with only that application
- Example domains:
  - Ethernet switch
  - Video codec
  - Proprietary controllers
- Designed using HDL
- ASICs are often used in large production volumes
  - FPGAs may get used in prototyping



HPC Platforms: Heterogeneous

- Practical world uses a combination of HPC platforms
- Different hardware elements serve different needs.
- Need new technologies for sharing data



Parallel architectures are often classified based on instruction and data.