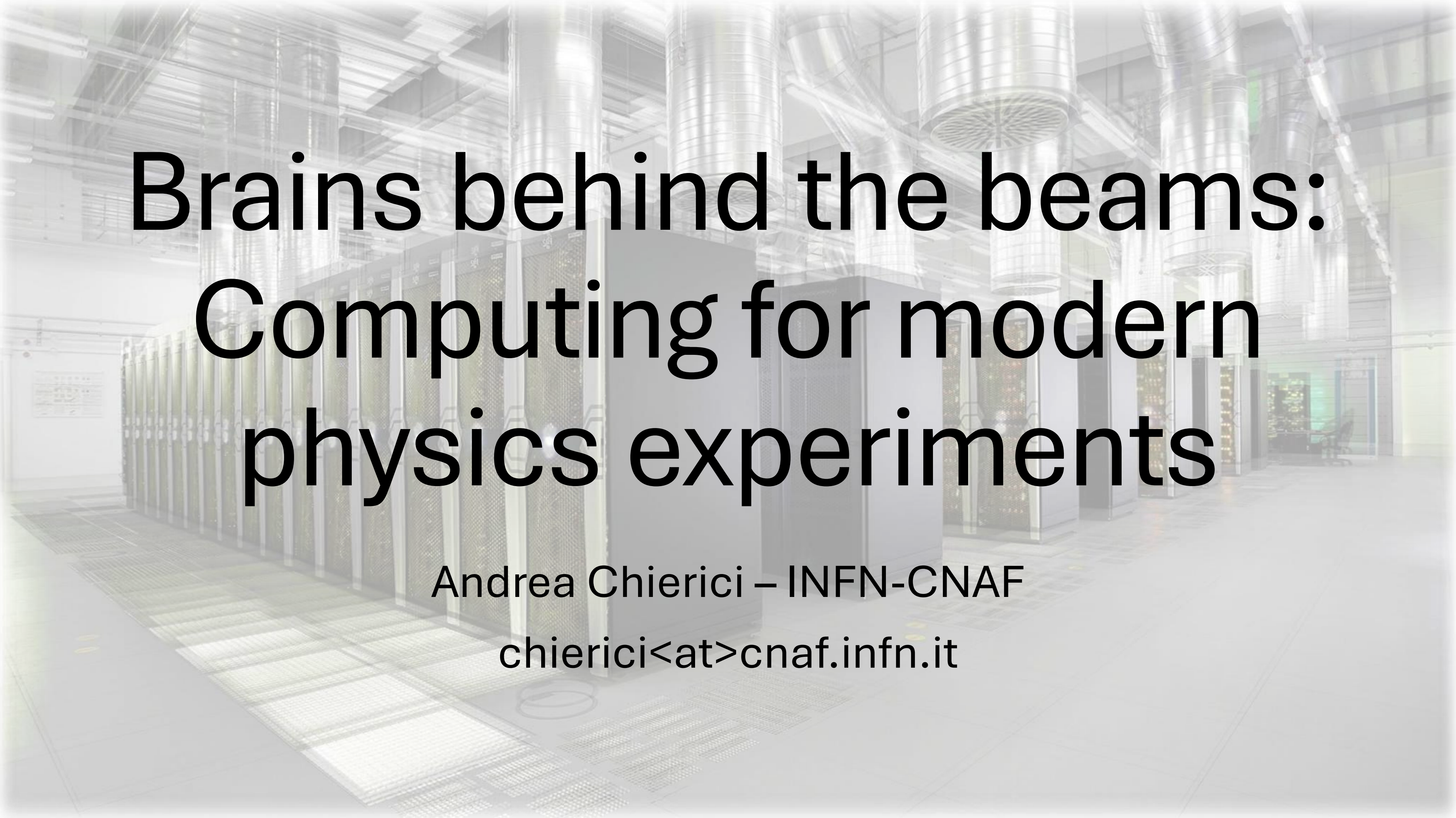


Brains behind the beams: Computing for modern physics experiments

Andrea Chierici – INFN-CNAF

chierici<at>cnaf.infn.it

Evolution of computing in particle physics

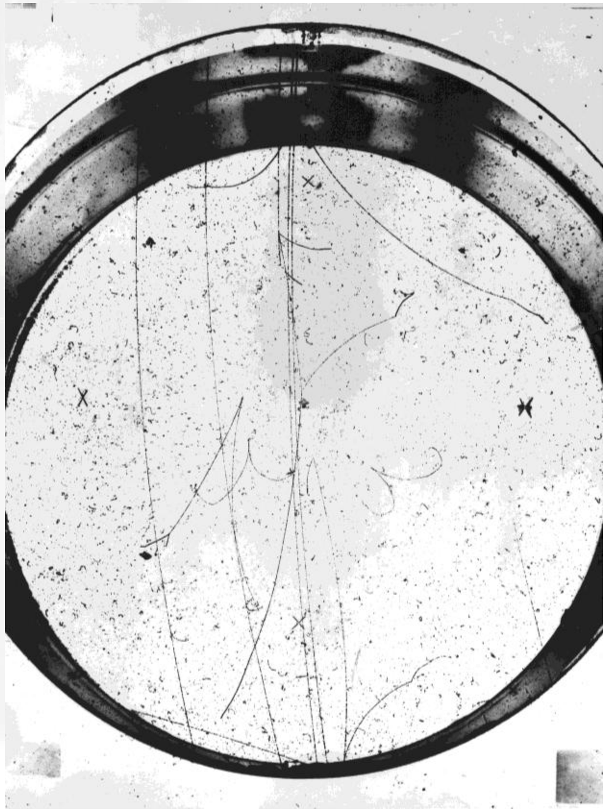
- The evolution of computing in particle physics has been marked by a close co-evolution with the advancements in computing technology itself
  - driven by the ever-increasing computational demands
  - From manual calculations and early electronic computers to distributed computing grids
  - In the future the potential of quantum computing
- Particle physics has consistently pushed the boundaries of what's computationally possible
  - HPC, Grid computing

Early days

- Calculations were performed manually, often by teams of "computers".
- Early electronic computers were developed to handle the growing complexity of calculations

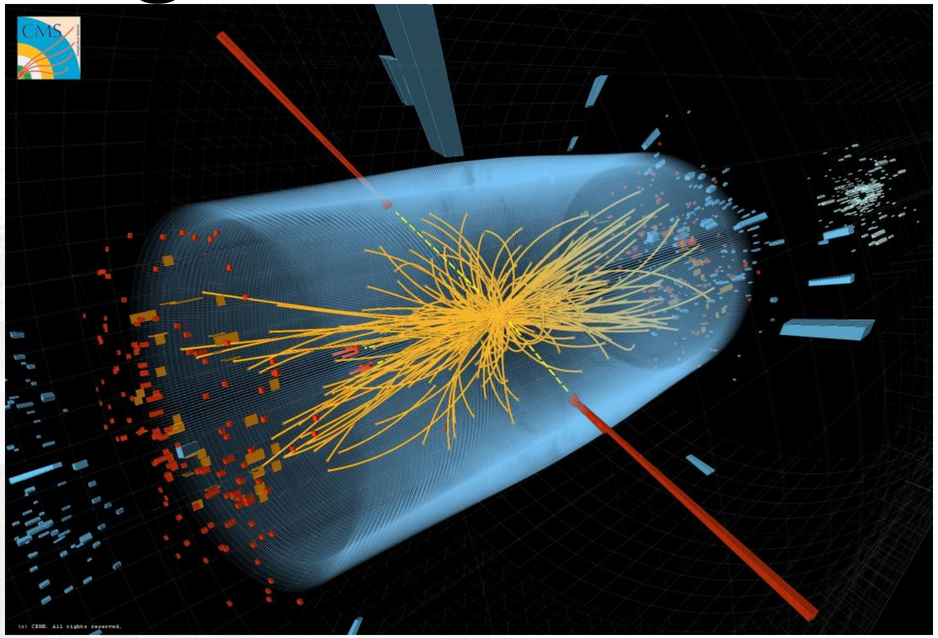
bubble chambers

- a vessel filled with a superheated transparent liquid used to detect electrically charged particles moving through it



The rise of distributed computing

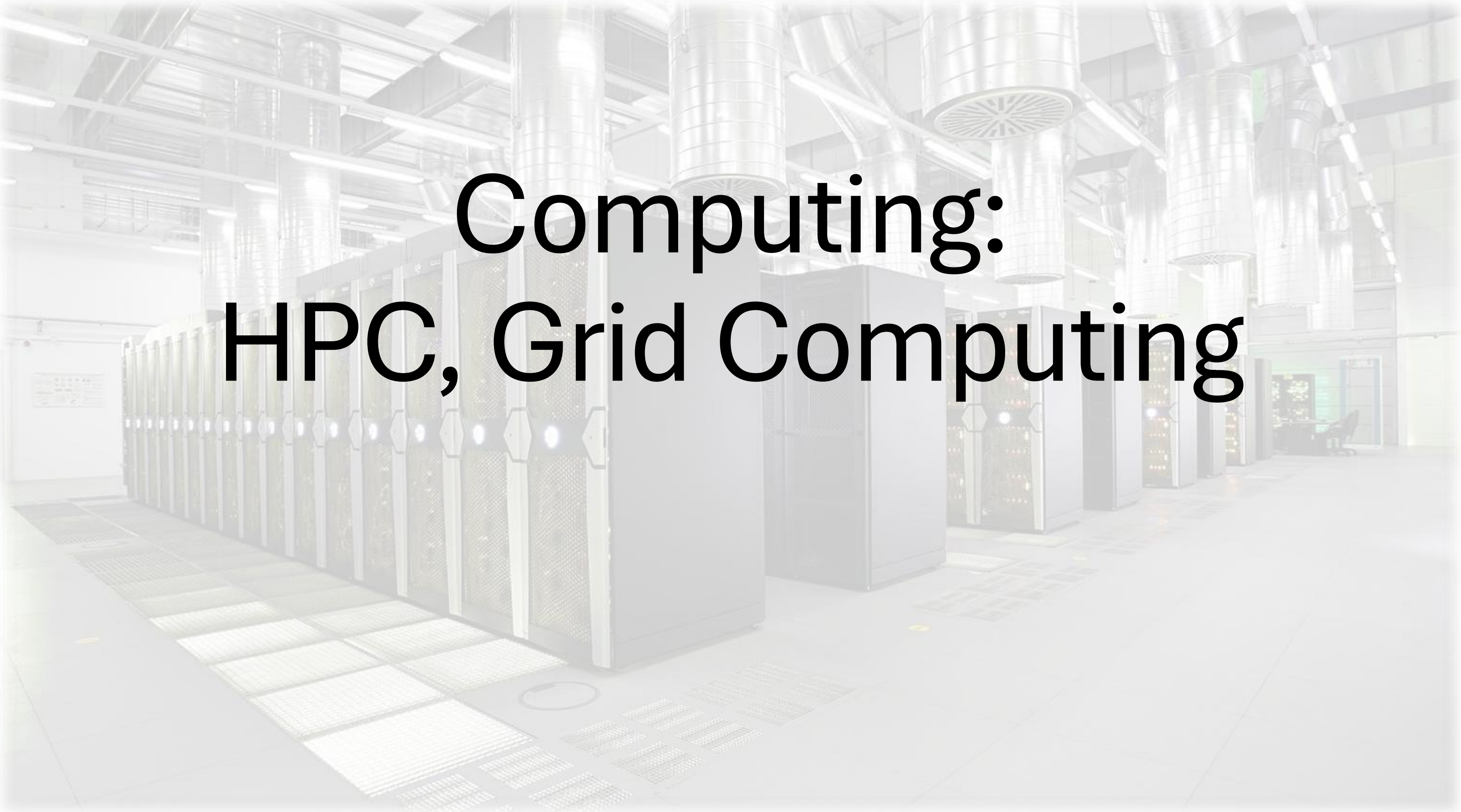
- Particle physics experiments generate vast amounts of data, requiring massive computational power.
- The Worldwide LHC Computing Grid (WLCG) emerged as a collaborative effort to distribute and analyze LHC data, integrating computer centers worldwide.
- This federated model, based on grid technologies, allowed physicists to access and process data from anywhere.



Today, what should I use?

- Which computer?

- Which storage?

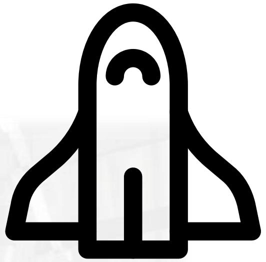


Computing:
HPC, Grid Computing

What is HPC — and why should we care?

- Not your average PC: HPC means solving really hard problems fast
- Used for climate modeling, astrophysics, material science
- Think “computational telescope”: it helps you see the invisible
- Spoiler: Your phone may be faster than 90s supercomputers

Classical computing vs HPC

- Classical computing is a  HPC is a 

- Classical computing = serial, one instruction at a time.
- HPC = parallel. Thousands of instructions simultaneously
- Scaling up ≠ just adding more PCs
- Requires coordination, special hardware, and clever software

Types of HPC systems

- Supercomputers
  - Extremely fast computers used for large-scale computations.
- Clusters
  - Groups of linked computers working together as a single system.
- Grid Computing
  - Distributed computing resources across multiple locations.

HPC technologies

- Parallel Computing
  - Simultaneous data processing using multiple processors.
- Distributed Computing
  - Computing tasks distributed across multiple machines.
- Cloud Computing
  - Using remote servers hosted on the internet to store, manage, and process data.

Divide, Conquer, and Simulate the Universe

- HPC divides large problems into smaller tasks
- Each task runs on a different CPU/GPU core
- Communication between tasks is the bottleneck
- Efficiency depends on problem type and architecture

Some definitions

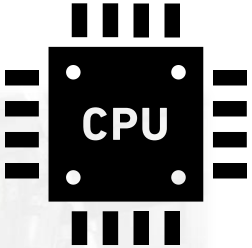
- GFLOPS: Billions of Floating-Point Operations per second
  - Max GFLOPS of a system can be calculated using:

$$GFLOPS = sockets \times \frac{cores}{sockets} \times clock \times \frac{FLOPS}{cycle} \quad (clock\ in\ Ghz)$$

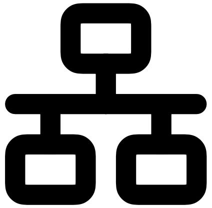
- TDP: Thermal Design Power is the maximum amount of heat generated by the CPU that the cooling system in a computer is required to dissipate in typical operation
- Vector processor: CPU that implements an instruction set designed to operate efficiently and effectively on large one-dimensional arrays of data called vectors
- This contrasts with scalar processors, whose instructions operate on single data items only.
- Vector processors can greatly improve performance on certain workloads, notably numerical simulation and similar tasks

Anatomy of an HPC System

- Compute Nodes: CPUs, GPUs, memory



- Interconnects: low-latency networks like InfiniBand

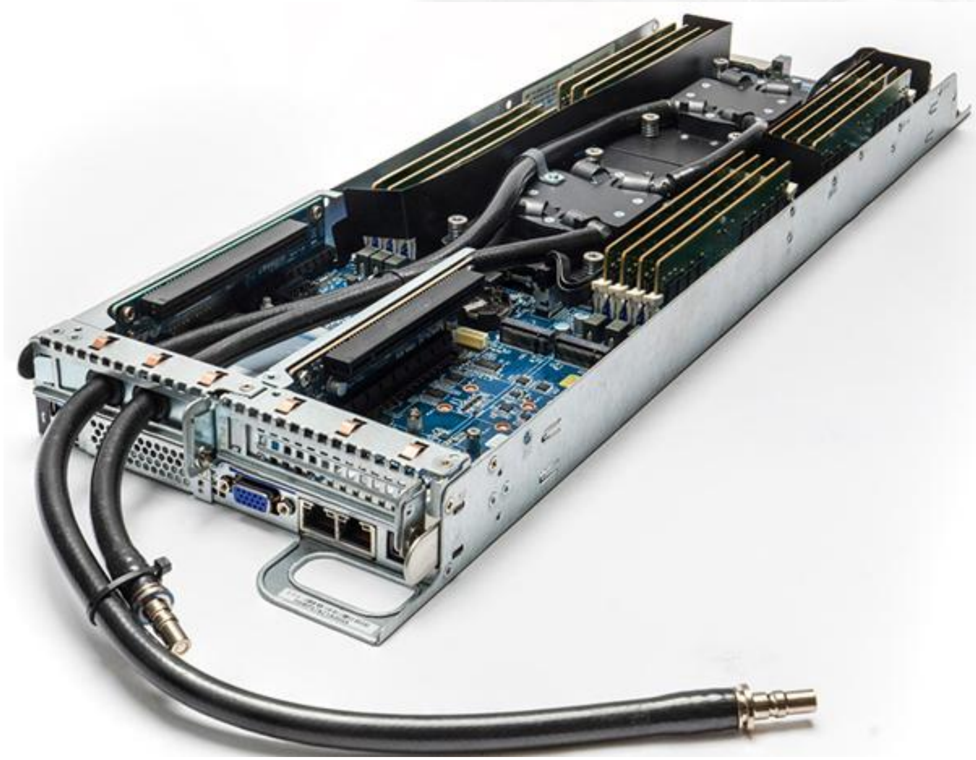


- Storage: parallel file systems (Lustre, BeeGFS)

- Cooling & Power: liquid, air, and money

Cooling digression

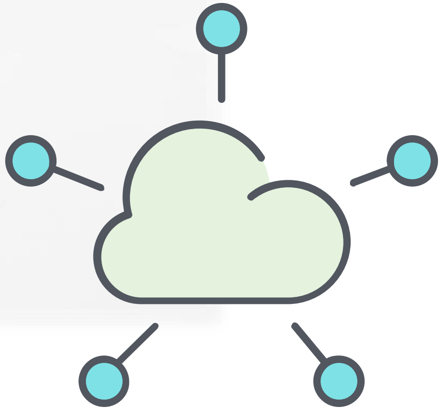
Direct liquid cooling



Immersion cooling

CPUs vs. GPUs

- CPUs: general-purpose, few strong cores
- GPUs: many simple cores, massive parallelism
- HPC loves GPUs for matrix-heavy tasks (e.g., simulations, AI)
- Used together for hybrid computing



The First Supercomputers

- Cray-1 (1976): 80 MFLOPS, 5.5 tons, Freon-cooled
- 5.5 tons including the Freon refrigeration
- Vector processors: ideal for numerical simulations
- From $8M dinosaurs to petaflop laptops
- Fun fact: Seymour Cray dug tunnels to think better

Quite a complicate machine

Microprocessors

1971

The Intel 8080
1974

The Intel 8086
1978

The Intel 80286
1982

The Intel 80486
1989

In the 80s we saw the replacement of vector supercomputers with Massively Parallel Processors (MPP) and clusters assembled from a larger number of slower performing microprocessors

Clusters

- A parallel computer system
  - comprising an integrated collection of independent nodes
    - each of which is a system in its own
    - capable of independent operation
    - derived from products developed and marketed for other stand-alone purposes

TOP500: The World Ranking

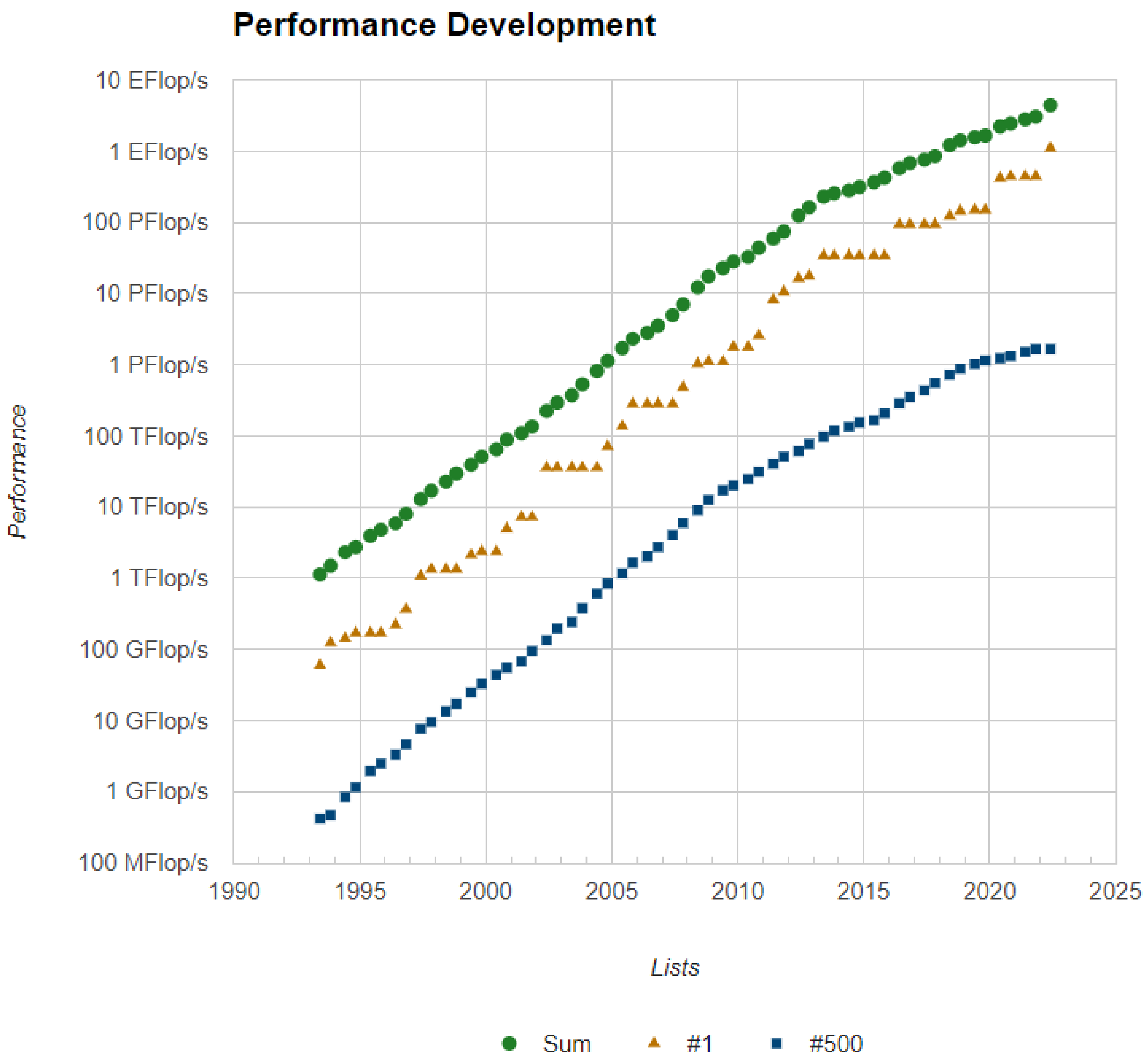
- List of the 500 most powerful supercomputers
  - https://top500.org
- Updated twice a year: ISC in June, SC in November
- Measured with Linpack (HPL) benchmark
- The project aims to provide a reliable basis for tracking and detecting trends in high-performance computing
- Italy's Leonardo is in the top positions

Top500.org (Jun 25)

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN EuroHPC/FZJ Germany	4,801,344	793.40	930.00	13,088
5	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
6	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A Italy	3,143,520	477.90	606.97	8,461

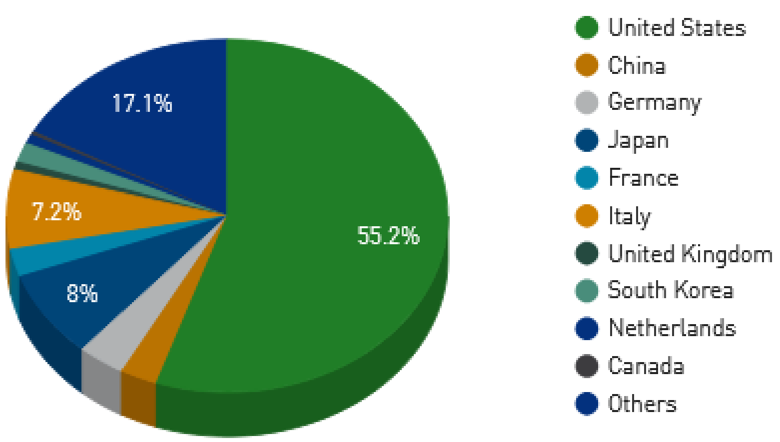
8	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90
9	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70
10	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	241.20

top500.org - stats



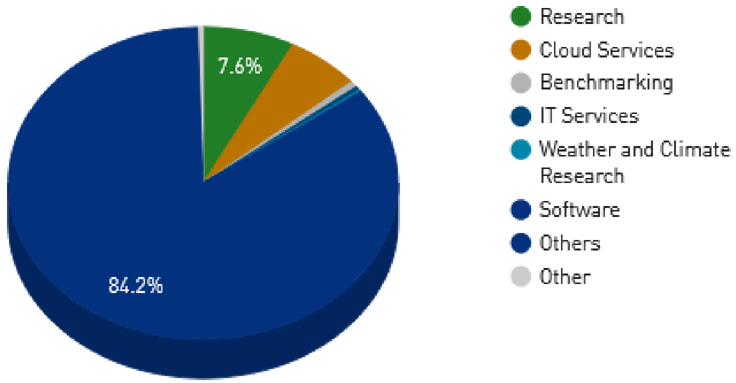
Top500.org - stats

Countries Performance Share

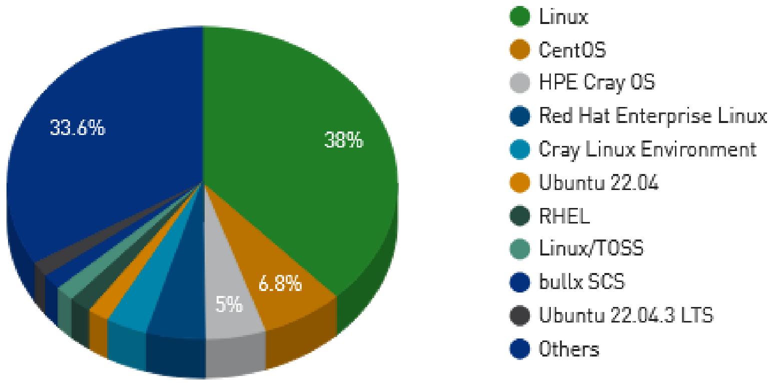


	Countries	Count
1	United States	172
2	China	63
3	Germany	41
4	Japan	34
5	France	24
6	Italy	14
7	United Kingdom	14
8	South Korea	13
9	Netherlands	10
10	Canada	9

Application Area Performance Share



Operating System System Share



El Capitan

- Hewlett Packard Enterprise El Capitan, is an exascale supercomputer, hosted at the Lawrence Livermore National Laboratory in Livermore, United States and becoming operational in 2024.
- El Capitan uses a combined 11,039,616 CPU and GPU cores consisting of 43,808 AMD 4th Gen EPYC 24C "Genoa" 24 core 1.8 GHz CPUs (1,051,392 cores) and 43,808 AMD Instinct MI300A GPUs (9,988,224 cores).
- Blades are interconnected by an HPE Slingshot 64-port switch that provides 12.8 terabits/second of bandwidth. Total cabling runs 145 km (90 mi).
- El Capitan uses an APU architecture, where the CPU and GPU share an internal on-chip coherent interconnect.

El Capitan

Active	Deployment: 2H 2023 Completion: 2024
Sponsors	U.S. Department of Energy
Operators	Lawrence Livermore National Laboratory and U.S. Department of Energy
Location	Livermore Computing Complex
Architecture	HPE Cray Shasta
Power	40 MW (Proj)
Operating system	TOSS
Space	TBA
Memory	TBA
Storage	TBA
Speed	1.742 exaFLOPS (Rmax) / 2.746 exaFLOPS (Rpeak)
Cost	US$600 million (estimated cost)
Purpose	Scientific research and development, stockpile stewardship[1]

Fugaku

- The supercomputer is built with the Fujitsu A64FX microprocessor.
  - Based on the ARM version 8.2A processor architecture
  - Fugaku was aimed to be about 100 times more powerful than the K computer
    - i.e. a performance target of 1 exaFLOPS
- The initial (June 2020) configuration of Fugaku used 158,976 A64FX CPUs joined together using Fujitsu's proprietary torus fusion interconnect.
- An upgrade in November 2020 increased the number of processors
  - **To reach 442 petaFLOPS**

Active	From 2021
Sponsors	MEXT
Operators	RIKEN
Location	RIKEN Center for Computational Science (R-CCS)
Architecture	158,976 nodes Fujitsu A64FX CPU (48+4 core) per node Tofu interconnect D
Operating system	Custom Linux-based kernel
Memory	HBM2 32 GiB/node
Storage	1.6 TB NVMe SSD/16 nodes (L1) 150 PB shared Lustre FS (L2)[1] Cloud storage services (L3)
Speed	442 PFLOPS (per TOP500 Rmax), after upgrade; higher 2.0 EFLOPS on a different mixed-precision benchmark
Cost	US$1 billion (total programme cost)[2][3]
Ranking	TOP500: 1, June 2020
Web site	www.r-ccs.riken.jp/en/fugaku
Sources	Fugaku System Configuration

Lumi

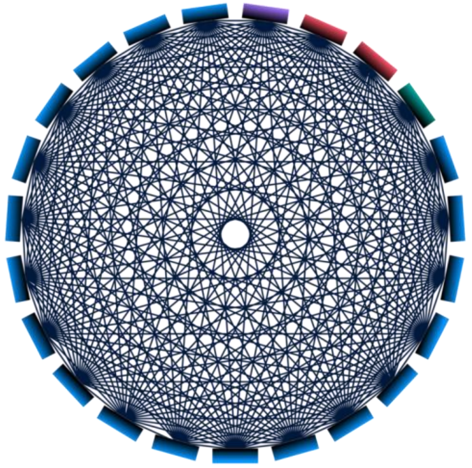
Active	June 13, 2022
Sponsors	European High-Performance Computing Joint Undertaking, LUMI Consortium
Location	Kajaani, Finland
Architecture	362,496 cores, AMD EPYC CPUs, 10,240 AMD Radeon Instinct MI250X GPUs (144,179,200 cores)[1][2]
Power	8.5 MW
Space	150 m^2
Memory	1.75 petabytes
Storage	117 petabytes
Speed	550 petaFLOPS (peak)
Cost	€144.5 million
Website	www.lumi-supercomputer.eu

- **LUMI** (**Large Unified Modern Infrastructure**) is a petascale supercomputer located at the CSC data center in Kajaani, Finland.
- The completed system will consist of around 362,496 cores, capable of executing more than 375 petaflops, with a theoretical peak performance of more than 550 petaflops, which would place it among the top five most powerful computers in the world
- The system is being supplied by HPE, providing an HPE Cray EX supercomputer with next generation 64-core AMD EPYC CPUs and AMD Radeon Instinct GPUs. LUMI is a GPU based system, and the majority of its computing power comes from its GPU cores, an architecture which was chosen primarily for its cost/performance advantage.

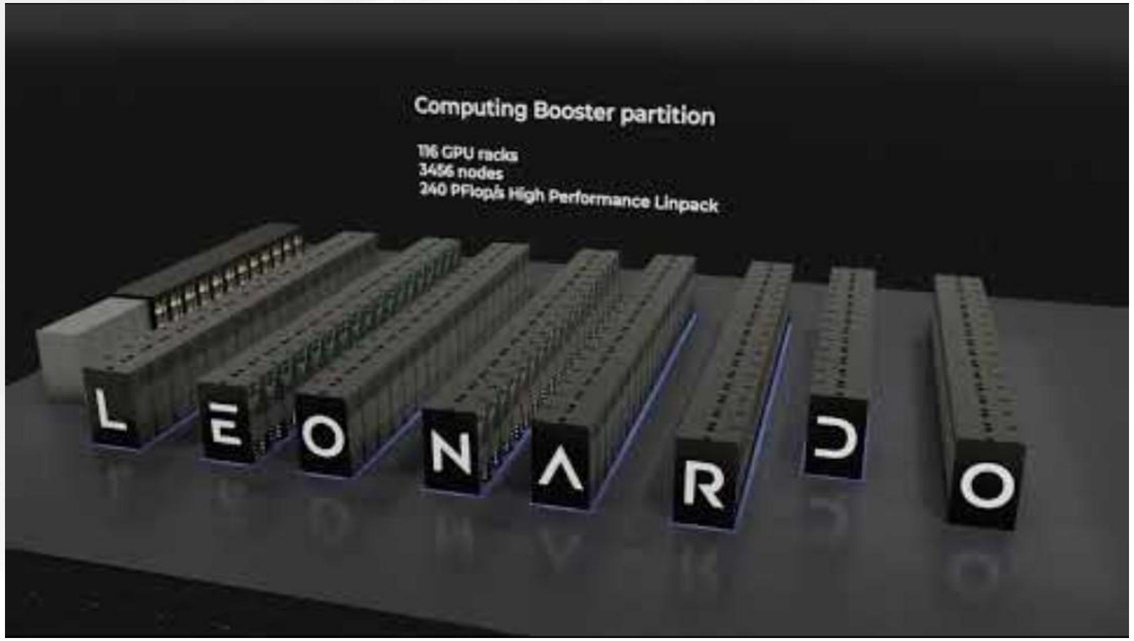
Leonardo

LEONARDO'S NUMBERS

155 SYSTEM RACKS	**4992** COMPUTING NODES	**250** PETAFLOPS	**2800** TB OF RAM
6 MW IN OPERATIONS	**110** PB OF STORAGE	**600** M² FOOTPRINT	**>95%** HEAT DISSIPATION VIA DLC



Booster Module nodes
I/O cell
Data-Centric cells
Hybrid cell (Booster + Data-Centric nodes)



Do you have a supercomputer at home?

- Gaming console technology is “similar” to El Capitan supercomputer
  - Multicore CPU
  - High memory bandwidth
  - GPU
  - Fast ssd storage

PS5

CPU: 8x Zen 2 Cores at 3.5GHz (variable frequency)

GPU: 10.28 TFLOPs, 36 CUs at 2.23GHz (variable frequency)

Memory: 16GB GDDR6/256-bit

Memory Bandwith: 448GB/s

Internal Storage: Custom 825GB SSD

I/O Throughput: 5.5GB/s (Raw), Typical 8-9GB/s (Compressed)

Expandable Storage: NVMe SSD Slot

External Storage: USB HDD Support

Optical Drive: 4K UHD Blu-ray Drive

XBOX SERIES X

CPU: 8x Cores @ 3.8 GHz (3.6 GHz w/ SMT) Custom Zen 2 CPU

GPU: 12 TFLOPS, 52 CUs @ 1.825 GHz Custom RDNA 2 GPU

Memory: 16 GB GDDR6 w/ 320b bus

Memory Bandwith: 10GB @ 560 GB/s, 6GB @ 336 GB/s

Internal Storage: 1 TB Custom NVME SSD

I/O Throughput: 2.4 GB/s (Raw), 4.8 GB/s (Compressed, with custom hardware decompression block)

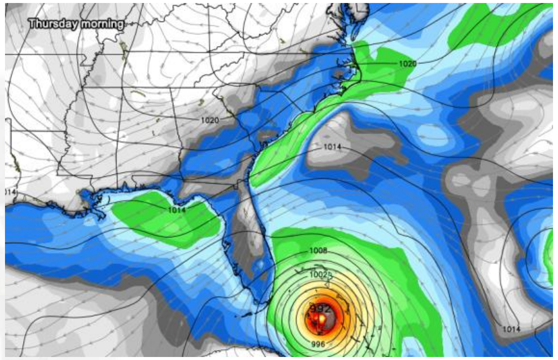
Expandable Storage: 1 TB Expansion Card (matches internal storage exactly)

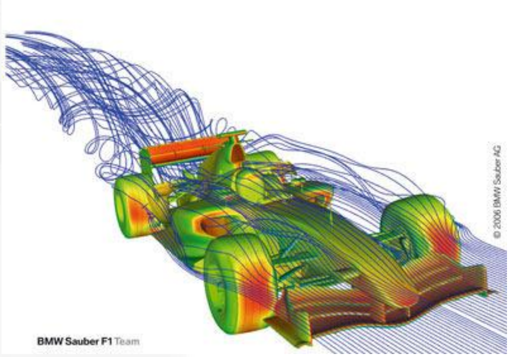
External Storage: USB 3.2 External HDD Support Support

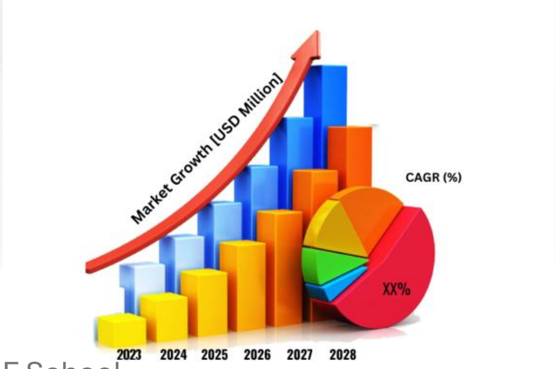
Optical Drive: 4K UHD Blu-Ray Drive

Applications

- Scientific Research
  - Simulations, data analysis, and modeling.
- Weather Forecasting
  - Predicting weather patterns and natural disasters.
- Engineering Simulations
  - Designing and testing new products.
- Financial Modeling
  - Analyzing market trends and risks.

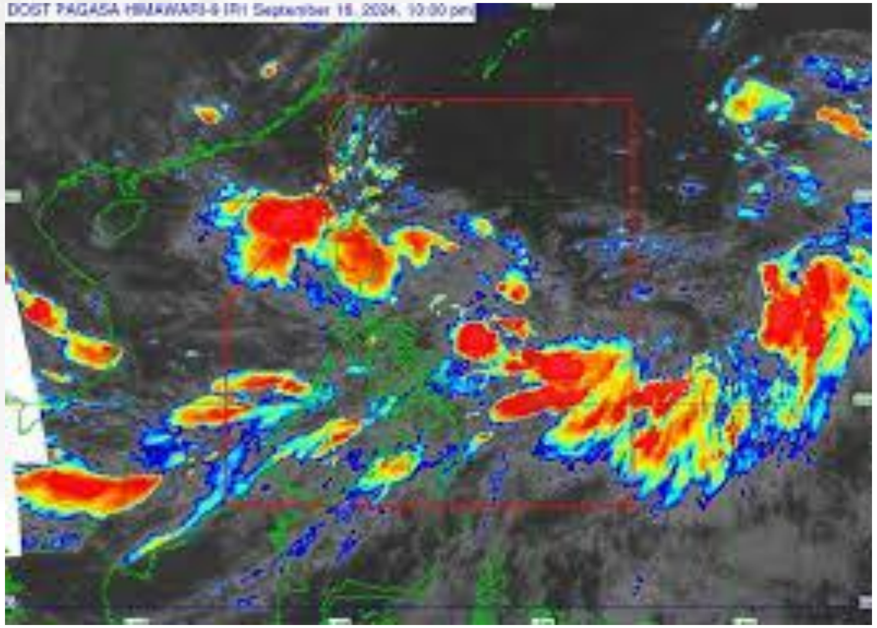






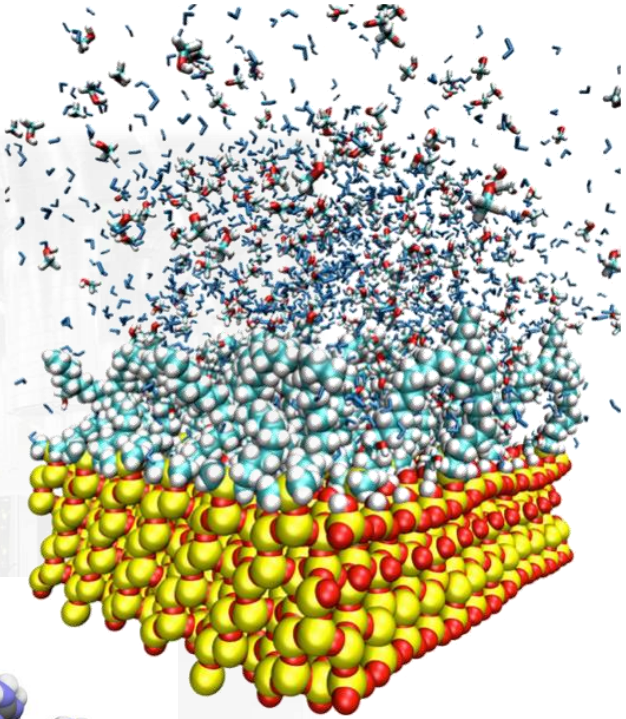
Applications: Weather and Climate

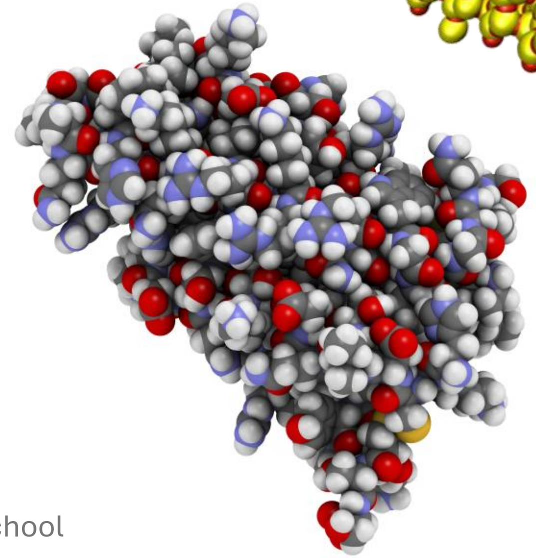
- Massive grid-based simulations
- Need fast compute + huge storage
- Used for forecasting, climate change modeling
- Time-critical and compute-hungry



Applications: Molecular Dynamics

- Simulates protein folding, drugs, viruses
- Used in bio, pharma, and materials science
- GPU acceleration critical for realism
- Popular tools: NAMD, GROMACS, LAMMPS



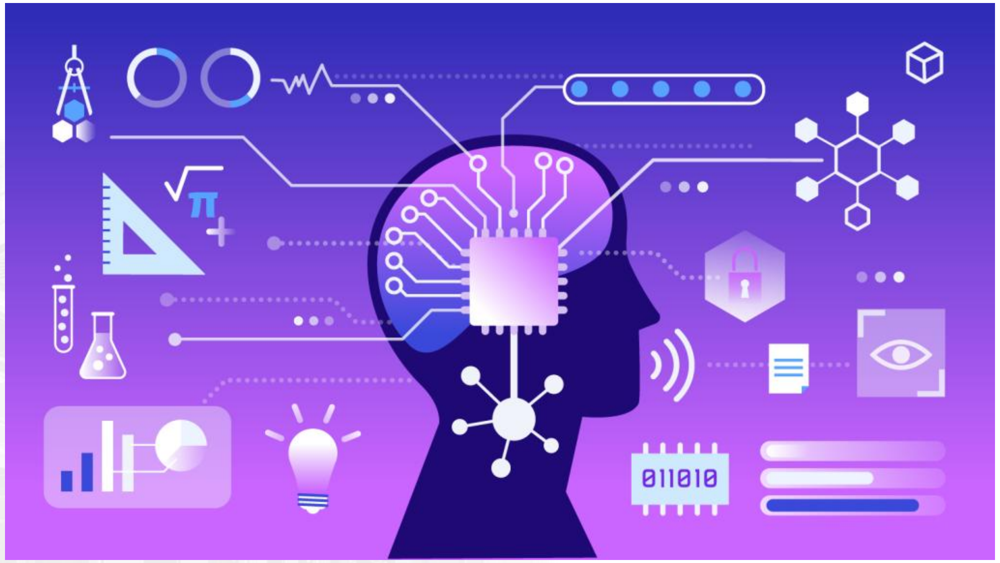


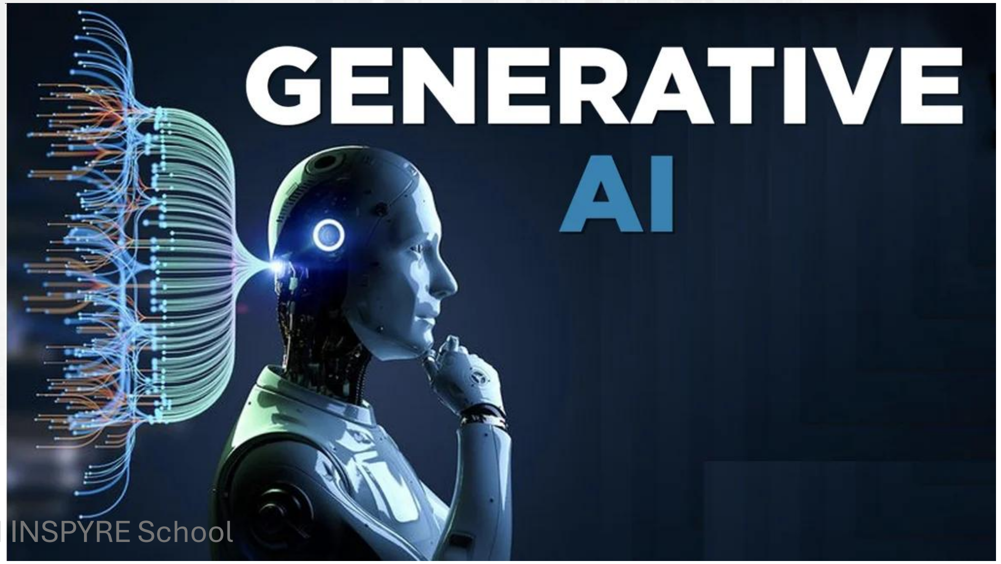
Applications: Astrophysics & Cosmology

- Simulating galaxy formation, dark matter
- Particle-based or grid-based solvers
- Extreme scales: time, space, memory
- Often hybrid CPU-GPU + MPI setups

Applications: AI Meets HPC

- Training large models (e.g., LLMs)
- HPC used for massive matrix multiplications
- GPU clusters lead the way
- AI/HPC convergence is the new norm



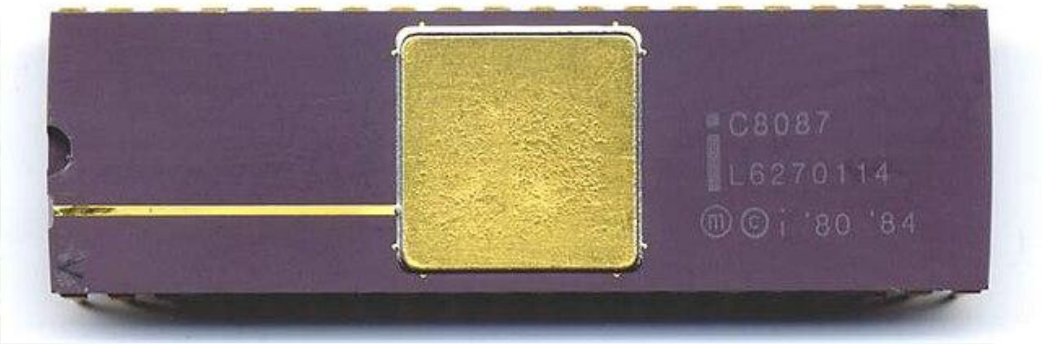


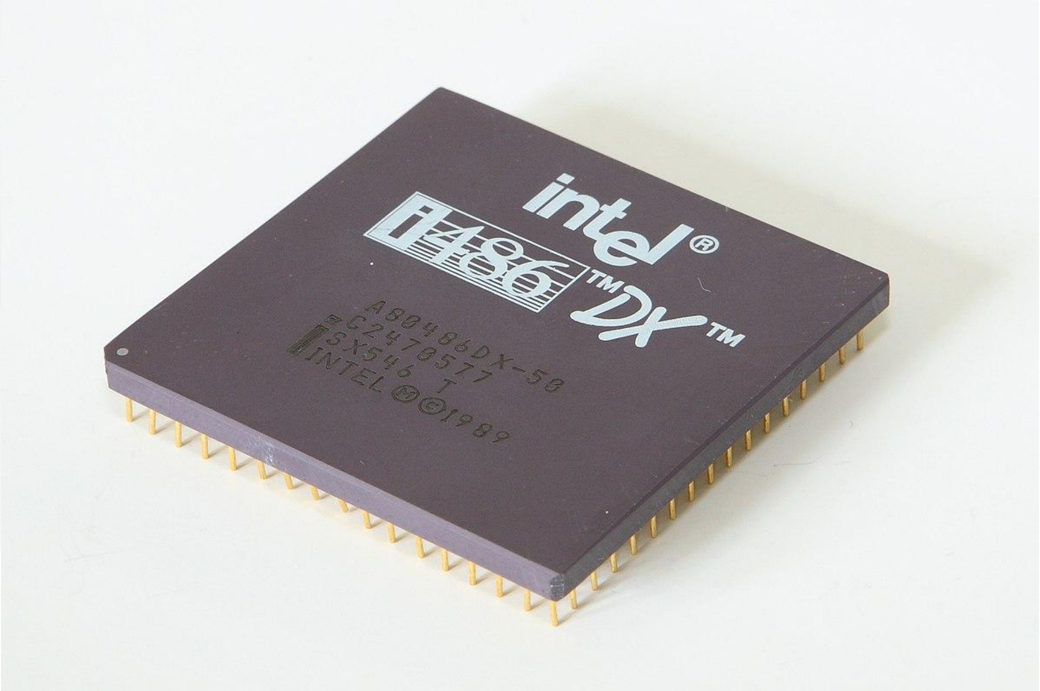
HPC Accelerators

Coprocessors to accelerate FLOPS

- The 8087 was introduced in 1980
  - First x87 floating point coprocessor for the 8086 line of microprocessors
  - Performance enhancements from 20% to 500%, depending to the workload

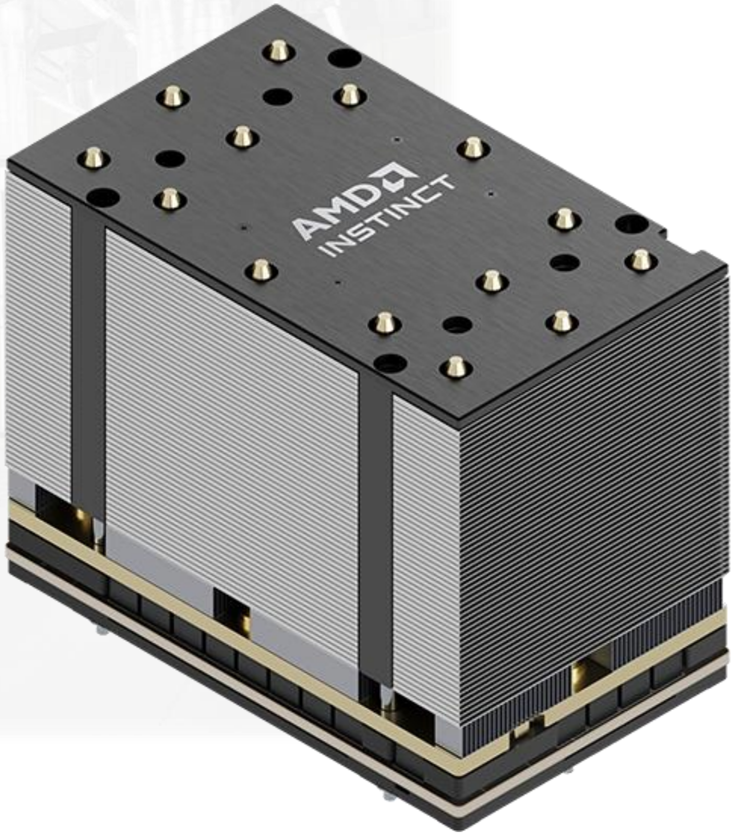
- Intel 80486dx, Pentium and later processors, include FP functionality in the CPU.





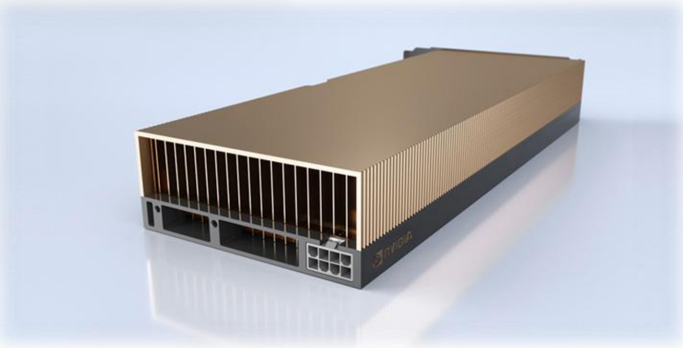
GPUs: Graphics processing Units

- GPUs (Graphics Processing Units) are heavily used in High Performance Computing (HPC) for several key reasons
  - Parallel Processing Capabilities
  - High Computational Throughput
  - Efficiency in Handling Specific Workloads
  - Energy Efficiency
  - Advancements in Software and Infrastructure



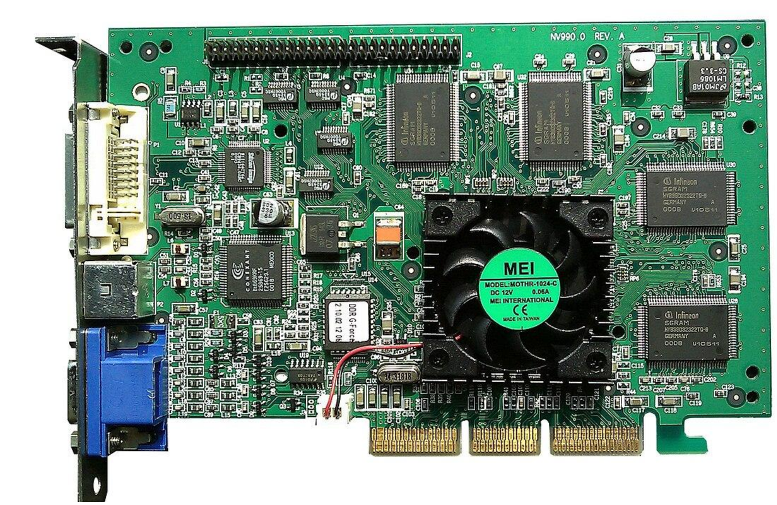
Evolution of the GPU

- 1° generation: Voodoo 3dfx (1996)
- 2° generation: GeForce 256/Radeon 7500 (1998)
- 3° generation: GeForce3/Radeon 8500 (2001)
  - The first GPU to allow limited programmability in vertex pipeline
- 4° generation: Radeon 9700/GeForce FX (2002)
  - First generation of fully programmable graphics cards
- 5° generation: GeForce 8800/HD2090 (2006) and the birth of CUDA



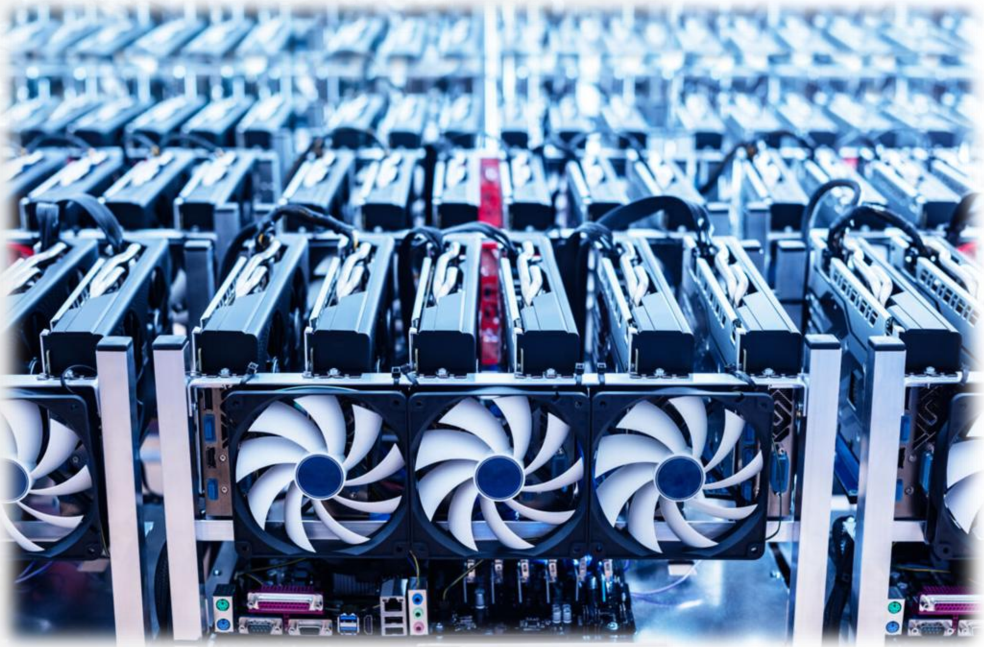
NVIDIA and the GPU Revolution

- Originally a graphics company
- CUDA (2006): unlocked general-purpose computing on GPUs
- From gaming to science: DGX, A100, H100
- Dominating HPC and AI training workloads

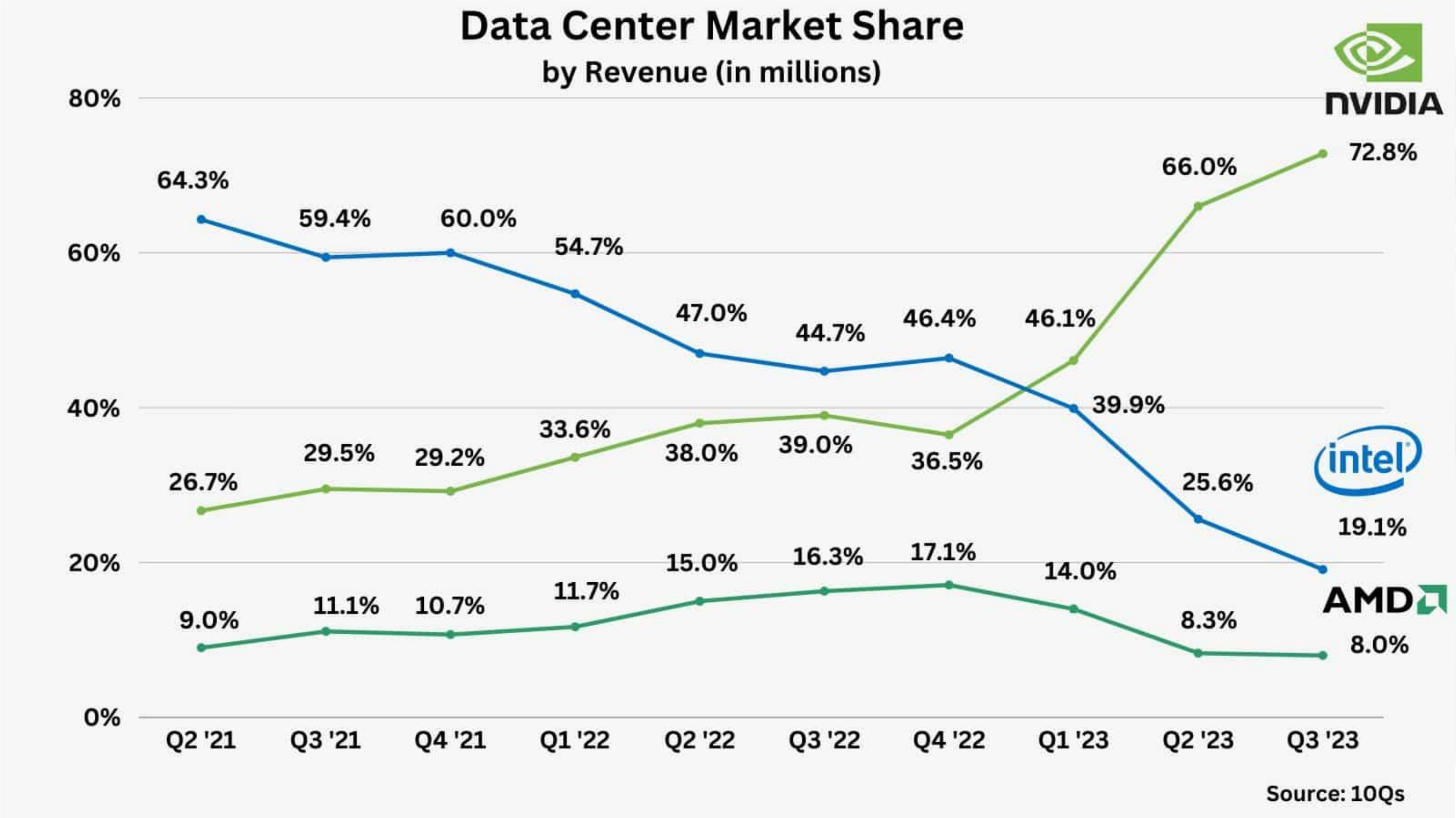


GPUs for mining

- **GPU mining** is the use of Graphics Processing Units (GPUs) to “mine” cryptocurrencies, such as Bitcoin.
- Miners receive rewards for performing computationally intensive work.

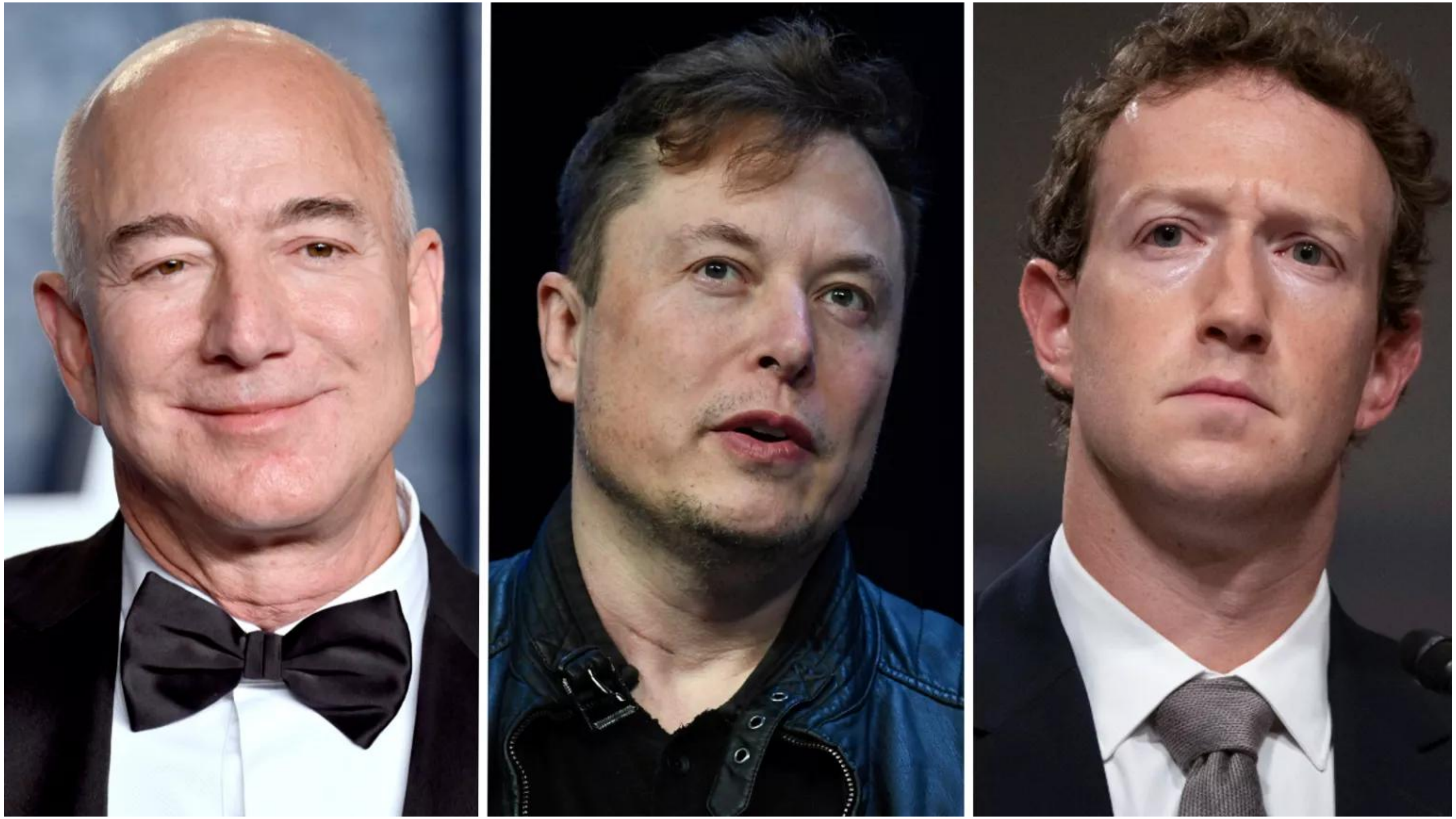


Data center market share



- **In 2024 and 2025, Nvidia performance is even better**

The famous ones



The new guy

Nvidia briefly reached a market capitalization of $4 trillion in Jul 2025, making it the first company in the world to reach the milestone and solidifying its position as one of Wall Street's most-favored stocks.

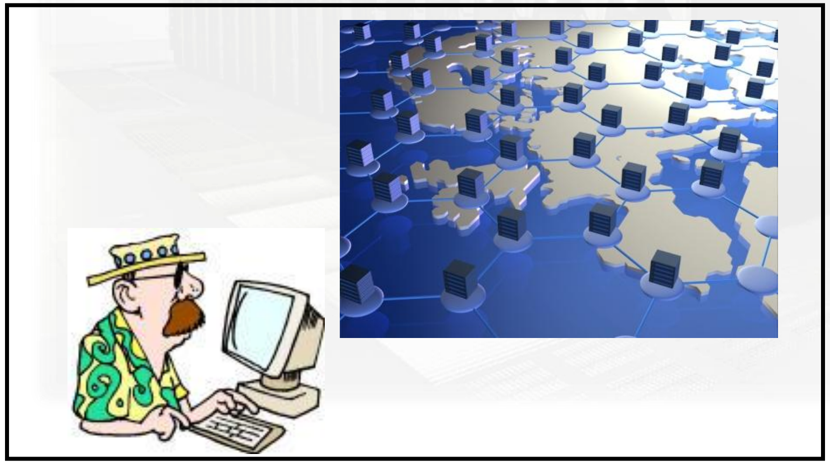
Grids and distributed systems

© Grant Faint

Grid: No centralized control

The user in general has full ownership of a desktop workstation.

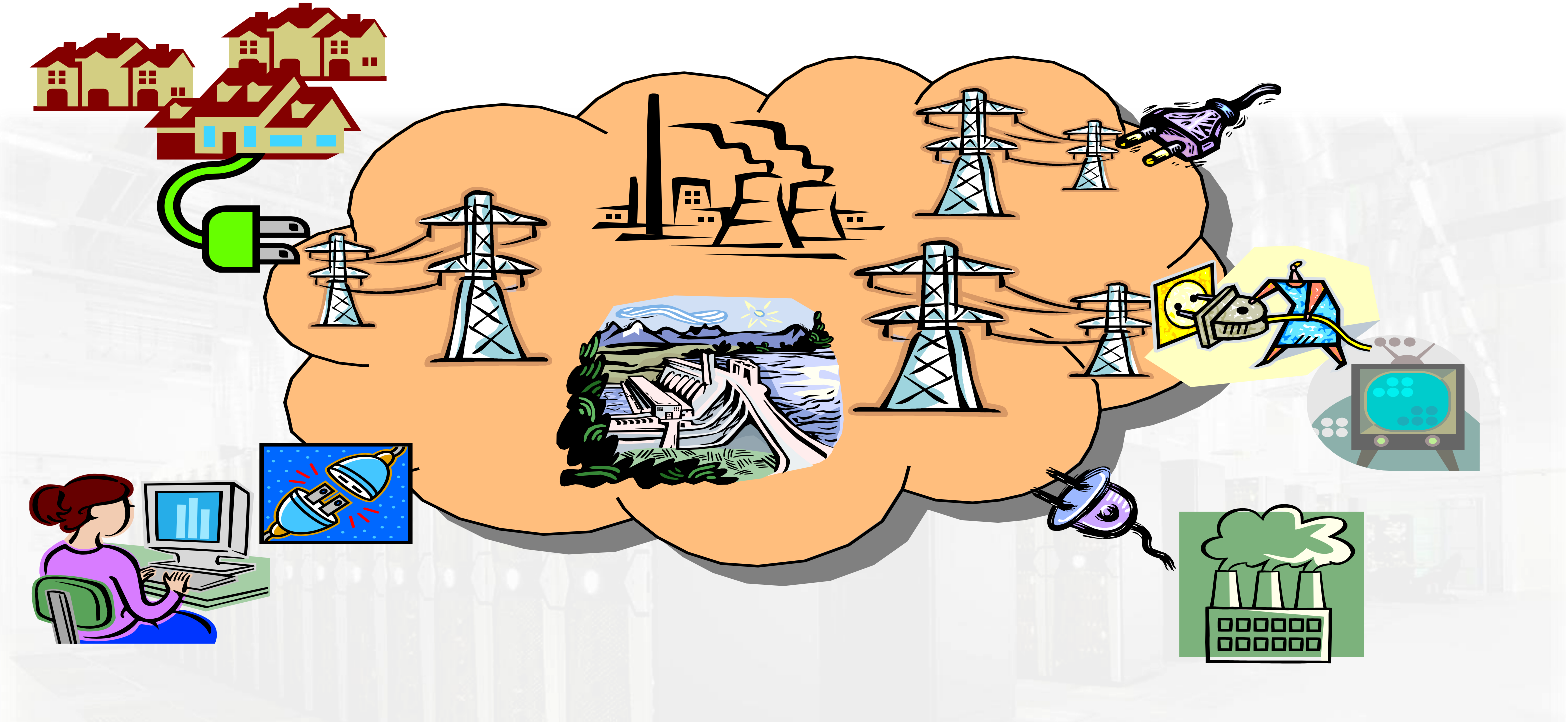
A Cluster is a shared resource – Only the administrator has full control of the system The physical layer is still well defined.



I submit my jobs to “the GRID” and they get processed: somehow, somewhere, after some time.

There is no GRID owner!

Power Grid Similarity



"We will probably see the spread of computer utilities, which, like present electric and telephone utilities, will service individual homes and offices across the country" **(Len Kleinrock, 1969)**

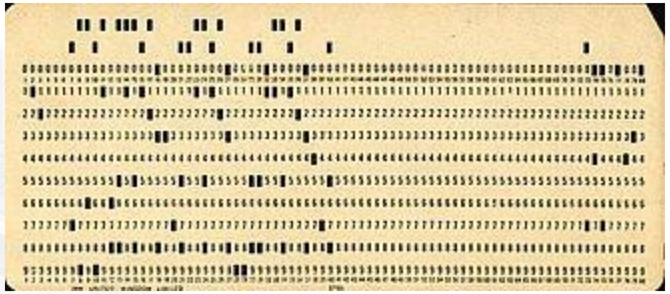
Storage

Evolution of storage

- The evolution of data storage in particle physics reflects a continuous push to handle increasingly large datasets generated by experiments.
- Early methods like punch cards and magnetic tapes gave way to more sophisticated systems like mass storage systems (MSS) with robotic tape libraries and object stores.
- The need for faster data access and analysis has driven the development of optimized data formats, alongside efforts to leverage distributed and cloud-based storage solutions.

Early stages

- Punch cards and magnetic tapes
  - These were the initial methods for storing data, offering limited capacity and requiring physical access for data retrieval.
- Floppy disks
  - These offered slightly improved storage capacity but were still limited and required physical access.
- Local storage
  - Data was primarily stored on local systems, which was manageable for smaller datasets



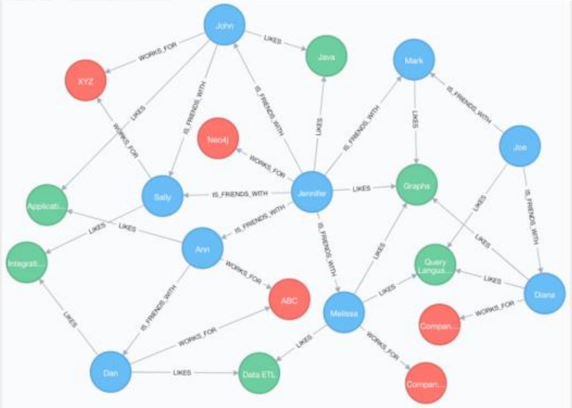
The Rise of Mass Storage Systems

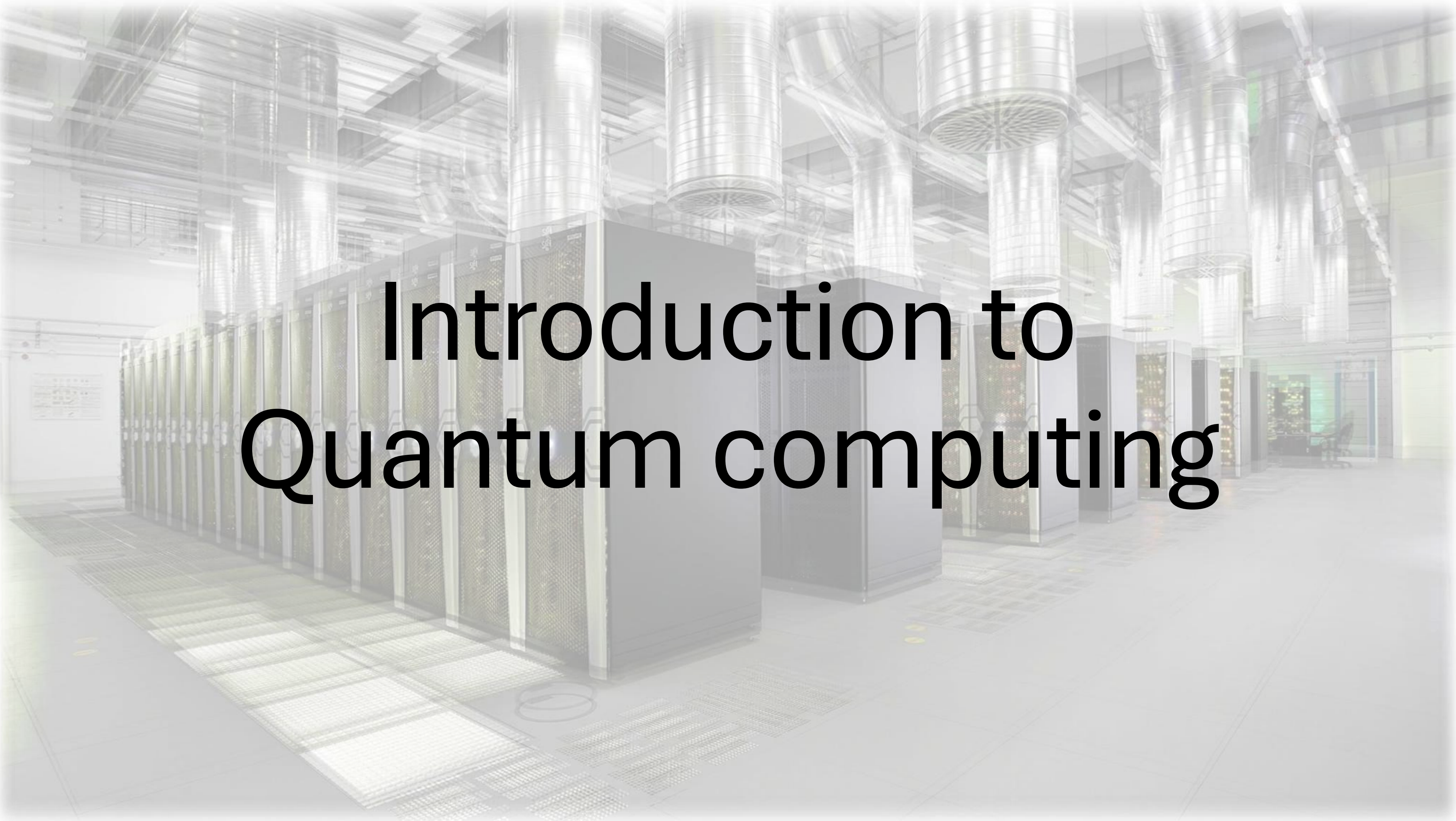
- Robotic Tape Systems
  - As data volumes grew, robotic tape systems became crucial for long-term data storage, especially in high-energy physics.

- Mass Storage Systems (MSS)
  - MSSs manage the robotic tape systems and provide a way to organize and retrieve data from the tapes.

Future Trends

- Cloud Storage
  - Cloud storage is becoming increasingly important for particle physics, offering scalability and flexibility.
- Federated Storage
  - Federated storage solutions are being explored to allow data to be accessed across different storage systems and locations.
- Graph Databases
  - Graph databases are being investigated as a way to represent and analyze complex relationships within particle physics data





Introduction to Quantum computing

What is Quantum Computing?

- Quantum computing is a computational paradigm that leverages quantum mechanical principles to process information in fundamentally different ways from classical computers.
- Key Difference
  - Classical computer: processes bits (0 or 1)
  - Quantum computer: processes qubits (0, 1, or both simultaneously)
- Why is it Important?
- Potential to solve problems that are computationally intractable for classical computers.

Qubits - The Basic Unit

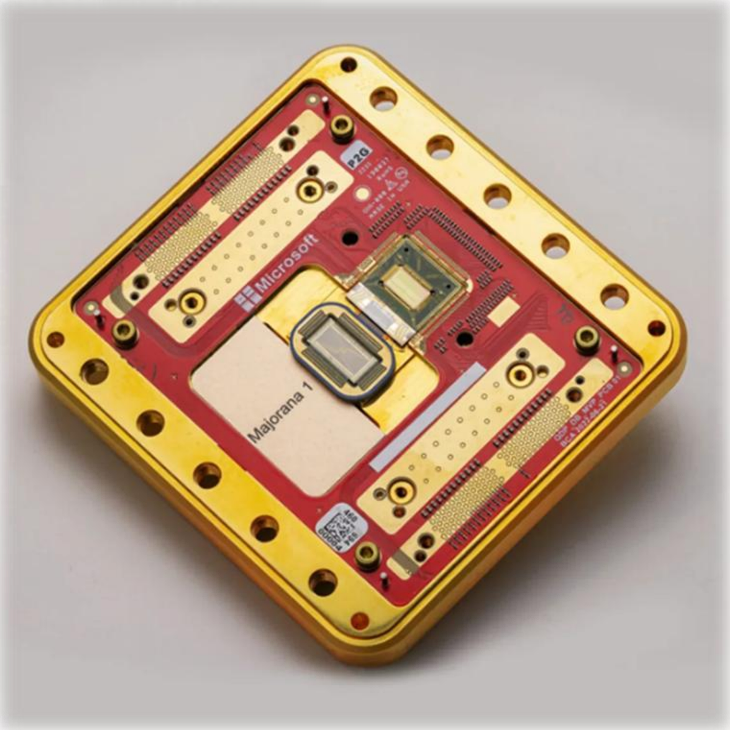
- Classical Bit vs Qubit
  - Classical bit: $|0\rangle$ or $|1\rangle$
  - Qubit: $\alpha|0\rangle + \beta|1\rangle$
- **Fundamental Properties**
  - Superposition: can be in both states simultaneously
  - Probability: $|\alpha|^2 + |\beta|^2 = 1$
  - Measurement: collapses to $|0\rangle$ or $|1\rangle$ with probabilities $|\alpha|^2$ and $|\beta|^2$
- A qubit can represent all possible combinations until measurement.

Key Quantum Principles

- **Superposition**
  - A qubit can be in multiple states simultaneously
  - N qubits can represent 2^N states simultaneously
  - **Example:** 3 qubits = 8 classical states represented together
- **Entanglement**
  - Quantum correlation between qubits
  - Measuring one qubit instantly affects the other
  - Foundation for many quantum algorithms
- **Interference**
  - Quantum states can interfere constructively or destructively
  - Allows amplifying correct solutions and canceling wrong ones

Limitations and Challenges

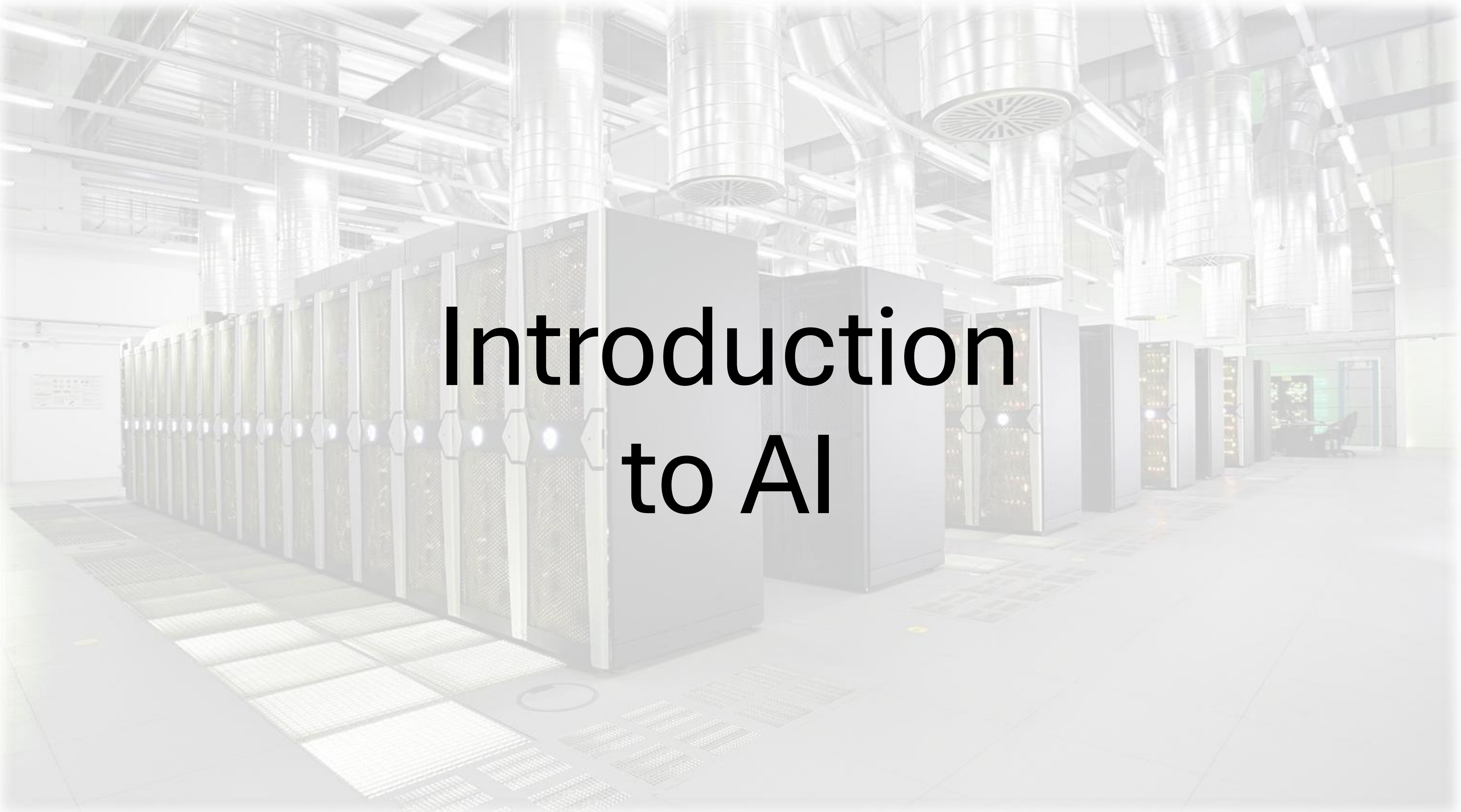
- **Technical Problems**
  - Decoherence: Qubits lose quantum properties quickly
  - Errors: High error rates (~0.1-1%)
  - Control: Difficulty in precise control
- **Engineering Challenges**
  - Environmental isolation
  - Quantum error correction
  - Scaling to thousands/millions of qubits
- **Theoretical Limits**
  - Not all problems benefit from quantum speedup
  - Some problems remain intractable



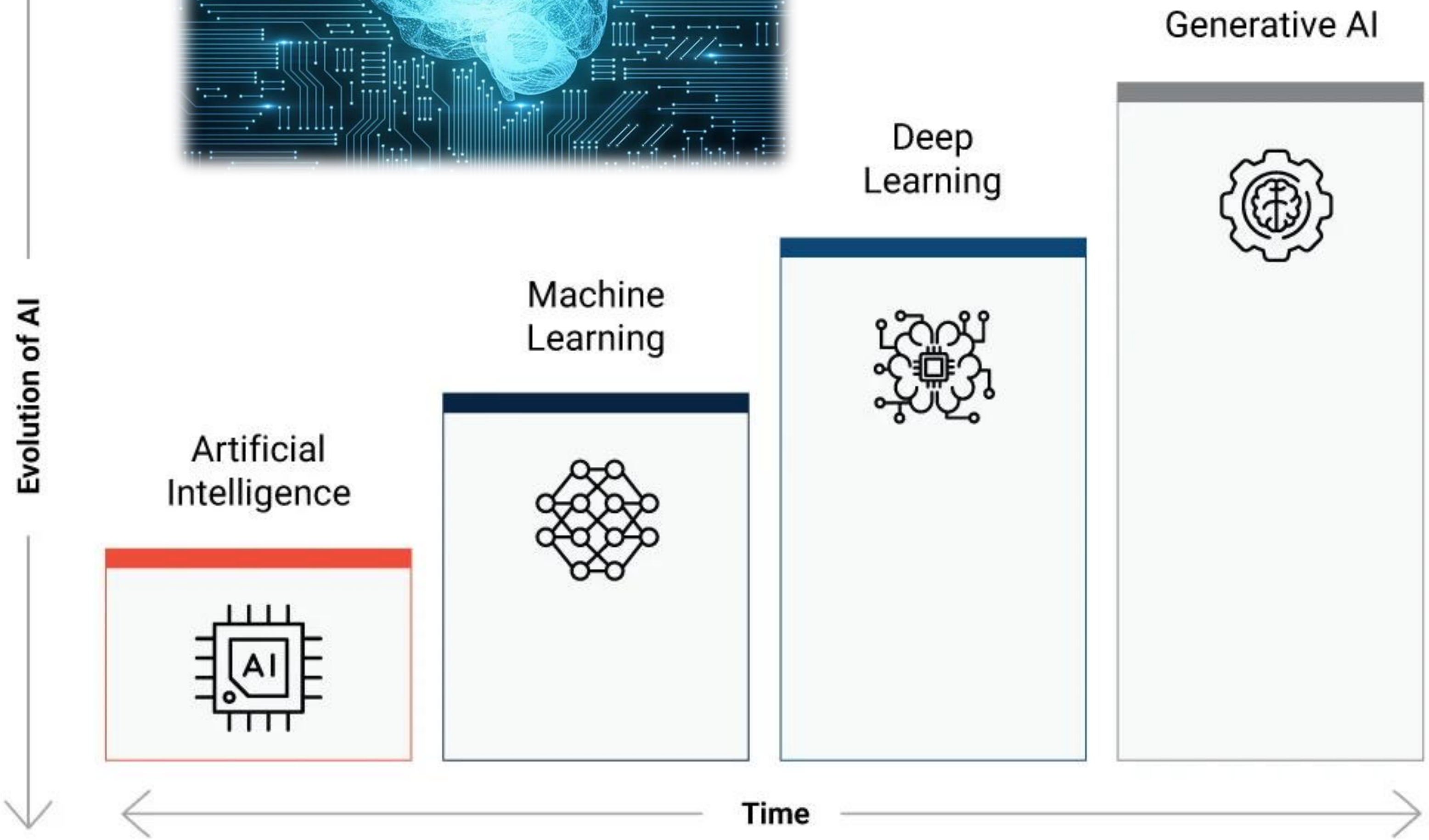
Future Applications in HPC

- Promising Sectors
  - Molecular Simulation: Drug discovery, Catalysis, Advanced materials
  - Optimization: Logistics, Portfolio optimization, Traffic flow
  - Machine Learning: Quantum neural networks, Pattern recognition, Feature mapping
  - Cryptography: Post-quantum cryptography, Quantum key distribution

Introduction to AI



Evolution of AI



Artificial Intelligence

- AI is gaining mass interest thanks to latest development in generative AI
- Most AI is built on the analysis of big data sets that contain too much information for any human to analyze on their own in a reasonable time.
- An AI model is built to identify patterns in those data-sets and then use those patterns to predict future or additional patterns.
- AI models use probability and statistical analysis in order to do so.
  - Some AI models are good enough at this to mimic human behaviors.

Machine Learning

- Machine learning is a branch of AI; it refers to the practice of feeding a program structured or labeled data in order to train the program how to identify that data without human intervention.
  - For example, a machine learning model for finding bottles of ketchup in photos of open refrigerators may start out unable to identify any condiments, let alone ketchup.
  - It is then fed millions of images of ketchup bottles in various refrigerators and is told that each one represents a ketchup bottle.
  - Eventually, it is able to automatically identify ketchup bottles even in photos it has never seen before.

Deep Learning (DL)

- Deep learning is a type of machine learning.
- Deep learning models are able to use probabilistic analysis to identify differences in raw data.
- A deep learning model could potentially learn what a bottle of ketchup is and how to distinguish it from other condiments from photos of open refrigerators alone, without being told what a bottle of ketchup is.
- Like other types of machine learning, deep learning requires access to large data sets. Even an advanced deep learning model would probably need to analyze millions of photos of open refrigerators to be able to identify ketchup.

Generative AI

- Generative AI is a type of AI model that can create content, including text, images, audio, and video.
- A generative AI model could, for example, receive a photo of an empty refrigerator and populate it with probable contents, based on photos it has been shown in the past.
  - While the content generated by such a model may be considered “new”, it is based on content that the model has been previously fed.
- Generative AI tools are increasingly popular. In particular, the large language model (LLM) ChatGPT, image generators DALL-E and Midjourney have captured the public's imagination and the business world's attention.
  - Other popular generative AI tools include Bard, Bing Chat, and Llama.

What is the meaning of GPT in chatGPT?

- Can someone answer?

Some key terms: training

- Training is the process of teaching an AI model how to perform a given task
  - Training is the first phase for an AI model
  - Training may involve a process of trial and error, or a process of showing the model examples of the desired inputs and outputs, or both.
  - Training an AI model can be very expensive in terms of compute power. But it is more or less a one-time expense.
    - Involves feeding AI models large data sets
  - Once a model is properly trained, it ideally does not need to be trained further.

Some key terms: inference

- Inference is the AI model in action, drawing its own conclusions without human intervention.
  - Almost any real-world application of AI relies on AI inference
  - Inference is ongoing. If a model is actively in use, it is constantly applying its training to new data and making additional inferences.
  - This takes quite a bit of compute power and can be very expensive.

Some key terms: tokens, parameters

- Tokens represent the smallest units of data that the model processes, such as words or characters in natural language processing.
- Parameters are variables within a model that dictate how it behaves and what results it produces.

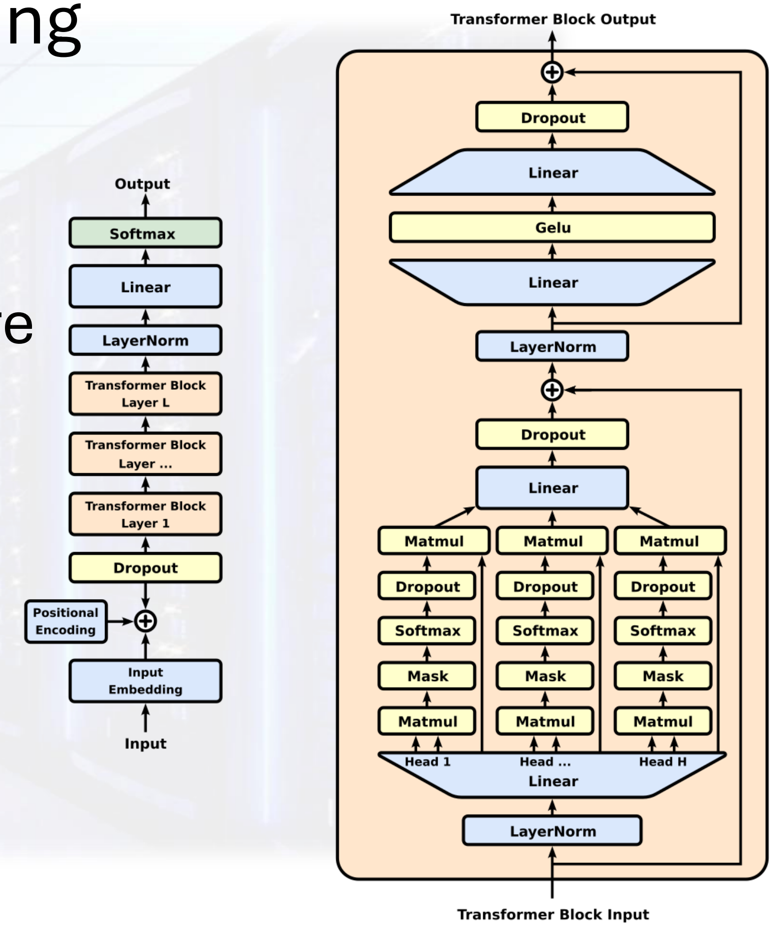
Some key terms: LLM

- A large language model (LLM) is a type of artificial intelligence (AI) program that can recognize and generate text, among other tasks.
- LLMs are trained on huge sets of data — hence the name "large"
- LLMs are built on machine learning: specifically, a type of neural network called a transformer model.
    - LLM is a computer program that has been fed enough examples to be able to recognize and interpret human language or other types of complex data.

What is the meaning of GPT in chatGPT?

- Now we can answer and understand the meaning

- Generative Pre-trained Transformer
  - based on the transformer deep learning architecture
  - pre-trained on large data sets of unlabeled text
  - able to generate novel human-like content



What hardware is required

- For ML/AI you just need a high compute processor with sufficient ram for your target dataset.
- Deep learning, on the other hand, is large scale training of million of parameters.
  - This is done via matrix calculations. **GPUs are specialized** for matrix calculations: the speed up is significant.
  - Modern deep learning was a lost cause before GPU adaptation.
  - Further, you can run multiple GPUs in tandem, allowing you to parallel train models.
- This has led most neural libraries to optimize for GPU based training.
  - On top of GPUs having significant speedups, most library optimization has GPUs in mind.
- You can perform inference with just a CPU, but at best you'll probably have a 2.5x slowdown than when you used a GPU

An example: llama3.1

- Meta has recently unveiled Llama 3.1, its most advanced open-source AI Model to date.
- This model stands out due to its 405 billion parameters, making it the largest open-source AI Model available
- The training process for Llama 3.1 leveraged over 16,000 Nvidia H100 GPUs
  - Llama 3.1 brings context window to 128k tokens
    - context window: the amount of text that can be reasoned about at once

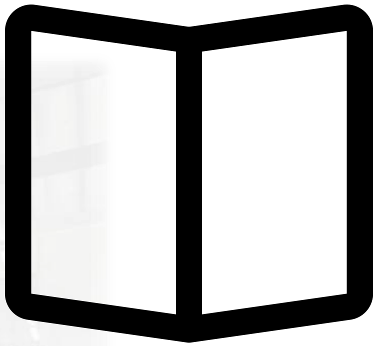
Summary and take-aways

Take aways

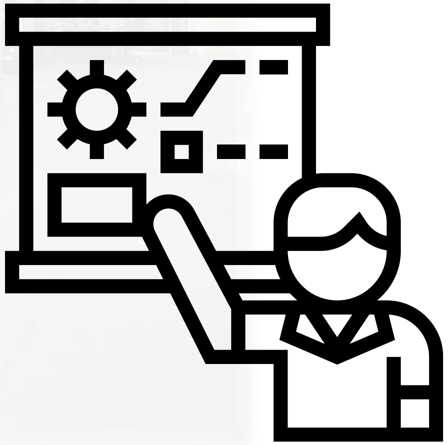
- Computing is fundamental for today's physics experiments
  - Storage is a fundamental part of modern computing
- Different applications have different computing needs that can be mapped on different computing infrastructures
  - HPC → High Performance Computing → Supercomputers
  - HTC → High Throughput Computing → Grids
- Federation of Computing and Storage are needed to address the extreme- scale experiments requirements

Further reading and resources

- top500.org and green500.org websites



- HPCwire and insideHPC news sites
- Courses: PRACE, EuroHPC



- CINECA training