High Performance Computing

Teachers: Daniele De Sensi, Salvatore Pontarelli

Information

E-mails:

desensi@di.uniroma1.it
salvatore.pontarelli@uniroma1.it

Class schedule:

- Monday 4pm-7pm
- Thursday 5pm-7pm

Google Classroom:

zx5soohr

Syllabus (1/2)

Part 1: Single-Node HPC ***[Pontarelli]***

- Introduction to HPC
- Modern CPU Architectures for HPC
- Vector processors
- High performance Programming
- Application Profiling
- Hardware Accelerators

Syllabus (2/2)

Part 2: Large-Scale HPC ***[De Sensi]***

- Introduction to scale-up and scale-out networks
- High-Performance Network Protocols (RDMA)
- Advanced GPU Programming
- Collective Communications in HPC
- Use cases:
  - High-Performance Linear Algebra
  - Distributed Deep Learning Training
  - CFD (Computational Fluid Dynamics)

Prerequisites

- Good knowledge of the C/C++ programming language

- Course *Programmazione dei sistemi embedded e multicore [can be added to the study plan]*

Key Applications of HPC

- **Scientific Research:** Climate modelling, astrophysics, material science.

- **Engineering:** Computational fluid dynamics, structural analysis.

- **Healthcare & Medicine:** Genomic sequencing, drug discovery.

- **Artificial Intelligence & Machine Learning:** Training large-scale models.

Simulation: The Third Pillar of Science

- Traditional scientific and engineering paradigm:

  1) Do theory or paper design.

  2) Perform experiments or build system.

- Limitations:

  Ø Too difficult -- build large wind tunnels.

  Ø Too expensive – to experiment with birds in a

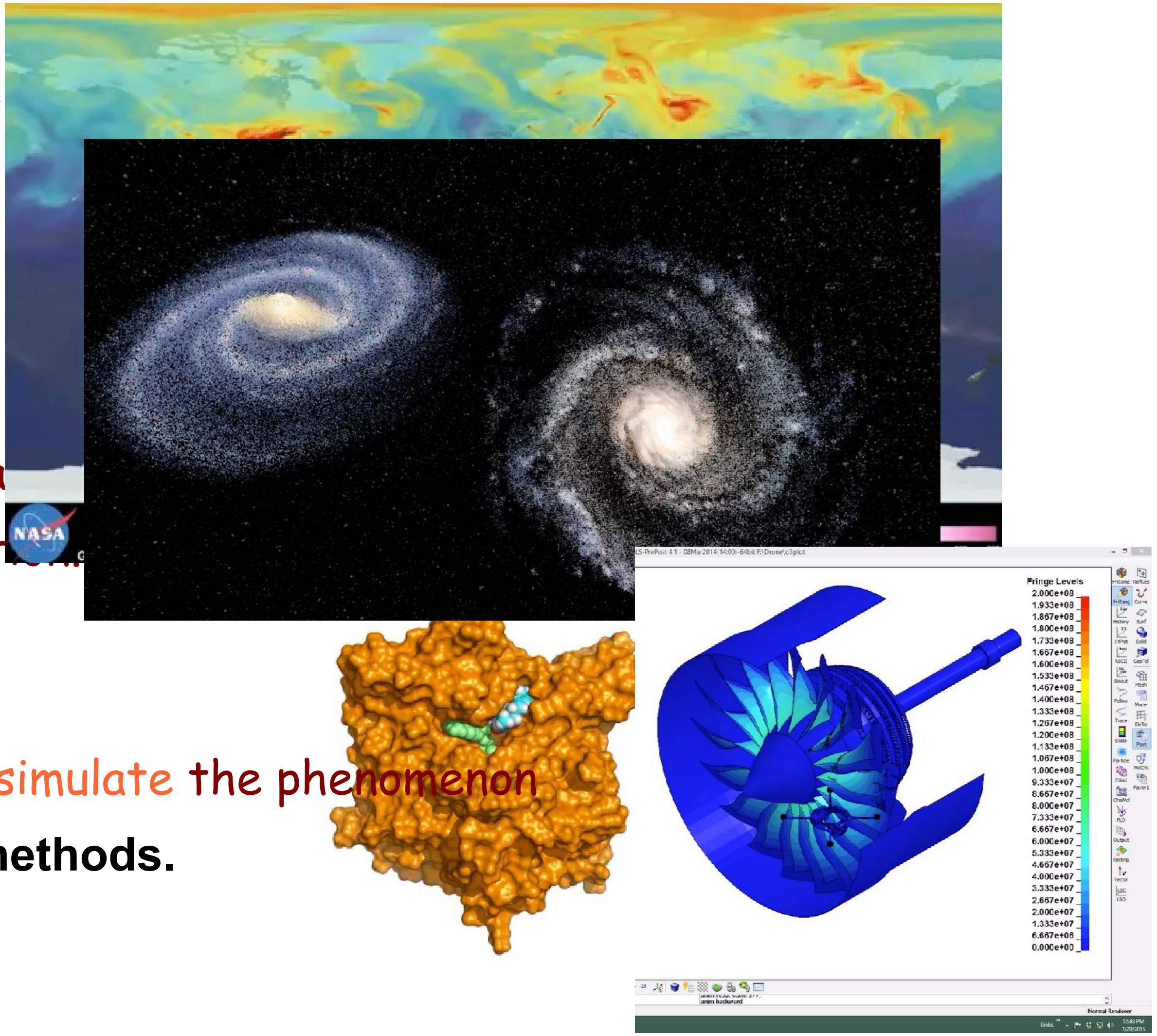
  Ø Too slow -- wait for climate or galactic evolution.

  Ø Too dangerous -- weapons, drug design.

- Computational science paradigm:

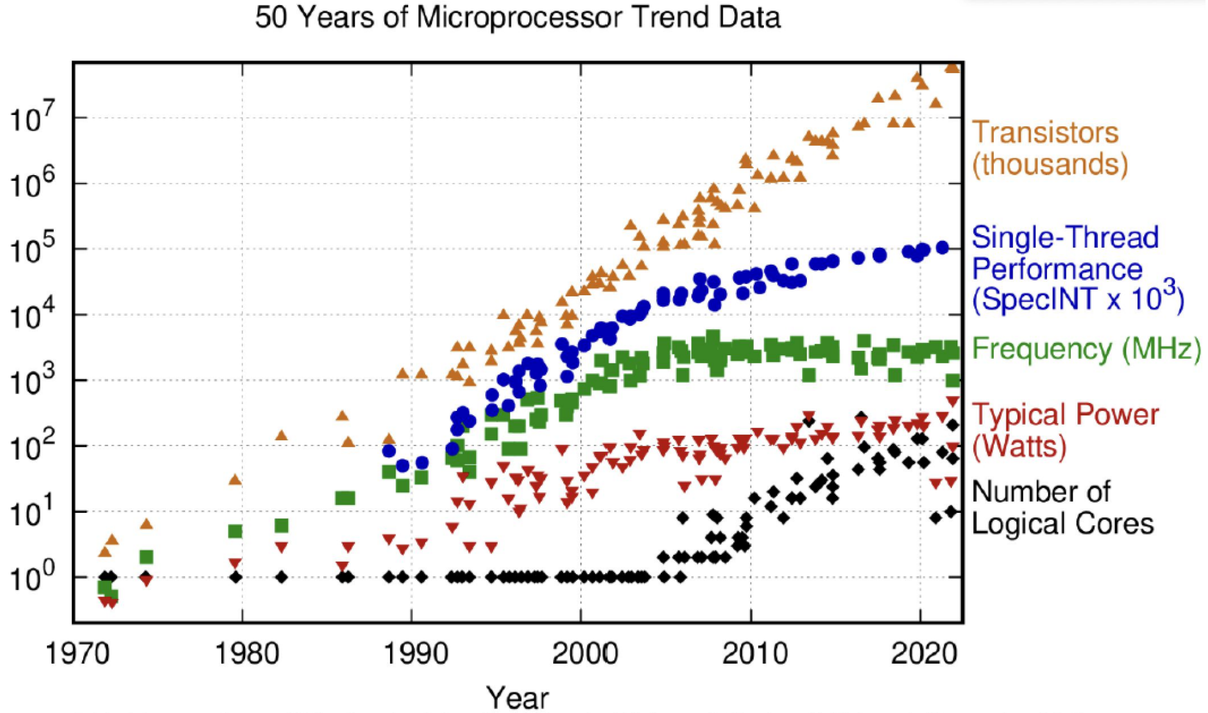
  3) Use high performance computer systems to simulate the phenomenon

  Base on known physical laws and numerical methods.



Moore's law

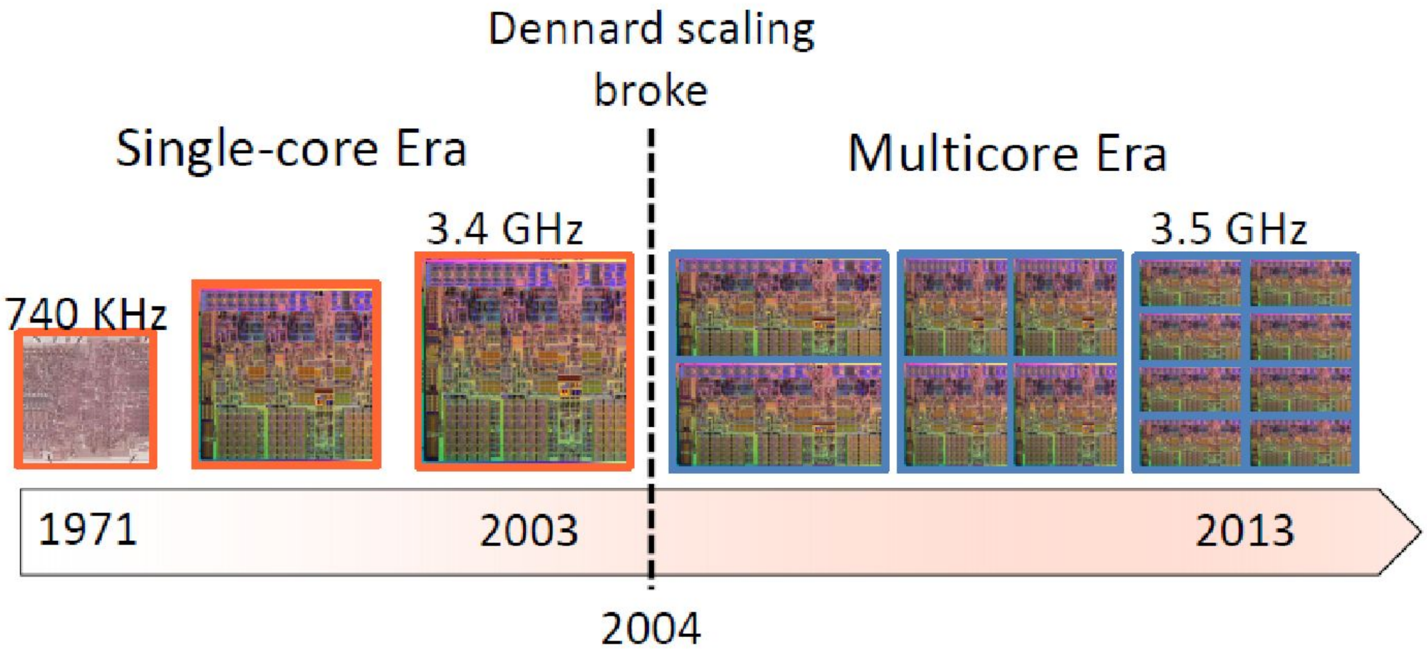
- In the past decade, the world has experienced one of the most exciting periods in computer development.
- Microprocessors have become smaller, denser, and more powerful.
- The result is that microprocessor-based supercomputing is rapidly becoming the technology of preference in attacking some of the most important problems of science and engineering.



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 2010-2021 by K. Rupp

Dennard Scaling Over

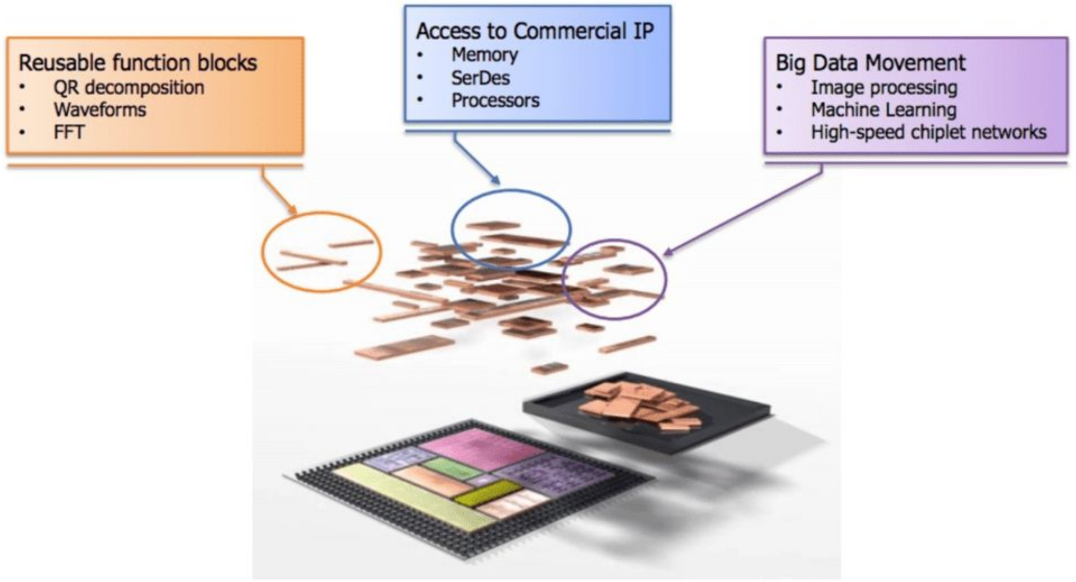
- Voltages couldn’t be further reduced. Why? Threshold voltage of transistors and noise.
- End of Increasing Frequency
- Slowing Improvement Single Thread (core) Performance
  - Sequential Abstraction
  - The move to parallelism



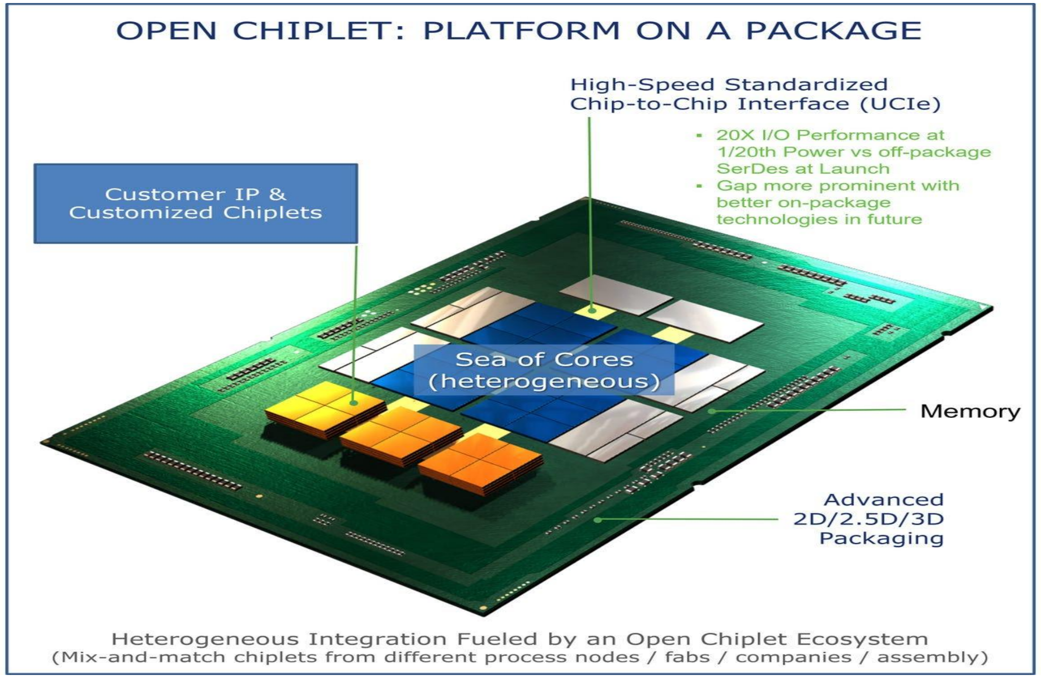
Chiplets: Integrating Multiple Functions

- Rather than fabricating a monolithic system-on-a-chip, chiplet technology combines multiple chips, each representing a portion of the desired functionality, possibly fabricated using different processes by different vendors and perhaps including IP from multiple sources

- Chiplet designs are part of the recent offerings from Intel and AMD
  - Amazon’s Graviton3 also uses a chiplet design with seven different chip dies
  - Advanced Query Accelerator (AQUA) for AWS Redshift, Amazon’s powerful and popular data warehouse service, relies on a package of custom ASICs and FPGA accelerators

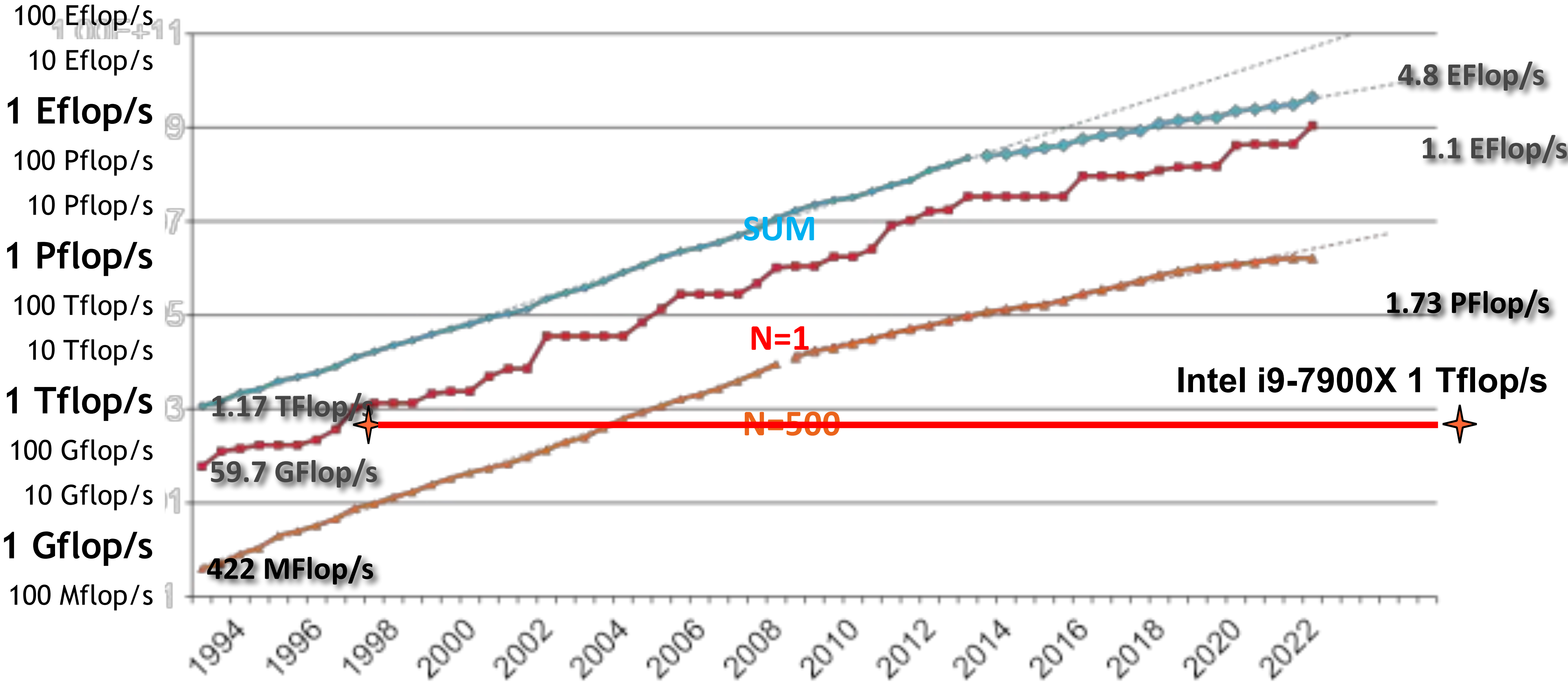


CHIPS modularity targets the enabling of a wide range of custom solutions


OPEN CHIPLET: PLATFORM ON A PACKAGE
High-Speed Standardized Chip-to-Chip Interface (UCIe)
20X I/O Performance at 1/20th Power vs off-package SerDes at Launch
Gap more prominent with better on-package technologies in future
Customer IP & Customized Chiplets
Sea of Cores (heterogeneous)
Memory
Advanced 2D/2.5D/3D Packaging

Heterogeneous Integration Fueled by an Open Chiplet Ecosystem
(Mix-and-match chiplets from different process nodes / fabs / companies / assembly)

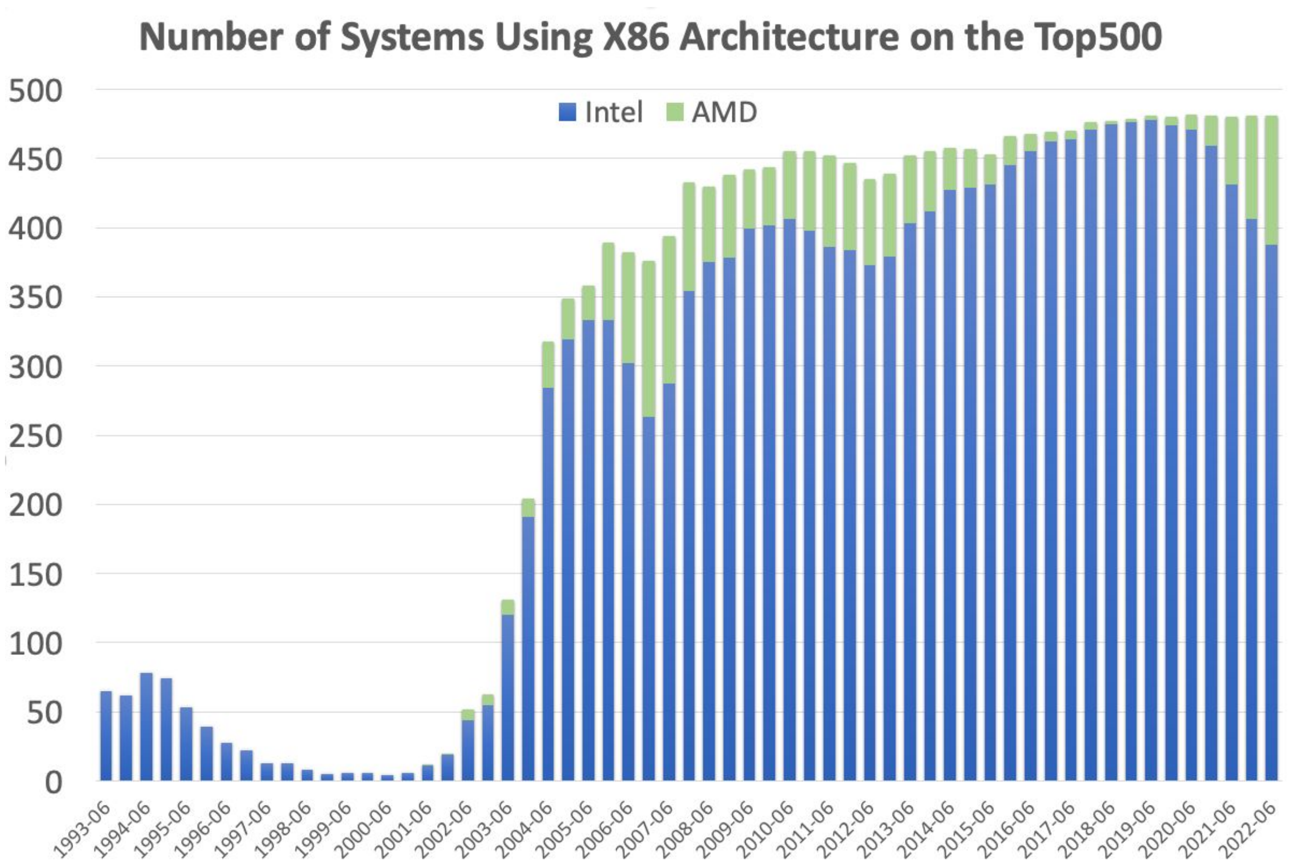
Performance Development of HPC over the Last 30 Years from the Top500



N.B. performance are computed on the HPL benchmark (solve a large system of linear equations) using FP64 / double-precision

HPC: the TOP500 list

- TOP500 list began in 1993
  - 65 systems used Intel’s i860 architecture
  - Remainder had specialized architectures, mainly vector based
- Most recent TOP500 list
  - 78% of systems used Intel processors
  - Another 19% used AMD processors
- **97% of the systems use x86-64 architecture**
    - Many use GPU accelerators



HPC: the TOP500 list

First phase (specialized architectures):

- Based on monolithic supercomputers. These were extremely expensive, custom-built machines, often from companies like Cray and Control Data Corporation (CDC).T hese early supercomputers relied on a small number of highly powerful, specialized vector processors. Due to the limited number of processors, they typically utilized a shared memory architecture.

Second phase (microprocessor revolution):

- Commodity microprocessors to eventually surpass the performance of their specialized counterparts for many HPC workloads. In 1994 the **Beowulf cluster** shown how to cluster multiple, inexpensive personal computers running on the Linux operating system to work in parallel as a single, powerful machine.

HPC: the TOP500 list

Third phase (accelerators):

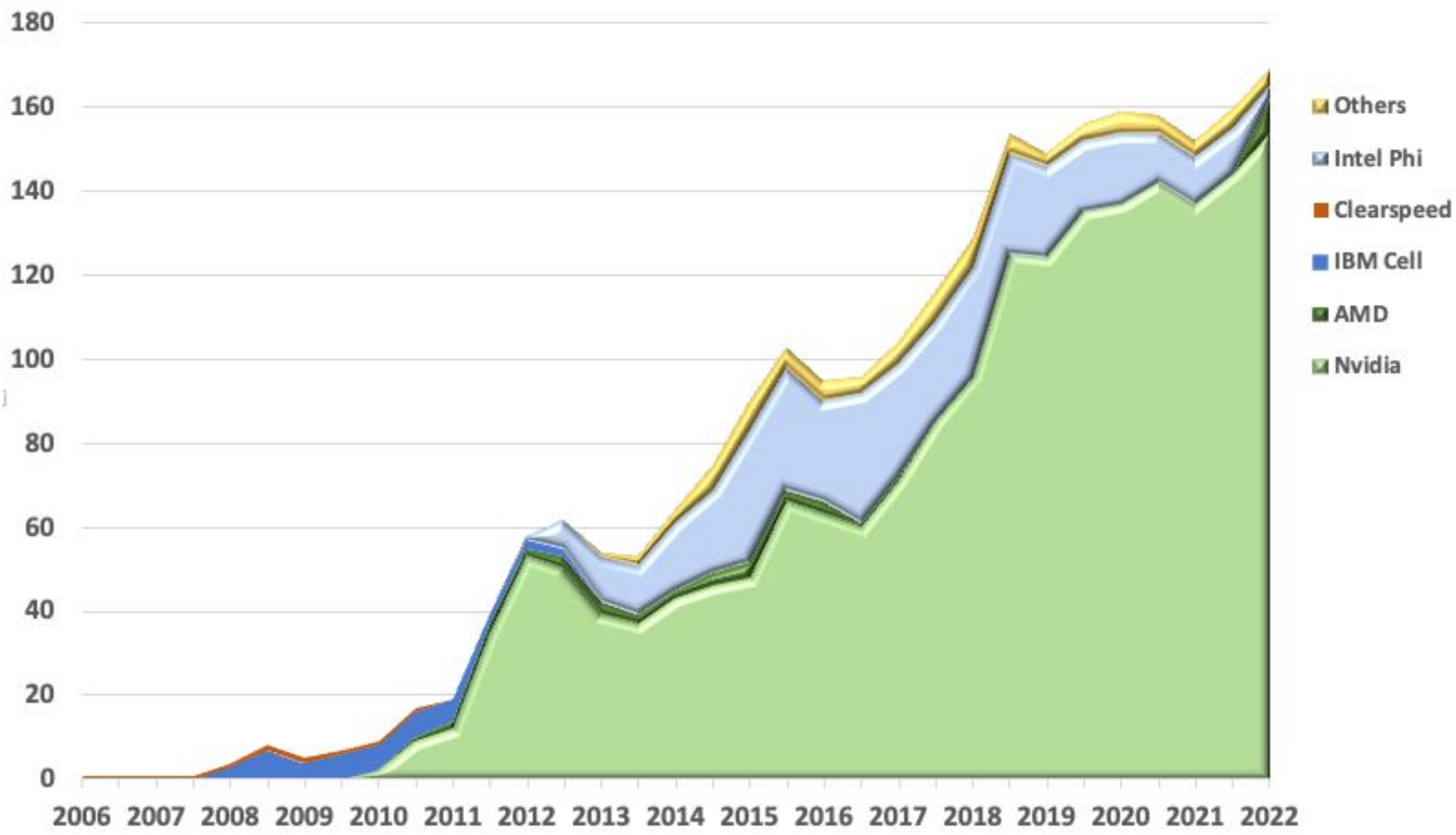
- Now we moved to heterogeneous architectures, where traditional general-purpose CPUs are augmented by specialized accelerators. The most dominant and transformative of these accelerators are Graphics Processing Units (GPUs), which have fundamentally reshaped the architecture and capabilities of modern supercomputers.

Four phase (the future):

- Come back to specialized architectures
- Even more heterogenous nodes: neuromorphic, quantum, optical, AI/ML
- Composable Disaggregated Infrastructure

HPC: Accelerators/Interconnects/OS

- Nvidia dominates accelerators
- Interconnects are mainly Ethernet/InfiniBand
  - 426 of the Top500
- Linux is standard everywhere

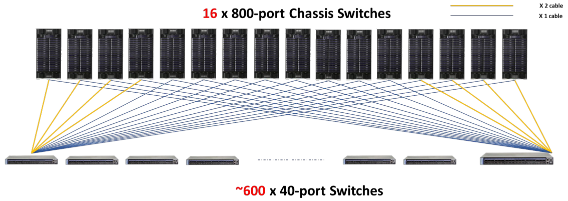


Today's HPC Environment for Scientific Computing

ORNL Frontier, 2 Eflop/s,
8.8 x 10^6 Cores, 9408 nodes, 30 MW
(node = 1-AMD CPU + 4-AMD GPUs)
> 98% of performance from GPUs

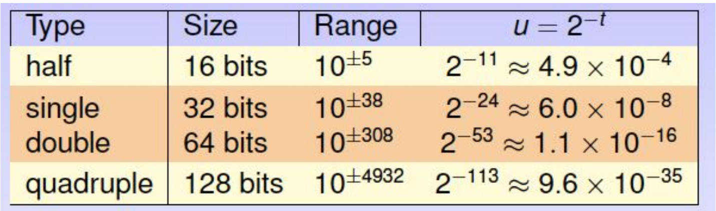
- Highly parallel
  - Distributed memory
  - MPI + Open-MP programming model

- Heterogeneous
  - Commodity processors + GPU accelerators



- Communication between parts very expensive compared to floating point ops

- Floating point hardware at 64, 32, 16, & 8 bit levels



Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

November 2022: The TOP 10 Systems (53% of the Total Performance of Top500)

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	DOE / OS Oak Ridge Nat Lab	Frontier, HPE Cray Ex235a, AMD 3rd EPYC 64C, 2 GHz, AMD Instinct MI250X, Slingshot 10	USA	7,733,248	1,102	65	21.1	52.2
2	RIKEN Center for Computational Science	Fugaku, ARM A64FX (48C, 2.2 GHz), Tofu D Interconnect	Japan	7,299,072	442.	82	29.9	14.8
3	EuroHPC /CSC	LUMI, HPE Cray EX235a, AMD 3rd EPYC 64C, 2 GHz, AMD Instinct MI250X, Slingshot 10	Finland	1,268,736	304.	72	2.94	52.3
4	EuroHPC/CINECA	BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 (108C), Quad-rail NVIDIA HDR100	Italy	1,463,616	175.	68	5.6	31.1
5	DOE / OS Oak Ridge Nat Lab	Summit, IBM Power 9 (22C, 3.0 GHz), NVIDIA GV100 (80C), Mellonox EDR	USA	2,397,824	149.	74	10.1	14.7
6	DOE / NNSA L Livermore Nat Lab	Sierra, IBM Power 9 (22C, 3.1 GHz), NVIDIA GV100 (80C), Mellonox EDR	USA	1,572,480	94.6	75	7.44	12.7
7	National Super Computer Center in Wuxi	Sunway TaihuLight, SW26010 (260C), Custom Interconnect	China	10,649,000	93.0	74	15.4	6.05
8	DOE / OS NERSC - LBNL	Perlmutter HPE Cray EX235n, AMD EPYC 64C 2.45GHz, NVIDIA A100, Slingshot 10	USA	706,304	64.6	71	2.59	27.4
9	NVIDIA Corporation	Selene NVIDIA DGX A100, AMD EPYC 7742 (64C, 2.25GHz), NVIDIA A100 (108C), Mellanox HDR	USA	555,520	63.4	80	2.64	23.9
10	National Super Computer Center in Guangzhou	Tianhe-2A NUDT, Xeon (12C) , MATRIX-2000 (128C) + Custom Interconnect	China	4,981,760	61.4	61	18.5	3.32

Components of an HPC System

Components of an HPC System

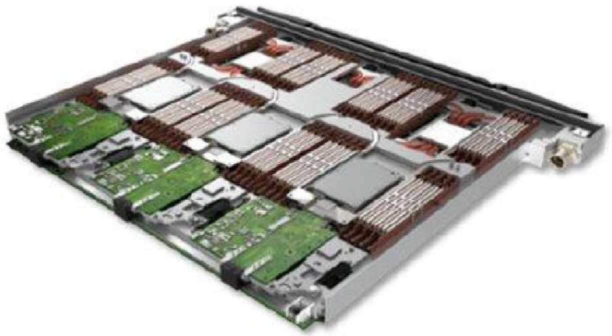
- **Processors (CPUs/GPUs):** The computational core.
- **Memory (RAM):** High-speed storage for active computations.
- **Interconnects:** High-speed networking for data transfer.
- **Storage:** Large-scale data storage solutions.
- **Software:** Operating systems, schedulers, parallel programming tools.

Today/future HPC Systems

- Heterogeneous computing architectures
    - Multicore CPUs
    - General Purpose GPU
    - In-network computation
    - Tensor Processing Units/Neural Processing Units
    - Field Programmable Gate Arrays

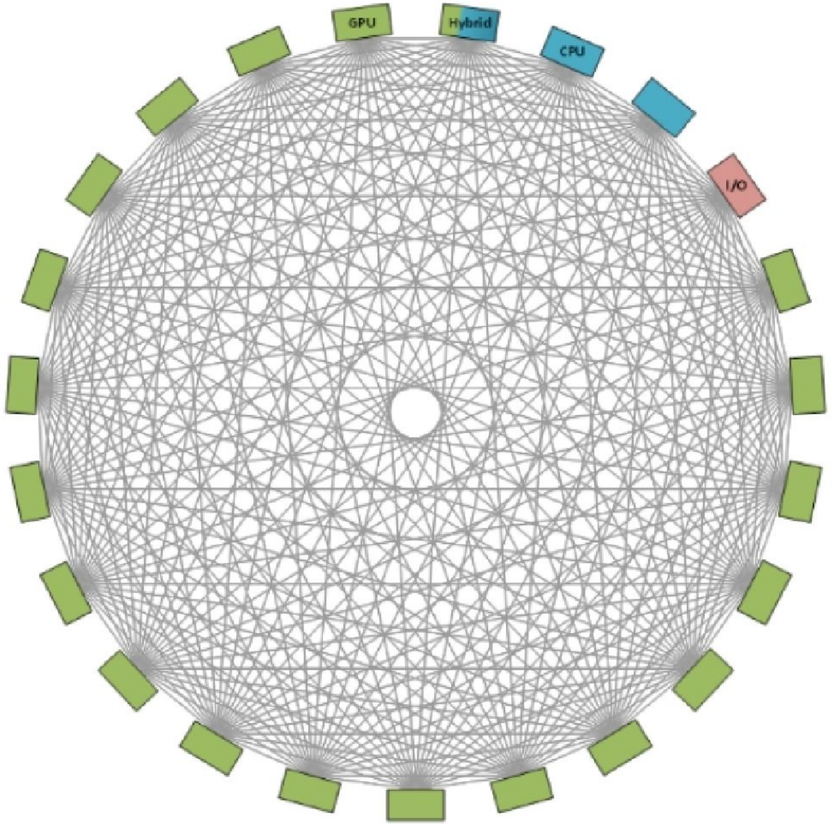
HPC Systems: nodes

- Heterogeneous nodes
  - General Purpose nodes (e.g. CPU based)
  - High-throughput computational nodes (e.g. GPU based)
  - Control/frontend nodes
  - Storage nodes
    - Fast storage
    - Capacity storage



HPC Systems: interconnects

- Heterogeneous interconnections:
  - intra-node: connects CPU to GPU/memory/Accelerator/local storage
  - intra-rack (node to node) communication
  - inter-rack communication



Inter-node communication

- **Scope:** Inside a compute node, between CPUs, GPUs, memory, and accelerators.
- **Technologies:**
  - **QPI / UPI (Intel)**, **Infinity Fabric (AMD)** → CPU–CPU interconnect.
  - **NVLink, NVSwitch (NVIDIA), UALink** → high-bandwidth GPU–GPU links.
  - **CXL, PCIe Gen5/Gen6** → CPU ↔ GPU, CPU ↔ accelerator, CPU ↔ memory expanders.

- **Characteristics:** Ultra-low latency (<100 ns), very high bandwidth (hundreds of GB/s), coherence support (e.g., CXL.mem).

Intra-rack interconnection

- **Scope:** Connects compute nodes within the same rack.
- **Technologies:**
  - **InfiniBand HDR/NDR (200–400 Gbps)** → dominant in HPC clusters.
  - **Slingshot (HPE/Cray)** → used in exascale systems (e.g., Frontier).
  - **Omni-Path (Intel, legacy)**.
  - **Ethernet (100/200/400 Gbps with RoCE or iWARP)** → in cost-sensitive HPC.
  - **UltraEthernet (UEC)** → Designed for next-gen HPC AI workload

- **Characteristics:** Optimized for low latency (~1–2 μs), high throughput, collective operations, and RDMA support

Inter-rack interconnection

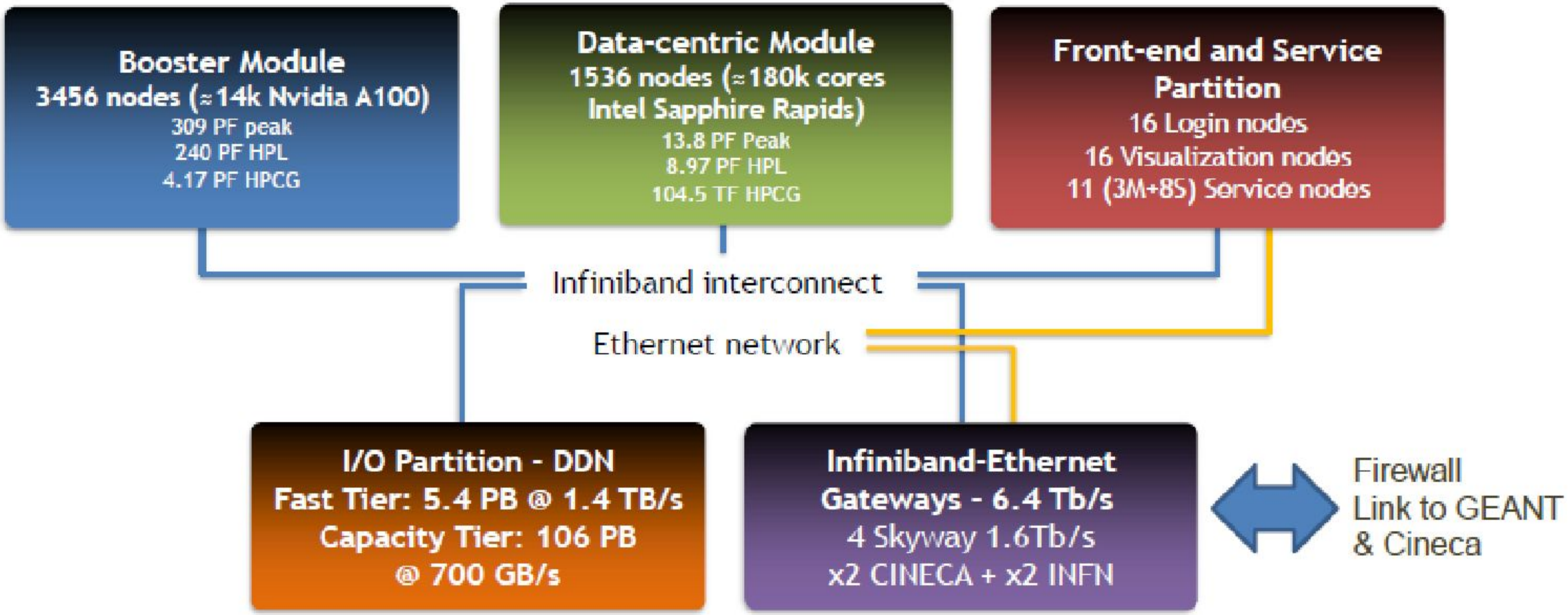
- **Inter-rack interconnection (rack-to-rack / large cluster scale)**
- **Scope:** Connects racks together into large supercomputers.
- **Technologies:**
  - Same as intra-rack (InfiniBand, Slingshot, high-speed Ethernet), but arranged in scalable **topologies**:
    - **Dragonfly, Fat-Tree, Torus, HyperX**.
  - Optical interconnects increasingly important for longer distances (fiber links, silicon photonics).
- **Characteristics:** Scalability (tens to hundreds of thousands of nodes), fault tolerance, congestion control.

Example of a Typical Supercomputer: Leonardo

Example of a Typical Supercomputer

- Based on **Atos BullSequana XH2000** with a heterogeneous design
  - CPU partition (Intel Xeon “Ice Lake”)
  - booster partition with **NVIDIA A100 GPUs**
- interconnected by **Mellanox HDR InfiniBand (200 Gbps)**.
- **Performance: 240 petaflops peak**
- **Storage & Networking:** (~100 PB capacity, ~1.2 TB/s bandwidth)

Leonardo



Booster Module
3456 nodes (≈14k Nvidia A100)
309 PF peak
240 PF HPL
4.17 PF HPCG

Data-centric Module
1536 nodes (≈180k cores Intel Sapphire Rapids)
13.8 PF Peak
8.97 PF HPL
104.5 TF HPCG

Front-end and Service Partition
16 Login nodes
16 Visualization nodes
11 (3M+8S) Service nodes

Infiniband interconnect

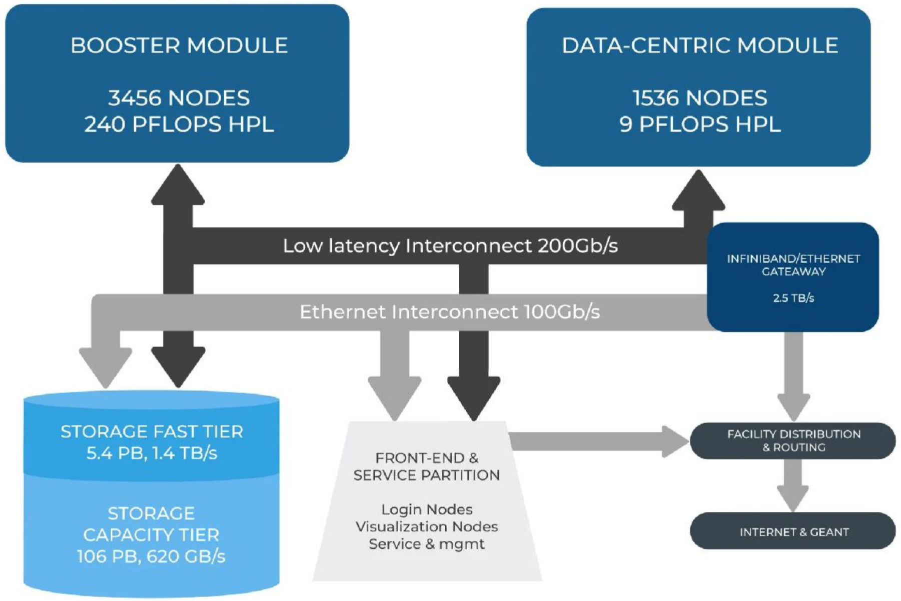
Ethernet network

I/O Partition - DDN
Fast Tier: 5.4 PB @ 1.4 TB/s
Capacity Tier: 106 PB @ 700 GB/s

Infiniband-Ethernet Gateways - 6.4 Tb/s
4 Skyway 1.6Tb/s
x2 CINECA + x2 INFN

Firewall
Link to GEANT
& Cineca

Leonardo infrastructure and login nodes



Atos BullSequana X430-E6

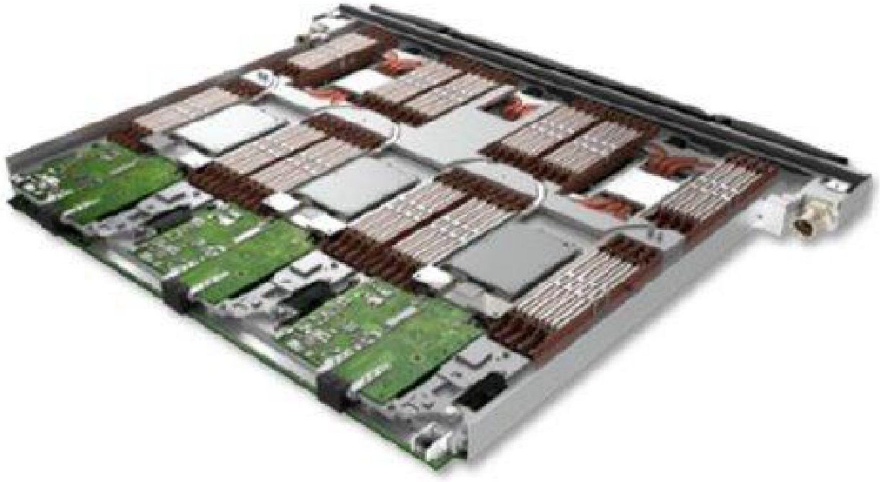
➢ Processors (dual-socket): 2x CPU Intel Ice Lake (64 cores/node), 2.4 GHz

➢ RAM: 512 (16x32) GB RAM DDR4 3200 MHz

➢ Disk: 14 TB HDD

➢ NO GPUs

Data Centric and General Purpose partition



BullSequana X2140 three-node CPU Blade

- ➢ 1536 nodes
- ➢ 2x CPU Intel Xeon 8480+, 56 cores Intel Sapphire Rapids (112 cores/node), 3.8 GHz
- ➢ RAM: 512 (16 x 32) GB DDR5 4800 MHz
- ➢ Disk: 1x SSD 3.84 TB M.2 NVMe
- ➢ 1x port HDR100 100Gb/s network interface

Peak performance: about 13 PFlops

Booster (GPU) module

Atos BullSequana X2135 “Da Vinci” blade

- ➢ 3456 nodes
- ➢ 1 x CPU Intel Xeon 8358, 32 cores Intel Ice Lake, 2.6 GHz
- ➢ RAM: 512 (8 x 64) GB DDR4 3200 MHz
- ➢ Accelerators: 4 x NVidia custom Ampere GPU A100 SXM4 64 GB, NVLink 3.0
- ➢ NVIDIA Mellanox HDR DragonFly+ 200Gb/s
- ➢ DISKLESS!!!

Peak performance per node: about 89,4 TFlops
Peak performance: about 309 PFlops

Storage

Fast Tier (5.4 PB, 1.4 TB/s)

NVMe storage (SSD disks)

➢ HOME, PUBLIC, FAST SCRATCH

Capacity Tier (106 PB, 744/620 GB/s)

HDD disks

➢ WORK, LARGE SCRATCH, DRES

Inter-node network topology

Dragonfly+ topology

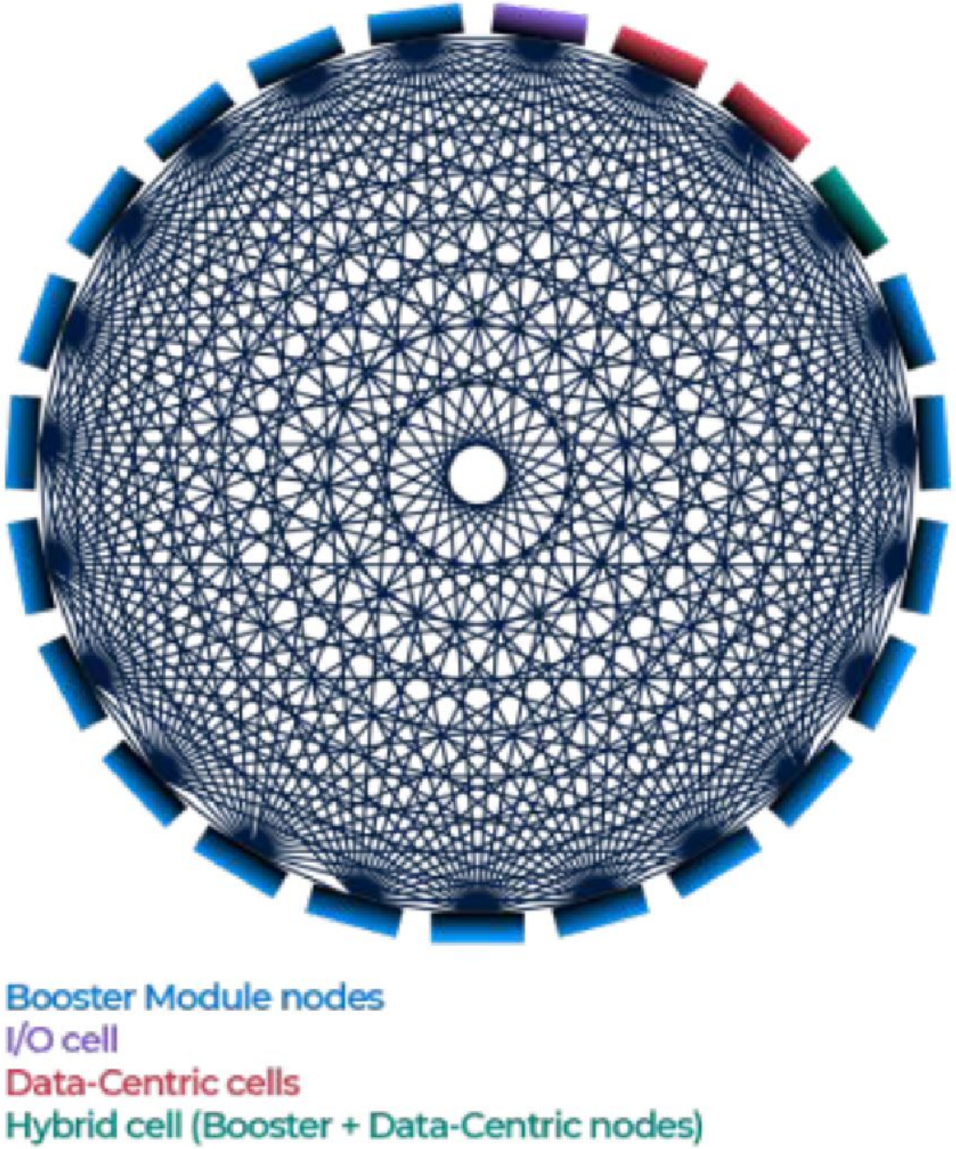
based on Nvidia Mellanox Infiniband HDR 200 Gb/s

➢ All nodes are divided into cells (180 nodes for cell)

➢ **Non-blocking**, two-layer Fat Tree within the cells

➢ All to all connection between cells

➢ Adaptive routing algorithm

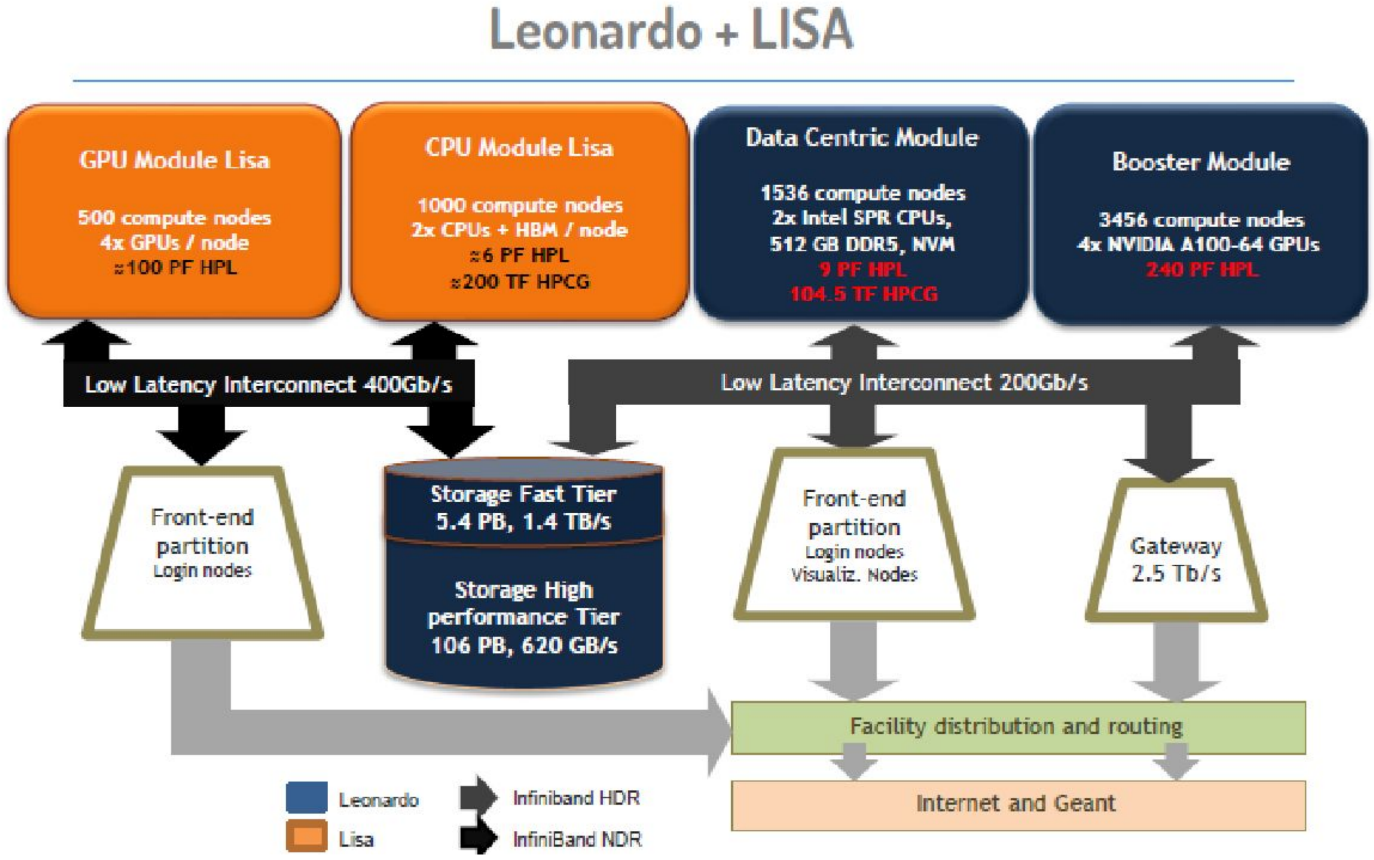


Leonardo expansion: LISA

LISA: Leonardo **I**mproved **S**upercomputing **A**rchitecture**:**

An AI partition of Leonardo dedicated for Generative AI

- LISA will operate alongside the existing Leonardo architecture as an independent but integrated partition



Leonardo + LISA

GPU Module Lisa

500 compute nodes
4x GPUs / node
≈100 PF HPL

CPU Module Lisa

1000 compute nodes
2x CPUs + HBM / node
≈6 PF HPL
≈200 TF HPCG

Data Centric Module

1536 compute nodes
2x Intel SPR CPUs,
512 GB DDR5, NVM
9 PF HPL
104.5 TF HPCG

Booster Module

3456 compute nodes
4x NVIDIA A100-64 GPUs
240 PF HPL

Low Latency Interconnect 400Gb/s

Low Latency Interconnect 200Gb/s

Front-end partition
Login nodes

Storage Fast Tier
5.4 PB, 1.4 TB/s

Storage High performance Tier
106 PB, 620 GB/s

Front-end partition
Login nodes
Visualiz. Nodes

Gateway
2.5 Tb/s

Facility distribution and routing

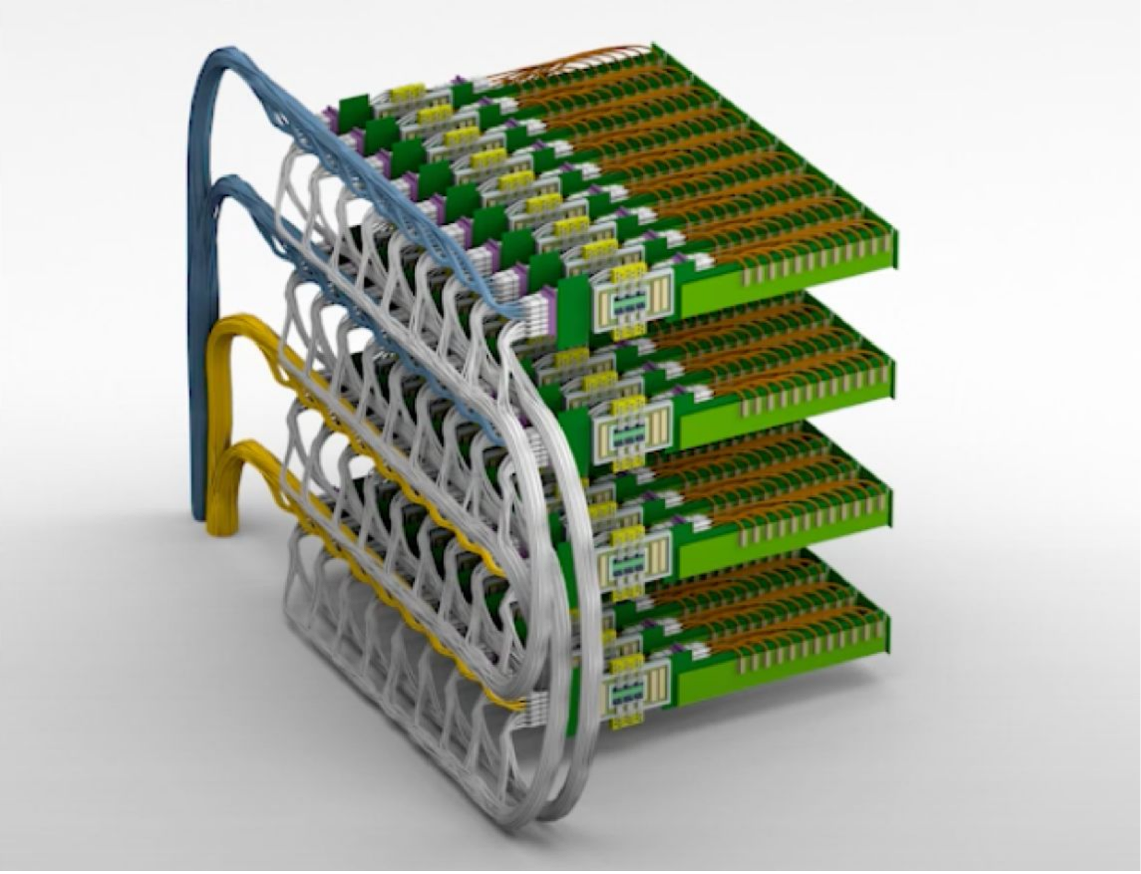
Internet and Geant

🟦 Leonardo
🟧 Lisa

Infiniband HDR
InfiniBand NDR

Example of a Typical Supercomputer

- Each node in the LISA partition is equipped with 8 NVIDIA H100 GPUs.
- Total GPUs: The system comprises 1,328 H100 GPUs, organized into 166 nodes,
- Each H100 GPU is equipped with 80 GB of high-bandwidth memory (HBM2e)
- The GPUs within each node are interconnected using NVIDIA NVLink, providing high bandwidth and low latency communication between GPUs.



Break