Il Supercomputer Leonardo al Tecnopolo di Bologna e gli Sviluppi Futuri del CINECA

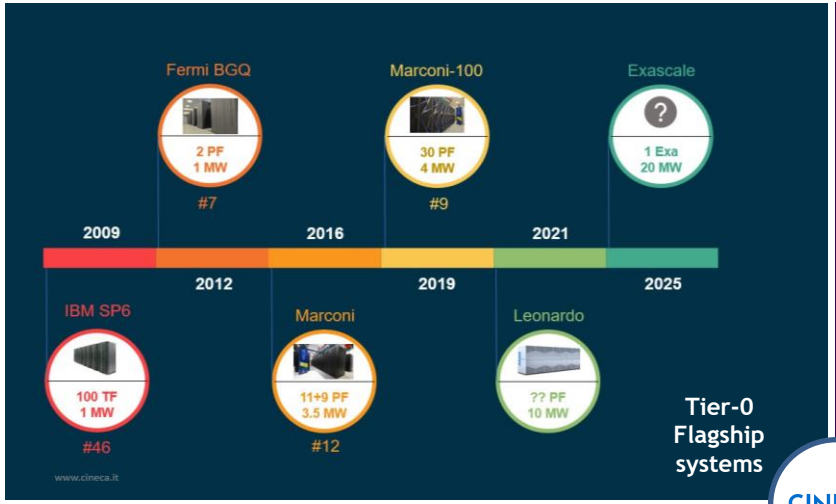
Dr. Daniele Cesarini - Workshop sul Calcolo nell'I.N.F.N.

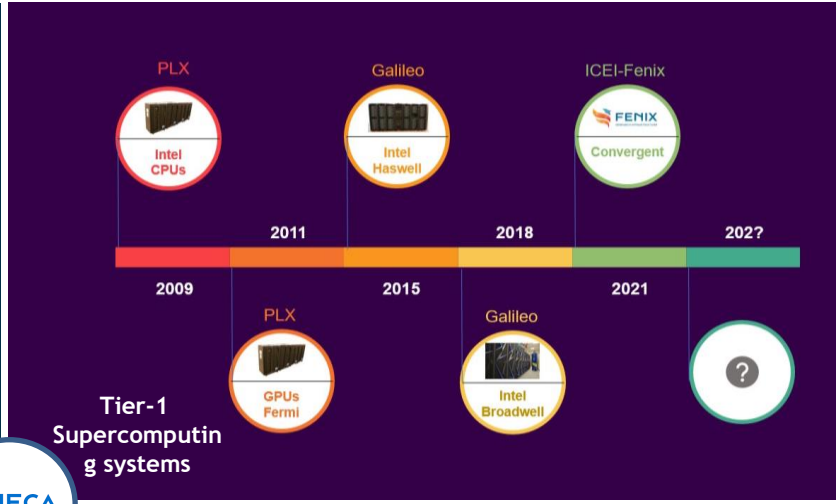
Hotel Ariston, Paestum, 23-27 Maggio 2022

Overview

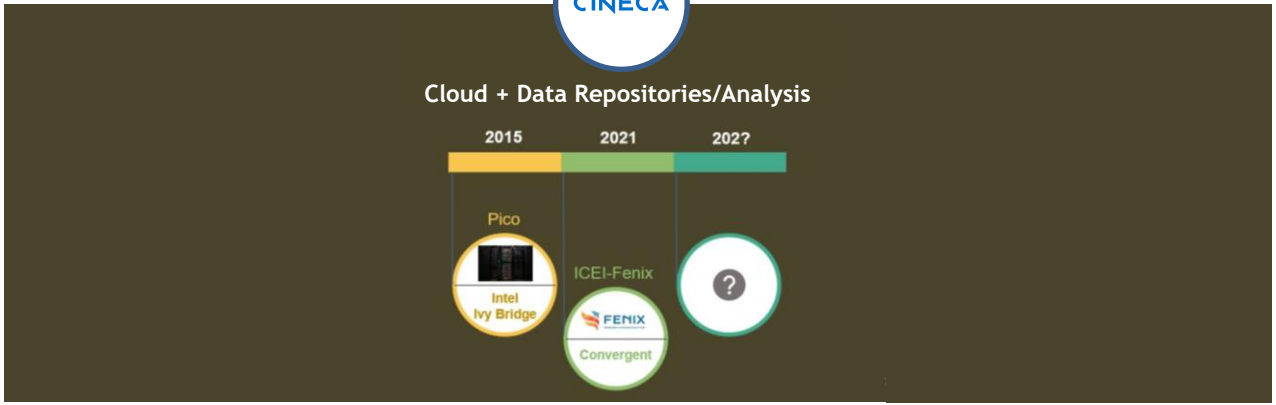
- CINECA HPC Infrastructures
- Leonardo
- Leonardo upgrade
- Galileo100 upgrade
- Data center interconnection between Casalecchio and Bologna Tecnopole
- New CINECA’s data center in Naples
- Quantum Computing at CINECA
- Meet Monte Cimone – First HPC-like RISC-V Cluster
- Leonardo facility at the Bologna Tecnopole

CINECA HPC Infrastructure

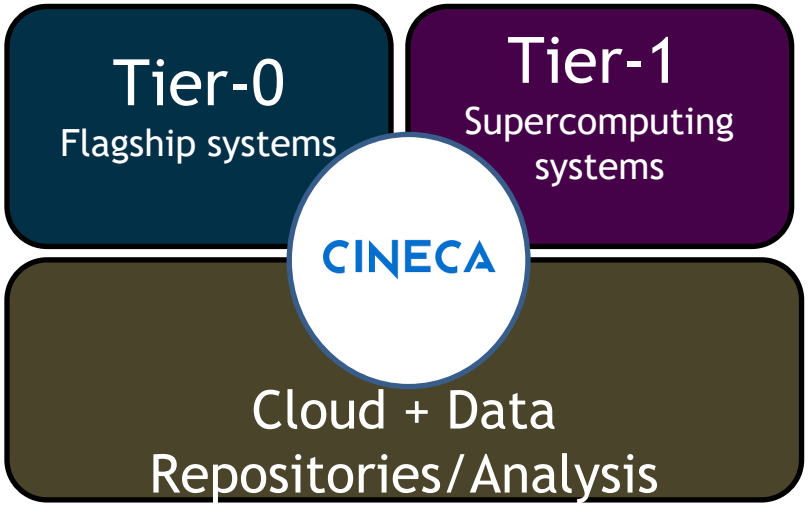


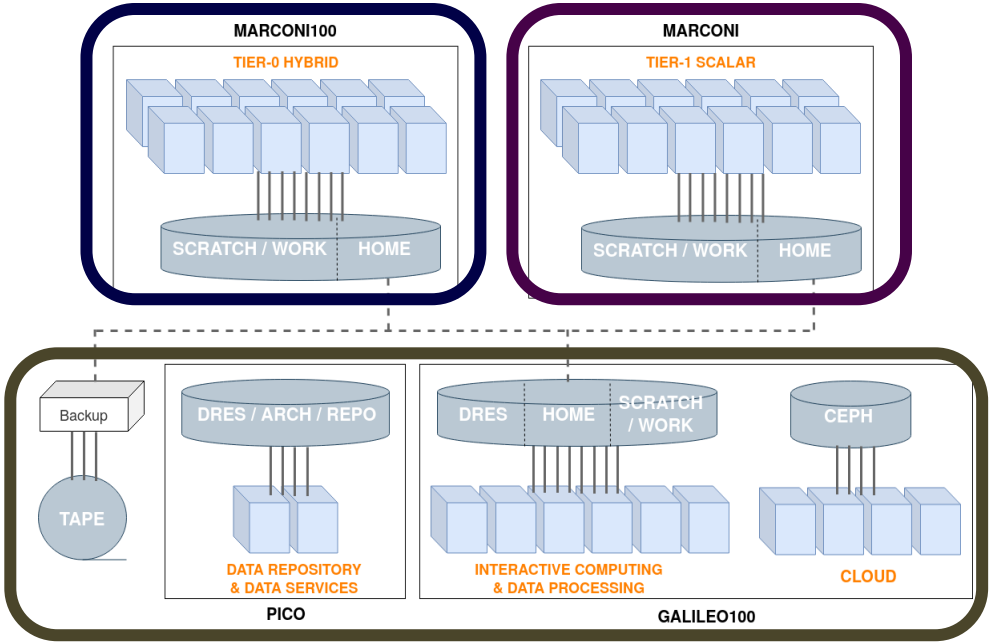


CINECA



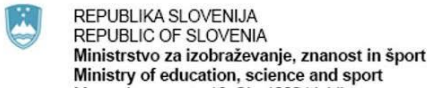
CINECA HPC Infrastructure





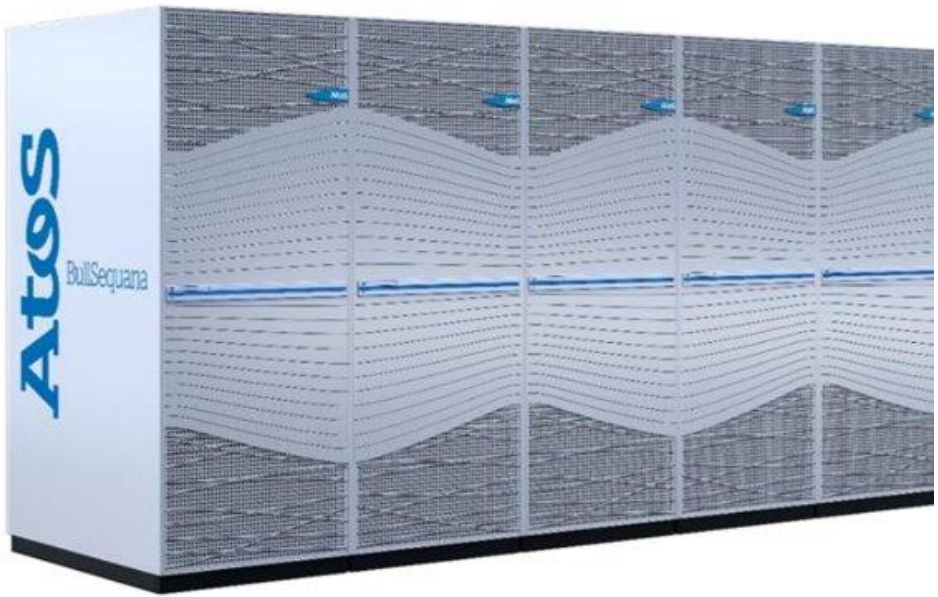
Technical Specifications

Leonardo Consortium

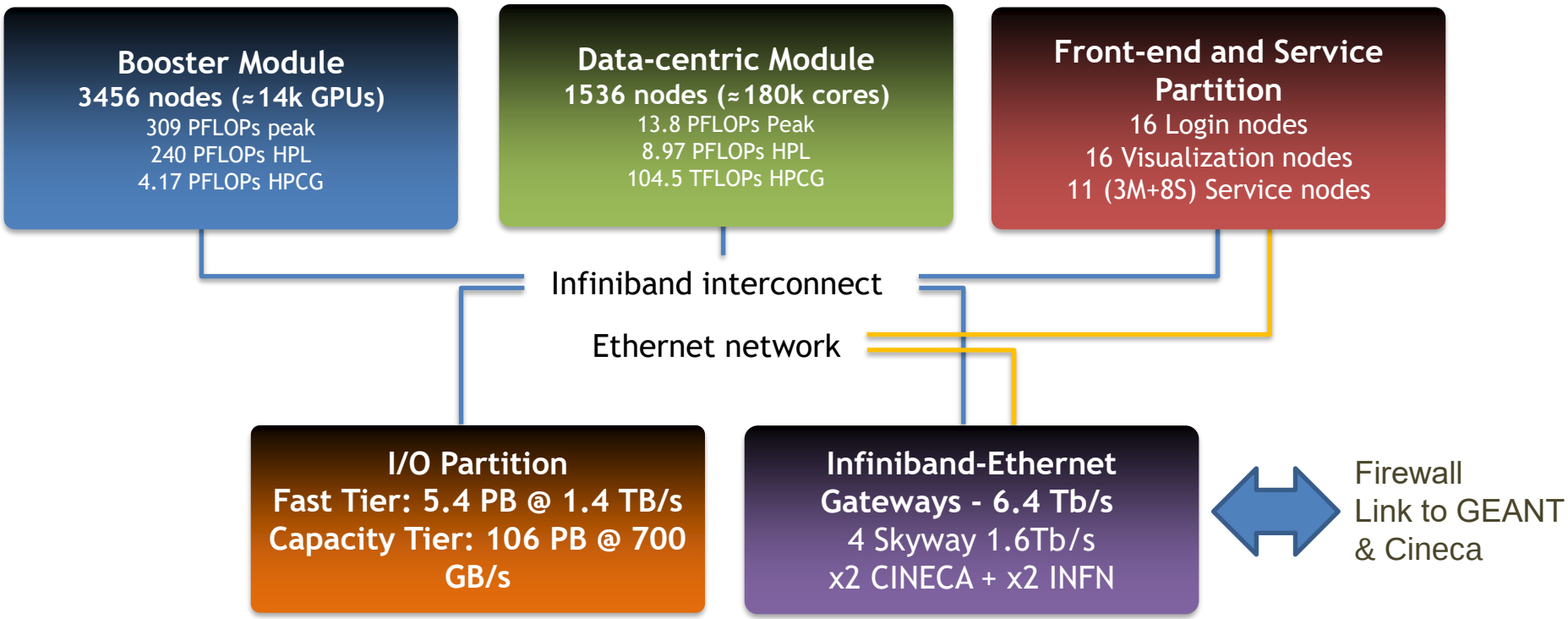


Leonardo Specifications

- Based on **BullSequana XH2000** platform technology
- Computing racks **95% Direct Liquid Cooled**
- **Warm water**: Inlet temperature of 37 degrees
- NVIDIA Mellanox **HDR 200** interconnect
  - Dragonfly+ topology
  - 1.11:1 intra-cell
  - 0.82:1 globally



Leonardo Specifications



Leonardo Compute Cabinet – BullSequana XH2000

PDU + Power Management Controller

Up to 6 x 15kW PSU shelves
DLC & hot-pluggable

Up to 32 compute blades
Technologies provided by:
- Intel
- Nvidia
- ARM
- AMD

DLC & hot-pluggable

Up to 3 Hydraulic chassis for Direct Liquid Cooling
- Inlet water temperature up to 40°C
- 95 per cent dispersion

Up to 2 Management network switches
hot-pluggable

Up to 10 Interconnect Network Switches
Technologies:
- BXI
- Mellanox HDR
- High-speed Ethernet

DLC & hot-pluggable

Flexible mid plane
Allows for:
- routing customization
- mixed interconnect bandwidths
- two possible topologies (Full Fat Tree and DragonFly+)

Front side of the cabinet

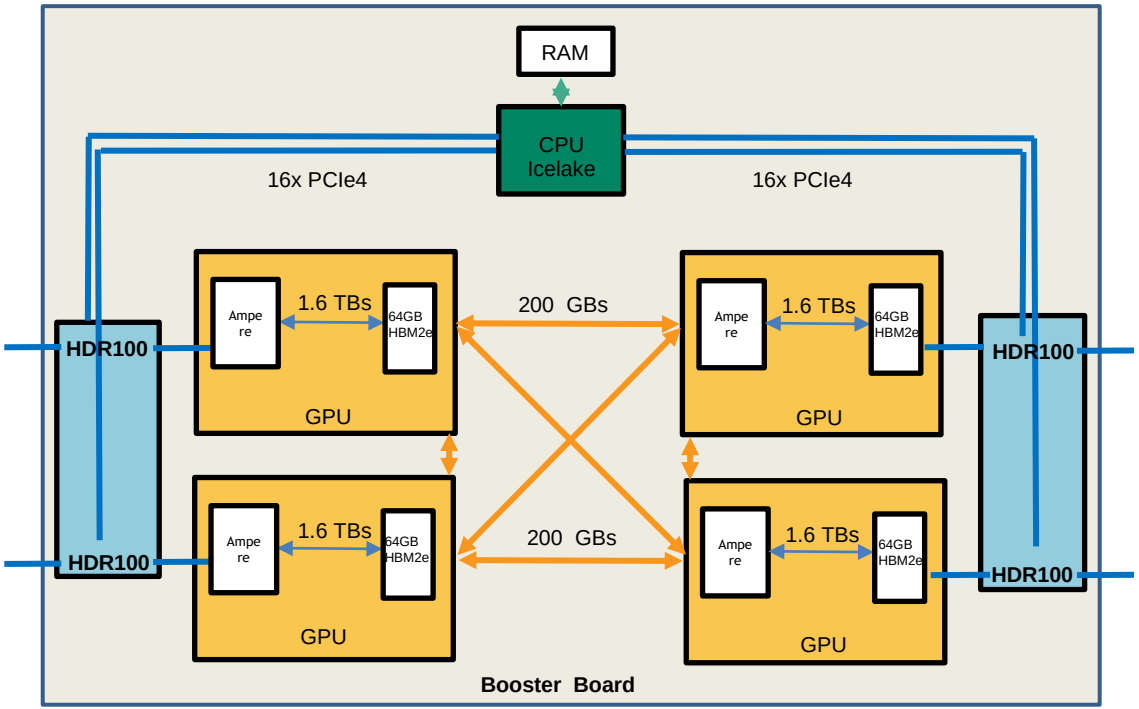
Rear side of the cabinet

BOOSTER: Da Vinci Blade

BullSequana X2135 “Da Vinci” single-node GPU Blade

- 1 x CPU Intel Xeon 8358 32 cores, 2,6 GHz
- 8 x 64 GB (512GB) RAM DDR4 3200 MHz
- 4 x NVidia custom Ampere GPU 64GB HBM2
- 2 x NVidia HDR100 dual port card

Booster Blade: Da Vinci



- Ad hoc board from ATOS
- 4 NVidia custom Ampere GPUs (SXM)
- CPU-GPU connection via PCIe4 16x connection through HDR Connect6 HCA
  - PCI passthrough
  - 16 PCI links towards CPU, 16 links towards GPU
  - Bandwidth: 32 GB/s
- Full NVLink GPU-GPU connection
  - 200 GB/s bi-directional
- No PCI switch between host and external network
  - Low latency
- Out-of-band telemetry information
- GPUdirect

Data Centric Blade

BullSequana X2140 three-node CPU Blade

For each node:

- 2 x CPU Intel Sapphire Rapids 56 core 350W
- 16 x 32 GB RAM (512 GB) DDR5 4800 MHz
- 1 x NVidia HDR100 single port card
- 1 x M.2 SSD 3,84 TB

IO Partition: Storage

Fast Tier: 5.4 PB @ 1.4 TB/s

- 31 x DDN EXAScaler ES400NVX2 appliances for NVMe storage:
  - 24 x 7,68 TB SSD NVMe with encryption support
  - 4 x InfiniBand HDR ports

Capacity Tier: 106 PB @ read 744 GB/s - write 620 GB/s

- 31 x DDN EXAScaler SFA799X appliances for HDD storage:
  - Controller node: 82 x 18 TB HDD SAS 7200 rpm and 4 x HDR100 ports
  - 2 x SFA18KX JBOD expansion per controller, each with 82 x 18 TB HDD SAS 7200 rpm (Declustered RAID)
  - 4 x InfiniBand HDR100 ports
- 4 x DDN EXAScaler SFA400NVX appliances for metadata
  - 21 x 7,68 TB SSD NVMe with encryption support
  - 4 x InfiniBand HDR100 ports

Network Topology

Based on Mellanox Networking Infiniband HDR hardware for switches, cables and NICs

Dragonfly+ topology

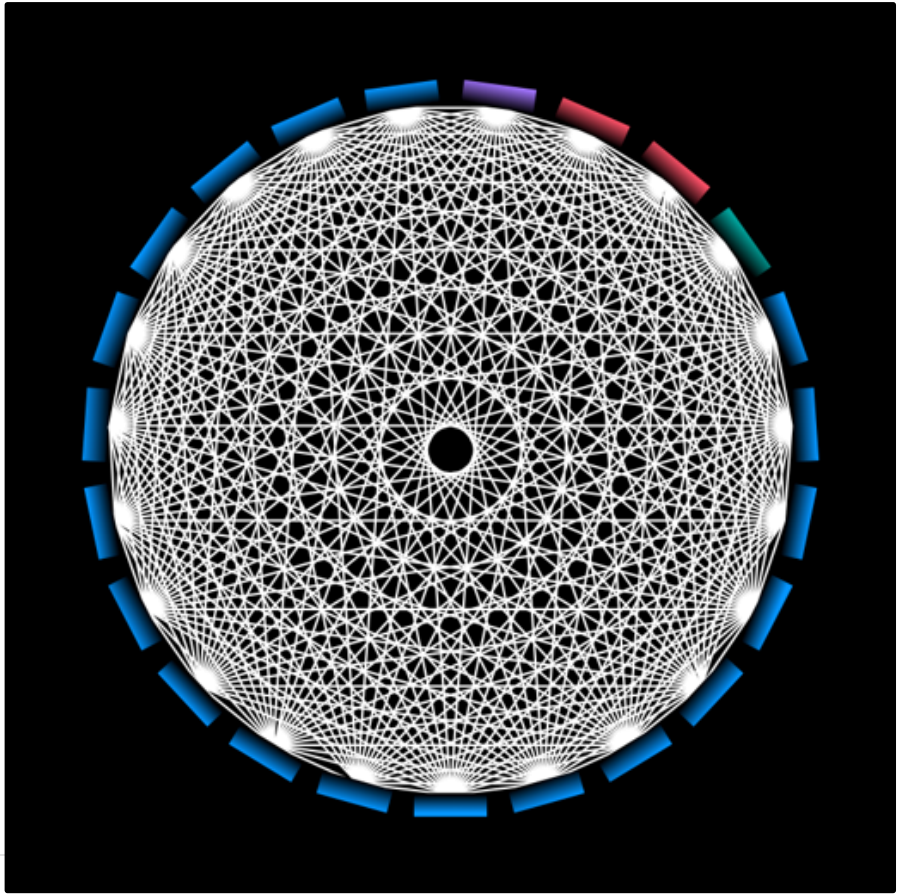
- All nodes are divided into cells
- Non-blocking, two-layer Fat Tree within the cells
- All to all connection between cells
- Mellanox QM8700 40-ports switches
- NIC Connect-X6

19 cells for Booster Module nodes

1 I/O cell

2 Data-Centric & General Purpose Cells

1 Hybrid cell, made of Booster and Data-Centric & General Purpose nodes



Booster Module Topology

19 Booster cells

6 Cabinets per cell

180 nodes per cell (30 nodes/blades per cabinet)

18 L2 switches per cell (3 switches per cabinet)

Switch: 22 uplinks, 18 downlinks = 0.82:1

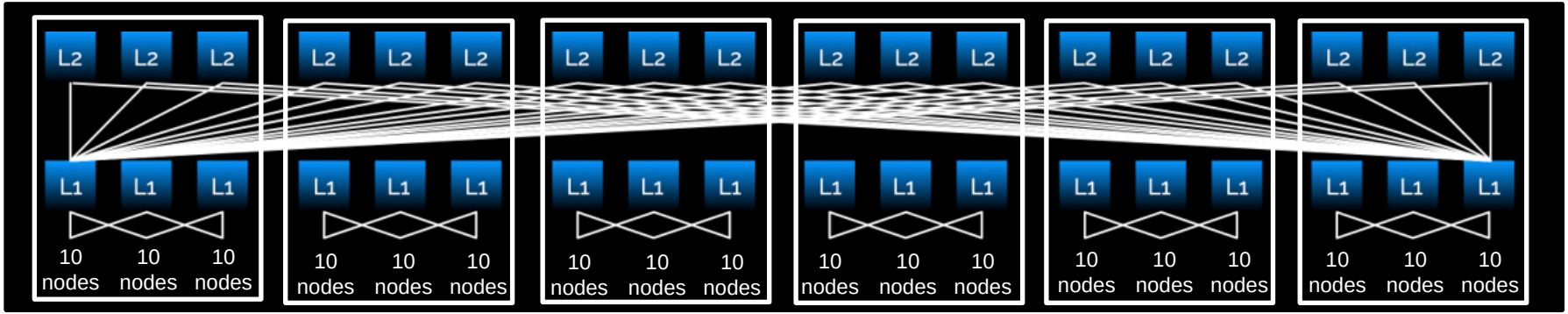
18 L1 switches per cell (3 switches per cabinet)

Switch: 18 uplinks, 20 downlinks = 1.11:1

Each node is connected to two different L1 switches

Drangonfly+ Booster Cell

In the diagram below, just the left and right L1 switch show their connections for the sake of clarity



Data Centric Module Topology

2 Data Centric & General Purpose cells

6 Cabinets per cell

576 nodes per cell (96 nodes into 32 blades per cabinet)

18 L2 switches per cell (3 L2 switches per cabinet)

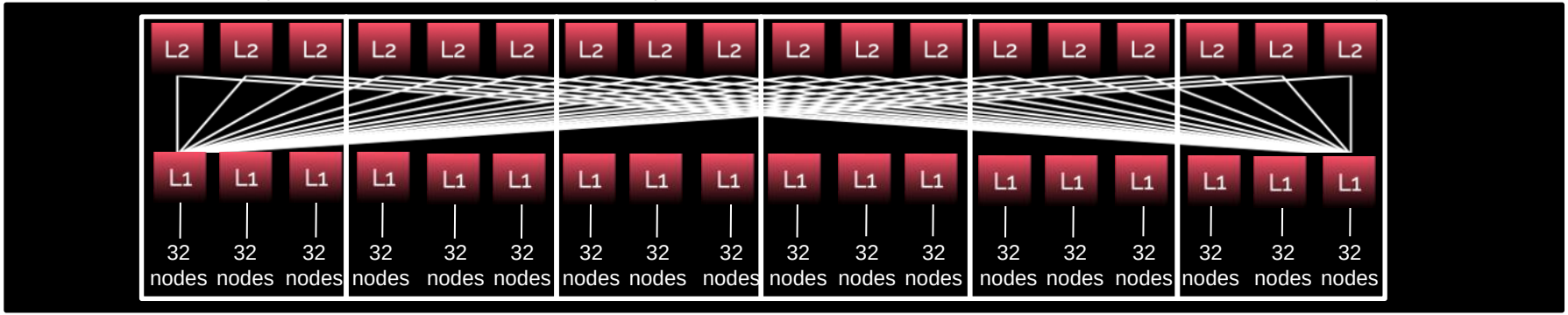
Switch: 22 uplinks, 18 downlinks = 0.82:1

18 L1 switches per cell (3 L1 switches per cabinet)

Switch: 18 uplinks, 16 downlinks = 0.89:1

Drangonfly+ DC & GP Cell

In the diagram below, just the left and right L1 switch show their connections for the sake of clarity



Hybrid Cell Module Topology

1 Hybrid cell

6 Cabinets per cell

384 CPU nodes (4 cabinet, 32 blades, 98 nodes)
+36 GPU nodes (2 cabinet, 18 nodes/blades) per cell

18 L2 switches per cell (3 L2 switch per cabinet)

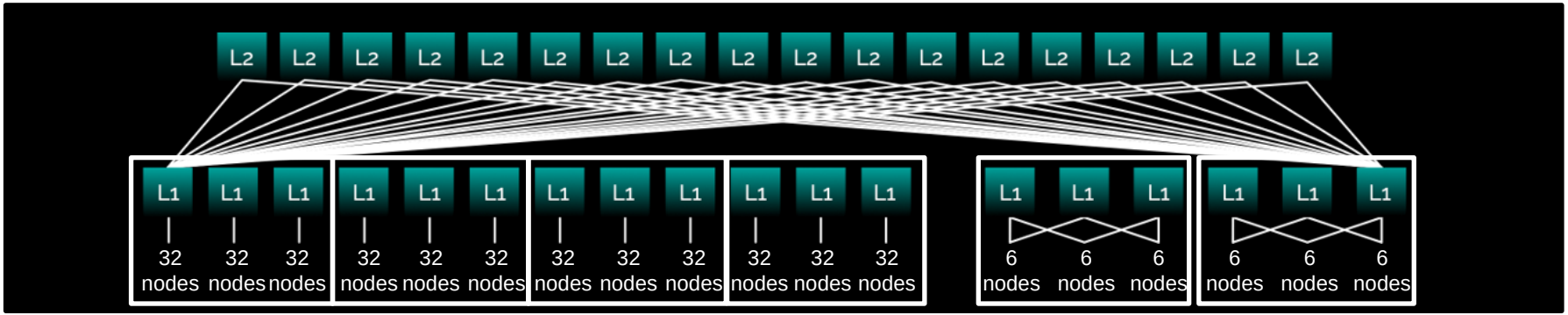
Switch: 22 uplinks, 18 downlinks = 0.82:1

18 L1 switches per cell (3 L1 switches cabinet)

Switch CPU: 18 uplinks, 16 downlinks = 0.89:1
Switch GPU: 18 uplinks, 12 downlinks = 0.67:1

Drangonfly+ Hybrid Cell

In the diagram below, just the left and right L1 switch show their connections for the sake of clarity



IO Cell Module Topology

1 Hybrid cell

14 cabinet (11 storage + 3 ancillaries)

HDR Non-Blocking for Storage Access

18 L2 switches per cell

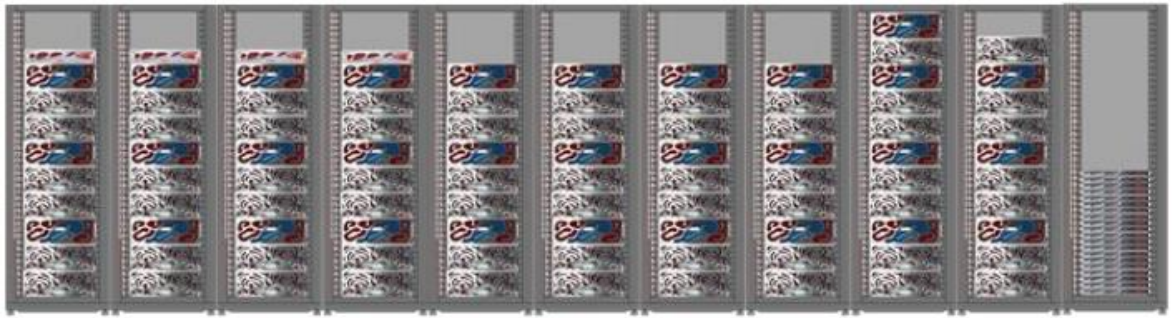
Switch: 22 uplinks, 18 downlinks = 0.82:1

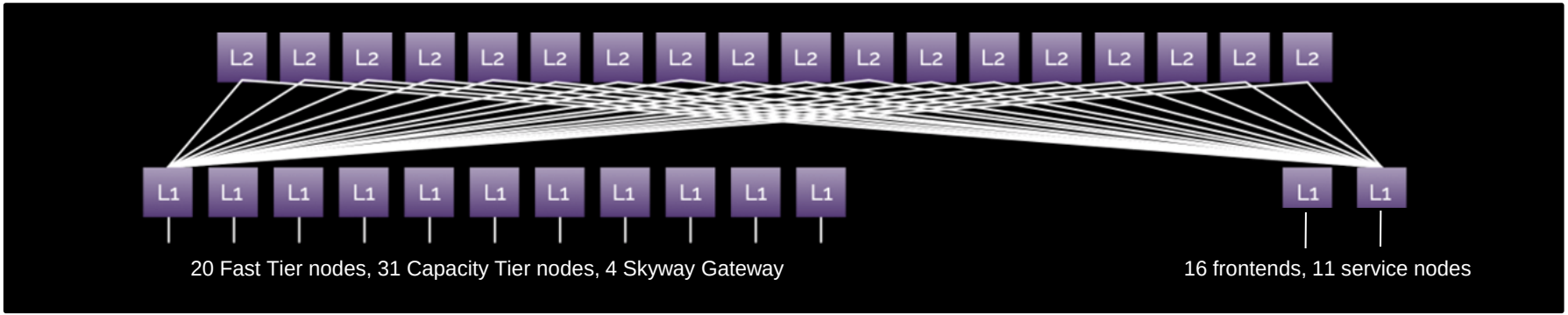
13 L1 switches per cell

Switch: 18 uplinks, downlinks:

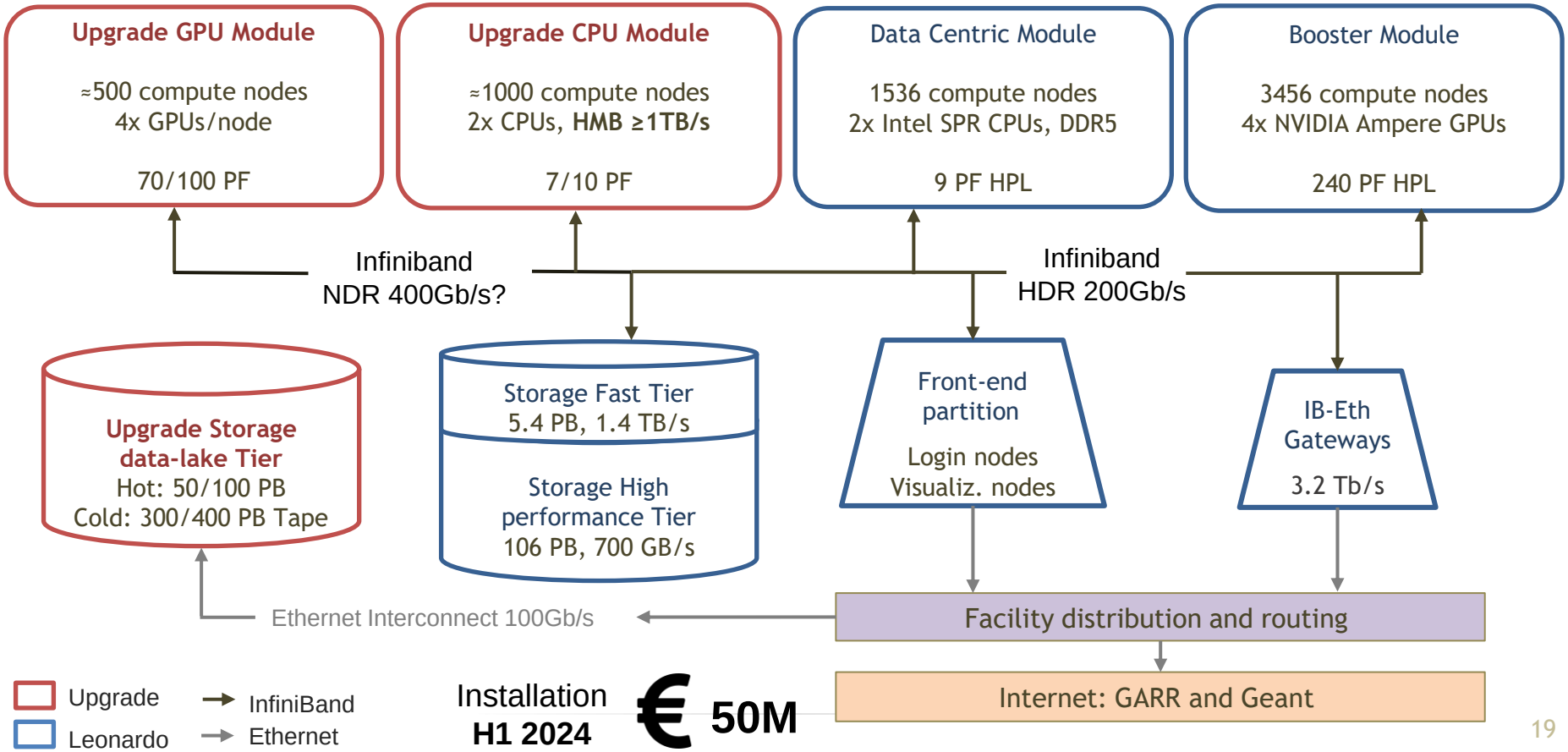
20-ports HDR Fast tier
4-ports HDR100 Metadata
31-ports HDR100 Capacity tier
32-ports HDR Skyway Gateway
96-ports HDR100 Frontend and Service

Drangonfly+ IO Cell





Upgrade Leonardo



Upgrade Galileo100

CINECA plans to **upgrade G100**

Leonardo will remain a **conventional** HPC system

G100 will be upgraded to become an **important cloud asset**

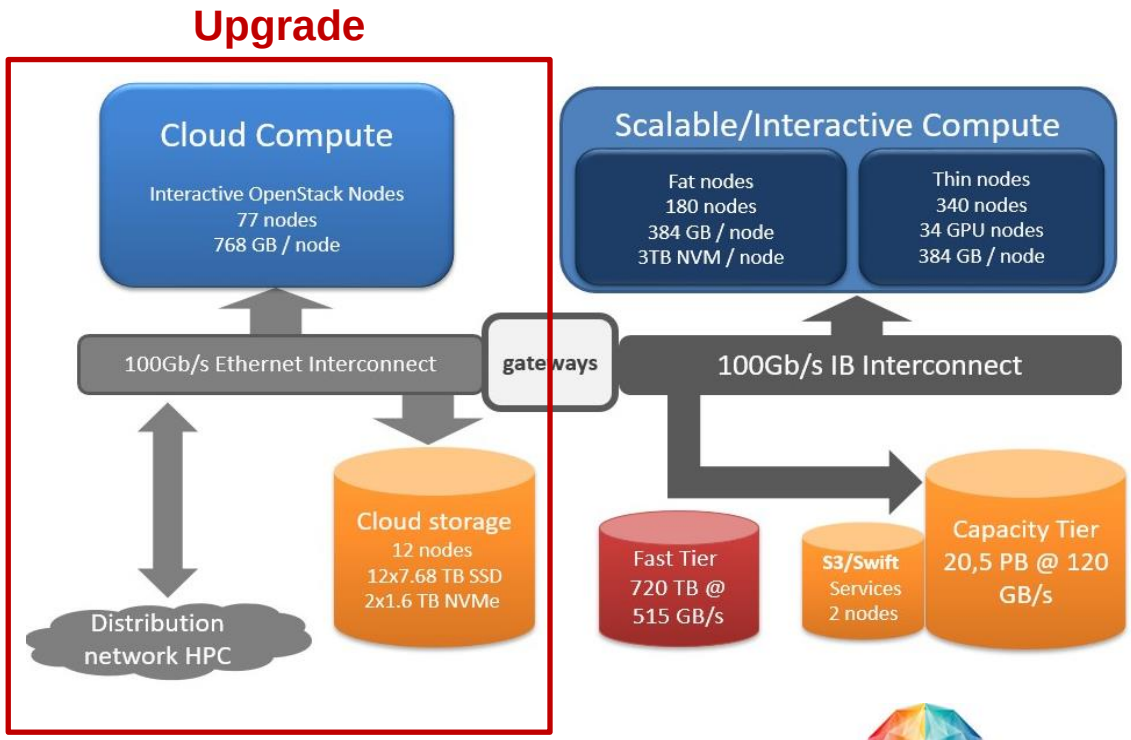
G100 will stay in the **Casalecchio area**

Time frame: **installation 2024**

Most of the investments will be used for the **cloud partition**

CINECA **investments** will be in the order of:

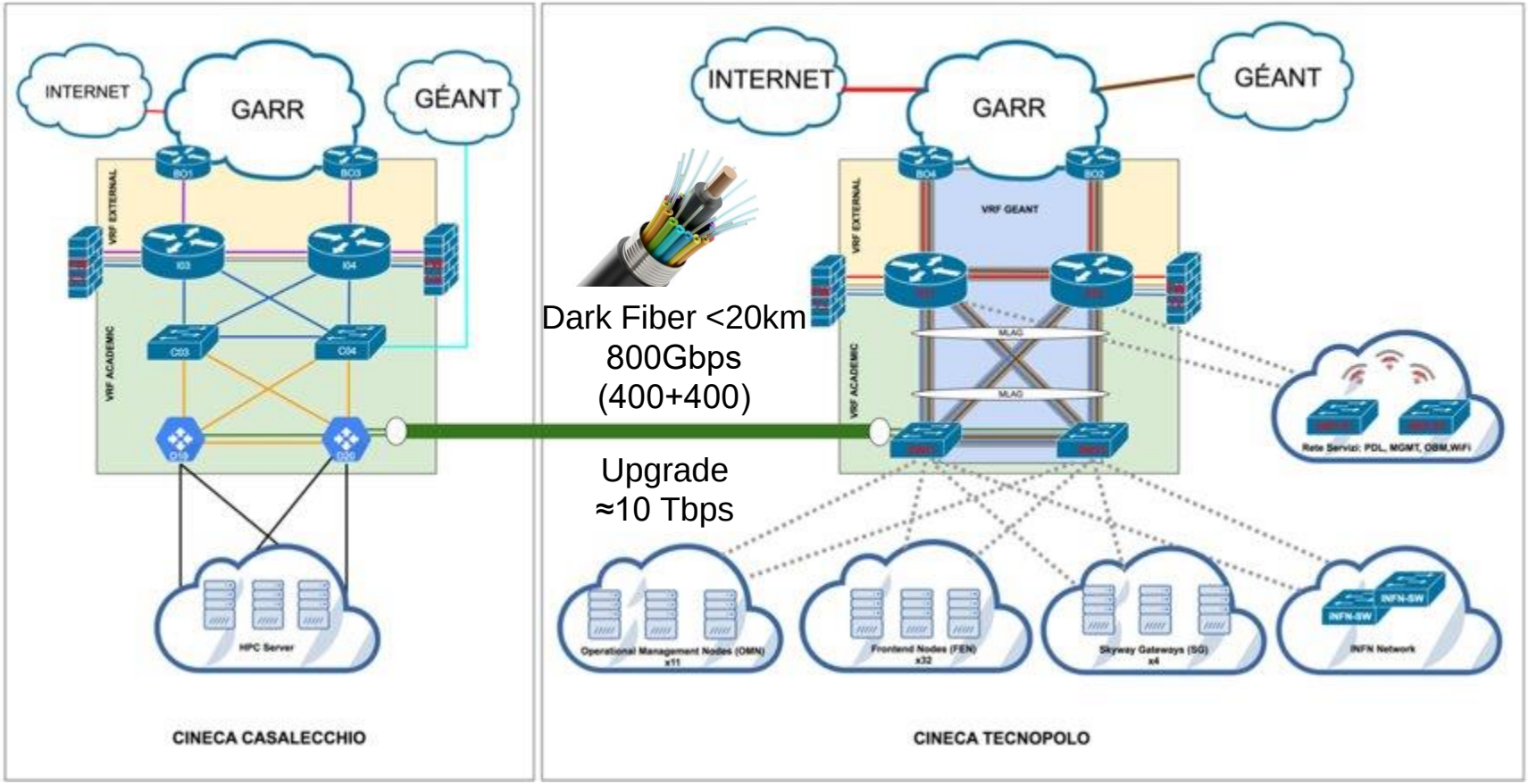
€ 7-10M



FENIX RESEARCH INFRASTRUCTURE

Human Brain Project

Casalecchio – Tecnopolo Interconnection



New CINECA's Data Center in Naples

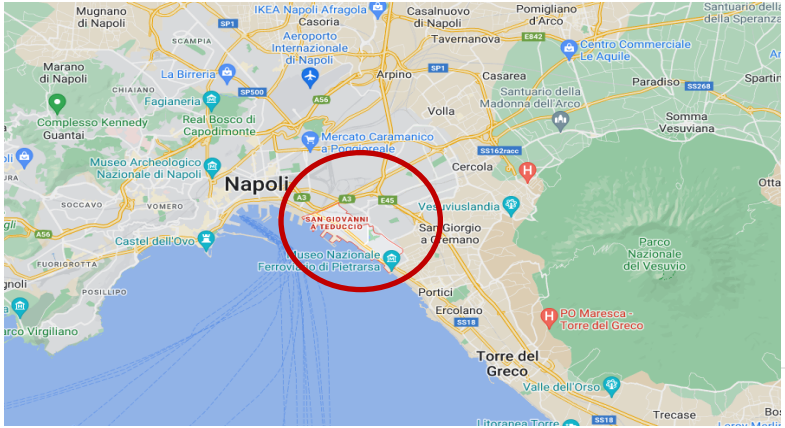
CINECA will open a new Tieri1 data center in **Naples area**

San Giovanni a Teduccio is the interested area

Complementary techonology with respect to Tecnopolo and Casalecchio HPC systems

Time Frame: **operative** 2024

CINECA's **investment** will be in the order: **€ 10M**



Around 15-20 positions will be opened in Naples! Drop me an email!

New data center solutions are currently under investigation…

Quantum Computing

CINECA plans to acquire a Quantum Computer

Initially the QC will be an experimental and dedicated system but the idea is to use QC as an **accelerator of Leonardo**

Some **QC technologies** are under investigation

It will be considered **QC European technologies**

Time frame: **installation H2-2023**

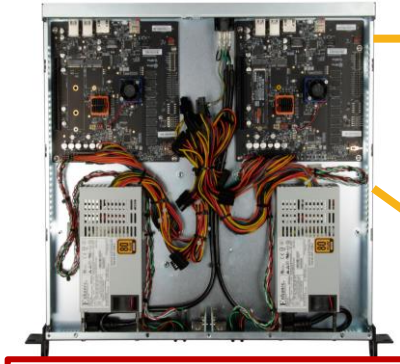
CINECA **investments** will be in the order of

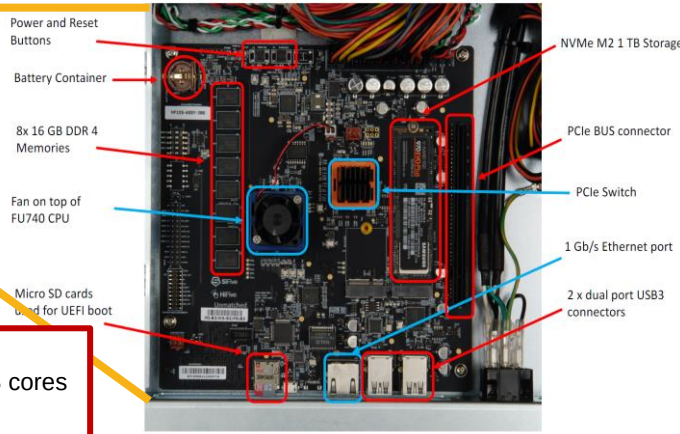
€ 10M

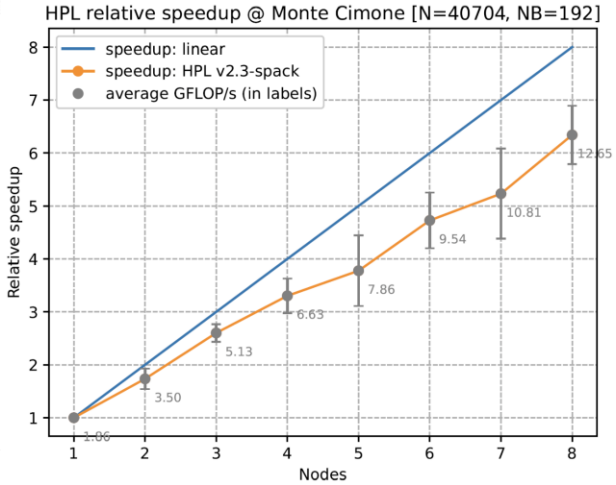
Meet Monte Cimone – First HPC-like RISC-V Cluster

Question: How mature is the RISC-V ecosystem? Is the **RISC-V ecosystem mature enough to build HPC production clusters**?

This work: We designed and built **Monte Cimone**, the **first physical prototype** and test-bed of a **complete RISC-V (RV64) compute cluster** integrating **compute, interconnect,** *a **complete software stack for HPC** and a full-featured **system monitoring infrastructure**.*







4x E4 RV007 1U Custom Server Blades:

- **2x SiFive U740 SoC** with **4x U74 RV64GCB** cores
- 16GB of **DDR4**
- 1TB node-local **NVME storage**
- **PCIe expansion** card **w/InfiniBand HCAs**
- **Ethernet + IB parallel networks**


CINECA
INFN
Tecnopolo di Bologna

Bologna TECNOPOLO

Bologna, Italy

Tecnopolo will be located close to the city center

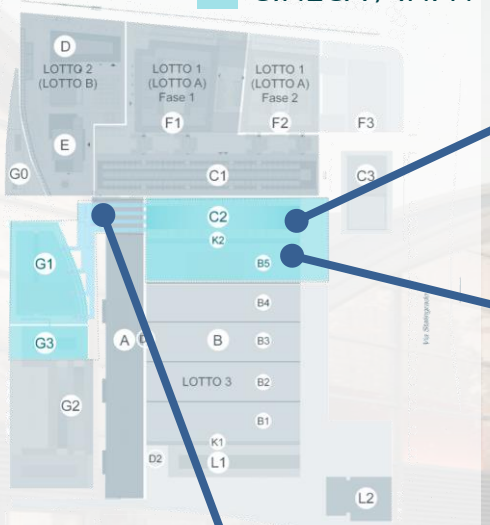
TECNOPOLO – Construction site

Manifattura Tabacchi

At the Tecnopolo in Bologna,
1950s structure designed by Pier Luigi Nervi

CINECA – INFN site

CINECA / INFN



Capannone Miscela C2 - LEONARDO

Botte B5 - INFN Data center

Ballette

Technological center G1

Technological Tunnels

C2

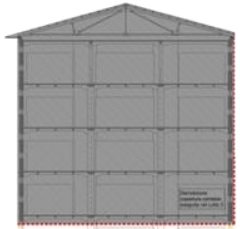
TECNOPOLO

CINECA / INFN

Capannone Miscela C2 - LEONARDO

Botte B5 - INFN Data center

Ballette



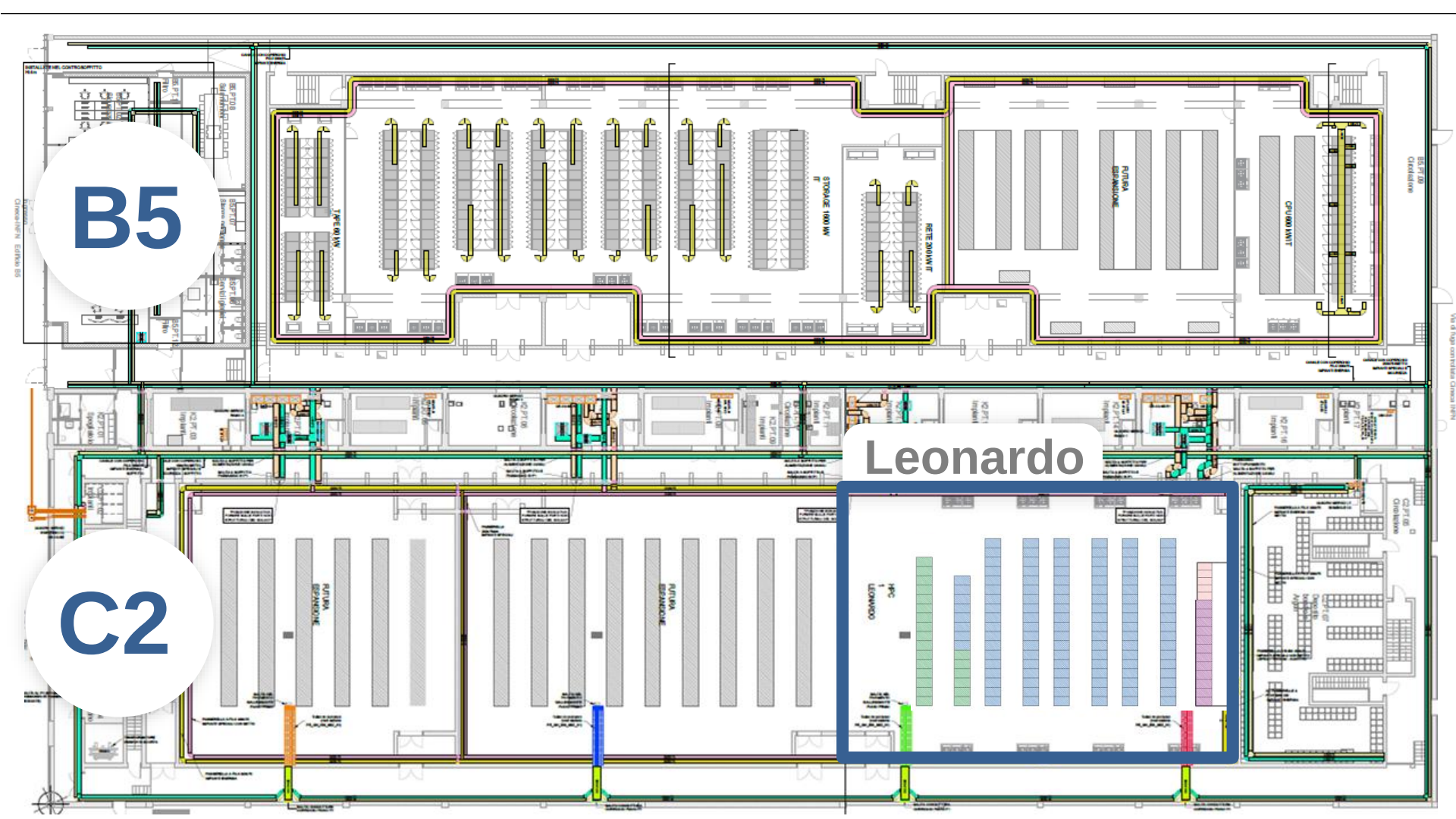
Technological center G1

Data Hall Leonardo

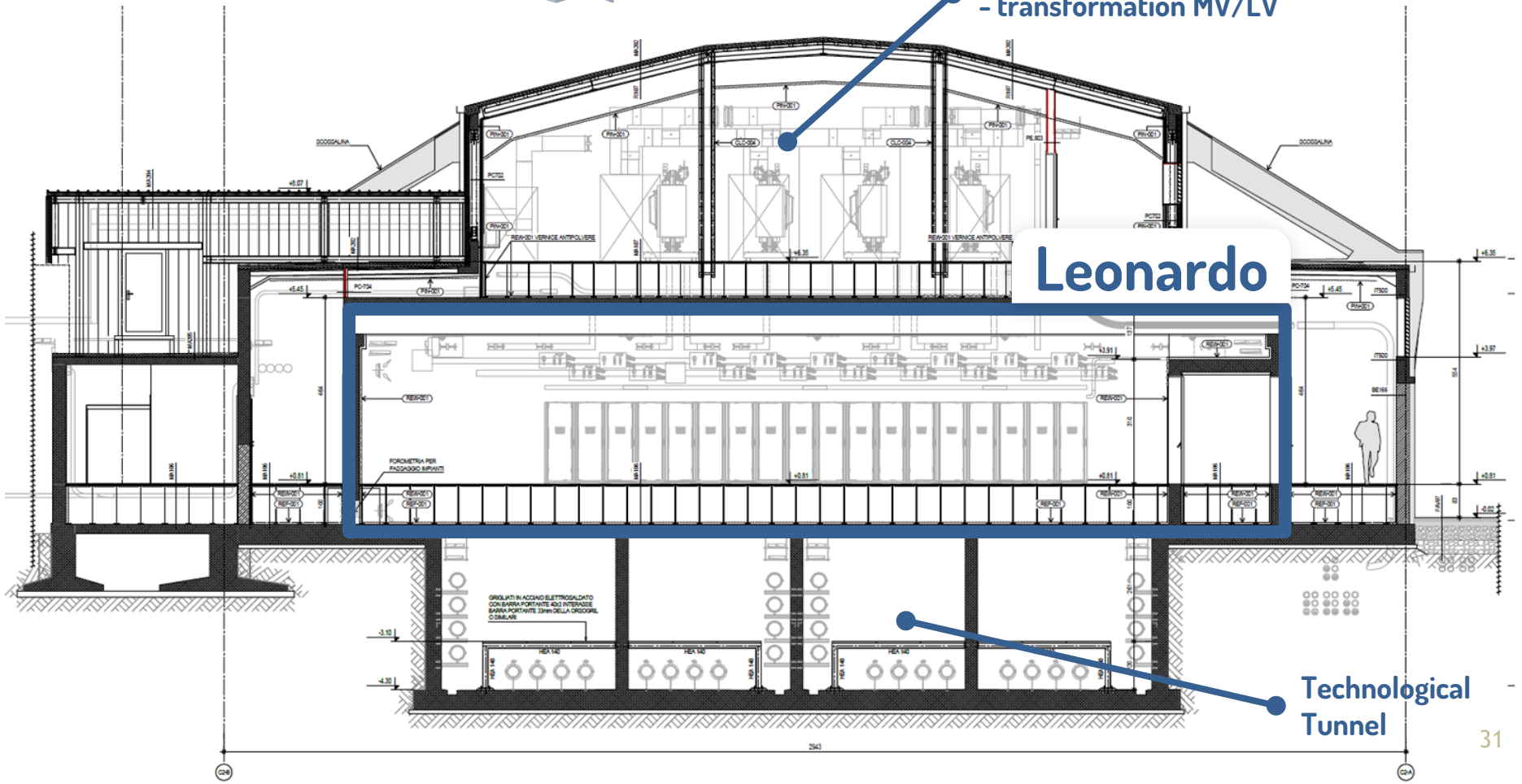
Capannone Miscela C2

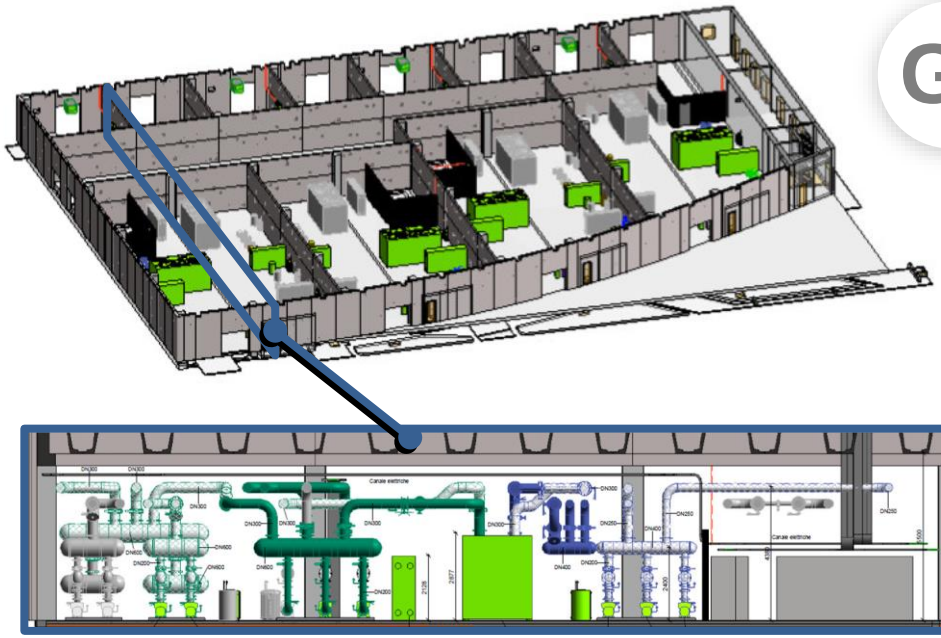
Technological tunnels

MEP connections

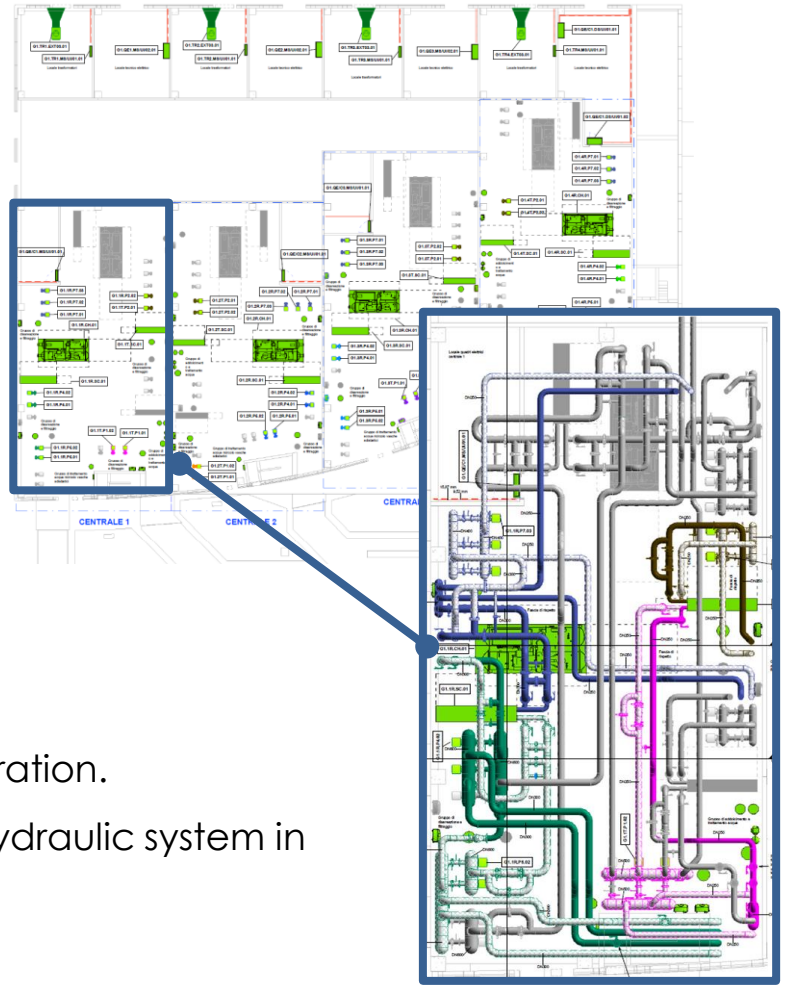

B5
C2
Leonardo
TAPE 60 kW
STORAGE 1000 kW IT
RETE 200 kW IT
FUTURA ESPANSIONE
CPU 600 kW IT
HPC 1 LEONARDO

C2 Section


Electrical Infrastructure
- transformation MV/LV
Leonardo
Technological
Tunnel



G1



G1 Building

Subdivided in **4 independent branches**, in a 3+1 configuration.

The water flows from each cooling plants, through the hydraulic system in the tunnels, to reach the white space.

DC – Infrastructure

Features

- **Concurrent Maintainability** and **Fault Tolerance** according to Rating 4 - TIA942 and Tier IV - Uptime Institute
- **MEP Infrastructure** designed to guarantee design performances at extreme external conditions as required by Uptime Institute: n = 20 years: +39,5°C / -12°C; based on ASHRAE Handbook – Fundamentals – 2017 / Bologna
- **Redundancy** Configuration: **3+1**, Electrical and Mechanical
- **PUE < 1,10** (year based measurement strategy compliant to Level 3 Green Grid/ASHRAE)
- Scalability, Modularity, Expandability for two different **expansion phases** (see table below);
- Stage 1 scheduled in 2022 - 2026 for **10 MW** ICT load and **1240 sqm** Rack Room;
- Stage 2 scheduled in 2026 - 2030 for **additional 10 MW** ICT load and **additional 2600 sqm** Rack Room;
- Mechanical and Electrical infrastructure able to comply with **2 different expansion strategies**.
    - **Stage 2a: Liquid Cooling** Expansion (16 MW Liquid Cooled + 4 MW Air Cooled)
    - **Stage 2b: Air Cooling** Expansion (8 MW Liquid Cooled + 12 MW Air Cooled)

Expansion phases

	Stage 1 (2020-2025) Liquid Cooled+Air Cooled	Stage 2a (2025-2030) Liquid Cooling Expansion	Stage 2b (2025-2030) Air Cooling Expansion
CINECA ICT Loads (Liq.Cooled+Air Cooled)	~10 MW 8 MW LC + 2 MW AC	~20 MW 16 MW LC + 4 MW AC	~20 MW 8 MW LC + 12 MW AC
Rack Room	1.240 sqm	3.840 sqm (1240+2600 sqm)	3.840 sqm (1240+2600 sqm)
Ancillary Spaces	900 sqm	900 sqm	900 sqm
Tot. Cooling Capacity (Med.Temp.W. 40-50°C)	8 MW No Redundancy	16 MW No Redundancy	8 MW No Redundancy
Tot. Cooling Capacity (Chilled Water 18-23°C)	6+2 MW 3+1 Redundancy	6+2 MW 3+1 Redundancy	12+4 MW 3+1 Redundancy
No-Break Power Cap.(UPS)	3+1 MW 3+1 Redundancy	6+2 MW 3+1 Redundancy	12+4 MW 3+1 Redundancy
Short-Break Power Cap.(GEN)	9+3 MW 3+1 Redundancy	9+3 MW 3+1 Redundancy	18+6 MW MW 3+1 Redundancy


Atos
LEONARDO
CINECA
LEONARDO


LEONARDO
EuroHPC
CINECA
Founded by
the European Union
L
E
O

CINECA

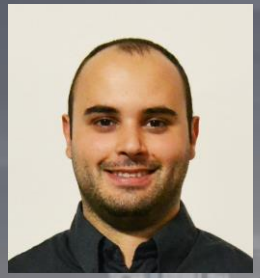
Ministero dell'Università e della Ricerca

EuroHPC Joint Undertaking

Regione Emilia-Romagna

Atos

INFN



Dr. Daniele Cesarini
HPC Specialist

Mail: d.cesarini@cineca.it

Thank you for your attention!