Generalitat de Catalunya
Departament de Recerca
i Universitats

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

RED ESPAÑOLA DE SUPERCOMPUTACIÓN

EuroHPC Joint Undertaking

MareNostrum 5

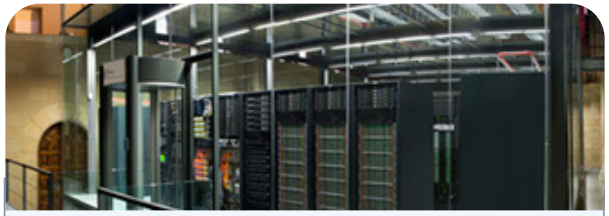
Dr. Sergi Girona
Operations Director

November 2023

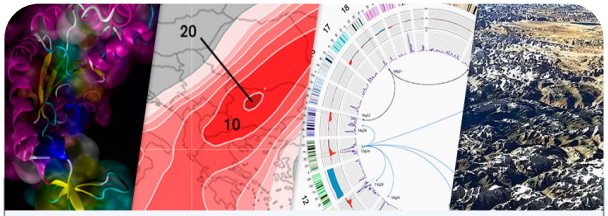
SC24 - HPCSYSPROS23 - HPC Systems Professionals Workshop

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS objectives



Supercomputing services
to Spanish and EU researchers



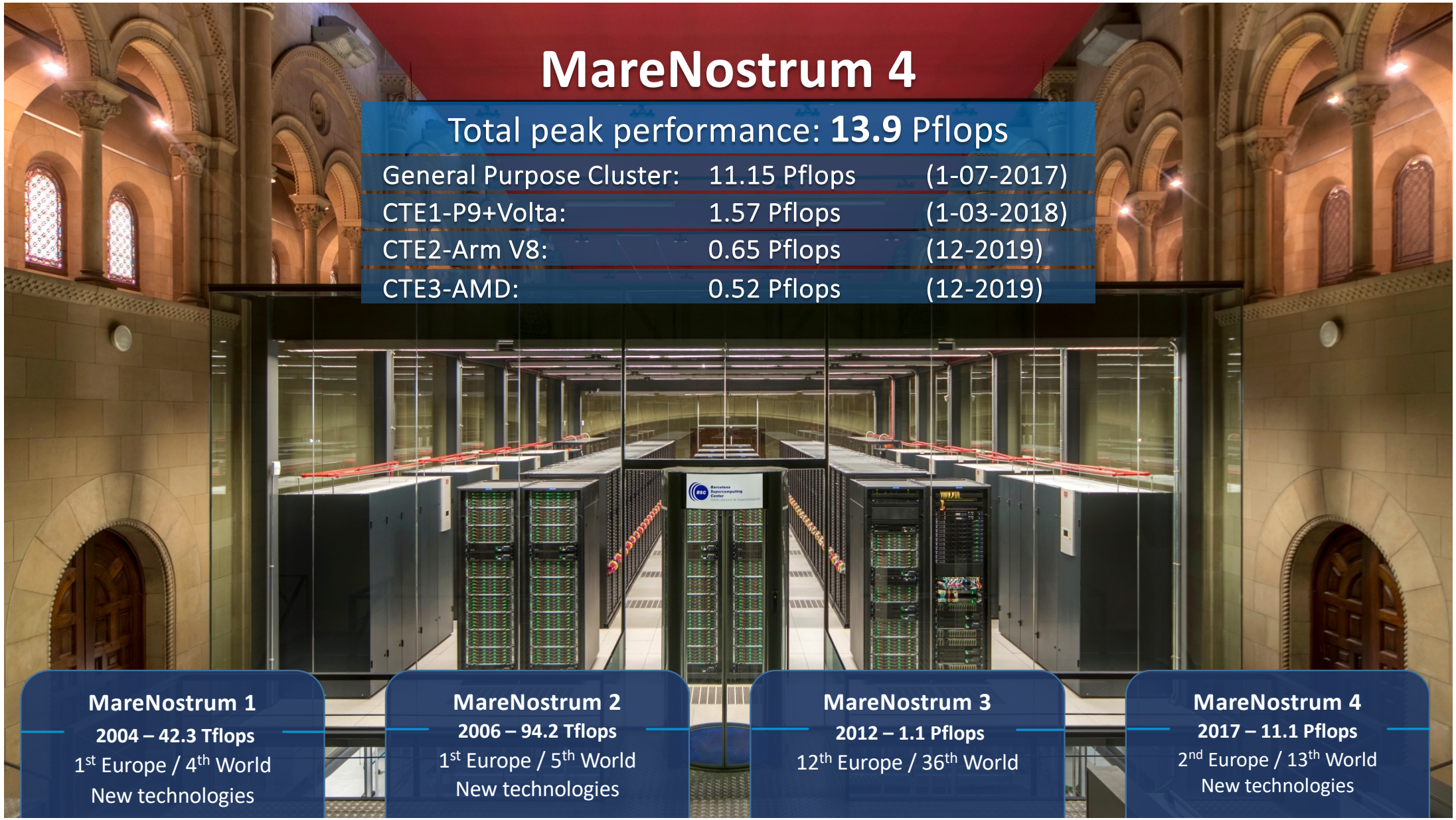
R&D in Computer, Life, Earth and
Engineering Sciences

PhD programme, technology
transfer, public engagement

BSC-CNS is a consortium that includes

- **Spanish Government**
- **Catalan Government** — Generalitat de Catalunya, Departament d'Empresa i Coneixement
- **Univ. Politècnica de Catalunya (UPC)** — Universitat Politècnica de Catalunya, BarcelonaTech

Barcelona Supercomputing Center
Centro Nacional de Supercomputación


MareNostrum 4
Total peak performance: 13.9 Pflops
General Purpose Cluster: 11.15 Pflops (1-07-2017)
CTE1-P9+Volta: 1.57 Pflops (1-03-2018)
CTE2-Arm V8: 0.65 Pflops (12-2019)
CTE3-AMD: 0.52 Pflops (12-2019)
MareNostrum 1
2004 – 42.3 Tflops
1st Europe / 4th World
New technologies
MareNostrum 2
2006 – 94.2 Tflops
1st Europe / 5th World
New technologies
MareNostrum 3
2012 – 1.1 Pflops
12th Europe / 36th World
MareNostrum 4
2017 – 11.1 Pflops
2nd Europe / 13th World
New technologies

Spanish Supercomputing Network (RES), since 2006

RED ESPAÑOLA DE SUPERCOMPUTACIÓN

www.res.es

Membership update: September 2022

HPC and data management resources for the scientific community

- 14 institutions
  - 16 supercomputers
  - 9 data management centres
- +22 PFlop/s combined capacity
- +20 PB storage in 2022 (and growing)
- +800 million CPU hours/year [2022]
- +1.000 regular users
- +200 scientific papers annually

- 3 HPC/A(calls per year
- 1 Data call per year
- Continuous call for AI small access
- Applications Support Teams

- Member of Spanish Unique Scientific and Technical Infrastructure network (**ICTS**)
- Access Committee and Users Committee
- EuroHPC National Competence Centre
- Coordinated by **BSC-CNS**

EuroHPC: towards European HPC technologies

EuroHPC-JU members:

Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden and Türkiye.

"A new legal and funding structure – the EuroHPC Joint Undertaking – shall acquire, build and deploy across Europe a world-class High-Performance Computing (HPC) infrastructure.

It will also support a research and innovation programme to develop the technologies and machines (hardware) as well as the applications (software) that would run on these supercomputers."

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EuroHPC Joint Undertaking

October 2022


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

MN5 Site preparation

- Public tender: CONOBR02019010OP
  - Awarded on 01/08/2019
  - Awarded Prize: 12.557.990 € (excluding VAT)
    - Including: project, construction and maintenance
  - Awardee: Climava SL
  - Formalisation on 26/11/2019

- Climava SL

  - Gisela Valderrama, Jaume Villa
  - https://www.climava.com

  - Global Technia Consulting
    - Lluis Gironella
    - https://www.b-global.tech

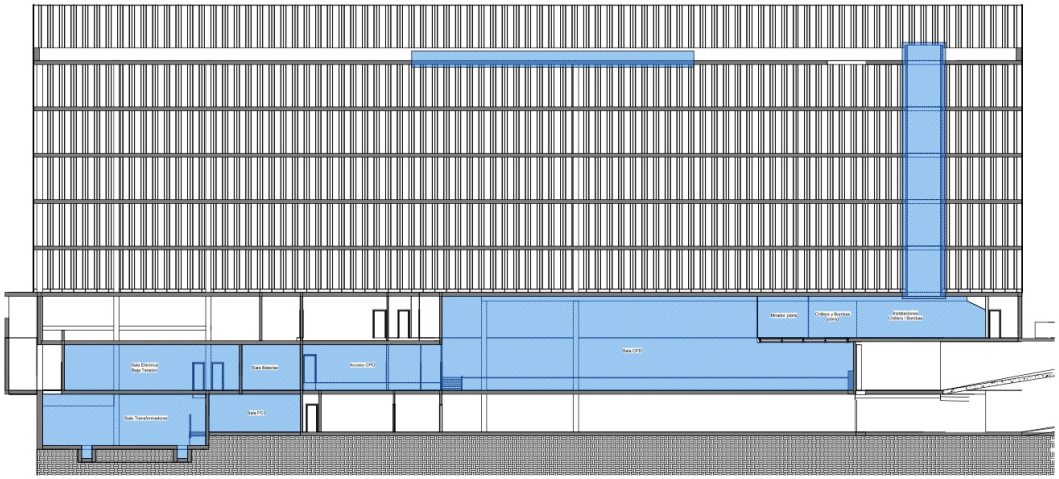
Expected date before covid19:	September 2020
Acceptance date:	April 2022

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

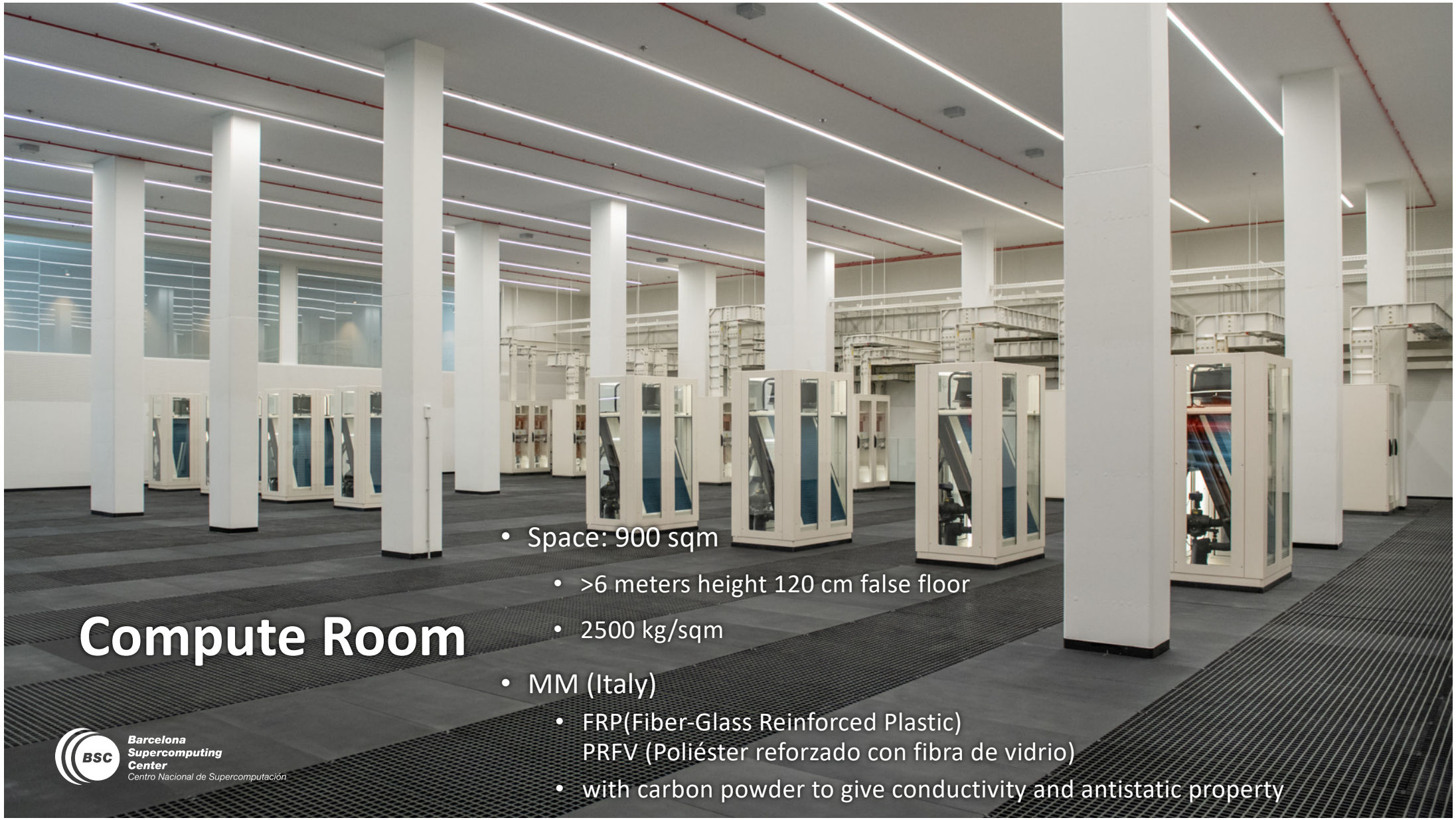
UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional

Space available for MN5

Floor		m^2	Total
P-3	Transformers	426	470
	Fire extinction	49	
P-2	Compute Room	847	1374
	Access to compute room	46	
	Batteries room	73	
	Low voltage room	408	
P-1	Chillers & Pumps room	466	711
	Riser / *"PATIO"*	9	
	Visitors area	236	
Roof		320	320
Total		rounded	2875

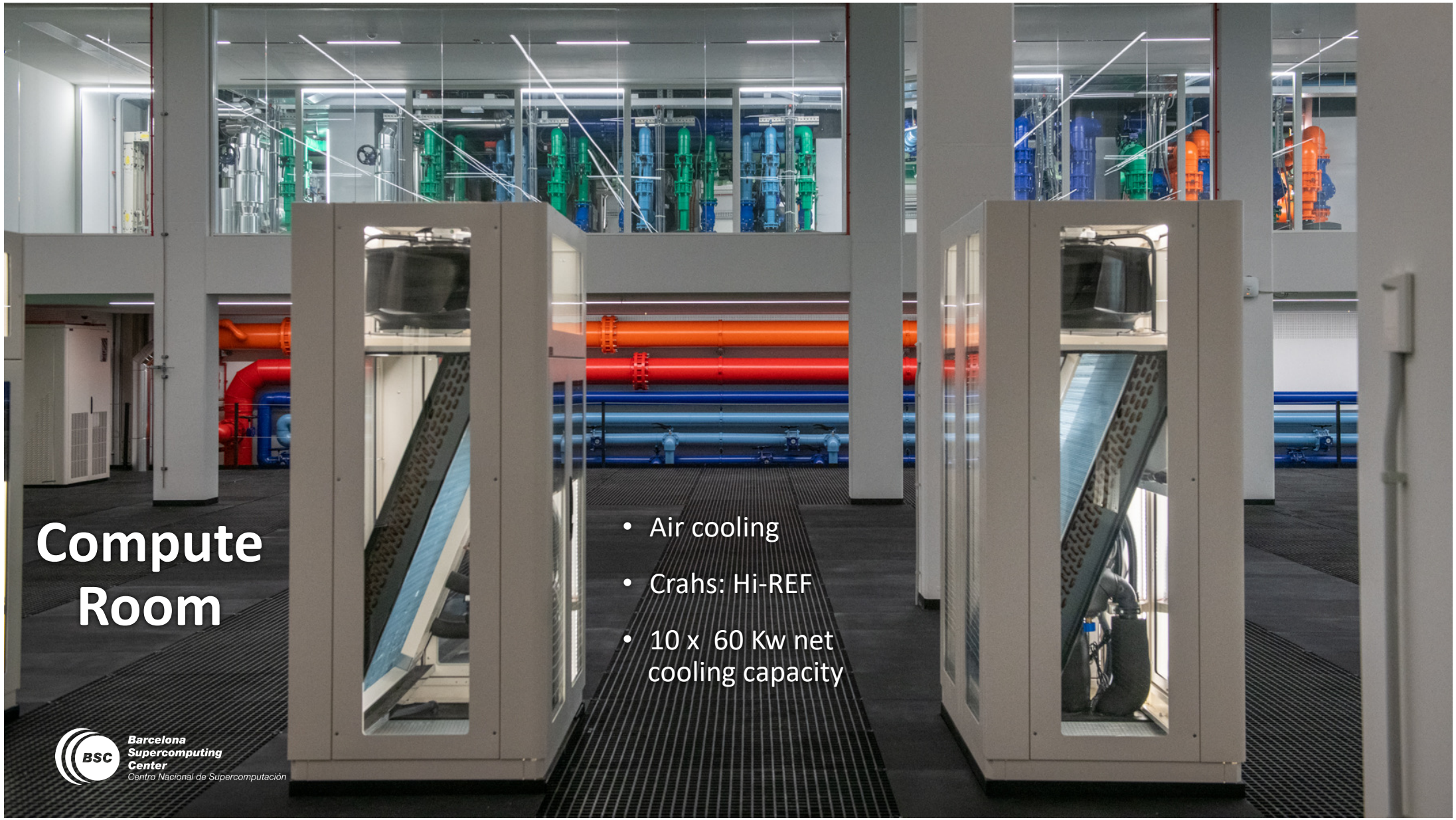


Barcelona Supercomputing Center
Centro Nacional de Supercomputación


Compute Room
• Space: 900 sqm
• >6 meters height 120 cm false floor
• 2500 kg/sqm
• MM (Italy)
• FRP(Fiber-Glass Reinforced Plastic)
PRFV (Poliéster reforzado con fibra de vidrio)
• with carbon powder to give conductivity and antistatic property
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación


Compute Room
• 3 water distribution loop
• Italsan
• PPR, Polypropylene
• About 4 km
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

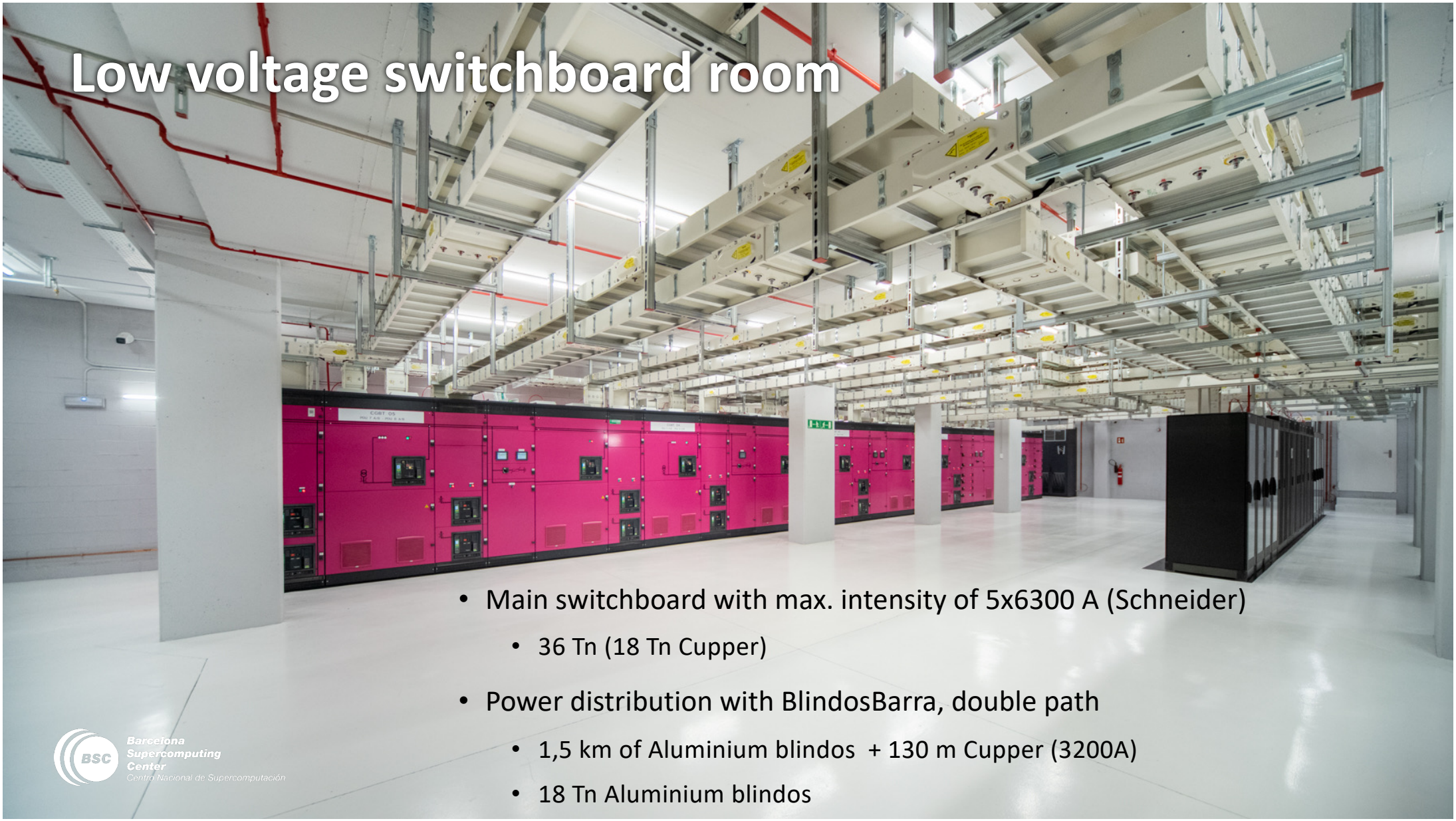

Compute Room
• Fire detection and extinction
• VESDA
• Water mist
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación


Compute Room
• Air cooling
• Crahs: Hi-REF
• 10 x 60 Kw net cooling capacity
Barcelona Supercomputing Center
Centro Nacional de Supercomputación


Compute Room
• Power distribution
• PDU: Schneider
• 8 x 2 x 3200 A/B (2 MW)
• 1x2x1600A UPS (1MW)
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación


Transformers
• 5 x TRANSFORMADOR 4150KVA VACUUM CAST FILLED DRY (ABB)
• 4150 kVA
• Primary: 25 kV
• Secondary: 420 V
• Frequency: 50Hz
• 3 phases
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Low voltage switchboard room

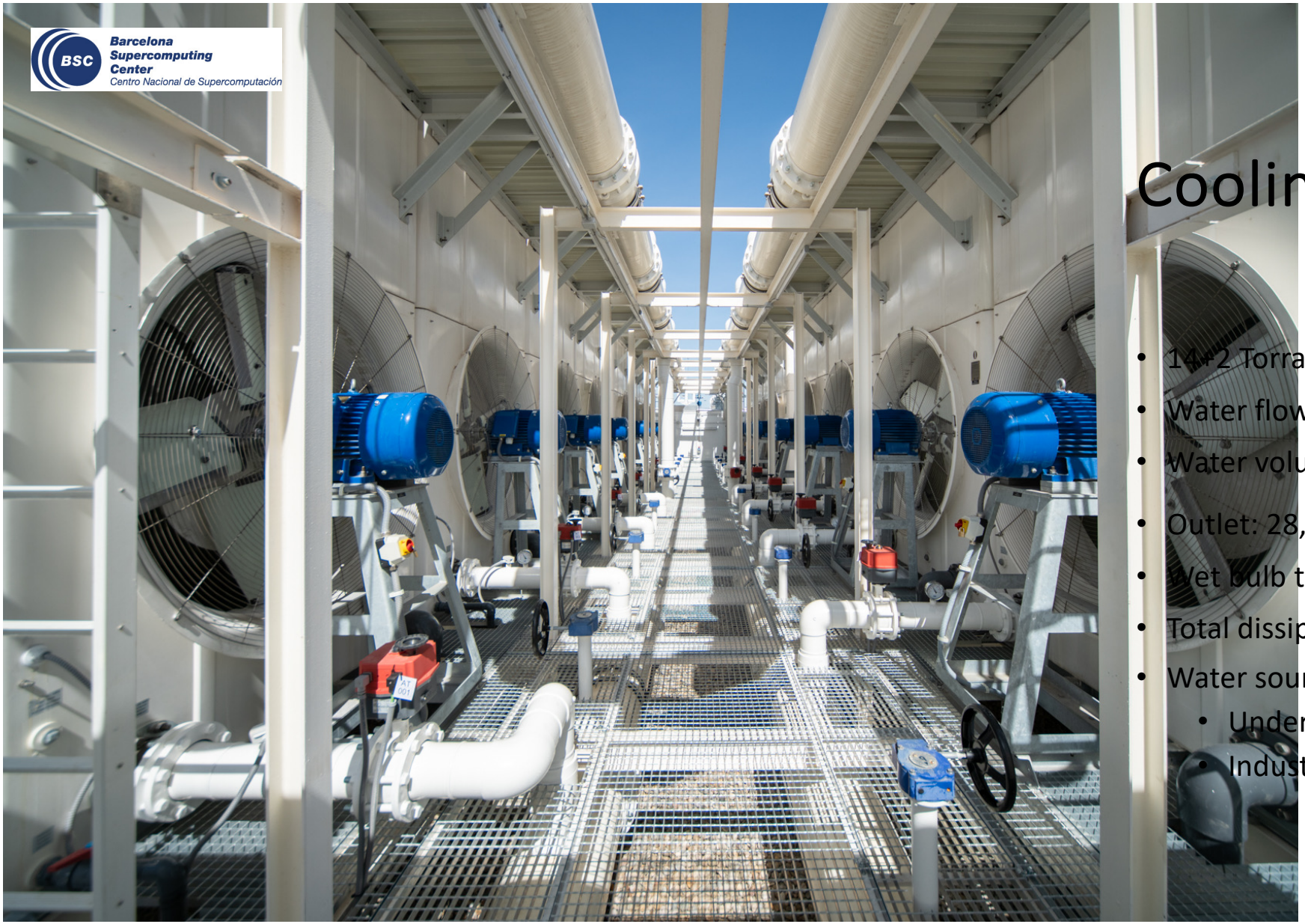


- Main switchboard with max. intensity of 5x6300 A (Schneider)
  - 36 Tn (18 Tn Cupper)
- Power distribution with BlindosBarra, double path
  - 1,5 km of Aluminium blindos + 130 m Cupper (3200A)
  - 18 Tn Aluminium blindos

UPS

- Huawei
- 2 x UPS 1MW, 2N. Lithium batteries
- 10 minutes durations


2020-10-09 12:53:06
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Cooling towers

- 14+2 Torraval CTFP-2436(SB)
- Water flow: 1500 m3/h
- Water volume: 60 m3
- Outlet: 28,1ºC , Inlet: 38,1ºC
- Wet bulb temperature: 25C
- Total dissipation power: 17300 kW
- Water source
  - Underground/phreatic water
  - Industrial water

Heat exchangers

- 6 (4+2) Heat exchangers T25-PFM
- Water flow: 1170 m3/h
- Water volume: 26 m3
- Temperatures
  - To tower: outlet: 28,1ºC , Inlet: 38,1ºC
  - To rack: outlet: 30ºC , Inlet: 40ºC
- Total dissipation power: 13500 kW

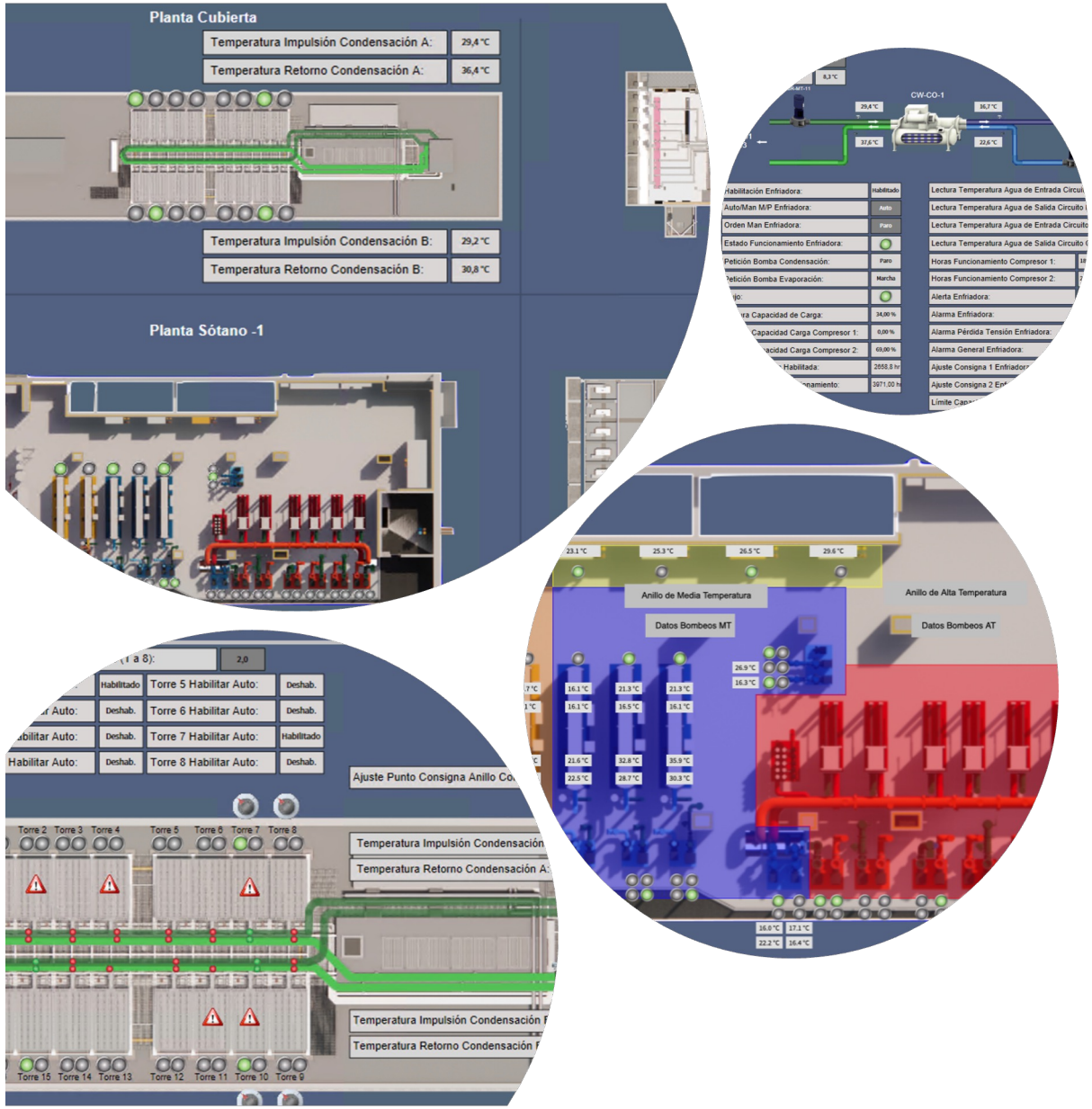
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Chillers

- 5 (2 MT + 1 LT+ 2) Chillers
- Water flow: 302 m3/h + 151 m3/h
- Water volume: 12 m3 + 8 m3
- Temperatures, separate loops
  - 16ºC – 26ºC
  - 8ºC – 14ºC
- To rack outlet: 30ºC , Inlet: 40ºC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación


Pumps
• Grundfos
• 36 in total
• 12 DLC
• 12 Medium
• 12 Low
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



BMS: Building Monitoring System

- Redundant Ethernet/TCP communications ring, with redundant Master Controllers.
- Fully bistable system, in case of loss of communications or failure of the management system, the infrastructure remains operational without any alteration.
- Option of operation in manual mode remotely controlled by an operator or 100% local manual from the plant itself.
- Management of alarms and warnings via SNMP (bidirectional).
- Storage of historical events, alarms and logs in event, alarm and log databases in SQL databases

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum 5. A European pre-exascale supercomputer

- **200 Petaflops** peak performance (200 x 10^{15})*
- **Experimental platform** to create supercomputing technologies “made in Europe”
- **217 M€** of Total Cost Ownership

EuroHPC Joint Undertaking

Hosting Consortium:

Spain Portugal Türkiye Croatia

European Commission

* *At the time of call for HE, peak performance expected of 200 Petaflops*
- *At the time of tender publications, minimum aggregated sustained HPL of 205 Petaflops*
- *Contract signed on July 2022, with a aggregated sustained performance HPL of 204,64 and peak performance of 314,22 PF*

Generalitat de Catalunya
Departament d'Empresa i Coneixement

Gobierno de España
Ministerio de Ciencia, Innovación y Universidades

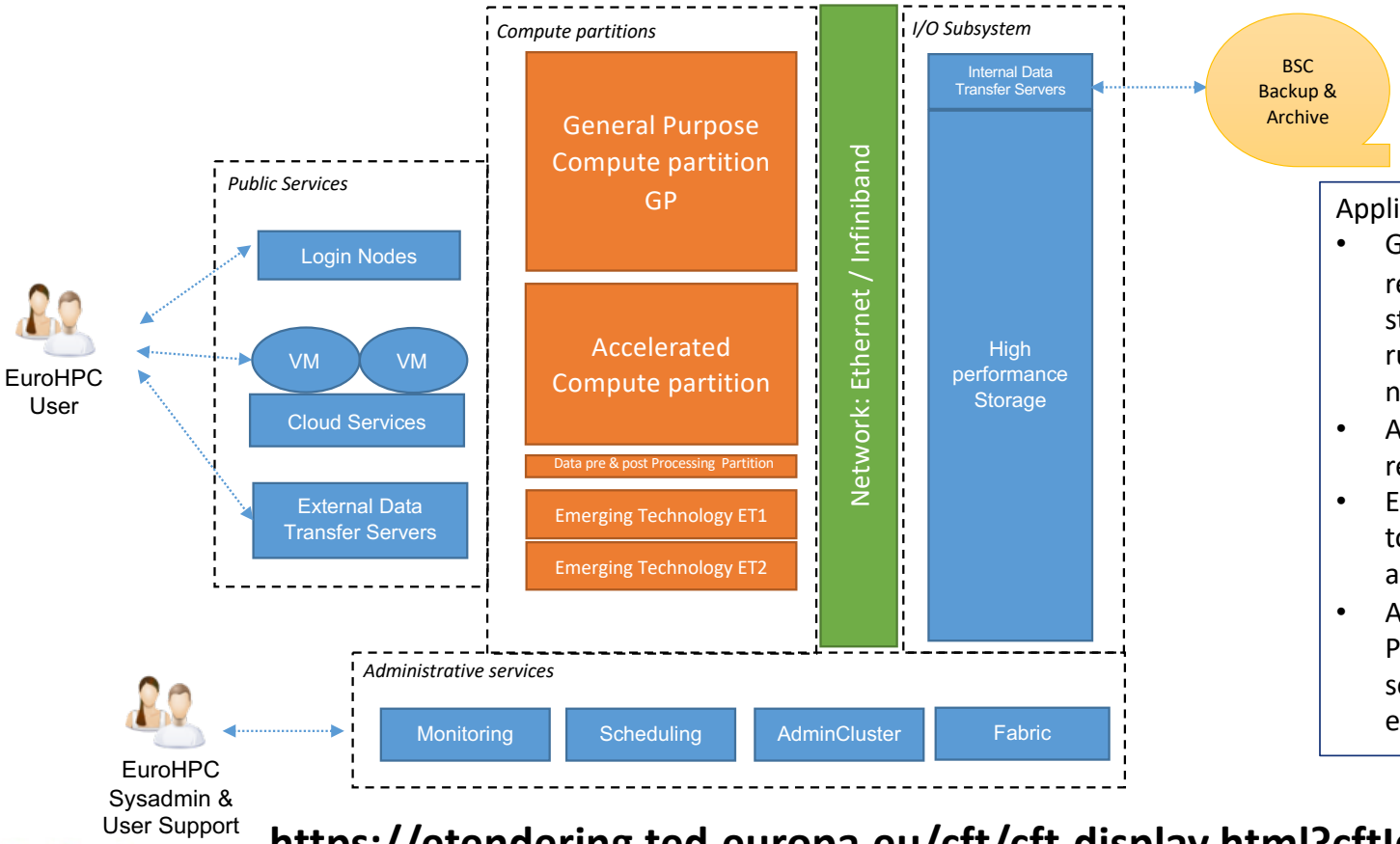
UPC Universitat Politècnica de Catalunya BarcelonaTech

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union’s Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal, and Türkiye

EuroHPC Joint Undertaking

MareNostrum5 concept



Hosting Consortium:

Spain Portugal Türkiye

Applications:

- General purpose partition, open to all researchers with MPI, OpenMP codes, standard HPC codes. Scalable machine to run codes with high scalability, thousands of nodes.
- Accelerated partition: Any GPU application ready to scale to thousands of GPUs
- Emerging technologies: prepare workloads to exascale era, exascale technology assessment
- Any domain with workflows mixing General Purpose and GPU, e.g. Earth science, Life science, Engineering, AI and AI driven executions.

https://etendering.ted.europa.eu/cft/cft-display.html?cftId=9758

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

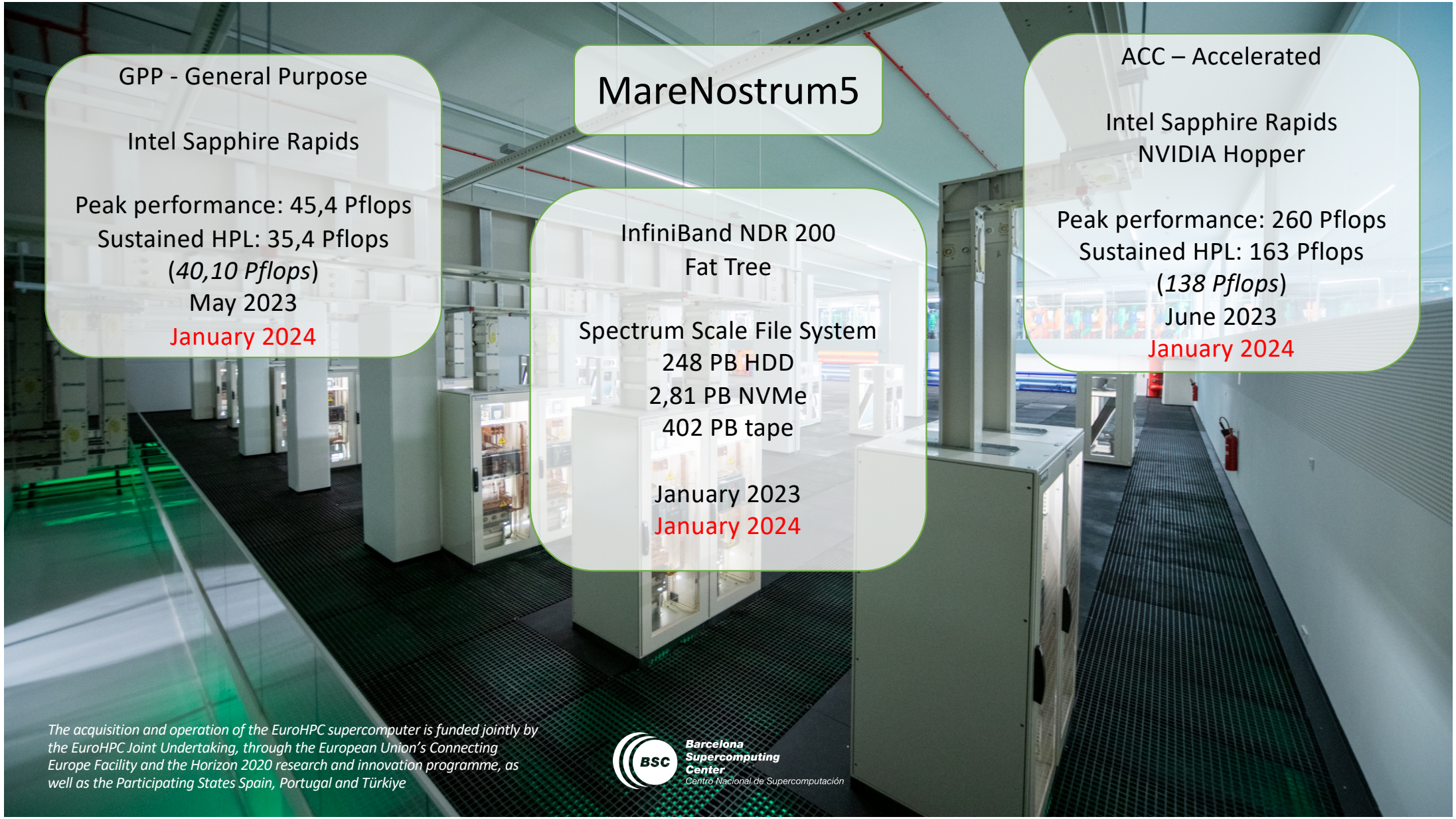
The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal, and Türkiye

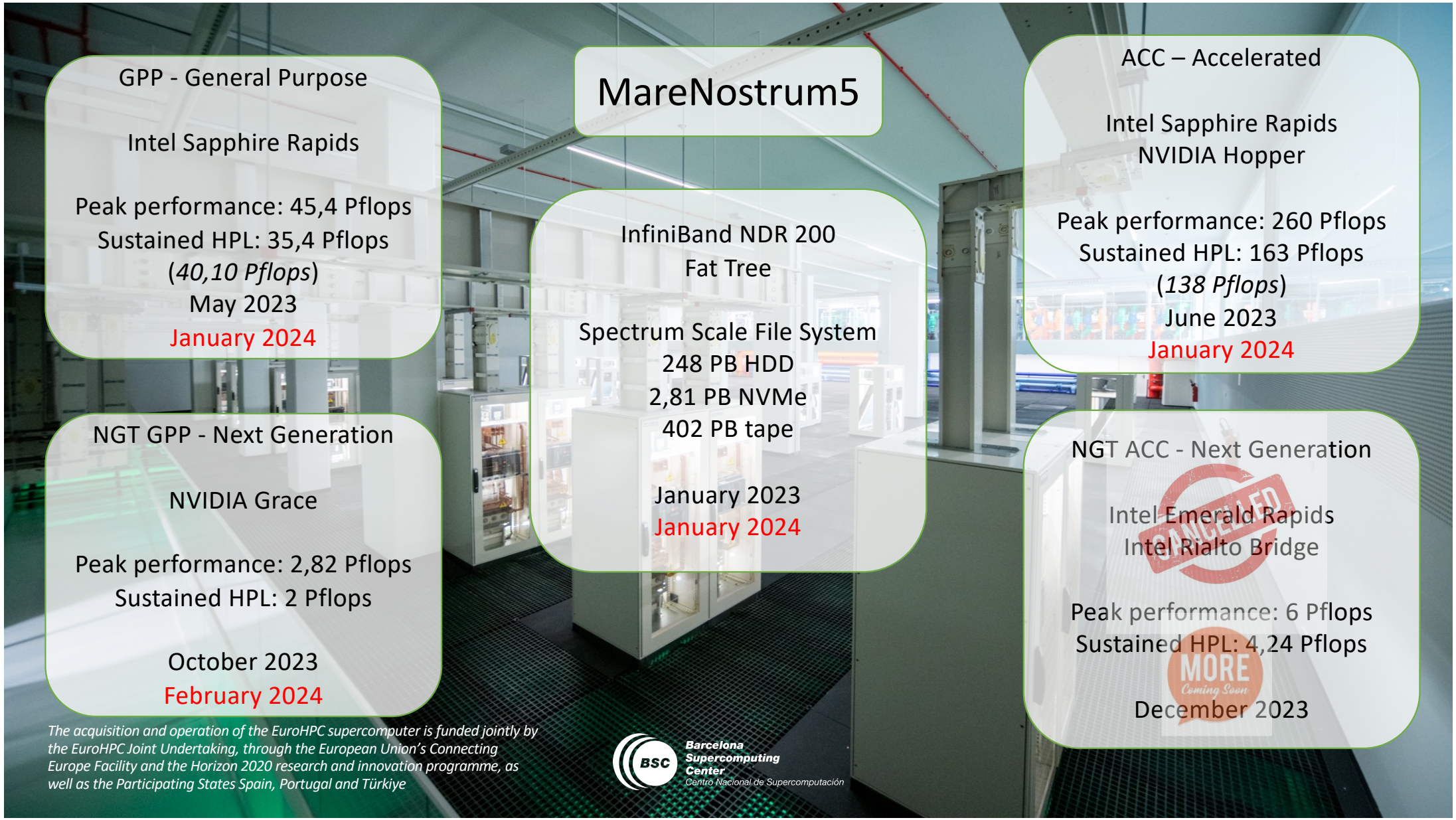
The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye

BSC **Barcelona Supercomputing Center**
Centro Nacional de Supercomputación

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye

BSC **Barcelona Supercomputing Center**
Centro Nacional de Supercomputación


GPP - General Purpose
Intel Sapphire Rapids
Peak performance: 45,4 Pflops
Sustained HPL: 35,4 Pflops
(40,10 Pflops)
May 2023
January 2024
MareNostrum5
InfiniBand NDR 200
Fat Tree
Spectrum Scale File System
248 PB HDD
2,81 PB NVMe
402 PB tape
January 2023
January 2024
ACC – Accelerated
Intel Sapphire Rapids
NVIDIA Hopper
Peak performance: 260 Pflops
Sustained HPL: 163 Pflops
(138 Pflops)
June 2023
January 2024
The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

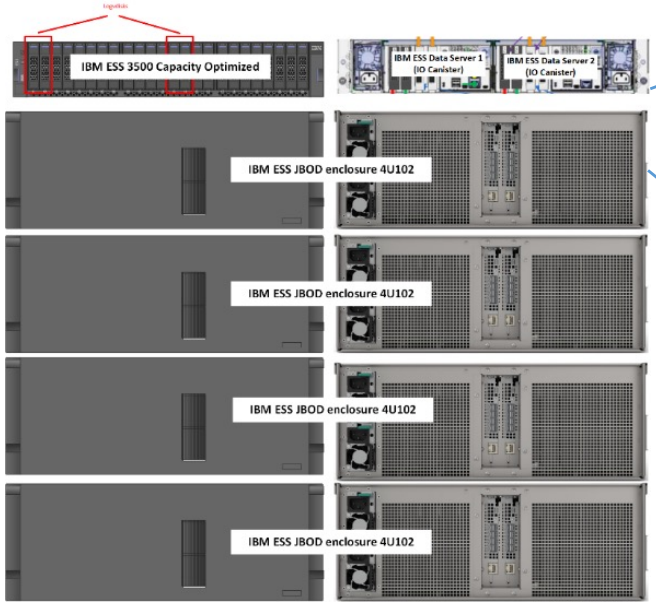

MareNostrum5
GPP - General Purpose
Intel Sapphire Rapids
Peak performance: 45,4 Pflops
Sustained HPL: 35,4 Pflops
(40,10 Pflops)
May 2023
January 2024
InfiniBand NDR 200
Fat Tree
Spectrum Scale File System
248 PB HDD
2,81 PB NVMe
402 PB tape
January 2023
January 2024
ACC – Accelerated
Intel Sapphire Rapids
NVIDIA Hopper
Peak performance: 260 Pflops
Sustained HPL: 163 Pflops
(138 Pflops)
June 2023
January 2024
NGT GPP - Next Generation
NVIDIA Grace
Peak performance: 2,82 Pflops
Sustained HPL: 2 Pflops
October 2023
February 2024
NGT ACC - Next Generation
CANCELLED
Intel Emerald Rapids
Intel Rialto Bridge
Peak performance: 6 Pflops
Sustained HPL: 4,24 Pflops
MORE
Coming Soon
December 2023
The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

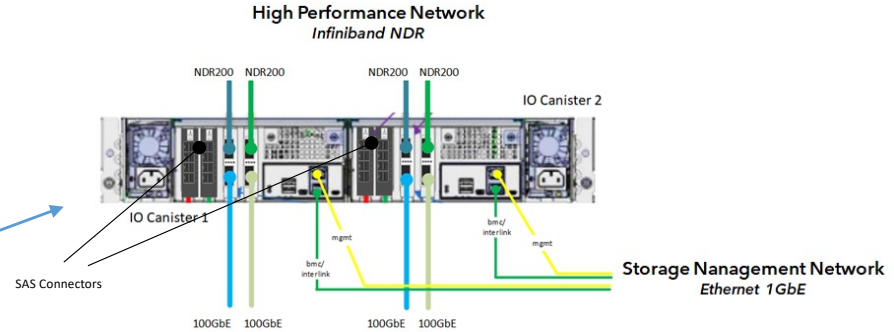
Storage global numbers

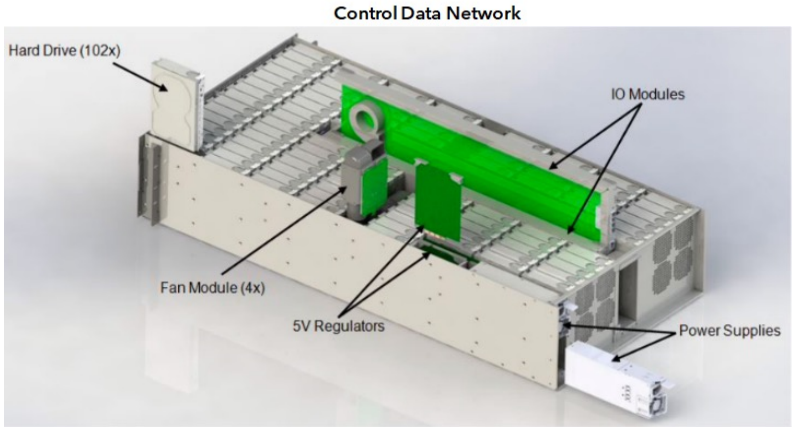
Net Capacity (HDD)	248 PB
Net Capacity metadata (Flash)	2.8 PB
Performance (HDD)	1.6 TB/s read and 1.2 TB/s write
Performance (Flash)	600 GB/s read or write
Racks	25
Power consumption	400-550 kW
HDDs	20400 x 18TB NL-SAS 3.5”
NVMe Flash	312 x 15.36 TB
Tape Library net capacity	402 PB

Storage: Data Module

- 50 x Data Module: ESS 3500 Capacity
  - 2x Data Servers: AMD Rome 48c and 512 GB RAM
  - 4x JBOD enclousures with 102 disk, 18 TB each



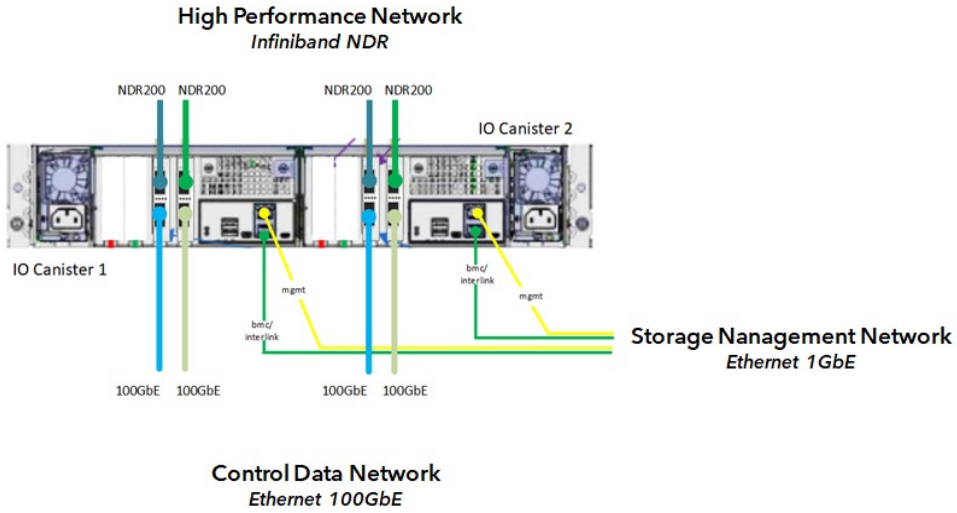


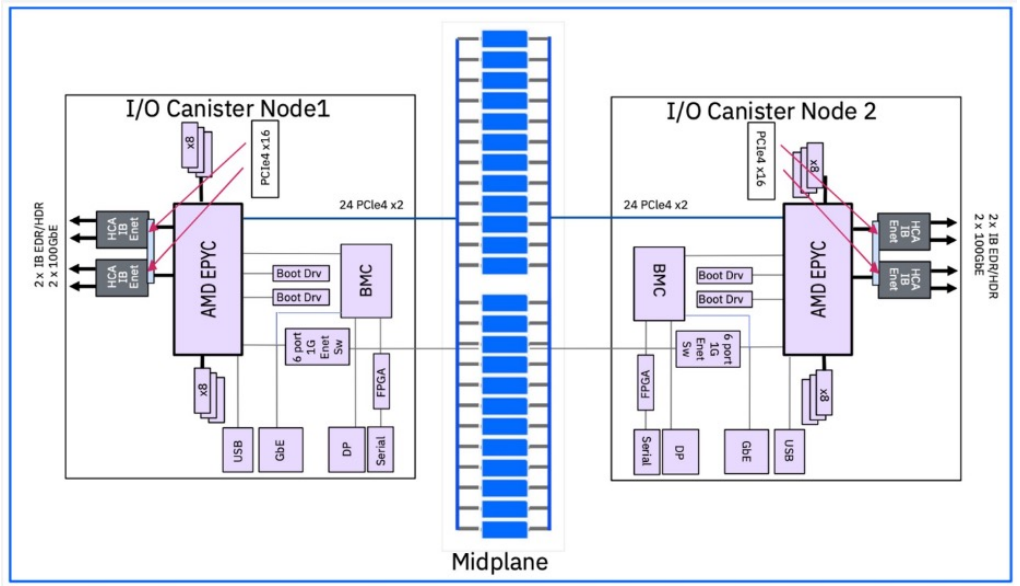


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Storage: Metadata Module

- 13x Metadata Modules: ESS 3500 Performance
  - 2x Metadata Servers: AMD Rome 48c and 512 GB RAM
  - 24x NVMe all-flash drives, 15.36 TB each



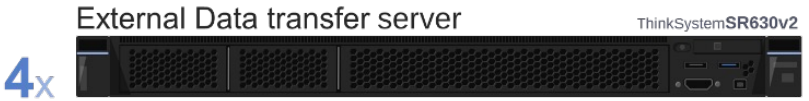


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

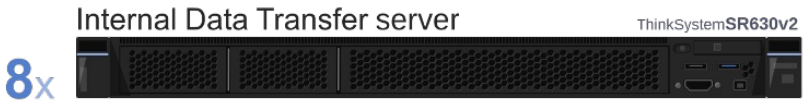
Storage Services

- 4x Export servers
  - Provide Access through NFS, CIFS and Object

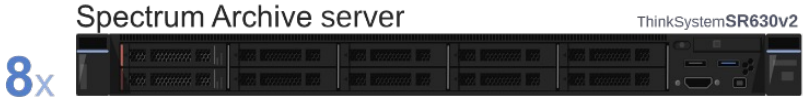
- 4x External Data Transfer
  - Provide Transfer data services from/to Internet



- 8x Internal Data Transfer
  - Provide Internal data transfer services between storages
  - Used by dtcommands

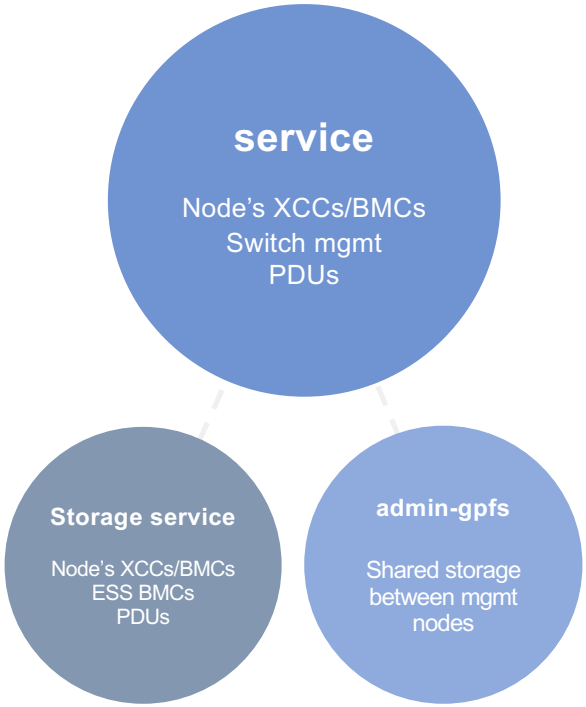


- 8x Archive servers
  - Implements HSM policies to migrate or recover data from tapes



Ethernet Network
System VLANs

Networks



Management

Cluster management, deployment, monitoring
Node OS
Scheduling

External BSC LAN

External access

Control Data Network

GPFS control traffic
Backup to the high-speed network

High performance Network

Infiniband NDR compute network

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Ethernet network: Switches

Nvidia**SN4600**

64 200GbE QSFP56 ports

128 100/50/25/10/1GbE

425ns latency

8.4B pps

Line-Rate switching

L2/L3

600W Typical

Nvidia**SN3700V**

32 200GbE QSFP56 ports

64 100GbE ports

128 25/10/1GbE ports

425ns latency

8.33B pps

Line-Rate switching

L2/L3

250W Typical

Nvidia**SN3700C**

32 100GbE QSFP28 ports

128 25/10/1GbE ports

425ns latency

4.76B pps

Line-Rate switching

L2/L3

242W Typical

Nvidia**SN2410**

8 100GbE QSFP28 ports

48 25/10/1GbE SFP+ ports

300ns latency

2.97B pps

Line-Rate switching

L2/L3

165W Typical

Nvidia**AS4610-54t**

48 1GbE RJ45 ports

4 10GbE SFP+ ports

4us latency

Line-Rate switching

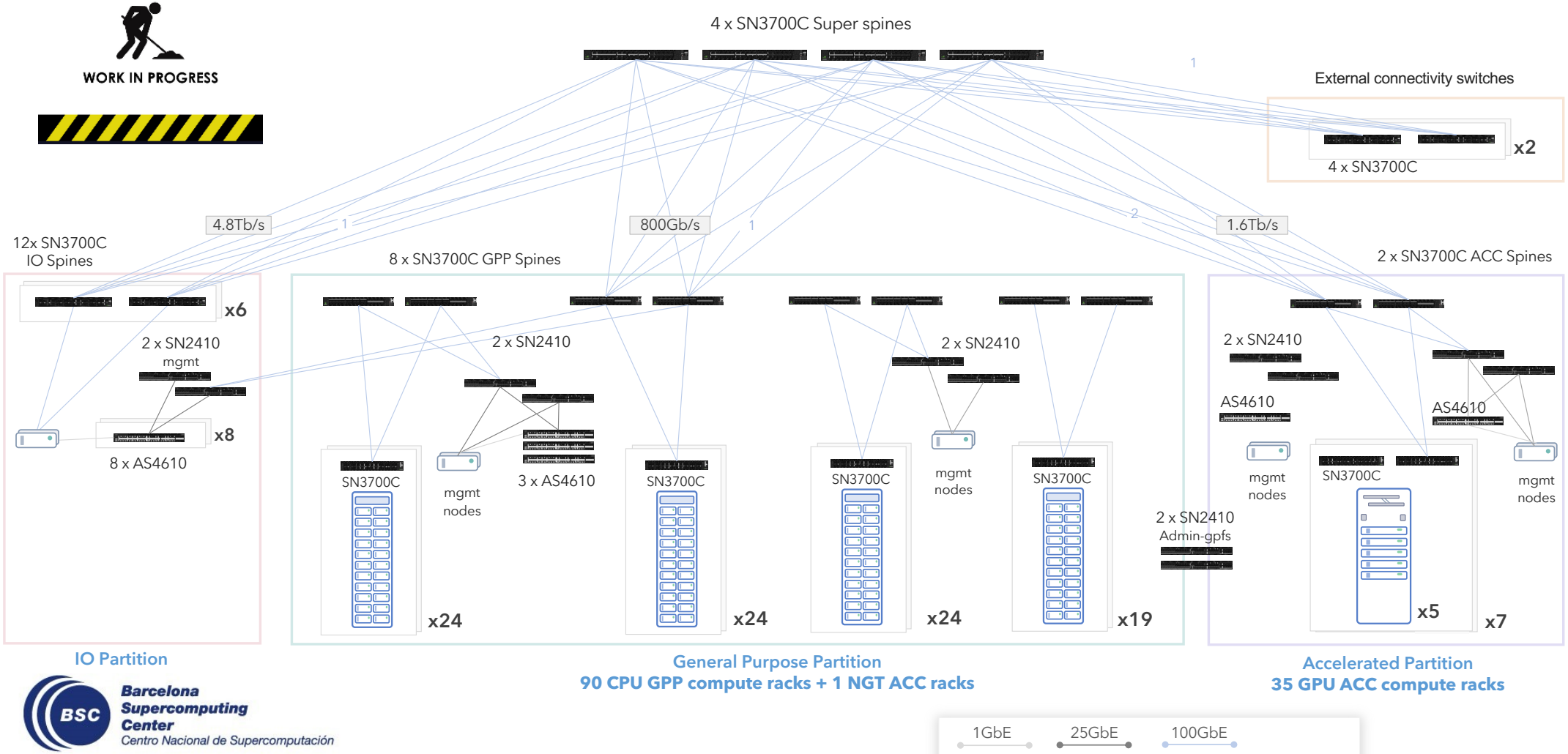
L2/L3

90W Typical

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Ethernet network overview

WORK IN PROGRESS



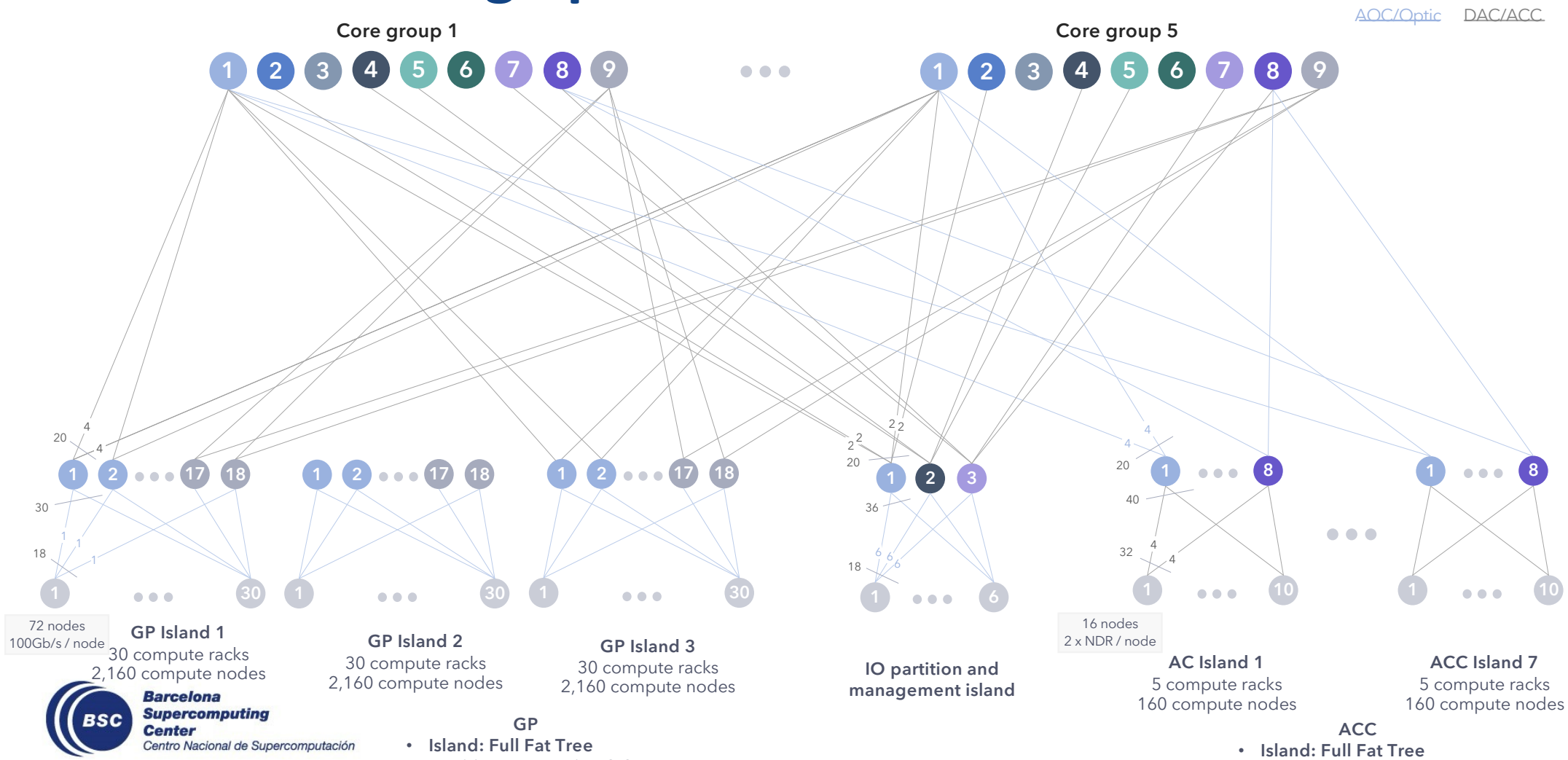
IO Partition

General Purpose Partition
90 CPU GPP compute racks + 1 NGT ACC racks

Accelerated Partition
35 GPU ACC compute racks

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

High performance network



GP Island 1
30 compute racks
2,160 compute nodes

GP Island 2
30 compute racks
2,160 compute nodes

GP Island 3
30 compute racks
2,160 compute nodes

IO partition and management island

AC Island 1
5 compute racks
160 compute nodes

ACC Island 7
5 compute racks
160 compute nodes

GP
- **Island: Full Fat Tree**
- **Partition: contention 3:2**

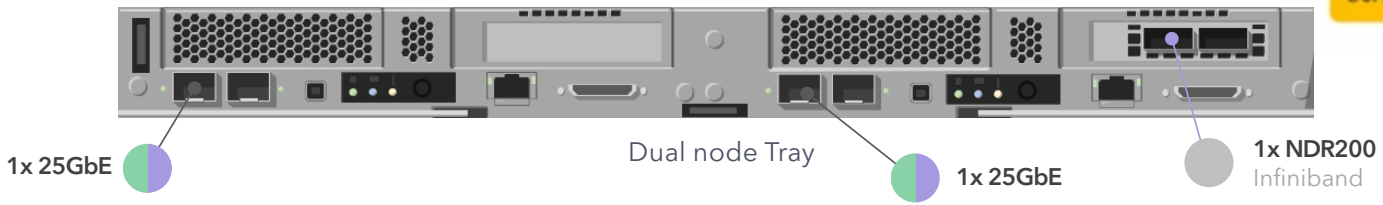
ACC
- **Island: Full Fat Tree**
- **Partition: contention 2:1**

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

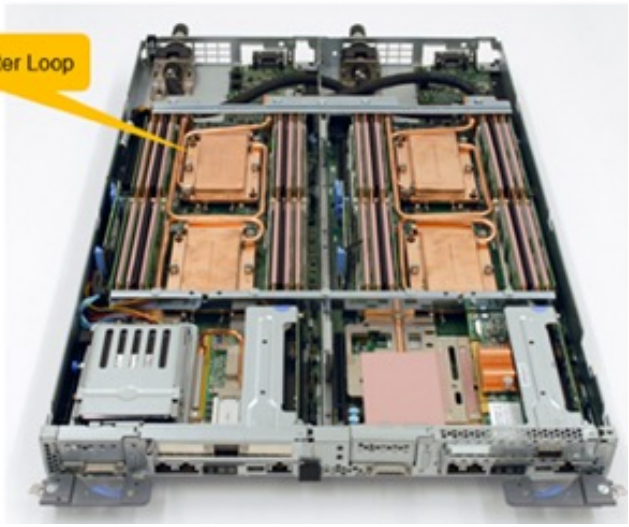
Compute partitions overview

		Racks	Cooling	Nodes		Provider	Processor/Accelerator		Memory	PFlops (HPL) Contractual achieved		Local Drive	High-Perf. Network
				Total	per rack								
Main	General Purpose	89	DLC +RDHX	6192	72 (6x6x2)	Lenovo	2x Intel Sapphire R. 8480+	56c @ 2GHz	>2GB/core 256GB DDR5	35.43	>205	960GB NVMe	1x NDR200 Shared by 2 nodes
				216					>8GB/core 1024GB DDR5	*40.10*			
		1		72			2x Intel Sapphire R. 03H-LC	56c @ 1.7GHz	> 0.5GB HBM/core 128GB HBM + 32GB DDR5	0.34			
	Accelerated	35	DLC	1120	32	Atos	2x Intel Sapphire R. 8460Y+	32c @ 2.3GHz	512GB	163		480GB NVMe	4x NDR200
							4x Nvidia Hopper 64GB HBM			*138*			
Next Gen	General Purpose	7	AC +RDHX	408	68	Atos	2x Nvidia Grace	72c @ 2.6GHz	240GB LPDDR5	2 --		128GB NVMe	1x NDR200
	Accelerated	1	DLC +RDHX	24	24	Lenovo	2x Intel Emerald R.	48c	512GB DDR5	4.24 --		960GB NVMe	2x NDR
							4x Intel Rialto Bridge 128GB HBM2E						

General Purpose Compute Node



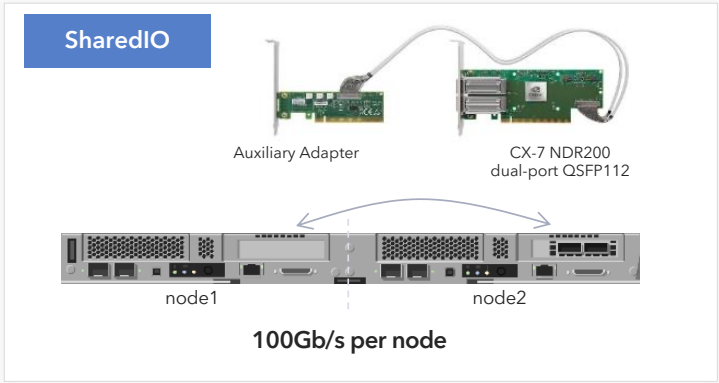
Dual Motherboard Tray



6,192x GPP Compute node (256GB RAM 16x16 GB 4800MHz DDR5)

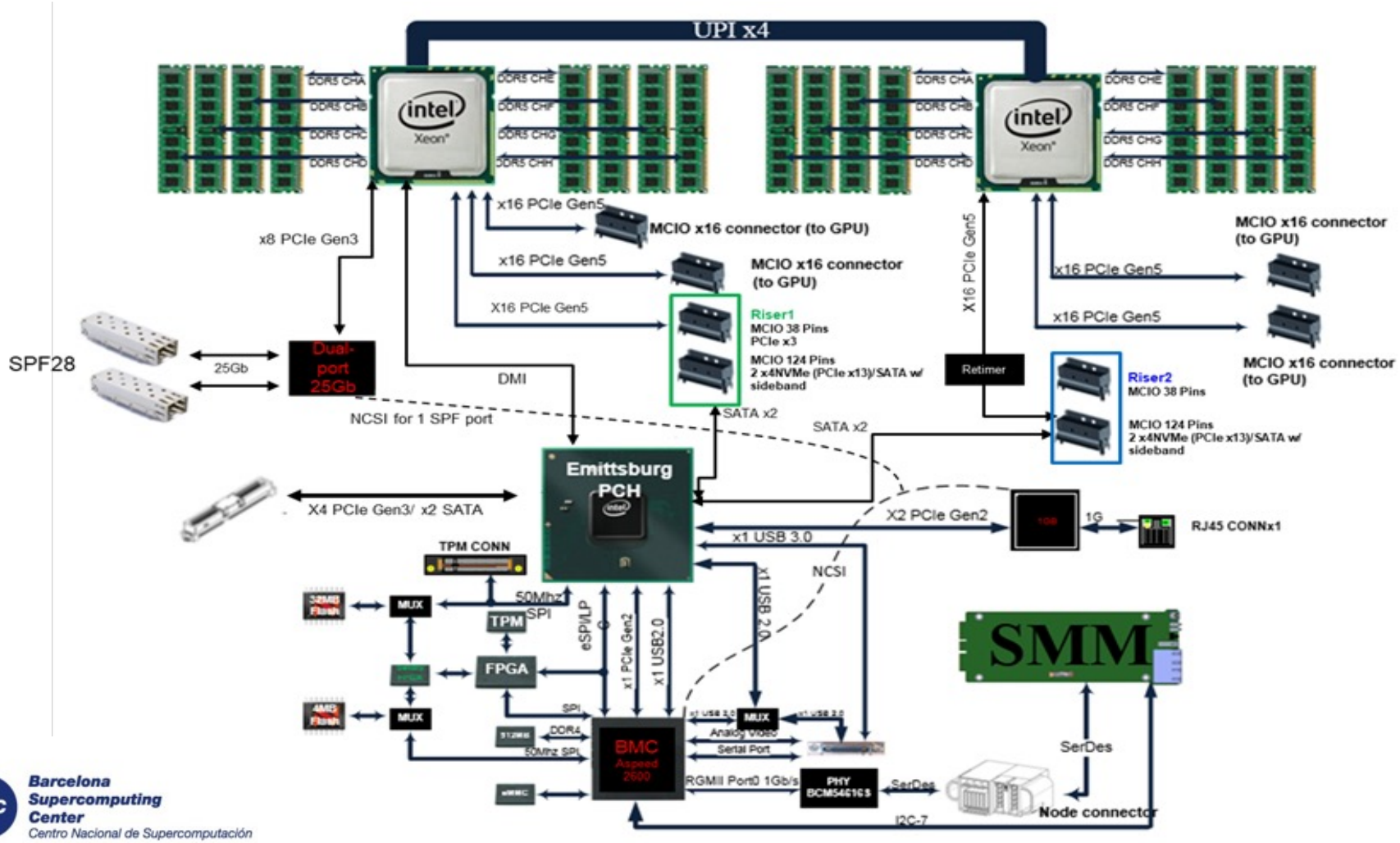
216x GPP Compute node (1TB RAM 16x64 GB 4800MHz DDR5)

72x GPP HBM Compute node (32GB RAM 2x16 GB + 128 GB HBM2)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

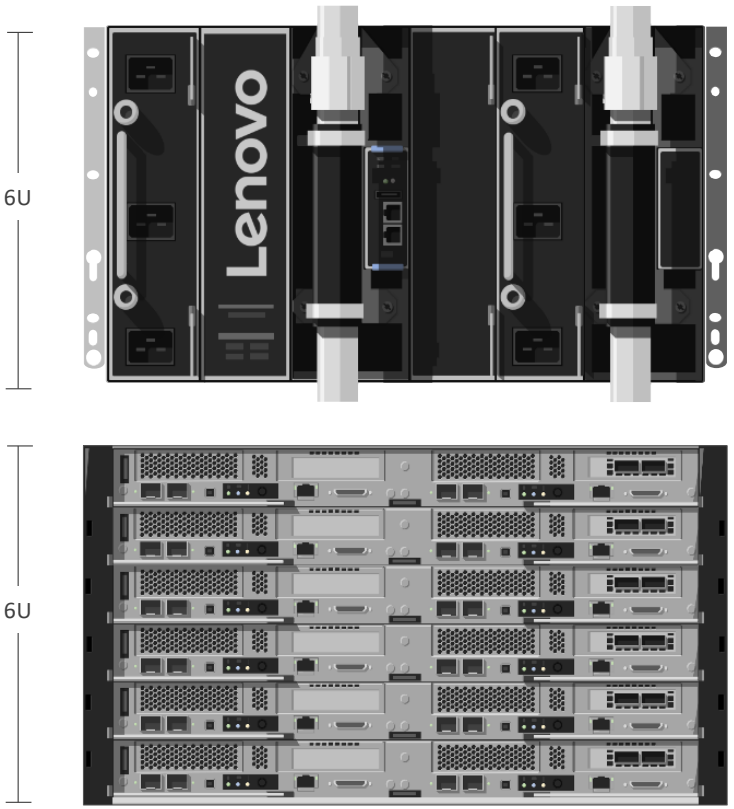
General Purpose Motherboard



General purpose chassis

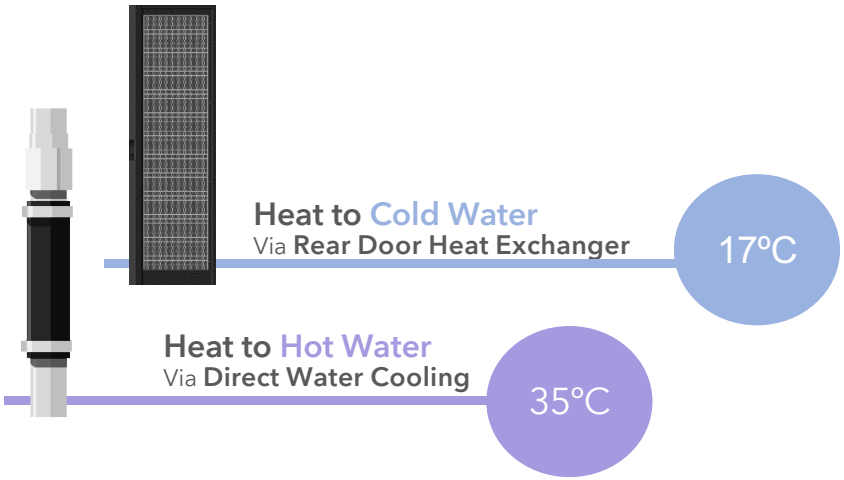
- 12 Nodes per chassis
- 2 x 7200W liquid cooled Power supplies
  - 3x internally 2x 2400W PSU each
- Normal consumption 85% of HPL

Chassis Power consumption HPL	Type of nodes
11.4 kW	256GB RAM
12 kW	1TB RAM
10.4 kW	HBM



ThinkSystemDW612 Chassis

General purpose rack



256GB rack - 69.6kW **HPL**

66.7 2.9

1TB rack - 72.8kW **HPL**

69.9 2.9

HBM rack - 63.5kW **HPL**

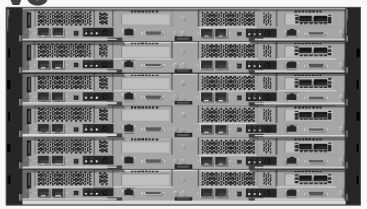
60.6 2.9

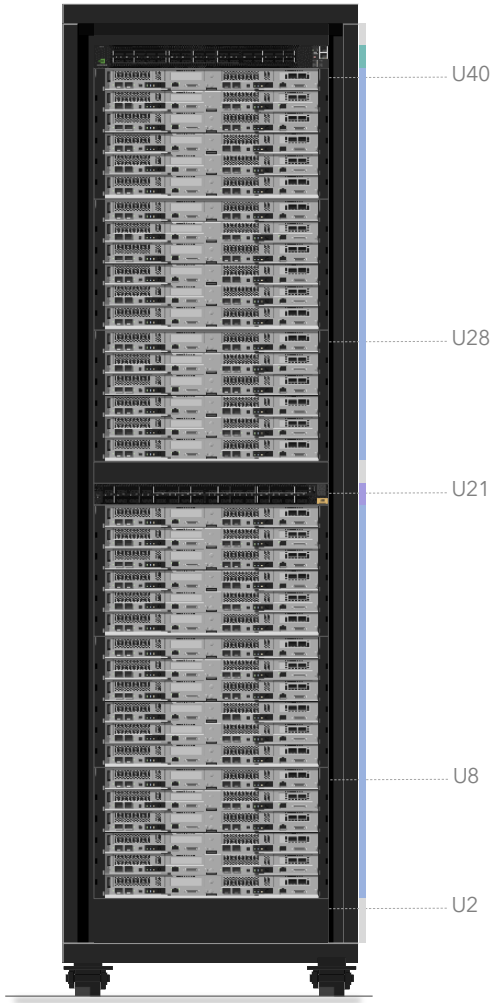
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

1x Mellanox **SN3700C**

1x Mellanox **QM9790**

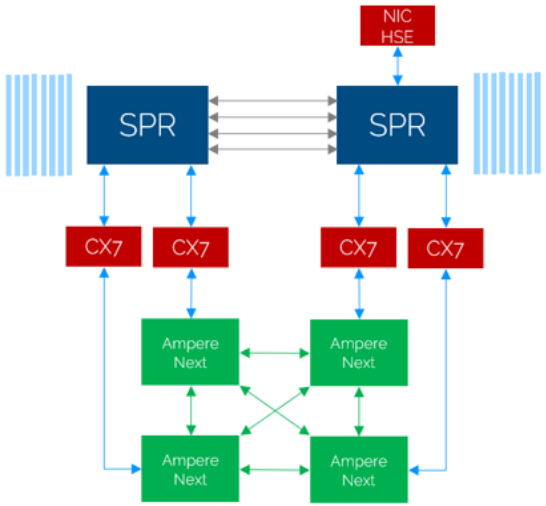
72x Lenovo **SD650 v3**



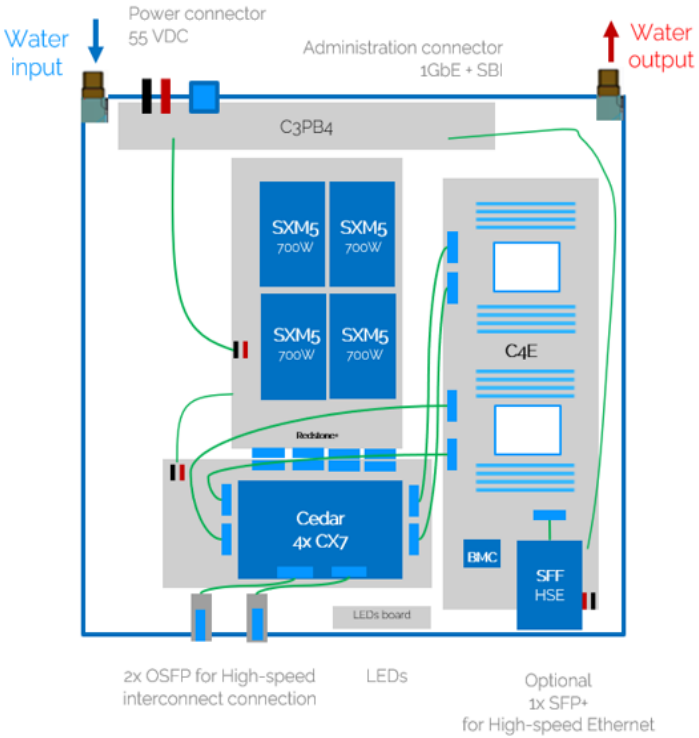


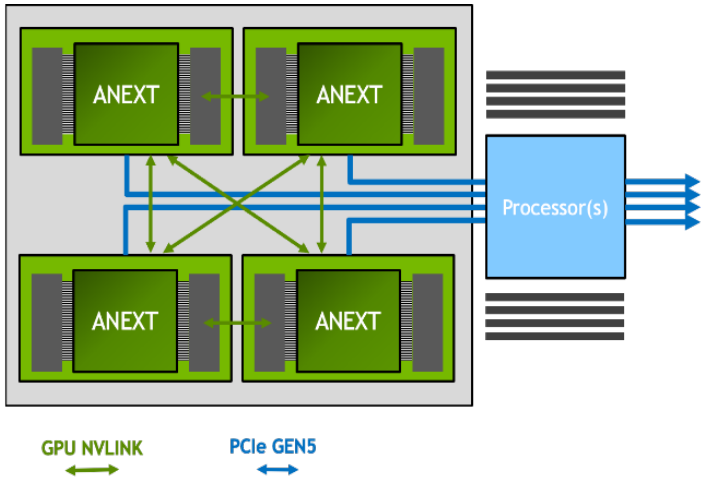
ACC Compute Node

Logical Architecture

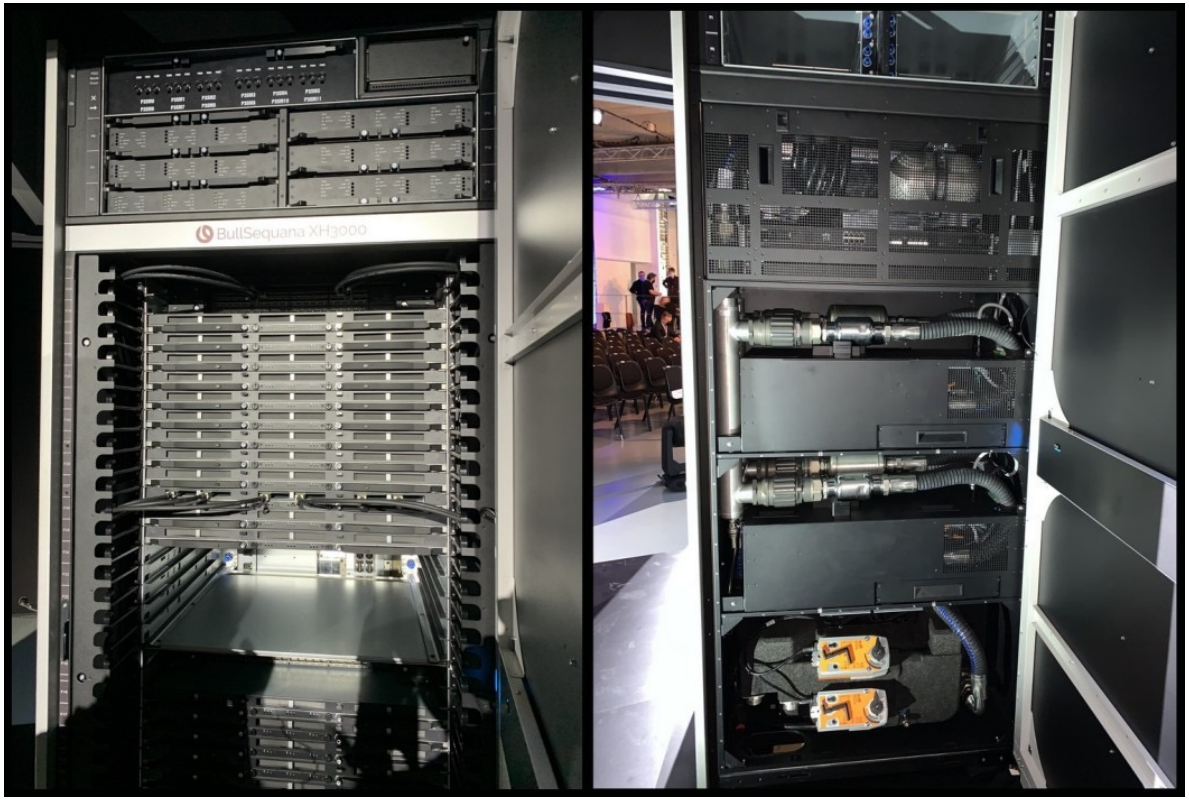


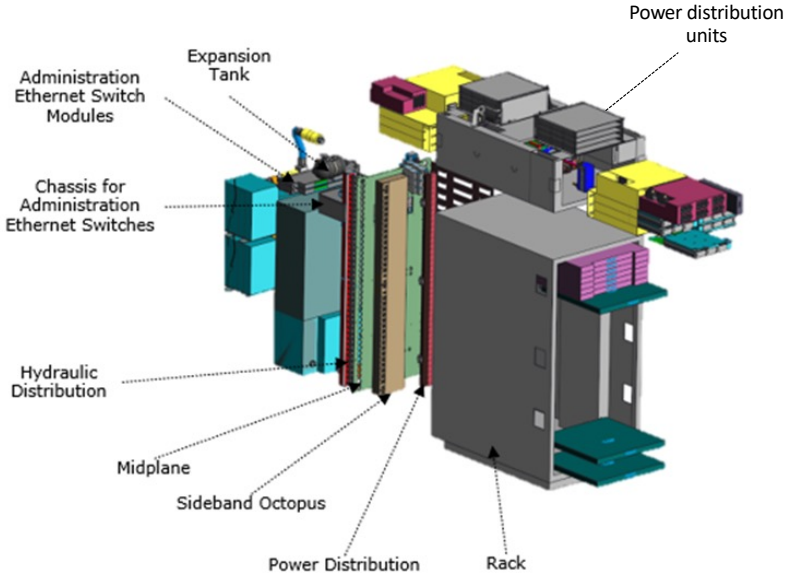
Physical Architecture

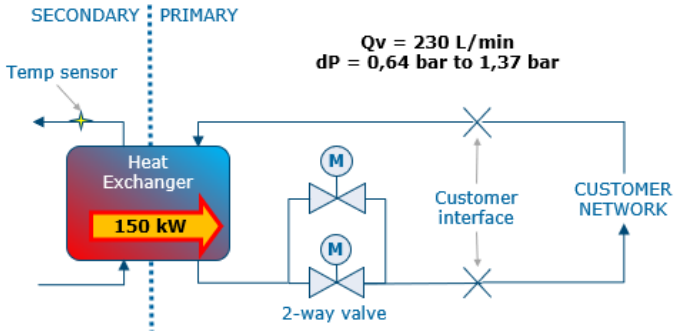




ACC Compute rack



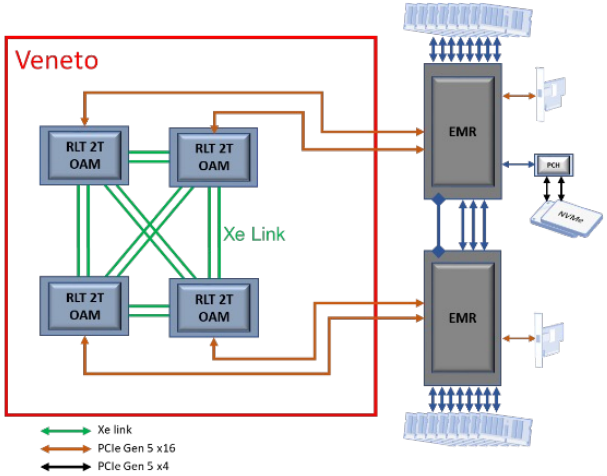


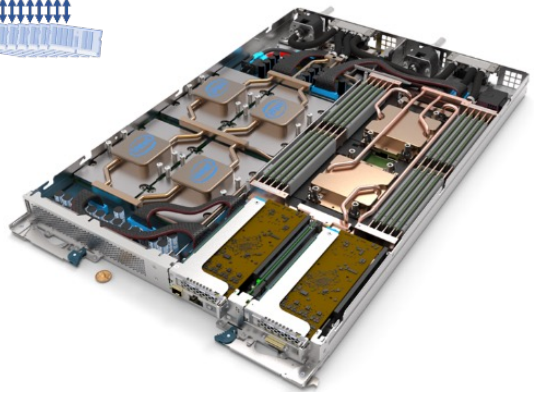


Next Generation Compute

- General purpose
  - 408 compute nodes
  - NVIDIA Grace processor
  - Air-cooled chassis
  - Some immersion cooling pods

- Accelerated
  - 24 compute nodes
  - Emerald Rapids + Intel Rialto





Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Other nodes

- 4x Logins per compute partition
  - Same as a compute node of that partition
- 10x Nodes data Pre & Post processing
  - 2x Intel Sapphire 8480+ 56c 2GHz
  - 2 TB Main memory
  - 2x 3.2 TB NVMe disk
  - 1x NDR200 Interface
- 18x Virtualization Servers
  - 2x Intel 6342 24c 2.8 GHz
  - 512 GB RAM

Login node

4x

Data Pre-post processing server ThinkSystem**SR650v3**

10x

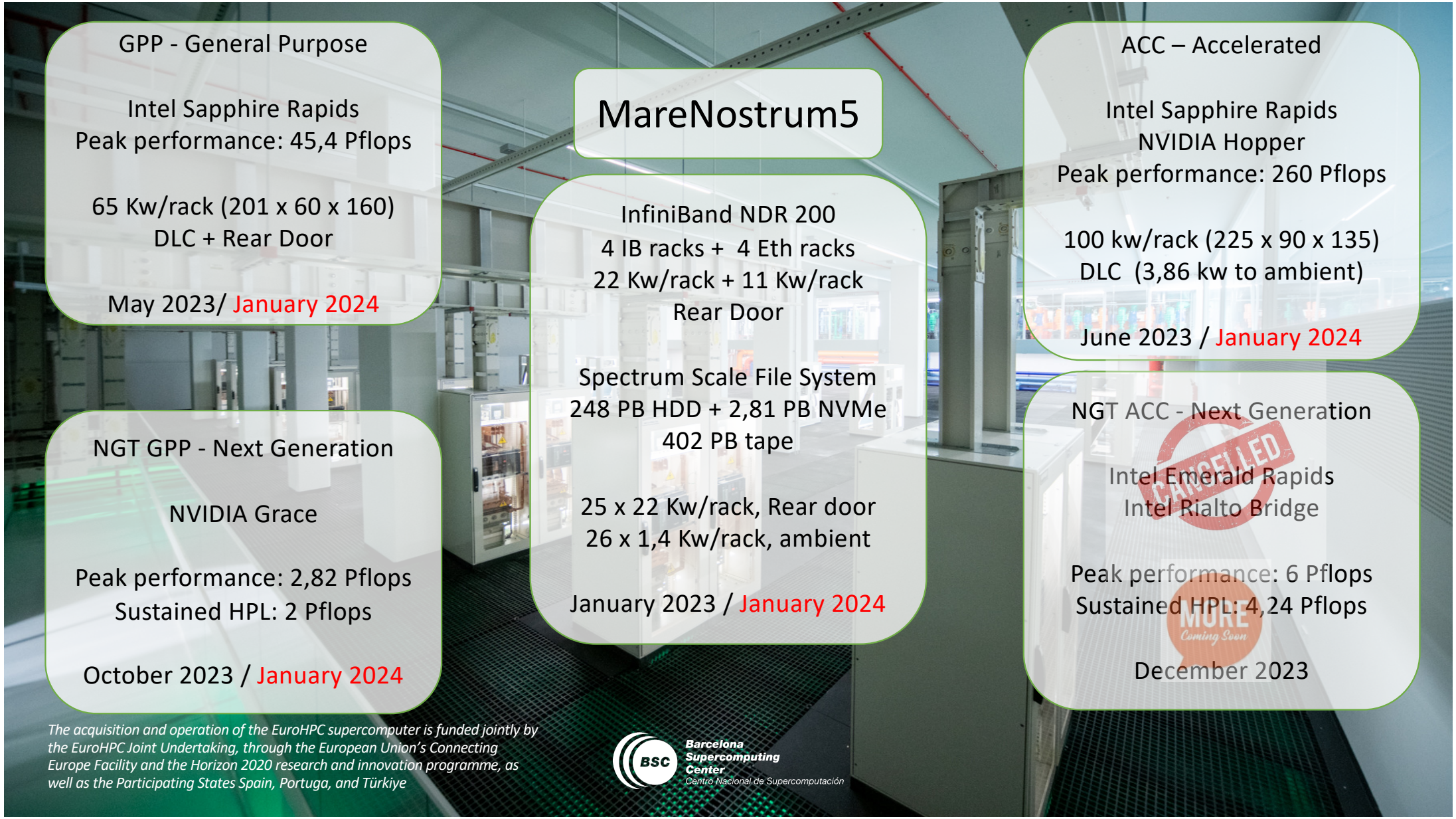
Virtualization server ThinkSystem**SR630v2**

18x

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum5 – Software stack

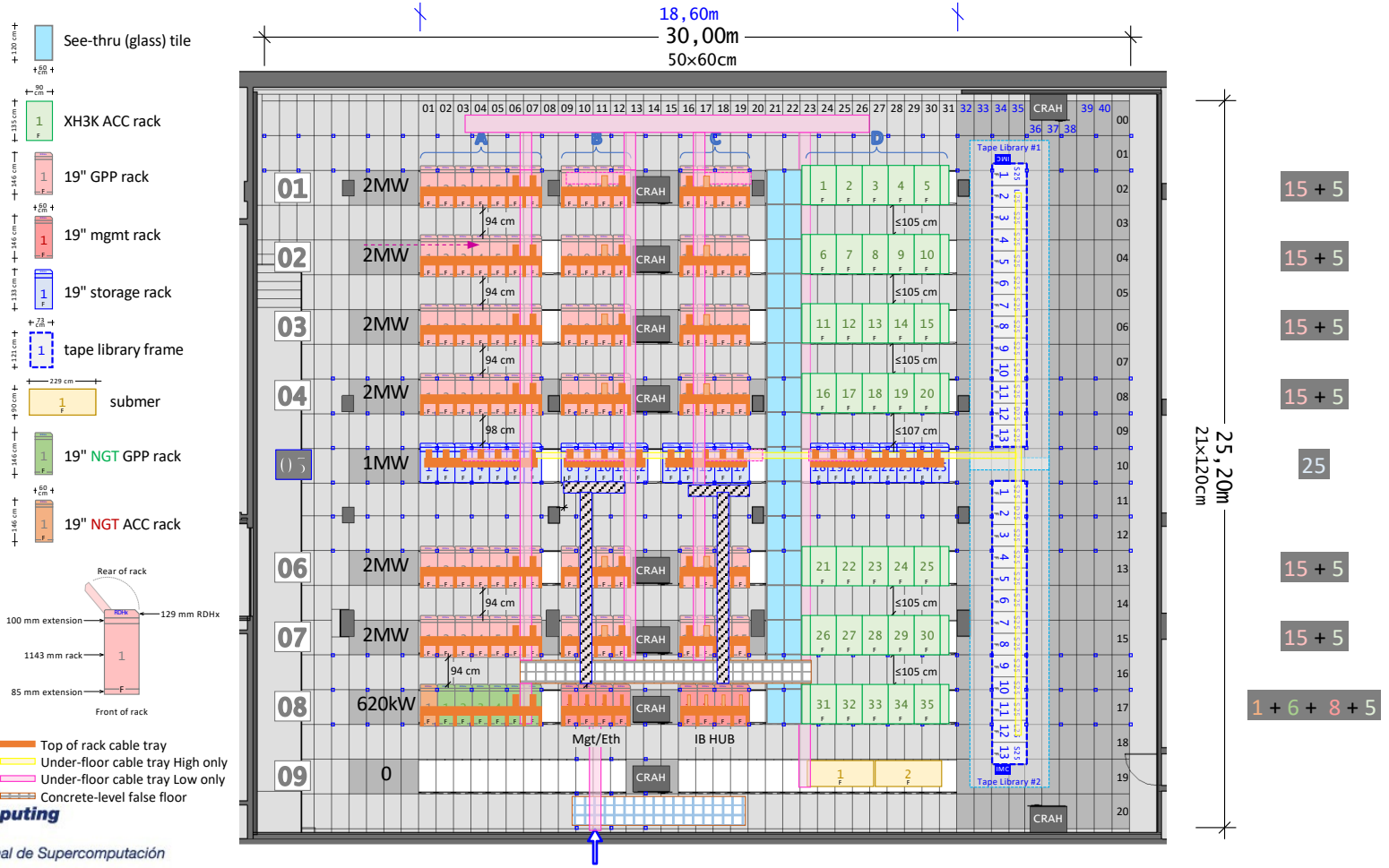
Software type	MN5
Operating system	Red Hat Enterprise Linux
Compiler Suite	Intel OneAPI HPC Toolkit Nvidia SDK (PGI)
Numerical libraries	Intel MKL Nvidia SDK
Debugging/profiler tools	BSC Performance tools ARM DDT Nvidia SDK Intel OneAPI HPC Toolkit (vtune, ...)
Resource and workload manager	SLURM Only one Slurm cluster, with different partitions
Energy Efficiency and Power Management	EAR



The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portuga, and Türkiye

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum5 DC layout



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Next Projects

- On-going
  - System installation
  - System and facility validation
  - Access to HPC systems
  - Osmosis Facility
- On construction or procurement
  - Installation of quantum systems
  - Utilization of phreatic water
- At legal/economical validation
  - Power station
- On background preparation
  - MareNostrum 6

Barcelona Supercomputing Center
Centro Nacional de Supercomputación


Thank you
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación
sergi.girona@bsc.es