Guia docent CAP-GIA

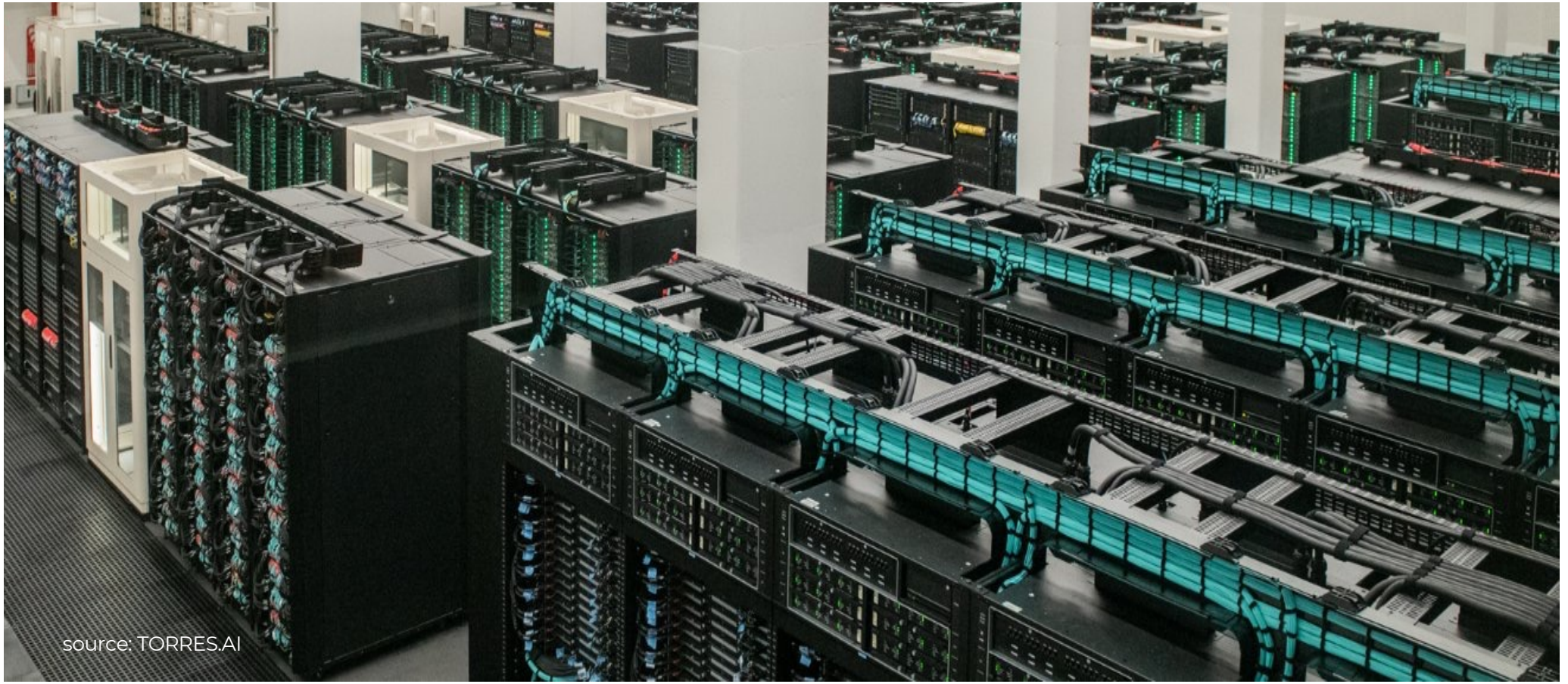
- **Calendari de sessions – tardor 2024 (Jordi Torres)**

Setmana 1: 09/09 - 13/09	Cloud Computing & Virtual Machines	Lab1 - Màquines Virtuals
Setmana 2: 16/09 - 20/09	Containers	Lab2 - Contenidors
Festiu: 23/09 - 27/09		
Setmana 3: 30/09 - 04/10	Arquitectura de Serveis	Lab3 - Serveis
Setmana 4: 07/10 - 11/10	Altes Prestacions & Supercomputació	Lab4 - MareNostrum 5 (Visita)
Setmana 5: 14/10 - 18/10	Altes Prestacions & AI - [PR1]	Lab5 - GPUs i CUDA
Setmana 6: 21/10 - 25/10	Computació pre-Exascale - [PR2]	Lab6 - Programació pre-Exascale
Setmana 7: 28/10 - 01/10	Arquitectures Big Data	Lab7 - Contenidors + HPC
Parcials: 04/11 - 08/11		
Setmana 8: 11/11 - 15/11	Arquitectures per Streaming	Lab8 - Middleware i Streams (Spark)
Setmana 9: 18/11 - 22/11	Seminaris	Presentacions Laboratori - [PK-Lab]
Setmana 10: 25/11 - 29/11	Sistemes de Fitxers	Lab9 - Sistemes de Fitxers Distribuits
Setmana 11: 02/12 - 06/12	Altes Prestacions & Deep Learning	Lab10 - Supercomputadors i AI
Setmana 12: 09/12 - 13/12	Paral·lelisme - [PR3]	Lab11 - Entrenament en Supercomputadors
Setmana 13: 16/12 - 20/12	Seminaris	Lab12 - Entrenament en Paral·lel

Teoria 4
SUPERCOMPUTING

Computació d'Altes Prestacions

Josep Ll. Berral – Jordi Torres · Grau IA – FIB



Introduction

Basic terms

- **High Performance Computing (HPC)** is really a collection of **multiple interrelated disciplines**

- HPC is a field of work that relates to all facets of **technology, methodology, and application** associated with achieving the greatest computing capability possible at any point **in time and technology**.

Basic terms

- It engages a class of electronic digital machines referred to as "**supercomputers**" to perform a wide array of computational problems or "applications" (alternatively "workloads") **as fast as is possible**.

- The action of performing an application on a supercomputer is widely termed "supercomputing" and is synonymous with HPC.

Supercomputing = HPC

What is Supercomputing?

- **Marenostrum5 Supercomputer (HPC cluster)**
  - thousands of nodes/servers (MN5 7600 servers)
  - 1 node/server of the MN5 cluster:
    - 2 sockets (each socket with 56 cores and AVX512)
    - Main memory with up to 256GB per node
    - Local Hard drive SSD of 400GB
    - High-speed access network of 100Gb/s per node

      →A Marenostrum5 node is approx. 20 times faster than a laptop

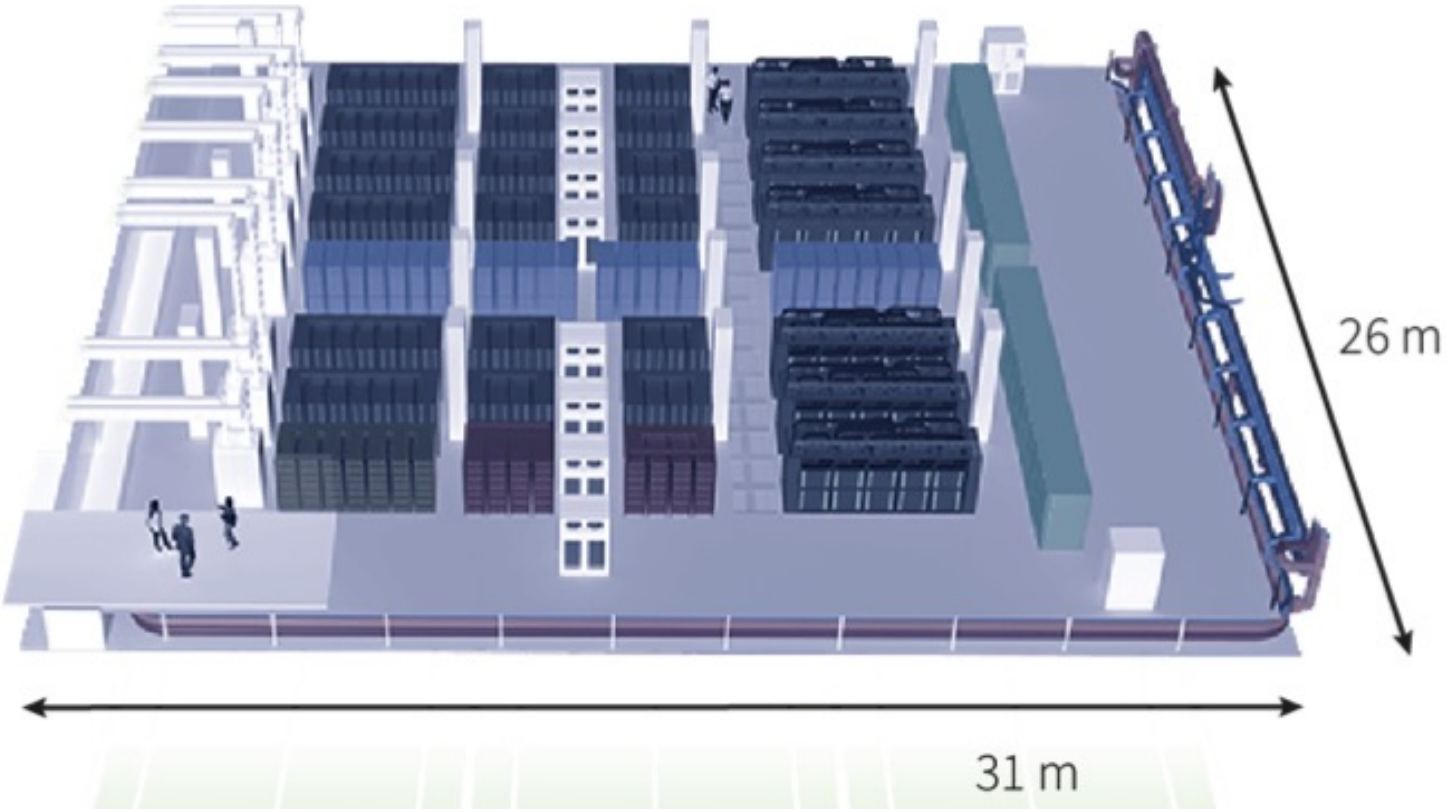
    - **1240 nodes additionally have 4 H100 GPUs**

- **Marenostrum5 storage**
  - GPFS filesystem with 200 PB
  - Tapes HSM storage with 400 PB

Computer hardware components in Marenostrum 5

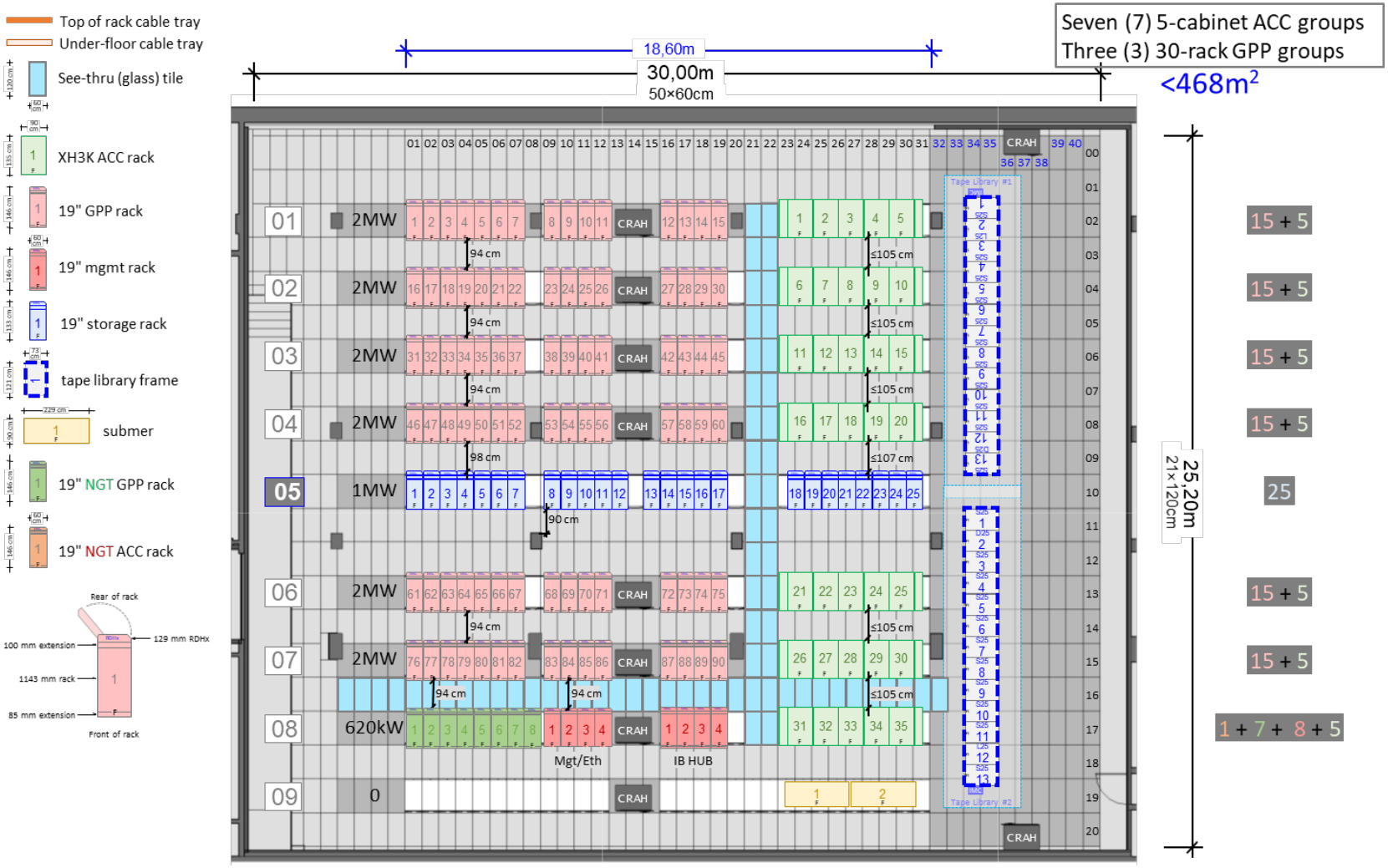
Marenostrum 5 general view

The supercomputer occupies a room with an area of 800 m², equivalent to about 3 tennis courts.



Services (e.g. refrigeration and electrical transformers) occupy almost three times as much: 2,000 m²

MN5 DC Layout



Simplified view of hardware components in a DC components

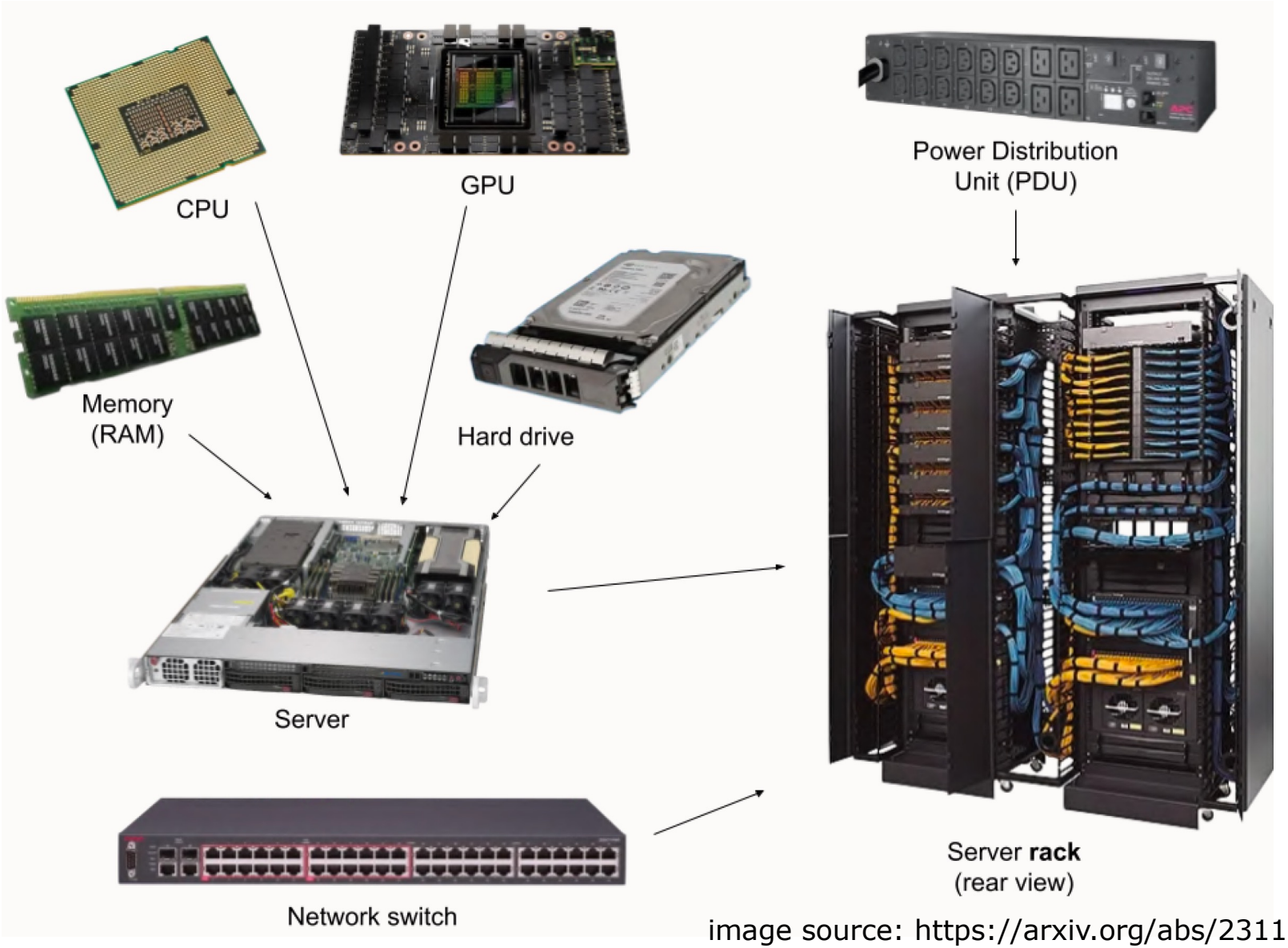


image source: https://arxiv.org/abs/2311.02651

MN5: 2 clusters

MareNostrum 5 is the only European supercomputer with two entries in the list of the 20 most powerful supercomputers in the world

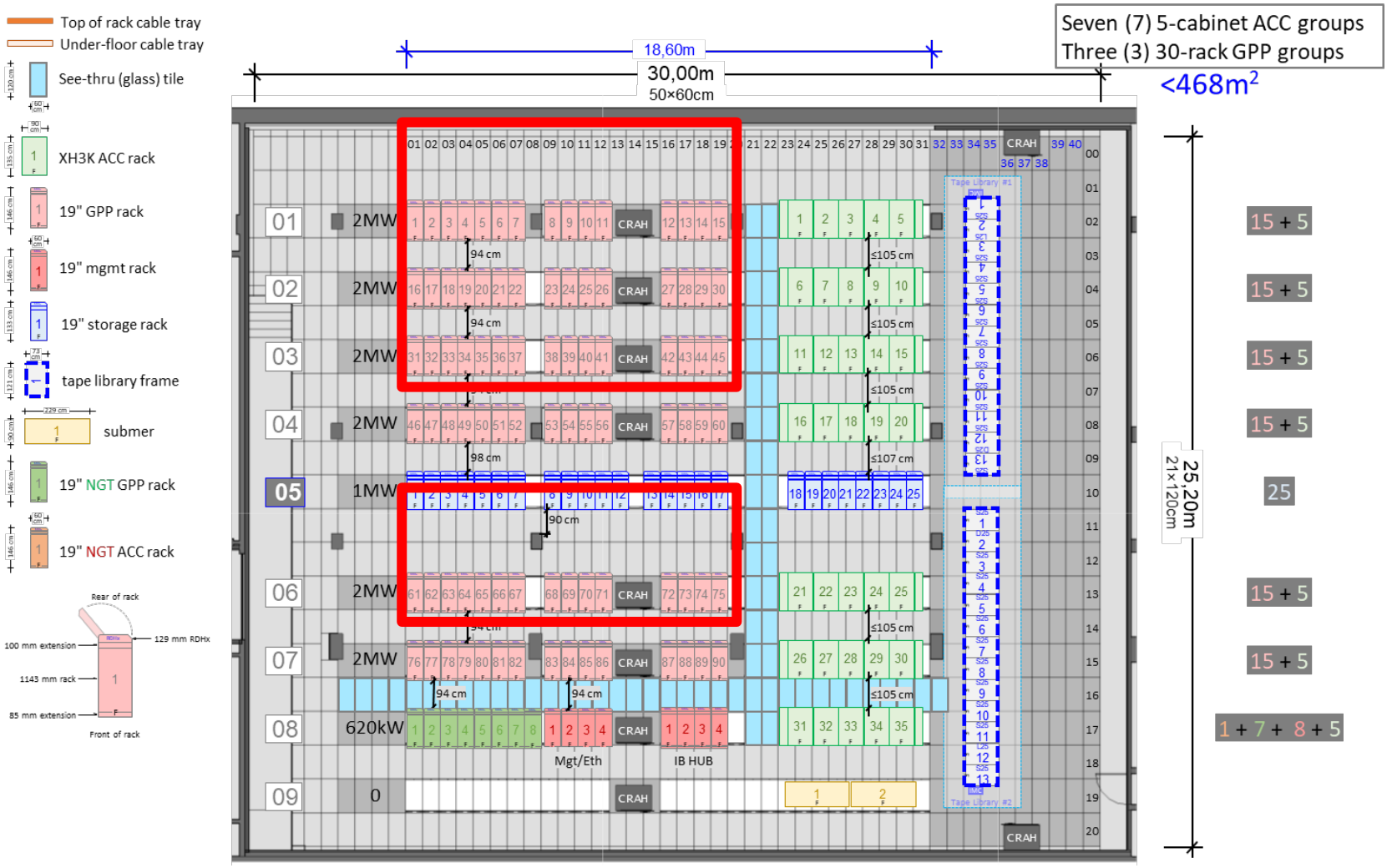
14 November 2023

The new supercomputer has a computer architecture that combines two different systems including the world's most powerful general purpose partition based on the popular x86 architecture.

The BSC is the only supercomputing center in Europe to have two entries in the top 20 of the Top500, including both the general purpose partition, the world's largest based on the well-known x86 computing architecture, and the accelerated partition, which is the third most powerful in Europe and the eighth most powerful in the world, enabling research to advance in areas as important as artificial intelligence and numerical simulation.

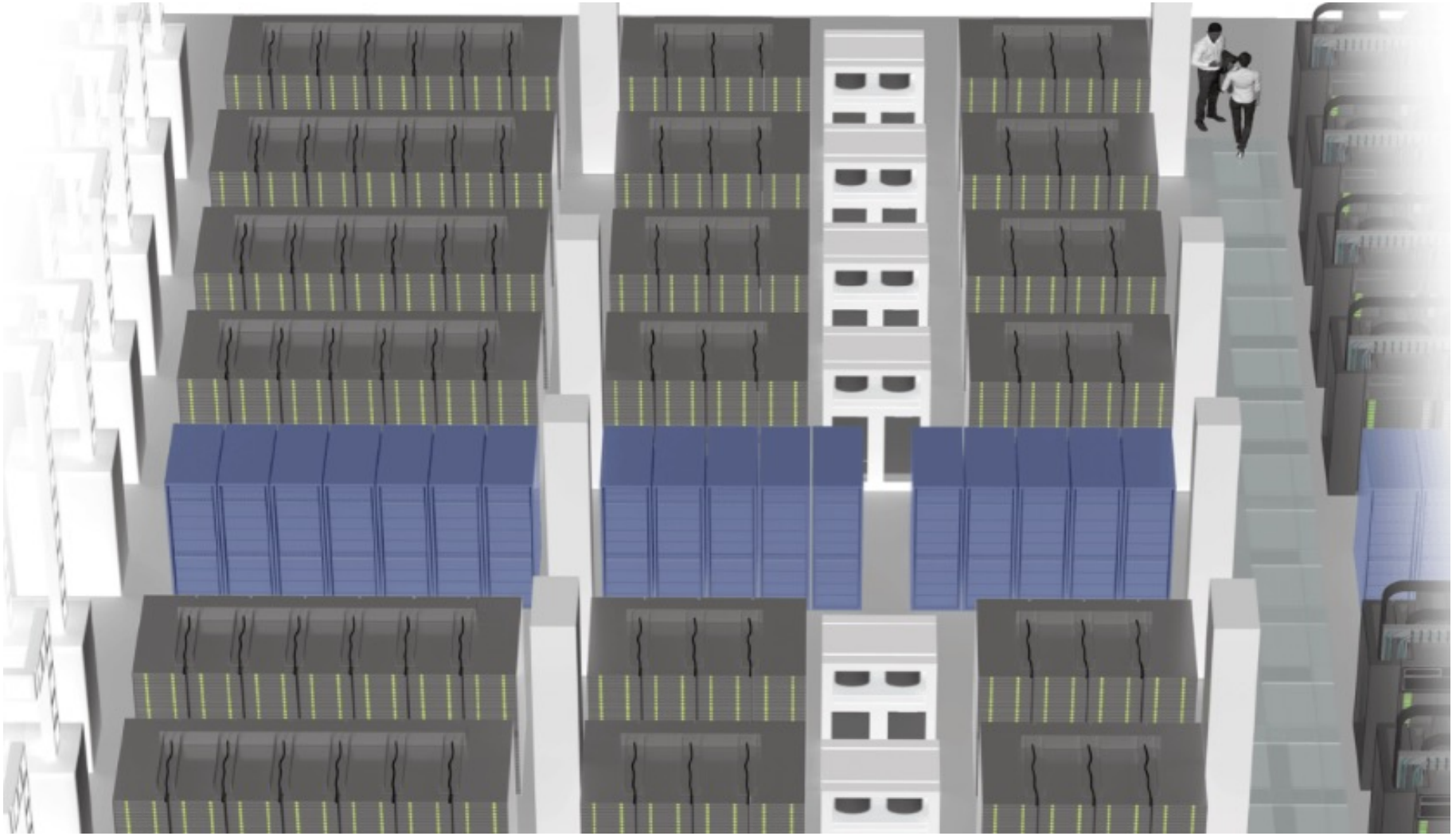
MN5 General Purpose Partition (GPP)

- General Purpose Partition



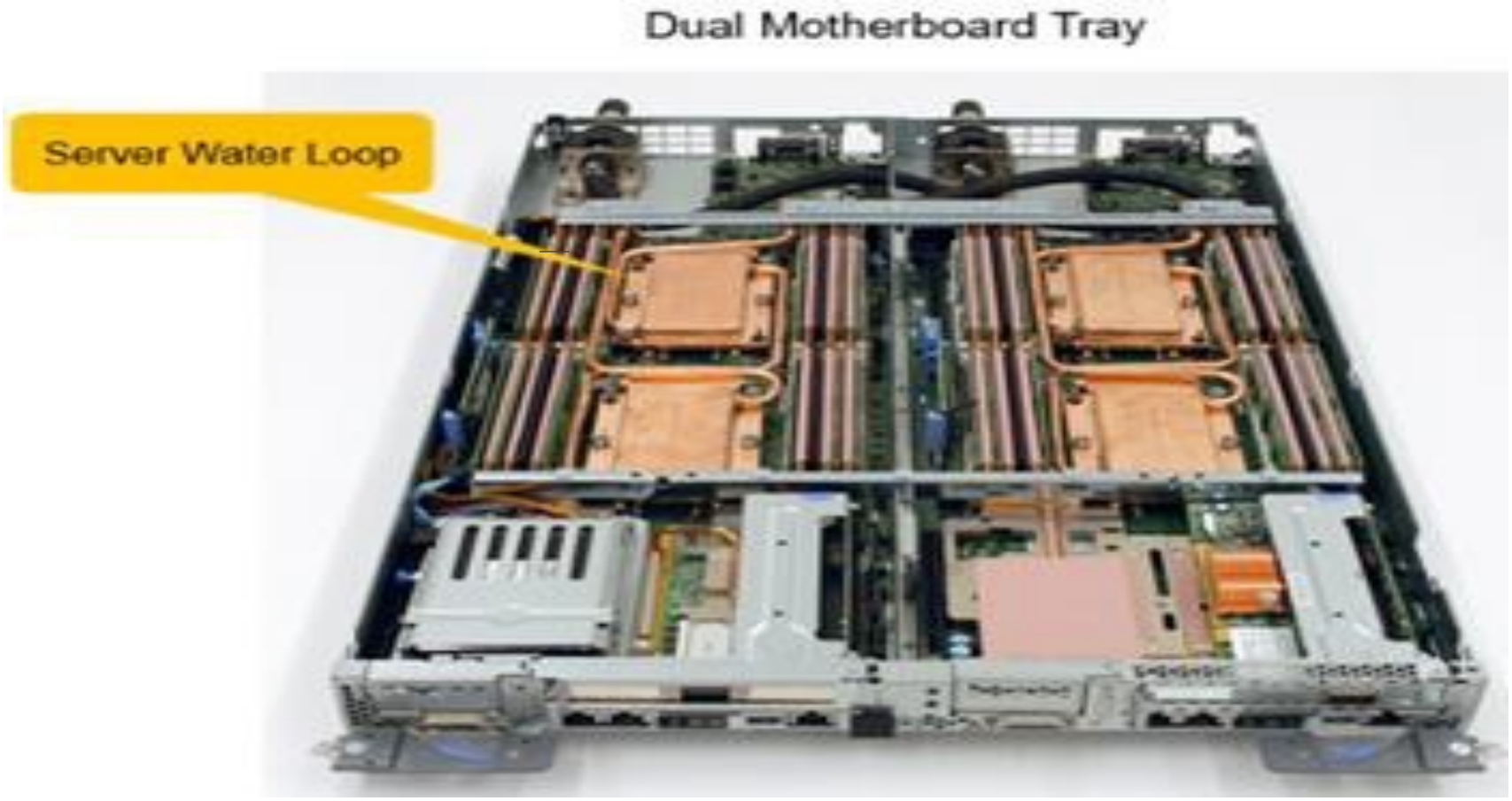
MN5-GPP

- **90 racks, 6,480 nodes and 12,960 chips (**Intel Sapphire Rapids **).**



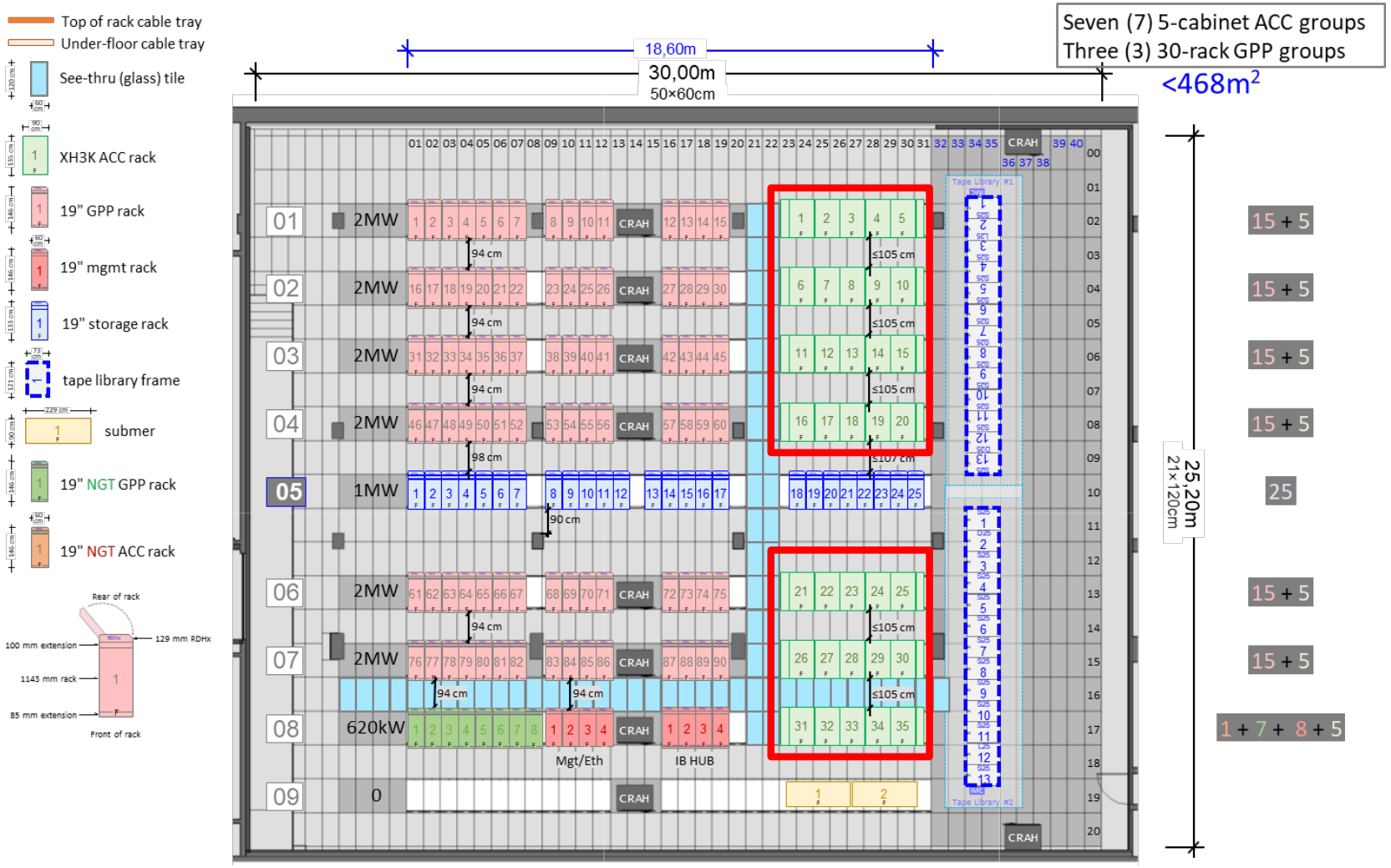
GPP Compute Node

Compute Node: 2x Intel 2x Intel Sapphire R.



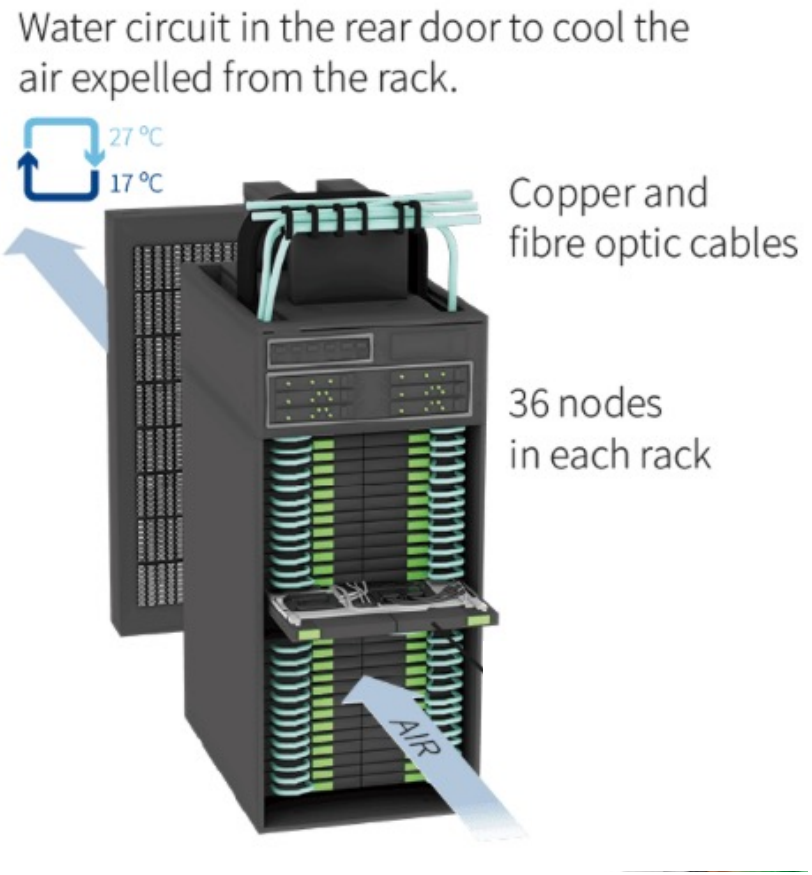
MN5 DC Layout: ACC

- **MN5 Accellerated partition**

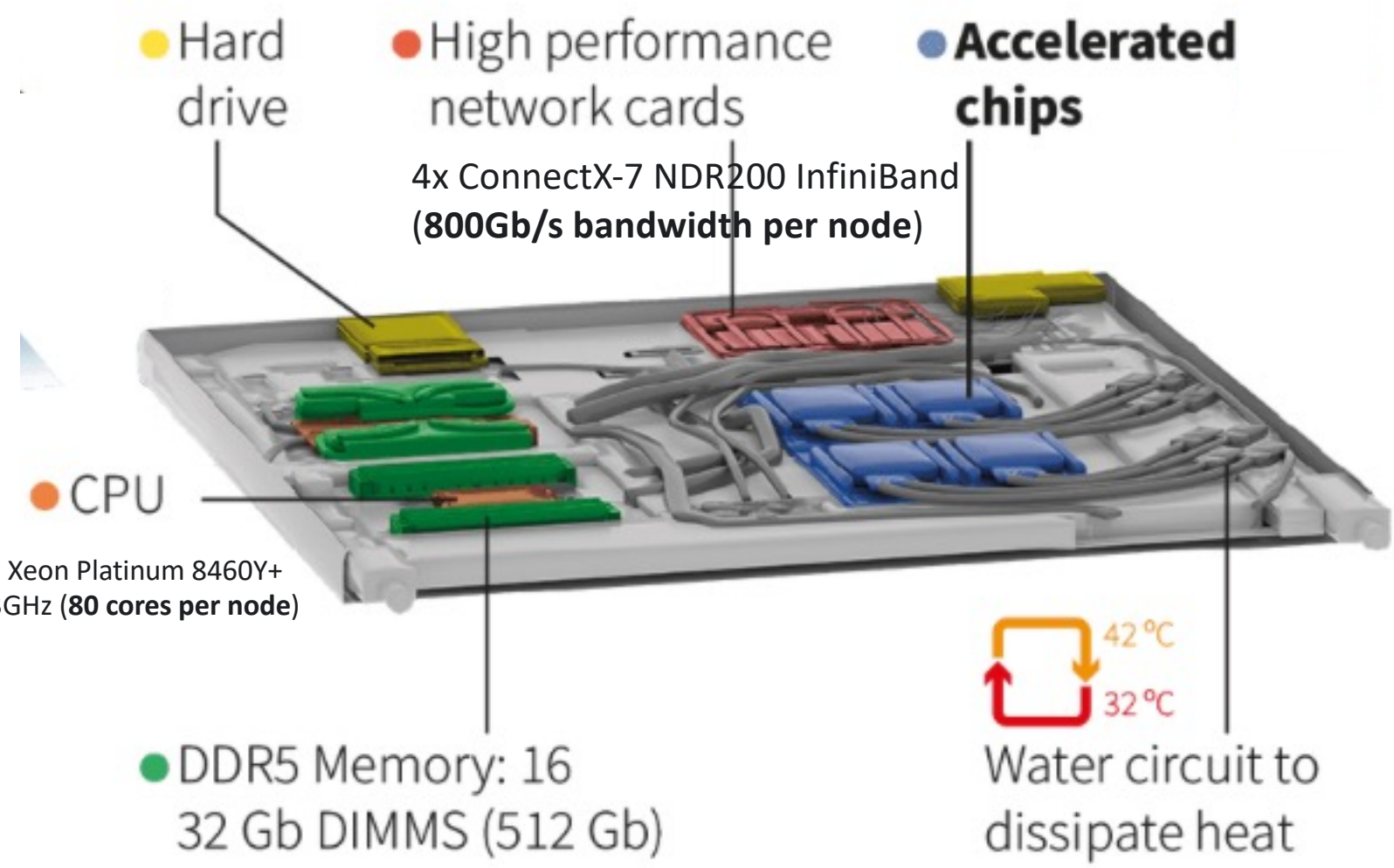


Accelerated partition (MN5-ACC)

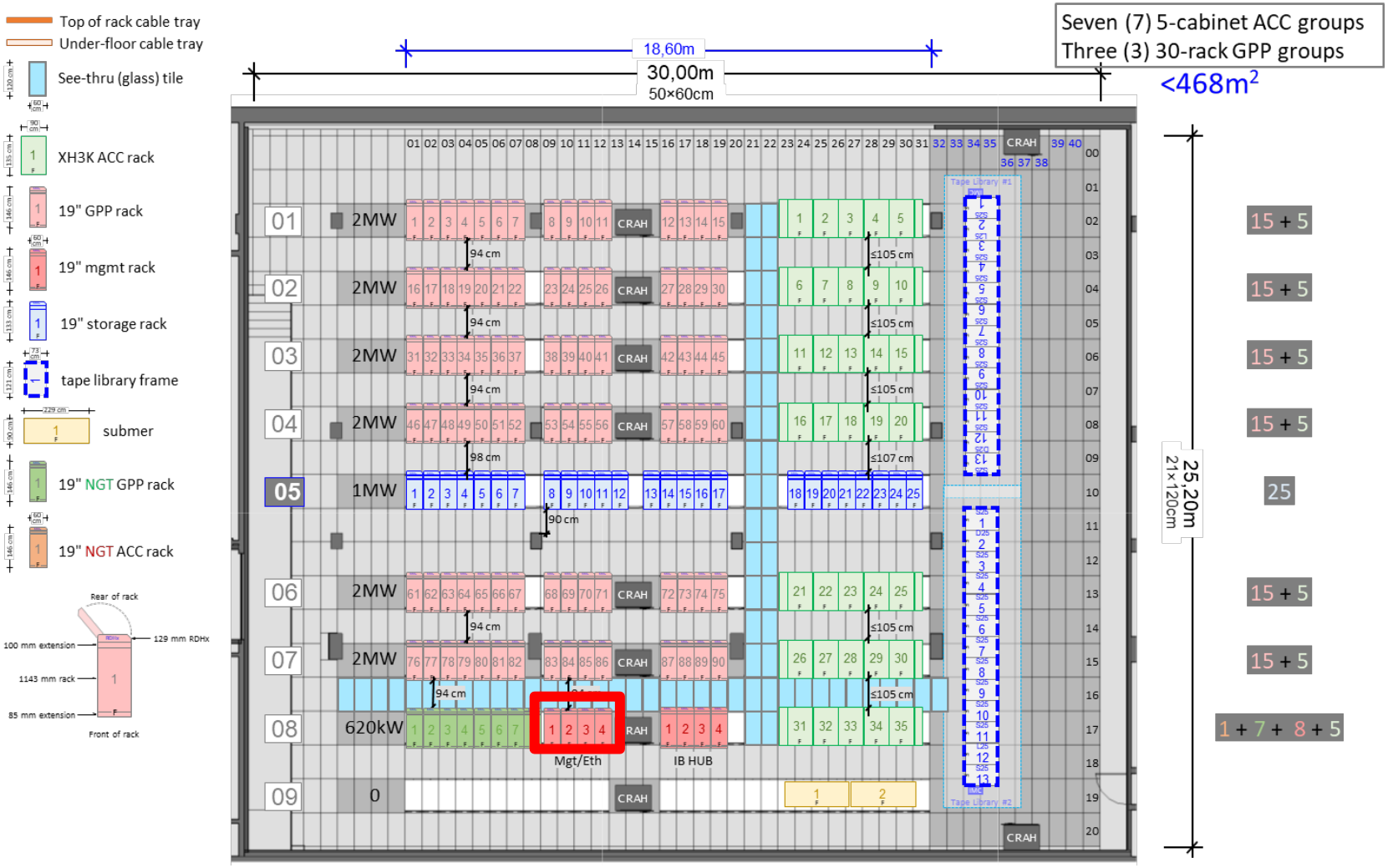
- **35 Racks, 1.120 nodes and 6720 chips (4,480 accelerated)** Intel Xeon Sapphire Rapids and NVIDIA Hopper



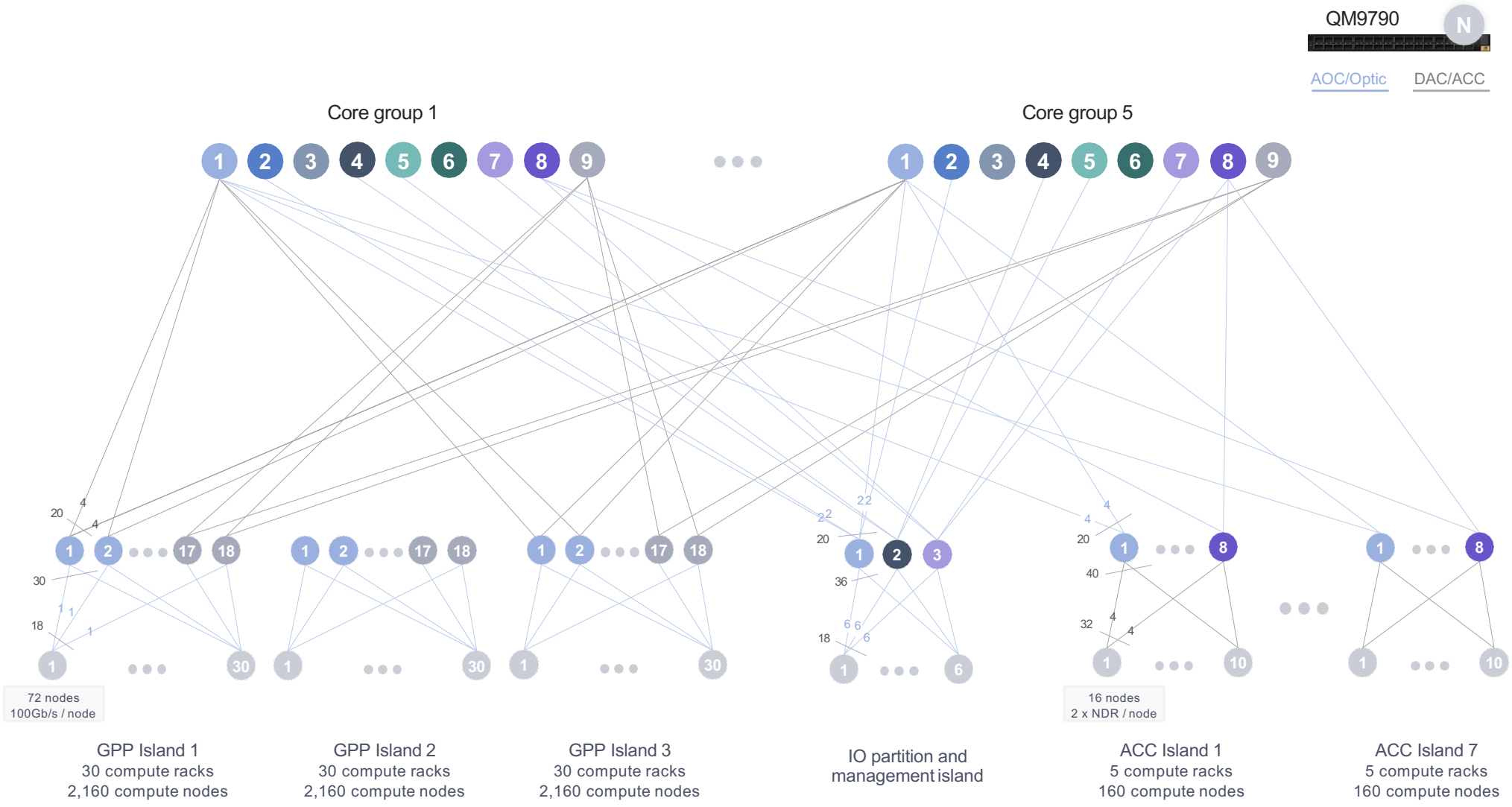
Accelerated node



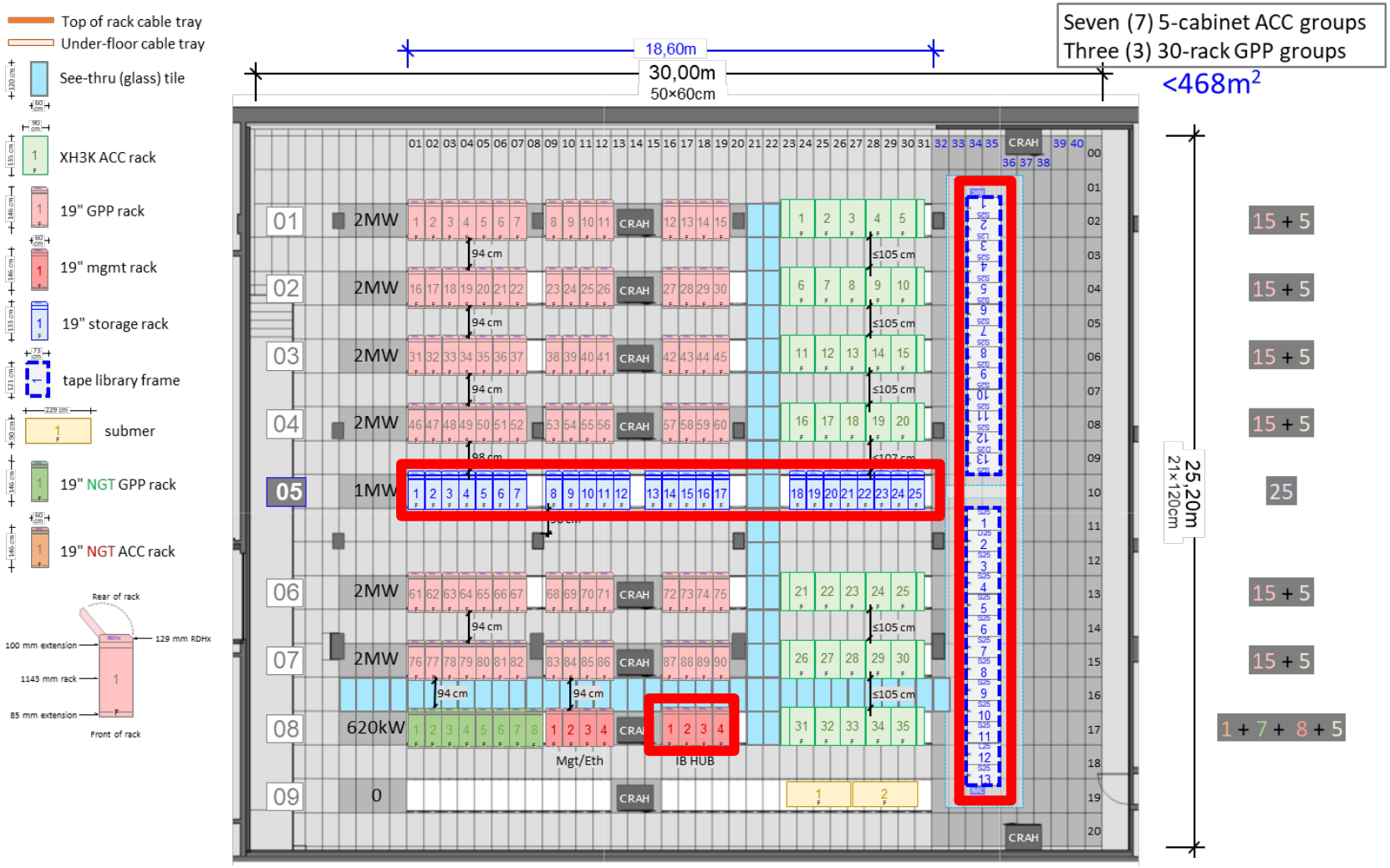
MN5 DC Layout: Infiniband



Infiniband network Overview

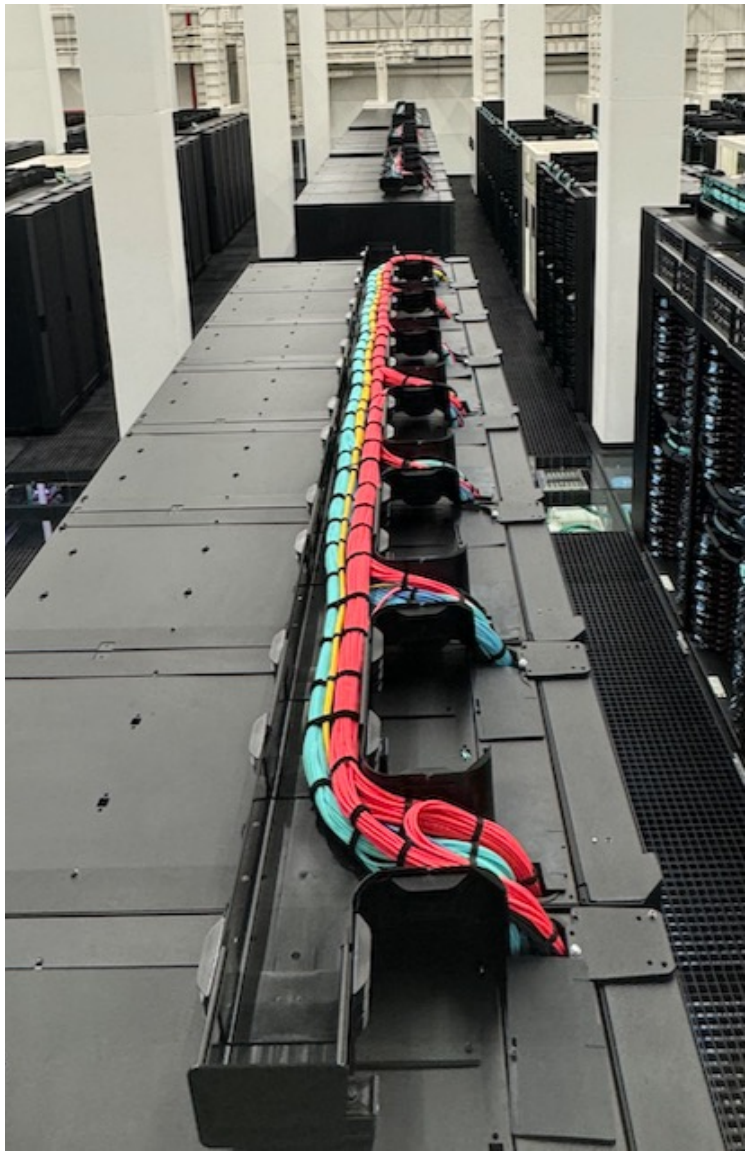


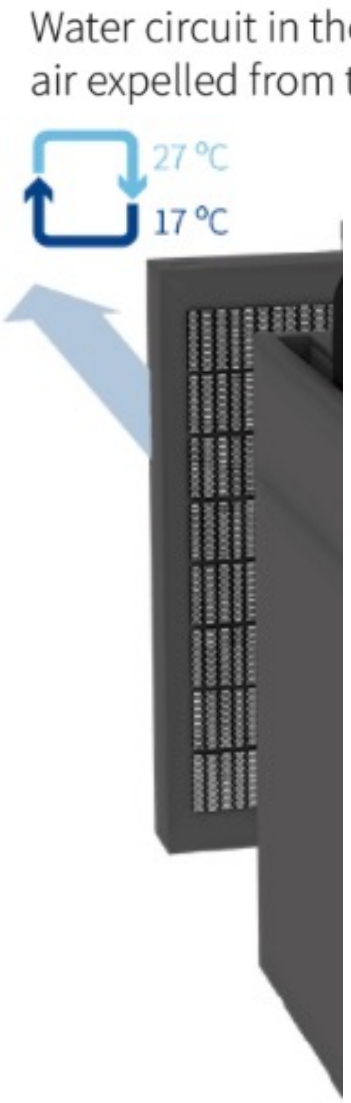
MN5 Storage


Top of rack cable tray
Under-floor cable tray
See-thru (glass) tile
XH3K ACC rack
19" GPP rack
19" mgmt rack
19" storage rack
tape library frame
submer
19" NGT GPP rack
19" NGT ACC rack
Rear of rack
100 mm extension
129 mm RDHx
1143 mm rack
85 mm extension
Front of rack
18,60m
30,00m
50×60cm
Seven (7) 5-cabinet ACC groups
Three (3) 30-rack GPP groups
<468m²
01 2MW
02 2MW
03 2MW
04 2MW
05 1MW
06 2MW
07 2MW
08 620kW
09 0
CRAH
94 cm
≤105 cm
Mgt/Eth
IB HUB
25,20m
21×120cm
15 + 5
15 + 5
15 + 5
15 + 5
25
15 + 5
15 + 5
1 + 7 + 8 + 5

Hard Drives

Top class storage system of 248PB net capacity based on SSD/Flash and hard disks





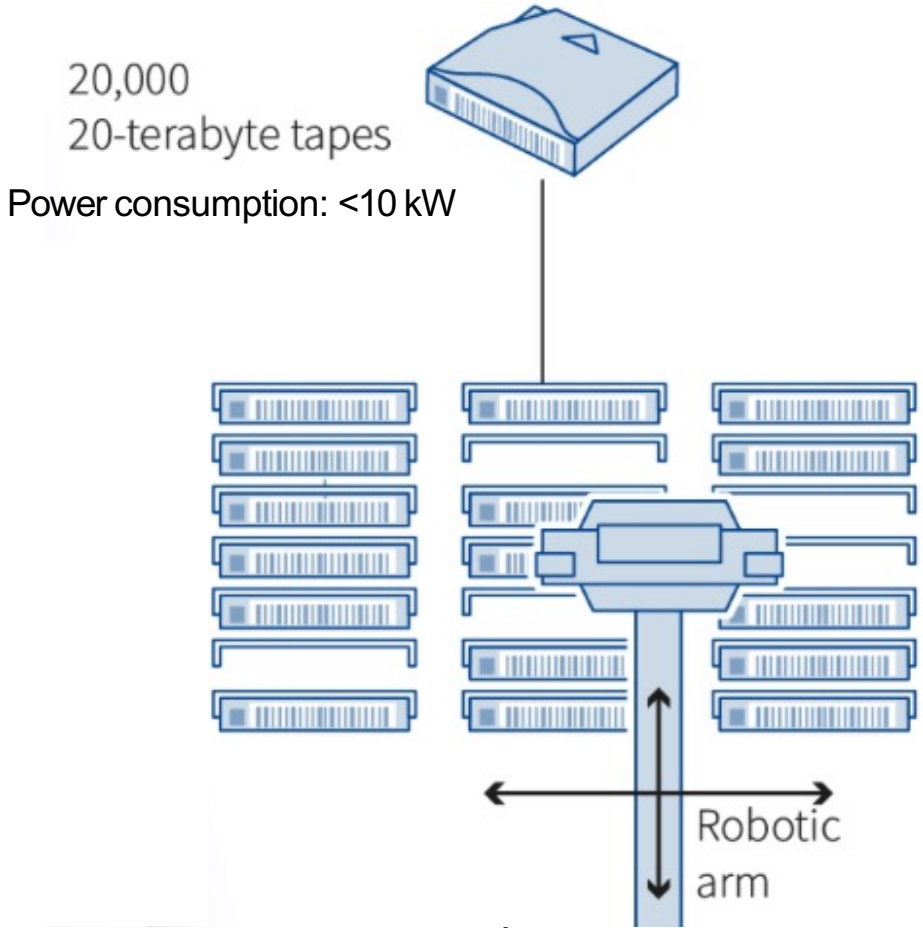
HDD: 20400 x 18 TB NL-SAS 3,5”
248 / 376 PB Net/Brut Capacity
1.6 TB/s read and 1.2 TB/s write

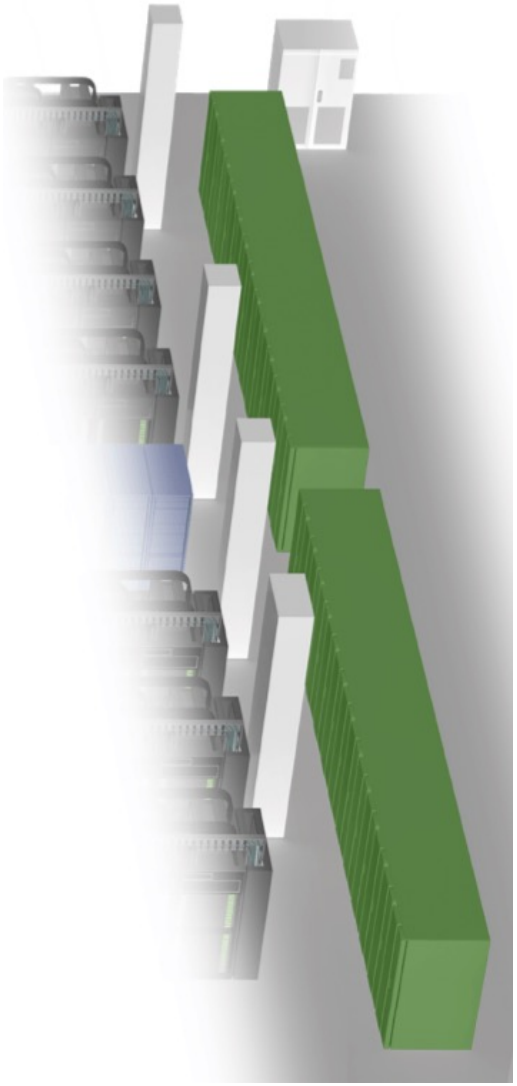
Flash: 312x 15.36 TB
2.8 / 4.8 PB Net/Brut Capacity
600 GB/s read or write

Power consumption: 400-550 kW

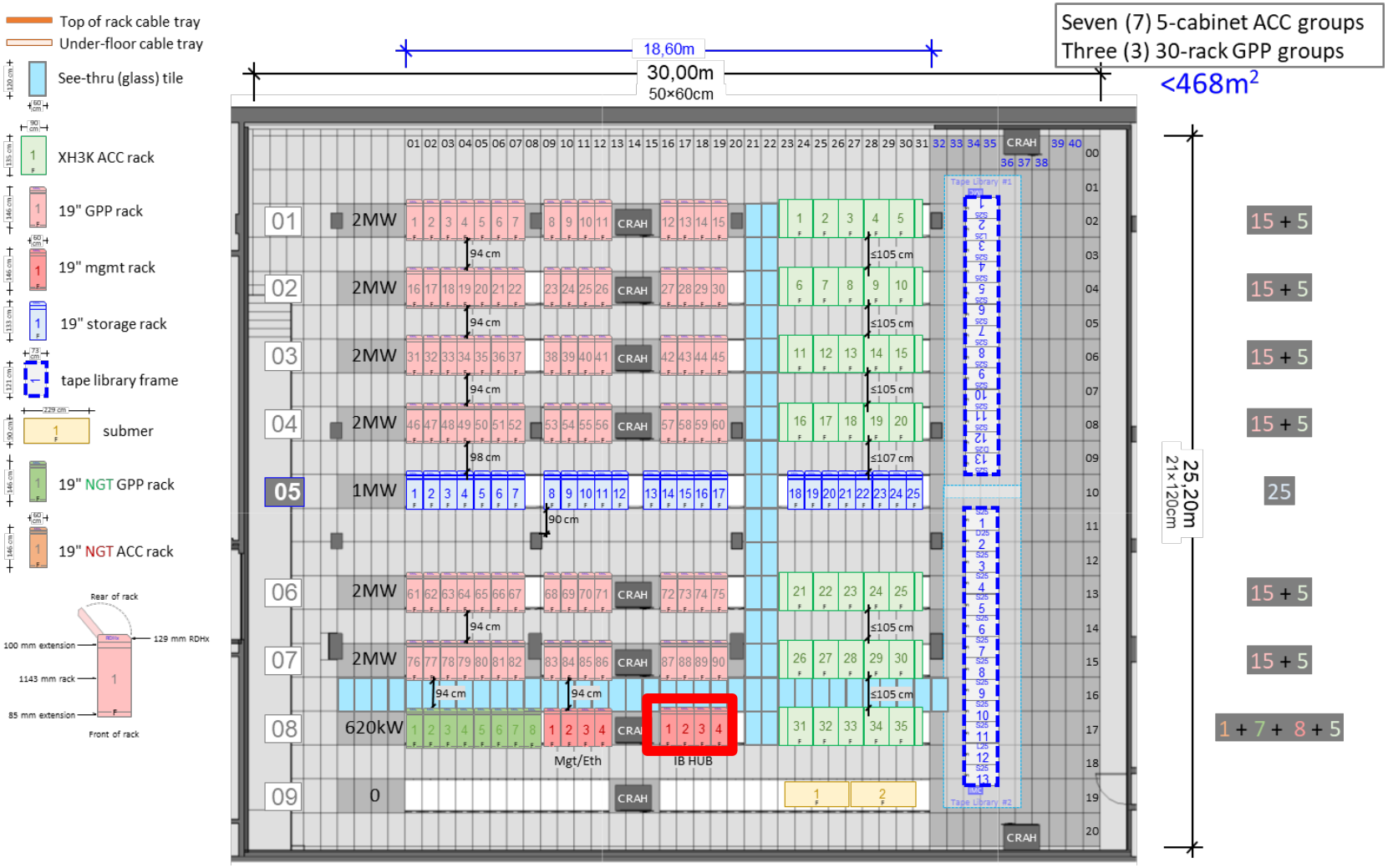
Magnetic tapes

- Slow to access, they are used because they consume less electricity. They store long-term data that is consulted less frequently.





MN5 Management



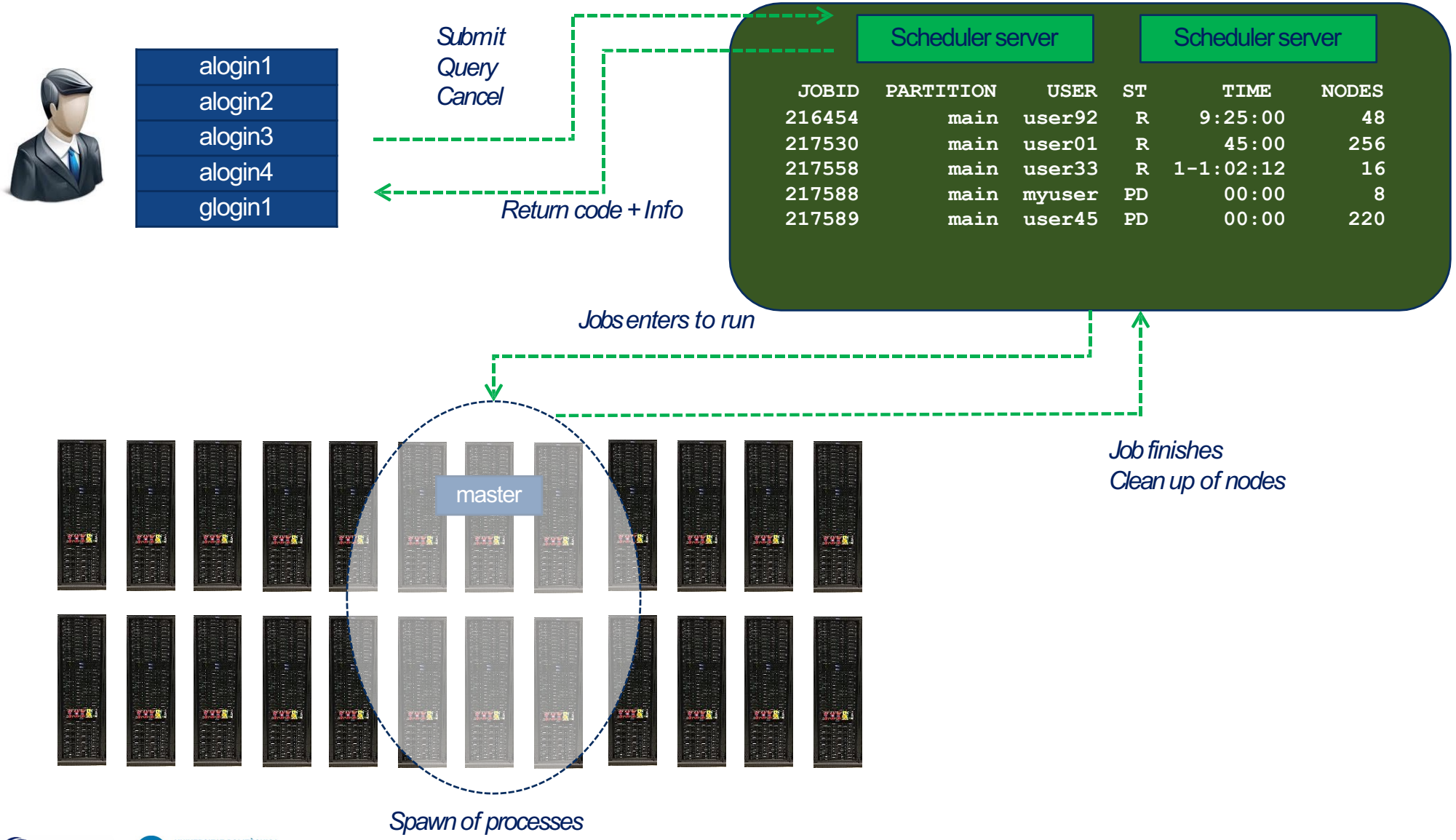
Batch system

Batch system

- **Marenostrum 5 uses SLURM as a batch system**
  - Simple Linux Utility for Resource Management

    Now: Slurm Workload Manager
- **Development started in LLNL in 2002 as simple resource manager for Linux clusters**
  - Open-source
  - Modular design
  - Around 100 plugins (extending core functionalities)

Batch Scheduler overview



Basic slurm job example

- **Sequential example**

```
#!/bin/bash
#SBATCH --job-name=seq_job
#SBATCH --chdir=.
#SBATCH --output=serial_%j.out
#SBATCH --error=serial_%j.err
#SBATCH --ntasks=1
#SBATCH --time=00:02:00

./serial_binary
```

Slurm Commands

- Jobs
  - sbatch

    Submit script for later execution (batch)

```
$ sbatch mpi-slurm-test.cmd
Submitted batch job 3159800
$
```

  - scancel

    Cancels a running or pending job

```
$scancel 2297588
```

Slurm Commands

- System information
  - squeue

    Report job and job step status

```
[root@phead1 ~]# squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
             30101      main measure_ bscXXXXX CG      11:08      2 p9r3n[15-16]
             30153      main fixAmpl_ bscXXXXX PD       0:00     32 (Resources)
             30204      main fR5_016- bscXXXXX PD       0:00     16 (Priority)
                                  .....

             30274      main ilsvrc20 bscXXXXX  R   18:49:28      1 p9r1n14
             30344      main cadSOpt5 bscXXXXX  R   11:22:47      1 p9r2n06
             30342      main cadSOpt5 bscXXXXX  R   11:22:51      1 p9r2n06
             30343      main cadSOpt5 bscXXXXX  R   11:22:51      1 p9r2n06
             30338      main cadSOpt4 bscXXXXX  R   11:24:34      1 p9r1n16
                                     .....
```

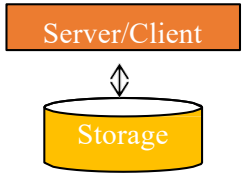
SA-MIRI Unix Group & Queues

- **Unix Cgroup: Linux Kernel mechanism to limit, isolate and monitor resource usage (CPU, memory, disk I/O, etc) of groups os processes.**
- **→ nct_308**
- **Queues (QoS):**
  - acc_debug,
  - acc_interactive,
  - acc_training,
  - gp_debug,
  - gp_interactive,
  - gp_training

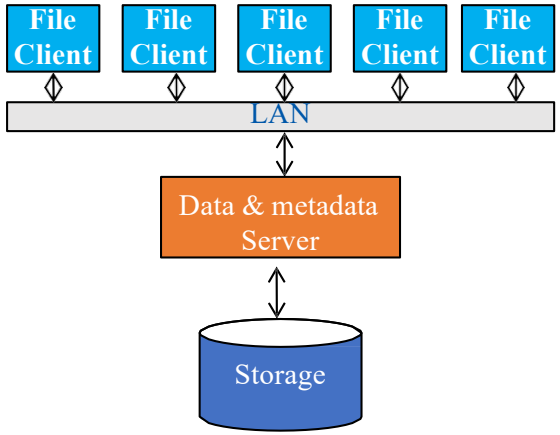
Storage in Marenostrum 5

File System architectures?

Local File system

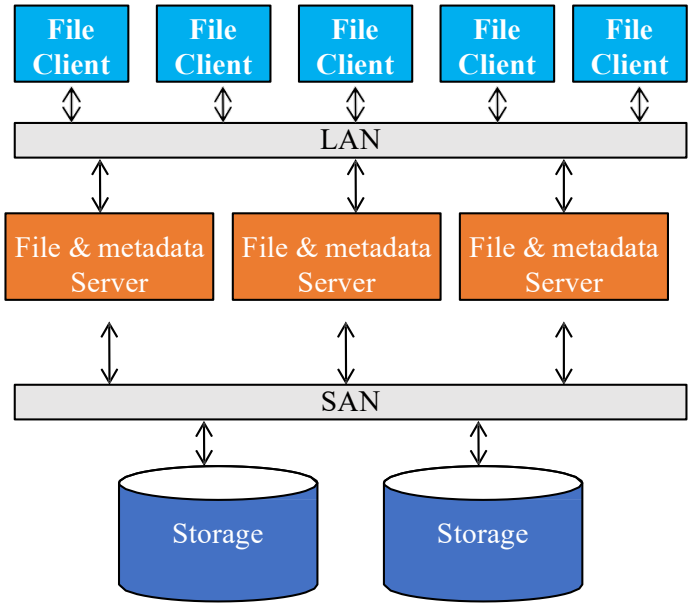


Distributed File system



Parallel File system

we require a single I/O to access all storage, so disks are distributed, allowing multiple clients to operate simultaneously with a single name space



Parallel File system description

- **Parallel File system description**
  - Multiple servers and storage devices
  - I/O Operations distributed and run in parallel
  - Files spread in multiple devices
  - Concurrent access to files from multiple clients
  - Single namespace
  - Examples: Lustre, BeeGFS IBM Storage Scale (GPFS).

- **General Parallel File System (GPFS)**
  - IBM high performance parallel file system.
  - Not GPFS anymore~~. Now IBM Spectrum Scale.~~
  - ~~Not IBM Spectrum Scale anymore~~. Now IBM Storage Scale.

General Parallel Filesystem (GPFS)

Relevant filesystems to your everyday use:

- **/apps/GPP/ and /apps/ACC/** (Applications and libraries)
  - Applications installed on MN.
- **/gpfs/home** (User's home)
  - Scripts, codes and documents.
- **/gpfs/projects** (Data, custom installations)
  - Execution folder, init data, shared data in your group.
- **/gpfs/scratch** (Temporary files, doesn't have backups)
  - Execution data, huge log/output files, temporary files.
- **/gpfs/tapes/hpc** (Archive): HSM (Hierarchical Storage Management)

MN5 HPC Storage

#ESS	Drive Capacity	Total # drives	Raw capacity	Net capacity	Read perf	Write perf
50	NL-SAS 18TB	20400	367PB	248 PB (8+3P)	1.6TB/s (IOR 100%read)	1.2TB/s (IOR 100%read)
13	NVMe 15.36TB	312	4.79PB	2.81PB (8+2P)	600GB/s 1Mio iops 4KB	600GB/s 500K iops 4KB

We place metadata operations in NVM (Non-Volatile-Memory, SDD, ...)

We divide all this storage into 5 file systems.

File System	Size	Data	Metadata	Backup
Projects	22PB	NL-SAS	NVMe	Yes
Scratch	176PB	NL-SAS	NVMe	No
home	306TB	NVMe+NL-SAS	NVMe	Yes
apps	535TB	NVMe+NL-SAS	NVMe	Yes
archive disk cache	44PB	NL-SAS	NVMe	No

Software environment: Modules, Compilers, Containers

source of these slides:
PATC courses, 2024
Operation Department
BSC-CNS

Module Environment (I)

- **System used by Marenostrum to manage all installed software**
- **Environment variables and software dependencies management**
- **Several versions of the same program coexisting at /apps/GPP and /apps/ACC/**
- **Default modules loaded will depend on the partition:**
  - Intel compiler, libraries and tools (intel/2023.2.0) - GPP
  - Intel MKL (mkl/2023.2.0) and Intel MPI (impi/2021.10.0) - GPP
  - BSC custom commands (bsc/1.0) - GPP and ACC

Module Environment (II)

- **Module commands:**

Command	Option	Example	Info
avail	[program]	module avail	List available modules
list	[program]	module list	List loaded modules
purge		module purge	Unload all modules
load/unload	<program[/version]>	module load/unload gcc/5.1.0	Load or unload a module
switch	<old> <new>	module switch intel gcc	Change a module by another

Compilers

- **Intel, GNU and NVIDIA compiler suites available via modules**

- **Several versions, managed by the module system**
  - Intel (licensed)
    2023.0, 2023.1, 2023.2.0, 2024.0
  - GCC (Free Software)
    11.4.0, 13.2.0
  - NVIDIA (licensed) - Only for ACC
    23.9, 23.11, 24.3

- **MPI compilation also managed by modules through wrappers**
  - Intel: mpicc (C), mpiicpc (C++), mpifort (FORTRAN), …
  - GCC: mpicc (C), mpicxx (C++), mpifort (FORTRAN), …
  - NVHPCX: mpicc (C), mpicxx (C++), mpifort (FORTRAN), …

- **Load optimization flags for compiling:**

module load opt

Compiler optimization drawbacks

- **Each software is different, but:**
  - Intel can get up to 20% performance increase in some applications
    - Linking mkl libraries usually boosts performance
  - Static compilation sometimes runs faster
- **Optimization drawbacks**
  - Over optimization may result in numeric error
- **Intel compilers might get a bit finicky with some compilations, you could try to use GCC instead and vice versa**

Containers?

- **What is Docker?:**
  - Docker is a platform for creating, deploying, and running applications in containers.
  - Uses a client-server architecture.
  - Simplifies application deployment across environments.
- **What is Singularity?:**
  - Singularity is designed for high-performance computing (HPC).
  - Allows containers to run without root privileges.
  - Focuses on security and scientific workflows.

- **Docker is better for general-purpose, Singularity excels in HPC.**

Supercomputer performance basics

Student notes

What do you mean by performance?

- **For HPC the most widely used metric is "flops".**

- **What is "flop/*s*"?**
  - Flop/s is a rate of execution, some number of floating-point operations per second.
  - Whenever this term is used, it will refer to 64-bit floating-point operations, and the operations will be either addition or multiplication.

Units of Measure

- Typical sizes are millions, billions, trillions…

Mega	Mflop/s = 10^6 flop/sec
Giga	Gflop/s = 10^9 flop/sec
Tera	Tflop/s = 10^{12} flop/sec
Peta	Pflop/s = 10^{15} flop/sec
Exa	Eflop/s = 10^{18} flop/sec
Zetta	Zflop/s = 10^{21} flop/sec
Yotta	Yflop/s = 10^{24} flop/sec

Units of Measure (cont.)

- Other HPC units are:
  - Bytes: size of data:

Mbyte = 2^{20} = 1048576 ~ 10^6 bytes
Gbyte = 2^{30} ~ 10^9 bytes
Tbyte = 2^{40} ~ 10^{12} bytes
Pbyte = 2^{50} ~ 10^{15} bytes
Ebyte = 2^{60} ~ 10^{18} bytes
Zbyte = 2^{70} ~ 10^{21} bytes
Ybyte = 2^{80} ~ 10^{24} bytes

Computer performance

- The principal defining property and value provided by HPC is delivered performance for an end-user application.

- **Expressions of "speed" or "how fast" are common, describing, perhaps vaguely, the relationships among time, work as computation actions, system size, and other factors.**

- **Performance is an intuitive notion of a machine going well: how fast it runs or the speed of an application.**

- **The Peak performance of a system is the maximum rate at which operations can be accomplished theoretically by the hardware resources of a supercomputer.**

Computer performance

- **Sustained performance is the actual or real performance achieved by a supercomputer system in running an application program. (cannot exceed peak performance)**
  - **It is considered a better indicator of the true value of a supercomputer than its specified peak performance.**

- **But because it is highly sensitive to variations in the workload, comparison of different systems only has meaning if they are measured running equivalent applications.**

- **Benchmarks are specific programs created for this purpose.**

Benchmarks

- Many different benchmarks reflect different classes of problems
- The Linpack or HPL benchmark is one such application used to compare supercomputers
- HPL is widely employed and referenced, and is the baseline for the Top 500 list that tracks the fastest computers in the world (at least those so measured) on a semiannual basis.

Scalability (PSD review)

- "Scaling" or alternatively "scalability" is a relationship of performance to some measure of the size (or "scale") of the HPC system.
  - It reflects the ability to achieve increased performance for an application by employing machines of ever-greater size.
  - Although there are many ways to quantify a system's size, a simple and widely used measure is the number of processor cores employed.

Evolution of supercomputers

Student notes

1976 - Cray-1 (250 MFLOPS peak)

- **Los Alamos National Laboratory (USA)**
- **Among the first to use Integrated Circuits**
- **Vector architecture**
  - SIMD programming model
- **Successors**
  - Cray X-MP - 800 MFLOPS
  - Cray-2 - 1.9 GFLOPS

Source: Àlex Ramirez

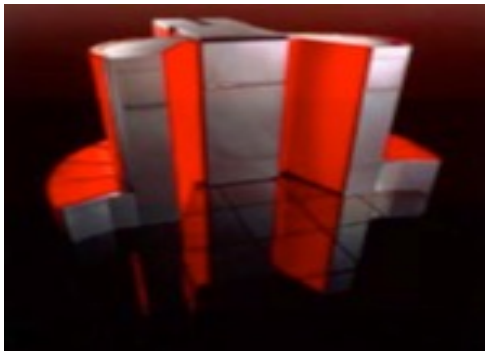
Evolution of the computing power of Supercomputers

Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
8,699,904	1,194.00	1,679.82	22,703

E P T G **FLOP/second**

1000000000000000000

June 2023
Frontier



1988
Cray Y-MP (8 processadors)

2008
Cray XT5 (15000 processadors)

1998
Cray T3E (1024 processadors)

Source: Mateo Valero

The Top500 list

- **About the TOP500 List**
  - Ranking of the world's 500 most powerful computers
  - The first version of what became today's TOP500 list started in Germany in June 1993.
  - The TOP500 list is compiled by Erich Strohmaier and Horst Simon of Lawrence Berkeley National Laboratory; Jack Dongarra of the University of Tennessee, Knoxville; and Martin Meuer of ISC Group, Germany.

- **June 2024: 63st list**

Rmax - Maximal LINPACK performance achieved

Rpeak - Theoretical peak performance

The Top500 list

- **Computer performance evaluated on a single benchmark**
  - High Performance Linpack (HPL)
- **List updated twice a year**
  - June, announced at ISC (Europe)
  - November, announced at Supercomputing (USA)
- **Warning**
  - It only ranks systems that submit their HPL score
    - **Proprietary or classified systems do not appear in the list**
- **Loads of historical data and statistics:**
  - http://www.top500.org

About the LINPACK benchmark

- **Supercomputers measure performance on a single benchmark**
  - High Performance Linpack (HPL for short)
  - Measures (useful) Floating Point Operations Per Second (FLOPS)

- **High-Performance Linpack Benchmark solves a dense NxN system of linear equations (Ax = b)**
  - Gaussian elimination method with partial pivoting

$$
\begin{aligned}
a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n &= b_1 \\
a_{12}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n &= b_2 \\
&\ldots \\
a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n &= b_n
\end{aligned}
$$

- **DAXPY routine consumes most of the CPU time**
  - Often hand-coded in assembly for higher efficiency

63st TOP 10 list

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**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
3	**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	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**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	531.51	7,107
6	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	1,305,600	270.00	353.75	5,194
7	**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	306.31	7,494
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 32C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR, EVIDEN EuroHPC/BSC Spain	663,040	175.30	249.44	4,159
9	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
10	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, Nvidia NVIDIA Corporation United States	485,888	121.40	188.65	

63st TOP 10 list

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**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
3	**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	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**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	531.51	7,107

63st TOP 10 list

6	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	1,305,600	270.00	353.75	5,194
7	**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	306.31	7,494
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 32C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR, EVIDEN EuroHPC/BSC Spain	663,040	175.30	249.44	4,159
9	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
10	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, Nvidia NVIDIA Corporation United States	485,888	121.40	188.65	

Next list: Supercomputing 2024

https://sc24.supercomputing.org

Green 500 list

- **Is it a better list than the Top500 list?**

Green 500 list

- The focus of performance-at-any-cost computer operations has led to the emergence of supercomputers that consume vast amounts of electrical power and produce so much heat that large cooling facilities must be constructed to ensure proper performance.

- To address this trend, the Green500 list puts a premium on energy-efficient performance for sustainable supercomputing.

- The Green500 list ranks the top 500 supercomputers in the world by energy efficiency.

What are data centers (DC)?

Introduction to Data Centers

- **Data Centers?**
  - Data centers are purpose-built facilities that host large-scale hardware, providing compute at scale.
  - Essential for modern internet services, scientific simulations, and AI development.

image source: https://arxiv.org/abs/2311.02651

Key features

- **Power Consumption and Cooling Systems:**
  - High power usage.
  - Advanced cooling systems required to manage extensive heat production.

- **Security and Reliability Measures:**
  - - Physical security with restricted access, surveillance, and 24/7 security staff.
  - - Redundant components like backup generators to ensure uptime

Simplified view of infrastructure components in a DC

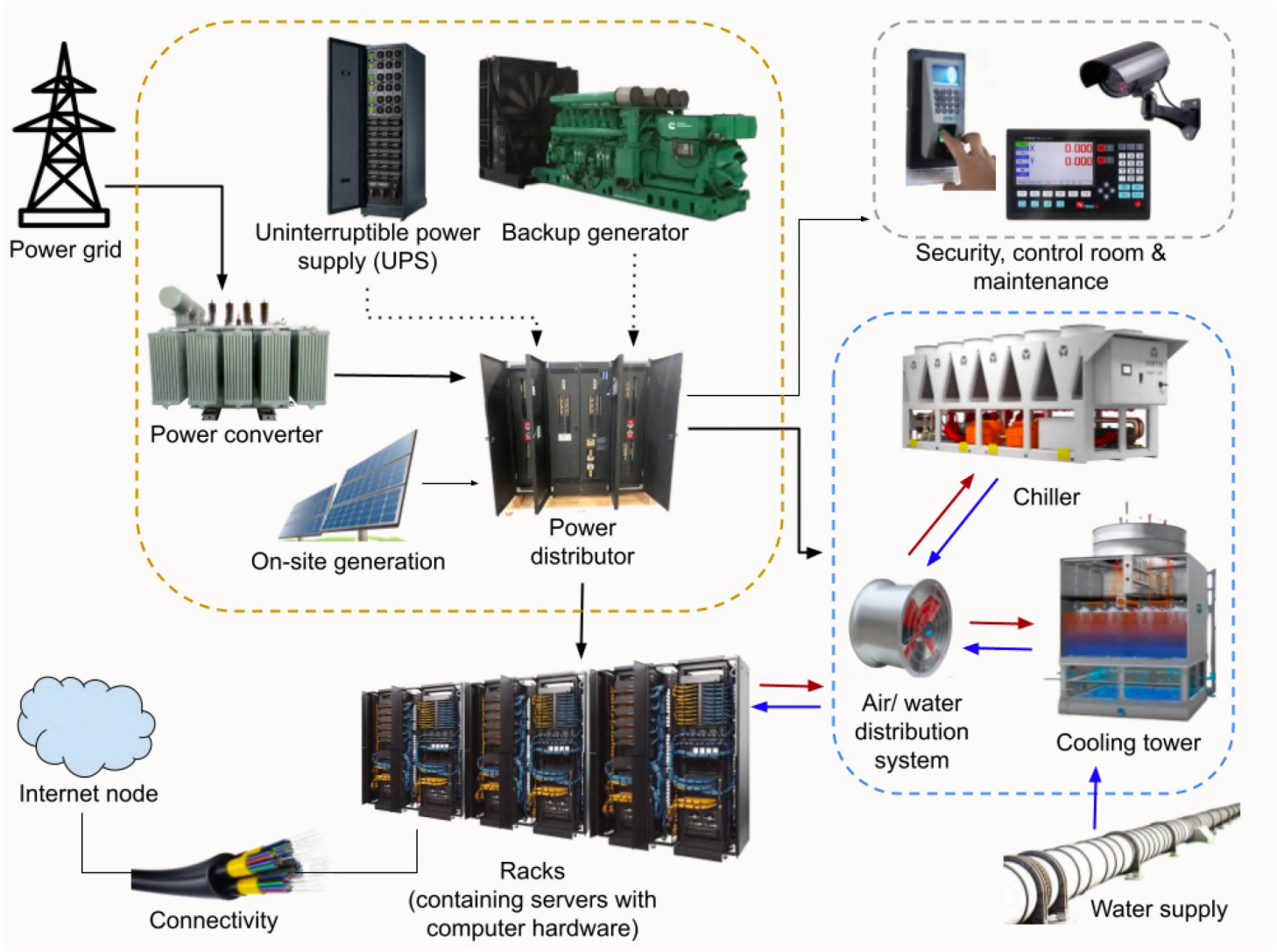


image source: https://arxiv.org/abs/2311.02651

Infrastructure Overview

- **Simplified Components (Based on Figure):**
  - Electrical and cooling infrastructure essential for reliable operation.
  - Data centers receive power from the local grid or on-site generation.

- **Power Flow and Distribution:**
  - Power flows from substations through converters and distributors to hardware.
  - Uninterruptible Power Supply (UPS) and generators provide backup in case of grid failure.
    - UPS temporarily supplies power during short-term failures.
    - Diesel generators activate if outages last longer, ensuring continuous operation.

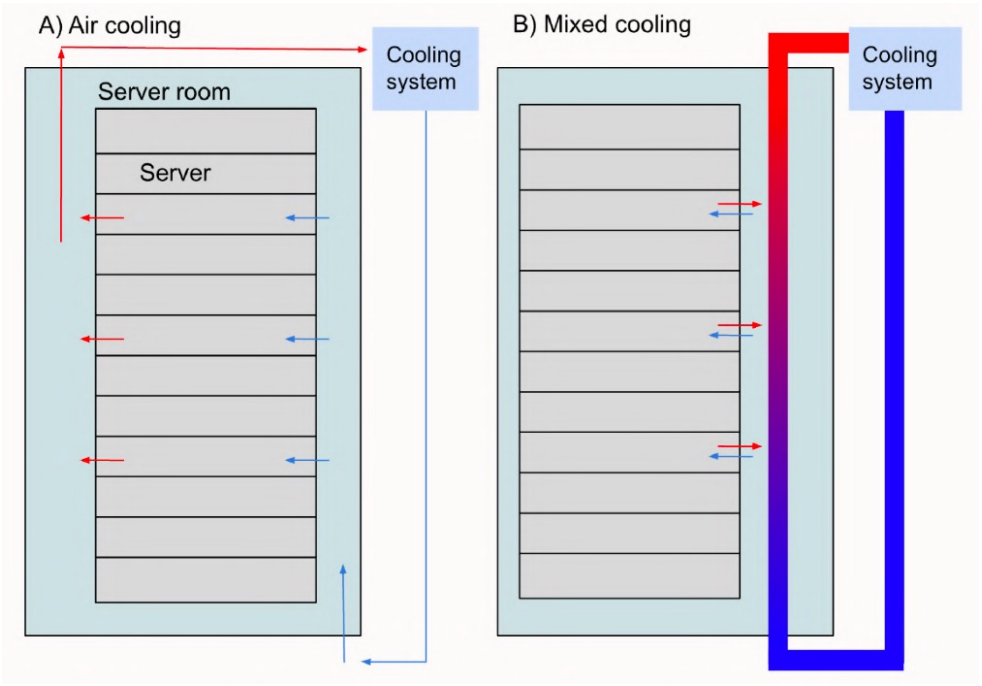
Cooling Systems

- **Air and Water Cooling Methods:**
  - Cooling systems use air, water, or special fluids to manage heat.
  - Cooling towers and chillers are commonly used to maintain optimal temperatures.

- **Heat Management:**
  - Critical to prevent hardware overheating, which can degrade performance and cause failures.
  - Water plays a crucial role in certain data center cooling systems that depend on evaporation, which presents one of the most common cooling solutions.

- **Challenges include managing power consumption, water for cooling needs…**

Cooling server racks



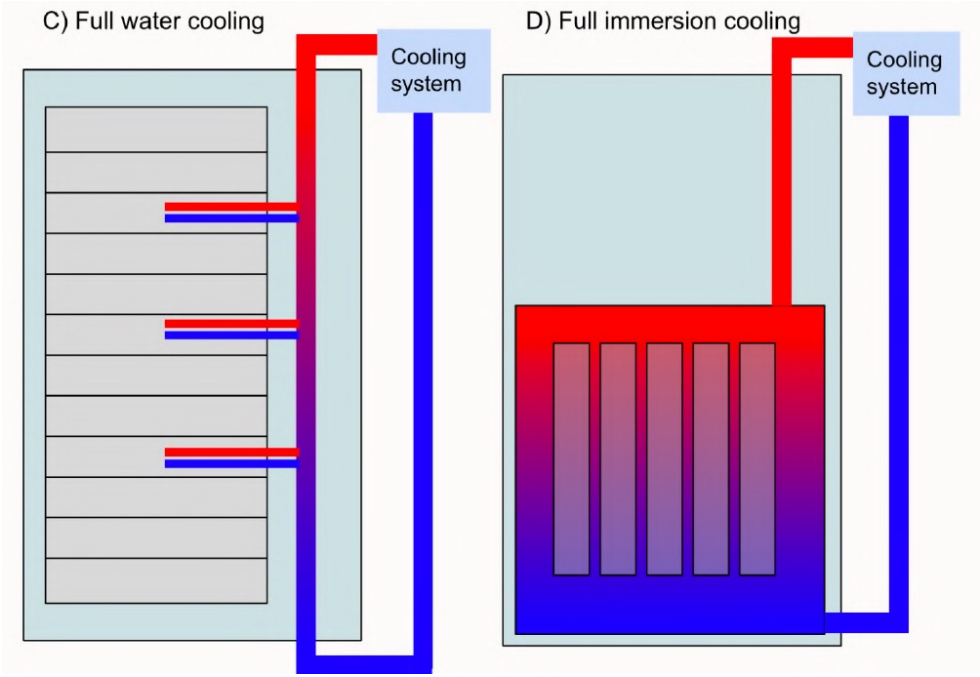


image source: https://arxiv.org/abs/2311.02651

PR01: Presentation

PR01: PRESENTATION

- Presentation of the section:
  HPC and AI convergence
- Types of presentation:
  - Slides presentation (5-10 minutes)
  - All students should submit a PDF version of slides to the Racó (as a PR01) Wednesday 14/10/2024
- Presentation day:
  - Monday 14/10/2024
  - One student will present (randomly) (*)

(*) For maximum grade of this homework the student must attend the class during the presentation day

PR: Student PResentations

- Good practical experience for students!

  and … a way to stimulate homework accomplishment

- Short presentation: “X” minutes + slides (to support presentation)

- 1 student **will be randomly chosen**
  - We'll sum 4 numbers from randomly chosen students and use the '%' function with the total number of students to find the winner in the list.

```
>>> nums_to_add = ...+...+...+...
>>> winner= nums_to_add % num_students +1
>>> print (winner)
```

PR: Student PResentations

- All students should submit a version of slides to the Racó (as a Homework) before the deadline
  - Given the design of the course to maximize student learning, **exercises must be completed with the understanding that any student may be selected to present their work to the class**, followed by a discussion.
  - **To receive full credit, students must attend the presentation class. Failure to attend will result in the exercise grade being halved.**

List of students and their assigned “lucky” number

1	ALVAREZ ARAGONÉS, RUBÉN
2	ÁLVAREZ GÓMEZ-CALCERRADA, LUCÍA
3	ANDREU LOPEZ, RAMON MANEL
4	ATIENZA VILELA, CARLA
5	AUBACH ALTES, ARTUR
6	BAIGES TRILLA, ROGER
7	BARNADAS CONANGLA, EDUARD
8	BARRENECHEA PEREA, PABLO
9	BENNÀSSAR MARTÍN, JOAN
10	BERNAUS CASADESÚS, JOAN
11	BIASIZZO SERRA, ENZO
12	BONET VILA, VIOLETA
13	BRICHS RALLÓ, MIREIA
14	CANTARERO CARRERAS, ADRIÀ
15	CARRIÓN BASTIDA, MARTA
16	CASANOVAS POIRIER, ANNA
17	CHEN, HAO
18	CHEN, PENGCHENG
19	CHEN, ZHIHAO
20	DURÁN LAPLAZA, NILS
21	FIGUERAS FERNÁNDEZ, ALBA
22	FLORES ALBÓ, ADRIÀ
23	FURRIOLS LLIMARGAS, LLUC
24	GIL CASAS, MARIA
25	GONZÁLEZ MONFORT, PABLO
26	GRANJA I BAYOT, JORDI
27	GUTIÉRREZ KITAJIMA, LUIS-KAZUTO
28	HIDALGO PUJOL, PAU
29	IBARS MINGUELLA, ALEIX
30	JEREZ CUBERO, ALBERTO
31	JUNCAROL PI, MARTA
32	LLOPART FERNANDEZ, NURIA
33	LÓPEZ GARCÍA, DANIEL
34	MACIÀ CODERA, NIL
35	MARGARIT FISAS, POL
36	MARTÍNEZ MARTÍNEZ, EVA
37	MEJIA ROTA, CESAR ELIAS
38	MEYA MORALES, MÁXIMO
39	MIRA GARCÍA, MAX
40	MONROY MIR, LOLA
41	MORA LADÀRIA, JAUME
42	NADAL PAR, MARTA
43	NAVARRO NAVARRO, ALEX
44	PEREZ PRADES, POL
45	PRAT MORENO, PAU
46	PUMARES BENAIGES, IRENE
47	RISSO MATAS, ABRIL MARÍA
48	RODRÍGUEZ SANSALONI, MIQUEL
49	ROPERO SERRANO, MIQUEL
50	SELVAS SALA, CAI
51	ZHOU, ZHIQIAN

Guia docent CAP-GIA

- **Calendari de sessions – tardor 2024**

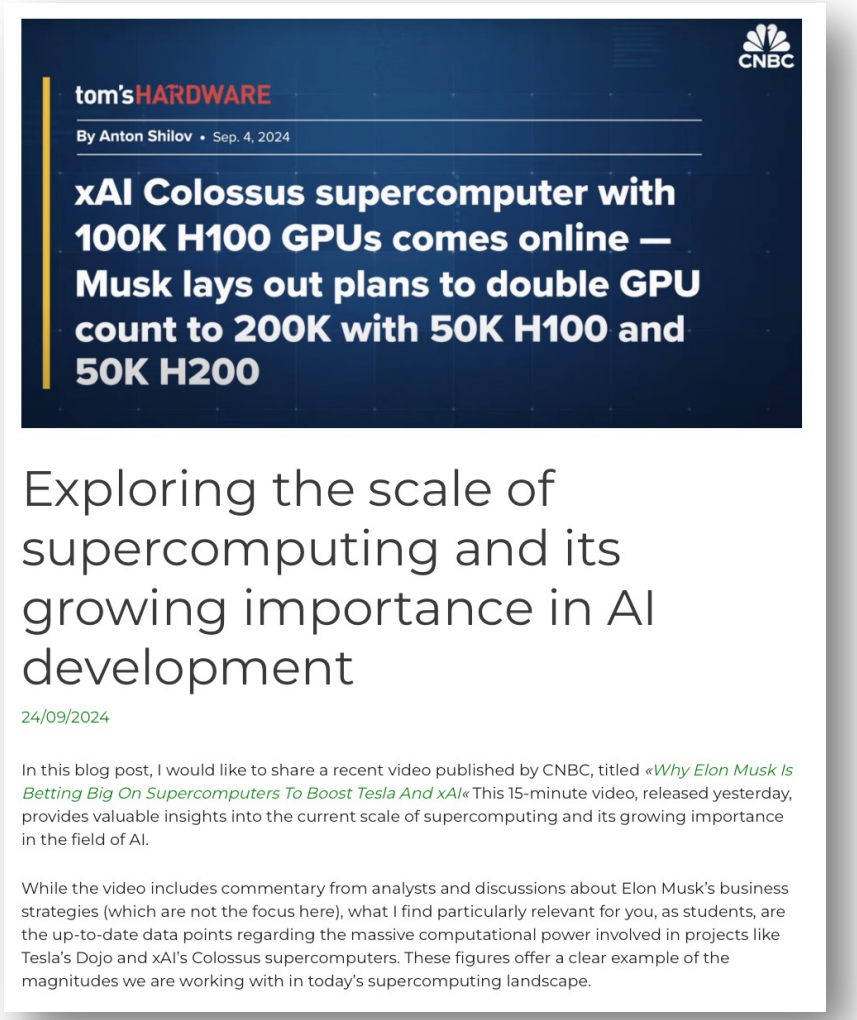
Setmana 1: 09/09 - 13/09	Cloud Computing & Virtual Machines	Lab1 - Màquines Virtuals
Setmana 2: 16/09 - 20/09	Containers	Lab2 - Contenidors
Festiu: 23/09 - 27/09		
Setmana 3: 30/09 - 04/10	Arquitectura de Serveis	Lab3 - Serveis
Setmana 4: 07/10 - 11/10	Altes Prestacions & Supercomputació	Lab4 - MareNostrum 5 (Visita)
Setmana 5: 14/10 - 18/10	Altes Prestacions & AI - [PR1]	Lab5 - GPUs i CUDA
Setmana 6: 21/10 - 25/10	Computació pre-Exascale - [PR2]	Lab6 - Programació pre-Exascale
Setmana 7: 28/10 - 01/10	Arquitectures Big Data	Lab7 - Contenidors + HPC

PR02: Presentation

PR02: PRESENTATION

- **Exploring the scale of supercomputing and its growing importance in IA development (*)**
- Type of presentation:
  - Slides presentation (**8-12** minutes)
  - All students should submit a PDF version of slides to the Racó FIB (Ex02 inbox)
- Deadline and presentation day:
  - Wednesday **21/10/2024** One student will present (randomly) (**)

(**) For maximum grade of this homework the student must attend the class during the presentation day



(*) https://torres.ai/exploring-the-scale-of-supercomputing-and-its-growing-importance-in-ai-development/

CAP-GIA Accounts

CAP-GIA Accounts

Subgrup 11				
	Estudiant 1	Estudiant 2	Usuari MN5-ACC	Usuari MN5-GPP
1	Maria Gil Casas	Anna Casanovas Poirier	nct01151	nct01151
2	Irene Pumares Benaiges	Marta Nadal Par	nct01152	nct01152
3	Alberto Jerez Cubero	Jordi Granja Bayot	nct01153	nct01153
4	Maximo Meya Morales	Alex Navarro Navarro	nct01154	nct01154
5	Pol Pérez Prades	Núria Llopart Fernandez	nct01155	nct01155
6	Abril Risso Matas	Marta Juncarol Pi	nct01156	nct01156
7	Roger Baiges Trilla	Pau Prat Moreno	nct01157	nct01157
8	Pau Hidalgo Pujol	Miquel Rodríguez Sansaloni	nct01158	nct01158
9	Adrià Cantarero Carreras	Luis Kazuto Gutiérrez Kitajima	nct01159	nct01159
10	Adrià Flores Albó	Cai Selvas Sala	nct01160	nct01160
11	Eduard Barnadas Conangla	Pablo Barrenechea Perea	nct01161	nct01161
12	Pablo González Monfort		nct01162	nct01162
13	Max Mira García	Violeta Bonet Vila	nct01163	nct01163

CAP-GIA Accounts

Subgrup 12				
	Estudiant 1	Estudiant 2	Usuari MN5-ACC	Usuari MN5-GPP
14	Miquel Ropero	Pol Margarit	nct01164	nct01164
15	Alba Figueras	Eva Martínez	nct01165	nct01165
16	Carla Atienza	Lucia Álvarez	nct01166	nct01166
17	Marta Carrión	Mireia Brichs	nct01167	nct01167
18	Hao Chen	Pengcheng Chen	nct01168	nct01168
19	Zhiqian Zhou	Zhihao Chen	nct01169	nct01169
20	Nil Macià	Aleix Ibars	nct01170	nct01170
21	Cesar Mejia Rota	Rubén Alvarez Aragones	nct01171	nct01171
22	Joan Bernaus	Joan Bennàssar	nct01172	nct01172
23	Ramon Andreu	Nils Duran	nct01173	nct01173
24	Artur Aubach	Lluc Furriols	nct01174	nct01174
25	Jaume Mora	Lola Monroy Mir	nct01175	nct01175
26	Enzo Biasizzo	Daniel Lopez	nct01176	nct01176

Connect to Marenostrum5

- **You can connect to MareNostrum 5 using the following Public login nodes:**
  - MareNostrum 5 GPP:
    - `glogin1.bsc.es`
    - `glogin2.bsc.es`
  - MareNostrum 5 ACC:
    - `alogin1.bsc.es`
    - `alogin2.bsc.es`
  - Storage 5:
    - `transfer1.bsc.es`
    - `transfer2.bsc.es`
    - `transfer3.bsc.es`
    - `transfer4.bsc.es`

Connect to Marenostrum5

- **All connections must be done through SSH:**
  - OpenSSH for Linux / macOS
    - for example, from Linux:

```
$ ssh nct01XXX@transfer1.bsc.es # Storage5
$ ssh nct01XXX@glogin1.bsc.es # MN5 GPP
$ ssh nct01XXX@alogin1.bsc.es # MN5 ACC
```

  - PuTTY for Windows

- **verify that you are inside the MN5 by running the 'pwd' command to see your current directory**

Changing the password

- **For security reasons, you must change the first password.**
- **To change the password:**
  1. Log in to the transfer1 machine (use the same username and password as in the cluster):

     ```
     mylaptop$> ssh {username}@transfer1.bsc.es
     ```

  2. Run the passwd command and set a new password:

     ```
     transfer1$> passwd
     ```

  3. The new password should become effective about 5 minutes after the change.

Shared Credentials

- **Shared users among:**
  - Data Transfer
  - MareNostrum 5 (GPP and ACC partition)
  - Some other HPC machines

- **Shared filesystem (GPFS)**

- **Same username & passwd**