
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS

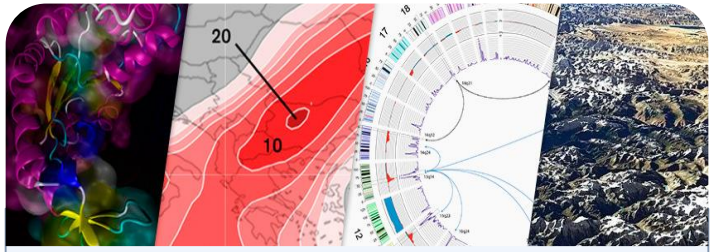
Sergi Moré
Sysadmin team leader

June 2023

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

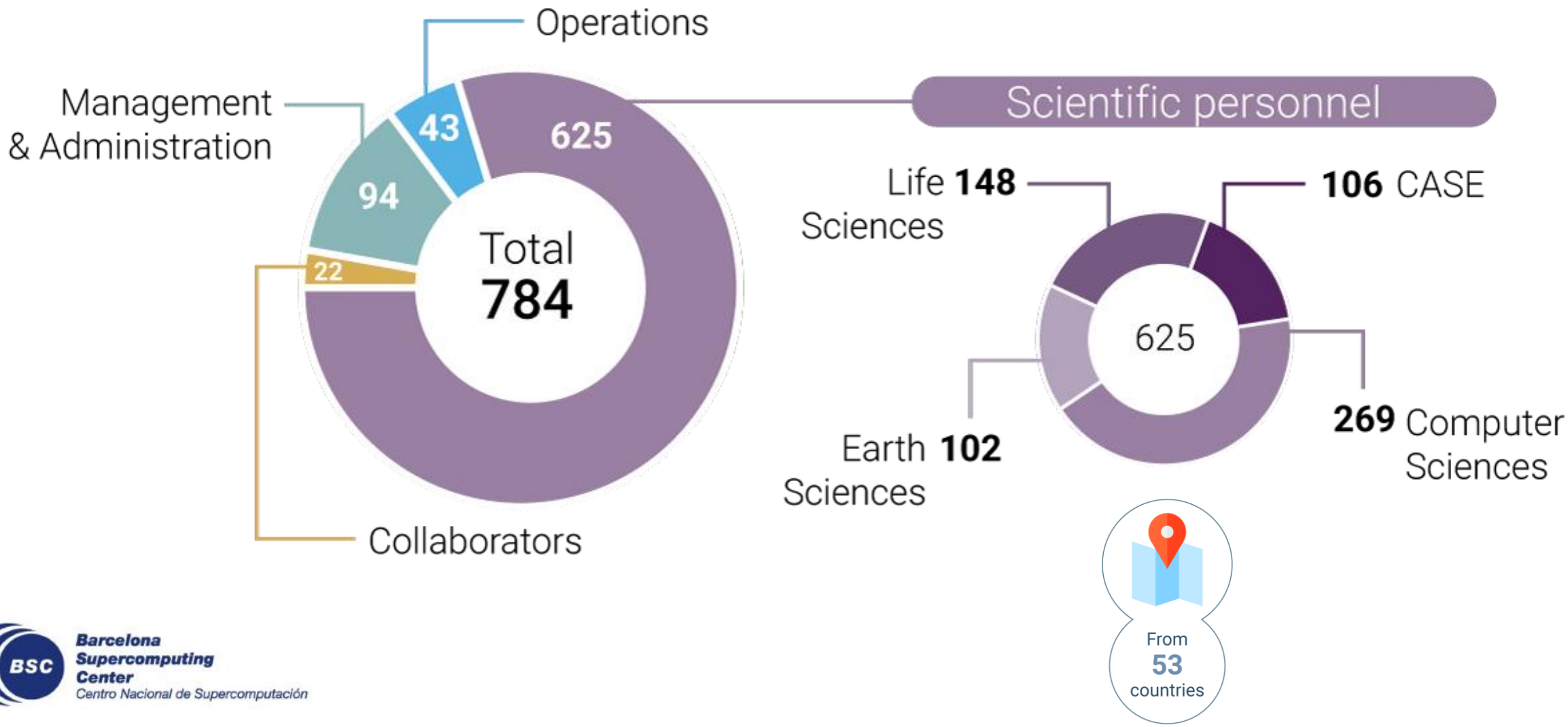
PhD programme, technology
transfer, public engagement

BSC-CNS is a consortium that includes

Spanish Government	**60%**	Gobierno de España · Ministerio de Ciencia, Innovación y Universidades
Catalan Government	**30%**	Generalitat de Catalunya · Departament d'Empresa i Coneixement
Univ. Politècnica de Catalunya (UPC)	**10%**	Universitat Politècnica de Catalunya · BarcelonaTech

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

People



HR EXCELLENCE IN RESEARCH

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 investment

EuroHPC Joint Undertaking

Hosting Consortium:

Spain **Portugal** **Turkey** **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*

Generalitat de Catalunya
Departament d'Empresa i Coneixement

Gobierno de España
Ministerio de Ciencia, Innovación y Universidades

UPC Universitat Politècnica de Catalunya BarcelonaTech

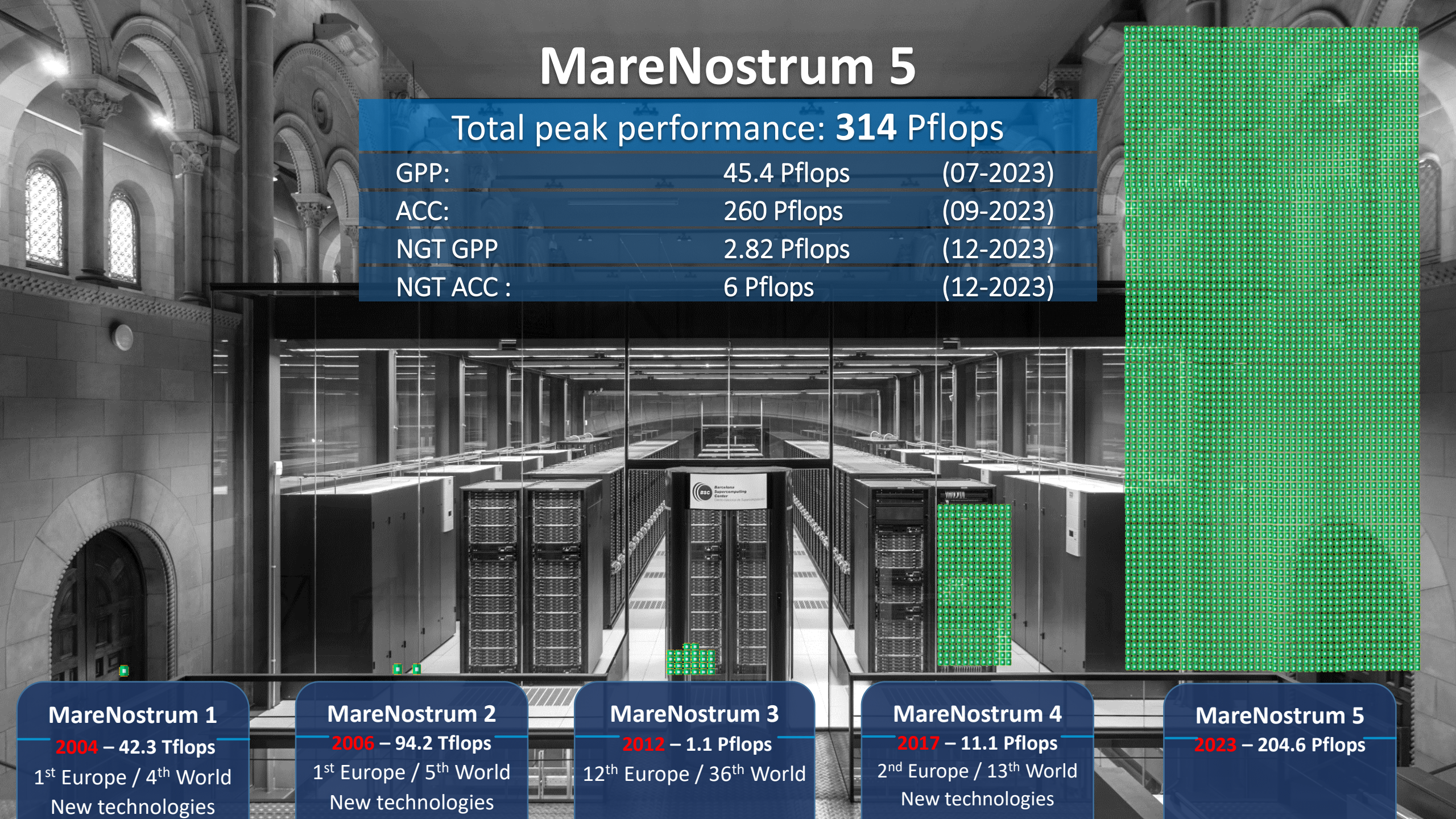
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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, Croatia, and Turkey

EuroHPC Joint Undertaking

MareNostrum 5

Total peak performance: **314** Pflops		
GPP:	45.4 Pflops	(07-2023)
ACC:	260 Pflops	(09-2023)
NGT GPP	2.82 Pflops	(12-2023)
NGT ACC :	6 Pflops	(12-2023)



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

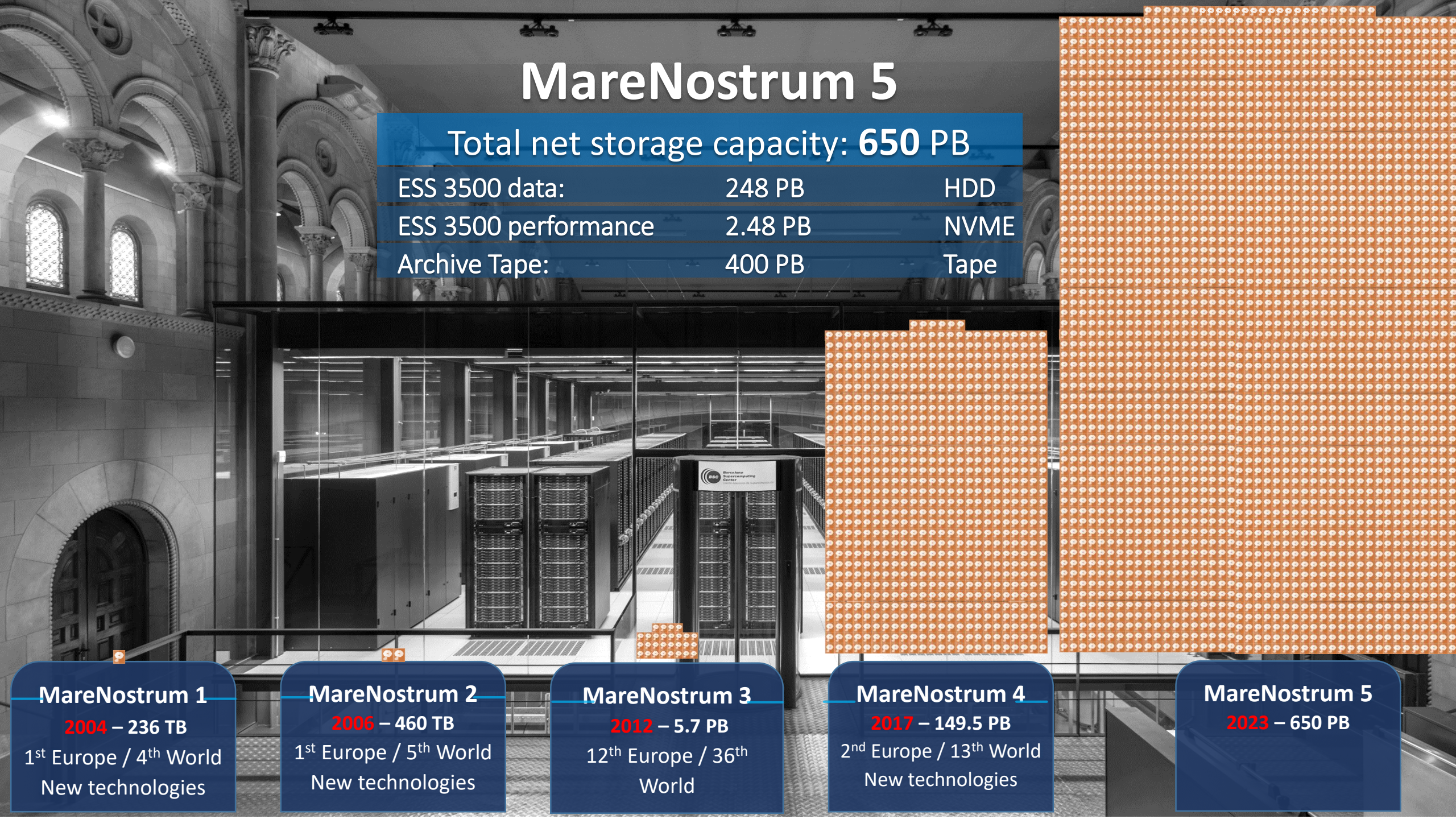
MareNostrum 4
2017 – 11.1 Pflops
2nd Europe / 13th World
New technologies

MareNostrum 5
2023 – 204.6 Pflops

MareNostrum 5

Total net storage capacity: **650** PB

ESS 3500 data:	248 PB	HDD
ESS 3500 performance	2.48 PB	NVME
Archive Tape:	400 PB	Tape



MareNostrum 1
2004 – 236 TB
1st Europe / 4th World
New technologies

MareNostrum 2
2006 – 460 TB
1st Europe / 5th World
New technologies

MareNostrum 3
2012 – 5.7 PB
12th Europe / 36th World

MareNostrum 4
2017 – 149.5 PB
2nd Europe / 13th World
New technologies

MareNostrum 5
2023 – 650 PB

MN5: IO Partition

ESS model	#ESS	Drive Capacity	Total # drives	Raw capacity	Net capacity	Read perf	Write perf
ESS 3500 Capacity model	50	NL-SAS 18TB	20400	367PB	248 PB (8+3P)	1.6TB/s (IOR 100%read)	1.2TB/s (IOR 100%read)
ESS 3500 Performance model	13	NVMe 15.36TB	312	4.79PB	2.81PB (8+2P)	600GB/s 1Mio iops 4KB	600GB/s 500K iops 4KB

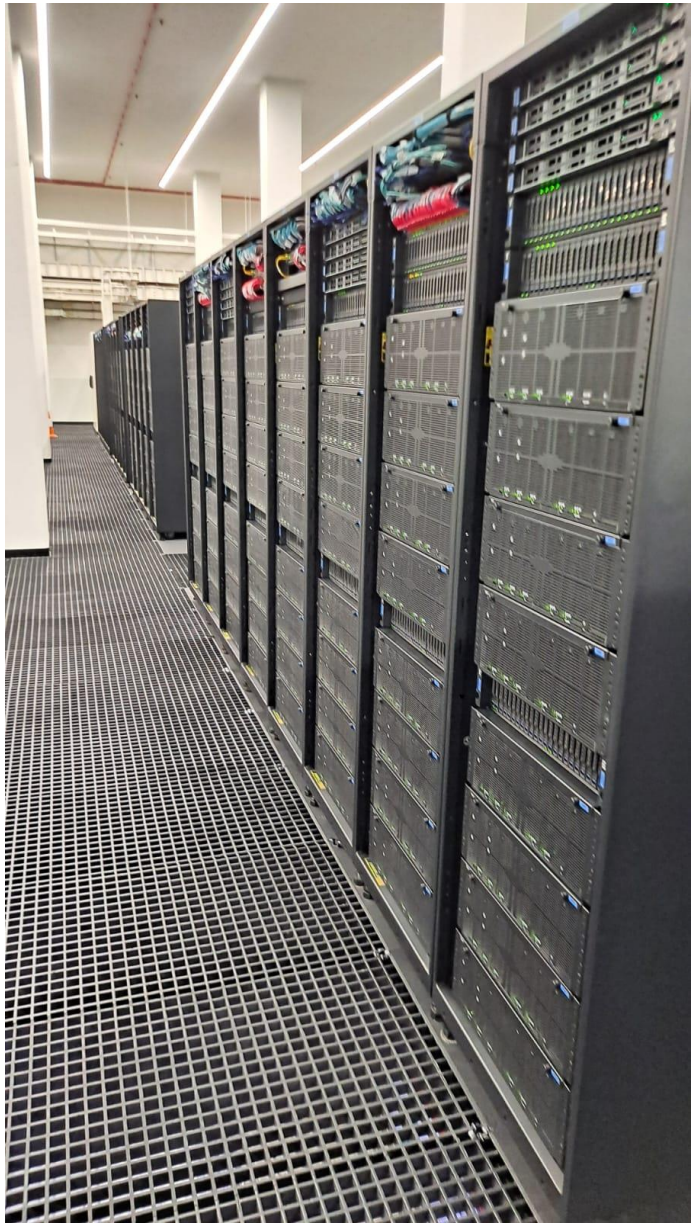
Total net storage capacity: **650** PB

Element	Element	Size
IBM TS4500	2	
Tape Enterprise	20100	400 PB
Drives	64	

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MN5 Filesystems

File System	Size	Data	Metadata	Backup
Projects	20PB	NL-SAS	NVMe	Yes
Scratch	184PB	NL-SAS	NVMe	No
Home	280TB	NVMe+NL-SAS	NVMe	Yes
apps	140TB	NVMe+NL-SAS	NVMe	Yes
archive disk cache	44PB	NL-SAS	NVMe	No



MN5 : Filesystems design

HDD pool. dataOnly NSDs

File System	Capacity	Proposed Capacity	vdisk size	#vdisks / DA	#vdisks / ESS	#vdisks
gpfs_projects	20 PB	22.43 PB	204 TiB (224 TB)	1	2	100
gpfs_scratch	184 PB	179.44 PB	204 TiB (224 TB)	8	16	800
gpfs_archive	44 PB	44.86 PB	204 TiB (224 TB)	2	4	200
gpfs_home	280 TB	231 TB = 156 TB (Flash) + 75 TB (HDD)	1.5 TB	1	2	100
gpfs_apps	184 TB	254 TB = 104 TB (Flash) + 150 TB (HDD)	1.5 TB	2	4	200
	TOTAL	Aprox 247.18 PB				

MN5 : Filesystems design

Flash pool. dataOnly NSDs

File System	Net Capacity	Capacity	vdisk size	#vdisks / server	#vdisks / ESS	#vdisks
gpfs_home	156 TB (312 TB)	312 TB	3.63 TiB (4 TB)	3	6	78
gpfs_apps	104 TB (208 TB)	208 TB	3.63 TiB (4 TB)	2	4	52

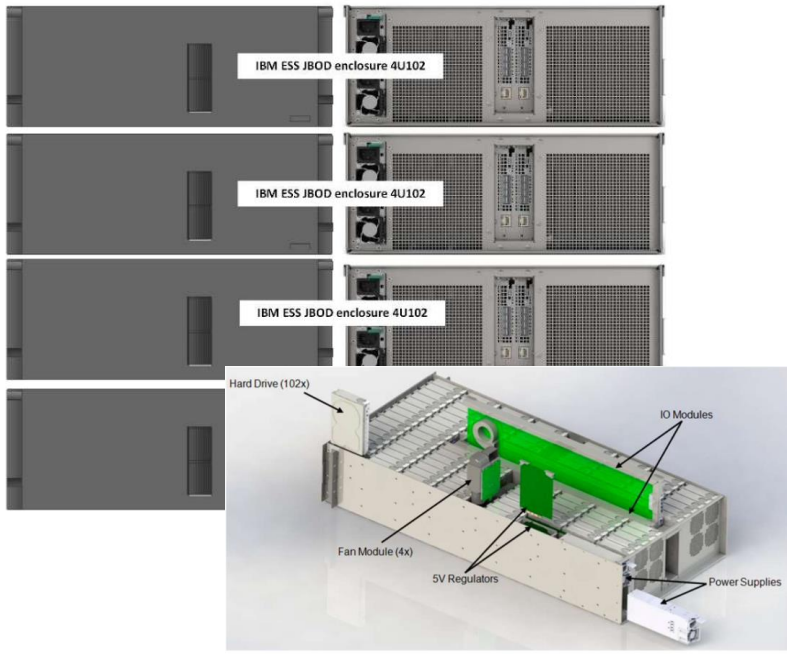
System pool. metadataOnly NSDs

File System	Net Capacity	Capacity	%metadata	vdisk size	#vdisks / server	#vdisks / ESS	#vdisks
gpfs_projects	21.99 PB	104 TB	0.47% (incl. rep)	3.63 TiB (4 TB)	1	2	26
gpfs_scratch	175.92 PB	1040 TB	0.59% (incl. rep)	3.63 TiB (4 TB)	10	20	260
gpfs_archive	440 PB	1040 TB	0.28% (incl. rep)	3.63 TiB (4 TB)	10	20	260
gpfs_home	231 TB	20.8 TB	9% (incl. rep)	727.59 GiB (800 GB)	1	2	26
gpfs_apps	254 TB	10.4 TB	4.1% (incl. rep)	363.79 GiB (400 GB)	1	2	26

MN5 HPC Storage

- ESS 3500
  - 5th generation Elastic Storage Server
  - 2 x ESS data servers
    - 1x48 core AMD EPYC Rome 2.2GHZ
    - 512 GB DDR4
    - PCI Gen 4
    - 2 x NDR 200
    - 2 x 100 Gbps Eth
- Performance model
  - 24 x NVMe PCI-attached flash drives
  - 368TB raw - 2U
- Capacity model
  - 4 x JBOD
    - 102 HDD
    - 4U
  - 7.344 PB raw - 18U

IBM ESS 3500



Elastic Storage Server Storage Enclosure 4U102

MN5 Archive Storage

- 2 x IBM TS4500
- 44 PB disk cache
- 400PB total tape capacity
- 20100 x 20TB Enterprise tapes
- 64 Drives
- 8 x Spectrum Archive servers

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MN5 Storage: Service and Management

- 8 x Internal data transfer
  - HPC file systems backup to current Tape infrastructure using Spectrum Protect.
  - Unattended data movement from HPC to Archive.
- 4 x Export servers
  - CES protocol nodes: Provide nfs/cifs/swift access to HPC file systems.
    - HBP Federated FTS3 Service

- 4 x External data transfer
  - Storage logins.
  - Transfer data from/to internet.
- Management:
  - 2 x EMS servers

    Deploy of ESS servers based on ansible Storage Monitoring.
  - 2 x xCAT servers

    Deploy service servers.
  - 5 x manager servers

    Filesystem and tocken management

Santa’s wish list

Installation – Lessons learn and possible improvements

- From xCAT to Ansible
- Monitoring
- File system and token managers
- Software release cycle
- Sync between GPFS filesytems

Installation – Lessons learn and possible improvements

- From xCAT to Ansible
  - Perfect if everything goes smoothly, but really hard to debug problems
  - Feeling that has been designed with a far smaller system in mind
    - Not enough level of parallelism. Improved in 6.1.6.1
    - Single operation at a time from the container
      - 2 to 3 hours for a config load (To much coffe time for our healthiness)

Installation – Lessons learn and possible improvements

- From xCAT to Ansible
  - Missing some xCAT features
    - OS rinstall
      - Server / partition filled up by mistake. /etc/fstab become corrupted and server did not boot anymore.
      - Solved in 30m with xCAT.
      - Can take weeks now.
    - PDU Management
  - essrpower
    - Could be good to have range support, xCAT like.

      # essrpower –n ess[01-50]s[1-2] –t on

```
[root@stgcabeza1 ~]# rpower ess01 pdustat
ess01e1: stg01pdu1 operational state for outlet 2 is on
ess01e1: stg01pdu2 operational state for outlet 2 is on
ess01e2: stg01pdu1 operational state for outlet 11 is on
ess01e2: stg01pdu2 operational state for outlet 11 is on
ess01e3: stg01pdu1 operational state for outlet 9 is on
ess01e3: stg01pdu2 operational state for outlet 9 is on
ess01e4: stg01pdu1 operational state for outlet 7 is on
ess01e4: stg01pdu2 operational state for outlet 7 is on
ess01e5: stg01pdu1 operational state for outlet 3 is on
ess01e5: stg01pdu2 operational state for outlet 3 is on
[root@stgcabeza1 ~]#
```

Installation – Lessons learn and possible improvements

- Monitoring
  - EMS servers do not scale
  - 8GB daily messages logs

```
[root@stgems1 log]# ls -lah messages-20230601
-rw------- 1 root root 8.6G Jun  1 03:37 messages-20230601
[root@stgems1 log]#
```

  - mmhealth and remote clusters
    - Please, allow us not to receive alerts from remote clusters. About 9000 remote clients.

Installation – Lessons learn and possible improvements

- File system and token managers
  - No specific servers appointed in original design.
  - EMS and ESS to busy to take such role
  - 5 servers appointed as managers

Installation – Lessons learn and possible improvements

- Software release cycle
  - System shipped with 6.1.4.1
  - 16/02/23 – Installed 6.5.1.1
  - 14/03/23 – Installed 6.1.6.0
  - 27/04/23 – Installed 6.1.6.1
  - 29/06/23 – Plan to install 6.1.8.0

Installation – Lessons learn and possible improvements

- Software release cycle
  - System shipped with 6.1.4.1
  - 16/02/23 – Installed 6.5.1.1
  - 14/03/23 – Installed 6.1.6.0
  - 27/04/23 – Installed 6.1.6.1
  - 29/06/23 – Plan to install 6.1.8.0
    - 6.1.6 out of support
    - Update everything but archive nodes.
    - Spectrum Archive does not yet support Storage Scale 5.1.8

4 upgrades in 4 months

System not yet accepted and already running unsupported versions

Installation – Lessons learn and possible improvements

- Software release cycle
  - System shipped with 6.1.4.1
  - 16/02/23 – Installed 6.5.1.1
  - 14/03/23 – Installed 6.1.6.0
  - 27/04/23 – Installed 6.1.6.1
  - 29/06/23 – Plan to install 6.1.8.0
    - 6.1.6 out of support
    - Update everything but archive nodes.
    - Spectrum Archive does not yet support Storage Scale 5.1.8

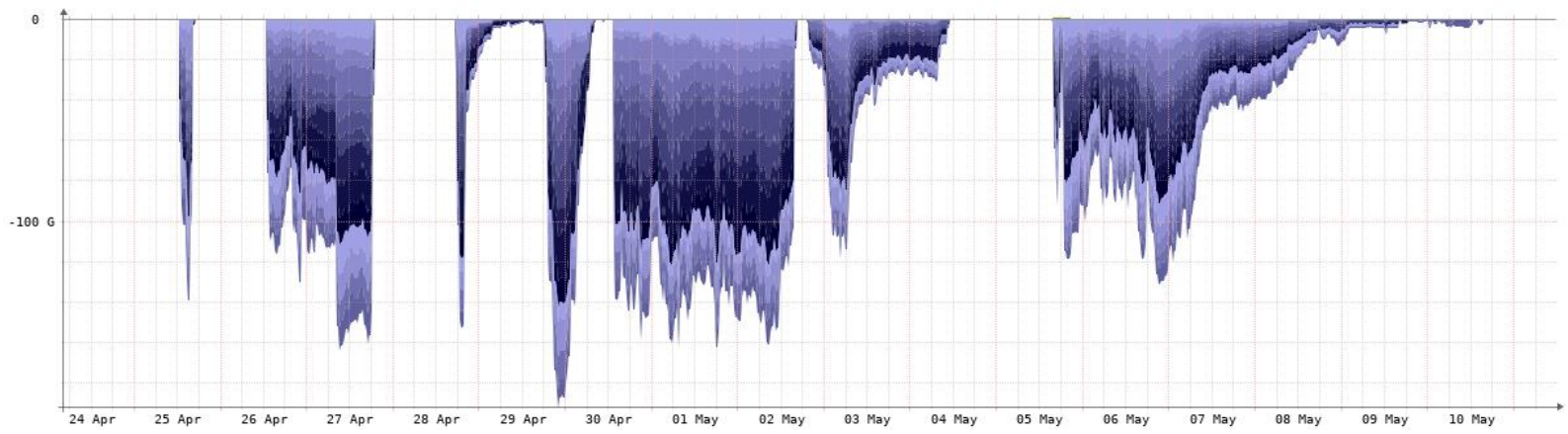
4 upgrades in 4 months

System not yet accepted and already running unsupported versions

- Goal to go to 6.1.9 for LTS
- In our ideal world, once in production, only security updates will happen.

Installation – Lessons learn and possible improvements

- Tool to sync GPFS file systems from different clusters
  - Need to sync data between new systems and legacy ones.
  - Currently doing massive rsyncs distributing task against directories
  - Copied 12PB in about ~10days
    - 8 servers – 400Gbps available – 160 threads.
    - About 100Gbps mean bandwidth usage.
  - Too manual stuff needed, prone to errors
  - Hard to debug errors and check consistency between the two copies



Installation – Lessons learn and possible improvements

- It is always good to have good people around you.
- Special thanks to Jordi Caubet and Luis Bolinches for all the work and effort they have put in MN5 project

jordi.caubet@es.ibm.com
luis.bolinches@fi.ibm.com

Thank you

sergi.more@bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación