Barcelona Supercomputing Center
Centro Nacional de Supercomputación

HPC User Support at MN5

David Vicente
Head of HPC User support team

October 2022

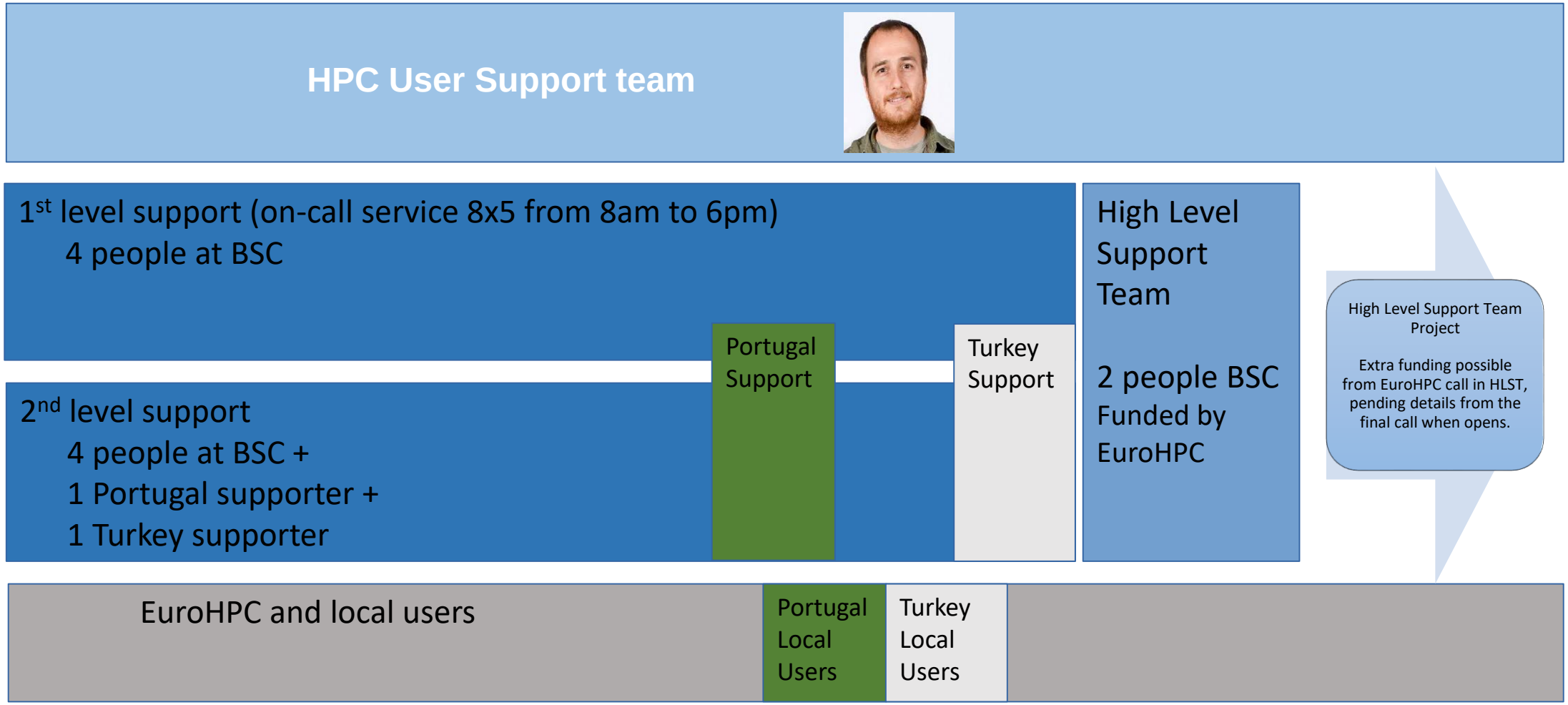
Operations team in charge of MN5

GROUP	HEAD		SERVICE/RESPONSIBILITY
System Administration	Javier Bartolomé		System Administration Networking and security Batch system and monitoring
User Support	David Vicente		Application Enabling First & Second HPC Level Support 3D Visualization High level HPC support
Facility Management	Miguel Armenta		Infrastructure, building and facilities
Data Management	Nadia Tonello		Data management support Cloud support
Infrastructure Access Policy Unit	Oriol Pineda		Access mechanisms to BSC Resources coordination of EuroCC Spain CC

User Support Structure

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

User Support Team for MN5



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

User Support tasks

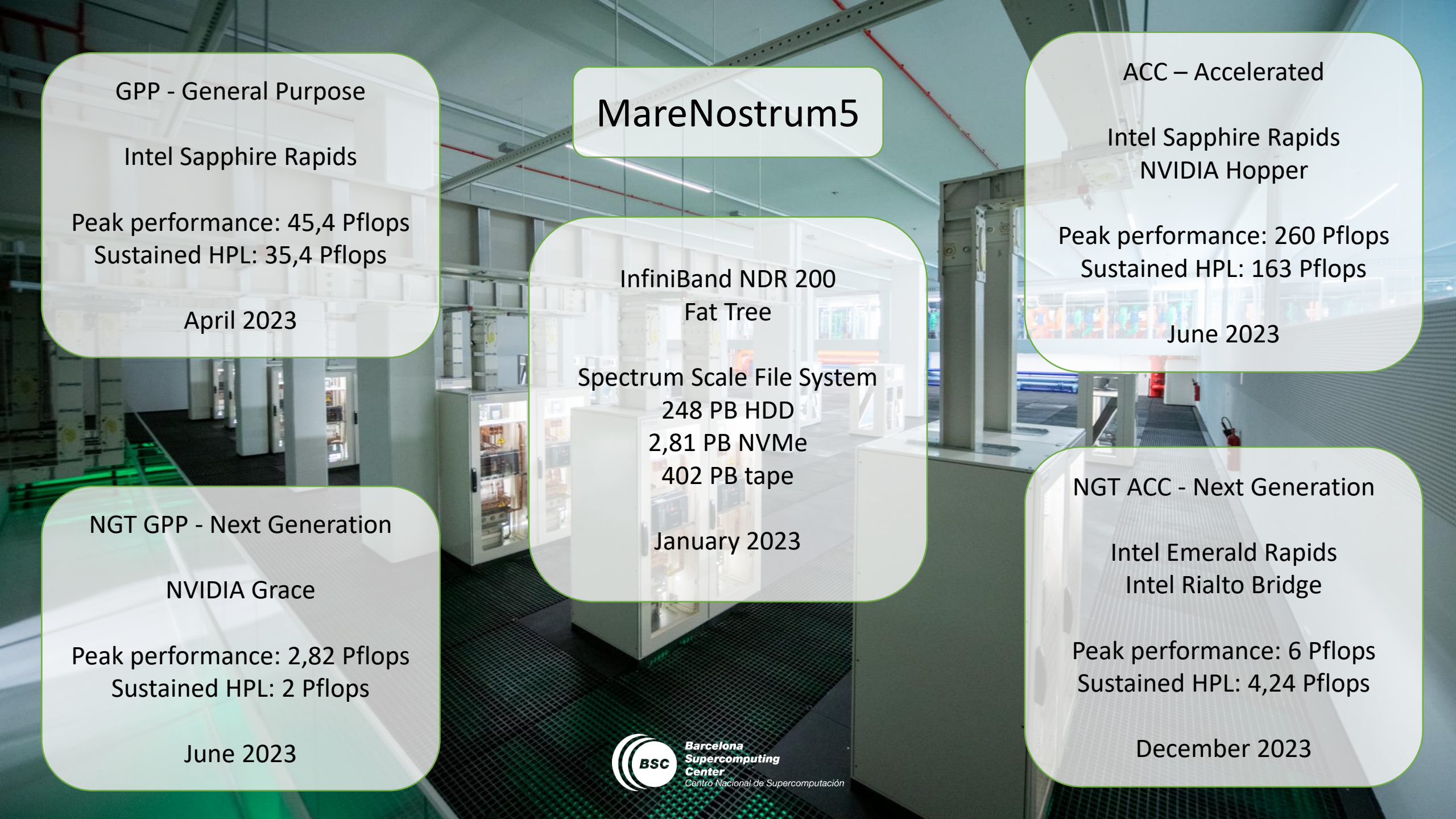
- 1st Level support tasks
  - User Account creation and cpu and disk accounting management
  - Ticketing system management
  - 1st level filter for requests and incidents
  - Contact with the users related to maintenance tasks or other notifications
  - Basic compilations and module management
  - Other tasks requiring basic HPC knowledge and short time duration
- 2nd Level support tasks
  - Compilation and configuration
  - Optimization and tuning of the codes for the MN5 architecture
  - Debugging and performance analysis of codes for large executions
  - Efficient use of the allocated resources
  - Data Management
  - Management of installation frameworks like EasyBuild or Spack

User Support tasks

- 3rd Level support tasks
  - Long development activities related to optimization, scalability for exascale machines, or new architectures optimization
  - Improvement of large used applications or linked with COEs (center of excellence)
  - The 3rd level activities will be managed by the HLST team. As this group has a limited human power, the tasks will be selected with an agreement between HLST-host site and EUROHPC.
  - Each of the tasks of HLST team can imply long developments of more than 6 months.

User Support Timeline for MN5

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



MareNostrum5

GPP - General Purpose

Intel Sapphire Rapids

Peak performance: 45,4 Pflops
Sustained HPL: 35,4 Pflops

April 2023

ACC – Accelerated

Intel Sapphire Rapids
NVIDIA Hopper

Peak performance: 260 Pflops
Sustained HPL: 163 Pflops

June 2023

InfiniBand NDR 200
Fat Tree

Spectrum Scale File System
248 PB HDD
2,81 PB NVMe
402 PB tape

January 2023

NGT GPP - Next Generation

NVIDIA Grace

Peak performance: 2,82 Pflops
Sustained HPL: 2 Pflops

June 2023

NGT ACC - Next Generation

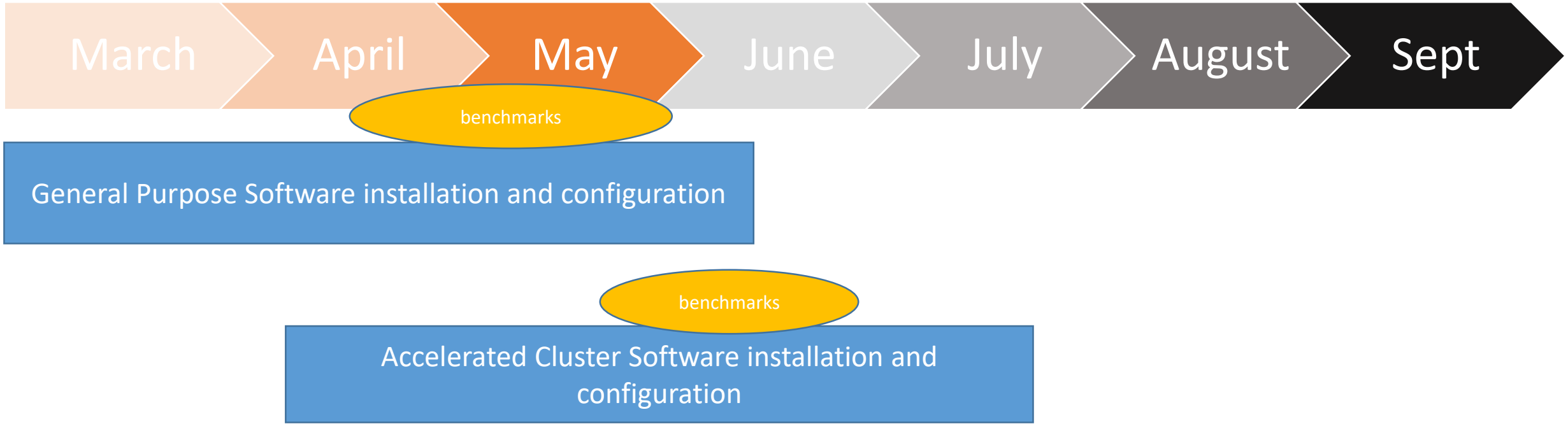
Intel Emerald Rapids
Intel Rialto Bridge

Peak performance: 6 Pflops
Sustained HPL: 4,24 Pflops

December 2023

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

User Support tasks



During the installation phase only on-site access will be permitted. So on-site presence will be required for some training and installation activities.

Compute partitions overview

		Cooling	Nodes Total	Technology	Processor/Accelerator		Memory	PFlops (HPL)		Local Drive	High-Perf. Network
	General Purpose	DLC +RDHX	>6000	Lenovo	2x Intel Sapphire R.		>2GB/core 256GB DDR5	35.43	>205	960GB NVMe	1x NDR200 Shared by 2 nodes
			>200				>8GB/core 1024GB DDR5				
			>50		2x Intel Sapphire R. HBM		> 0.5GB HBM/core 128GB HBM + 32GB DDR5	0.34			
	Accelerated	DLC	> 1000	Atos	2x Intel Sapphire R.		512GB	163		480GB NVMe	4x NDR200
					4x Nvidia Hopper 64GB HBM						
Next Gen	General Purpose	AC +RDHX	> 400	Atos	Nvidia Grace	144c @ > 2.4GHz	240GB LPDDR5	2		128GB NVMe	1x NDR200
	Accelerated	DLC +RDHX		Lenovo	2x Intel Emerald R.		512GB DDR5	4.24		960GB NVMe	2x NDR
					4x Intel Rialto Bridge ≧128GB HBM						

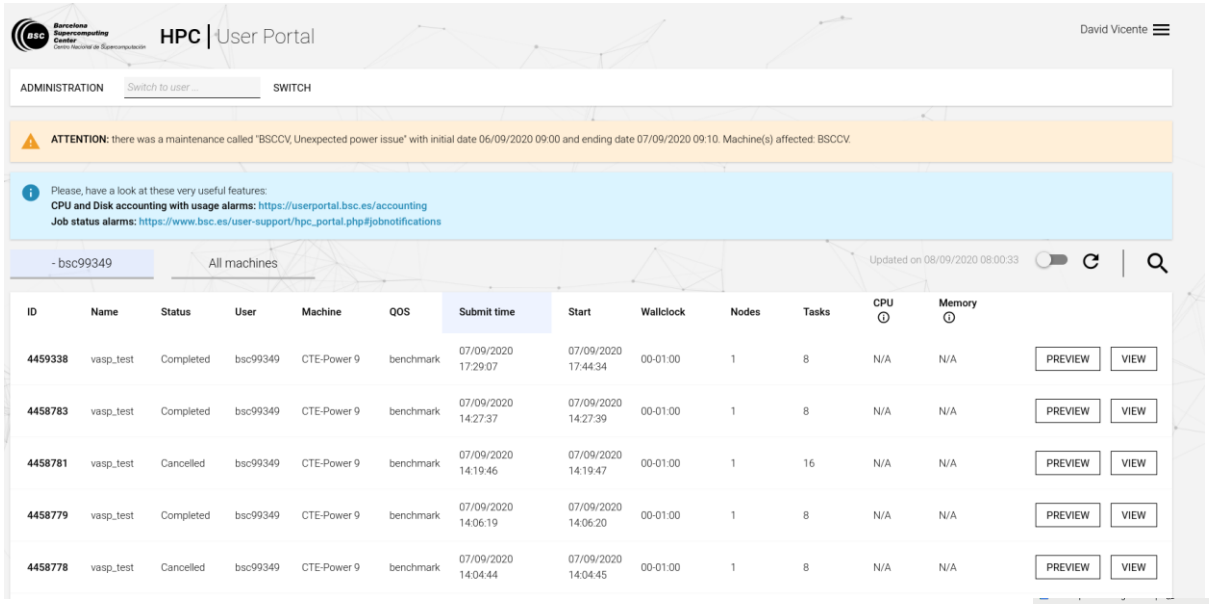
User Support Tools

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

User Support tools

- Requirements from users :
  - Accounting
  - Job Monitoring
  - Job status
  - System usage
  - E-mail Alarms for disk quota and cpu-h quota
  - E-mail Alarms for jobs (starting, end)
- Extra Requirements from the User support team to provide high level support
  - System monitoring
  - Job Monitoring
  - Job behavior status
  - Power consumption Monitoring
  - Full view of the system usage per queue and account
  - Usage per application

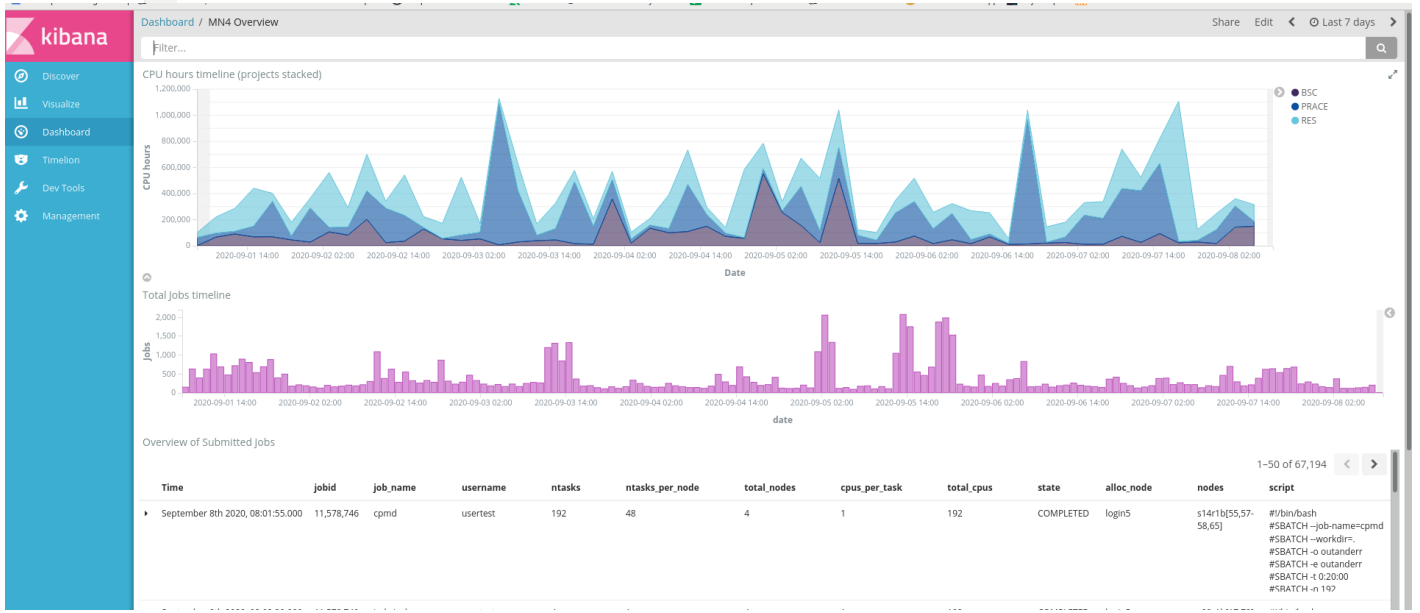
BSC Solution : userportal.bsc.es + Kibana



Userportal is a BSC development providing the final HPC users with the information about their jobs and other functionalities to improve their HPC usability and performance.

For the internal monitoring of the job status, BSC is using the Slurm plugin for Elastic search, using a Kibana visualization* system to interact with the information.

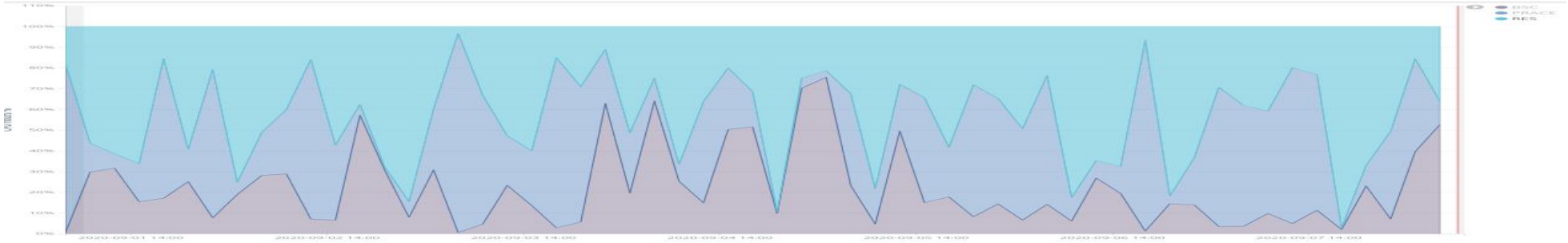
*Only admins/support have access to this system.



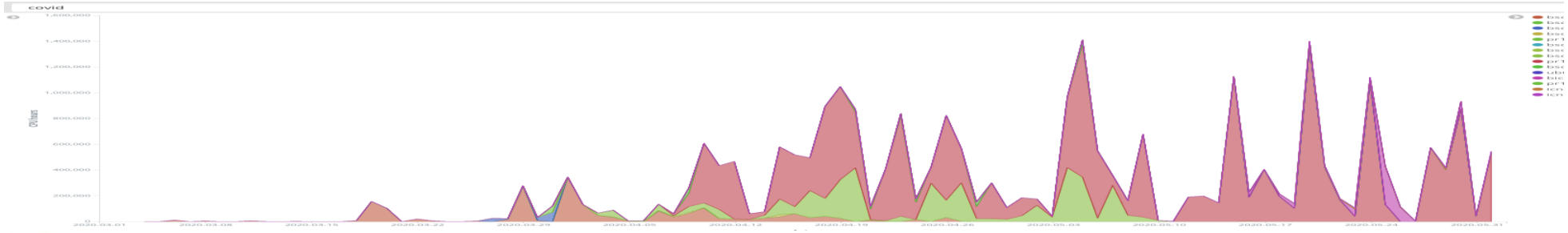
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Kibana functionalities

- The ElasticSearch has all the slurm information about the jobs submitted to our main HPC cluster MareNostrum4. It permits to generate queries in Kibana to extract any information related to accounting per group, account, qos, etc.
- For example the usage in CPU-h of the 3 main areas (RES,BSC and PRACE) during the last 7 days :



- Or thanks to the powerful search engine from kibana, we are able to parser all the job scripts to find all the cpu hours consumed in the system by specific application or tag, for example COVID



UserPortal
User functionalities

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Job status (I)

Job 4459338

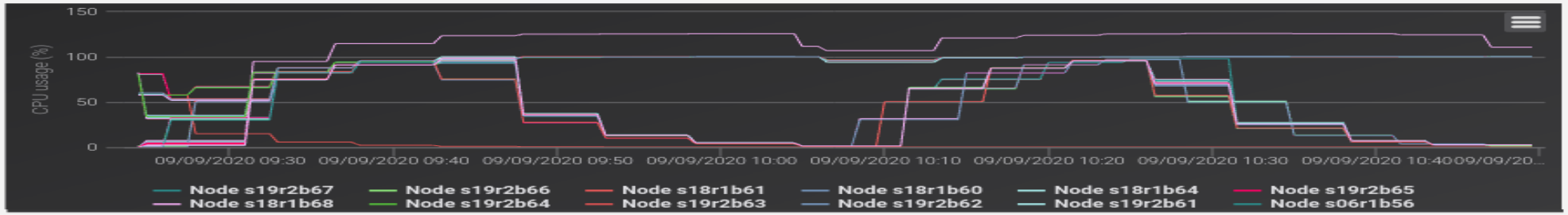
Job details

Machine: CTE-Power 9
ID: 4459338
Name: vasp_test
Status: Completed
Load status: Ok
Submit time: 07/09/2020 17:29:07
Start time: 07/09/2020 17:44:34
End time: 07/09/2020 17:56:38
Wallclock: 1 hour
Run time: 12 minutes, 4 seconds
Submit node: p9login1
Is batch? Yes
Batch node: p9r2n09
Last updated: 07/09/2020 18:07:10

Max(Memory %): N/A
Avg(CPU %): N/A
Nodes: 1
Tasks: 8
Number of CPUs: 160
Shared: No
Dependency: None
Features: None
General Resources: gpu:4
User: bsc99349
Account: bsc99
Partition: main
QOS: benchmark
Reservation: None

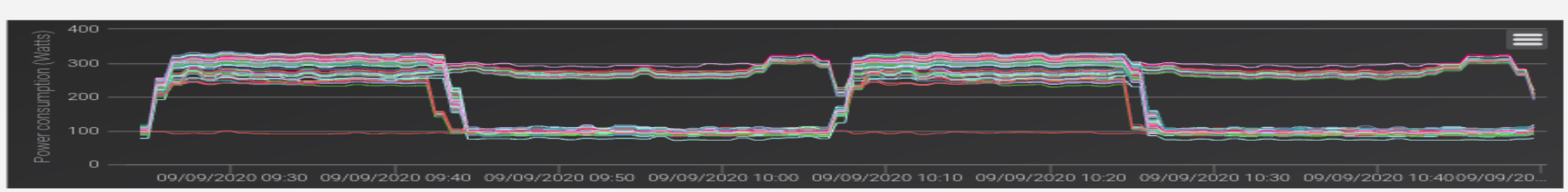
Job status (II)

CPU usage

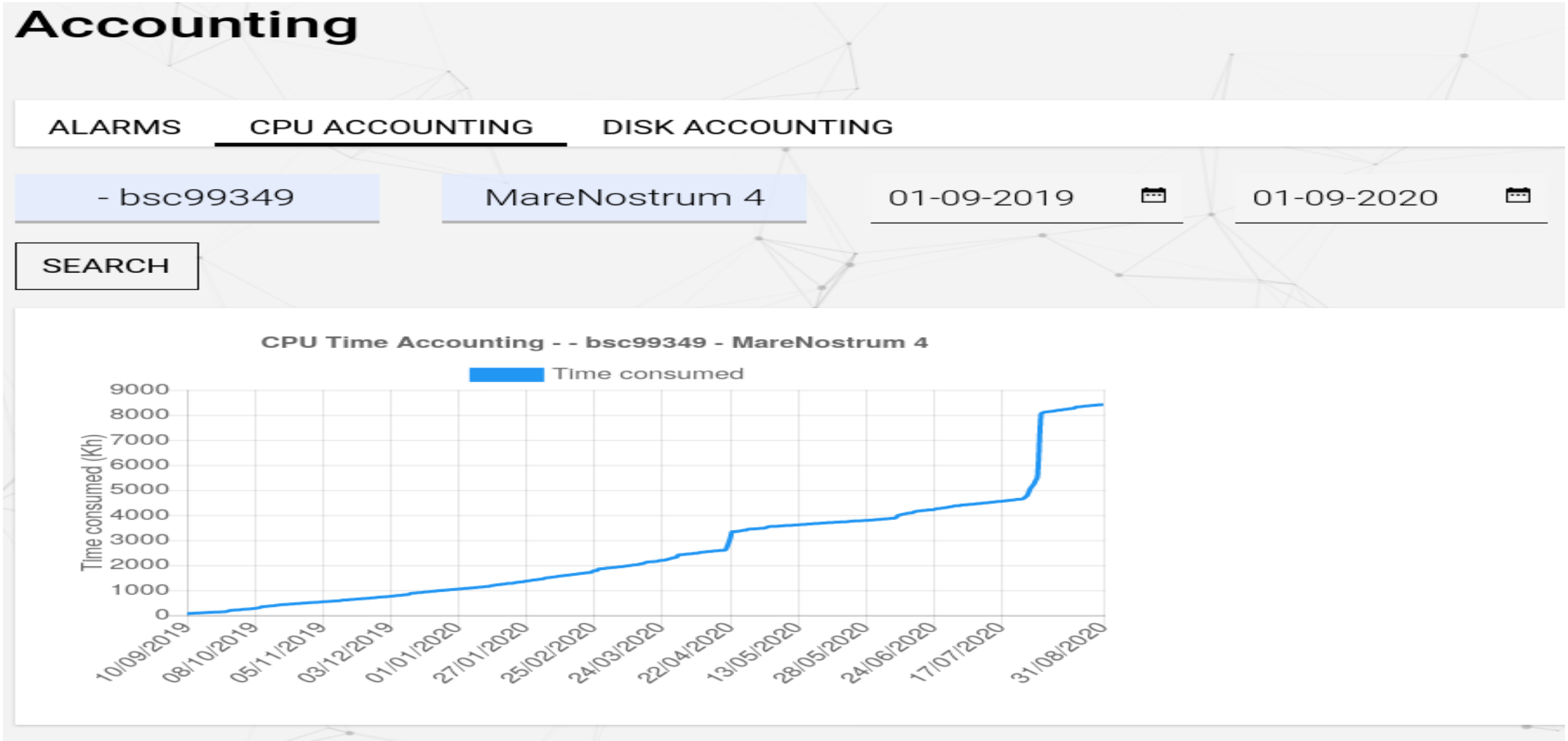


Memory usage

Power consumption

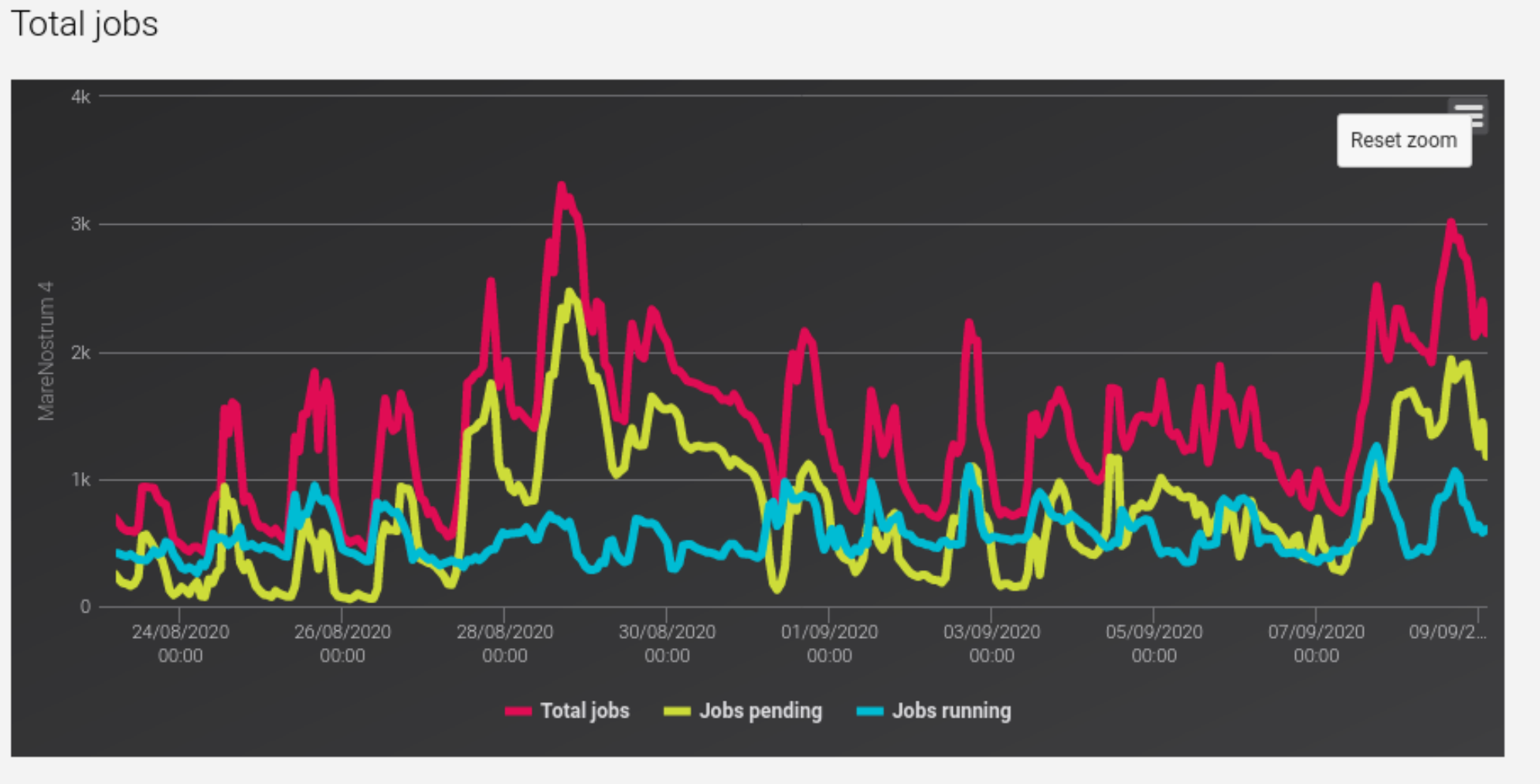


Accounting

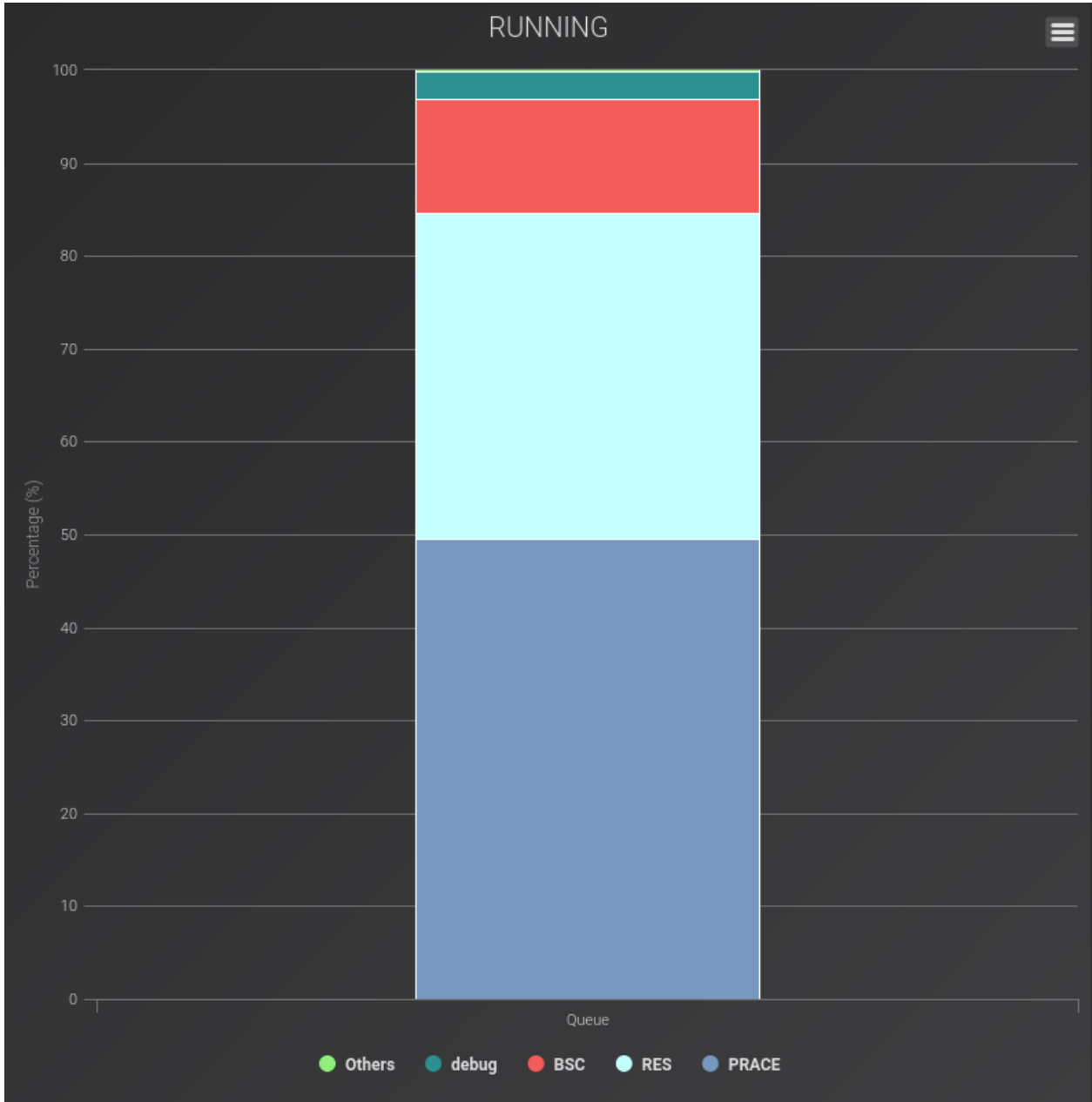

Accounting
ALARMS
CPU ACCOUNTING
DISK ACCOUNTING
- bsc99349
MareNostrum 4
01-09-2019
01-09-2020
SEARCH
CPU Time Accounting - - bsc99349 - MareNostrum 4
Time consumed
Time consumed (Kh)
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
10/09/2019
08/10/2019
05/11/2019
03/12/2019
01/01/2020
27/01/2020
25/02/2020
24/03/2020
22/04/2020
13/05/2020
28/05/2020
24/06/2020
17/07/2020
31/08/2020

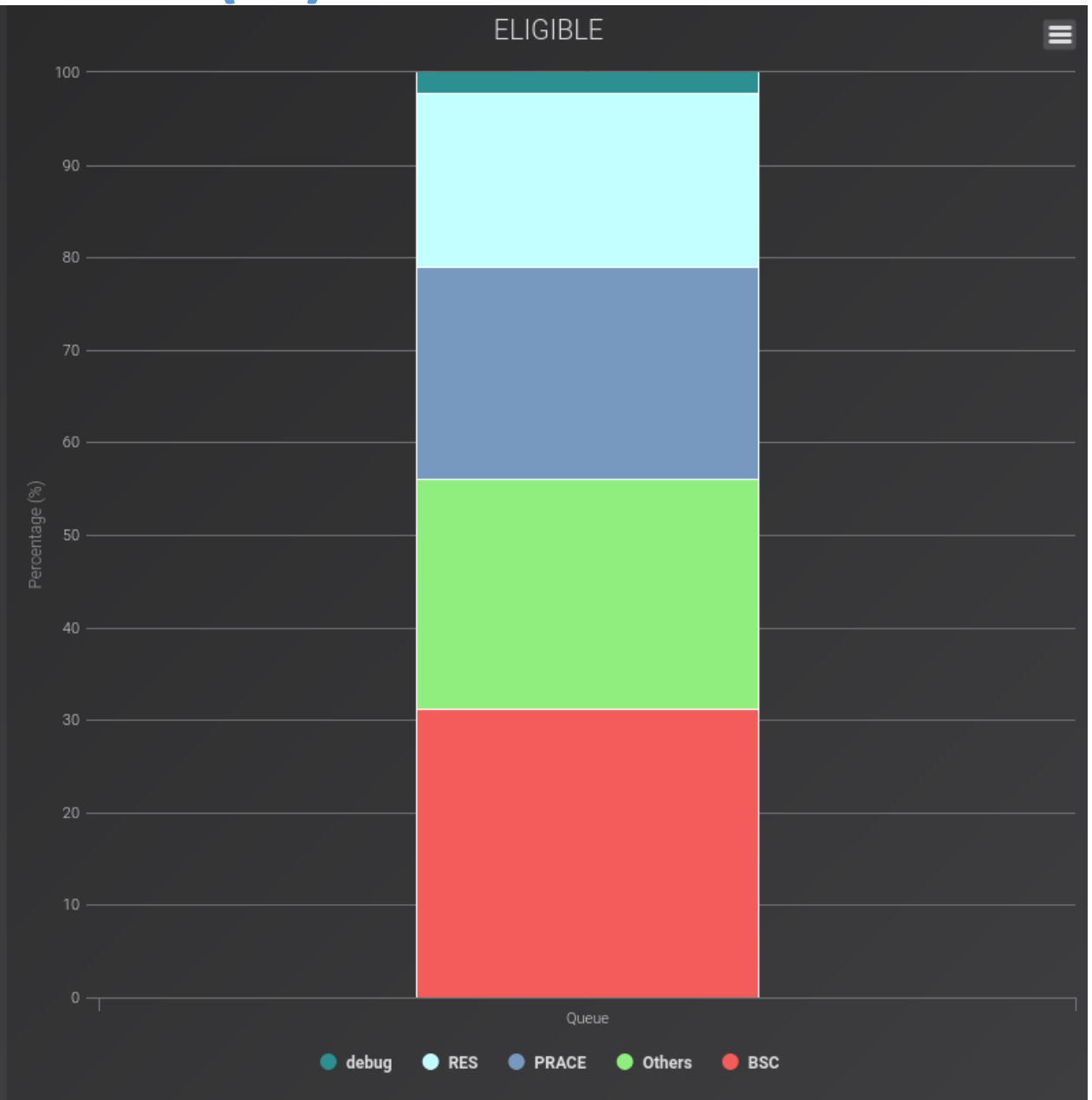
Machines Stats (I)

Machines Stats (II)



Machines Stats (III)



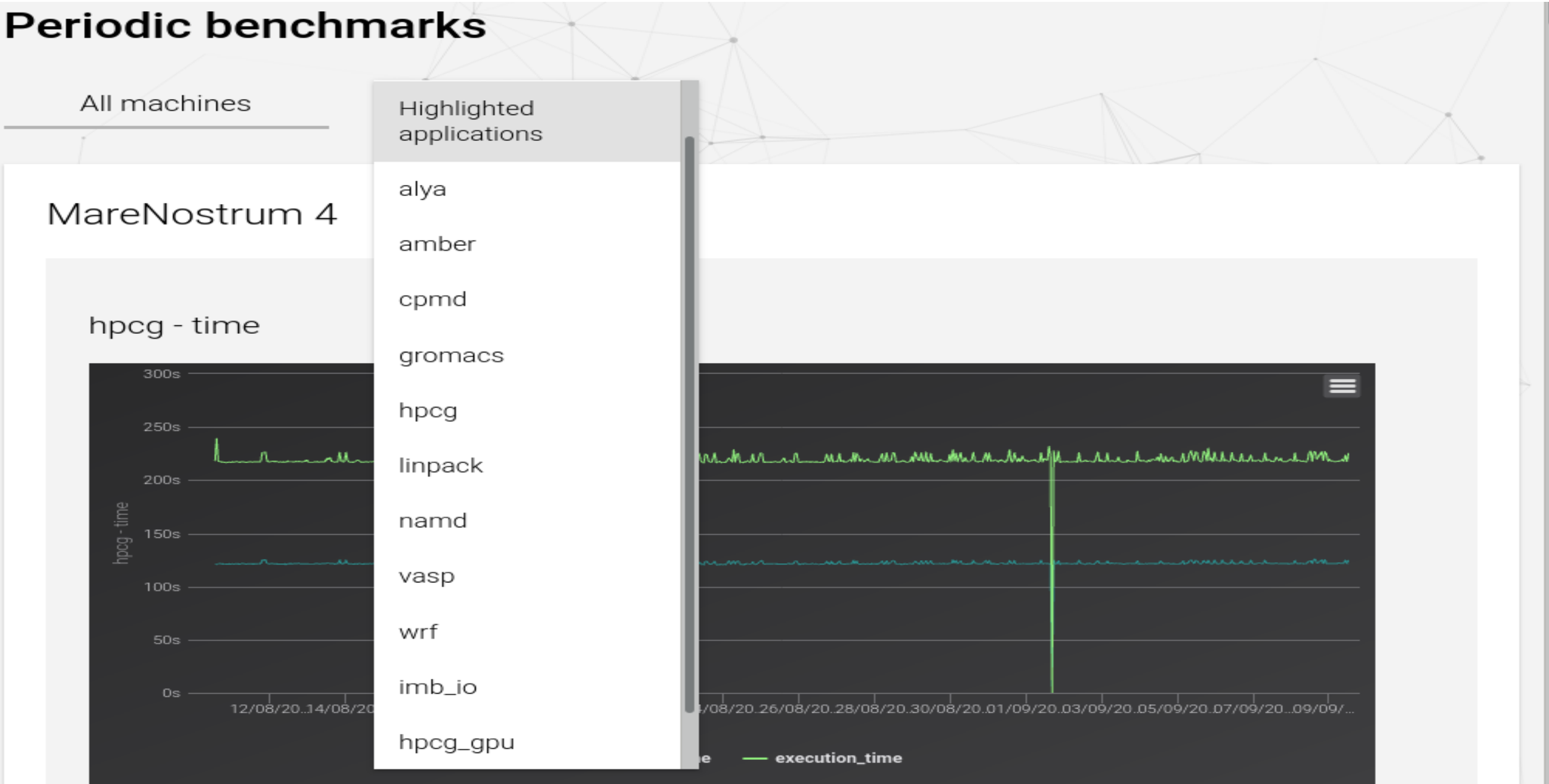


Maintenance

● BSCCV ● CTE-KNL ● CTE-Power 9 ● MareNostrum 4 ● MinoTauro ● Nord 3 ● StarLife ● Storage

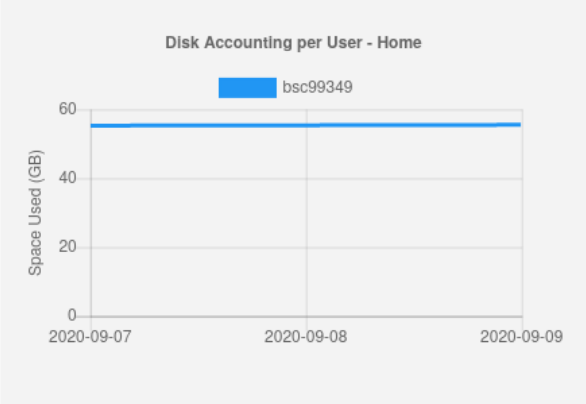
Topic	Type	Initial	Final	Duration	Machines
BSCCV, Unexpected power issue	Generic	06/09/2020 09:00	07/09/2020 09:10	1 day, 10 minutes	BSCCV
Unexpected power issues affecting MN4	Power	19/08/2020 12:55	19/08/2020 17:30	4 hours, 35 minutes	MareNostrum 4
Unexpected power issues affecting CTE-Power9	Power	19/08/2020 12:55	19/08/2020 17:30	4 hours, 35 minutes	CTE-Power 9
Unexpected power issues affecting CTE-KNL	Power	19/08/2020 12:55	19/08/2020 17:30	4 hours, 35 minutes	CTE-KNL

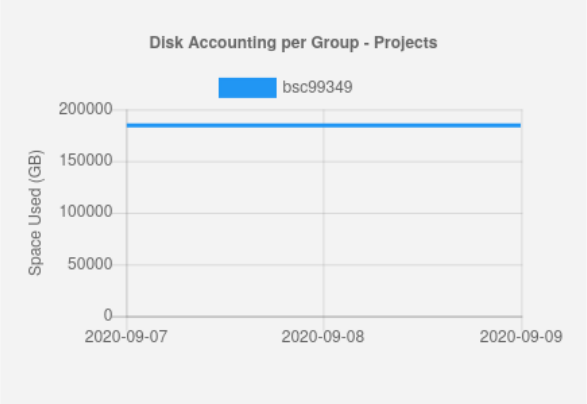
Periodic Benchmarks

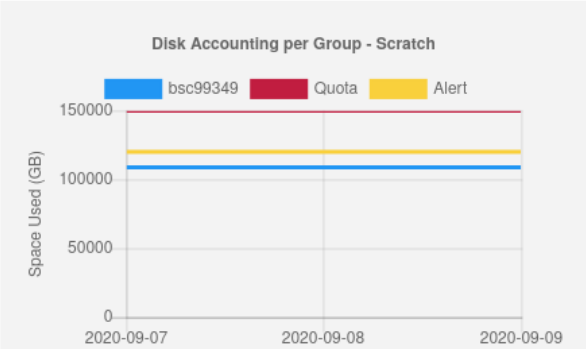


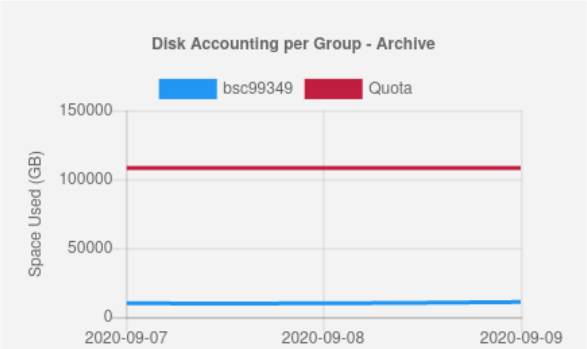
Alarms for Disk and CPU usage

- bsc99349 | 2020-09-07 | End date | SEARCH









Type of alarm	Storage partition	
Disk	Scratch	
Trigger value		EDIT | DELETE
80%		

ADD NEW ALARM

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Power Monitoring

Power monitoring

All machines

MareNostrum 4

Top 200 nodes by power consumption (last hour)

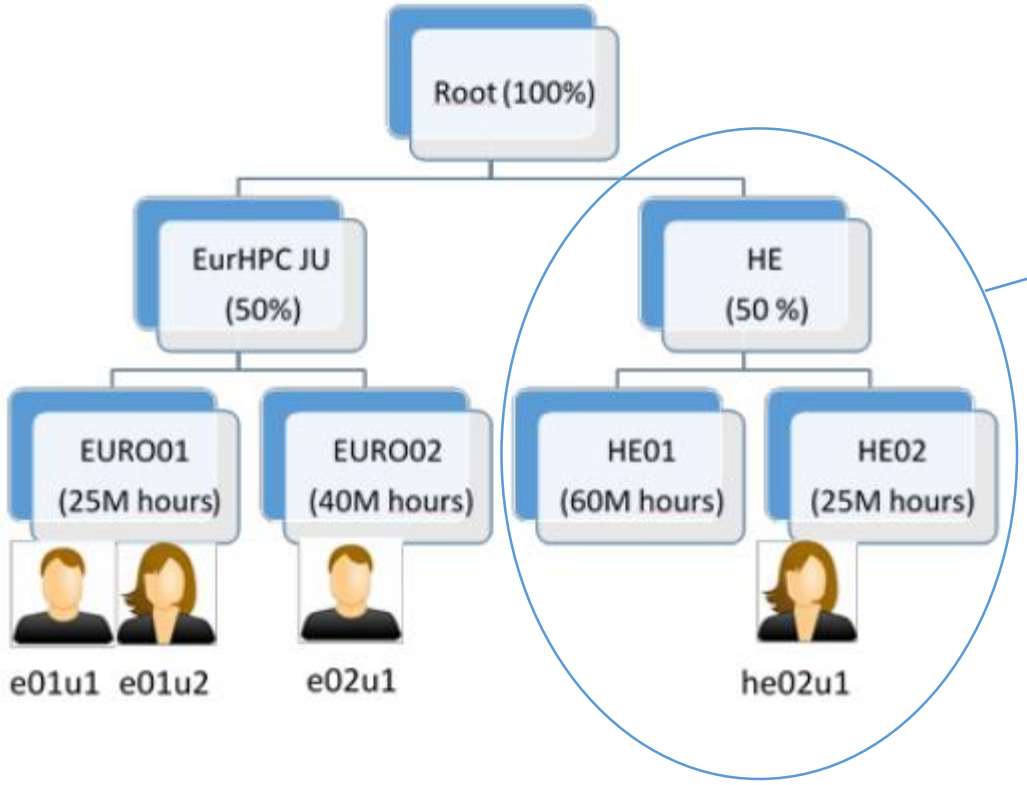
Cluster	309.84	70.76	75	546	Total mean: 0.97 MW
Hostname	Mean PW	Std Dev PW	Min PW	Max PW	Job ID
s03r2b07	526.60	8.81	508	546	**7754746**
s05r1b06	516.78	8.81	501	536	**7754746**
s05r1b12	516.17	9.13	500	536	**7754746**
s05r1b05	515.69	8.36	501	535	**7754746**
s05r1b15	512.78	8.83	496	532	**7754746**
s03r2b02	512.07	9.09	494	533	**7754746**
s03r2b05	510.98	8.65	493	529	**7754746**
s03r2b24	509.22	8.90	491	529	**7754746**
s05r1b21	507.35	8.72	493	526	**7754746**
s05r1b09	506.72	8.32	492	526	**7754746**
s05r1b13	504.92	8.54	490	525	**7754746**

Distribution of hours, how do you want to manage your users and hours ?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Hour distribution in MN5

- To ensure the proper distribution of the machine we allocate only 80% of the total computing hours, leaving the rest for draining nodes, node errors, etc.
- At batch system level we use Fair-Sharing to ensure the proper hours distribution :



Example with only 2 levels, possible to split it in an extra level to define Hosting entity share between countries

Group/User Creation information

The information required for a group creation	The information required for a user creation

• Leader of the group
  • Full Name
  • E-mail
  • Telephone
  • Institution
  • Nationality
• Name of the group
• Disk Quota
	• Full name
	• Group
	• E-mail
	• Telephone
	• Institution
	• Nationality

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thanks!

david.vicente@bsc.es