CINECA

HPCMD performance monitoring tool @ MARCONI cluster

Susana N. Bueno Mínguez s.buenominguez@cineca.it

HPC User Support and Production team

June 2023

HPCMD on Marconi cluster overview

CINECA

- HPCMD or HPC monitoring daemon is a software **tool designed to measure performance data of running jobs on HPC compute nodes**, to compute derived metrics, and to write the results
- Computes the **job performance in GFLOPS**;
- Supports performance metrics from OPA network and GPFS file systems, to obtain network and disk I/O bandwidths
- Integrates with SLURM scheduler, allowing the SLURM job detection and enabling the correlation of performance metrics with each job and to gather also other information as the jobid, the requested number of nodes, threads, etc.
- It also computes derived metrics and **writes the data to syslog lines**, that can be collected via rsyslog and finally stored in a database for subsequent analysis and visualization.

HPCMD on Marconi cluster overview

CINECA

- **hpcmd daemon** is installed as a **systemd service on all Marconi-SKL compute nodes on the skl_fua_prod partition**
- Performs **measurements** of several command line tools**, at regular and synchronized intervals**, based on a **system-wide configuration**

HPCMD on Marconi cluster overview

Command line tools used to query performance data

CINECA

perf

- query the Performance Monitor Unit (PMU) events, core counters of processors and also software events counted by the Linux kernel
- data aggregated by socket

ps

- calculates statistics about the threads running on different cores
- information about RSS memory

numastat

- query the amount of memory used per socket

ip

- collect information about generic IP network traffic

HPCMD on Marconi cluster overview

Other used tools

CINECA

HPCMD **integrates with SLURM** to complement performance data with job information:

- **scontrol, squeue, sacct** commands
- SLURM **configuration files**

opainfo

- to query OmniPath (OPA) network metrics;
- per-node data

mmpmon

- to query GPFS file systems metrics
- per-node data

HPCMD on Marconi cluster overview

CINECA

Interference with other performance tools

- It can be suspended by the user for the duration of the job to run some special category of jobs e.g. those using tools as **Intel VTUNE, Intel Advisor, PAPI, perf...** as the hpcmd tool continuously queries hardware counters through the linux perf tool and those cannot be simultaneously accessed by a second tool.

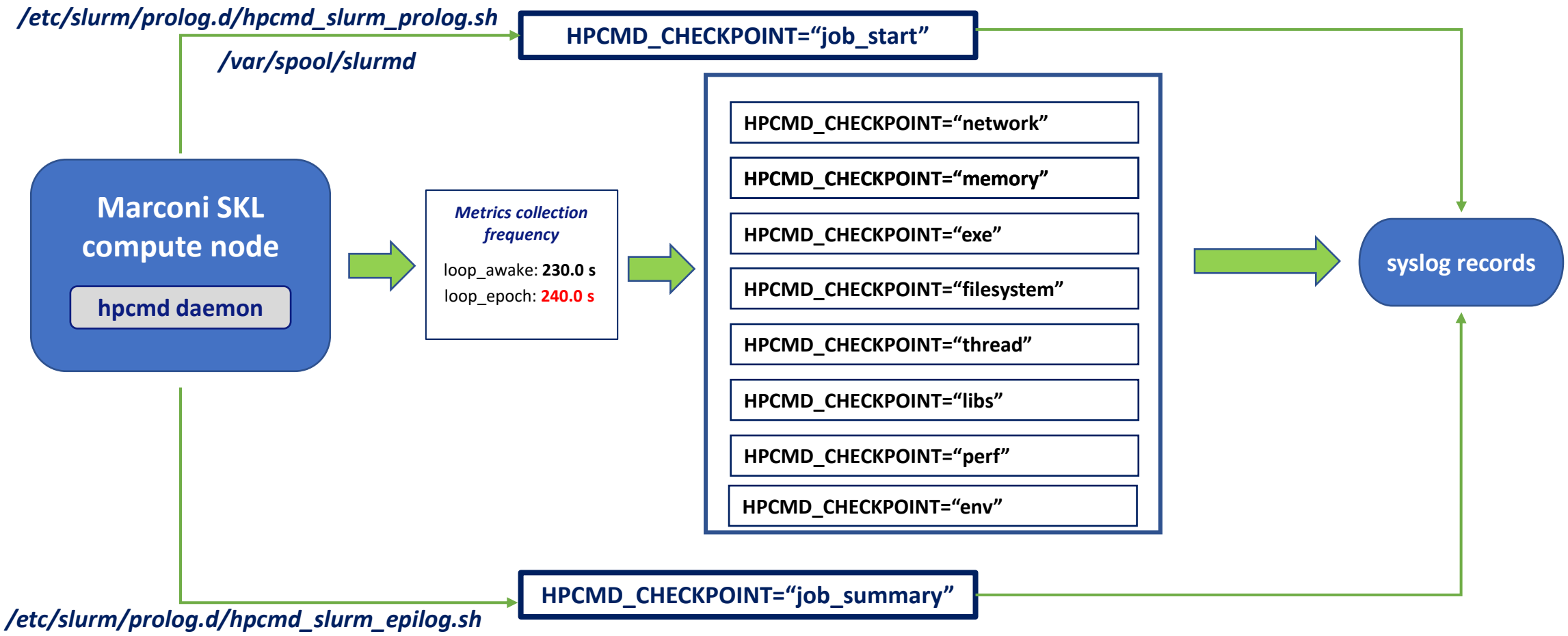
- To suspend the service insert "hpcmd_suspend" after "srun" in the batch job script before executing the application:

srun hpcmd_suspend <your_exe>

- Once the job has finished, the hpcmd systemd service will be automatically enabled for subsequent jobs so no action is needed by the user.

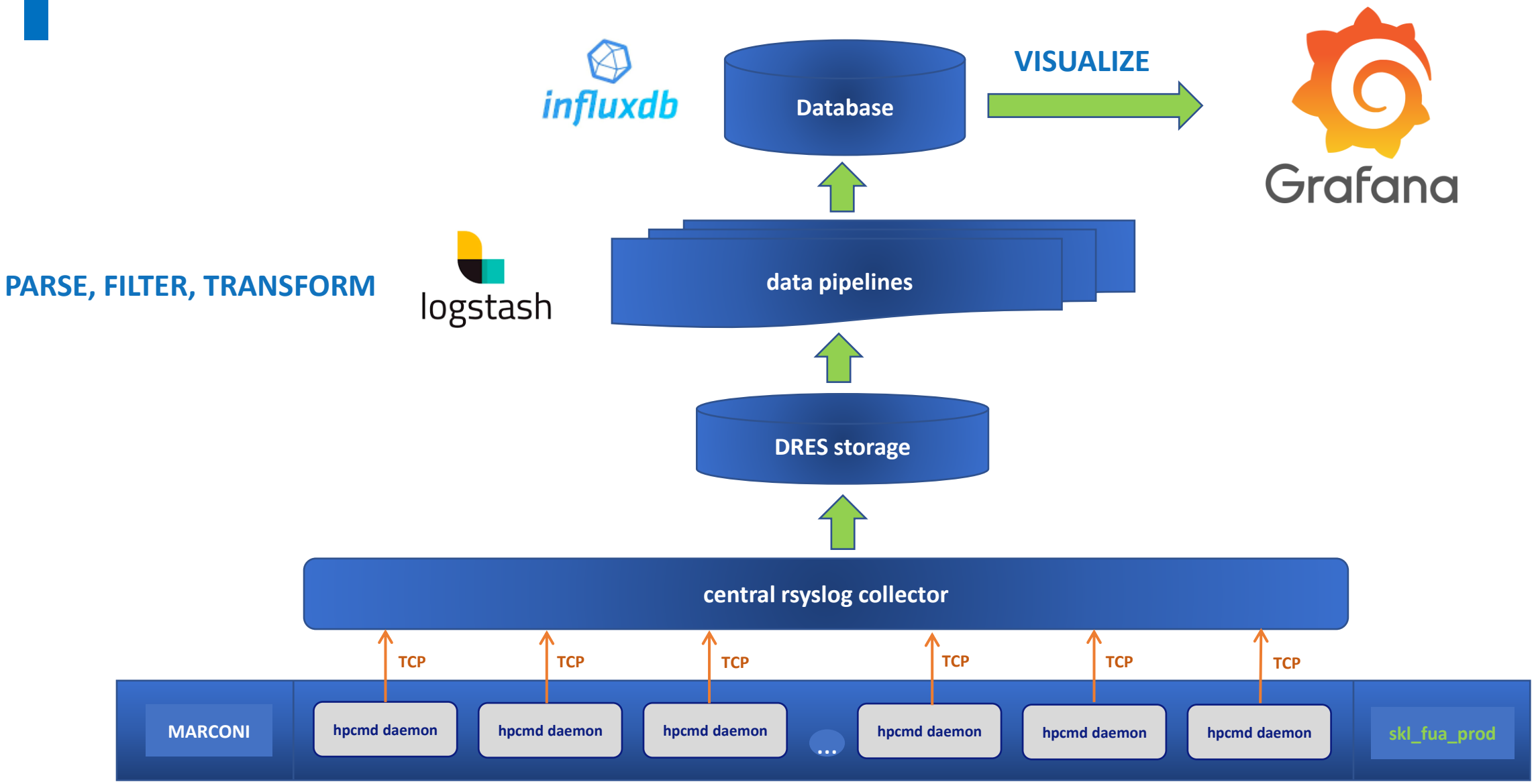
HPCMD data generation

CINECA



HPCMD data collection and transportation

CINECA



HPCMD data visualization

CINECA

- The **web interface** **[pre-production]** that allow users to consult and visualize data collected for their executed jobs on Marconi cluster can be reached at the following address:

  https://hpcmd.hpc.cineca.it

- This is **based on Grafana**, an *open source analytics & monitoring solution (www.grafana.com).*

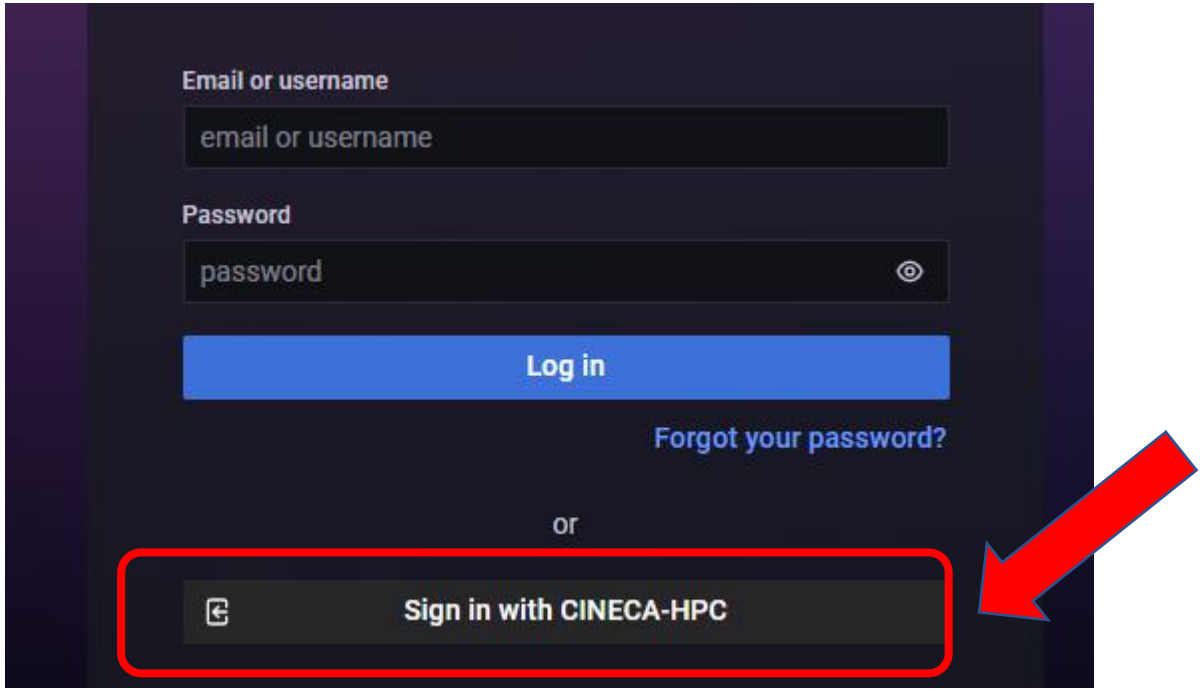
- All **users with active projects on Marconi cluster** can request to be allowed to get access by the association to **FUSIO_hpcmd_ud** project by writing to superc@cineca.it.

- **Please be aware that 2FA is enabled and, if not done yet, you will need to activate it and configure the OTP**: How to activate the 2FA and configure the OTP

HPCMD data visualization

CINECA

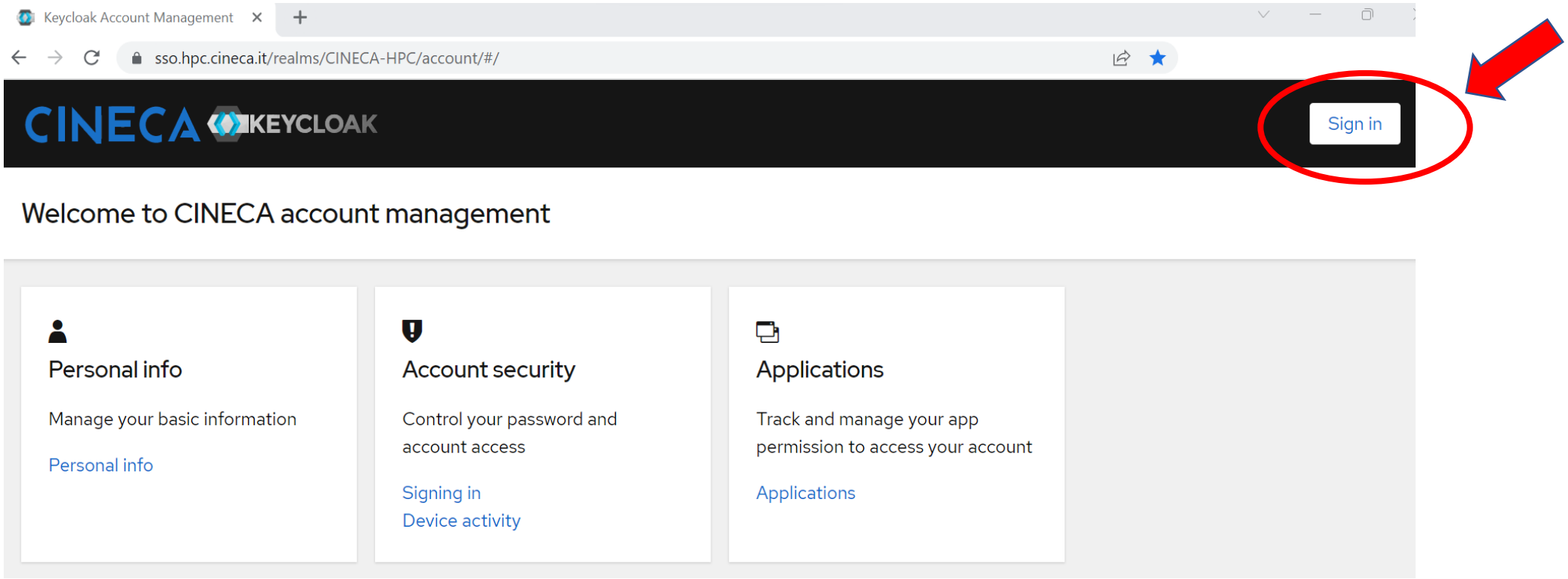
- Once the association to the project will be effective and the 2FA will be active you will be able to login to the site by following the **"Sign in with CINECA-HPC"** button and **using your HPC credentials (the same used to login to Marconi cluster)**:



How to activate the 2FA authentication and configure the OTP

Authenticate on our **new Identity Provider** at: https://sso.hpc.cineca.it

using username and password you use to connect to CINECA clusters



How to activate the 2FA authentication and configure the OTP

<u>At the first login you will be forced to</u>:

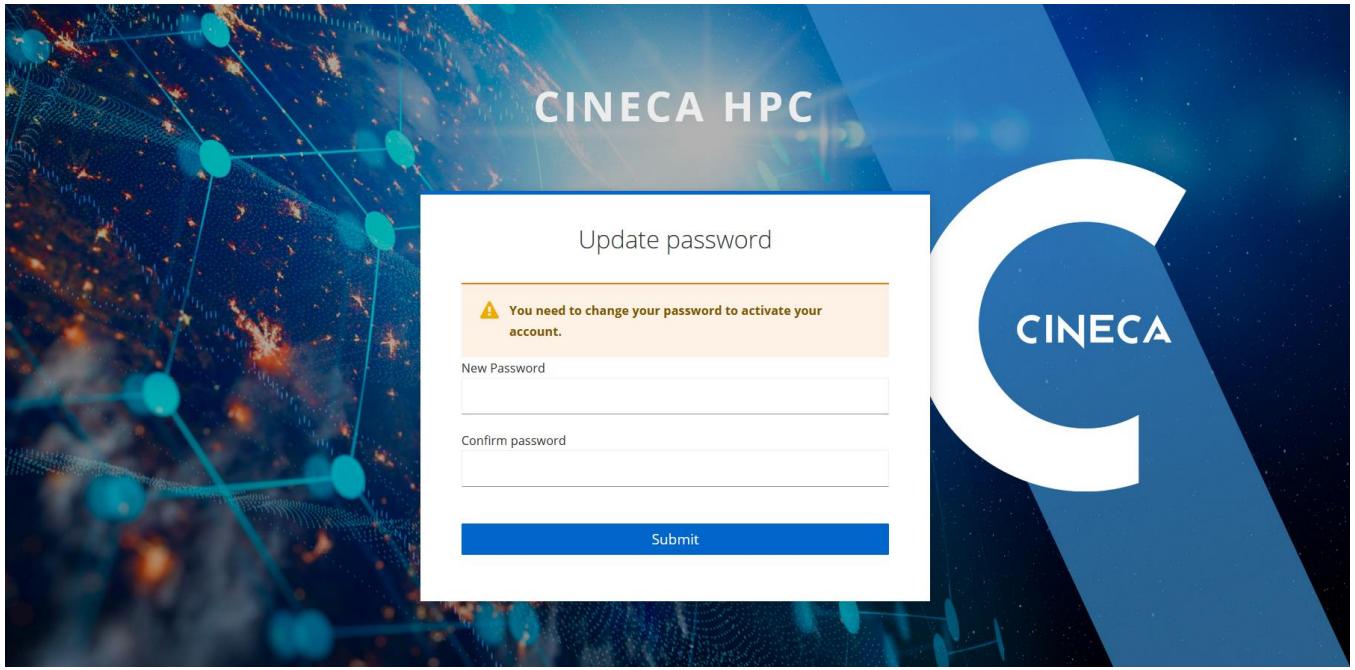
- ❑ **verify your email**
- ❑ change the password
- ❑ configure your One-Time Password (OTP) code

An e-mail containing a link will be sent to the e-mail address indicated into the UserDB site:

Subject **"CINECA HPC Single Sign On: verify your email"**

How to activate the 2FA authentication and configure the OTP

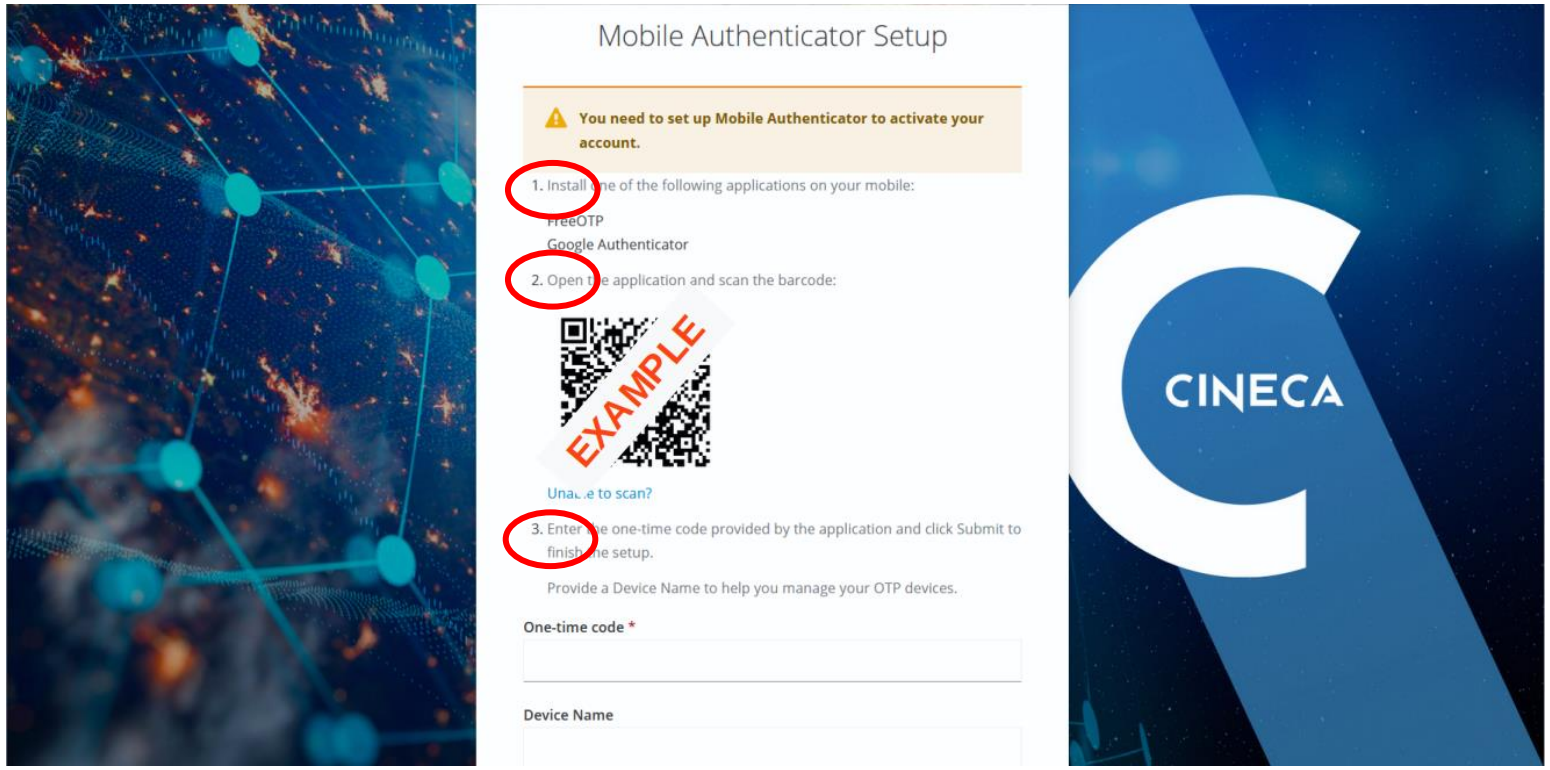
- Following the link received in the e-mail you will be forced to **change the password:**



- **The new defined password will replace the password used to login to CINECA cluster**

How to activate the 2FA authentication and configure the OTP

- Next step after the definition of the new password is the **configuration of the 2FA** following these simple steps:



How to activate the 2FA authentication and configure the OTP

First step: install on your mobile an **App to generate authentication codes:**

- **FreeOTP**
- **Google Authenticator**
- other

If problems in configuring the 2FA on your smartphone contact us at: **superc@cineca.it**

Second step: once installed, you can use your authenticator to **scan the QR** code shown in the page.

The OTP will be automatically configured on your authenticator.

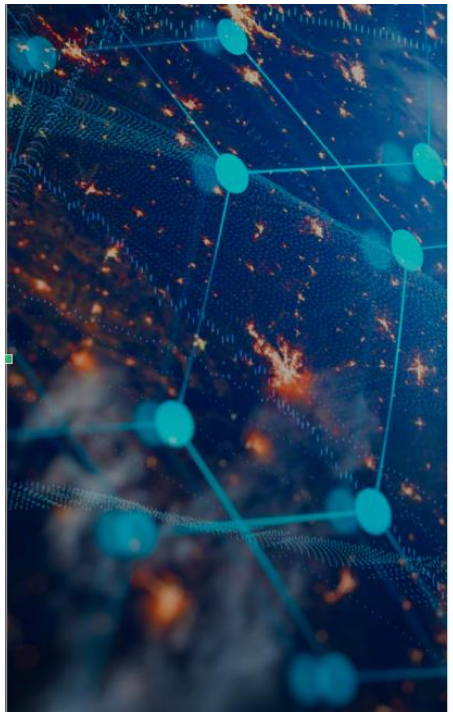
Third step: you will be asked to insert the 6 digits code that appears on the App to **verify the correct configuration**. If you have multiple OTP defined in the App, the correct one has the name "**CINECA HPC: <your username>**".

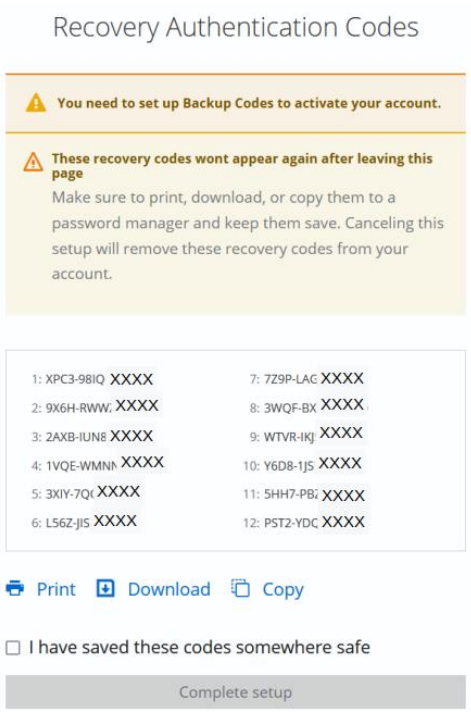
How to activate the 2FA authentication and configure the OTP

- Once verified the correct configuration the following page will show you the **Recovery codes**.

  Please **save these codes somewhere** by downloading, printing or copying in a text file

These codes are requested to the user in case of problems in the OTP configuration (issue with the app or smartphone lost) so they are very important.



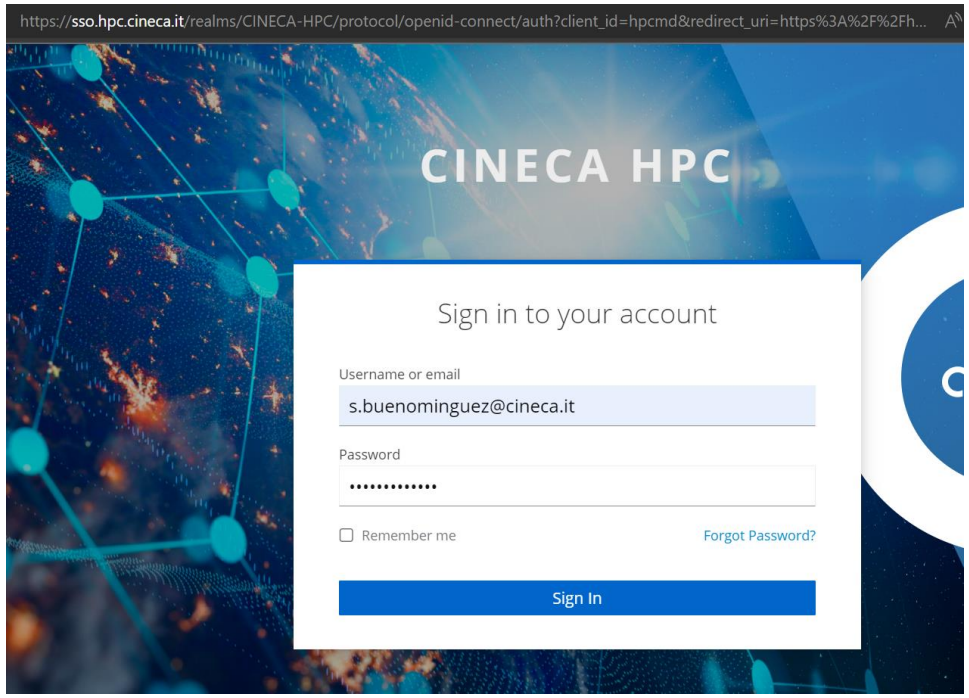


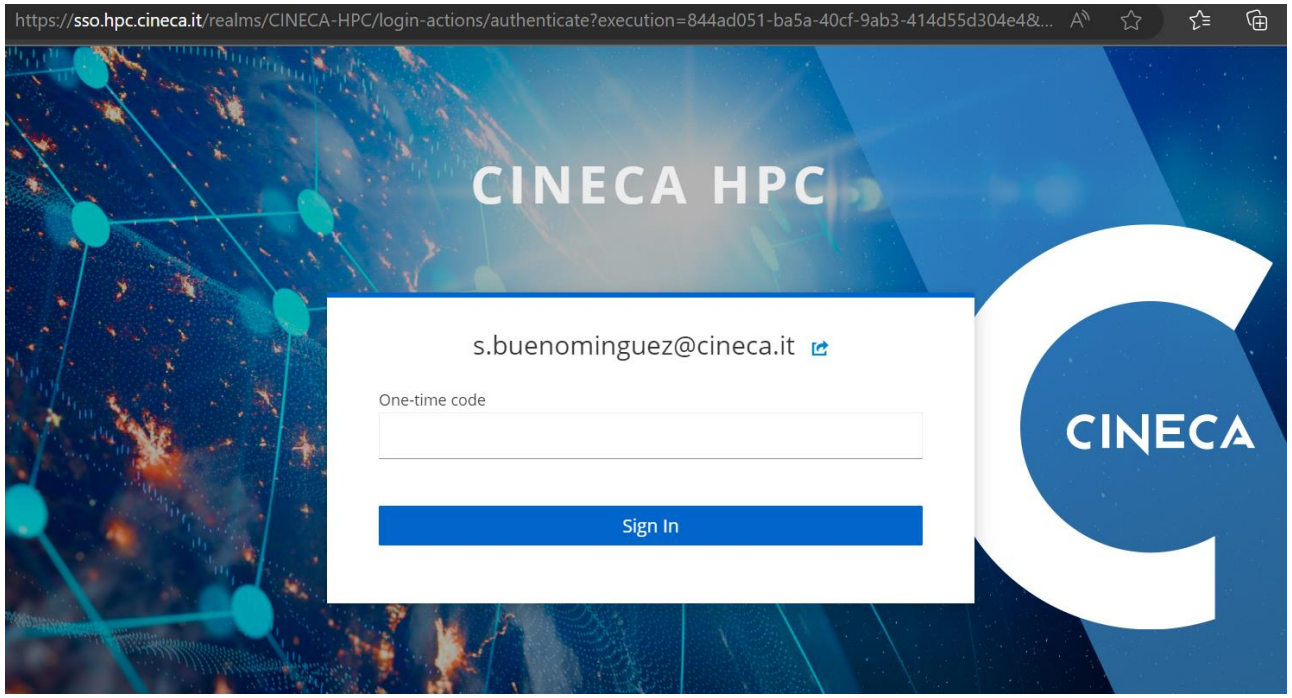
- **Now 2FA and OTP are enabled and configured.**

HPCMD data visualization

CINECA

- Once the <u>association to the project will be effective and the 2FA will be active</u> you will be able to login to the site by following the **"Sign in with CINECA-HPC"** button and **using your HPC credentials (the same used to login to Marconi cluster)**:





HPCMD data visualization

Home page

General / home_page

IMPORTANT NOTE

This user interface is **under construction**: metrics, dashboards, panels or functionalities may be changed or included in the future

Welcome to the USER INTERFACE

Tracked jobs

Complete data is available for jobs that:

- have started **since May 16th** - and -
- have been executed on the **skl_fua_prod** partition - and -
- ended with an **elapsed time of at least 12 minutes** - and -
- have passed at least **4 hours** since the end of job

Partial data available **after 4 hours (max.)** since its generation as syslog record

Missing jobs

- some jobs might have been excluded from the user job's list if any anomaly was detected in the recorded data

USER DASHBOARDS

Raw Data collected

Job's raw data & User activity

Command line tools info

Collected metrics visualization for ended jobs

Collected Metrics Viz

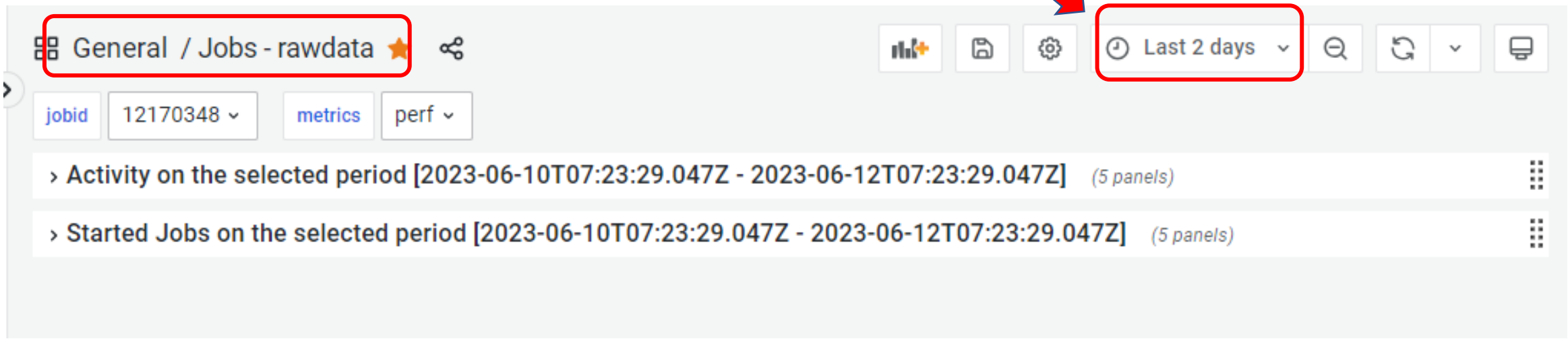
Executed applications (ended jobs)

Applications

HPCMD data visualization

User activity & Raw data dashboard

Select a time interval that comprises the start time of your job of interest



HPCMD data visualization

User activity & Raw data dashboard

Activity on the selected period [2023-06-10T20:36:38.265Z - 2023-06-11T20:36:38.265Z]

Started Jobs

3

Started Jobs

Time	jobid	jobname	nnodes	cores	ntasks	ntasks_per_node	sockets
2023-06-11 04:04:00	1216		48	48	48	0	2
2023-06-11 08:26:15	1216		48	48	48	0	2
2023-06-11 04:32:00	1216		64	48	72	0	2

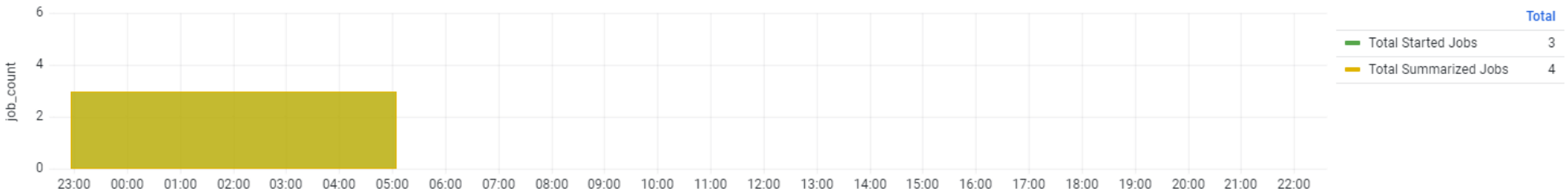
Ended Jobs

4

Ended Jobs

Time	jobid	exe	exit_code	nnodes	jobstart	jobend
2023-06-10 23:34:29	121		0:15	48	2023-06-10T08:34:13.000Z	2023-06-10T23:34:25.000Z
2023-06-11 05:27:44	121		1:0	64	2023-06-10T07:06:10.000Z	2023-06-11T05:27:43.000Z
2023-06-11 16:03:25	121		0:15	48	2023-06-11T04:03:10.000Z	2023-06-11T16:03:20.000Z
2023-06-11 16:46:33	121		0:15	48	2023-06-10T22:16:07.000Z	2023-06-11T16:46:26.000Z

Started jobs for user



HPCMD data visualization

User activity & Raw data dashboard

Available metrics: perf, GPFS, network, memory, exe

jobid 1215 metrics perf

legend

jobid: list of jobs that have a start time in the selected time period --- metrics: list of the metrics collected that are available for each job

› Activity on the selected period [2023-06-04T20:46:34.651Z - 2023-06-11T20:46:34.651Z] (5 panels)

⌄ Started Jobs on the selected period [2023-06-04T20:46:34.651Z - 2023-06-11T20:46:34.651Z]

Job start details for job 12153556 (username

Time	awake	cores	cpus_per_task	epoch	jobid	jobname	jobstart	loadedmodules	mhost	nnodes	nodeid	ntasks	ntasks_per_noc	opmode	sockets	userid
2023-06-08 16:44:08	230	48	null	240	1215		1686235211	profile/base:in...	r130c03s01	192	0	0	0	systemd	2	

HPCMD "perf" metric's raw values for job

Time	BR-MISS-RATIO	CACHE-MISS-RATI	FP-SCALAR	FP-VECTOR	GFLOPS	IPC	branch-misses	branches	cache-misses	cache-references	cpu	cycles	fp_128d	fp_128
2023-06-08 16:47:50	0.0146	0.959	2773035991145	0	12.1	1.60	34093733159	2338915208040	81432615190	84952842164	S0	11375475244897	0	0
2023-06-08 16:47:50	0.00698	0.908	1563636354730	0	6.80	1.48	19627365466	2811124741776	49118538915	54117549862	S0	11382505917842	0	0
2023-06-08 16:47:50	0.00408	0.830	951144550267	0	4.14	1.42	12408591070	3044231598149	31940577622	38470910737	S0	11379277840454	0	0
2023-06-08 16:47:50	0.00223	0.763	494675447019	0	2.15	1.37	7212864152	3238804370834	18444095744	24171305187	S0	11403919476280	0	0
2023-06-08 16:47:50	0.00152	0.600	292870110829	0	1.27	1.36	5063396952	3334637567475	11679010466	19456546176	S0	11395002612463	0	0
2023-06-08 16:47:50	0.00163	0.648	363287633939	0	1.58	1.37	5419256417	3316882362947	13034081528	20124117786	S0	11402837133863	0	0
2023-06-08 16:47:50	0.00121	0.527	215008951715	0	0.935	1.36	4144498692	3412026342779	8252993378	15659005858	S1	11405610357353	0	0

This dashboard allows also the visualization of raw data for jobs that are still on RUNNING state - partial data - in this case «job summary» information will not be available yet and «no data» label will be shown

HPCMD data visualization

Command line tools info

General / Command line tools info

› Active commands in the current system configuration (7 panels)

› Inactive commands in the current system configuration (1 panel)

Active commands in the current system configuration

Performance analysis

- "perf" Performance analysis tools for Linux
- Executed command:

```
$ perf stat -x , -a -e r5302c7,cache-references,r5380c7,cycles,r5340c7,r5308c7,cache-misses,branches,r5304c7,branch-misses,r5320c7,major-faults,r5301c7,instructions,r5310c7,minor-faults --per-socket sleep 230.0
```

'perf stat' command counts events specified with -e arguments

```
'fp_arith_inst_retired.scalar_double': 'fp_d','r5301c7': 'fp_d',
'fp_arith_inst_retired.scalar_single': 'fp_s','r5302c7': 'fp_s',
'fp_arith_inst_retired.128b_packed_double': 'fp_128d','r5304c7': 'fp_128d',
'fp_arith_inst_retired.128b_packed_single': 'fp_128s','r5308c7': 'fp_128s',
'fp_arith_inst_retired.256b_packed_double': 'fp_256d','r5310c7': 'fp_256d',
'fp_arith_inst_retired.256b_packed_single': 'fp_256s','r5320c7': 'fp_256s',
'fp_arith_inst_retired.512b_packed_double': 'fp_512d','r5340c7': 'fp_512d',
'fp_arith_inst_retired.512b_packed_single': 'fp_512s','r5380c7': 'fp_512s'
```

- Related plots reported on the perf dashboard

Filesystem

- collect I/O statistics per mounted filesystem from the point of view of GPFS servicing application I/O requests
- Executed command:

```
$ mmpmon -p -i <mounted_filesystem>
<mounted_filesystem> home, work, scratch
```

HPCMD data visualization

Collected metrics: exe dashboard

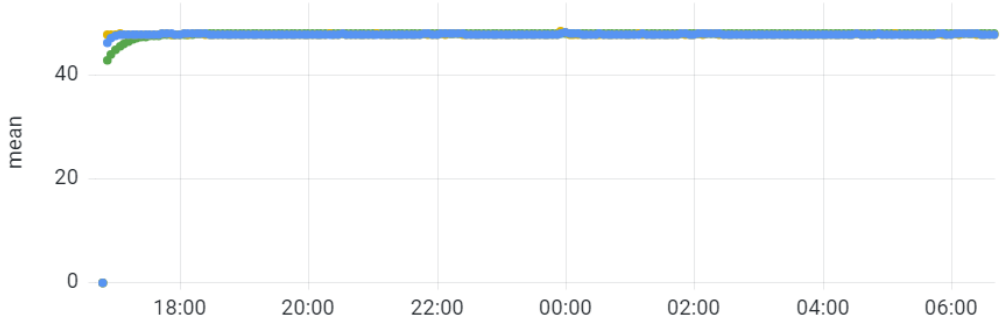
metrics / Executable Info

2023-06-08 16:40:05 to 2023-06-09 06:40:18

Jobid = ''

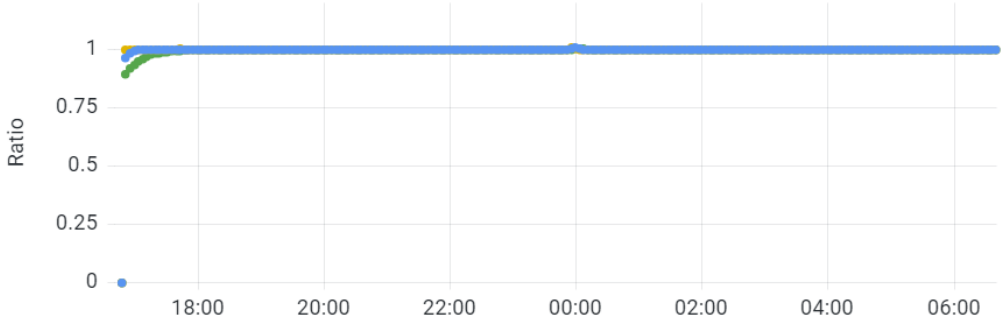
partition	exit_code	jobstart	jobend	timelimit	elapsed (h)	nnodes	groupid	userid
skl_fua_prod	0:15	2023-06-08 18:40:05	2023-06-09 08:40:18	14:00:00	14.0	192	interactive(25200)	

Linux Load Averages



	Min	Max	Mean	Last *
1min	0	48.2	47.7	48.1
5min	0	48.6	47.8	48.0
15min	0	48.4	47.8	48.0

Linux Load Averages - Ratio



	Min	Max	Mean	Last *
1min	0	1.00	0.994	1.00
5min	0	1.01	0.996	1.00
15min	0	1.01	0.996	1.00

Total threads

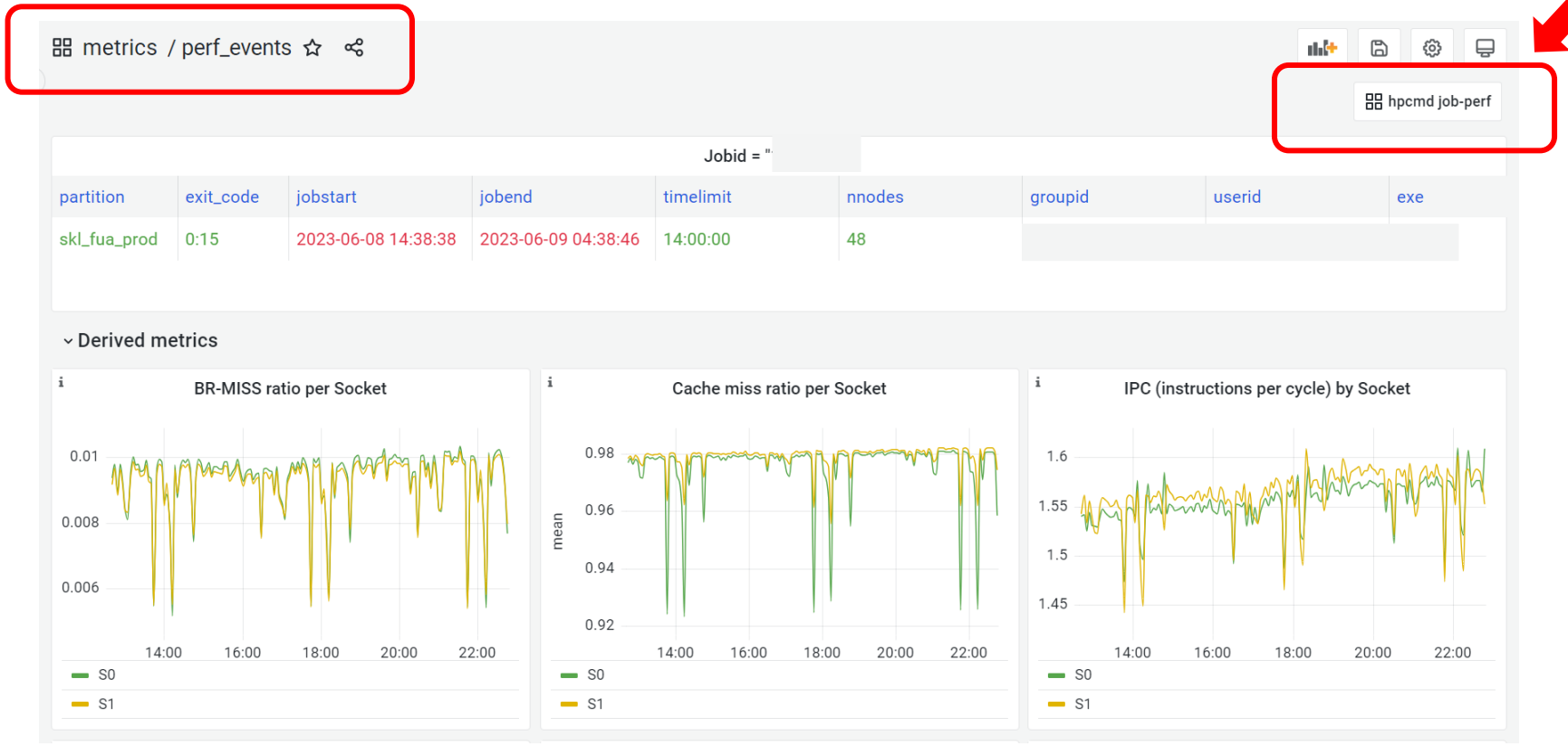
Empty cores for job

100

HPCMD data visualization

Collected metrics: perf events dashboard

Follow this link to open a new dashboard reporting **GFLOPS** measured for current job



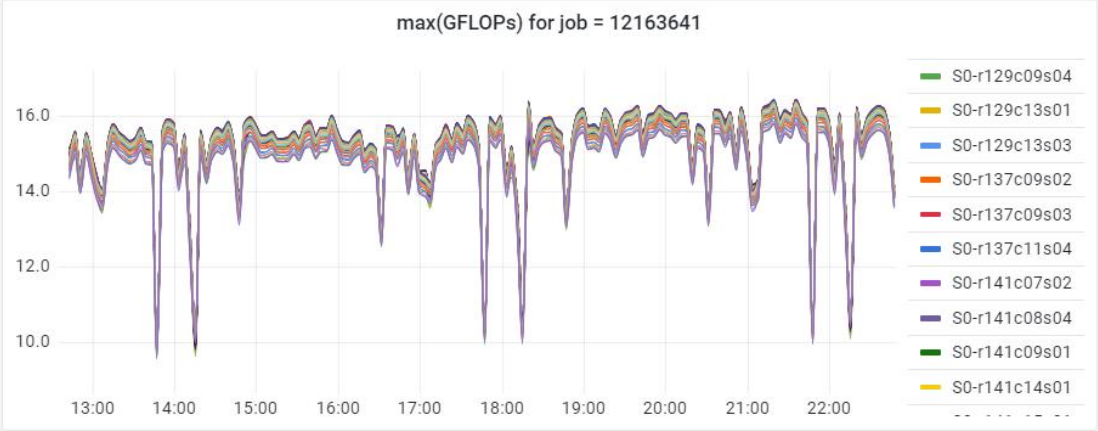
HPCMD data visualization

Collected metrics: job performance dashboard

metrics / hpcmd_job_perf

Jobid = ' '

partition	exit_code	jobstart	jobend	timelimit	nnodes	groupid	userid	exe
skl_fua_prod	0:15	2023-06-08 14:38:38	2023-06-09 04:38:46	14:00:00	48	interactive(25200)		BIT1



Performance on every socket for job = 12163641

max(GFLOPS)	avg(GFLOPS)	Socket	hostname
14.6	14.6	S0	r129c09s04
14.4	14.4	S1	r183c09s04
14.6	14.6	S0	r129c13s01

HPCMD data visualization

Collected metrics: gpfs dashboard

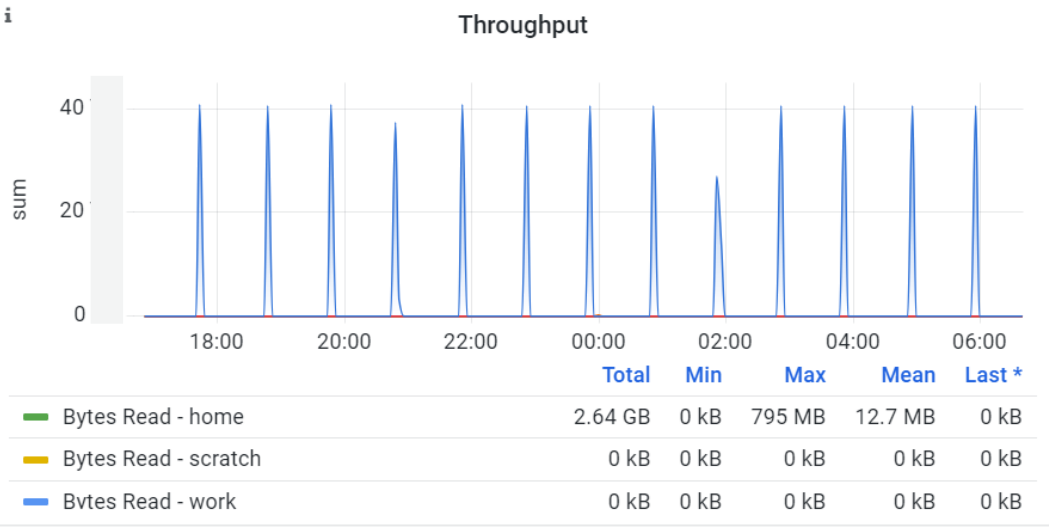
metrics / gpfs

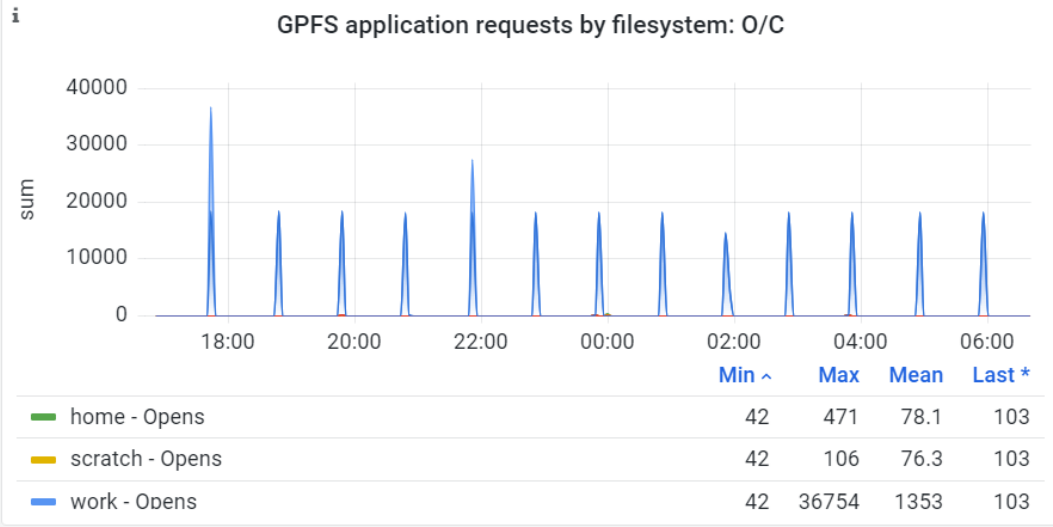
2023-06-08 16:40:05 to 2023-06-09 06:40:18

Jobid = "1

partition	exit_code	jobstart	jobend	timelimit	nnodes	groupid	userid	exe
skl_fua_prod		2023-06-08 18:40:05	2023-06-09 08:40:18	14:00:00		interactive(25200)		

GPFS activity for Job =





GPFS application requests by filesystem: R/W

GPFS inodes updates by filesystem

HPCMD data visualization

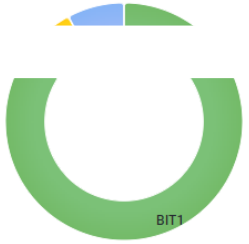
Applications: metrics summary for ended jobs

General / Jobs - applications

Last 2 days | 5s

exe

Number of Executed applications in the selected period (summarized jobs)



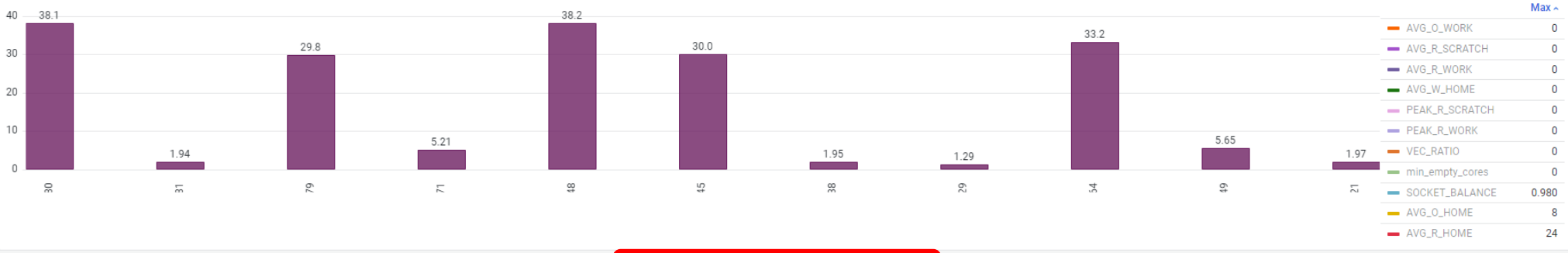
Value
11
1
1

Elapsed time for application in the selected period (ended jobs)

exe	jobid	node-h	core-h	elapsed-h	Nodes
		2689	129083	14.0	192
		1536	73731	24.0	64
		672	32260	14.0	48
		576	27654	12.0	48
		888	42636	18.5	48

1 2 3 — 1 - 5 of 11 rows

Jobs - metrics summary



Jobs - table of metrics summary

AVG_W_SCRATCH	AVG_W_WORK	FP-SCALAR	FP-VECTOR	GF	MEM_PEAK	PEAK_LOAD	PEAK_O_HOME	PEAK_O_SCRATCH	PEAK_O_WORK	PEAK_R_HOME	PEAK_R_NET	PEA
0	0	30953408687776	483	1.97	138499	48	3	1	1	2746	3329229	
260083	0	1566759882334971	0	38.2	37436	48	1	104	1	35	4880248	
64945	0	45153735193708	0	5.65	17390	72	290	872	1	431	4692378	
160773	0	1429111155013583	80	29.8	45530	48	1	196	1	52	2066569	

HPCMD on Marconi cluster documentation

- Documentation available on HPC User's Guide online documentation, in the EUROfusion users dedicated section:

  - General info about this tool:

https://wiki.u-gov.it/confluence/display/SCAIUS/HPC+performance+monitoring+tool+on+Marconi+cluster

  - Dedicated section to data management and visualization:

https://wiki.u-gov.it/confluence/display/SCAIUS/HPCMD+Data+management+and+visualization

CINECA

www.cineca.it