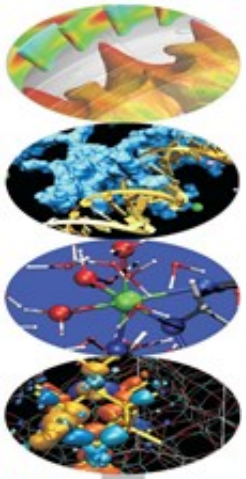
CINECA SCAI
SuperComputing Applications and Innovation

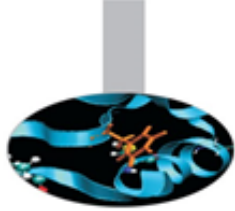


55th Ticket Meeting

HPC User Support @ CINECA
July, 15th 2021

CINECA

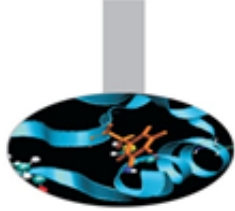
Content



- Status of the clusters - main events affecting production [June 16 – July 14]
- Examination of active tickets on HPC-US-SECOND queue
  - escalated to SchedMD support
  - escalated to Intel support
  - escalated to NVIDIA support
  - other active tickets
- Ticket statistics on queue [June 14 – July 12]
  - HPC-US-FIRST
  - HPC-US-SECOND
- Module usage on Marconi-SKL

Status of the clusters June 16 – July 14

Main events affecting production



June 17: quota imposed on $CINECA_SCRATCH area on Marconi100 cluster
https://www.hpc.cineca.it/center_news/marconi100-scratch-almost-full-quota-imposed
https://www.hpc.cineca.it/center_news/marconi100-scratch-quota-removed

June 24: scheduled maintenance operations on Marconi100 cluster on June 29th
https://www.hpc.cineca.it/center_news/scheduled-maintenance-marconi100-june-29th
https://www.hpc.cineca.it/center_news/reminder-scheduled-maintenance-marconi100-tomorrow-june-29th
https://www.hpc.cineca.it/center_news/marconi100-back-production-9

July 7: job level GPU usage and accounting report available on Marconi100 cluster
https://www.hpc.cineca.it/center_news/marconi100-job-level-gpu-usage-and-accounting-now-available

July 13: interruption of production on dedicated EUROfusion nodes on Marconi cluster due to power outage
https://www.hpc.cineca.it/center_news/eurofusion-problems-marconi-skl-compute-nodes

Tickets escalated to **SchedMD** support

Ticket	Subject	Requestors	Created	Host	Supp. Level	Comments
16506	Allocation of GPU jobs in M100	Leo.Ma@ukaea.uk	2021-03-17 16:43:09	M100	2nd SchedMD	The allocation of 1 GPU card job to nodes is pretty random, and not packing tightly; It may be a good idea to pack jobs with some ordering, and occupy less nodes --- After a long exchange of messages with SchedMD, we found the way to have four 1-gpu jobs packed on the same node; It is simply needed to discard the directive: #SBATCH --ntasks-per-node=1 or modify it as follows: #SBATCH --ntasks=1 If not set the ntasks-per-node value will be set to the default (1). A "bug" (or a desired "feature", the SLURM support is still investigating on the matter) is triggered by the --ntasks-per-node directive, forcing the selection of the first free cores per node, and slurm (correctly) reacts choosing a different node to enforce the cpu-gres binding. SchedMD reported: «Fixing this requires a notable change in the select plugin. I am afraid that a proper fix will be available only in slurm 21.08» slurm 21.08 will be released next August, and we will most likely install it on M100 at the following scheduled maintenance (end of August).

Tickets escalated to **SchedMD** support

Ticket	Subject	Requestors	Created	Host	Supp. Level	Comments
18895	Unclear error message on MARCONI ...	joerg.riemann@ipp.mpg.de	2021-06-24 14:54:01	SKL	2nd SchedMD	The user report randomly observed errors in his jobs as in «srun: error: Unable to create step for job 9552876: Memory required by task is not available». After some extensive checks we made, we got an additional information reported for the second mpirun call: “srun: error: Unable to create step for job 3130006: Memory required by task is not available” It appears that the first mpirun (or, better, the first srun called by the first mpirun) isn't able to free the allocated memory fast enough for give room to the second one. Hence the observed behavior is caused by a slurm bug that was addressed to the slurm support team: “Second mpirun in a job fails after Slurm upgrade”. SchedMD suggested a workaround while waiting for the bug to be fixed in 20.11.8 version. We indicated the user to put a "sleep 5" between the two mpirun commands in the jobscript.

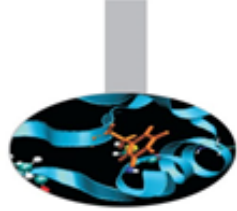
Tickets escalated to **Intel** support

Ticket	Subject	Creation date	Last Updated by Intel	Comments
3932	Pointers to module arrays not working with SIMD (nilsm@ipp.mpg.de)	18/09/2019 17:24:02	**01/07/2021**	Intel originally found a bug on the compiler front end and solved it. They provided us an archive with the sources, BUILD script and the outputs of the reproducer that did not showed the issue. The first bug had "hidden" a second bug on the vectorizer that was triggered by the definition of the macro "USE_ARR_IN_MODULE" in the reproducer. We reported this issue to Intel that confirmed the problem and opened a new bug: Intel bug report CMPLRIL0-33599 Intel support has provided a resolution for the problem reported in the second bug: “”” We cannot use simd for a loop that has F90 pointer assignment inside. For every iteration of the loop, it is updating the same dope vector for f4a. That means there is a loop carried dependency preventing vectorization. If it is vectorized with veclen =2, for iteration 1 and 2, it is storing into the same location of the dope vector and the address code of f4a is picked up incorrectly. In order to generate the right code, it can run in a non-vector mode. However, attached is fixed.f90 which shows the right way to do it. We need to declare a structure of arrays to store the F90 pointers. In that way, every iteration is storing into different dope vectors. And, of course, the result from fixed.f90 is different because I used different computations in the loop.”””

Other active tickets on HPC-US-SECOND

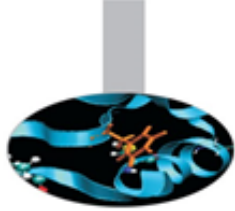
Ticket	Subject	Requestors	Created	Host	Supp. Level	Comments
16019	Marconi SKL inter socket connection	serhiy.mochalskyy@ipp.mpg.de	2021-03-01 09:10:02	SKL	2nd Intel Lenovo	Information about interconnection between two sockets of the same node has been retrieved and also published on the Users Guide. Additional information retrieved to estimate the correct value of the Inter-Socket bandwidth on SKL in terms of GB/s. Performed the osu_mbr_mr benchmark (uni-directional test) that supposes to test the inter socket bandwidth using different number of MPI pairs for simultaneous data transfer. To perform this or another benchmark (for example, the Intel IMB benchmark has similar test to check the aggregate bandwidth between two sockets), so to compare all results and the theoretical value. A report has been prepared with results obtained from different tests.

Other active tickets on HPC-US-SECOND



Ticket	Subject	Requestors	Created	Host	Supp. Level	Comments
18205	paraview on M100	mattwi@fysik.dtu.dk	2021-05-28 11:46:01	M100	Int.	The user reported issues when running paraview version available as module through a RCM session. To fix this is required to recompile the paraview version (in progress), we will include the installation of the nvidia plugin index required also by this user.
18851	Problem with the nvidia compiler and -lnvc on m100	nilsm@ipp.mpg.de	2021-06-23 11:00:02	M100	2nd	The user report a strange behaviour referring to CUDA devices detection when adding (or not) the «-lnvc» flag for the compilation of a test code available on the latest hpc-sdk compiler suite on Marconi100 cluster. Investigations are underway.
19157	Error on marconi	axel.koenies@ipp.mpg.de	2021-07-07 11:32:02	SKL	2nd	The user reported two jobs running euterpe code that wrote no output and asked for a budget refund of the lost computing resources. We performed checks for those jobs and we did not find any filesystem issue involved. We informed the user about this behaviour previously observed with euterpe code that has been fixed in other situations by recompiling the code; also, about our policy that do not foresees the budge refund due to system issues etc.

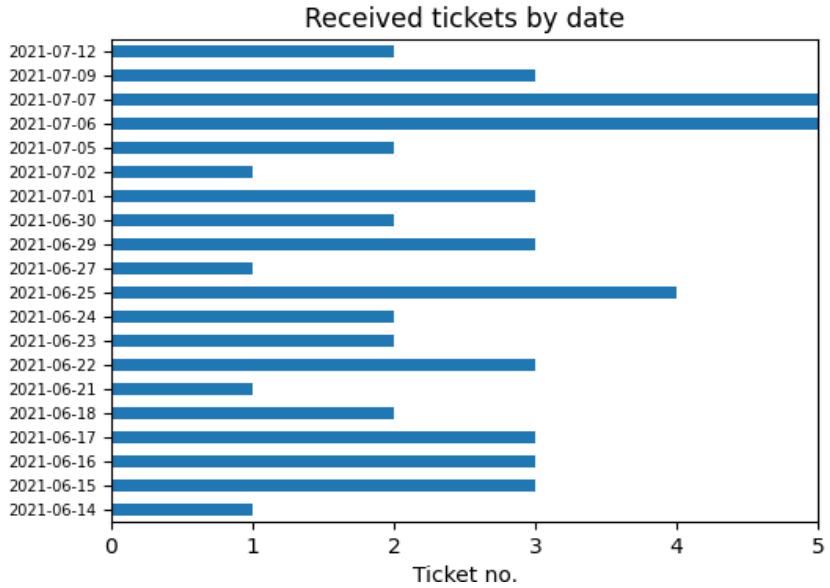
Other resolved tickets on HPC-US-SECOND

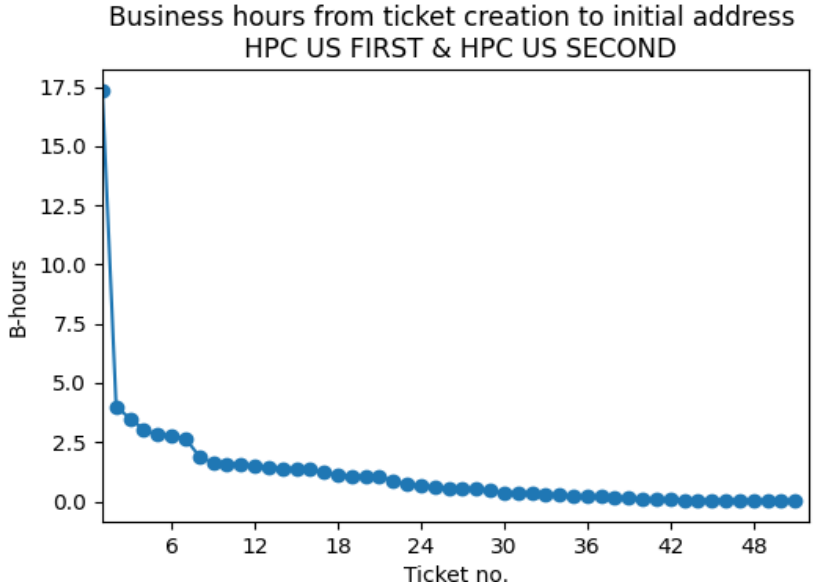


Ticket	Subject	Requestors	Created	Host	Supp. Level	Comments
18400	user guide marconi100	nicola.varini@epfl.ch	2021-06-07 10:48:06	M100	2nd	The user repors and error when following a example reported in our guide (that worked correctly before the cluster upgrade operations) due to the SLURM_LOCALID variable not being setup when using mpirun; this works correctly using srun instead. When using mpirun it is possible to use the OMPI_COMM_WORLD_RANK variable that gives the same results as SLURM_LOCALID used with srun.
18443	partition m100	nicola.varini@epfl.ch	2021-06-08 06:28:01	M100	2nd	The user has asked for the possibility of running benchmarking jobs requiring 100 nodes for a maximun walltime of 30 minutes. This user has received also the EUROfusion authorization to run this right after next scheduled maintenance operations on M100 cluster on June 29th will be completed. These benchmarks were previously suspended as the user announce to be not available this day.
18600	NVHPC 21.5 @ M100	thomas.hayward@ipp.mpg.de	2021-06-11 21:48:01	M100	2nd	The user has asked for the installation of the nvidia hpc-sdk module version 21.5. The 21.3 release fixed many of their compilation problems with the ORB5 code but as reported by the user “”” Unfortunately, there seems to have been a small regression in the atomics in 21.3, which an nvidia compiler engineer tells me is fixed in the 21.5 release.”””. The new release has been now installed on the cluster.

Ticket Statistics June 14 – July 12

HPC-US-FIRST & HPC-US-SECOND

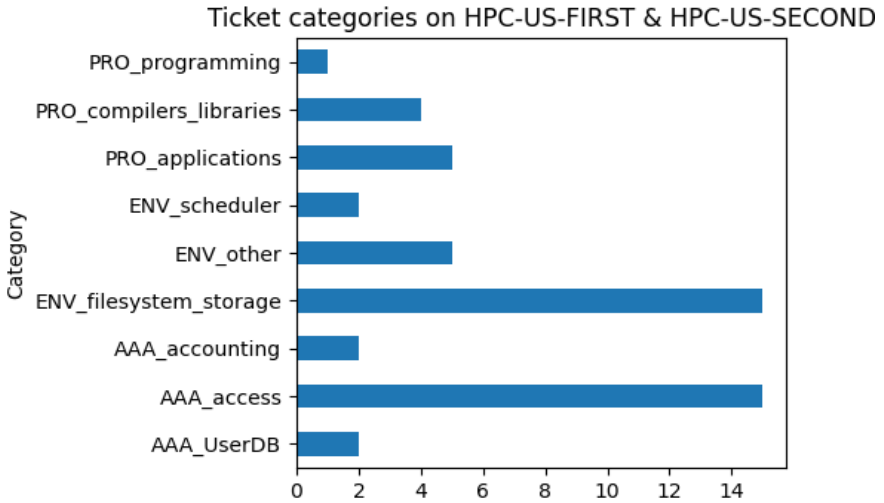


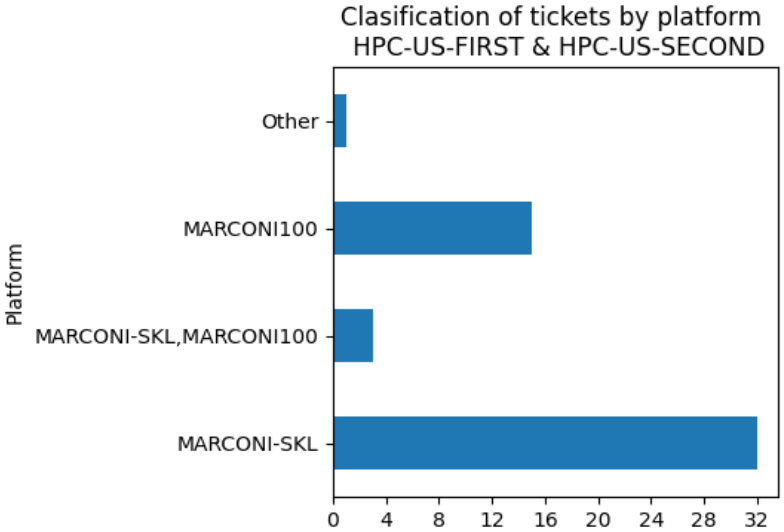


TOTAL	→ 51 tickets
- HPC-US-FIRST	→ 29
- escalated to HPC-US-SECOND	→ 22

Ticket statistics June 14 – July 12

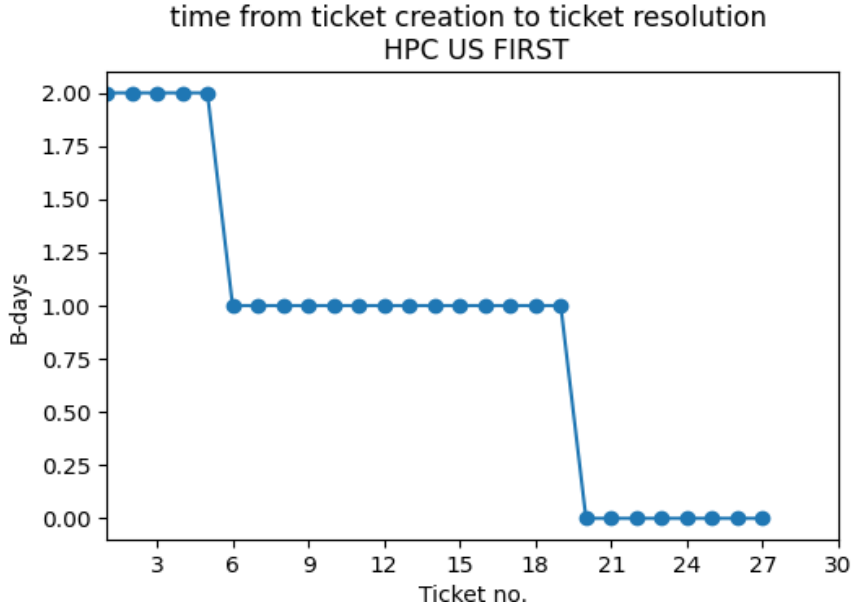
HPC-US-FIRST & HPC-US-SECOND



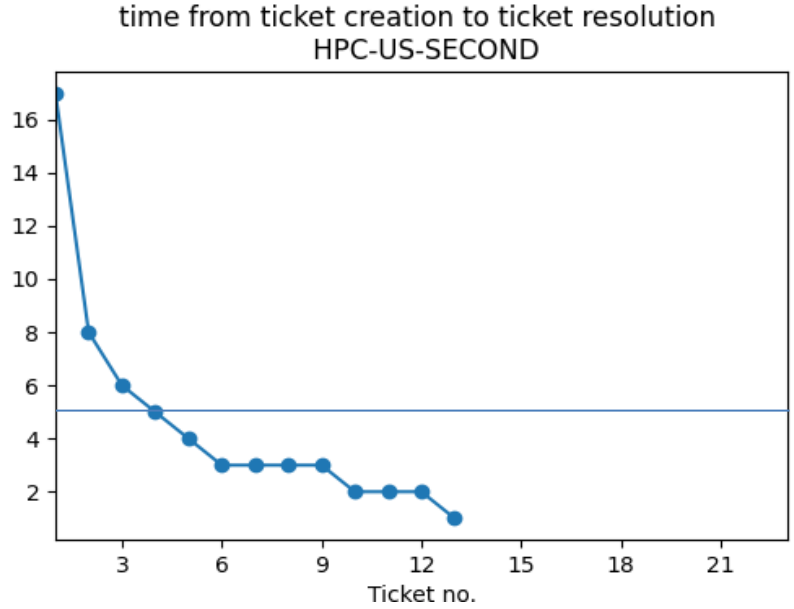


Ticket statistics June 14 – July 12

HPC-US-FIRST & HPC-US-SECOND

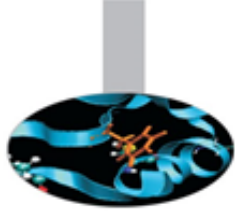


27 / 29 resolved tickets



13 / 22 resolved tickets

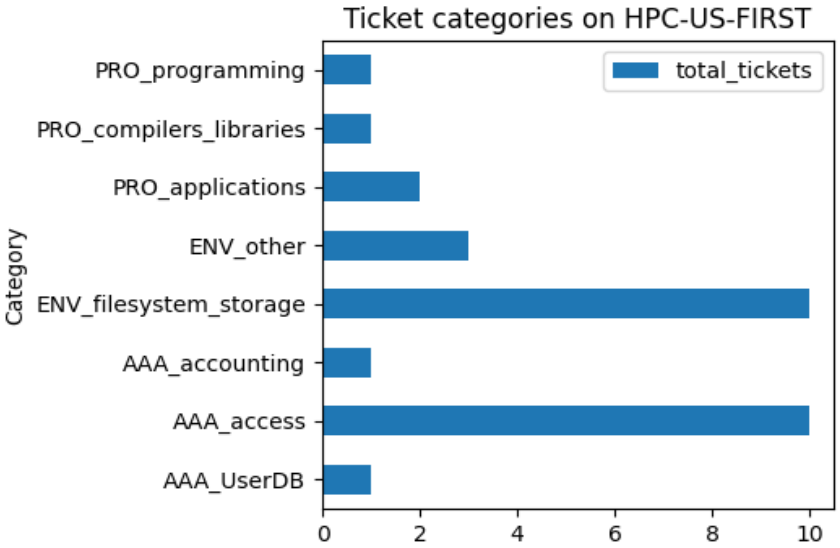
Ticket statistics June 14 – July 12

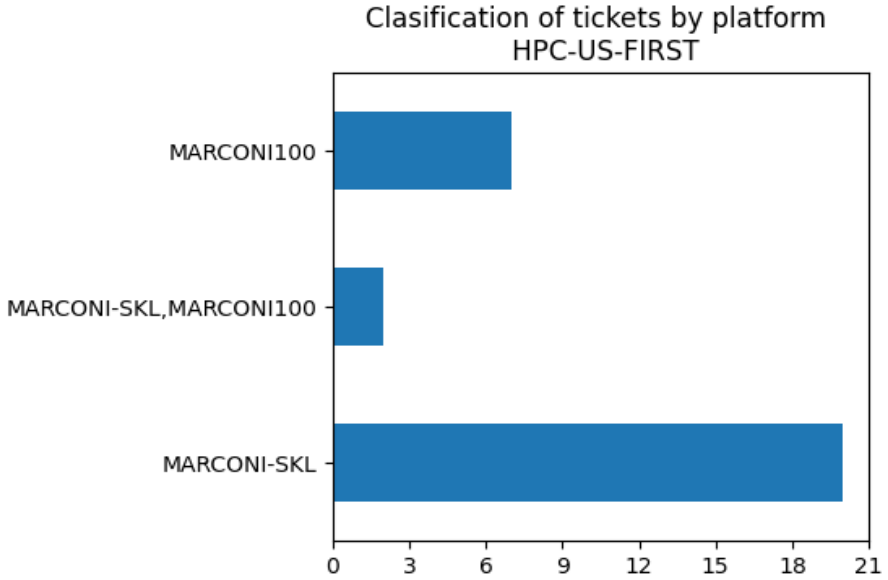


Ticket SECOND	Days	Creation time	Requestor	Subject	Notes
18693	17	2021-06-16 14:10:02	joseluis.velasco@ciemat.es	login to joseluis.velasco@login.Marconi.cineca.it	The user asked for a password reset for his username; the ticket was reopened as a second reset was requested by the user.
18814	8	2021-06-22 10:02:02	rgemi@dtu.dk	Problem with Python library	The user reported an issue with the hfd5 library missing in his python environment that was then fixed.
18998	6	2021-06-29 19:50:02	jgonzalez62@us.es	qos_special	The user asked for the possibility to run jobs that take up to 72 hours to complete and that require 1024 cores. We asked for and obtained authorization to allow this user to use of the qos_special on Marconi to run his jobs.

Ticket statistics June 14 – July 12

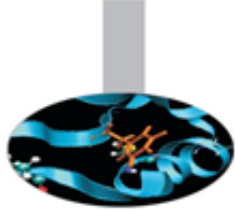
HPC-US-FIRST





Ticket statistics June 14 – July 12

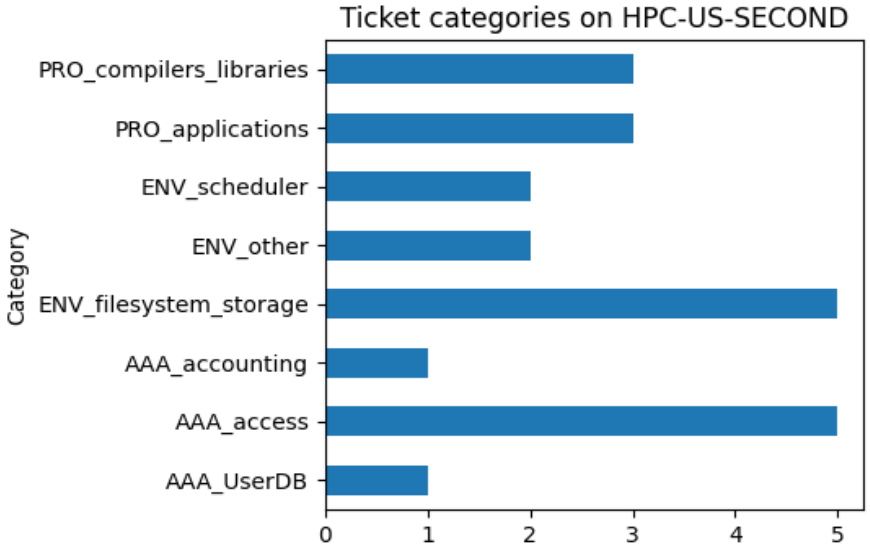
HPC-US-FIRST

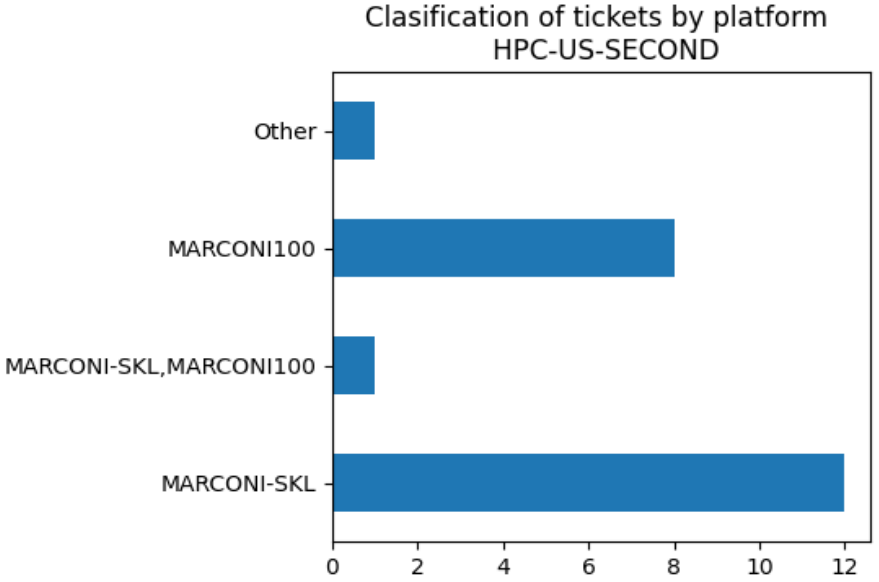


HPC-US-FIRST		By status		By platform		
		open	resolved	SKL	SKL&M100	M100
Information	ENV_filesystem_storage		1			1
	PRO_compilers_libraries		1	1		
Problem	AAA_access		3			3
	AAA_accounting		1			1
	ENV_filesystem_storage		4	4		
	ENV_other		2	1		1
	PRO_applications	1	1	2		
	PRO_programming		1			1
Service Request	AAA_UserDB		1	1		
	AAA_access	1	6	6	1	
	ENV_filesystem_storage		5	4	1	
	ENV_other		1	1		
total	**29**	**2**	**27**	**20**	**2**	**7**

Ticket statistics June 14 – July 12

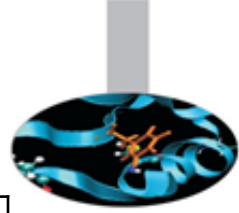
HPC-US-SECOND





Ticket statistics June 14 – July 12

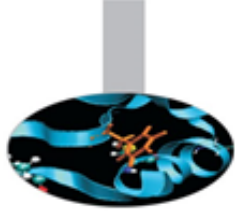
HPC-US-SECOND



HPC-US-SECOND		By status		By platform			
		open	resolved	SKL	SKL&M100	M100	Other
Information	AAA_access	1				1	
	ENV_filesystem_storage		1			1	
	PRO_applications	1		1			
Problem	AAA_access	1		1			
	ENV_other	1	1	1		1	
	PRO_applications	1	1	1		1	
	PRO_compilers_libraries	2	1	1		2	
Service Request	AAA_UserDB		1				1
	AAA_access	1	2	2	1		
	AAA_accounting		1			1	
	ENV_filesystem_storage	1	3	4			
	ENV_scheduler		2	1		1	
total	**22**	**9**	**13**	**12**	**1**	**8**	**1**

Module usage on Marconi-SKL

Batch jobs June 2021



category	modulename	partition	qos	module_load_count	elapsed_timelimit_sum
compiler	python	skl_fua_prod	normal	6	2,49
			skl_qos_fuabprod	6	2,98
library	fftw	skl_fua_prod	normal	6	2,49
			skl_qos_fuabprod	6	2,98
	hdf5	skl_fua_prod	normal	92	18,24
			skl_qos_fuabprod	6	2,98
	netcdf	skl_fua_prod	normal	6	2,49
			skl_qos_fuabprod	6	2,98
	netcdff	skl_fua_prod	normal	6	2,49
			skl_qos_fuabprod	6	2,98
	petsc	skl_fua_prod	normal	92	18,24
			skl_qos_fuabprod	6	2,98
	popt	skl_fua_prod	normal	104	29,58
			skl_qos_fuabprod	12	6,41
	scalapack	skl_fua_prod	normal	86	15,75