Analyzing real code performance efficiency on the MARCONI supercomputer with the *hpcmd* tool

Serhiy Mochalskyy

2nd Annual Meeting of EUROfusion HPC ACHs

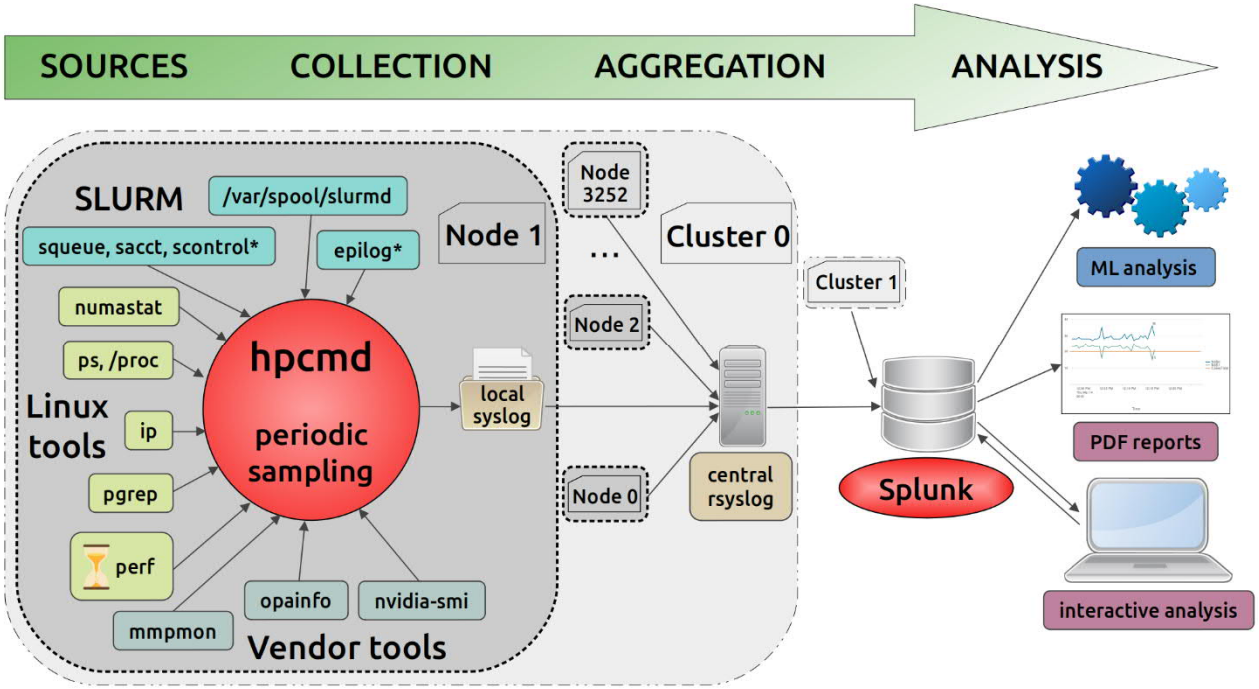
November 27-28, 2024

Advanced Computing Hub of MPG
Max-Planck-Institut für Plasmaphysik
Boltzmannstr. 2, D-85748 Garching, Germany

Hpcmd tool overview

- **The tool was developed at the Max Planck Computing and Data Facility (MPCDF).**
- ***Hpcmd* is a software daemon designed to measure performance data of running jobs on HPC compute nodes by executing periodically Linux *perf* and similar tools to gather metrics from performance counters. [https://gitlab.mpcdf.mpg.de/mpcdf/hpcmd]**
- **Intel Skylake and newer processors are fully supported to compute the performance in GFLOPS or to obtain the memory bandwidth in GB/s (Linux kernel >4.0 is required).**
- ***hpcmd* supports performance metrics from CPUs, GPUs, OmniPath and InfiniBand networks, as well as GPFS file systems.**
- ***Hpcmd* fully integrates with the SLURM batch system, enabling to correlate performance metrics with each job and to gather also other information as the jobid, the requested number of nodes, threads, etc.**

Hpcmd tool overview



picture taken from [https://gitlab.mpcdf.mpg.de/mpcdf/hpcmd]

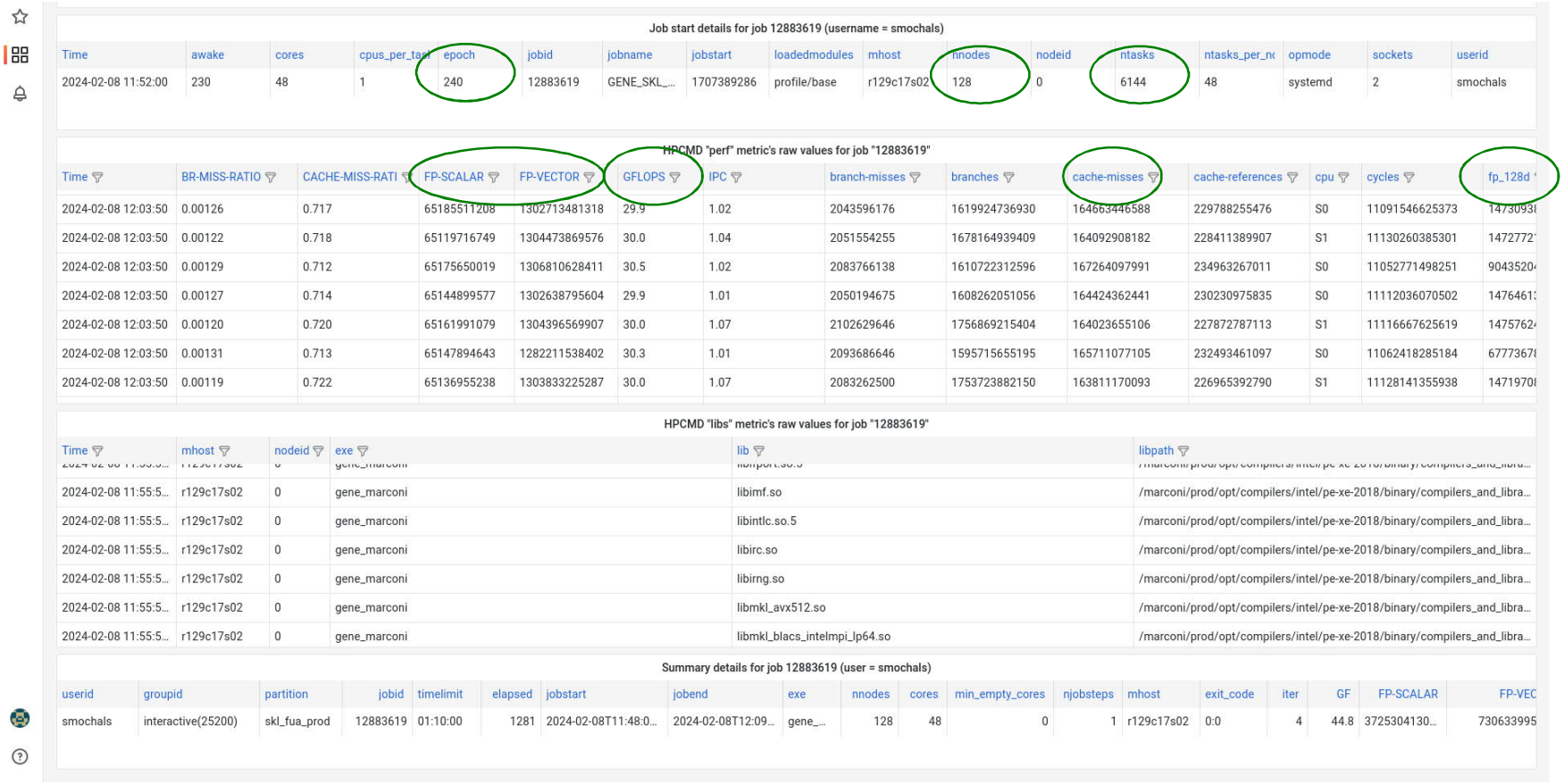
GFLOPS – total number of flops;
FP_SCALAR – number of floating point scalar operation;
FP_VECTOR – number of floating point vector operation;
fp_128d – number of double precision 128 bits (AVX) register operations;
fp_128s – number of single precision 128 bits (AVX) register operations;
…
fp_512d – number of double precision 512 bits (AVX512) register operations;
fp_512s – number of single precision 512 bits (AVX512) register operations;
Cache-misses;
IPC – instructions per cycle;
… .

- **The tool computes derived metrics and writes the data to syslog lines, that can be collected and stored in a database for subsequent analysis and visualization.**

Hpcmd data visualization using Grafana software

Marconi: https://hpcmd.hpc.cineca.it/

MPCDF: https://hpc-reports.mpcdf.mpg.de/raven



Job start details for job 12883619 (username = smochals)

Time	awake	cores	cpus_per_task	epoch	jobid	jobname	jobstart	loadedmodules	mhost	nnodes	nodeid	ntasks	ntasks_per_no	opmode	sockets	userid
2024-02-08 11:52:00	230	48	1	240	12883619	GENE_SKL_...	1707389286	profile/base	r129c17s02	128	0	6144	48	systemd	2	smochals

HPCMD "perf" metric's raw values for job "12883619"

Time	BR-MISS-RATIO	CACHE-MISS-RATI	FP-SCALAR	FP-VECTOR	GFLOPS	IPC	branch-misses	branches	cache-misses	cache-references	cpu	cycles	fp_128d
2024-02-08 12:03:50	0.00126	0.717	65185511208	1302713481318	29.9	1.02	2043596176	1619924736930	164663446588	229788255476	S0	11091546625373	1473093
2024-02-08 12:03:50	0.00122	0.718	65119716749	1304473869576	30.0	1.04	2051554255	1678164939409	164092908182	228411389907	S1	11130260385301	1472772
2024-02-08 12:03:50	0.00129	0.712	65175650019	1306810628411	30.5	1.02	2083766138	1610722312596	167264097991	234963267011	S0	11052771498251	9043520
2024-02-08 12:03:50	0.00127	0.714	65144899577	1302638795604	29.9	1.01	2050194675	1608262051056	164424362441	230230975835	S0	11112036070502	1476461
2024-02-08 12:03:50	0.00120	0.720	65161991079	1304396569907	30.0	1.07	2102629646	1756869215404	164023655106	227872787113	S1	11116667625619	1475762
2024-02-08 12:03:50	0.00131	0.713	65147894643	1282211538402	30.3	1.01	2093686646	1595715655195	165711077105	232493461097	S0	11062418285184	6777367
2024-02-08 12:03:50	0.00119	0.722	65136955238	1303833225287	30.0	1.07	2083262500	1753723882150	163811170093	226965392790	S1	11128141355938	1471970

HPCMD "libs" metric's raw values for job "12883619"

Time	mhost	nodeid	exe	lib	libpath
2024-02-08 11:55:5...	r129c17s02	0	gene_marconi	libimf.so	/marconi/prod/opt/compilers/intel/pe-xe-2018/binary/compilers_and_libra...
2024-02-08 11:55:5...	r129c17s02	0	gene_marconi	libintlc.so.5	/marconi/prod/opt/compilers/intel/pe-xe-2018/binary/compilers_and_libra...
2024-02-08 11:55:5...	r129c17s02	0	gene_marconi	libirc.so	/marconi/prod/opt/compilers/intel/pe-xe-2018/binary/compilers_and_libra...
2024-02-08 11:55:5...	r129c17s02	0	gene_marconi	libirng.so	/marconi/prod/opt/compilers/intel/pe-xe-2018/binary/compilers_and_libra...
2024-02-08 11:55:5...	r129c17s02	0	gene_marconi	libmkl_avx512.so	/marconi/prod/opt/compilers/intel/pe-xe-2018/binary/compilers_and_libra...
2024-02-08 11:55:5...	r129c17s02	0	gene_marconi	libmkl_blacs_intelmpi_lp64.so	/marconi/prod/opt/compilers/intel/pe-xe-2018/binary/compilers_and_libra...

Summary details for job 12883619 (user = smochals)

userid	groupid	partition	jobid	timelimit	elapsed	jobstart	jobend	exe	nnodes	cores	min_empty_cores	njobsteps	mhost	exit_code	iter	GF	FP-SCALAR	FP-VEC
smochals	interactive(25200)	skl_fua_prod	12883619	01:10:00	1281	2024-02-08T11:48:0...	2024-02-08T12:09...	gene_...	128	48	0	1	r129c17s02	0:0	4	44.8	3725304130...	730633995

Hpcmd tests on MARCONI

Triad test

```
start = MPI_WTIME()
  do num_iter =1, total_rep
    do i=1,total_iter
      a(i) = b(i) + c(i) * d(i)
    enddo
    if(a(total_iter/2) == 100) print *, "Test =", a(total_iter/2)
  enddo
finish = MPI_WTIME()
```

Linpack test

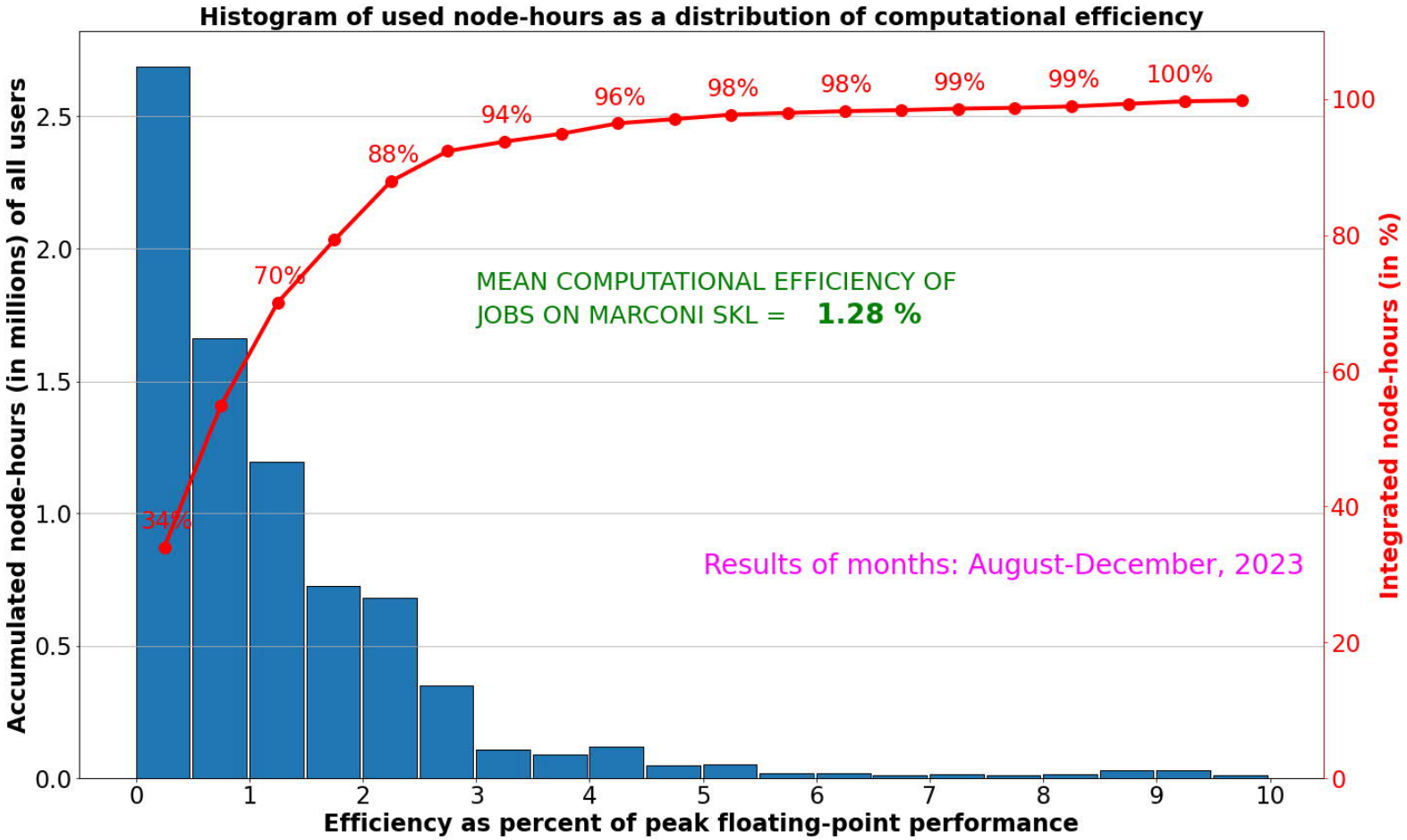
SMP Linpack benchmark:
problem size = 100000
5 repetitions

Linpack provides = 2147 GFLOPs
***hpcmd* shows = 2169 GFLOPs**

Table 1 Results of FLOP rates for different performance counters

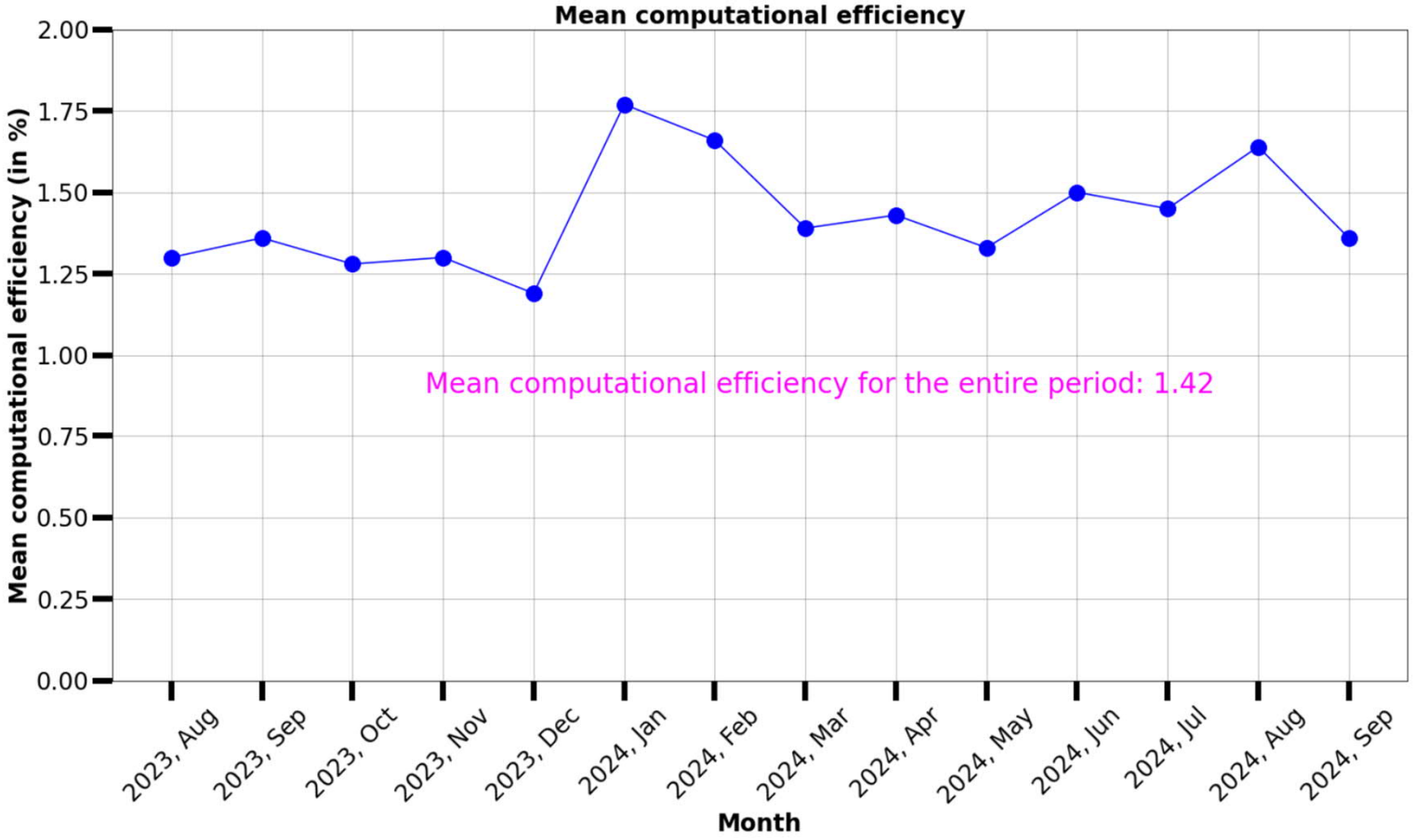
Test #	Run number	MPI	Array type (byte)	Compilation option	Major metric	GFLOPS analytical	GFLOPS *hpcmd*
1	9953449	1	Real*8	-O0	Fp_d	**0.305**	**0.305**
2	9955117	1	Real*4	-O0	Fp_s	**0.308**	**0.307**
3	9953581	1	Integer	-O0	-	-	-
4	9953600	1	Real*8	-O2 -mavx	Fp_128d	**0.988**	**0.988**
5	9953773	1	Real*4	-O2 -mavx	Fp_128s	**1.96**	**1.96**
6	9953642	1	Real*8	-O2 -xCORE-AVX2	Fp_256d	**1.03**	**1.02**
7	9953708	1	Real*4	-O2 -xCORE-AVX2	Fp_256s	**2.07**	**2.08**
8	9953658	1	Real*8	-O2 -xCOMMON-AVX512	Fp_512d	**1.1**	**1.09**
9	9953702	1	Real*4	-O2 -xCOMMON-AVX512	Fp_512s	**2.18**	**2.17**
10	9954904	48	Real*8	-O2 -xCOMMON-AVX512	Fp_512d	**12.5**	**12.47**
11	9954925	48	Real*4	-O2 -xCOMMON-AVX512	Fp_512s	**26.6**	**26.6**
12	9954929	48	Real*8	-O2 -xCORE-AVX2	Fp_256d	**12.17**	**12.13**
13	9954933	48	Real*4	-O2 -xCORE-AVX2	Fp_256s	**25.03**	**25**
14	9954999	48	Real*8	-O2 -mavx	Fp_128d	**11.84**	**11.8**
15	9954991	48	Real*4	-O2 -mavx	Fp_128s	**24.8**	**24.8**
16	9955558	48	Real*8	-O0	Fp_d	**10.11**	**9.12**
17	9955678	48	Real*4	-O0	Fp_s	**14.57**	**14.5**
18	9955475	192	Real*8	-O2 -xCOMMON-AVX512	Fp_512d	**50.41**	**50.69**

Hpcmd results – year 2023 (August-December)



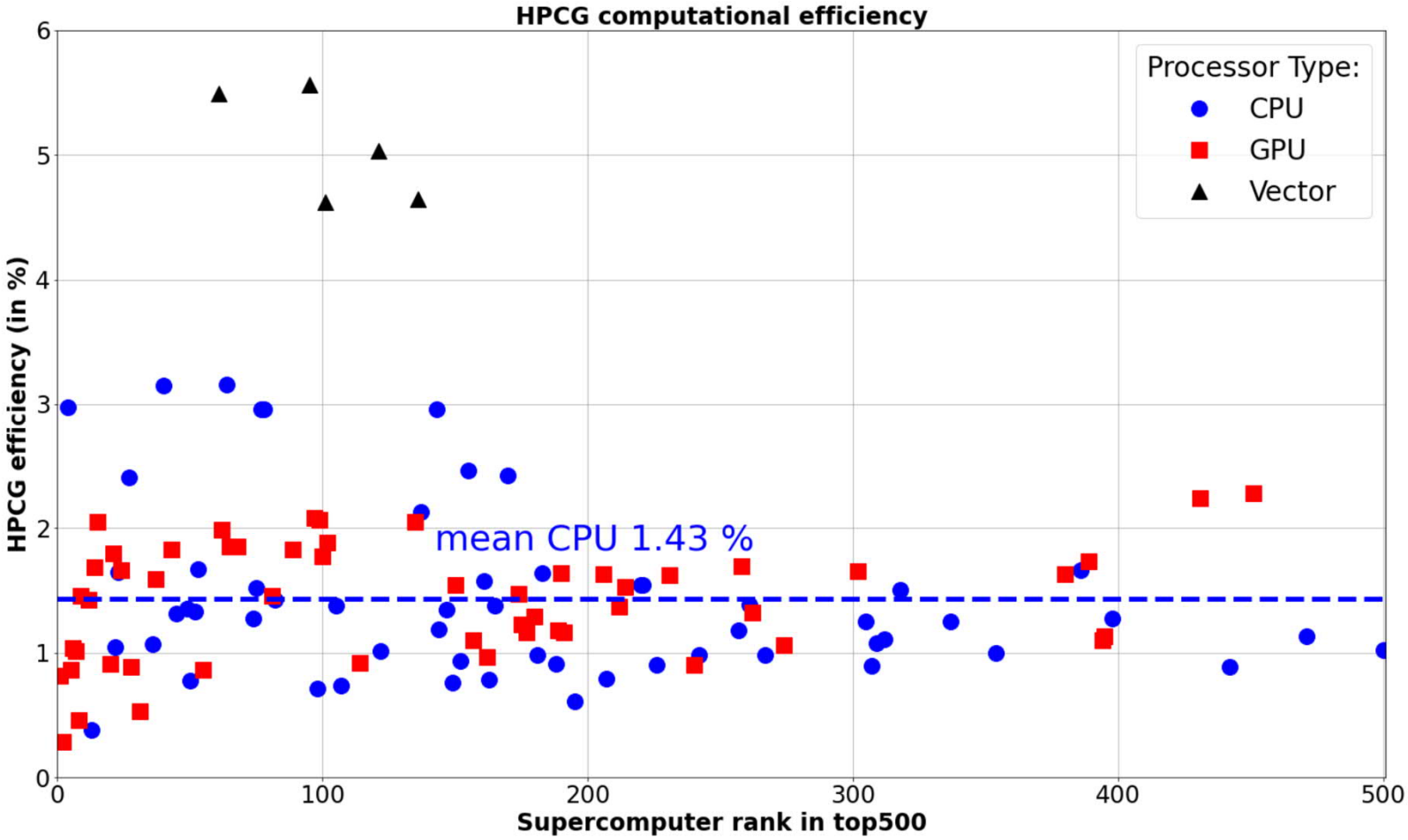
➢ **50% of supercomputer usage has computational efficiency below 1%.**

Hpcmd results — years 2023-2024



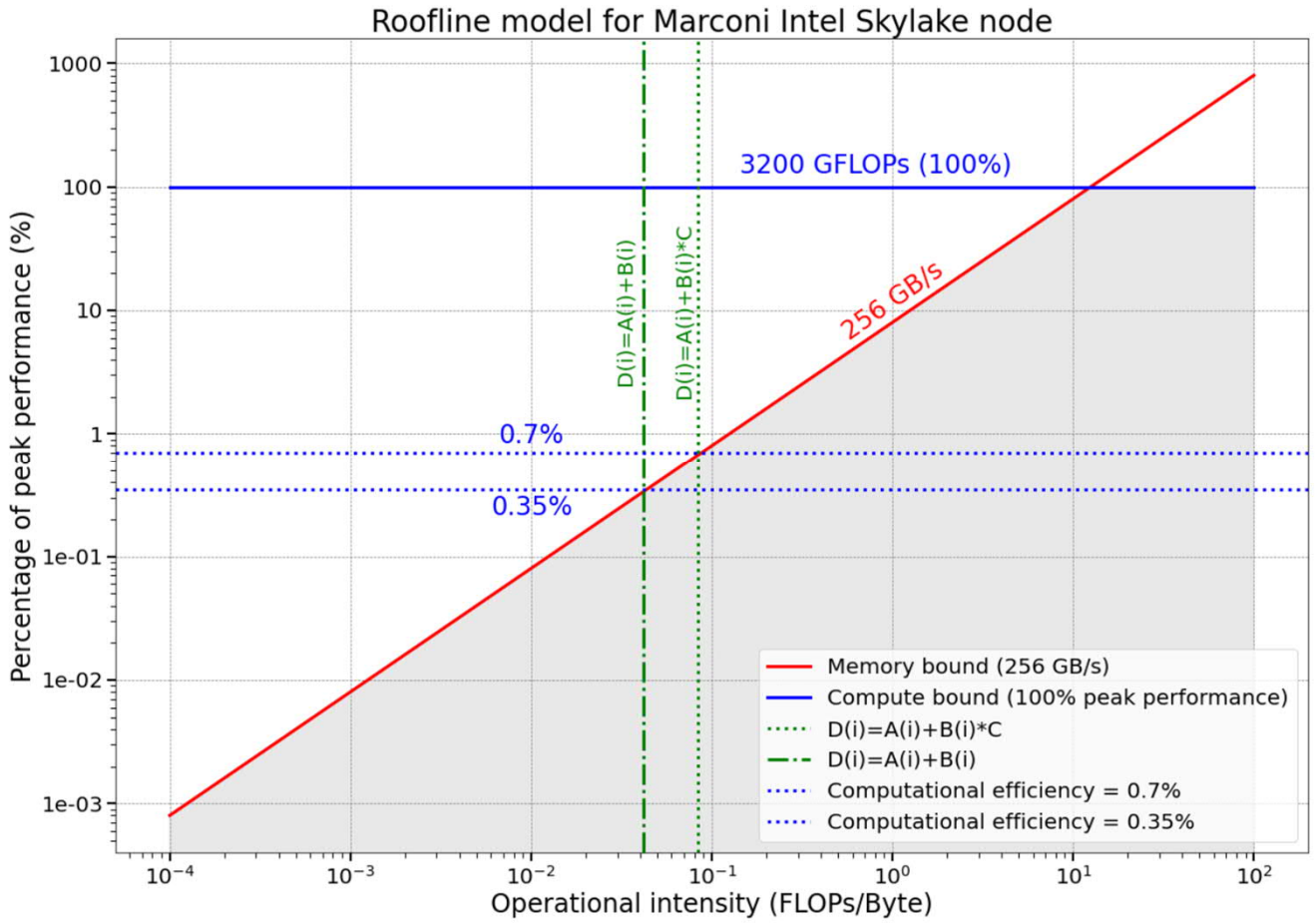
- **The computational efficiency of the entire supercomputer is relatively constant.**

HPCG benchmark of top500 list



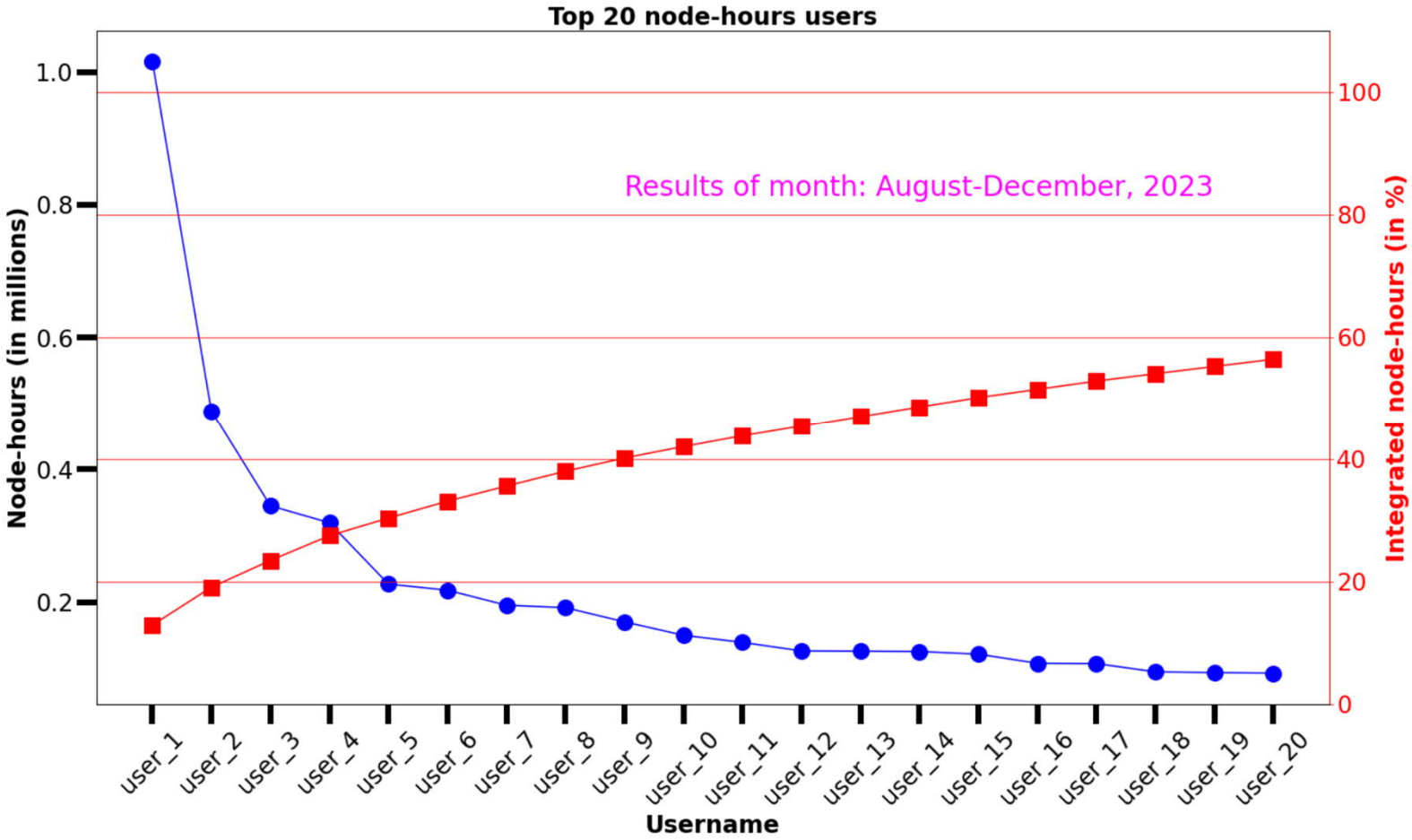
➢ **For CPU-based only systems the mean computational efficiency is 1.43%.**

Roofline model Marconi SKL



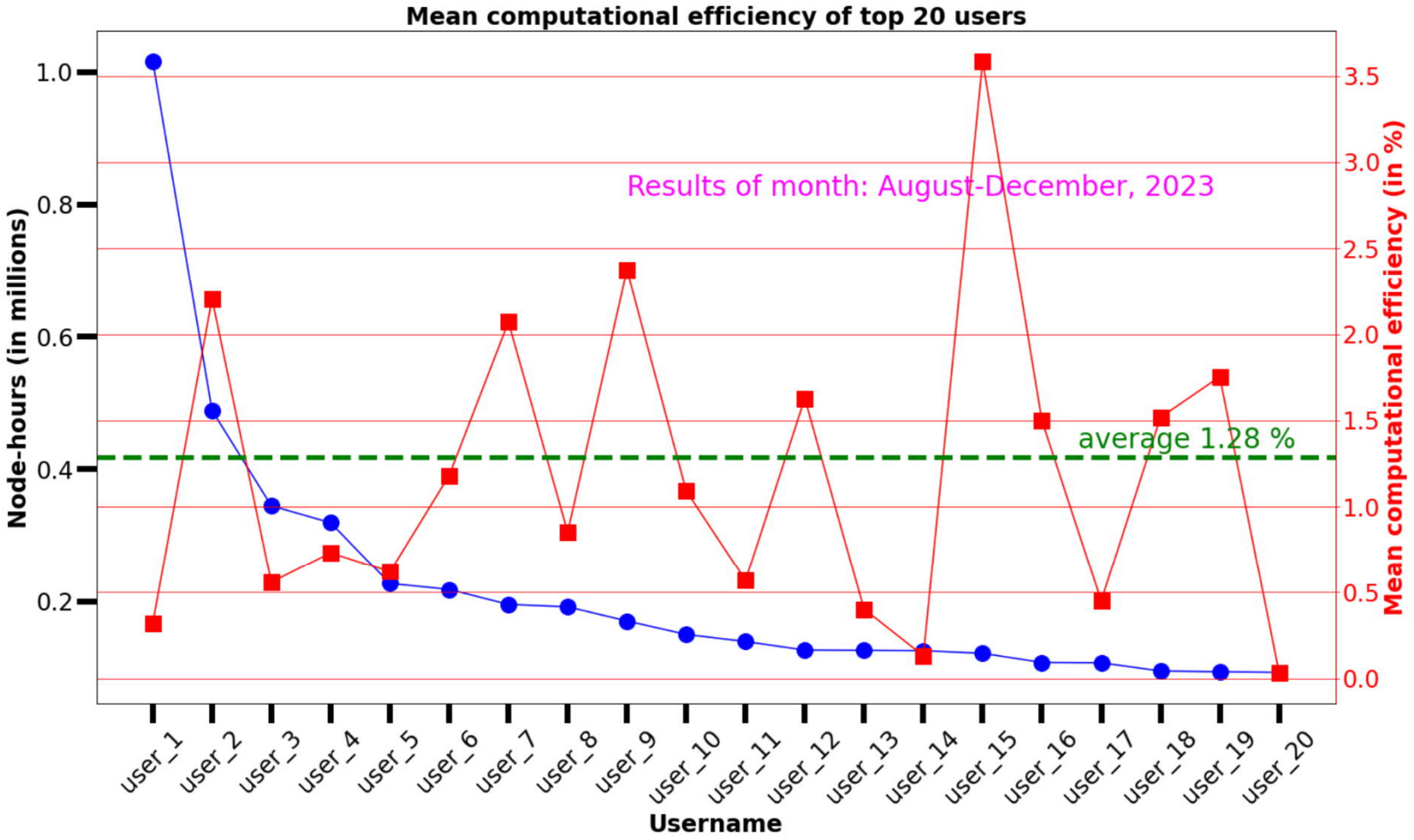
- **Stream triad = 0.7%**
- **Stream add = 0.35%**
- **Stream full = 0.3%**

Hpcmd results — year 2023 (August-December)



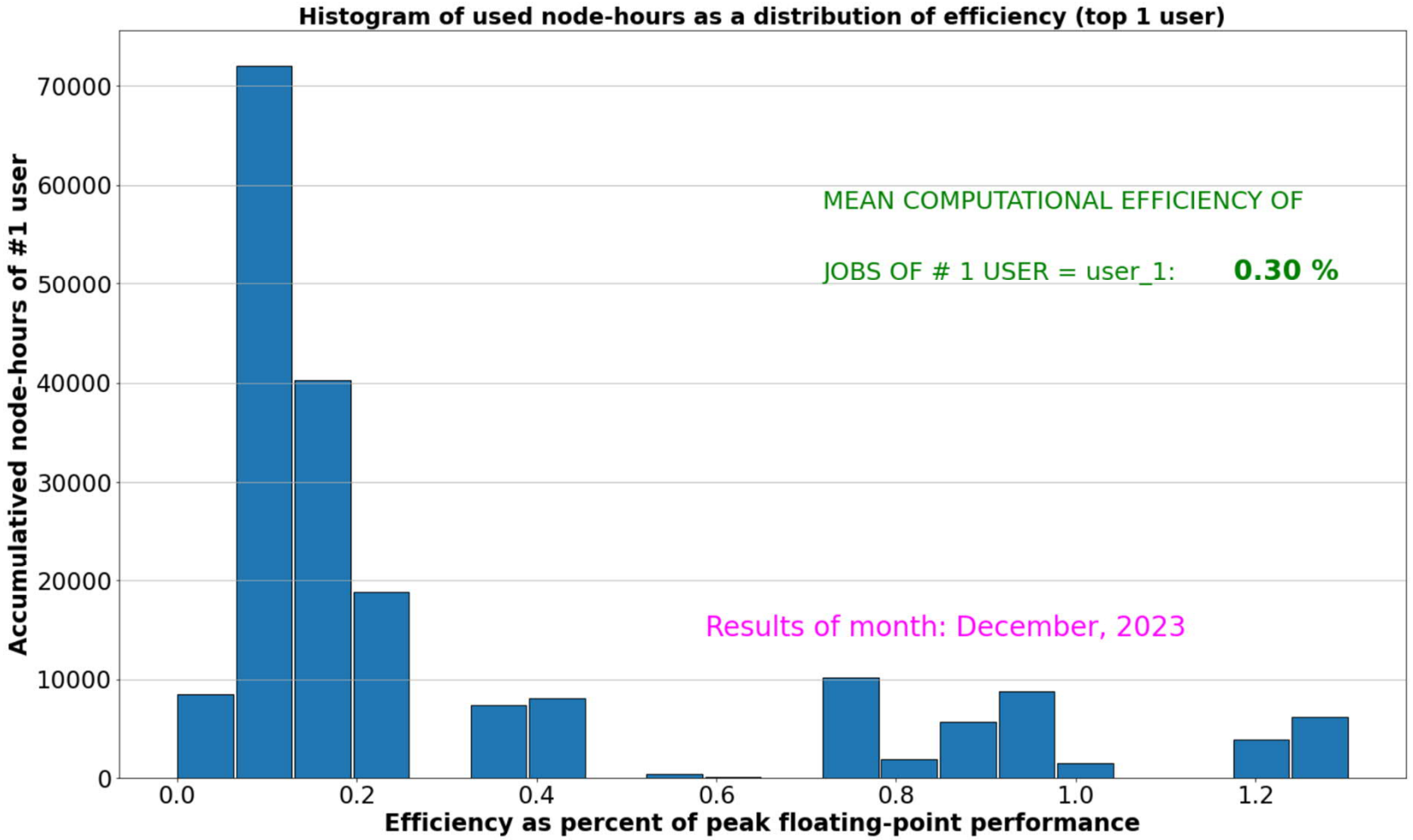
➢ **The top 15 users in terms of node-hours utilize ~50% of the machine's capacity.**

Hpcmd results – year 2023 (August-December)

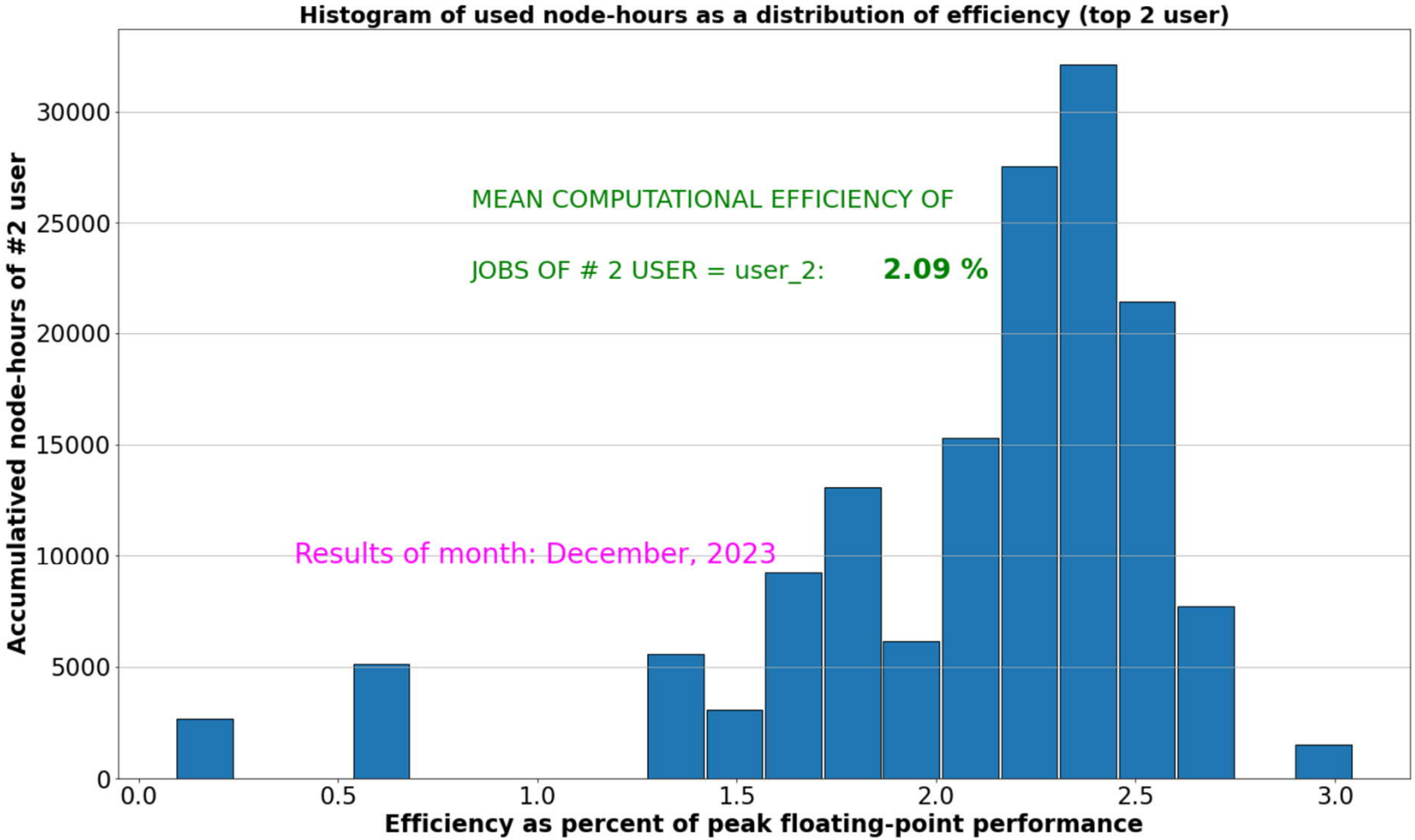


➢ **Out of the top 20 users, 10 have computational efficiency below 1%.**

Hpcmd results December 2023 – top 1 node-hours user



Hpcmd results December 2023 – top 2 node-hours user



Hpcmd results December 2023 – top 1 node-hours user

- Mean computational efficiency 0.3%.
- ~200 000 node-hours, 12% of the whole supercomputer.
- 200 000 / 24 (hours) / 31 (days) = 269 nodes: full time run over one month.
- 200 000 node-hours ≈ 100 000 EUR

Out[138]:

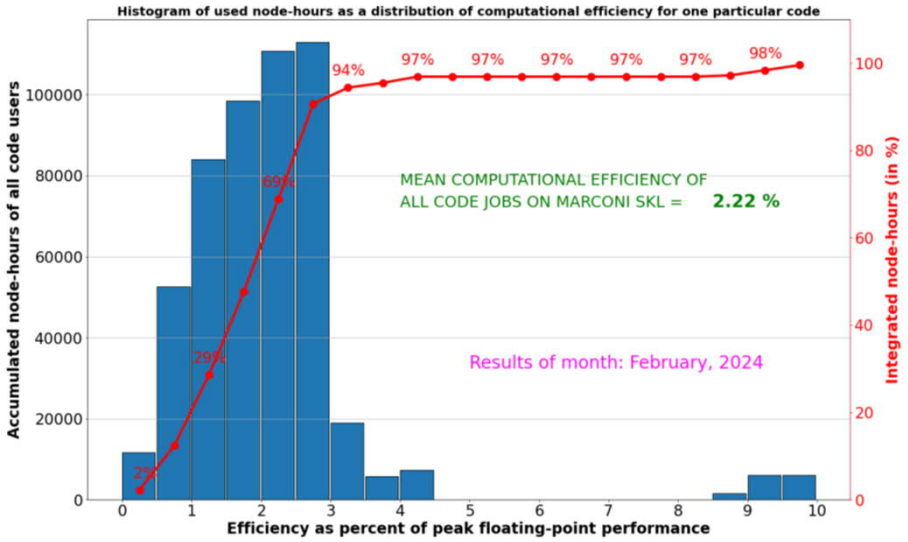
ne-es	cache-references	cpu	cycles	fp_128d	fp_128s	fp_256d	fp_256s	fp_512d	fp_512s	fp_d	fp_s	instructions	jobid
274	10520285280	S0	11407792328262	0	0	0	0	0	0	52551291029	125096718911	15423799040218	12718037
278	11259772726	S0	11409101485245	0	0	0	0	0	0	57335154957	138570819092	15445213386620	12718037
274	11908953318	S0	11405559628294	0	0	0	0	0	0	61059890286	144020040952	15429373716676	12718037
232	12225599152	S0	11407434390021	0	0	0	0	0	0	43481301021	143930550257	15395523151205	12718037
032	10496045040	S1	11409594719373	0	0	0	0	0	0	53021043098	125943706498	15537448163187	12718037
...	...	...	...	...	...	...	...	...	...	...	...	...	...
264	10128242643	S1	10392611616589	0	0	0	0	0	0	39232354636	126250394956	14224743717220	12718037
223	10649188819	S1	10393989827028	0	0	0	0	0	0	48028581965	156605977337	14242581796545	12718037
298	19043923945	S1	10382810764810	0	0	0	0	0	0	109510186117	353275861071	14413717614307	12718037
276	17207799909	S0	10386605729781	0	0	0	0	0	0	94935826500	308846497616	14335958968278	12718037
296	14039047011	S1	10384271814478	0	0	0	0	0	0	66342968860	216403731177	14275780632033	12718037

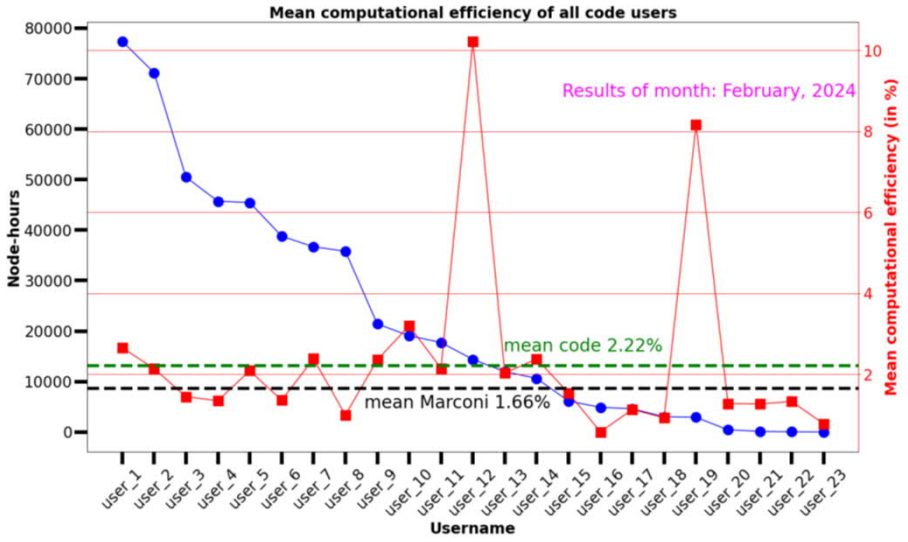
_W_HOME	AVG_W_NET	AVG_W_SCRATCH	AVG_W_WORK	FP-SCALAR	FP-VECTOR	GF	MAX_SLURMSTEPD_CPUTIME	MAX_SLURMSTEPD_RSS
0	518758	0	372031	1268034168920966	0	26419624	0	0

- Large number of scalar operations and 0 vector.

Computational efficiency of individual projects/codes

hpcmd results February, 2024

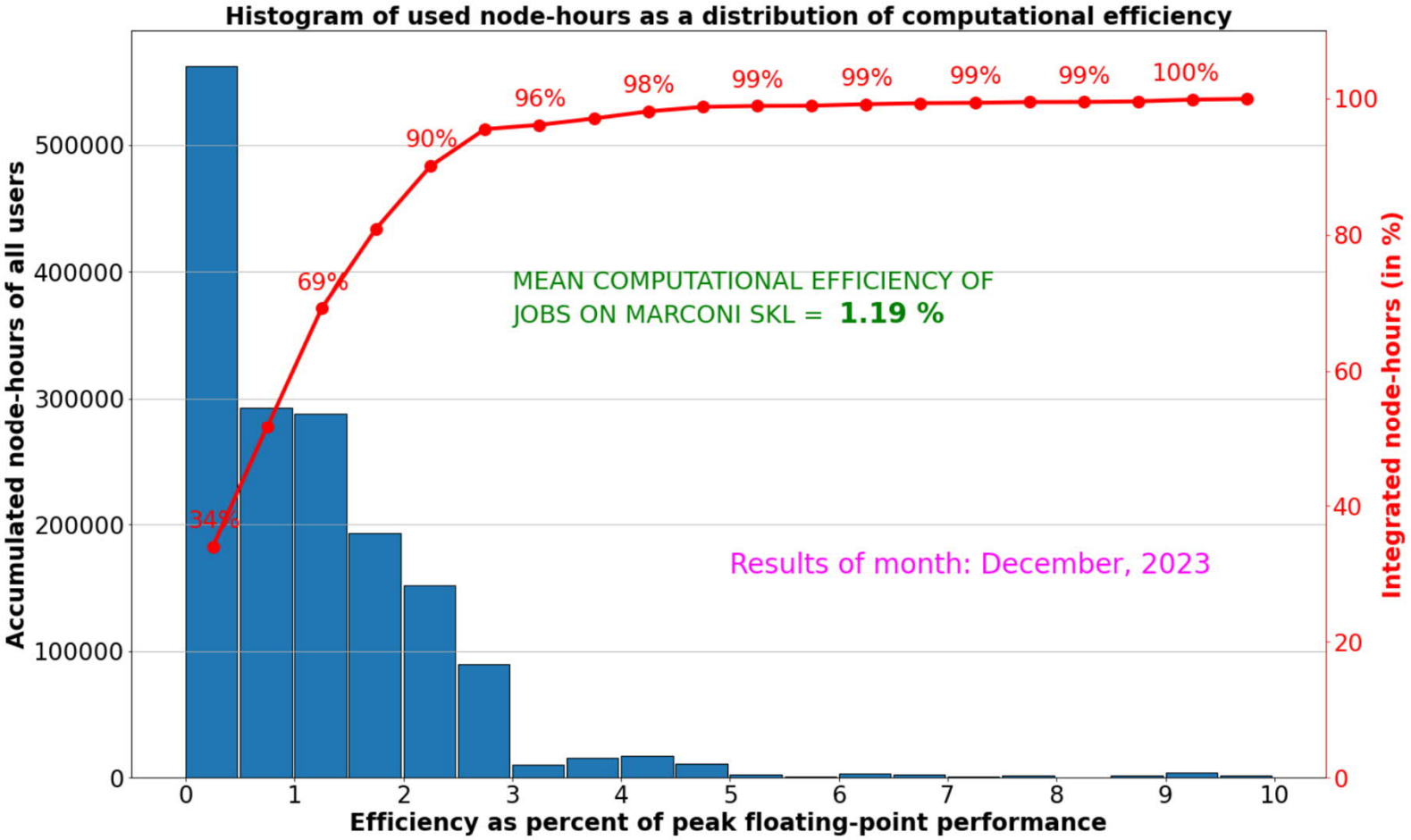




- We can **measure** the **computational efficiency of** any **project**, **group**, or even **division**.
- For **this particular project**, the mean computational efficiency was relatively high at **2.22%** compared to the mean on **Marconi**, which was **1.66%**.

Thank you for our attention!

Hpcmd results December 2023



➢ **50% of supercomputer usage has computational efficiency below 1%.**

Top 1 node-hours user – price estimation

200 000 / 24 (hours) / 31 (days) = 269 nodes: full time run over one month.

Intel® Xeon® Platinum 8160 Prozessor : ~ 5K EUR = 10k EUR node / 60 months = 167 EUR month one node.

All other hardware infrastructure including network, cooling, support ~10k EUR node / 60 months = 167 EUR month one node

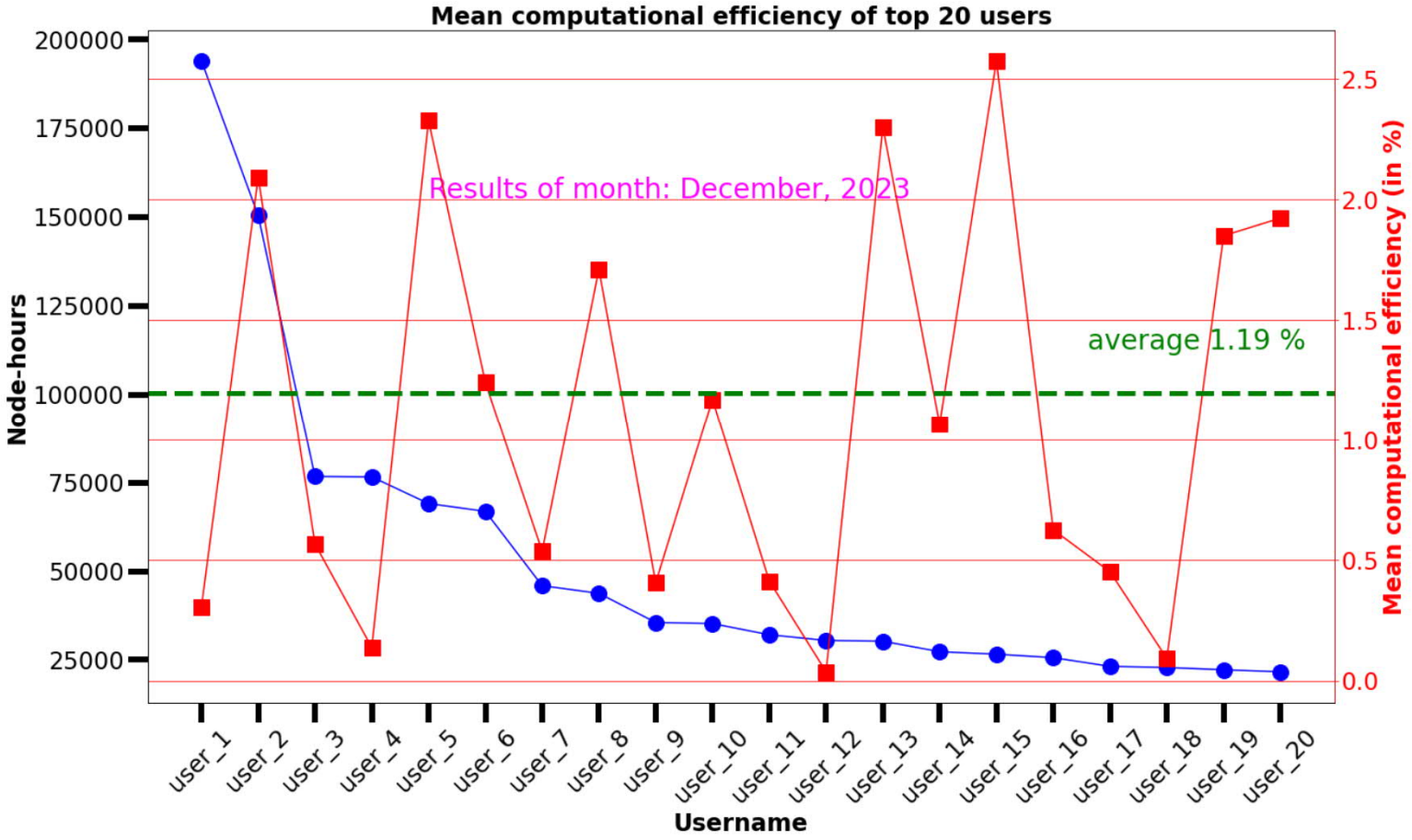
Uses_1: ~200 000 node-hours per month = 269 nodes = 269 * (167 + 167) = ~90k EUR per month

Electricity: one node ~250 W
200 000 node-hours * 250 W = 50 MW hours (for one month) – without cooling
Cooling about the same as node consumption i.e. + 50 MW
Italy 100 EUR per 1 MW hour.
50 MW hours = 50 * 100 = 5000 EUR (nodes); 50 MW hours = 50 * 100 = 5000 EUR (cooling)

Total very approximately => 90 + 5 + 5 =100k EUR per month ~1.2M EUR per year.

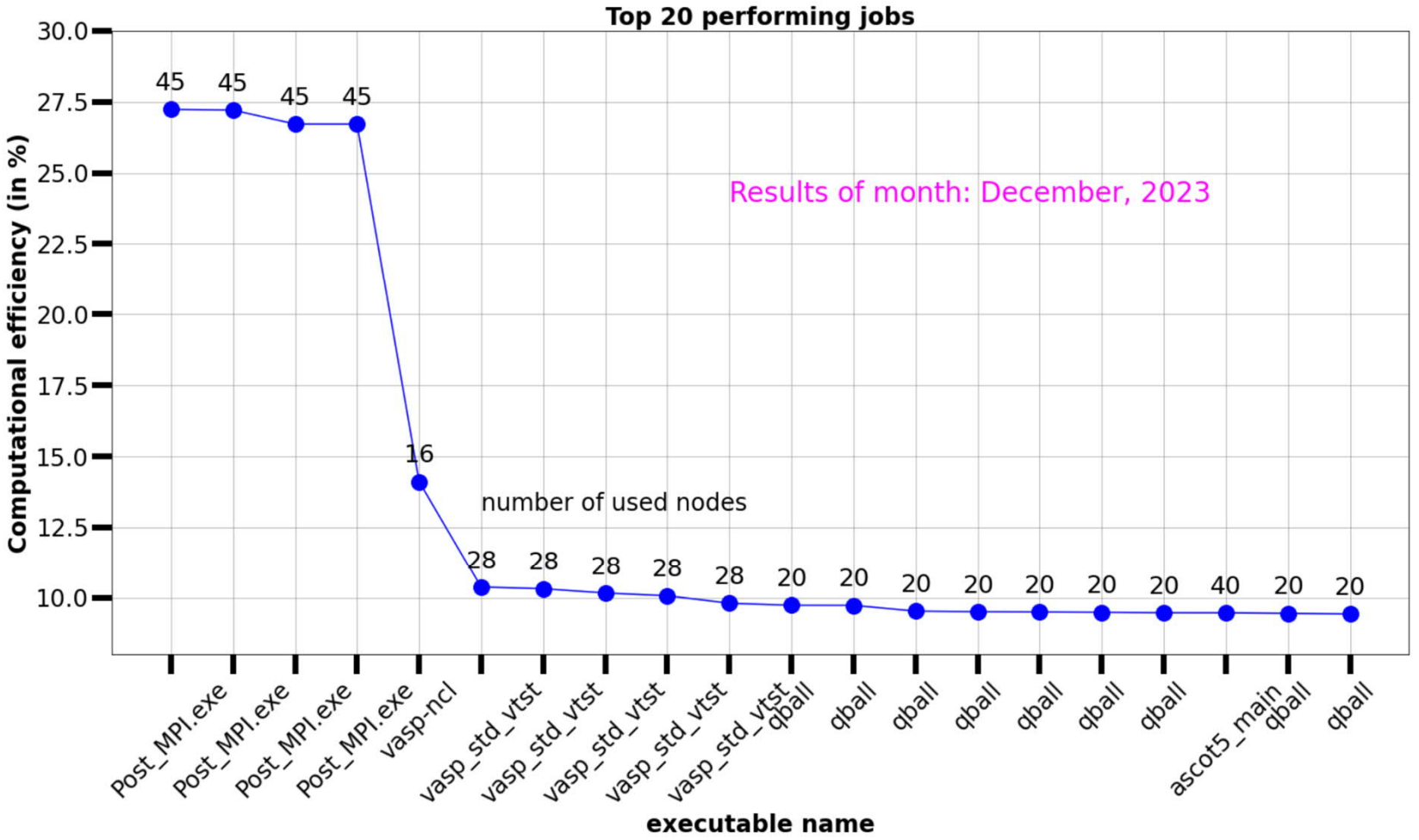
Cineca data:
Marconi SKL price for core-hour (including everything) = 0.012 EUR
Price for node-hour = 0.012*48 = 0.576 EUR
Price for 200 000 node hours = 0.576 EUR * 200 000 = 115 200 EUR per month
115 200 EUR per month * 12 months = 1.38 M EUR per year.

Hpcmd results December 2023



- **Out of the top 20 users, 10 have computational efficiency below 1%.**
- **Top 1 user has computational efficiency 0.3%.**

Hpcmd results December 2023 – top 20 computational efficiency jobs



- Many high-performance jobs involve the use of commercial libraries.
- The maximum efficiency goal is ~9%.

Hpcmd results December 2023 – max job's efficiency vs used nodes

Used nodes Interval	Number of jobs	Percentage of total jobs (%)	Max computational efficiency	Weighted Computational Efficiency Percentage (%)	Integrated node-hours	Percentage of node-hours (%)
[1, 1]	**6813**	**57.03**	8.92	1.15	20393	1.23
[2, 9]	2469	20.67	8.95	1.15	183888	11.13
[10, 17]	685	5.73	14.11	1.32	94531	5.72
[18, 33]	823	6.89	10.38	**1.36**	232031	14.04
[34, 65]	923	7.73	**27.23**	1.27	**551346**	**33.36**
[66, 129]	161	1.35	3.20	1.35	312958	18.94
[130, 256]	70	0.59	4.44	**0.71**	256831	15.54
[>=257]	2	0.02	0.00	**0.00**	753	0.05
Total	**11946**	100.01			**1,652,731**	100.01

Hpcmd results December 2023 – some statistics

users with most of submitted jobs

user_name	count
user_1	3059
user_2	1221
user_3	946
user_4	325
user_5	266