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

Serhiy Mochalskyy

E-TASC General Meeting

November 12th, 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
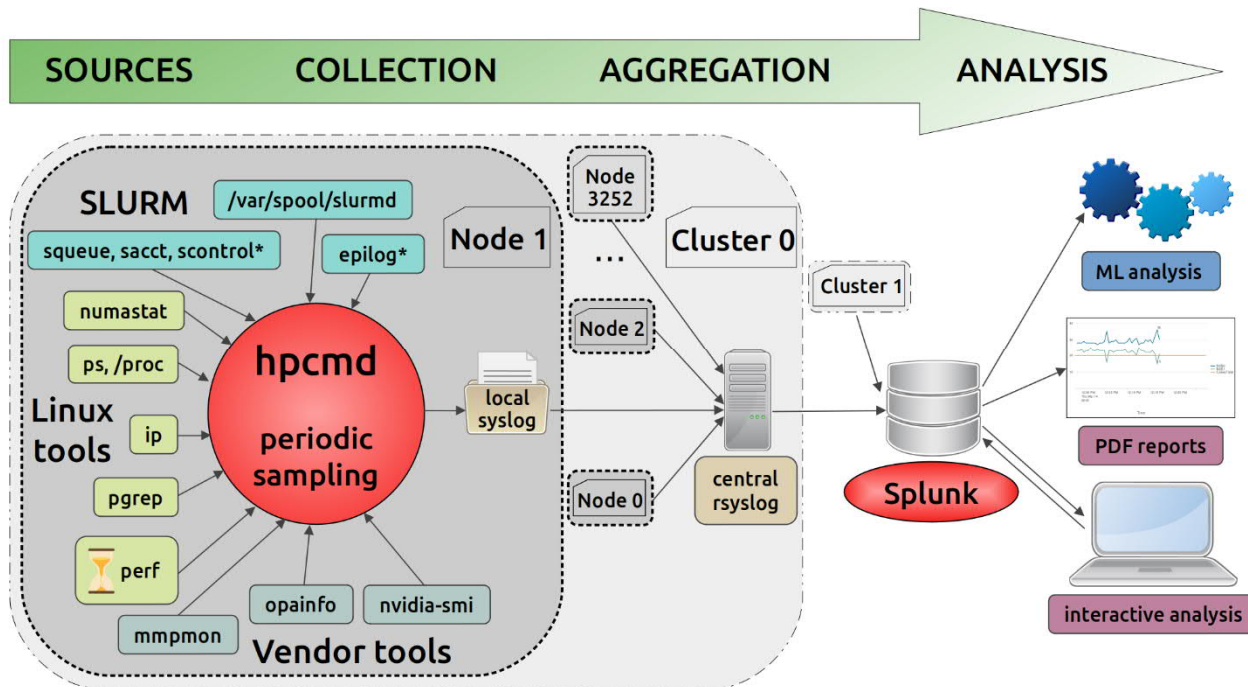
picture taken from [https://gitlab.mpcdf.mpg.de/mpcdf/hpcmd]

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

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

Histogram of used node-hours as a distribution of computational efficiency

MEAN COMPUTATIONAL EFFICIENCY OF
JOBS ON MARCONI SKL =   **1.28 %**

Results of months: August-December, 2023
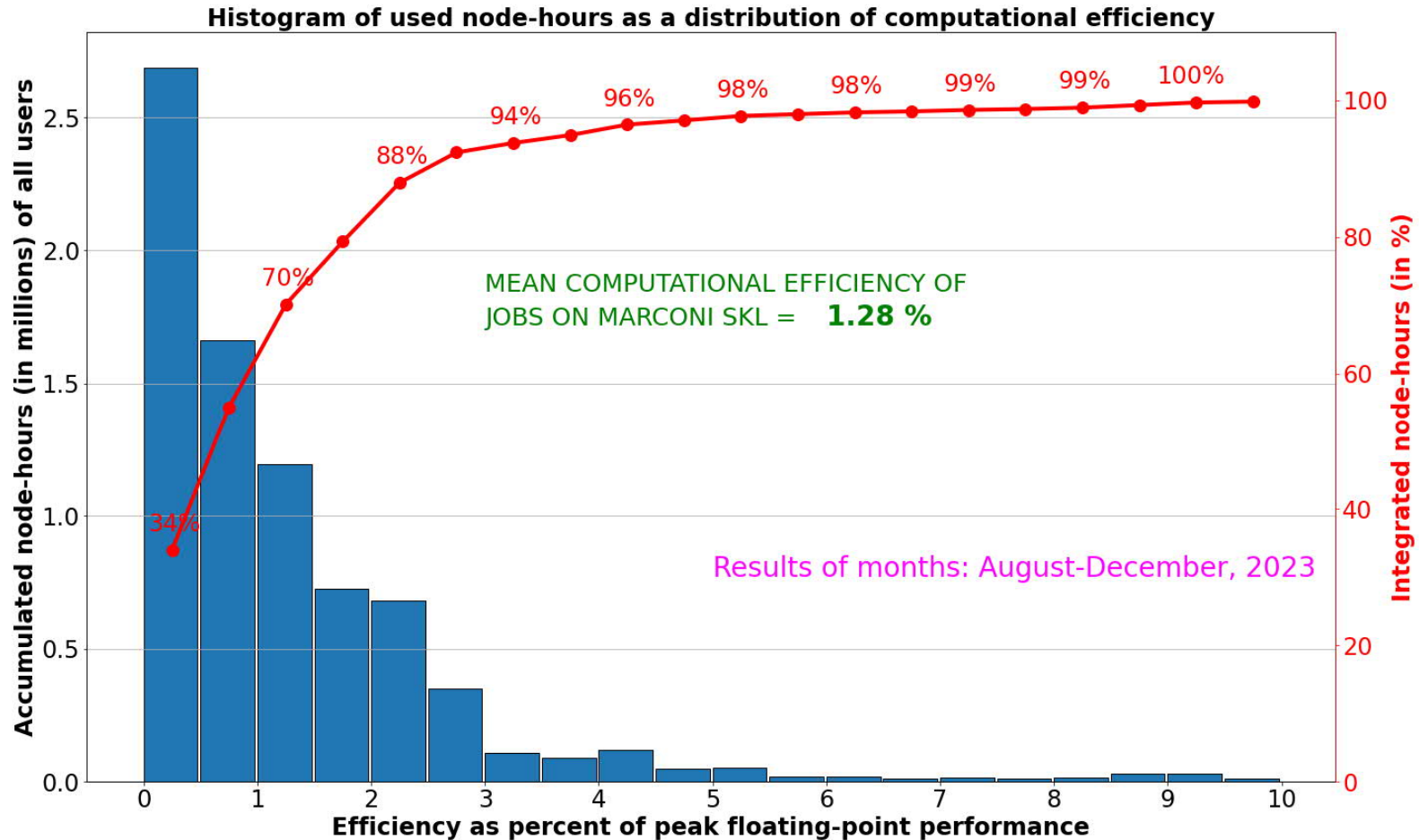
➤ **50% of supercomputer usage** has computational **efficiency below 1%.**

Hpcmd results – years 2023-2024
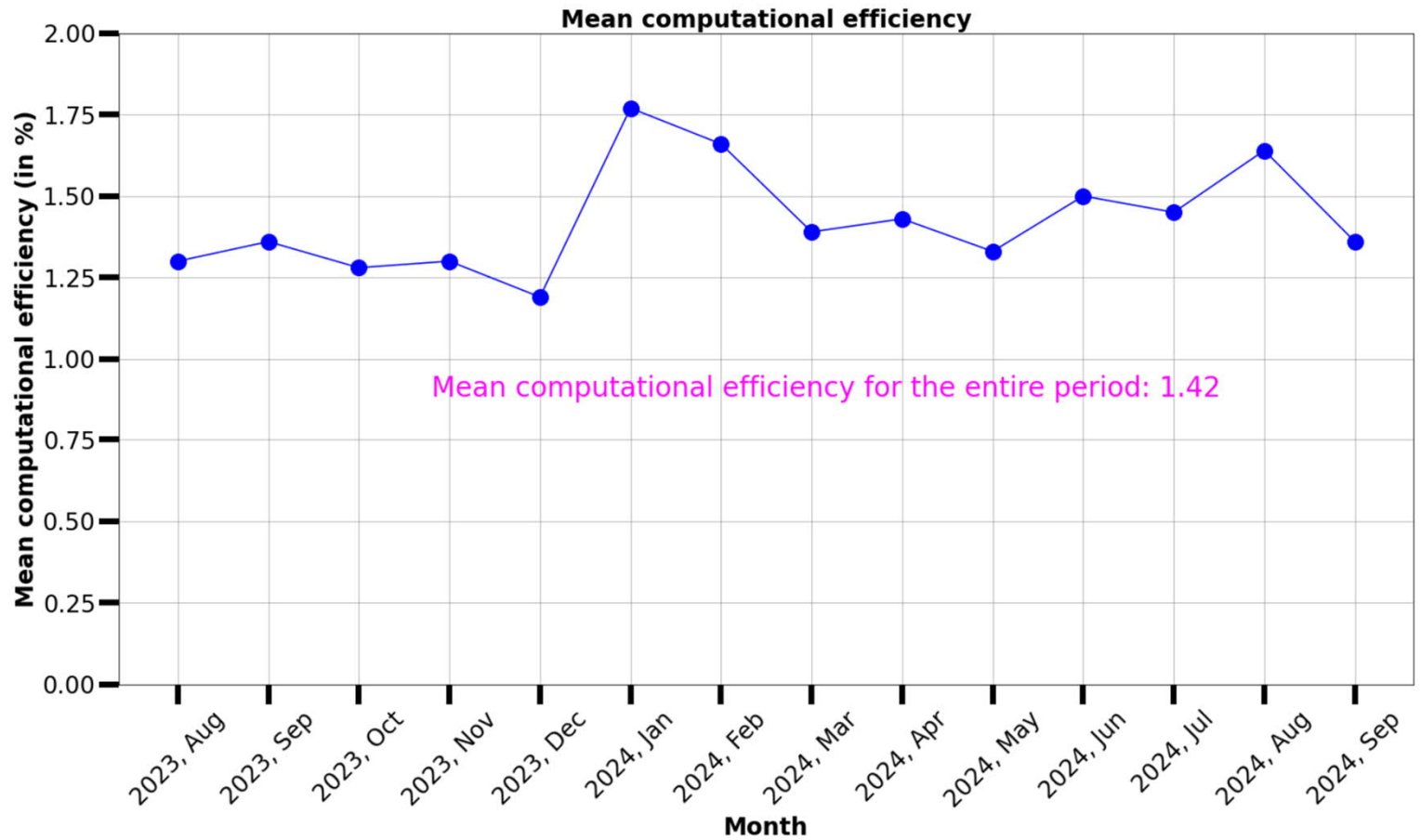
> The **computational efficiency** of the entire supercomputer is relatively **constant**.

HPCG benchmark of top500 list

HPCG computational efficiency

mean 1.58 %
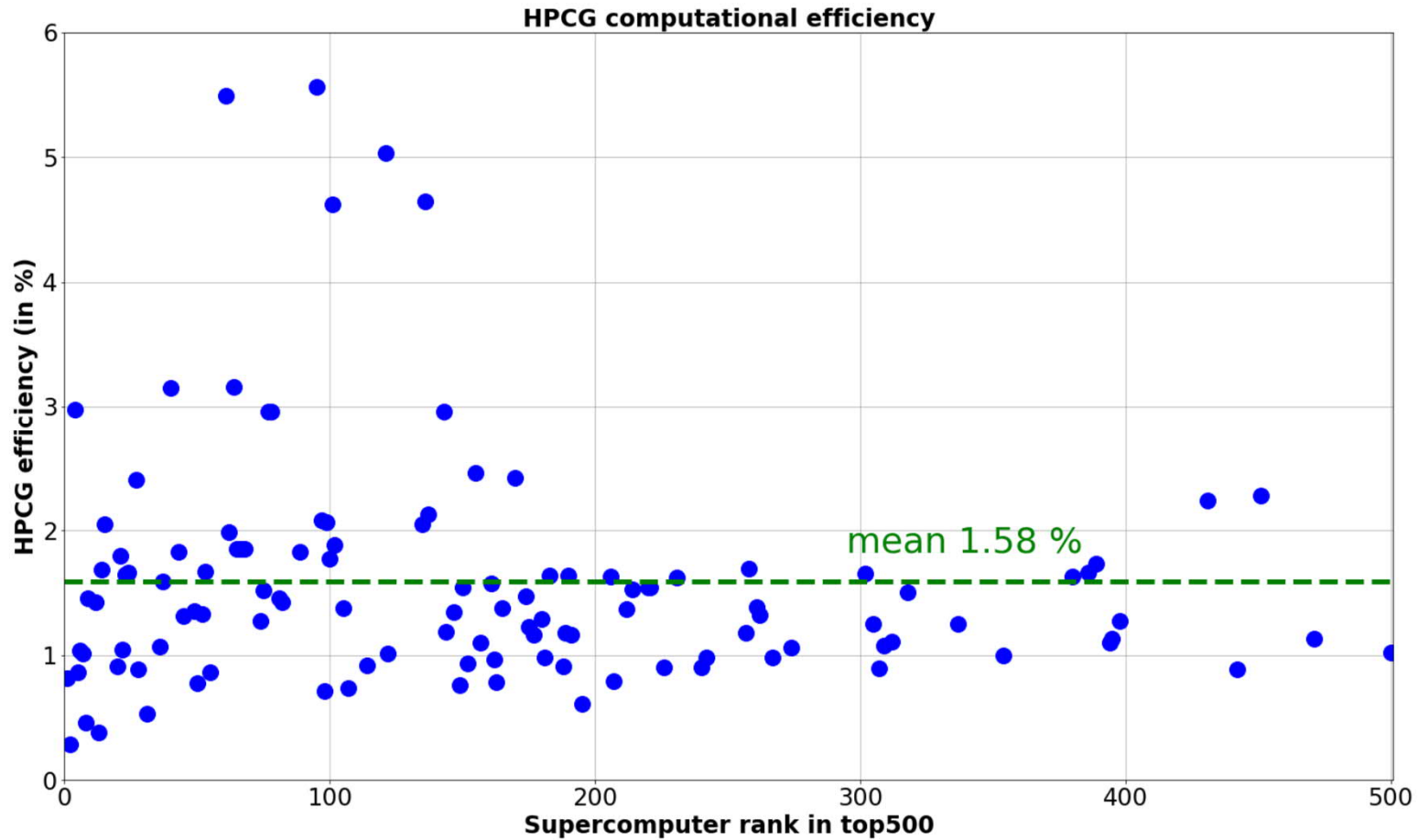
- ➢ The mean computational efficiency is **1.58%**.
- ➢ For **CPU-based** only **systems** the mean computational efficiency is **1.72%**.

Hpcmd results – year 2023 (August-December)

Top 20 node-hours users

Results of month: August-December, 2023
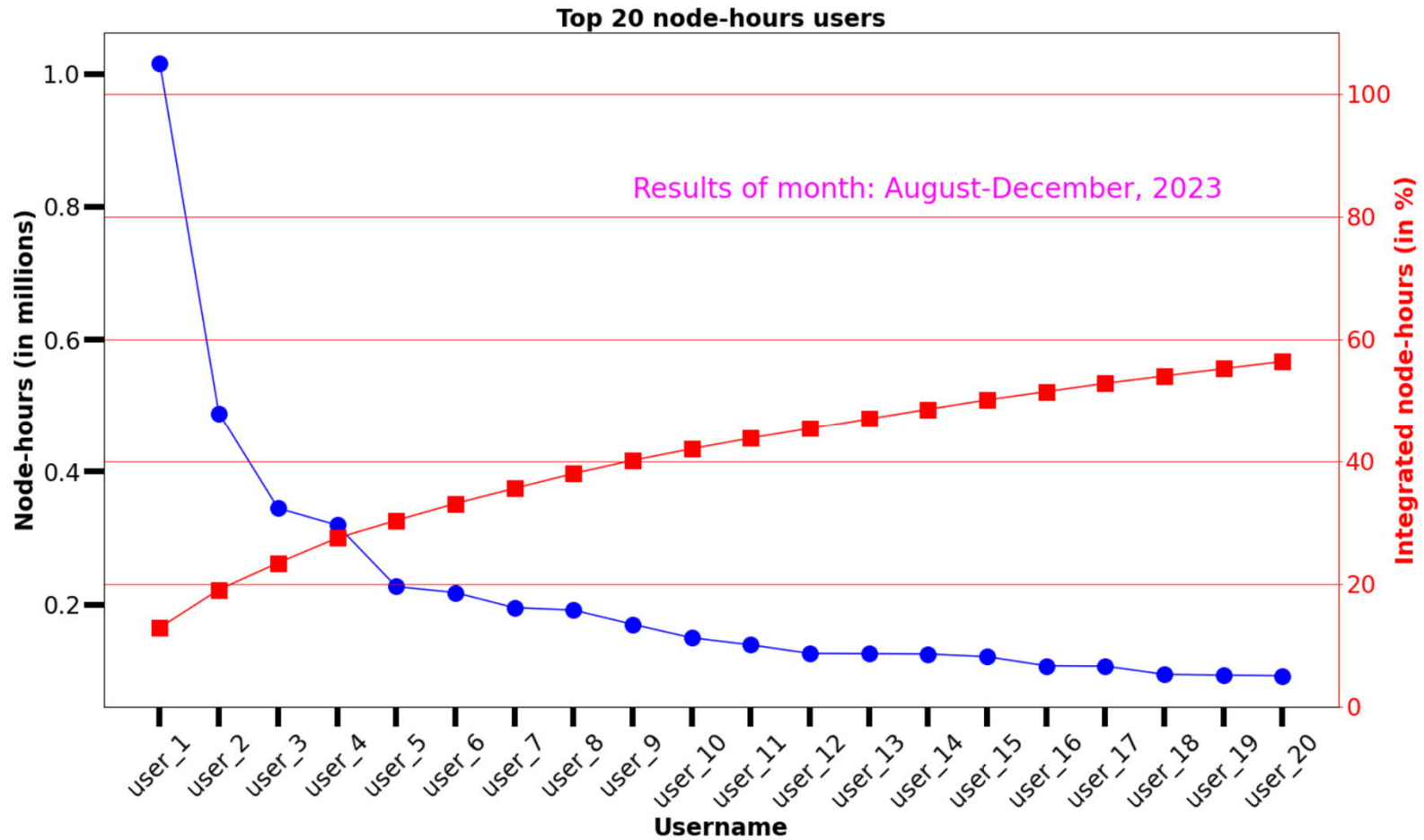
> **The top 15 users** in terms of node-hours **utilize ~50%** of the machine's capacity.

Hpcmd results – year 2023 (August-December)
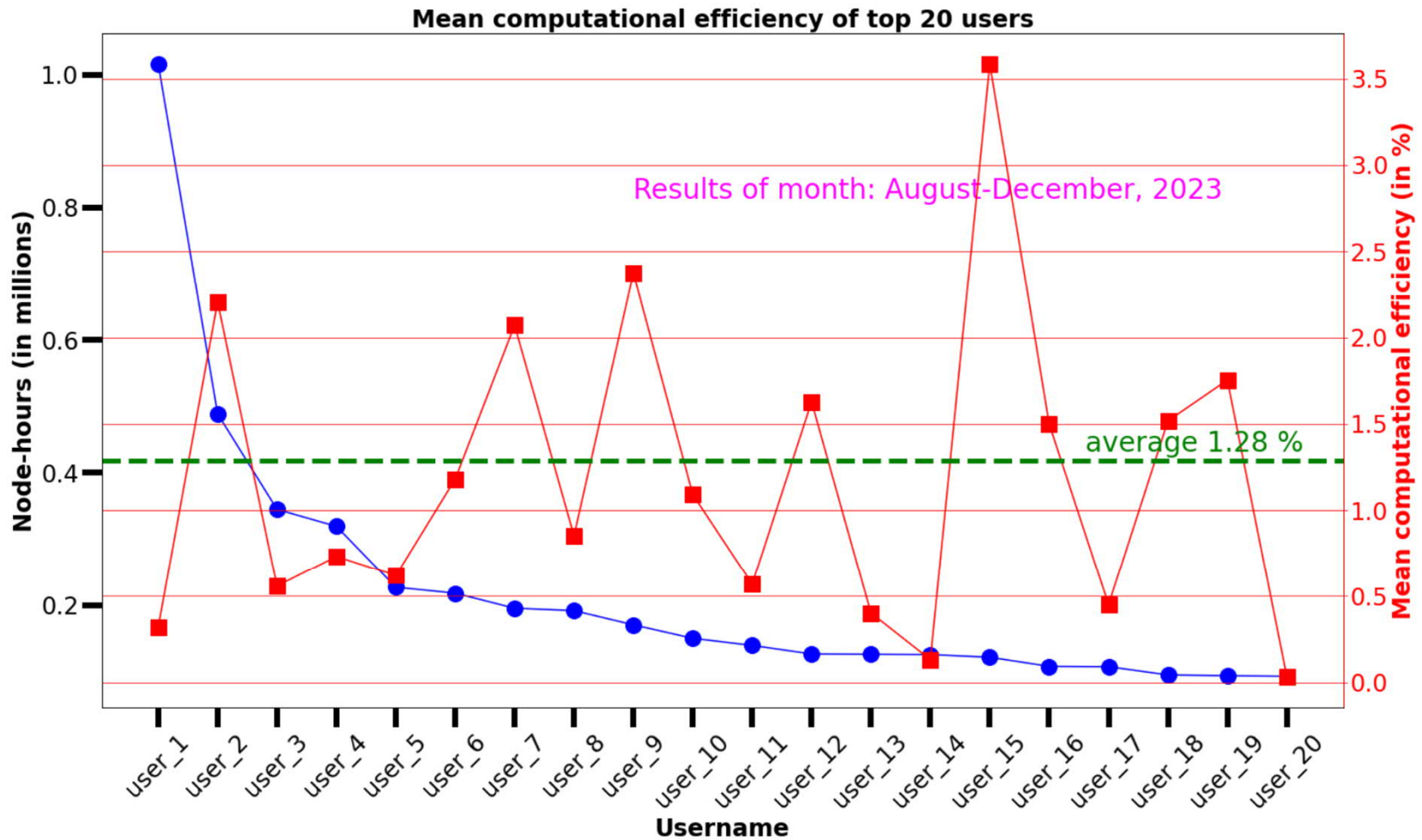
Mean computational efficiency of top 20 users

Results of month: August-December, 2023

average 1.28 %

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

Hpcmd results December 2023 − top 1 node-hours user

Histogram of used node-hours as a distribution of efficiency (top 1 user)

MEAN COMPUTATIONAL EFFICIENCY OF

JOBS OF # 1 USER = user_1:   **0.30 %**

Results of month: December, 2023

Histogram of used node-hours as a distribution of efficiency (top 2 user)

MEAN COMPUTATIONAL EFFICIENCY OF

JOBS OF # 2 USER = user_2:   **2.09 %**
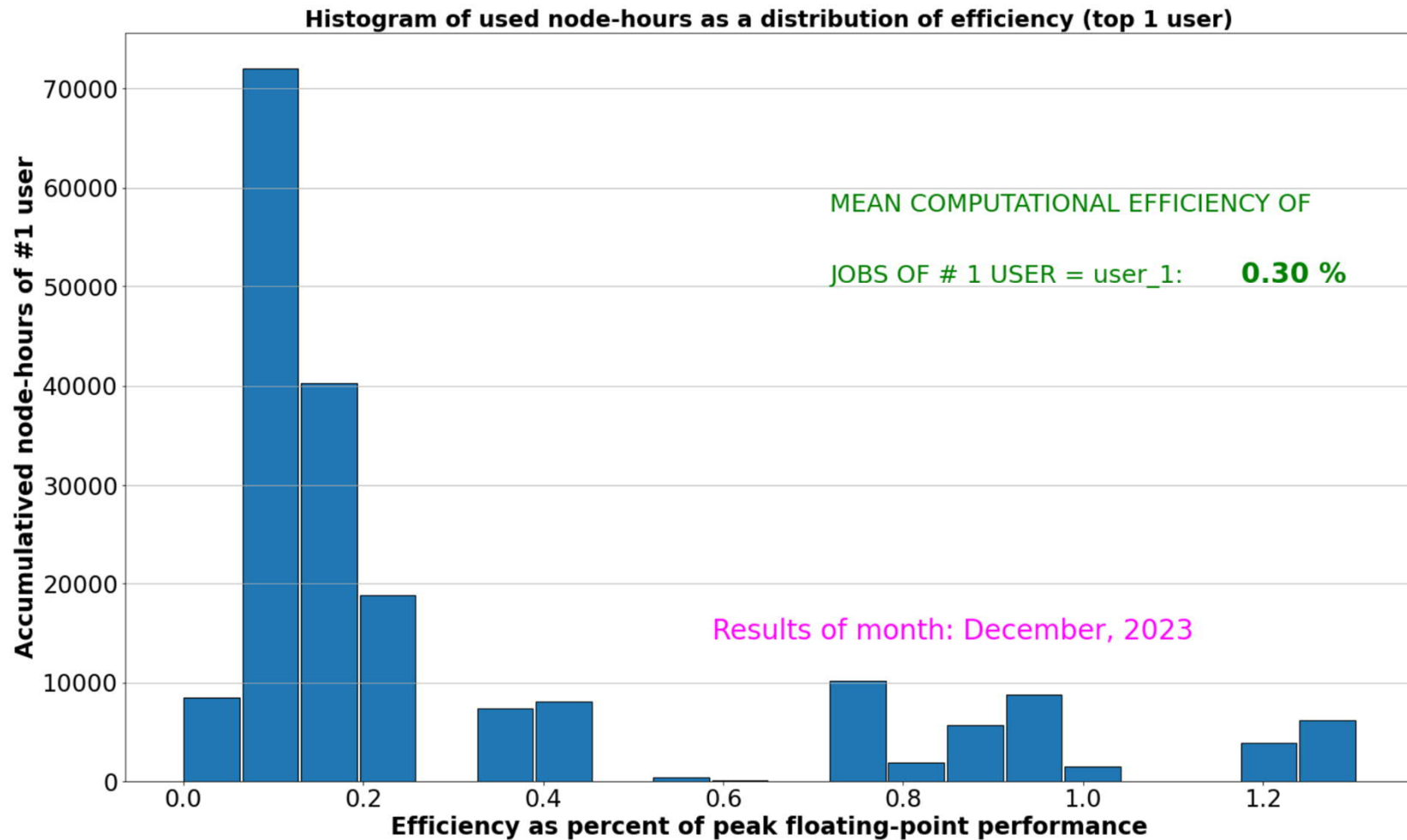
Results of month: December, 2023
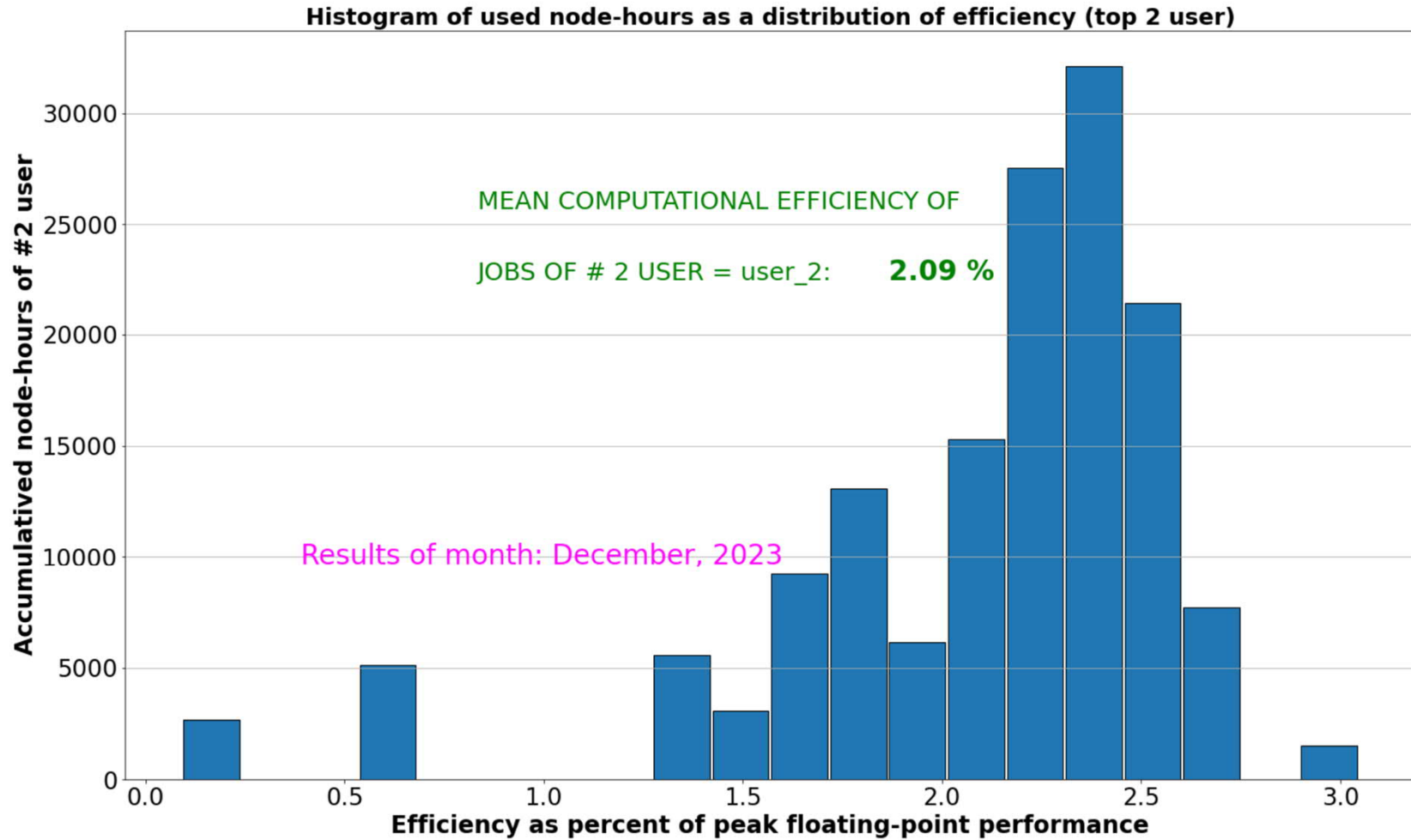
Hpcmd results December 2023 — top 1 node-hours user

> ➢ **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]:

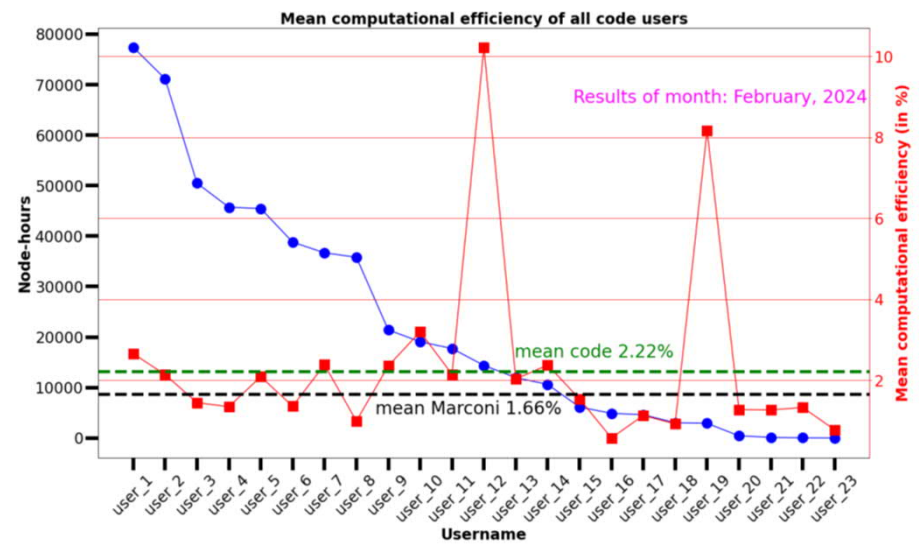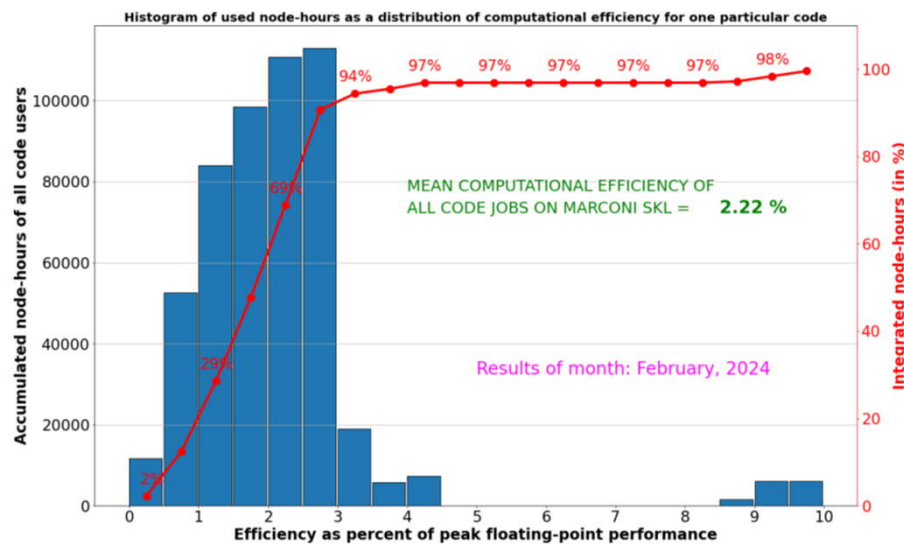| ne- es	cache- references	cpu	cycles	fp_128d	fp_128s	fp_256d	fp_256s	fp_512d	fp_512s	fp_d	fp_s	instructions	jobid
274	10520285280	S0	11407792328262	0	0	0	0	0	0	52551291029	125096718911	15423799040218	12718037
578	11259772726	S0	11409101485245	0	0	0	0	0	0	57335154957	138570819092	15445213386620	12718037
274	11908953318	S0	11405959628294	0	0	0	0	0	0	61059890286	144020040952	15429373716676	12718037
32	12225599152	S0	11407434390021	0	0	0	0	0	0	43481301021	143930550257	15395523151205	12718037
32	10496045040	S1	11409594719373	0	0	0	0	0	0	53021043098	125943706498	15537448163187	12718037
...	...	...	...	...	...	...	...	...	...	...	...	...	...
64	10128242643	S1	10392611616589	0	0	0	0	0	0	39232354636	126250394956	14224743717220	12718037
23	10649188819	S1	10393988827028	0	0	0	0	0	0	48028581985	156605977337	14242581796545	12718037
98	19043923945	S1	10382810764810	0	0	0	0	0	0	109510156117	353275861071	14413717614307	12718037
76	17207799909	S0	10386605729781	0	0	0	0	0	0	94835826500	308846497616	14335958968278	12718037
96	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

Histogram of used node-hours as a distribution of computational efficiency for one particular code

MEAN COMPUTATIONAL EFFICIENCY OF
ALL CODE JOBS ON MARCONI SKL = **2.22 %**

Results of month: February, 2024

Mean computational efficiency of all code users

Results of month: February, 2024

mean code 2.22%

mean Marconi 1.66%

> ➤ **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!

Histogram of used node-hours as a distribution of computational efficiency

MEAN COMPUTATIONAL EFFICIENCY OF JOBS ON MARCONI SKL = **1.19 %**

Results of month: December, 2023
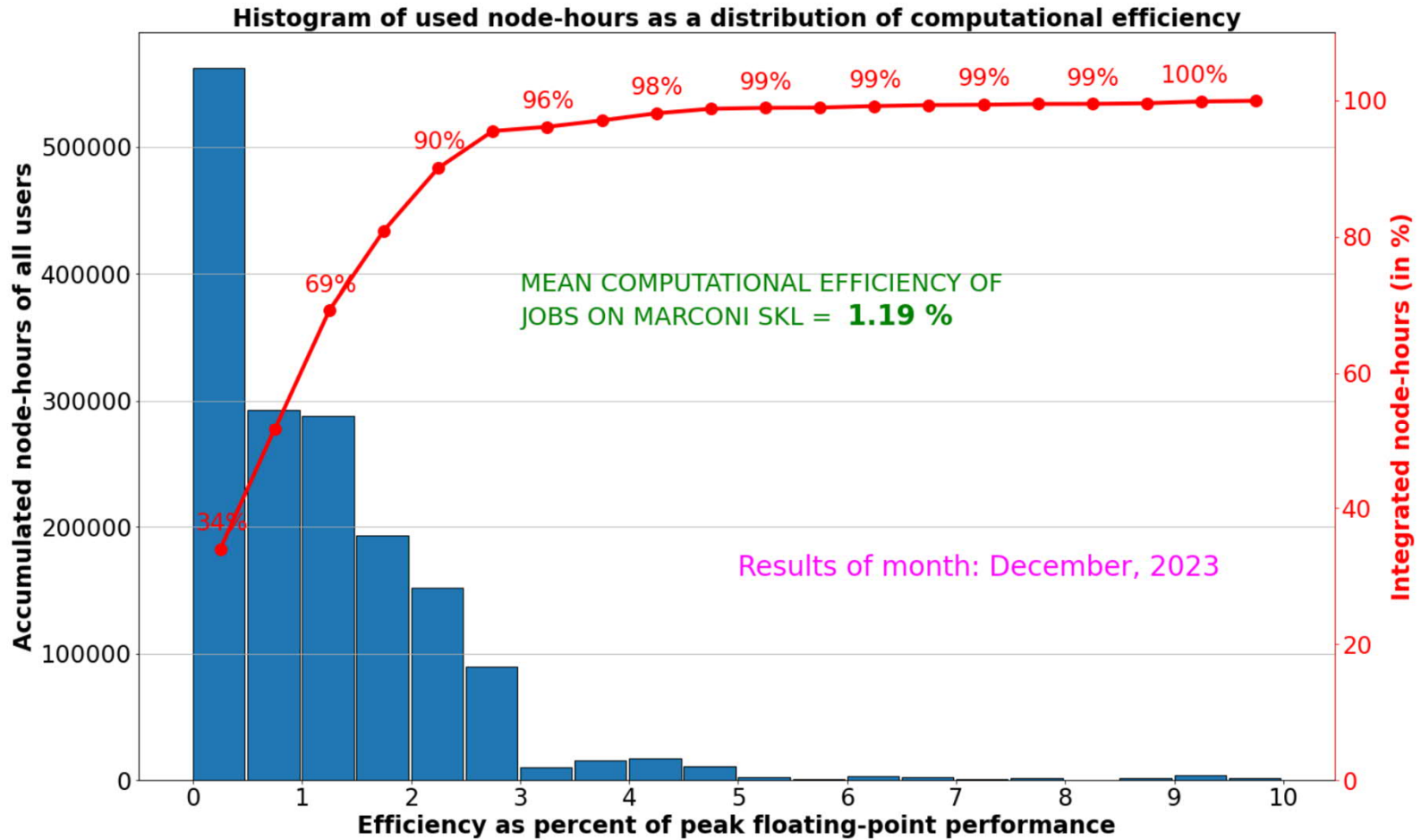
➢ **50% of supercomputer usage has computational efficiency below 1%.**

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

Mean computational efficiency of top 20 users

Results of month: December, 2023

average 1.19 %
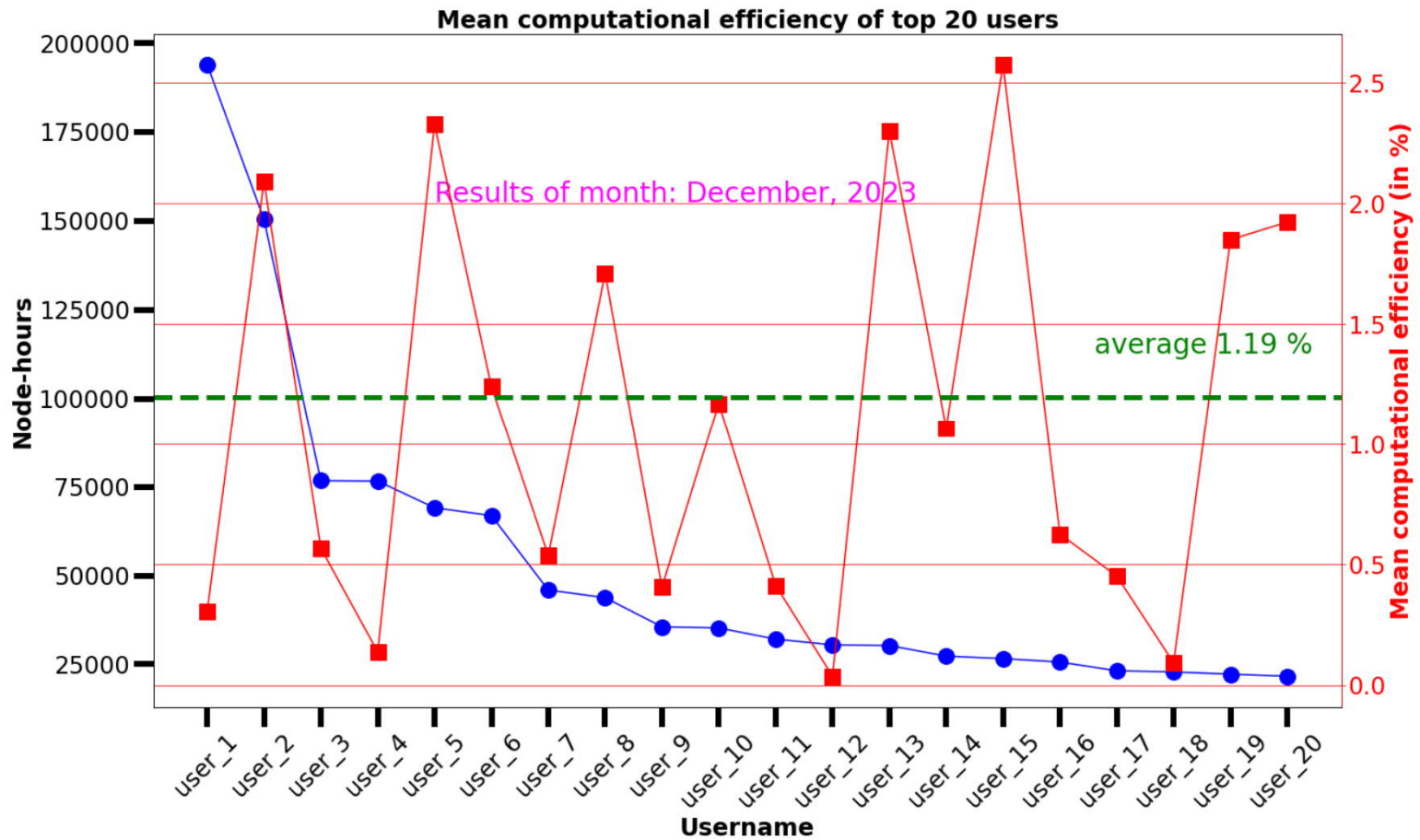
> ➤ **Out of the top 20 users, 10 have computational efficiency below 1%.**
> ➤ **Top 1 user has computational efficiency 0.3%.**

Hpcmd results December 2023 — top 20 computational efficiency jobs
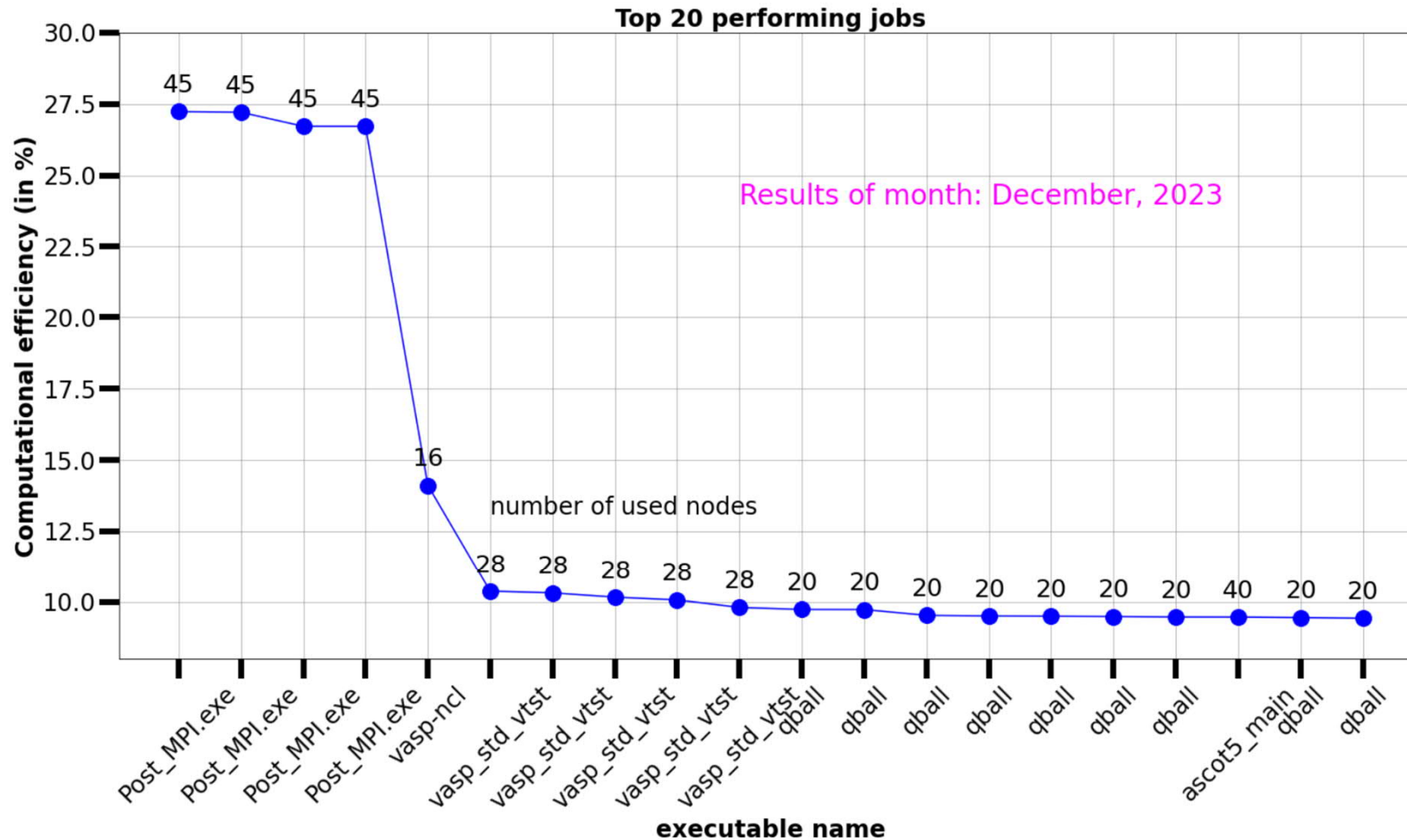
> ➢ **Many high-performance jobs involve the use of commercial libraries.**
> ➢ **The maximum efficiency goal is ~9%.**

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

Number of executed jobs of user: user_1

Results of month: December, 2023

Total number of jobs: 3059

execution time of each job during top day of top 1 user

Results of month: December, 2023

Total number of jobs of this day: 989