Comparative analysis of the performance and energy efficiency of JINR Tier-1 worker nodes.

Author: Aleksei Golunov (JINR, MLIT)

Co-author: Alexander Golunov (JINR, LHEP)

The 11th International Conference "Distributed Computing and Grid-technologies in Science and Education“
“GRID 2025”, Dubna

Goals

The main goals and objectives of the MLIT are the development and modernization of the MICC within the framework of the Large Infrastructure Project, which requires significant financial and organizational resources for the preparation, creation or reconstruction of the Basic MICC Installation.

These tasks are aimed at modernizing and developing main hardware and software components of the computing complex, creating a modern software platform that allows solving a wide range of research and applied problems, in accordance with the seven-year plan of JINR.

	2024	2025	2026	2027	2028	2029	2030
Tier-1 GRID site							
Performance HEPSPEC06	350000	400000	500000	550000	650000	750000	850000
Total CPU cores	22000	23000	30000	32000	38000	45000	50000

Tier-1 farm HEP-SPEC06 results Q1_2023 (published).

Hostname	Host type	nCPU	CPU model	Clock MHz	L2+L3 cache	RAM per host	RAM per core	HEP-SPEC06 per host	HEP-SPEC06 per core	Host count	Core Count SUM	HEP-SPEC06 SUM
wna000-wna083	Supermicro B1DRi	32	Intel E5-2683v4	2100	4M+40M	128GB	4GB	433.15	13.54	84	2688	36384.6
wna084-wna097	Supermicro SBI-4429P-T2N	32	Intel Gold 6226R	2900	16M+22M	384GB	12GB	610.00	19.10	14	448	8540
wna100-wna179	Supermicro SBA-4119S-T2N	64	AMD EPYC 7702P	2000	32M+256M	256GB	4GB	1100.00	17.19	80	5120	88000
wna276-wna415	Supermicro B1DRi	32	Intel E5-2683v4	2100	4M+40M	128GB	4GB	431.88	13.50	140	4480	60463.2
wna416-wna471	Supermicro SBI-4429P-T2N	32	Intel Gold 6226R	2900	16M+22M	384GB	12GB	610.00	19.10	56	1792	34160
wna472-wna485	Supermicro SBI-4429P-T2N	32	Intel Gold 6226R	2800	16M+22M	192GB	6GB	625.31	19.50	14	448	8754.34
wna486-wna565	Supermicro SBA-4119S-T2N	64	AMD EPYC 7702P	2000	32M+256M	256GB	4GB	1093.98	17.09	80	5120	87518.4
										468	20096	323,820.54

CentOS release 7.9
GCC: gcc (GCC) 4.4.7 20120313 (Red Hat 4.4.7-3)
C++: g++ (GCC) 4.4.7 20120313 (Red Hat 4.4.7-3)
FC: GNU Fortran (GCC) 4.4.7 20120313 (Red Hat 4.4.7-3)
*FLAGS: -O2 -pthread -fPIC -m32
SPECall_cpp2006 with 32-bit binaries.
SPEC2006 version 1.1

- All worker nodes optimized for running 16-core pilots jobs
- >50% of all CPU cores provided by AMD EPYC CPU powered servers
- For ~86% of all CPU cores available up to 4GB RAM per core

Tier-1 farm HEP-SPEC06 results Q1_2025 (not published).

Hostname	Host type	nCPU	CPU model	Clock MHz	L2+L3 cache	RAM per host	RAM per core	HEP-SPEC06 per host	HEP-SPEC06 per core	Host count	Core Count SUM	HEP-SPEC06 SUM
wna000-wna083	Supermicro B1DRi	32	Intel E5-2683v4	2100	4M+40M	128GB	4GB	433.15	13.54	84	2688	36384.6
wna084-wna097	Supermicro SBI-4429P-T2N	32	Intel Gold 6226R	2900	16M+22M	384GB	12GB	610.00	19.10	14	448	8540
wna100-wna179	Supermicro SBA-4119S-T2N	64	AMD EPYC 7702P	2000	32M+256M	256GB	4GB	1100.00	17.19	78	4992	85812.48
wna276-wna415	Supermicro B1DRi	32	Intel E5-2683v4	2100	4M+40M	128GB	4GB	431.88	13.50	140	4480	60463.2
wna416-wna471	Supermicro SBI-4429P-T2N	32	Intel Gold 6226R	2900	16M+22M	384GB	12GB	610.00	19.10	55	1760	33616
wna472-wna485	Supermicro SBI-4429P-T2N	32	Intel Gold 6226R	2800	16M+22M	192GB	6GB	625.31	19.50	13	416	8112
wna486-wna565	Supermicro SBA-4119S-T2N	64	AMD EPYC 7702P	2000	32M+256M	256GB	4GB	1093.98	17.09	78	4992	85313.28
										462	**19776**	**318,241.56**

- 6 servers with unrecoverable h/w issues (mostly motherboard).
- Hard to purchase the same server (EOL).
- Hard to initiate warranty replacement (“restrictions”).

1) To achieve pledges in 2025 we need additional resources (min):
- 23000 – 19776 = **3224 cores**
- 400000 - 318,241.56 = **81.758.44 HS06**

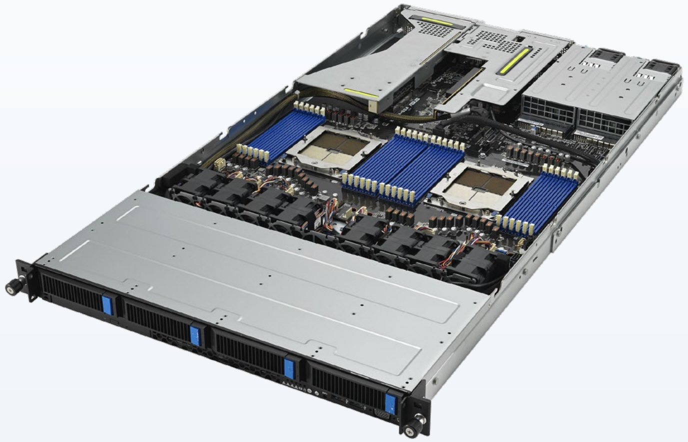
2) To keep power & cooling redundancy we must use <=~25KW additional power consumption for this task

3) To have possibility to share CPU resources with other experiments decided to have >4GB RAM per CPU core ratio (optimal 12GB/core)

AMD EPYC 9004 series CPU (only models with >=64 cores)

Model	Cores	Threads	Base freq. GHz	Max boost (GHz)	All-core boost (GHz)	Default TDP (W)	L3 cache (MB)	DDR5 channels	Max DDR5 freq. (MHz)	PCIE GEN 5 Lanes	CPU per-host	Optimal for 16-core pilots
9754	128	256	2.25	3.10	3.10	360	256	12	4800	128	2P/1P	Yes
9754S	128	128	2.25	3.10	3.10	360	256	12	4800	128	2P/1P	Yes
9734	112	224	2.20	3.00	3.00	340	256	12	4800	128	2P/1P	Yes
9654	96	192	2.40	3.70	3.55	360	384	12	4800	128	2P/1P	Yes
9654P	96	192	2.40	3.70	3.55	360	384	12	4800	128	1P	Yes
9684X (AMD 3D V-CACHE)	96	192	2.55	3.70	3.42	400	1152	12	4800	128	2P/1P	Yes
9634	84	168	2.25	3.70	3.10	290	384	12	4800	128	2P/1P	No
9554	64	128	3.10	3.75	3.75	360	256	12	4800	128	2P/1P	Yes
9554P	64	128	3.10	3.75	3.75	360	256	12	4800	128	1P	Yes
9534	64	128	2.45	3.70	3.55	280	256	12	4800	128	2P/1P	Yes

Tests & comparisons the same CPU in different server platforms (2024)



Test server Nerpa Nord D5015 (nerpa-lit.jinr-t1.ru)
Chassis: 1U ASUS RS700A-E12-RS4U
Cooling: 8 x high performance air fan (~28.5k RPM)
CPU: 2 x AMD EPYC 9654 (96C/192T, 2.4Ghz, 360W)
Memory: 24 x 64GB DDR5 ECC REG 4800Mhz
PSU: 2*2600W (N+1)
Storage: 3*1.92TB NVMe SSD
Network: 2*10Gb/s TP (built-in)
BMC: Aspeed AST2600

Test server Niagara R1208S (niagara-lit.jinr-t1.ru)
Chassis: 1U Supermicro AS -1125HS-TNR
Cooling: 8 x high performance air fan (~31k RPM)
CPU: 2 x AMD EPYC 96C, 2.4Ghz (96C/192T, 2.4Ghz, 360W)
Memory: 24 x 96GB DDR5 ECC REG 4800Mhz
PSU: 2*1600W (N+1)
Storage: 4*1.9TB NVMe SSD
Network: 2*10Gb/s TP OCP 3.0
BMC: Aspeed AST2600

AMD 9004 based servers HEP-SPEC06 test benchmark

Hostname	Host type	nCPU	CPU model	Clock MHz	L2+L3 cache	RAM per host	RAM per core	HEP-SPEC06 per host	HEP-SPEC06 per core	Power Consumption (peak)	CPU Temp (peak)	BIOS Performance profile
Nerpa-lit.jinr-t1.ru	ASUS RS700A-E12-RS4U	192	AMD EPYC 9654	2400	96M+384M	1536GB	8GB	4493	23.40	1300W	66 C	Default
Nerpa-lit.jinr-t1.ru	ASUS RS700A-E12-RS4U	192	AMD EPYC 9654	2400	96M+384M	1536GB	8GB	4618	24.05	1380W	74 C	Low latency
Nerpa-lit.jinr-t1.ru	ASUS RS700A-E12-RS4U	192	AMD EPYC 9654	2400	96M+384M	1536GB	8GB	5178	26.96	1600W	85 C	HPC
Niagara-lit.jinr-t1.ru	Supermicro AS -1125HS-TNR	192	AMD EPYC 9654	2400	96M+384M	2304GB	12GB	4032.56	21.00	1028W	59 C	Default
Niagara-lit.jinr-t1.ru	Supermicro AS -1125HS-TNR	192	AMD EPYC 9654	2400	96M+384M	2304GB	12GB	4474.32	23.30	1248W	65 C	HPM

- All tests were performed on the same rack in the same operating environment.
- Chassis fan speed: manual 100%
- Server Module air temperature sensors: 17C.
- Server chassis air inlet temperature: 18C.
- Server System Board sensor: 23C.

Nerpa Nord D5015 BIOS HPC Profile works properly if:

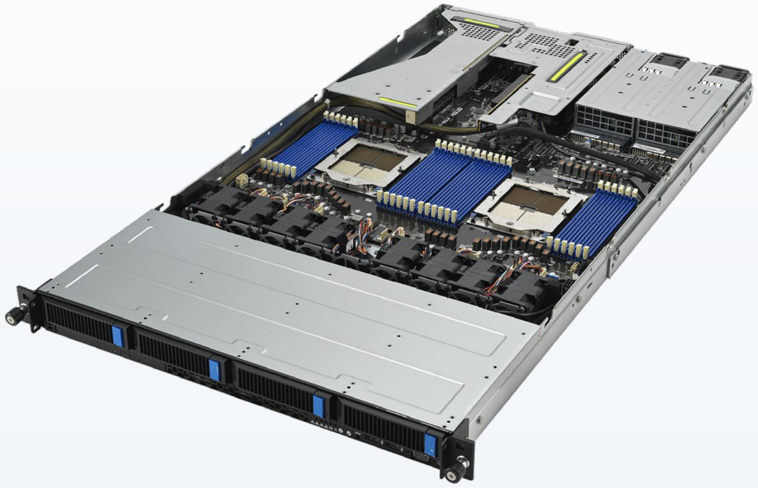
1) Discrete storage controller (NVMe is OK)
2) Server System Board sensor: <=25C

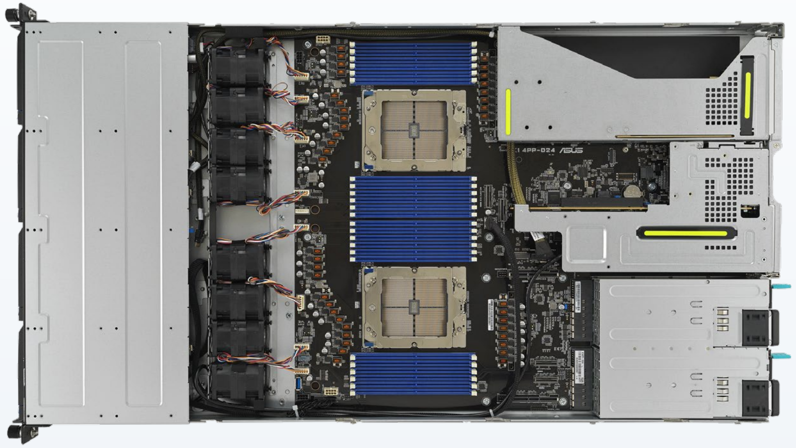
Nerpa is half the price, easier to buy, has better performance profiles.

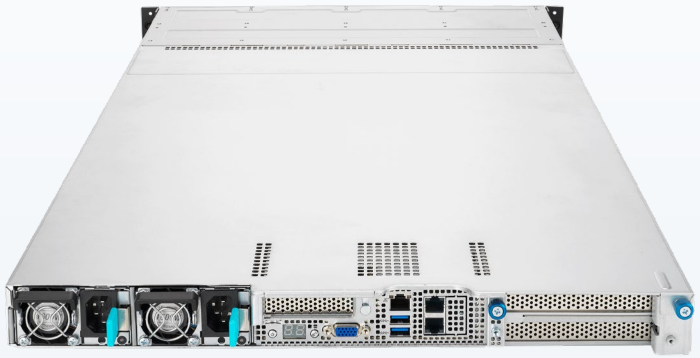
Minimal requirements for NERPA server count & performance:

1) Cores: 3224 / 192 = ~16.791(6) = 17
2) HEP-SPEC06 : (81758.44 - 4474.32) / 17 = 4,546.12

Test server Nerpa Nord D5015 (config v2, 2025)





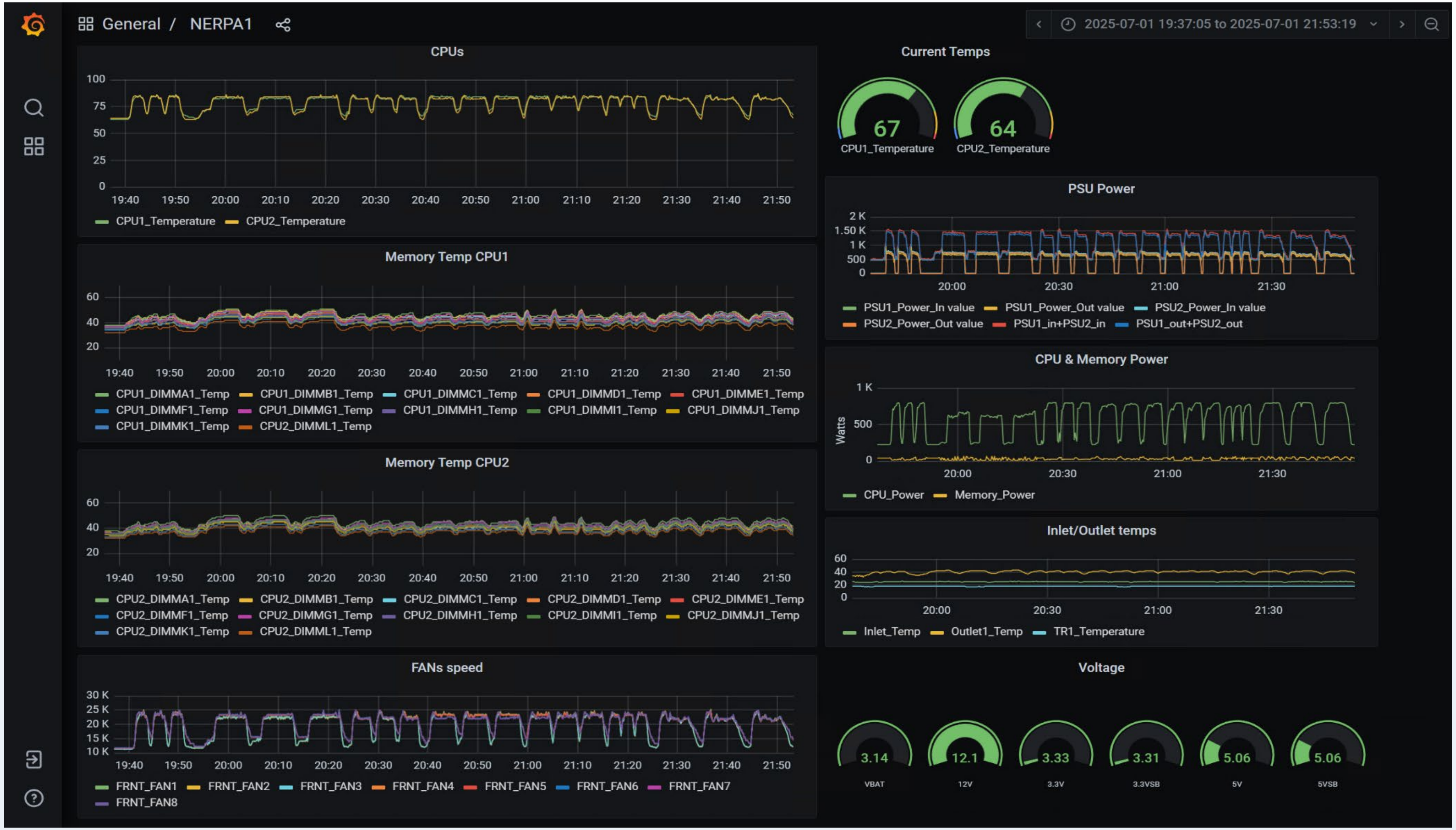


Chassis: 1U ASUS RS700A-E12-RS4U
CPU: 2 x AMD EPYC 96C, 2.4Ghz
Memory: 24 x 96GB 4800Mhz
PSU: 2*2600W (N+1)
Storage: 2*3.84TB NVMe SSD
Network: 2*10Gb/s TP (built-in)
BMC: Aspeed AST2600

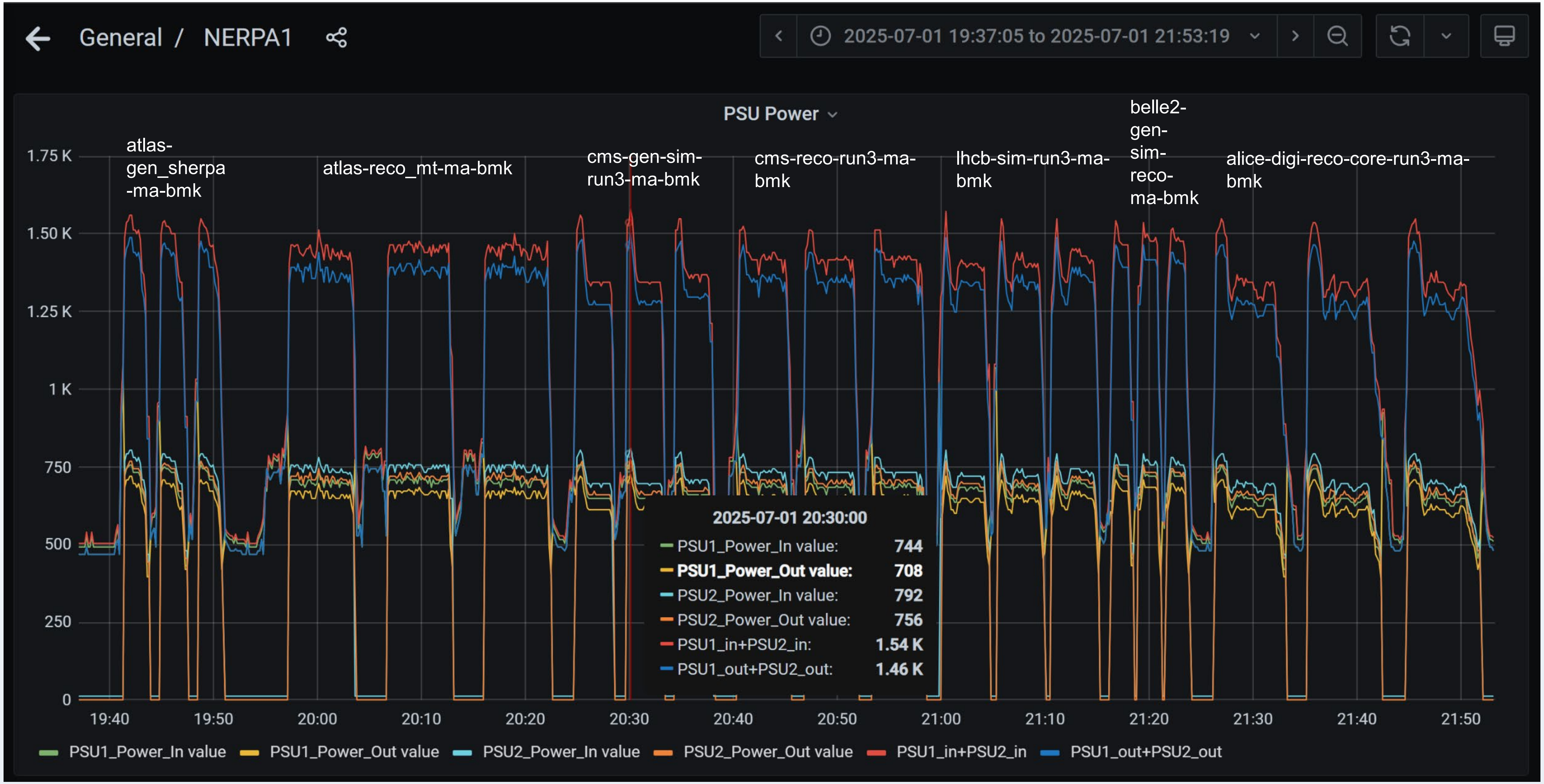
- **OS:** AlmaLinux 9.6 (5.14.0-570.24.1.el9_6.x86_64)
- **Benchmark:** HEPScore23 (HEP Workloads Project Apptainer + Python3)
https://w3.hepix.org/benchmarking/how_to_run_HS23.html
Includes:
• atlas-gen_sherpa-ma-bmk [v2.0_x86_64]
• atlas-reco_mt-ma-bmk [v2.0_x86_64]
• cms-gen-sim-run3-ma-bmk [v1.0_x86_64]
• cms-reco-run3-ma-bmk [v1.1_x86_64]
• lhcb-sim-run3-ma-bmk [v1.0_x86_64]
• belle2-gen-sim-reco-ma-bmk [v2.0_x86_64]
• alice-digi-reco-core-run3-ma-bmk [v2.1_x86_64]
- **Monitoring:** read BMC sensors remotely (Ipmitool –I –H …)

Create NERPA D5015 h/w monitoring instance on m4mon.jinr.ru

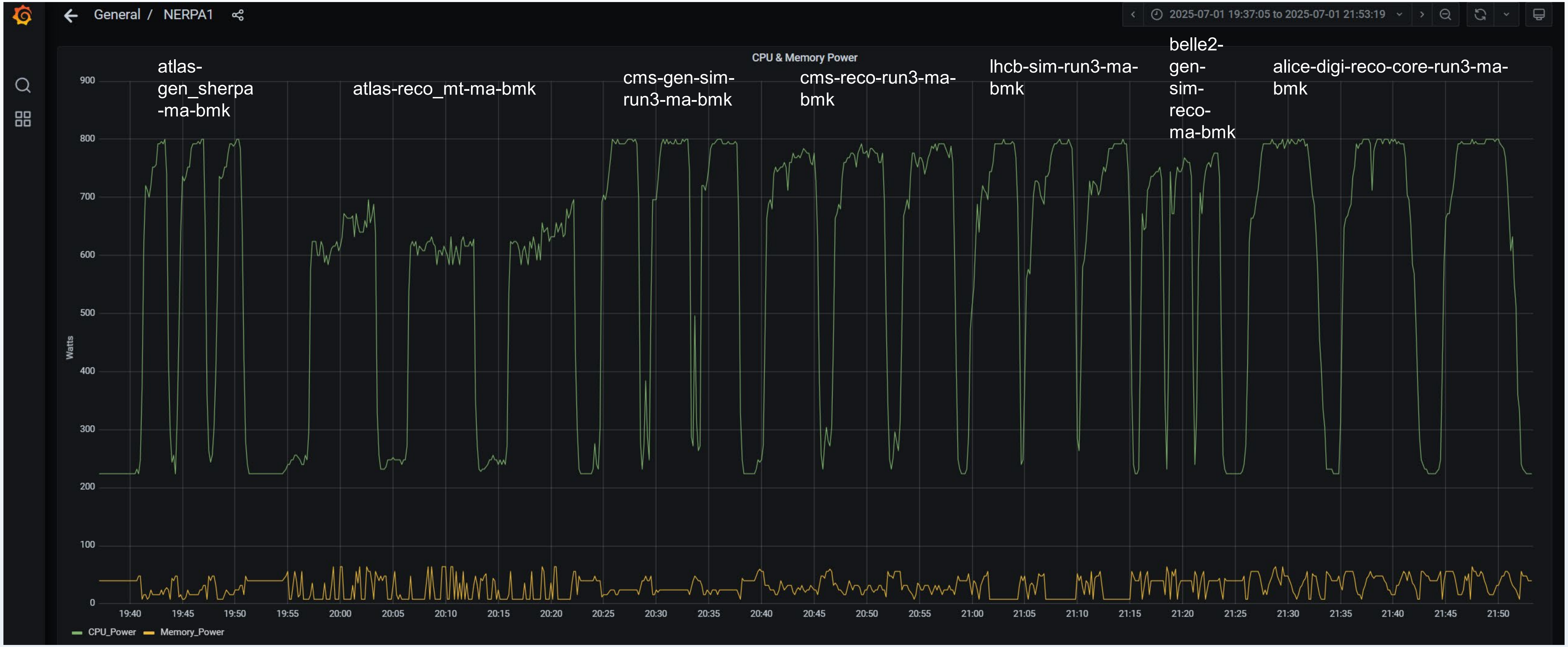
```
nerpa-lit:~ # ipmitool sdr
CPU1 Temperature             | 63 degrees C      | ok
CPU2 Temperature             | 64 degrees C      | ok
TR1 Temperature              | 18 degrees C      | ok
TR2 Temperature              | no reading        | ns
CPU1_DIMMA1_Temp             | 37 degrees C      | ok
CPU1_DIMMB1_Temp             | 37 degrees C      | ok
CPU1_DIMMC1_Temp             | 36 degrees C      | ok
CPU1_DIMMD1_Temp             | 36 degrees C      | ok
CPU1_DIMME1_Temp             | 35 degrees C      | ok
CPU1_DIMMF1_Temp             | 33 degrees C      | ok
CPU1_DIMMG1_Temp             | 36 degrees C      | ok
CPU1_DIMMH1_Temp             | 36 degrees C      | ok
CPU1_DIMMI1_Temp             | 35 degrees C      | ok
CPU1_DIMMJ1_Temp             | 34 degrees C      | ok
CPU1_DIMMK1_Temp             | 35 degrees C      | ok
CPU1_DIMML1_Temp             | 35 degrees C      | ok
CPU2_DIMMA1_Temp             | 37 degrees C      | ok
CPU2_DIMMB1_Temp             | 36 degrees C      | ok
CPU2_DIMMC1_Temp             | 35 degrees C      | ok
CPU2_DIMMD1_Temp             | 35 degrees C      | ok
CPU2_DIMME1_Temp             | 35 degrees C      | ok
CPU2_DIMMF1_Temp             | 34 degrees C      | ok
CPU2_DIMMG1_Temp             | 36 degrees C      | ok
CPU2_DIMMH1_Temp             | 36 degrees C      | ok
CPU2_DIMMI1_Temp             | 35 degrees C      | ok
CPU2_DIMMJ1_Temp             | 34 degrees C      | ok
CPU2_DIMMK1_Temp             | 33 degrees C      | ok
CPU2_DIMML1_Temp             | 32 degrees C      | ok
Inlet Temp                   | 24 degrees C      | ok
Outlet1 Temp                 | 34 degrees C      | ok
+VCORE0_CPU1                 | 1.12 Volts        | ok
+VCORE1_CPU1                 | 1.10 Volts        | ok
+VCORE0_CPU2                 | 1.12 Volts        | ok
+VCORE1_CPU2                 | 1.12 Volts        | ok
+VSOC_CPU1                   | 1.02 Volts        | ok
+VSOC_CPU2                   | 1.02 Volts        | ok
+VDDIO_CPU1                  | 1.15 Volts        | ok
+VDDIO_CPU2                  | 1.12 Volts        | ok
+12V                         | 12.12 Volts       | ok
+5V                          | 5.06 Volts        | ok
+3.3V                        | 3.33 Volts        | ok
+5VSB                        | 5.06 Volts        | ok
+3.3VSB                      | 3.31 Volts        | ok
VBAT                         | 3.14 Volts        | ok
FRNT_FAN1                    | 11440 RPM         | ok
FRNT_FAN2                    | 11310 RPM         | ok
FRNT_FAN3                    | 11440 RPM         | ok
FRNT_FAN4                    | 11700 RPM         | ok
FRNT_FAN5                    | 11440 RPM         | ok
FRNT_FAN6                    | 11700 RPM         | ok
FRNT_FAN7                    | 11570 RPM         | ok
FRNT_FAN8                    | 11570 RPM         | ok
REAR_FAN1                    | no reading        | ns
REAR_FAN2                    | no reading        | ns
PSU1 Power In                | 492 Watts         | ok
PSU1 Power Out               | 456 Watts         | ok
PSU2 Power In                | 12 Watts          | ok
PSU2 Power Out               | 0 Watts           | ok
CPU_Power                    | 224 Watts         | ok
Memory_Power                 | 40 Watts          | ok
```



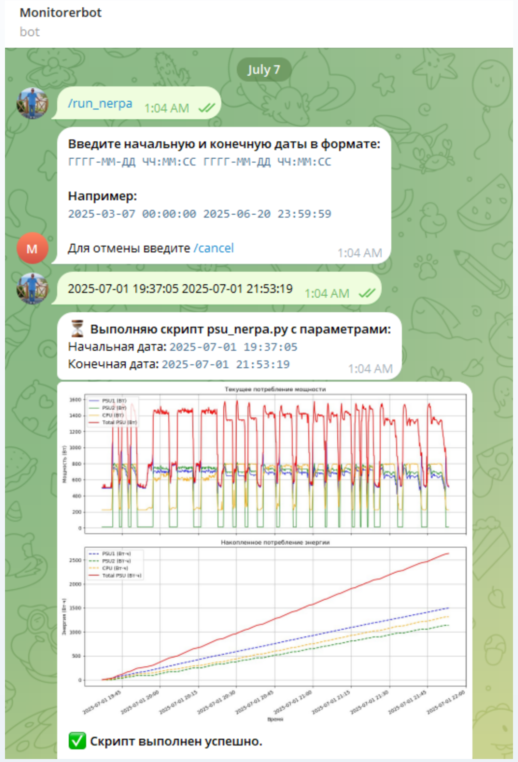
Test h/w monitoring instance m4mon.jinr.ru: PSU power

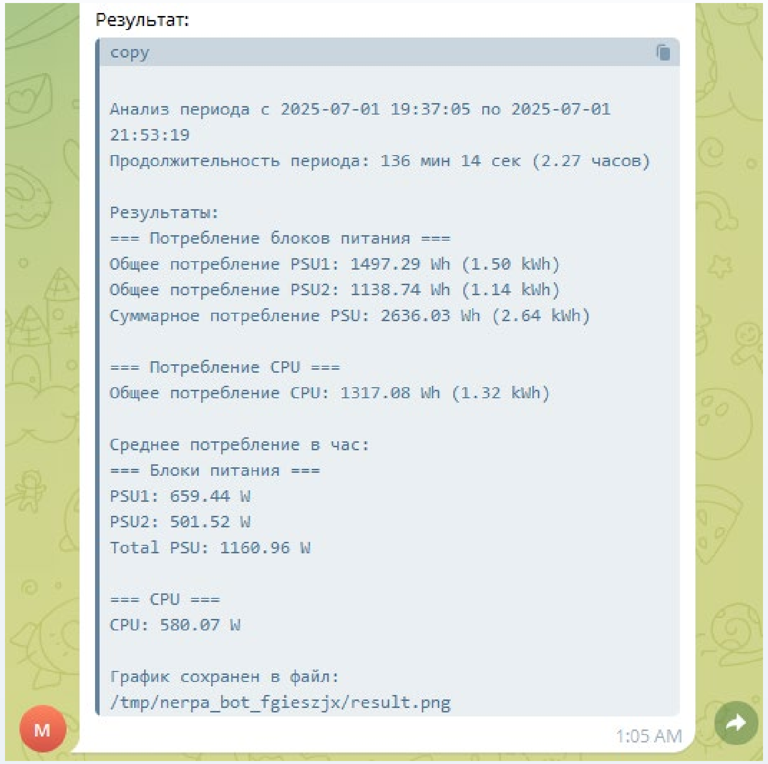


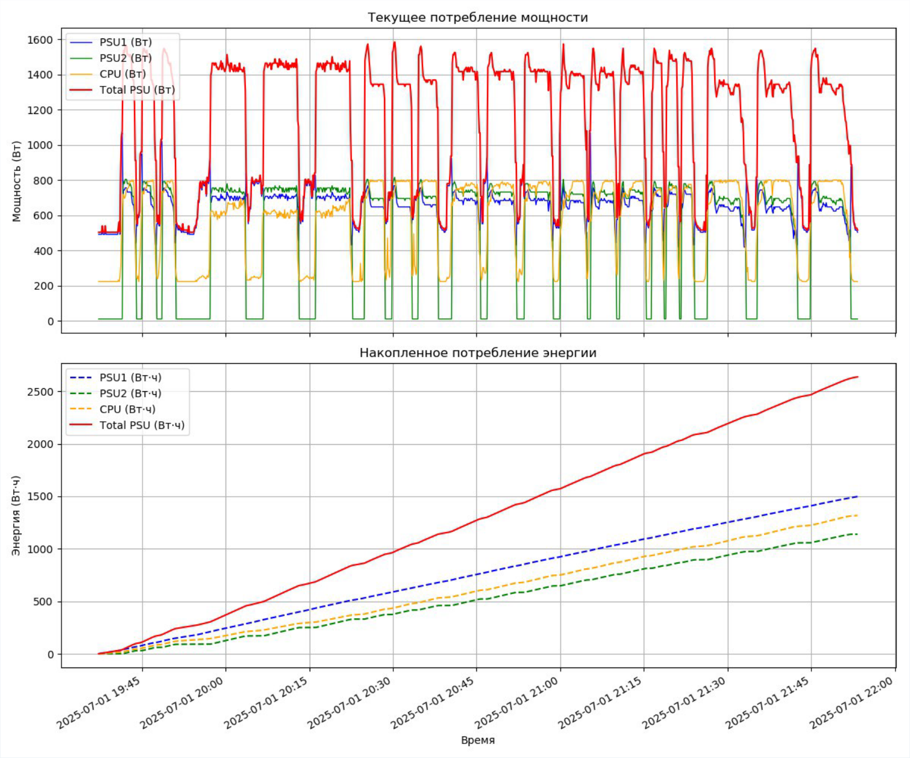
Test h/w monitoring instance m4mon.jinr.ru: CPU & RAM power



Implemented Telegram bot to collect results



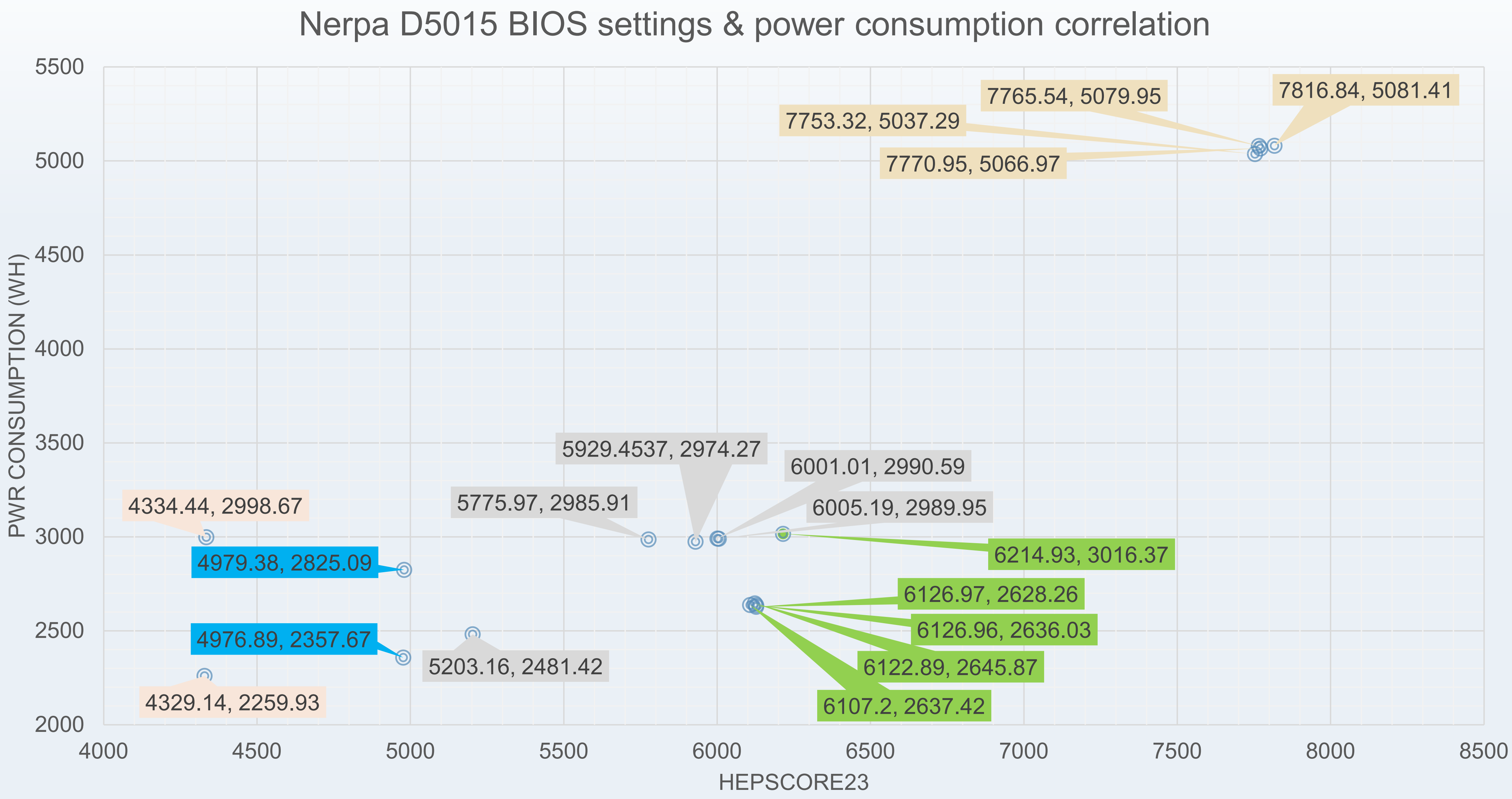




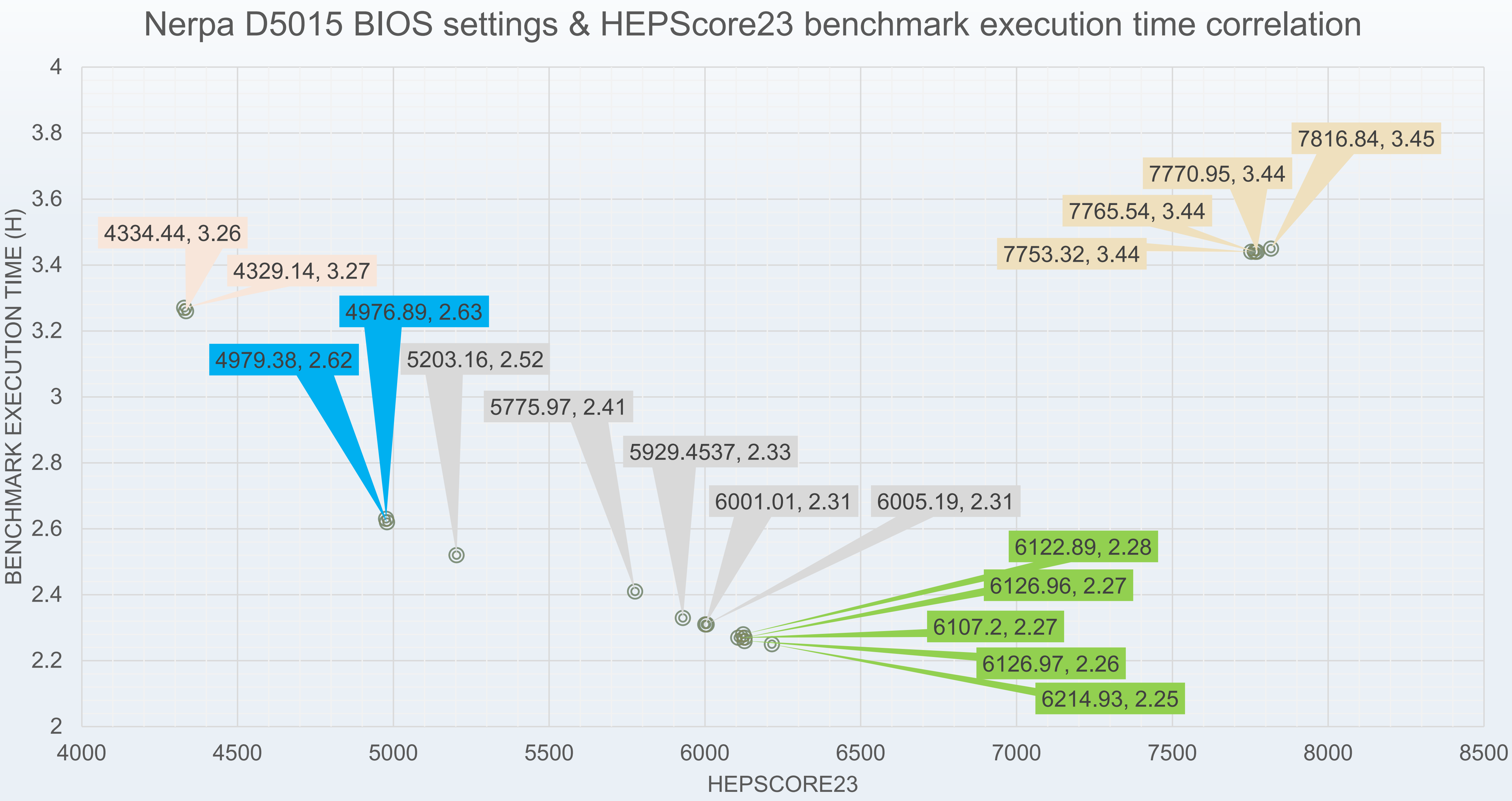
Input:
HS23 Suite start time &
HS23 Suite end time

Executing python3 script on monitoring server & send text result to Telegram

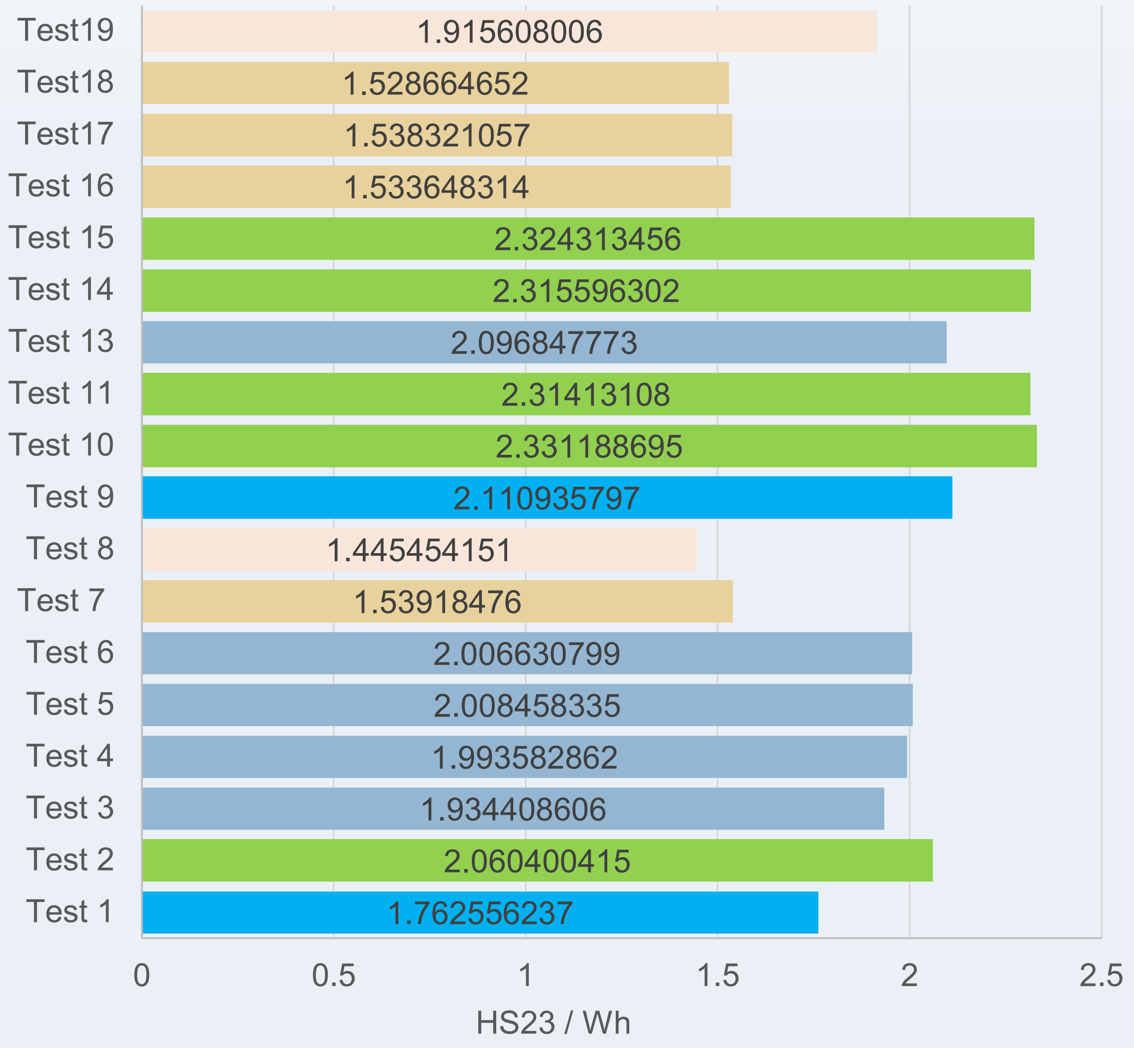
Executing Matplotlib script on monitoring server to plot graph & send .png result to Telegram



Nerpa D5015 BIOS settings & power consumption correlation
PWR CONSUMPTION (WH)
HEPSCORE23
5500
5000
4500
4000
3500
3000
2500
2000
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
7765.54, 5079.95
7816.84, 5081.41
7753.32, 5037.29
7770.95, 5066.97
5929.4537, 2974.27
6001.01, 2990.59
5775.97, 2985.91
6005.19, 2989.95
4334.44, 2998.67
6214.93, 3016.37
4979.38, 2825.09
6126.97, 2628.26
6126.96, 2636.03
4976.89, 2357.67
5203.16, 2481.42
6122.89, 2645.87
4329.14, 2259.93
6107.2, 2637.42


Nerpa D5015 BIOS settings & HEPScore23 benchmark execution time correlation
BENCHMARK EXECUTION TIME (H)
HEPSCORE23
4
3.8
3.6
3.4
3.2
3
2.8
2.6
2.4
2.2
2
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
4334.44, 3.26
4329.14, 3.27
4976.89, 2.63
4979.38, 2.62
5203.16, 2.52
5775.97, 2.41
5929.4537, 2.33
6001.01, 2.31
6005.19, 2.31
6122.89, 2.28
6126.96, 2.27
6107.2, 2.27
6126.97, 2.26
6214.93, 2.25
7816.84, 3.45
7770.95, 3.44
7765.54, 3.44
7753.32, 3.44

Nerpa D5015 BIOS settings & power consumption correlation



Test 19 Power Efficient profile, fan generic speed, SMT off

Test 16-18 HPC profile, fan full speed, SMT on

Test 14-15 HPC profile, fan generic speed, SMT off

Test 13 Latency Otimized profile, fan generic speed, SMT off

Test 12 (failed) No profile, manual 3600Mhz Engine boost, SMT off

Test 10-11 HPC profile, fan generic speed, SMT off

Test 9 Default profile, fan generic speed, SMT off

Test 8 Power Efficient profile, fan full speed, SMT off

Test 7 HPC profile, fan full speed, SMT on

Test 5-6 No profile, manual 3550Mhz Engine boost, fan full speed, SMT off

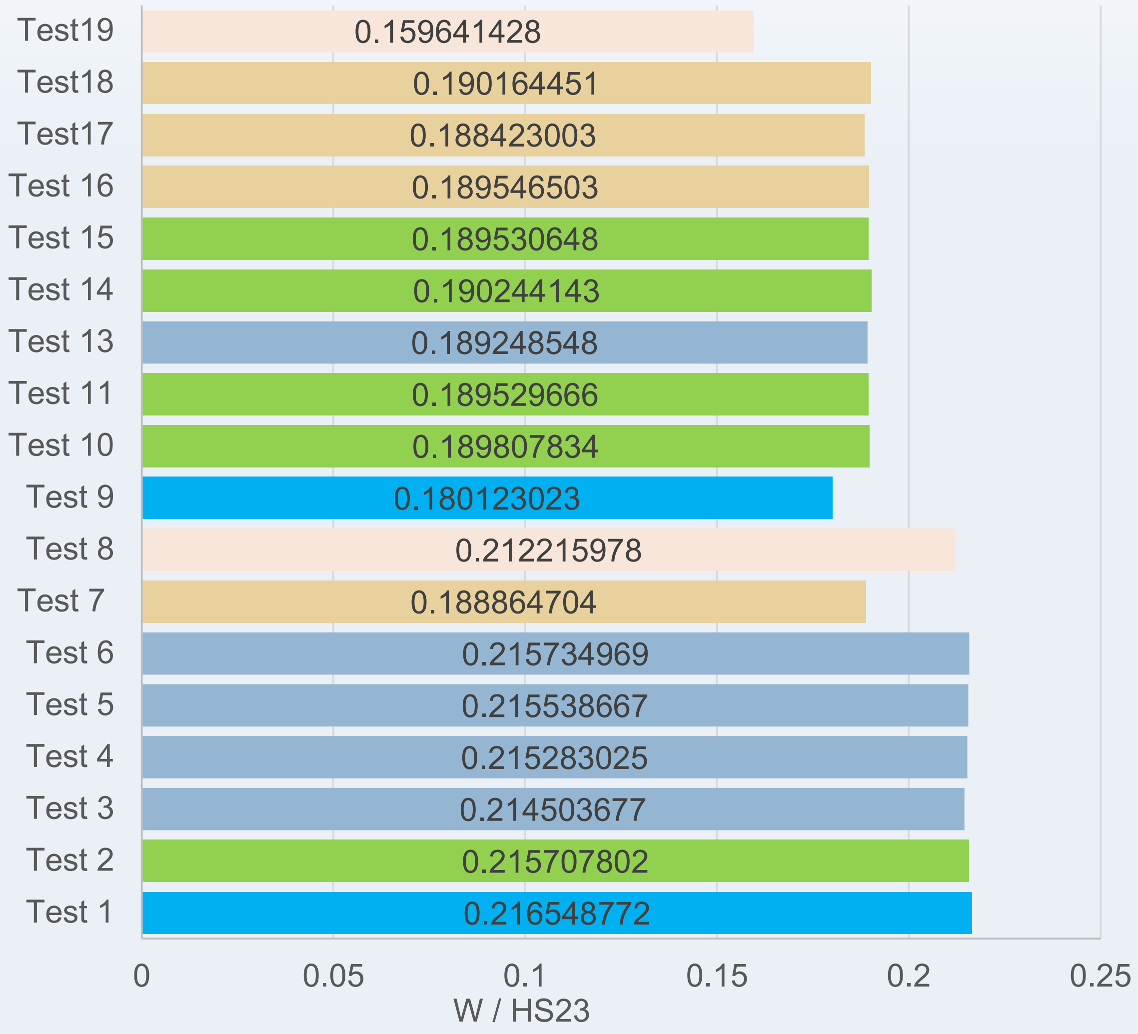
Test 4 No profile, manual 3500Mhz Engine boost, fan full speed, SMT off

Test 3 No profile, manual 3400Mhz Engine boost, fan full speed, SMT off

Test 2 HPC profile, fan full speed, SMT off

Test 1 Default profile, fan full speed, SMT off

Nerpa D5015 BIOS settings & power consumption correlation



Test 19 Power Efficient profile, fan generic speed, SMT off

Test 16-18 HPC profile, fan full speed, SMT on

Test 14-15 HPC profile, fan generic speed, SMT off

Test 13 Latency Otimized profile, fan generic speed, SMT off

Test 12 (failed) No profile, manual 3600Mhz Engine boost, SMT off

Test 10-11 HPC profile, fan generic speed, SMT off

Test 9 Default profile, fan generic speed, SMT off

Test 8 Power Efficient profile, fan full speed, SMT off

Test 7 HPC profile, fan full speed, SMT on

Test 5-6 No profile, manual 3550Mhz Engine boost, fan full speed, SMT off

Test 4 No profile, manual 3500Mhz Engine boost, fan full speed, SMT off

Test 3 No profile, manual 3400Mhz Engine boost, fan full speed, SMT off

Test 2 HPC profile, fan full speed, SMT off

Test 1 Default profile, fan full speed, SMT off

Conclusion

- Useful software tools were implemented to store, determine and compare server power consumption while running multicore software tasks (may be also useful for optimize parameters of NICA’s experiments tasks…)
- Found the way to achieve goals in 2025
- Determined the best BMC/BIOS configuration for our new worker nodes
- SMT still not efficient CPU mode for HEP jobs

Thank you for your attention!

Email: golunov@jinr.ru