HEPSCORE and power measurements with AMD Bergamo

Validating IPMI power measurements

Max Efficiency vs Max Performance profiles

Simon George (RHUL)

5 Feb 2025

Equipment

- Lenovo ThinkSystem SR645 V3
    - 2x AMD EPYC 9754 128-Core Processor (512 threads) – "Bergamo"
        - Base/Boost clock 2.25/3.1 GHz, 360W TDP
    - 2.3 TB memory (24 x 96 GB DDR5; 4.5 GB/thread)
    - 12 TB NVMe SSD
- Also tested a similar system with
    - 2x AMD EPYC 9654 (384 threads) – "Genoa"
        - Base/Boost clock 2.4/3.7 GHz, 360W TDP
- APC AP8853 PDU
    - 230V x 32A Zero-U PDU

ThinkSystem SR645 V3 AMD Power profiles

- UEFI provides selection between operating modes
- https://lenovopress.lenovo.com/lp1267-tuning-uefi-settings-for-performance-and-energy-efficiency-on-amd-servers
- **Maximum Efficiency**: Maximizes the performance / watt efficiency with a bias towards power savings.
- **Maximum Performance**: Maximizes the absolute performance of the system without regard for power savings. Most power savings features are disabled, and additional memory power / performance settings are exposed.
- **Custom Mode**: Allow user to customize the performance settings. Custom Mode will inherit the UEFI settings from the previous preset operating mode. For example, if the previous operating mode was the Maximum Performance operating mode and then Custom Mode was selected, all the settings from the Maximum Performance operating mode will be inherited.
- Observed differences in **CPU frequency** and consequently **HEPSCORE** and **power** use

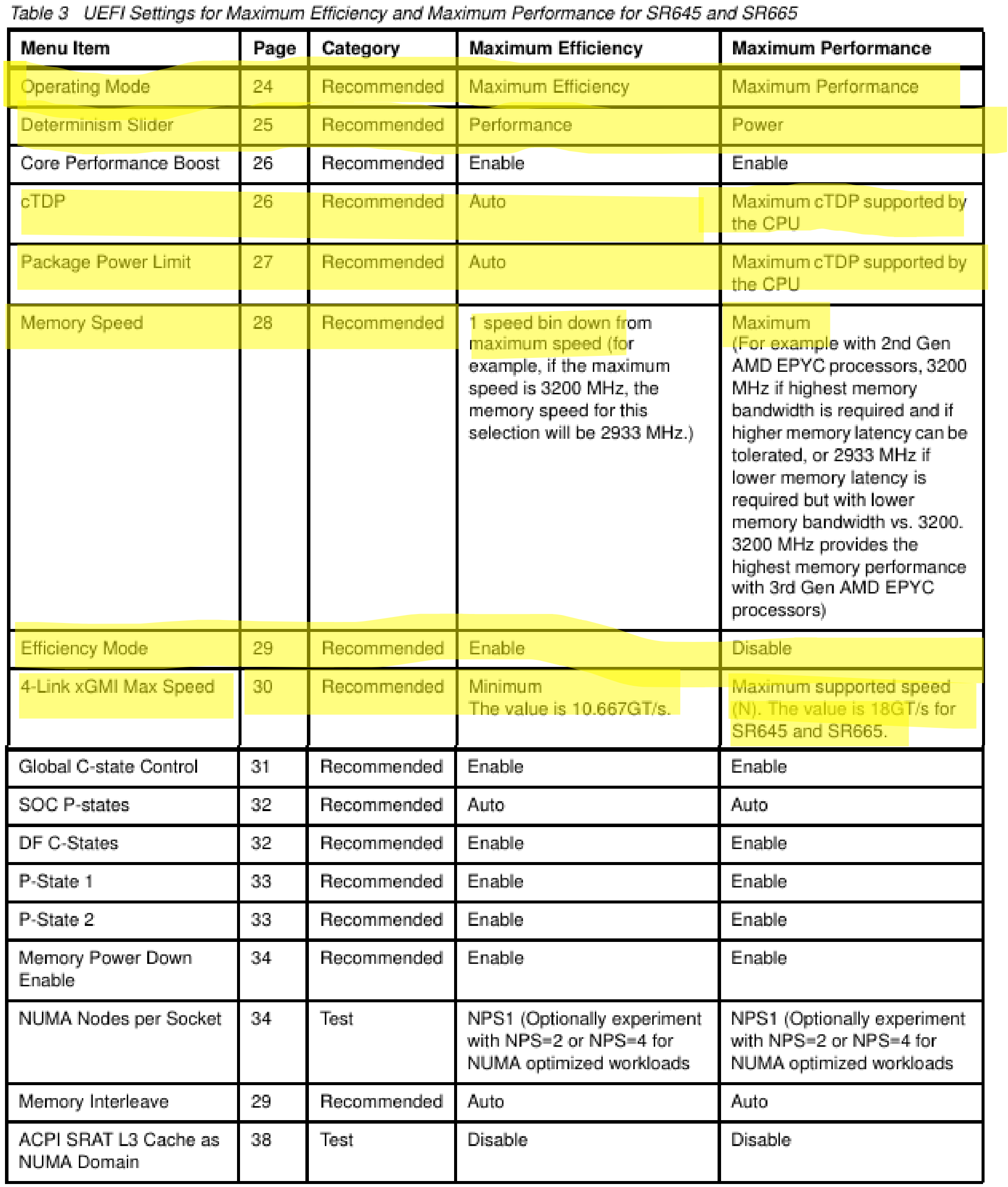
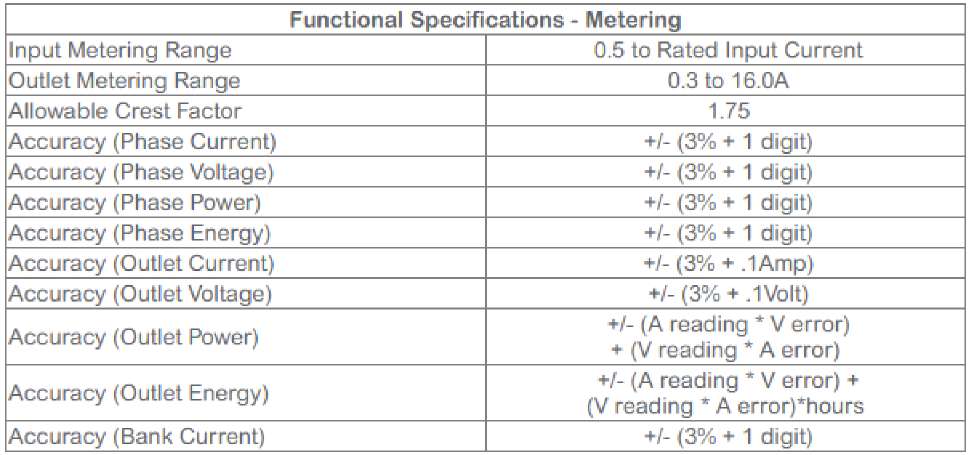


Table 3 UEFI Settings for Maximum Efficiency and Maximum Performance for SR645 and SR665

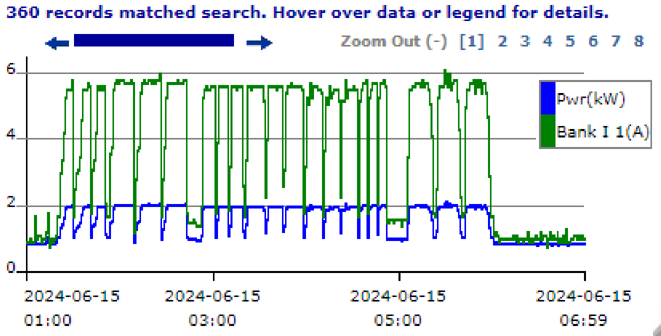
Menu Item	Page	Category	Maximum Efficiency	Maximum Performance
Operating Mode	24	Recommended	Maximum Efficiency	Maximum Performance
Determinism Slider	25	Recommended	Performance	Power
Core Performance Boost	26	Recommended	Enable	Enable
cTDP	26	Recommended	Auto	Maximum cTDP supported by the CPU
Package Power Limit	27	Recommended	Auto	Maximum cTDP supported by the CPU
Memory Speed	28	Recommended	1 speed bin down from maximum speed (for example, if the maximum speed is 3200 MHz, the memory speed for this selection will be 2933 MHz.)	Maximum (For example with 2nd Gen AMD EPYC processors, 3200 MHz if highest memory bandwidth is required and if higher memory latency can be tolerated, or 2933 MHz if lower memory latency is required but with lower memory bandwidth vs. 3200. 3200 MHz provides the highest memory performance with 3rd Gen AMD EPYC processors)
Efficiency Mode	29	Recommended	Enable	Disable
4-Link xGMI Max Speed	30	Recommended	Minimum The value is 10.667GT/s.	Maximum supported speed (N). The value is 18GT/s for SR645 and SR665.
Global C-state Control	31	Recommended	Enable	Enable
SOC P-states	32	Recommended	Auto	Auto
DF C-States	32	Recommended	Enable	Enable
P-State 1	33	Recommended	Enable	Enable
P-State 2	33	Recommended	Enable	Enable
Memory Power Down Enable	34	Recommended	Enable	Enable
NUMA Nodes per Socket	34	Test	NPS1 (Optionally experiment with NPS=2 or NPS=4 for NUMA optimized workloads	NPS1 (Optionally experiment with NPS=2 or NPS=4 for NUMA optimized workloads
Memory Interleave	29	Recommended	Auto	Auto
ACPI SRAT L3 Cache as NUMA Domain	38	Test	Disable	Disable

APC AP8853 PDU

- Monitoring capability
  - 2 banks, each with 21 sockets max 16A
  - Per-bank power monitoring (not per socket)
  - Test server was the only connection to bank 1
  - Configurable frequency of data logging, set to 1 minute
- Accuracy
  - "Rack PDUs (AP8XXX) have an accuracy of +/- 3% of reading, +/- 1 least significant digit across the entire power and temperature range."
  - "Note: Accuracy is not defined below 0.5A."
  - https://www.apc.com/uk/en/faqs/FA156074/
  - **Not clear if this is a statistical or systematic accuracy**
- Format and units
  - Download text file at end of test
  - Needs a little clean up before importing as a table e.g. into python/pandas
  - Actually measures/reports current (A) but it also reports the input voltage (241.0V) and the overall PDU load (kW) and current (A) vs time so I can convert to power drawn in W.
    - 241 +/- 1 W from 281 measurements



Functional Specifications - Metering	
Input Metering Range	0.5 to Rated Input Current
Outlet Metering Range	0.3 to 16.0A
Allowable Crest Factor	1.75
Accuracy (Phase Current)	+/- (3% + 1 digit)
Accuracy (Phase Voltage)	+/- (3% + 1 digit)
Accuracy (Phase Power)	+/- (3% + 1 digit)
Accuracy (Phase Energy)	+/- (3% + 1 digit)
Accuracy (Outlet Current)	+/- (3% + .1Amp)
Accuracy (Outlet Voltage)	+/- (3% + .1Volt)
Accuracy (Outlet Power)	+/- (A reading * V error) + (V reading * A error)
Accuracy (Outlet Energy)	+/- (A reading * V error) + (V reading * A error)*hours
Accuracy (Bank Current)	+/- (3% + 1 digit)



360 records matched search. Hover over data or legend for details.

Commands used to run benchmark

- Download and run

```
wget -O run_HEPscore.sh https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite/-/raw/3.0rc11/examples/hepscore/run_HEPscore.sh?ref_type=tags
chmod +x run_HEPscore.sh
./run_HEPscore.sh -v 3.0rc11 -b 'f,l,m,s,p' -s UKI-LT2-RHUL
```

- This collects power data by periodically running

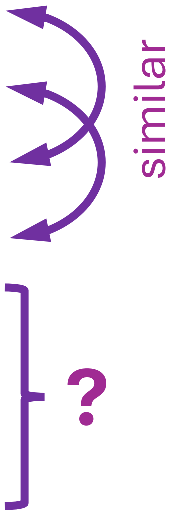
```
ipmitool dcmi power reading
```

Results

Host	AMD CPU model	timestamp	Profile	HEPSCORE	CPU freq (GHz) q75	CPU load q75	Power (W) q75	W/HS
node213	9654	2024-05-08T14:03	Efficiency	5807	2.60	384	1003	0.173
node213	9654	2024-05-03T10:57	Efficiency	5811	2.60	384	1006	0.173
node213	**9654**	**2024-05-02T16:15**	**Performance**	**7192**	**3.69**	**384**	**1326**	**0.184**
node221	9754	2025-01-27T22:30	Efficiency	6978	2.45	513	1022	0.146
node221	9754	2025-01-28T09:57	Efficiency	6967	2.45	513	1022	0.147
node214	9754	2024-05-16T09:39	Efficiency	7050	2.45	512	1015	0.144
node221	**9754**	**2025-01-27T17:21**	**Performance**	**8130**	**3.10**	**512**	**1306**	**0.161**
node221	**9754**	**2024-06-15T01:17**	**Performance**	**8227**	**3.10**	**512**	**1309**	**0.159**
node214	**9754**	**2024-05-17T11:30**	**Performance**	**8341**	**3.10**	**512**	**1351**	**0.162**

Comparison with published data

Site	AMD CPU model	SMT/Ncores/RAM	# Meas	Profile	HEPSCORE (mean)
UKI-LT2-RHUL	9654	Enabled/384/1.5TiB	2	Efficiency	5809
UKI-LT2-RHUL	9654	Enabled/384/1.5TiB	1	Performance	7192
IHEP	9654	Enabled/384/1TiB	26	?	6001
JP-KEK-CRC-02	9654	Enabled/384/820GiB	3	?	7268
UKI-LT2-RHUL	9754	Enabled/512/2.1TiB	3	Efficiency	6998
UKI-LT2-RHUL	9754	Enabled/512/2.1TiB	3	Performance	8232
UKI-SCOTGRID-GLASGOW	9754	Enabled/512/1TiB	5	?	7450

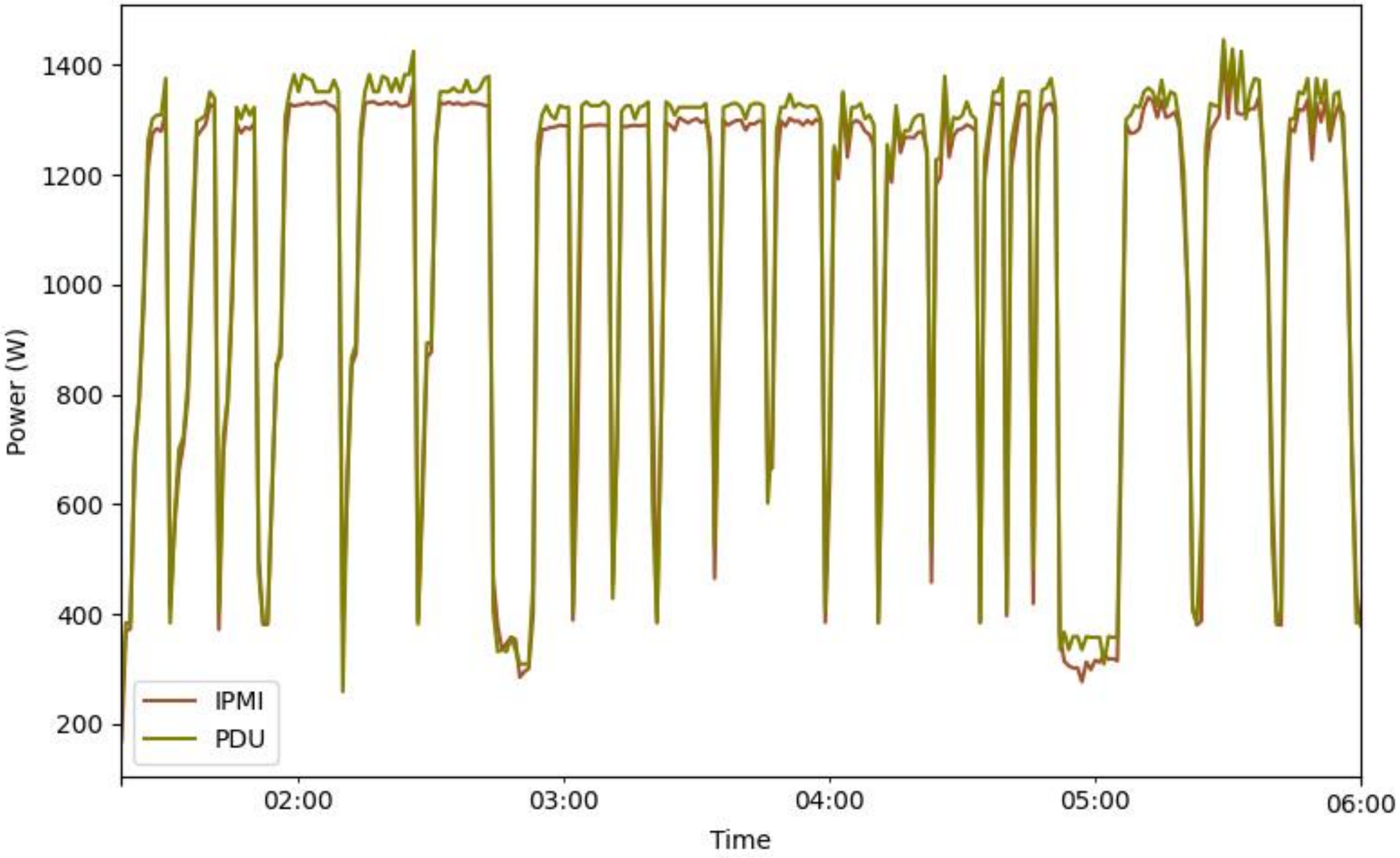


Power analysis

- Data from PDU and IPMI recorded every minute
- Checked clocks are correct on PDU and IPMI (ntp)
- PDU measurement at 44 secs past each minute
- IPMI measurement at 36 secs past each minute
- These points were aligned to the same minute for plotting
- Might expect a small lag in IPMI data during changes in power, but changes are so fast we don’t really see that, and the steady state power when each benchmark is running should not be affected
- PDU bank 1 *current* transformed to *power*, using voltage calculated from total PDU power/current at the same time

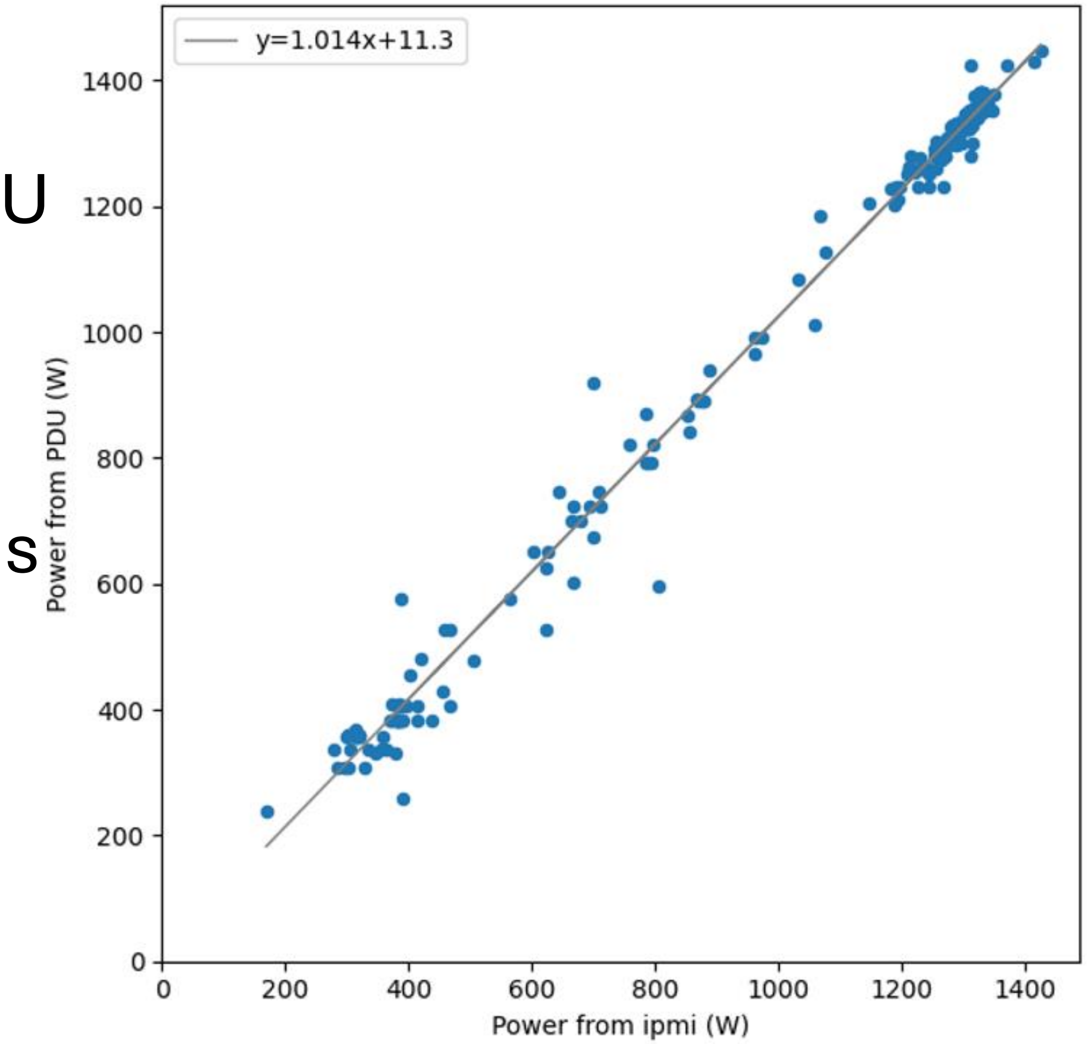
Power measurements PDU vs IPMI

- Taken with 2x AMD 9754 (node221)
- Each benchmark run is clear
- Excellent correlation between PDU and IPMI
  - Large rises and fall match
  - PDU systematically slightly higher than IPMI in steady state



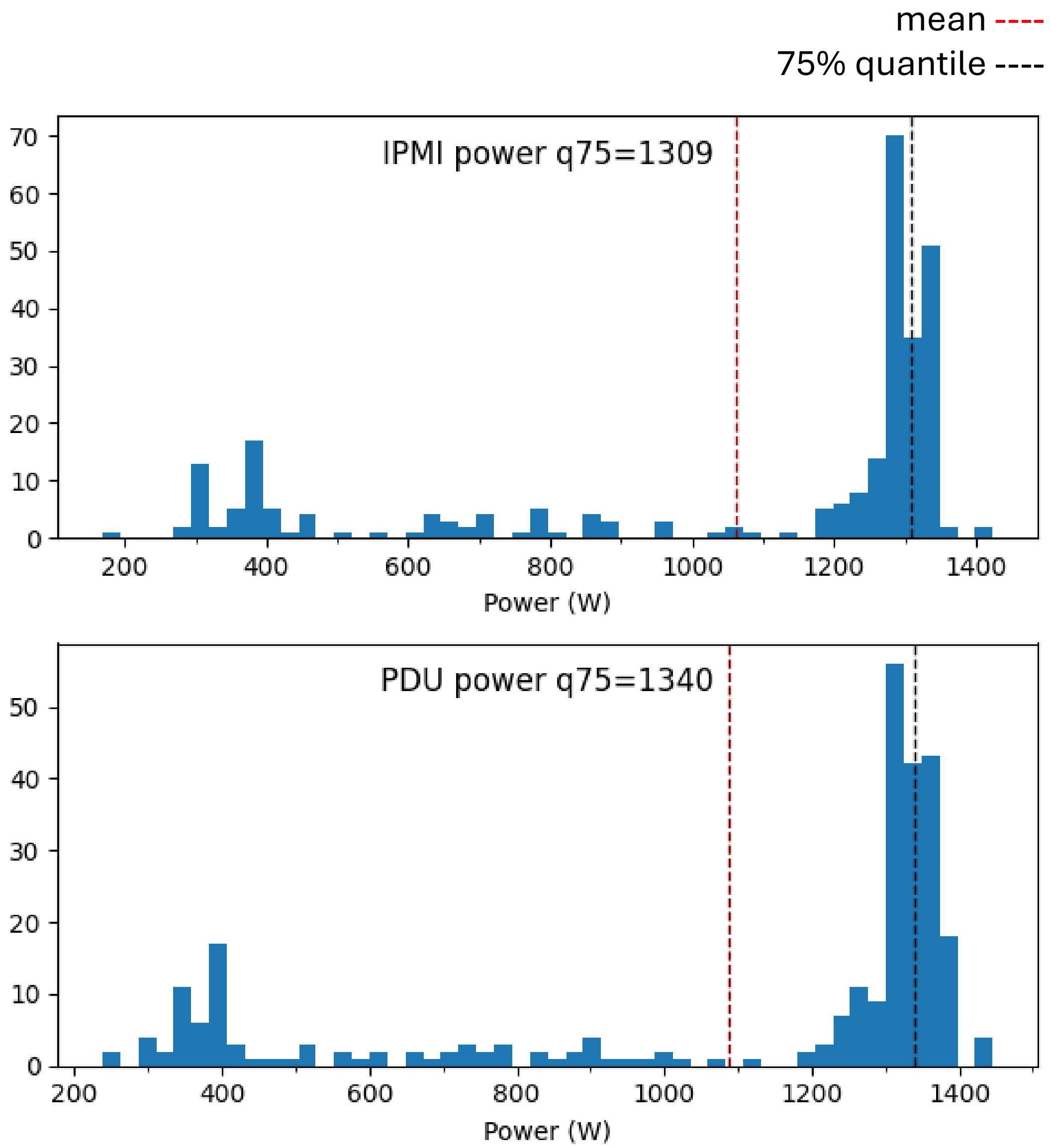
Correlation

- Good correlation between PDU and IPMI, especially in the most relevant area of high power use
- Mean of each PDU/IPMI ratio is 1.025



Power spectra

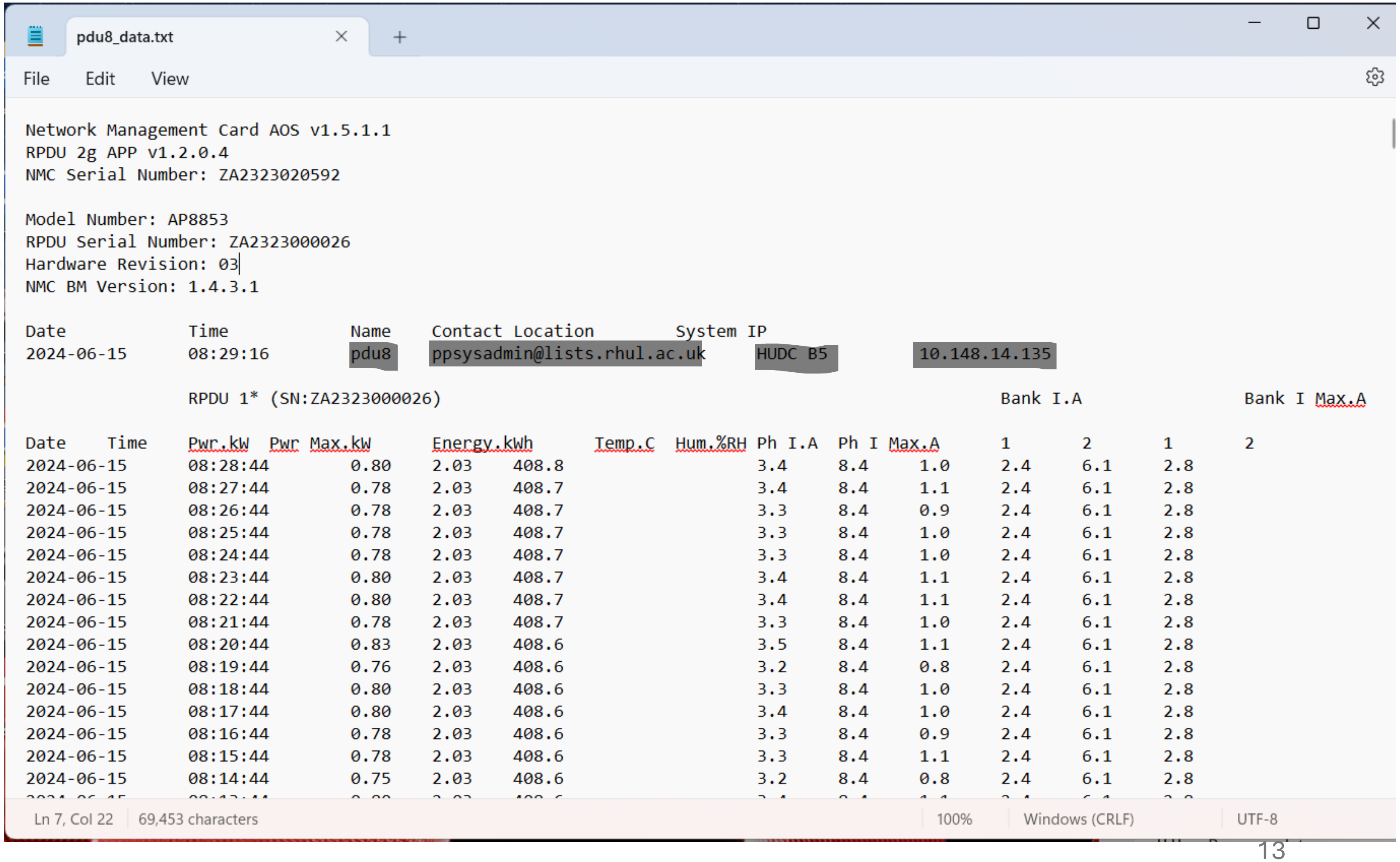
- q75 is a decent estimator of average power under load
- Mean is not
- PDU q75 2.3% higher than IPMI
- This is within claimed 3% accuracy of PDU



Additional material

Format of data from PDU

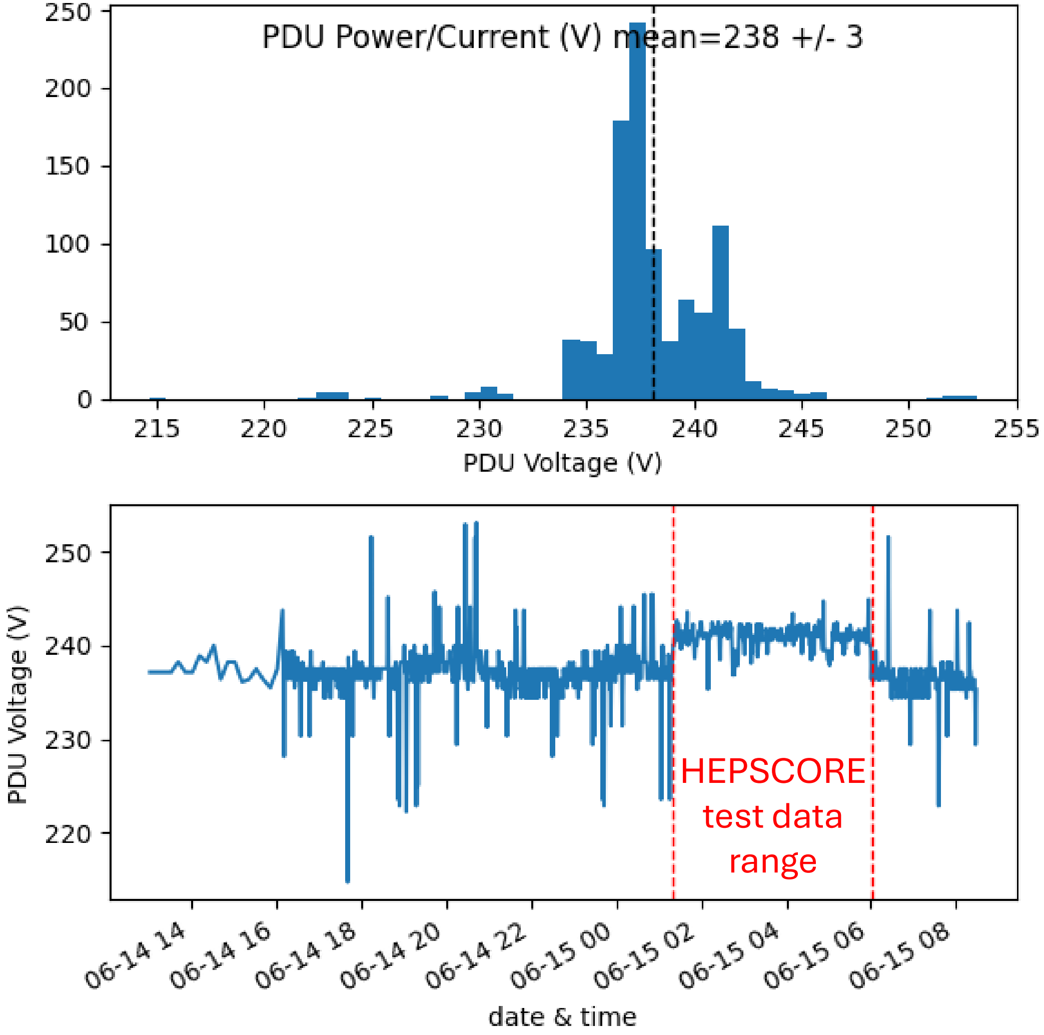
Remove first 14 lines then it is a simple tab-separated table



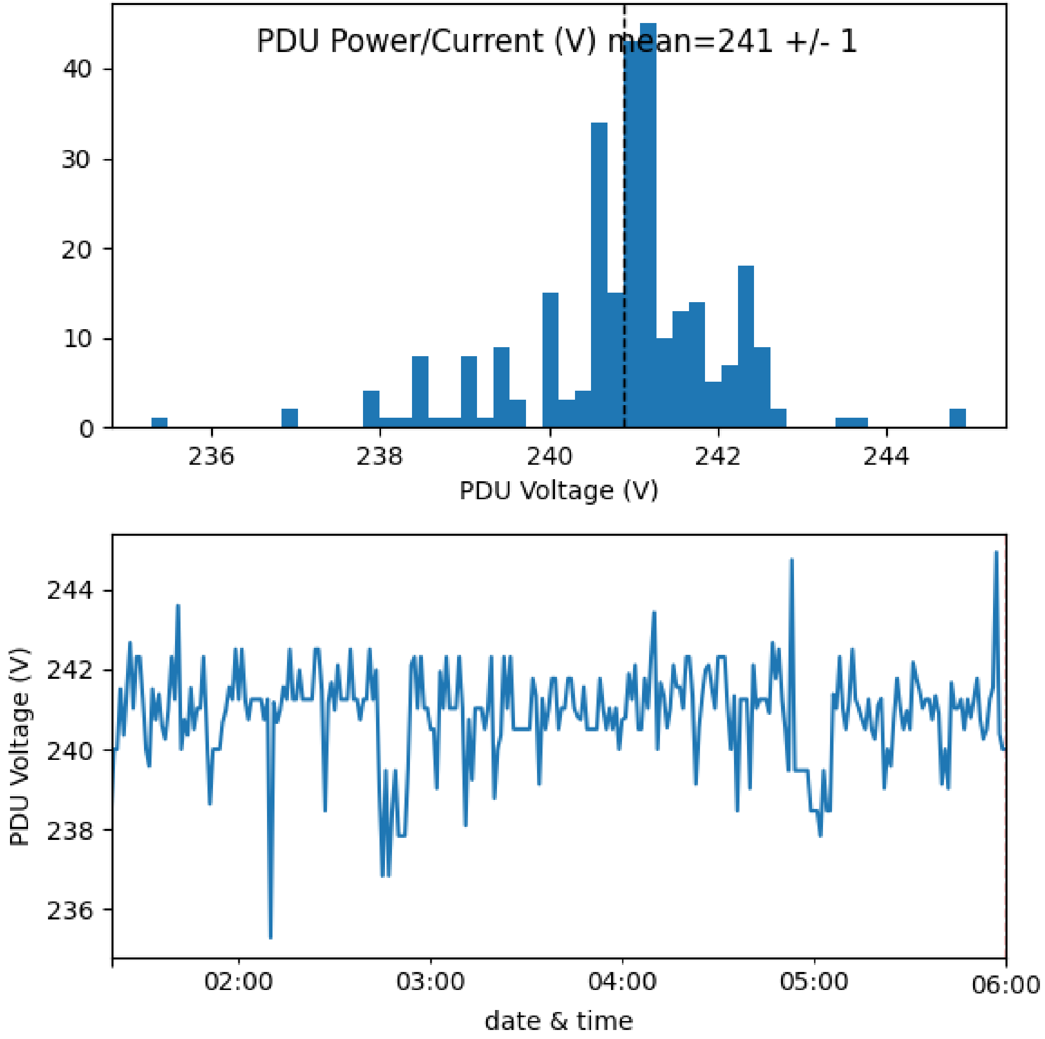
```
Network Management Card AOS v1.5.1.1
RPDU 2g APP v1.2.0.4
NMC Serial Number: ZA2323020592

Model Number: AP8853
RPDU Serial Number: ZA2323000026
Hardware Revision: 03
NMC BM Version: 1.4.3.1

Date            Time            Name    Contact Location        System IP
2024-06-15      08:29:16        pdu8    ppsysadmin@lists.rhul.ac.uk     HUDC B5         10.148.14.135

                RPDU 1* (SN:ZA2323000026)                                                       Bank I.A                Bank I Max.A

Date      Time  Pwr.kW  Pwr Max.kW      Energy.kWh      Temp.C  Hum.%RH Ph I.A  Ph I Max.A      1       2       1       2
2024-06-15      08:28:44        0.80    2.03    408.8                   3.4     8.4     1.0     2.4     6.1     2.8
2024-06-15      08:27:44        0.78    2.03    408.7                   3.4     8.4     1.1     2.4     6.1     2.8
2024-06-15      08:26:44        0.78    2.03    408.7                   3.3     8.4     0.9     2.4     6.1     2.8
2024-06-15      08:25:44        0.78    2.03    408.7                   3.3     8.4     1.0     2.4     6.1     2.8
2024-06-15      08:24:44        0.78    2.03    408.7                   3.3     8.4     1.0     2.4     6.1     2.8
2024-06-15      08:23:44        0.80    2.03    408.7                   3.4     8.4     1.1     2.4     6.1     2.8
2024-06-15      08:22:44        0.80    2.03    408.7                   3.4     8.4     1.1     2.4     6.1     2.8
2024-06-15      08:21:44        0.78    2.03    408.7                   3.3     8.4     1.0     2.4     6.1     2.8
2024-06-15      08:20:44        0.83    2.03    408.6                   3.5     8.4     1.1     2.4     6.1     2.8
2024-06-15      08:19:44        0.76    2.03    408.6                   3.2     8.4     0.8     2.4     6.1     2.8
2024-06-15      08:18:44        0.80    2.03    408.6                   3.3     8.4     1.0     2.4     6.1     2.8
2024-06-15      08:17:44        0.80    2.03    408.6                   3.4     8.4     1.0     2.4     6.1     2.8
2024-06-15      08:16:44        0.78    2.03    408.6                   3.3     8.4     0.9     2.4     6.1     2.8
2024-06-15      08:15:44        0.78    2.03    408.6                   3.3     8.4     1.1     2.4     6.1     2.8
2024-06-15      08:14:44        0.75    2.03    408.6                   3.2     8.4     0.8     2.4     6.1     2.8
```

PDU voltage stability

Full PDU data range



HEPSCORE test data range


PDU Power/Current (V) mean=241 +/- 1
PDU Voltage (V)
PDU Voltage (V)
02:00
03:00
04:00
05:00
06:00
date & time