Hewlett Packard
Enterprise

HPE Cray EX225a (MI300a) Blade Power Capping and HBM Page Retirement

Steven J. Martin, HPE

May 4, 2025

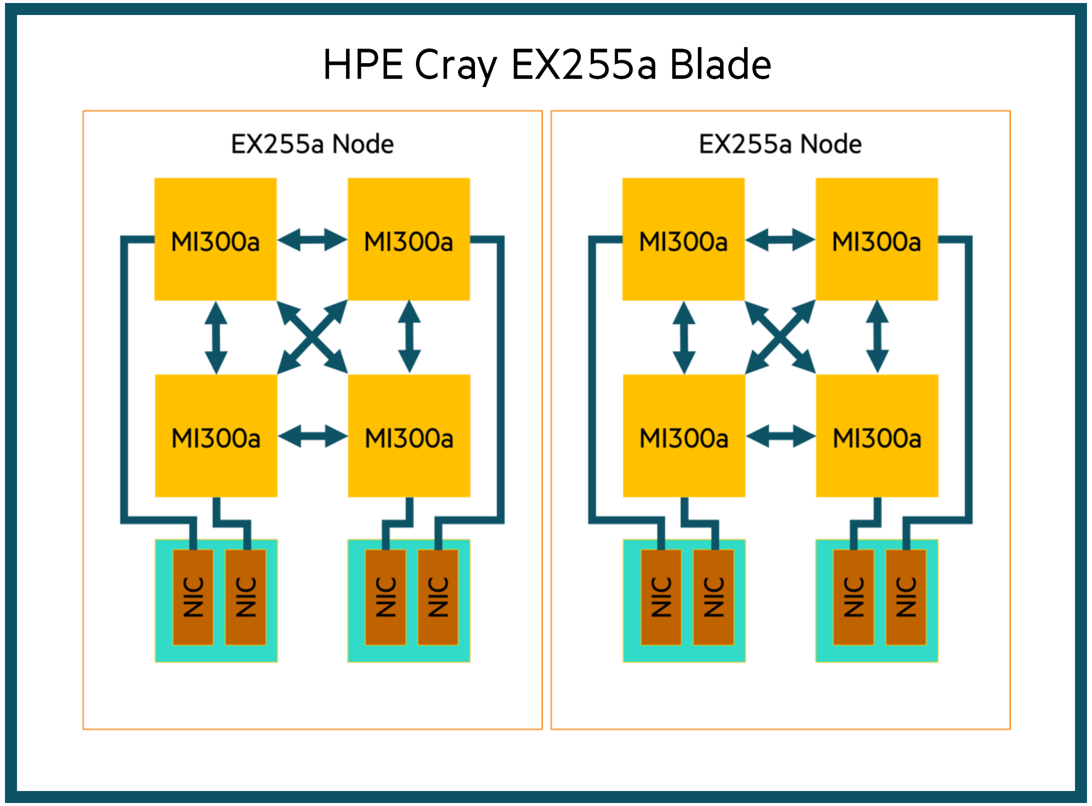
Agenda

- HPE Cray Supercomputing EX255a one slide overview
- HPE Cray EX225a (MI300a) blade power capping
- HPE Cray EX225a (MI300a) blade HBM page retirement

EL CAPITAN
Lawrence Livermore National Laboratory
U.S. DEPARTMENT OF ENERGY
NNSA
Hewlett Packard Enterprise
AMD
Red Hat
EL CAPITAN
ADVANCED SIMULATION & COMPUTING

HPE Cray EX255a Overview

- Two nodes per blade
- 4 AMD Instinct™ MI300a APUs per node
  - AMD MI300a
    - 24 AMD ‘Zen 4’ x86 CPU cores
    - 228 GPU compute units
    - 128 GB HBM3
- Quad injection Slingshot 200 per node
- Blade in LLNL El Capitan System



HPE Cray EX225a (MI300a) Power Capping

Enforcement of cabinet-level max power constraints has driven a new power-capping design for HPE Cray EX255a nodes

HPE Cray EX225a (MI300a) Power Capping: Objective

- Manage total cabinet power to HPE Cray EX4000 Cabinet constraints
  - North America: 400 KW
  - Non-North American regions: 346 KW
- Minimize performance impact of power capping to application performance
- Minimize complexity of solution
- Maintain solution Reliability Availability and Serviceability (RAS)

Node Power Cap Control

- Power capping implemented per node
  - Simplifies implementation, improves reliability, tunable response
  - Same Redfish APIs as other HPE Cray EX nodes
  - Out-of-band implementation independent of in-band controls
  - In-band rocm-smi controls complementary, lowest value is enforced
    - In-band control is closed-loop per socket and independent of node-level power

- APUs are configured for 550W TDP
  - EX packaging limits maximum current and cooling capacity – observed limited benefit above 550W
  - APUs not capped below 550W unless node power is greater than power cap target
  - Avoiding static APU capping that would lower performance even if not all the APUs are fully loaded
  - Maximize performance when APU load is unbalanced

- When power capping is enforced
  - All APUs are capped at the same power target
  - APU caps updated at most once every two seconds to avoid thrashing APU internal control logic

Node Power Cap Control: Redfish API

- New for EX225a:
  - Node level capping enabled by default and cannot be disabled
  - Default per node power cap is 2894 watts
    - Includes NIC power and conversion overhead
    - North America with all blades populated
    - Non-North American regions with 7-blades per chassis
  - Cap range is 2619 to 2910 watts

- Redfish API is the same as other HPE Cray EX Nodes
  - Curl call in a bash script would look something like:

```
curl -s ${CURLOPTS} --connect-timeout 20 \
    -X PATCH -H 'Content-Type: application/json’ \
    https://${EP}/redfish/v1/Chassis/${node}/Controls/NodePowerLimit -d \
    "{\"ControlMode\": \"Automatic\", \"SetPoint\": ${LIM}}";
```

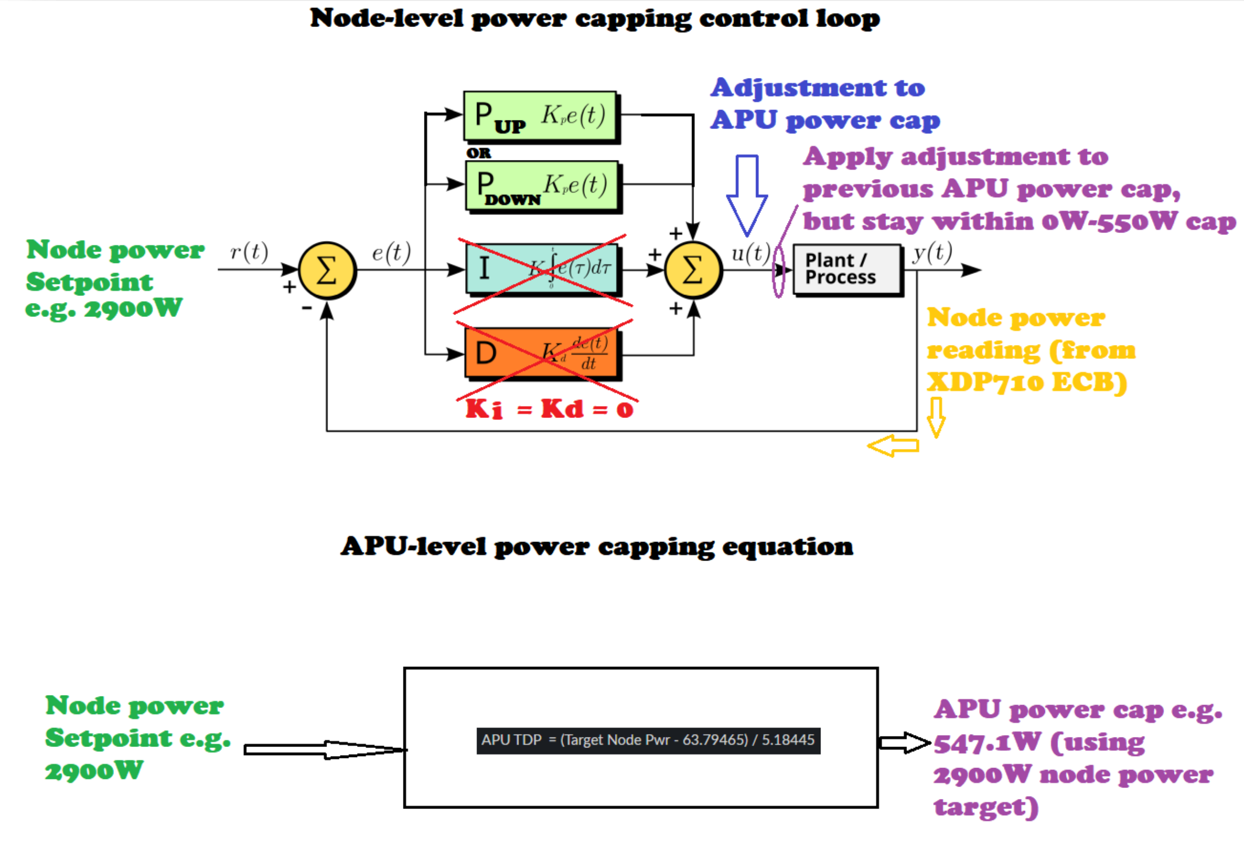
  - Default tools for setting power caps in HPCM and CSM are unchanged

Node Power Caping: Standard PID Control Loop Design

- Overview:

  A proportional–integral–derivative (PID) controller is a feedback-based control loop mechanism commonly used to manage machines and processes that require continuous control and automatic adjustment. Wikipedia

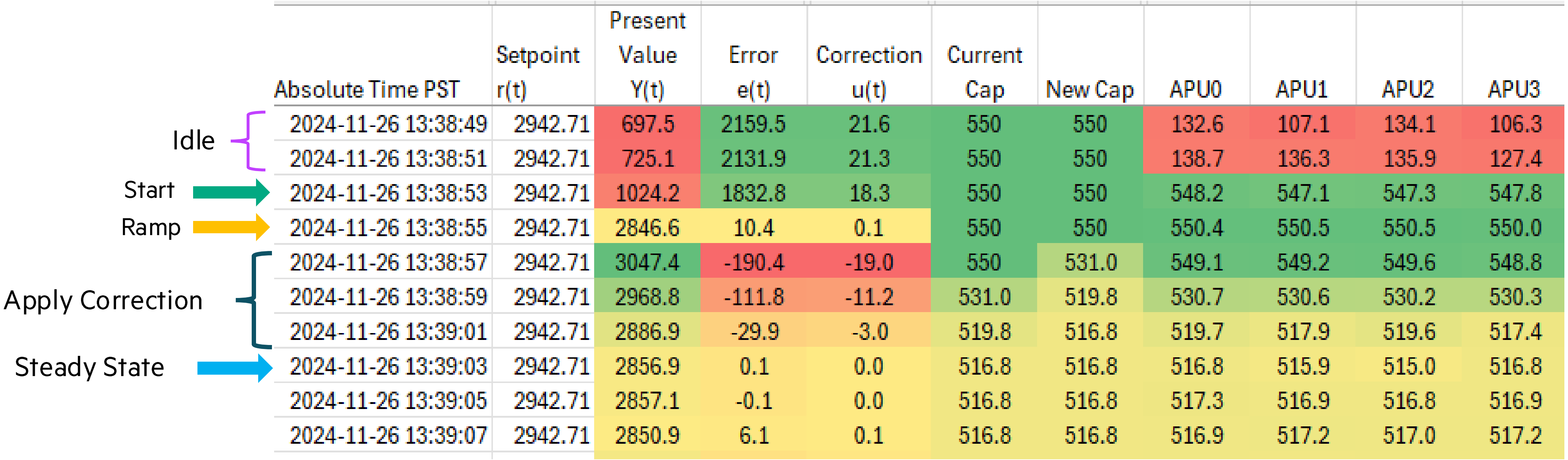
- Proportional function with two coefficients allowing
  - Fast clamp down when power is exceeded
  - Slow ramp back up when power is low
- Integral and differential coefficient not used
- Loop cycle time 2 seconds
- Controlled parameter is APU power cap



Node-level power capping control loop
Node power Setpoint e.g. 2900W
$r(t)$
+
-
Σ
$e(t)$
P_{UP} $K_p e(t)$
OR
P_{DOWN} $K_p e(t)$
I $K_i \int_0^t e(\tau)d\tau$
D $K_d \frac{de(t)}{dt}$
Ki = Kd = 0
+
+
+
Σ
$u(t)$
Adjustment to APU power cap
Apply adjustment to previous APU power cap, but stay within 0W-550W cap
Plant / Process
$y(t)$
Node power reading (from XDP710 ECB)
APU-level power capping equation
Node power Setpoint e.g. 2900W
APU TDP = (Target Node Pwr - 63.79465) / 5.18445
APU power cap e.g. 547.1W (using 2900W node power target)

PID Control Debug Data Running DGEMM-Stress at LLNL

- Each row shows one iteration of the PID control loop
- Starting with idle node (APUs running ~130W), APUs at default per APU cap of 550W
- DGEMM starts and APUs ramp to 550W measured node power increases and exceed the setpoint
- When setpoint exceeded, a correction factor is generated, and the APU power cap set < 550W
- Steady state is reached after 3 iterations (6 seconds)



	Absolute Time PST	Setpoint r(t)	Present Value Y(t)	Error e(t)	Correction u(t)	Current Cap	New Cap	APU0	APU1	APU2	APU3
Idle	2024-11-26 13:38:49	2942.71	697.5	2159.5	21.6	550	550	132.6	107.1	134.1	106.3
	2024-11-26 13:38:51	2942.71	725.1	2131.9	21.3	550	550	138.7	136.3	135.9	127.4
Start	2024-11-26 13:38:53	2942.71	1024.2	1832.8	18.3	550	550	548.2	547.1	547.3	547.8
Ramp	2024-11-26 13:38:55	2942.71	2846.6	10.4	0.1	550	550	550.4	550.5	550.5	550.0
Apply Correction	2024-11-26 13:38:57	2942.71	3047.4	-190.4	-19.0	550	531.0	549.1	549.2	549.6	548.8
	2024-11-26 13:38:59	2942.71	2968.8	-111.8	-11.2	531.0	519.8	530.7	530.6	530.2	530.3
	2024-11-26 13:39:01	2942.71	2886.9	-29.9	-3.0	519.8	516.8	519.7	517.9	519.6	517.4
Steady State	2024-11-26 13:39:03	2942.71	2856.9	0.1	0.0	516.8	516.8	516.8	515.9	515.0	516.8
	2024-11-26 13:39:05	2942.71	2857.1	-0.1	0.0	516.8	516.8	517.3	516.9	516.8	516.9
	2024-11-26 13:39:07	2942.71	2850.9	6.1	0.1	516.8	516.8	516.9	517.2	517.0	517.2

Performance Results Large Systems at LLNL

- Applications with performance drop < 1 % [in the range of run-to-run variation]
  - The performance numbers in the table are an average of five runs
  - Customer node power setpoint finalized at 2857W (taking Rabbit hardware into consideration)
  - HPL17 and HPL22 (interim implementations) were available and not the higher performance HPL version

Application	Uncapped	Capped	Cap Setting	Performance
HPL22	2.2697E+07	2.2477E+07	2874	0.97%
Pennant nc13	9.2788E+11	9.2919E+11	2874	-0.14%
qmcpack	3.8267E+05	3.8413E+05	2874	-0.38%
HPL17	1.7178E+07	1.7180E+07	2874	-0.01%
Pennant	9.1776E+11	9.2356E+11	2863	-0.63%
HACC	3.2172E+09	3.2367E+09	2857	-0.61%
laghos	5.1663E+04	5.1666E+04	2857	-0.01%
hipBone	1.6251E+06	1.6137E+06	2857	0.71%
Kripke	na	8.2342E+12	2857	Did not cap
LAMMPS	na	3.9803E+09	2857	Did not cap
quicksilver	na	3.7804E+11	2857	Did not cap
amg_solve+	na	1.5034E+13	2857	Did not cap
amg_setup+	na	1.4445E+12	2857	Did not cap

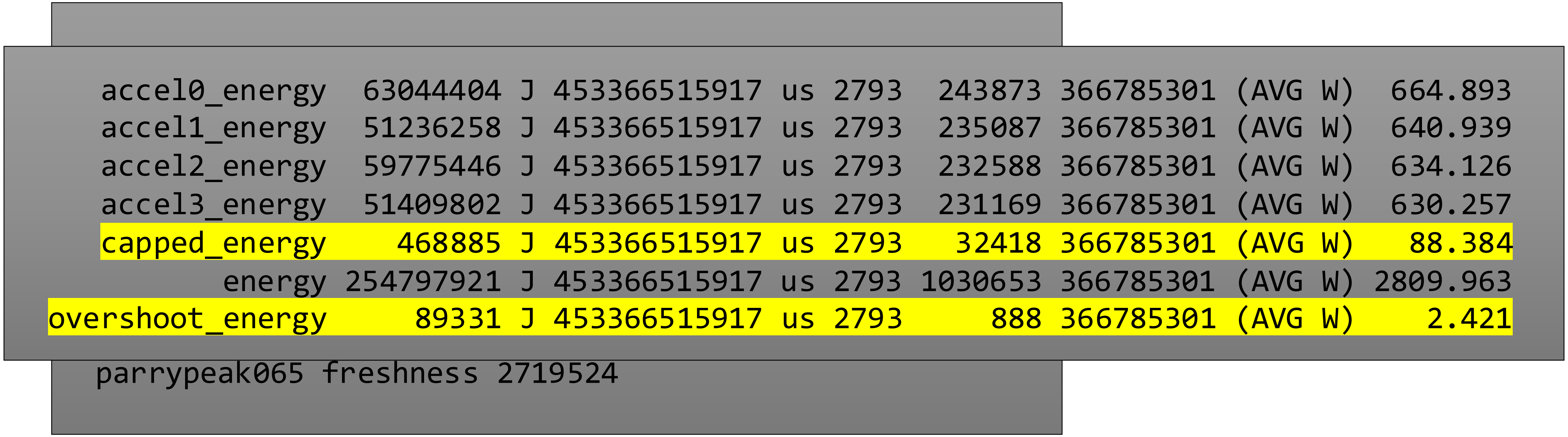
New PM Counters for Capped Energy and Overshoot Energy

- **capped_energy** accumulates when power capping actively enforced
  - Accumulates whenever APUs are capped below 550 watts
- **overshoot_energy** accumulates when power is over the power cap target
- Enabled visibility into power capping enforcement

```
parrypeak065 freshness 2719524
parrypeak065 accel0_energy 62800531 J 452999730616 us
parrypeak065 accel1_energy 51001171 J 452999730616 us
parrypeak065 accel2_energy 59542858 J 452999730616 us
parrypeak065 accel3_energy 51178633 J 452999730616 us
parrypeak065 capped_energy   436467 J 452999730616 us
parrypeak065 energy       253767268 J 452999730616 us
parrypeak065 overshoot_energy 88443 J 452999730616 us
parrypeak065 freshness 2719524
```

New PM Counters for Capped Energy and Overshoot Energy

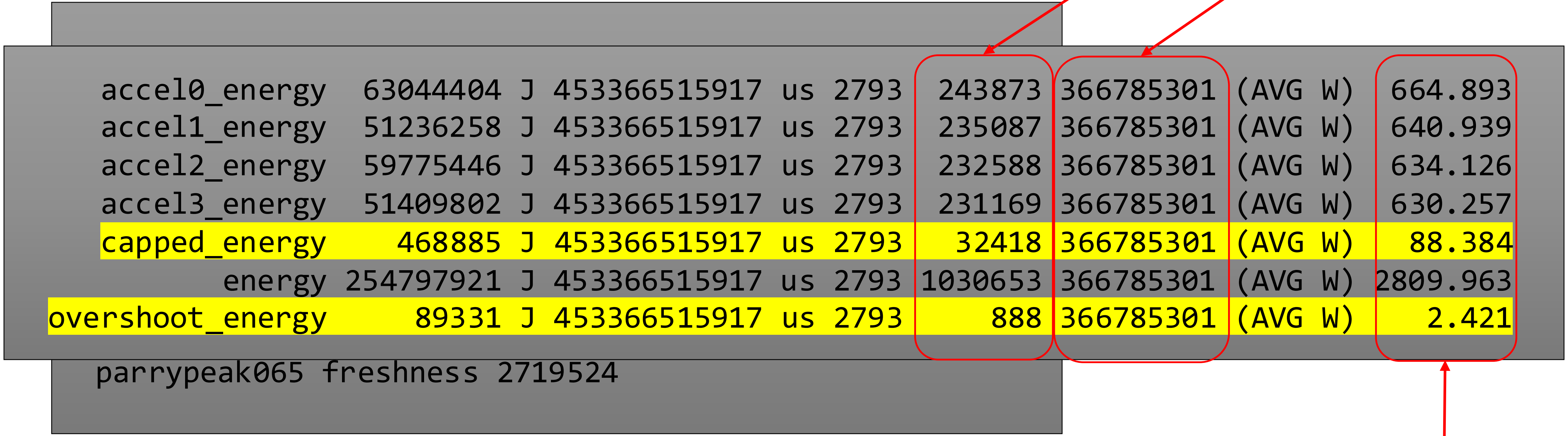
- **capped_energy** accumulates when power capping actively enforced
  - Accumulates whenever APUs are capped below 550 watts
- **overshoot_energy** accumulates when power is over the power cap target
- Enabled visibility into power capping enforcement



```
   accel0_energy  63044404 J 453366515917 us 2793  243873 366785301 (AVG W)  664.893
   accel1_energy  51236258 J 453366515917 us 2793  235087 366785301 (AVG W)  640.939
   accel2_energy  59775446 J 453366515917 us 2793  232588 366785301 (AVG W)  634.126
   accel3_energy  51409802 J 453366515917 us 2793  231169 366785301 (AVG W)  630.257
   capped_energy    468885 J 453366515917 us 2793   32418 366785301 (AVG W)   88.384
          energy 254797921 J 453366515917 us 2793 1030653 366785301 (AVG W) 2809.963
overshoot_energy     89331 J 453366515917 us 2793     888 366785301 (AVG W)    2.421
   parrypeak065 freshness 2719524
```

New PM Counters for Capped Energy and Overshoot Energy

- **capped_energy** accumulates when power capping actively enforced
  - Accumulates whenever APUs are capped below 550 watts
- **overshoot_energy** accumulates when power is over the power cap target
- Enabled visibility into power capping enforcement



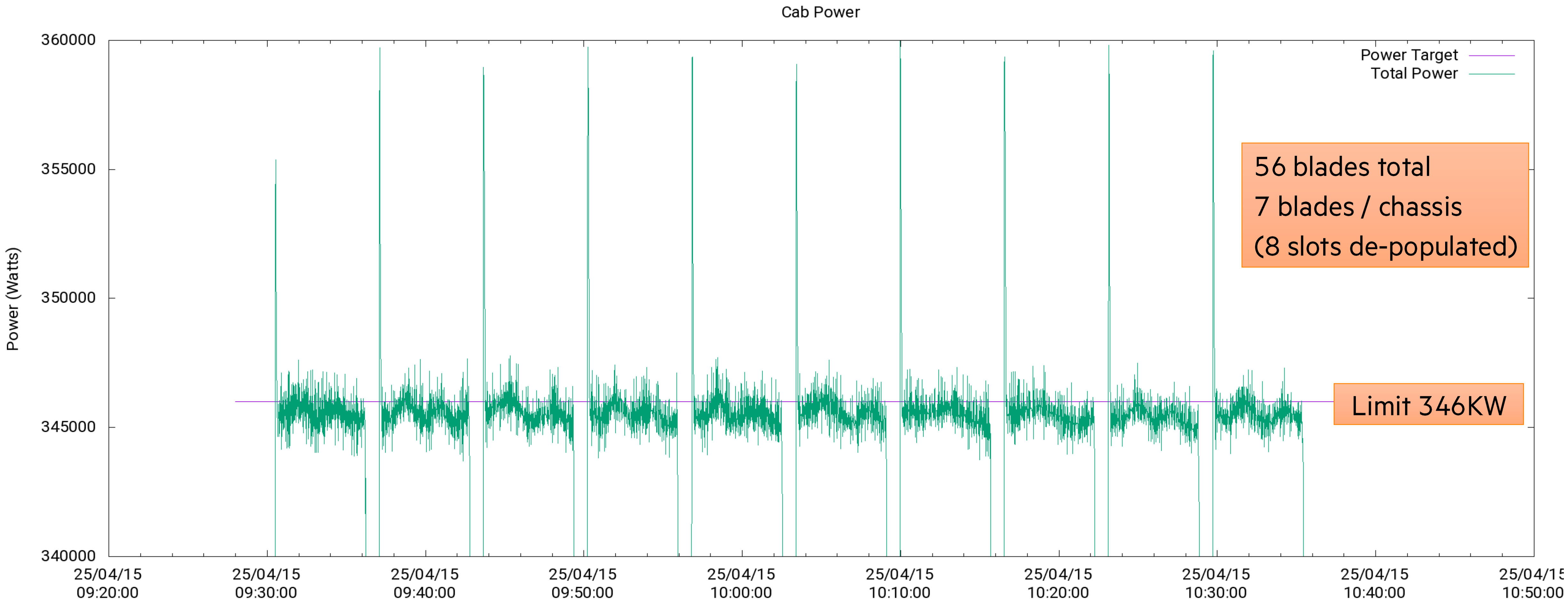
```
    accel0_energy   63044404 J 453366515917 us 2793  243873 366785301 (AVG W)  664.893
    accel1_energy   51236258 J 453366515917 us 2793  235087 366785301 (AVG W)  640.939
    accel2_energy   59775446 J 453366515917 us 2793  232588 366785301 (AVG W)  634.126
    accel3_energy   51409802 J 453366515917 us 2793  231169 366785301 (AVG W)  630.257
    capped_energy     468885 J 453366515917 us 2793   32418 366785301 (AVG W)   88.384
           energy  254797921 J 453366515917 us 2793 1030653 366785301 (AVG W) 2809.963
 overshoot_energy      89331 J 453366515917 us 2793     888 366785301 (AVG W)    2.421
    parrypeak065 freshness 2719524
```

Delta energy

Delta time
36.6 seconds

Average power of job run

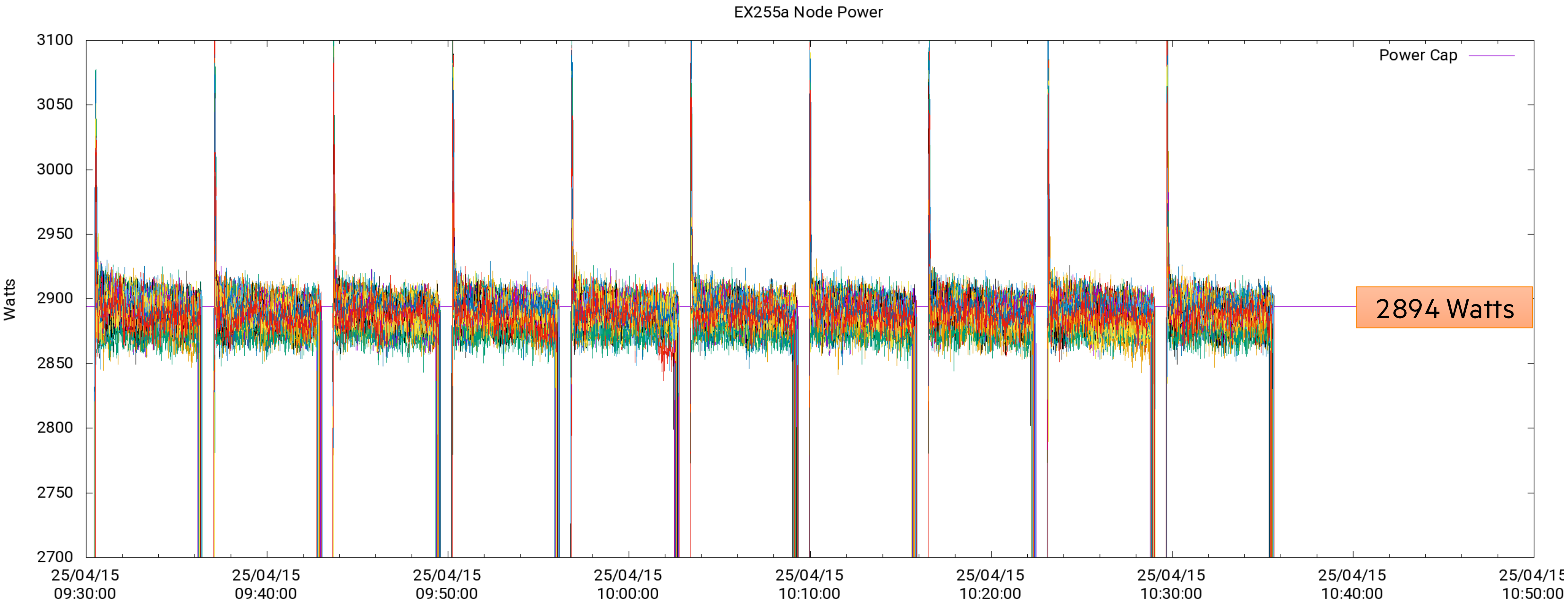
Non-North American Cabinet Power: DGEMM 112 Nodes Power Cap 2894



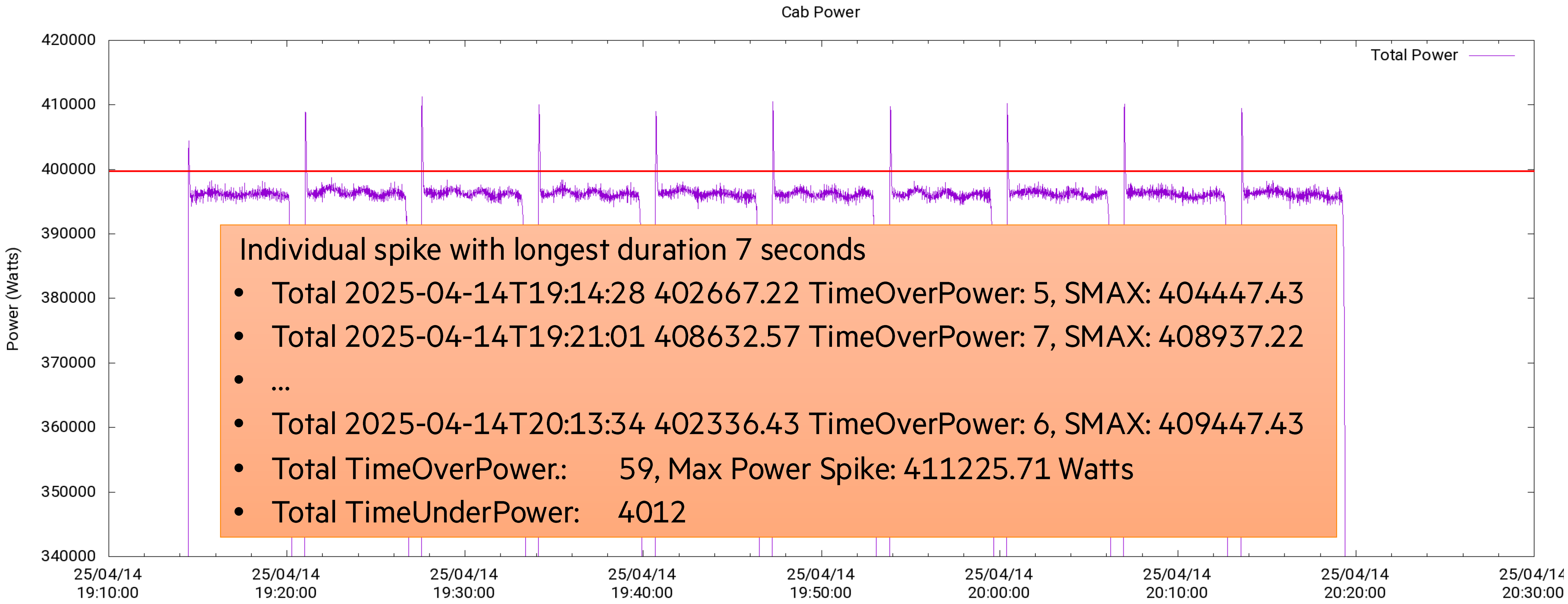
56 blades total
7 blades / chassis
(8 slots de-populated)

Limit 346KW

Non-North American Node Power: DGEMM 112 Nodes Power Cap 2894



Cabinet Power: DGEMM 10 Runs 128 Nodes Power Cap 2910 (Max)

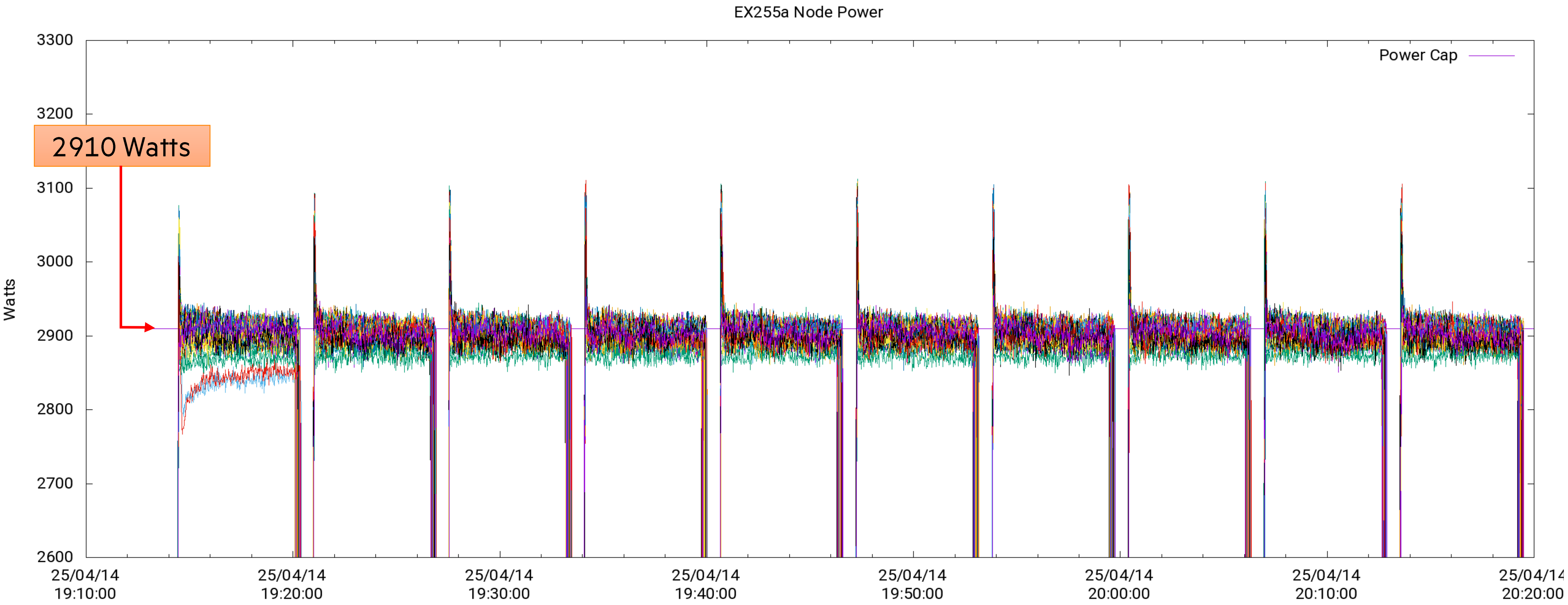


Individual spike with longest duration 7 seconds

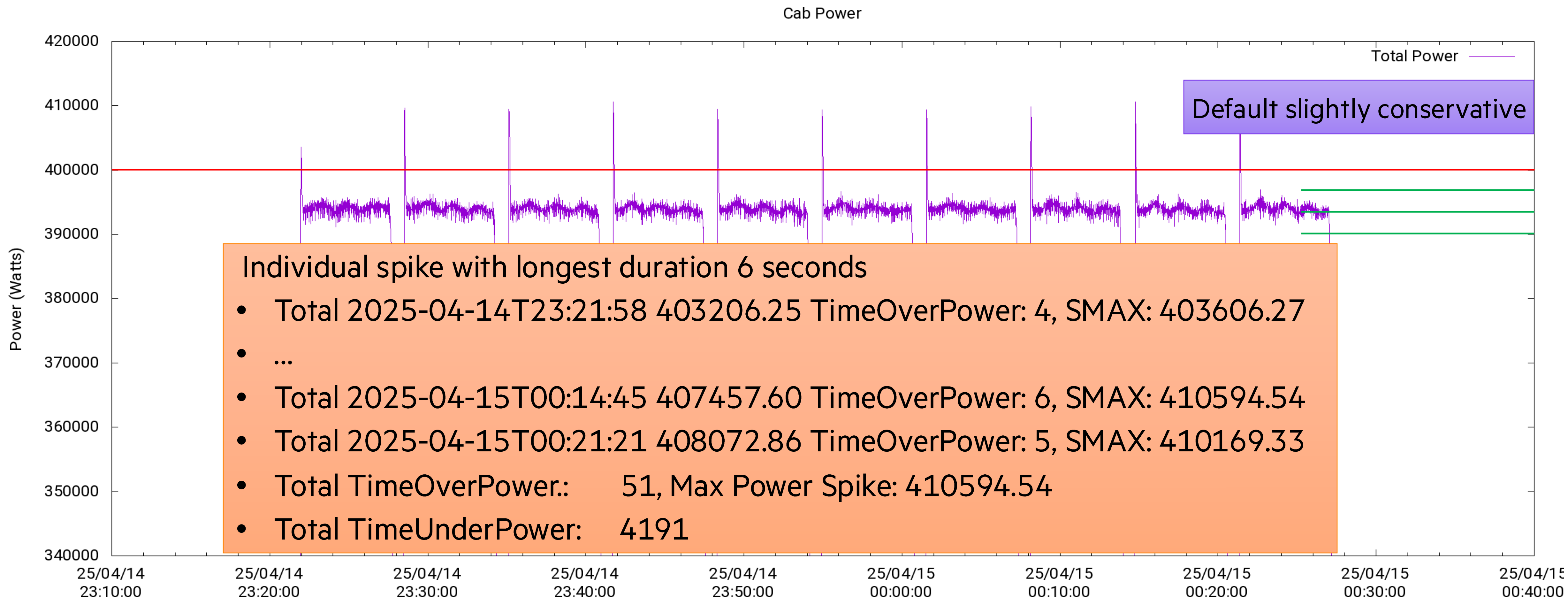
- Total 2025-04-14T19:14:28 402667.22 TimeOverPower: 5, SMAX: 404447.43
- Total 2025-04-14T19:21:01 408632.57 TimeOverPower: 7, SMAX: 408937.22
- …
- Total 2025-04-14T20:13:34 402336.43 TimeOverPower: 6, SMAX: 409447.43
- Total TimeOverPower.: 59, Max Power Spike: 411225.71 Watts
- Total TimeUnderPower: 4012

Design target is less than 10 seconds above the 400KW limit

Node Power: DGEMM 10 Runs 128 Nodes Power Cap 2910 (Max)



Cabinet Power: DGEMM 10 Runs 128 Nodes Power Cap 2894 (Default)

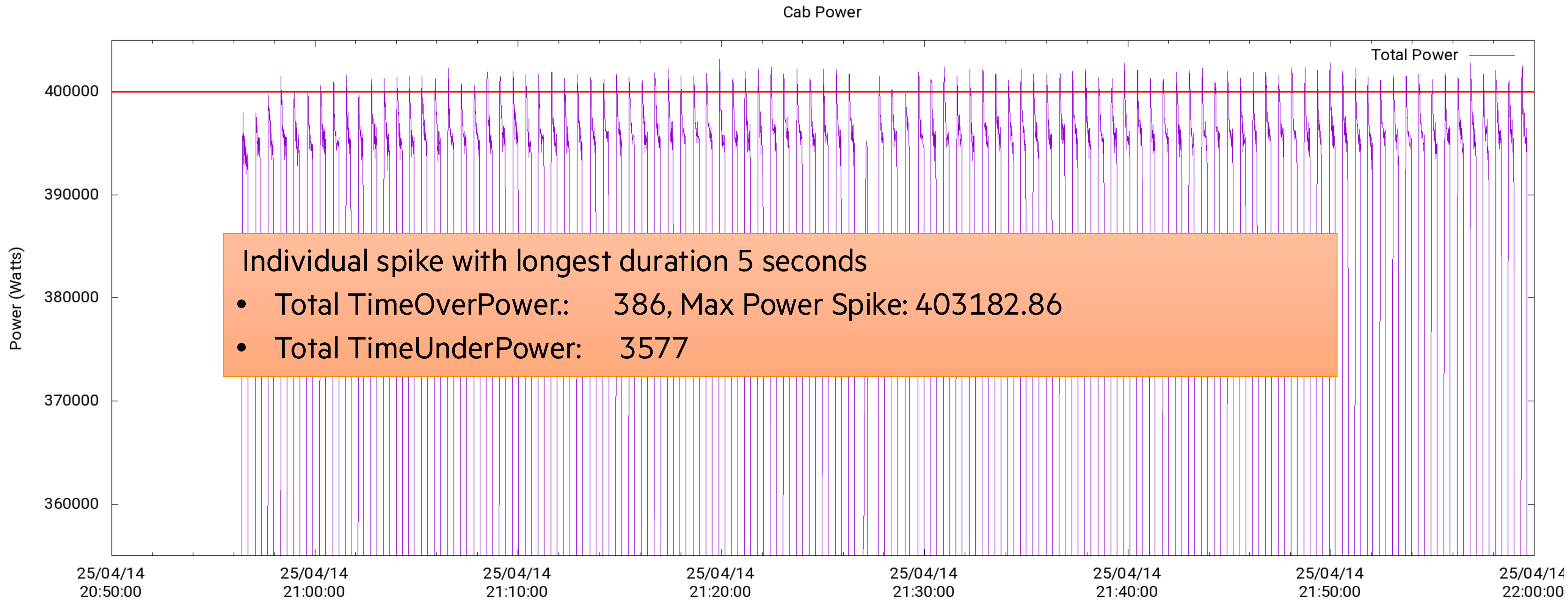


Default slightly conservative

Individual spike with longest duration 6 seconds

- Total 2025-04-14T23:21:58 403206.25 TimeOverPower: 4, SMAX: 403606.27
- ...
- Total 2025-04-15T00:14:45 407457.60 TimeOverPower: 6, SMAX: 410594.54
- Total 2025-04-15T00:21:21 408072.86 TimeOverPower: 5, SMAX: 410169.33
- Total TimeOverPower.: 51, Max Power Spike: 410594.54
- Total TimeUnderPower: 4191

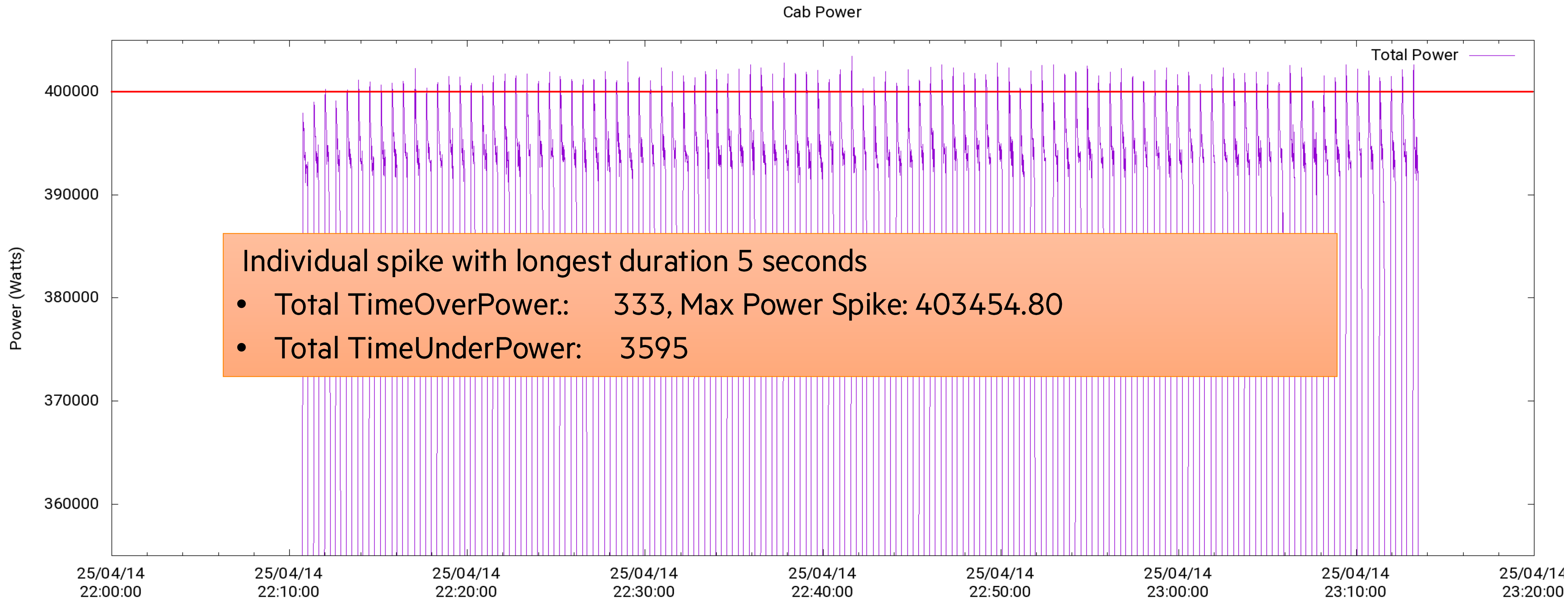
Cabinet Power: Pennant 100x128N Runs Nodes Power Cap 2910 (Max)



Individual spike with longest duration 5 seconds

- Total TimeOverPower.: 386, Max Power Spike: 403182.86
- Total TimeUnderPower: 3577

Cabinet Power: Pennant 100x128N Runs Power Cap 2894 (Default)

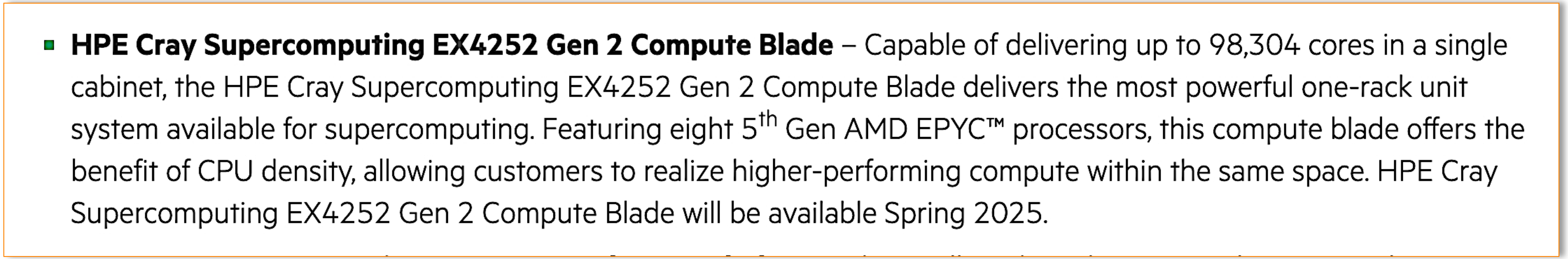


Individual spike with longest duration 5 seconds

- Total TimeOverPower.: 333, Max Power Spike: 403454.80
- Total TimeUnderPower: 3595

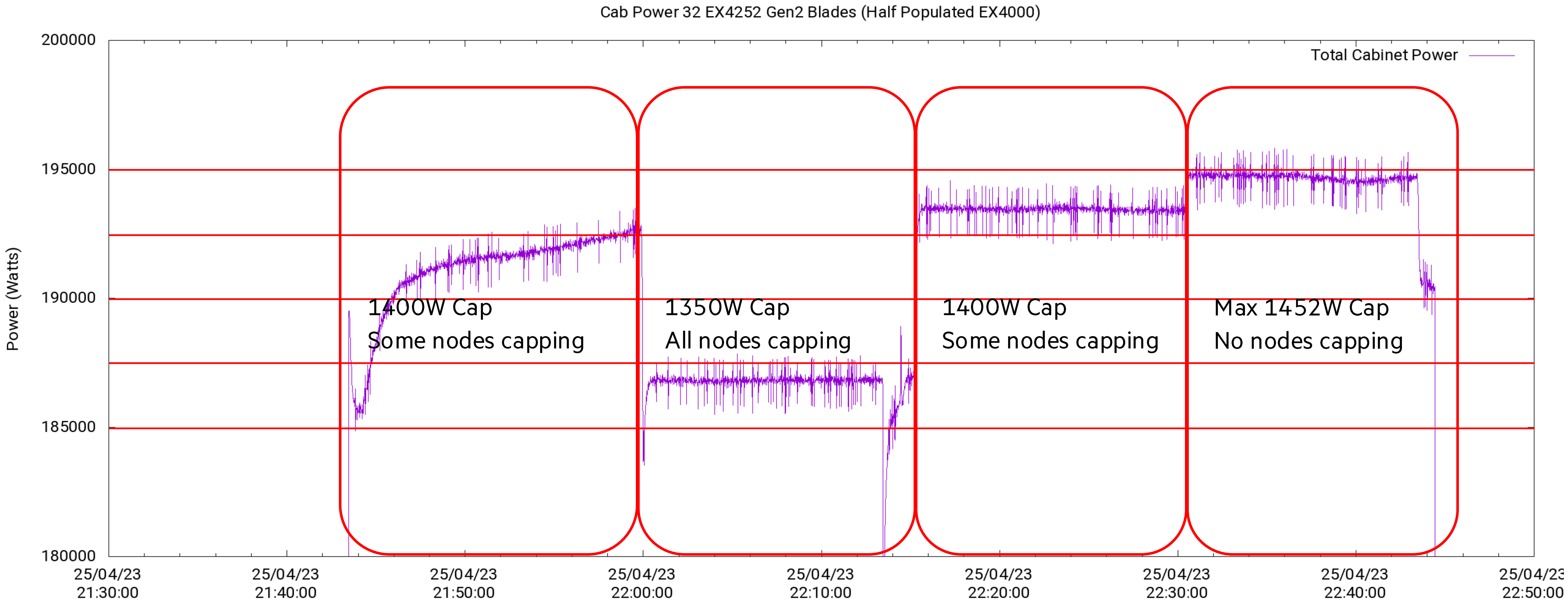
EX4252 Gen2 Power Capping Plan (Early, Subject to Change)

- Leverage design from EX225a
- Node level capping enabled by default and cannot be disabled
- Default power cap is 1452 watts
- Power cap range is 700 to 1452 watts



- **HPE Cray Supercomputing EX4252 Gen 2 Compute Blade** – Capable of delivering up to 98,304 cores in a single cabinet, the HPE Cray Supercomputing EX4252 Gen 2 Compute Blade delivers the most powerful one-rack unit system available for supercomputing. Featuring eight 5th Gen AMD EPYC™ processors, this compute blade offers the benefit of CPU density, allowing customers to realize higher-performing compute within the same space. HPE Cray Supercomputing EX4252 Gen 2 Compute Blade will be available Spring 2025.

EX4252 Gen2 (32 Blades) Cap at 1400, 1350, 1400, 1452 Watts, Cabinet Power



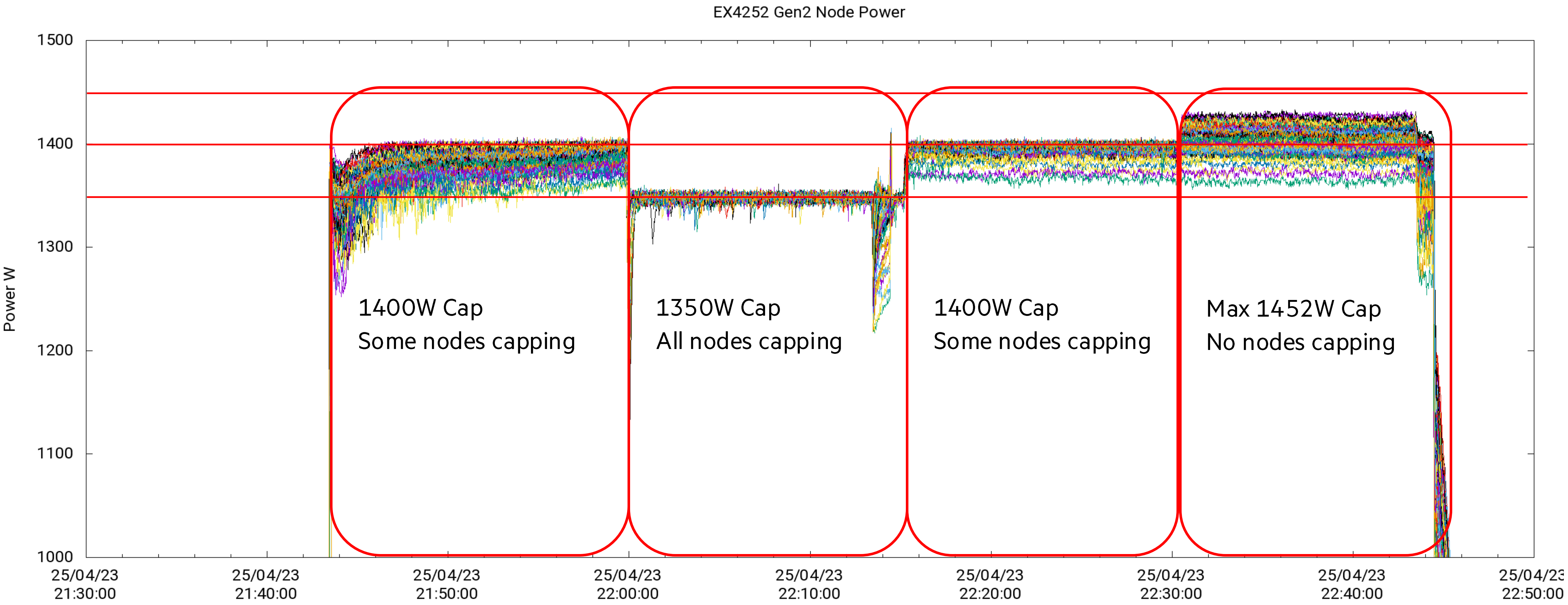
1400W Cap
Some nodes capping

1350W Cap
All nodes capping

1400W Cap
Some nodes capping

Max 1452W Cap
No nodes capping

EX4252 Gen2 (124 Nodes) Node Power Cap at 1400, 1350, 1400, 1452 Watts



1400W Cap
Some nodes capping

1350W Cap
All nodes capping

1400W Cap
Some nodes capping

Max 1452W Cap
No nodes capping

HPE Cray EX225a (MI300a) Power Capping Summary

- Challenge
  - Customer codes could cause EX4000 cabinet power to exceed 400KW design point
  - Real problem because the codes represent real job characteristics

- Solution
  - Node level power control feature
  - Limiting node power at a point that when aggregated is less then 400KW
    - [346KW outside of North America]
  - APU limits only imposed when total node power is over the power cap

- Results
  - Power cap feature able to control peak and steady state power to remain below breaker trip characteristics
  - Worst case performance impact is much less than the 5%
  - No performance impact to most applications tested
  - Leverage design for EX4252 Gen2

HPE Cray EX225a (MI300a) HBM Page Retirement

The HPE Cray EX225a blade persists High Bandwidth Memory (HBM) page retirement data across reboots, and power cycles.

> Google AI Overview:
> In the context of AMD's High Bandwidth Memory (HBM) technology, page retirement refers to a process where memory pages (sections of memory) that have experienced uncorrectable errors are marked and removed from service to prevent data corruption. This process is essential for maintaining the reliability of HBM, which is used in GPUs and other applications where high bandwidth and low latency are crucial.

HPE Cray EX225a (MI300a) HBM Page Retirement – Hardware Requirement

- The HPE Cray EX255a blade uses HBM for main memory
- AMD MI300a APUs support High Bandwidth Memory (HBM) page retirement
  - MI300a sockets do not provide non-volatile storage for page retirement data
  - Linux OS and driver code handle page retirement at runtime, and call into BIOS to persist data

EX255a node controller provides data store required to persist data across reboots or power cycles

Data is persisted to non-volatile storage by the BIOS

HBM Page Retirement: Data Available from Linux

- File layout for */sys/kernel/debug/ras/fmpm/entries*
  - Column 1 : fru_idx socket (APU) index with HBM error/poison fault
  - Column 2: fru_id PPIN of socket
  - Column 3: Retired page number, starting at 0
  - Column 4: Timestamp
  - Column 5: MCA_IPID
  - Column 6: MCA_ADDR
  - Column 7: System Physical address

- *grep 0x0 /sys/kernel/debug/ras/fmpm/entries* # ignore other cruft in the files

```
2  0x20cf28b705ecc018     0        2024-06-25T19:32:34     0x0000609600590f00      0x0000000004ddeac0      0x000000602e8c4700
2  0x20cf28b705ecc018     1        2024-06-25T21:38:11     0x0000609600590f00      0x0000000004b4e60a      0x0000005d9e58c3a0
2  0x20cf28b705ecc018     2        2024-06-26T13:55:45     0x0000609600590f00      0x0000000004c6a6e2      0x0000005eba7c63a0
…
2  0x20cf28b705ecc018     15       2024-07-11T05:30:29     0x0000609600590f00      0x0000000004b48a0a      0x0000005d988898a0
```

- Socket (APU) replacement recommended after 16 retirement events, contact HPE Service

Thank you

steven.martin3@hpe.com