



**Hewlett Packard
Enterprise**

Updated Node Power Management For HPE Cray EX255a and EX254n Blades

Brian Collum, HPE

May 2024



Abstract

Cray EX nodes have always supported a form of power capping that would allow customers to lower power usage of specific nodes as desired. With the introduction of the HPE Cray EX254n (NVIDIA Grace Hopper) and HPE Cray EX255a (AMD MI300a), this became critical as the overall rack power pushed beyond the maximum supported at some customers sites. With the HPE Cray EX254n in particular, the total TDP of the modules exceeds the maximum that can be delivered by the Cray EX infrastructure. This drove the decision to have to set a power limit on the Grace Hopper modules by default (a first for Cray EX). This presentation will walkthrough design goals of the blades, how power capping is implemented in the firmware, and how to configure the power limit in a running system. The presentation will also go through how to view the current limits configured via the node controllers Redfish API and in-band tools, where applicable, and how the in-band tools interact with the out-of-band configurations.



Agenda

- Redfish APIs
 - Node Power Limit
 - Accelerator Power Limit
- In-band APIs
 - PM Counters
 - nvidia-smi
 - rocm-smi
- HPE Cray EX255a
- HPE Cray EX254n



Redfish – Node Power Limit

- Configures the maximum input power the node to draw from the chassis
- Spikes may be seen above SetPoint value may be seen occasionally but on average will be below

```
curl https://<nC>/redfish/v1/Chassis/Node0/Controls/NodePowerLimit
{
  ...
  "ControlMode": "Automatic",
  ...
  "Sensor": {
    "DataSourceUri": "/redfish/v1/Chassis/Node0/Sensors/ChassisVoltageRegulator0InputPower",
    "Reading": 614
  },
  "SetPoint": 2500,
  "SettingRangeMax": 2852,
  "SettingRangeMin": 1600,
  "Status": {
    "Health": "OK"
  }
}
```

Example from EX254n

Redfish – Accelerator Power Limit

- Configures the maximum power to be used by the given Accelerator
- Available on GPU Compute Blades
- Separate from the NodePowerLimit, but NodePowerLimit takes precedence

```
curl https://<nC>/redfish/v1/Chassis/Node0/Controls/Accelerator0PowerLimit
{
  ...
  "ControlMode": "Automatic",
  ...
  "SetPoint": 400,
  "SettingRangeMax": 900,
  "SettingRangeMin": 100,
  "Status": {
    "Health": "OK"
  }
}
```

Example from EX254n

In-band Tools

- PM Counters
 - Read-only
 - `/sys/cray/pm_counters/power_cap`
 - `/sys/cray/pm_countes/accel[0-3]_power_cap`
- nvidia-smi
 - <https://developer.nvidia.com/system-management-interface>
- rocm-smi
 - <https://rocm.docs.amd.com/en/latest/>

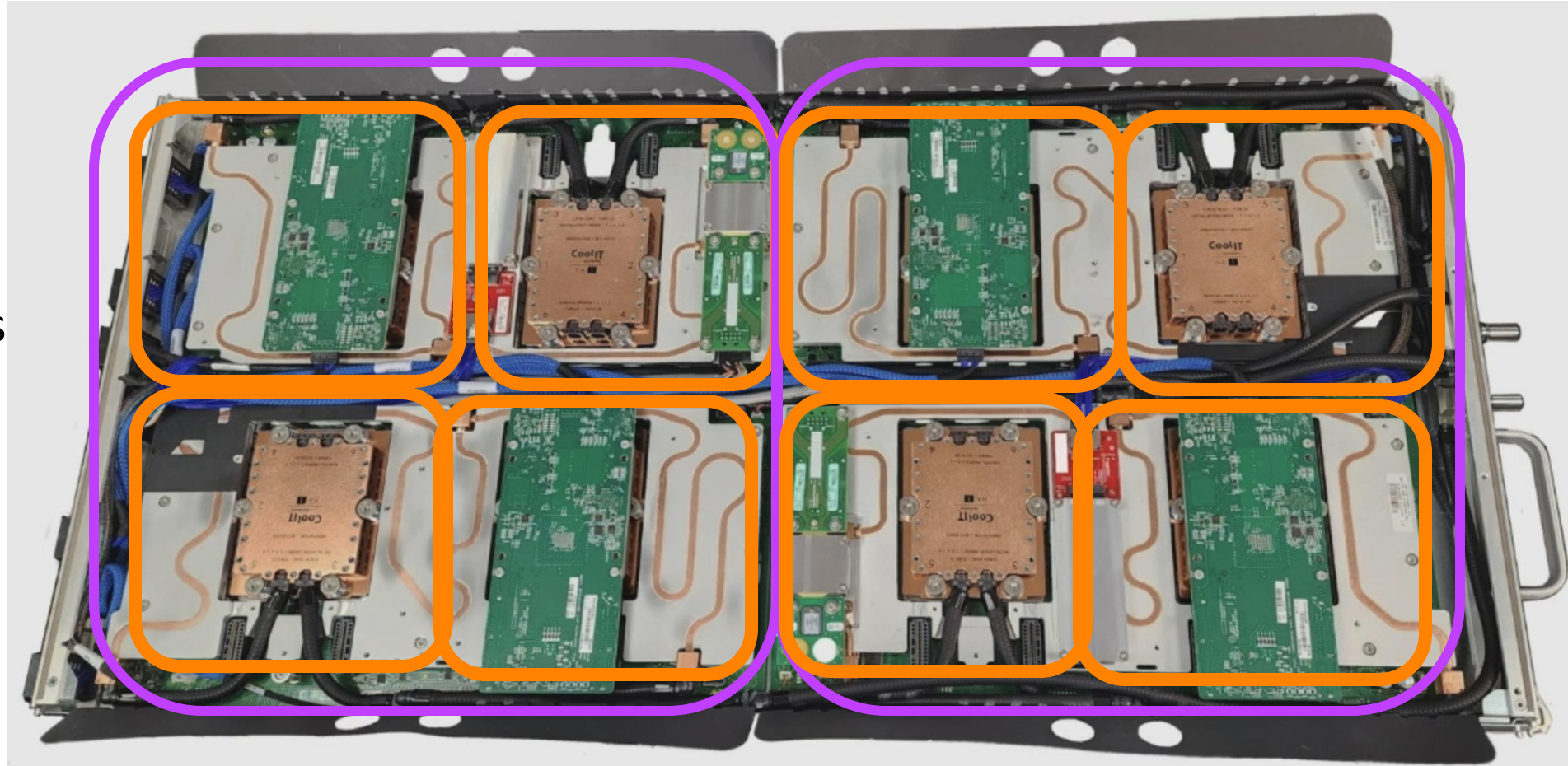


HPE Cray EX255a

- 4x AMD MI300a APUs

Nodes

APUs



HPE Cray EX255a – Redfish NodePowerLimit

- Evenly distributed across the four APUs
- Power Limit is configured per APU via AMD's APML 'Set Package Power Limit'
- When disabled, the limit is removed from the APU

$$APUPowerLimit = \frac{((\text{NodePowerLimit} * EfficiencyFactor) - OtherNodePower)}{NumberOfAPUs}$$



HPE Cray EX255a – View In-Band APU Limit

- Can read current limit with

```
# rocm-smi --showmaxpower
```

```
===== ROCm System Management Interface =====  
===== Power Cap =====  
GPU[0]          : Max Graphics Package Power (W): 550.0  
GPU[1]          : Max Graphics Package Power (W): 550.0  
GPU[2]          : Max Graphics Package Power (W): 550.0  
GPU[3]          : Max Graphics Package Power (W): 550.0  
===== End of ROCm SMI Log =====
```

- Does not reflect OOB limit
- Lower limit wins (OOB vs. In-band)!



HPE Cray EX255a – Set In-Band APU Limit

- Can set module limit to a value lower than OOB limit with

```
rocm-smi --setpoweroverdrive <power-limit>
```

- Should be able to read it back as well if successful

```
# rocm-smi --setpoweroverdrive 277
```

```
# rocm-smi --showmaxpower
```

```
===== ROCm System Management Interface =====
===== Power Cap =====
GPU[0]      : Max Graphics Package Power (W) : 277.0
GPU[1]      : Max Graphics Package Power (W) : 277.0
GPU[2]      : Max Graphics Package Power (W) : 277.0
GPU[3]      : Max Graphics Package Power (W) : 277.0
=====
===== End of ROCm SMI Log =====
```

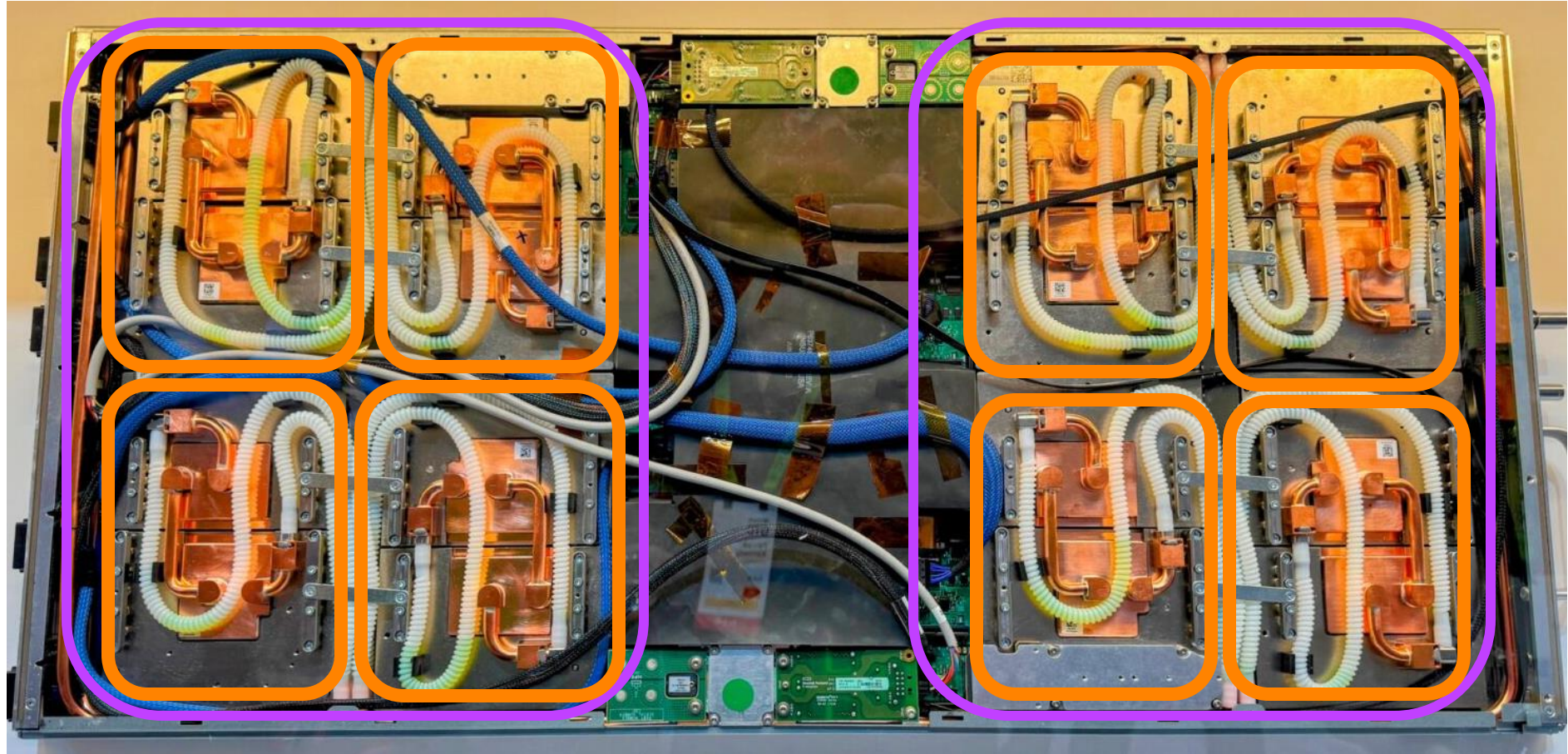
- Lower limit wins (OOB vs. In-band)!



HPE Cray EX254n

- 4x NVIDIA GH200 Superchip Modules

Nodes Modules



HPE Cray EX254n – Redfish NodePowerLimit

- Evenly distributed across the four modules
- Power Limit is configured per module via NVIDIA SMBPBI's 'Set Total Module Power Limit'
- When disabled, the maximum supported value is configured on the modules
 - This is the only Cray EX blade that sets a limit when NodePowerLimit is disabled
 - i.e. effectively NodePowerLimit.SetPoint == 2852W (660.12W per module)
- Power is managed by the GPU, preference is given to the CPU and the GPU is dynamically capped to keep module power under the configured limit
- Takes precedence over Accelerator[0-3]PowerLimit

$$ModulePowerLimit = \left(\frac{((\text{NodePowerLimit} * EfficiencyFactor) - OtherNodePower)}{NumberOfModules} \right) * LimitAccuracy$$



HPE Cray EX254n – View In-Band Module Limit

- Can read current limit with

```
nvidia-smi -q
```

- Shows a LOT of data, but in there is this for each module:

Module Power Readings

Power Draw	: 168.99 W
Current Power Limit	: 660.12 W
Requested Power Limit	: 1000.00 W
Default Power Limit	: 1000.00 W
Min Power Limit	: 200.00 W
Max Power Limit	: 1000.00 W

Requested In-Band
Limit

Difference indicates
out of band limit is in
effect

- Lower limit wins (OOB vs. In-band)!



HPE Cray EX254n – Set In-Band Module Limit

- Can set module limit to a value lower than OOB limit with

```
nvidia-smi --power-limit <power-limit> --scope 1
```

- After setting, able to read it back

Module Power Readings

Power Draw	: 168.99 W
Current Power Limit	: 550.00 W
Requested Power Limit	: 550.00 W
Default Power Limit	: 1000.00 W
Min Power Limit	: 200.00 W
Max Power Limit	: 1000.00 W

Requested In-Band
Limit



- Lower limit wins (OOB vs. In-band)!



HPE Cray EX254n – Accelerator[0-3]PowerLimit

- Power Limit applied to individual Hopper GPU
- Power Limit is configured per module via NVIDIA SMBPBI's 'Set Total GPU Power Limit'
- When disabled, the maximum supported value is configured on the GPU
 - i.e. effectively Accelerator[0-3]PowerLimit.SetPoint == 900W
- Module Limit (from NodePowerLimit) takes precedence, so higher value here has little meaning



HPE Cray EX254n – View In-Band GPU Limit

- Can read current limit with

```
nvidia-smi -q
```

- Shows a LOT of data, but in there is this for each GPU:

```
GPU Power Readings
```

```
Power Draw : 92.29 W
```

```
Current Power Limit : 800.00 W
```

```
Requested Power Limit : 900.00 W
```

```
Default Power Limit : 900.00 W
```

```
Min Power Limit : 100.00 W
```

```
Max Power Limit : 900.00 W
```

Requested In-Band
Limit

Difference indicates
out of band limit is in
effect

- Note: Can be misleading if Module Power Limit is set lower.
- Lower limit wins!



HPE Cray EX254n – Set In-Band GPU Limit

- Can set module limit to a value lower than OOB limit with
`nvidia-smi --power-limit <power-limit>`

- Should be able to read it back as well if successful

GPU Power Readings

Power Draw : 92.29 W

Current Power Limit : 500.00 W

Requested Power Limit : **500.00 W**

Default Power Limit : 900.00 W

Min Power Limit : 100.00 W

Max Power Limit : 900.00 W

Requested In-Band
Limit



- Lower limit wins!



Management Systems

- HPCM and CSM both support reading/setting
- CLI and API support
- Utilize same Redfish APIs

- **HPCM:**

- `mpower -n <nids> --get-limit`
- `mpower -n <nids> --set-limit <power>`

- **CSM:**

- `cray capmc get_power_cap create --nids <nids>`
- `cray capmc set_power_cap create --nids <nids> \`
 `--control "Node Power Limit" <power>`



Thank you

