

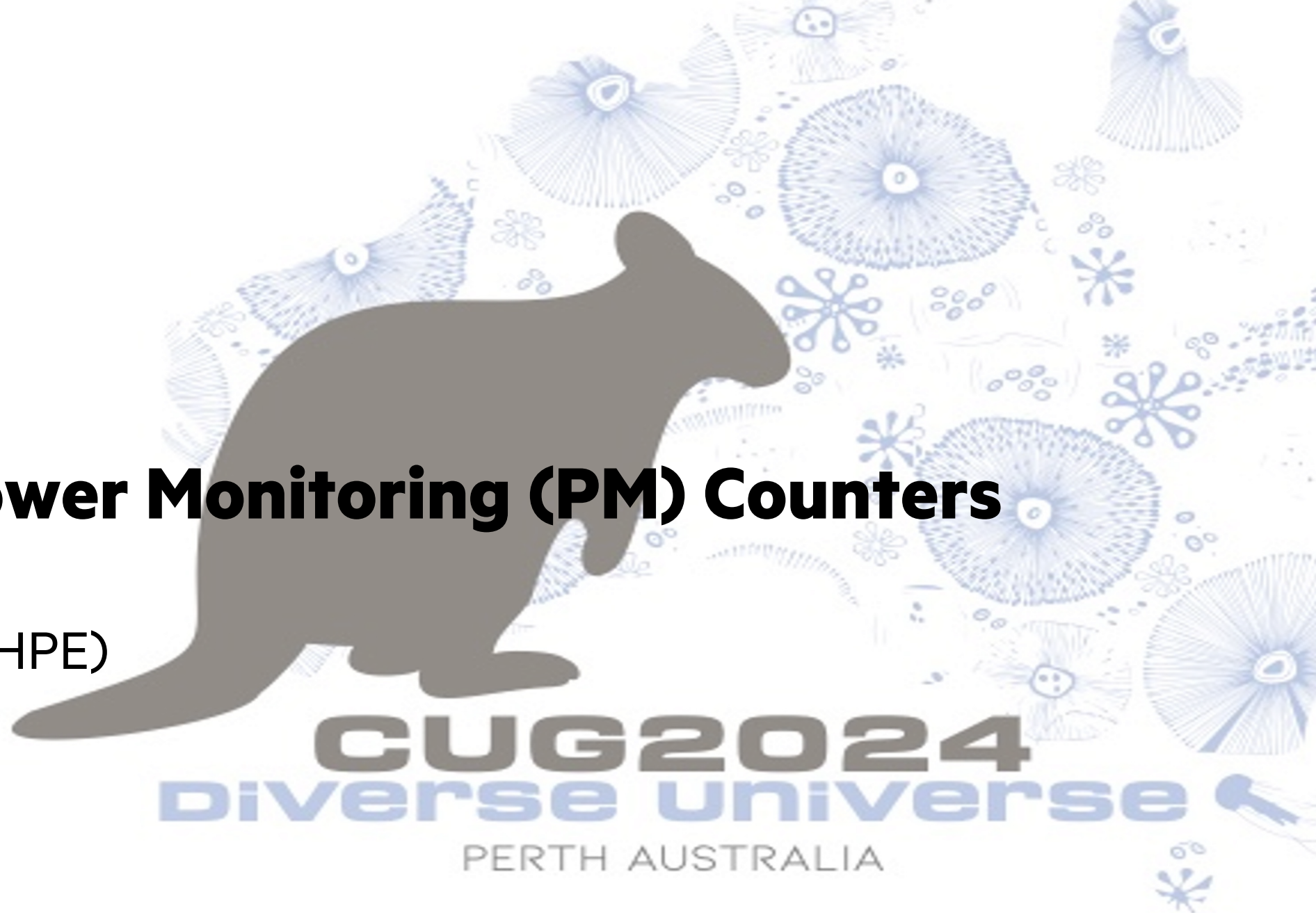


Hewlett Packard
Enterprise

HPE Cray Power Monitoring (PM) Counters

Steven J. Martin (HPE)

May, 2024



HPE Cray Power Monitoring (PM) Counters -- Agenda

- Brief history and overview of PM Counters basics
- Quick look at PM Counters on HPE Cray Supercomputer blades in full production
- More detailed look at latest HPE Cray EX Supercomputer blades announced at SC23
- Opportunities and challenges in supporting PM Counters on NGI (Next Generation Infrastructure)




Brief History of PM Counters

- PM Counters were first developed for the Cray XC30 Supercomputer
 - Initially supported two-socket CPU only nodes and then nodes with accelerators (GPUs)
- PM Counters designed to enable low overhead user access to power and energy data
 - No runtime jitter or performance impact when counter are not accessed
 - Very low impact when read at reasonable intervals

- PM Counters: Version 1

- From 2014 CUG slides
- Total node power and energy
- Accelerator (1 GPU) power and energy
- Power capping data
- Basic meta-data



```
/sys/cray/pm_counters/accel_energy:24675886 J
/sys/cray/pm_counters/accel_power:22 W
/sys/cray/pm_counters/accel_power_cap:0 W
/sys/cray/pm_counters/energy:71224823 J
/sys/cray/pm_counters/freshness:4516770
/sys/cray/pm_counters/generation:9
/sys/cray/pm_counters/power:62 W
/sys/cray/pm_counters/power_cap:425 W
/sys/cray/pm_counters/startup:1396011015159068
/sys/cray/pm_counters/version:1
```

Brief History of PM Counters – Early Papers

- Cray XC30 power monitoring and management
 - https://cug.org/proceedings/cug2014_proceedings/includes/files/pap130.pdf
 - https://cug.org/proceedings/cug2014_proceedings/includes/files/pap130-file2.pdf
- User-level Power Monitoring and Application Performance on Cray XC30 supercomputers
 - https://cug.org/proceedings/cug2014_proceedings/includes/files/pap136.pdf
 - https://cug.org/proceedings/cug2014_proceedings/includes/files/pap136-file2.pdf
- First Experiences With Validating and Using the Cray Power Management Database Tool
 - https://cug.org/proceedings/cug2014_proceedings/includes/files/pap148.pdf
 - https://cug.org/proceedings/cug2014_proceedings/includes/files/pap148-file2.pdf
 - <https://arxiv.org/pdf/1408.2657>
- Measurement and interpretation of micro-benchmark and application energy use on the cray xc30
 - <https://www.nersc.gov/assets/AustinWrightE2SC14.pdf>

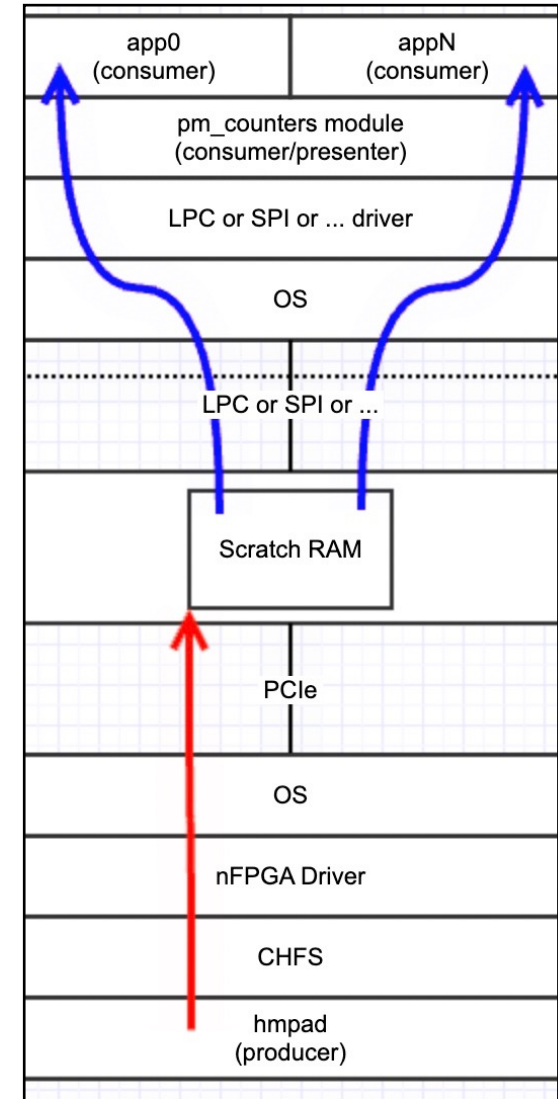
PM Counters Versions

- PM counters version 1 initial support
 - Total node power, energy, and power cap
 - Accelerator power, energy, and power cap
 - Meta-data: freshness, startup, generation, and version
- PM counters version 2 added
 - Aggregate sensors
 - cpu_energy, memory_energy, cpu_power, and memory_power
 - raw_scan_hz meta-data
- PM counters version 3 added
 - Timestamps in us (microsecond) on all telemetry files
 - Enabling more accurate data calculations
 - CPU Thermal data in Celsius



Overview of PM Counters

- PM counter data is produced by the node controller out-of-band
 - Data written into shared memory space at 10 Hz update rate
 - Data structure overwritten on each update
 - PM counter data is consumed in-band via reading files in sysfs
 - The kernel module **bpmcdmod** provides the /sys/cray/pm_counters functionality
 - Use '**modprobe bpmcdmod**' to load the module if it's not loaded
 - **bpmcdmod** caches data, so oversampling counters has minimal performance impact
- PM counter data is supported by the HPE Cray Programming Environment (CPE)
 - cray_pm - provide access to Cray Power Management (PM) counters
 - see '**man 5 cray_pm**' and '**PAPI_native_avail**' utility



PM counters support – Workload Management Energy Monitoring

- ATOM (Application Task Orchestration and Management) energy reports
 - HPE Portable Batch System Installation Guide: HPCM on HPE Cray EX Systems
 - (2.0.3) (S-8056), Part Number: S-8056, Published: February 2024
 - https://support.hpe.com/hpesc/public/docDisplay?docId=dp00004061en_us&docLocale=en_US
 - When configured, it captures per-node energy data whenever applications or workload manager jobs run
- Slurm:
 - slurm.conf: acct_gather_energy/pm_counters
 - https://slurm.schedmd.com/slurm.conf.html#OPT_AcctGatherEnergyType
 - Energy consumption data is collected from the Baseboard Management Controller (BMC) for HPE Cray systems



PM counters support/use in Open source

- pm_mpi_lib and pat_mpi_lib:
 - Apache-2.0 license
 - https://github.com/cresta-eu/pm_mpi_lib
 - https://github.com/cresta-eu/pat_mpi_lib
 - Same as pm_mpi_lib, but uses Cray Performance and Analysis Tools
- Power Measurement Toolkit (PMT)
 - <https://git.astron.nl/RD/pmt>
 - The Netherlands Institute for Radio Astronomy (ASTRON) and École Polytechnique Fédérale de Lausanne (EPFL)



PM Counters on Generally Available HPE Cray EX Blades

EX425, EX420, EX4252, EX235n, EX235a



HPE Cray EX425, EX420, EX4252 -- CPU Only Blades

- HPE Cray EX425:
 - AMD Rome or Milan
- HPE Cray EX420:
 - Intel Sapphire Rapids
- HPE Cray EX4252:
 - AMD Genoa or Bergamo
- These three blades all support the same set of files
 - CPU: `cpu[0-1]_temp, cpu_energy, cpu_power,`
 - Memory: `memory_energy, memory_power`
 - Node: `energy, power, power_cap`
 - Meta-data: `freshness, generation, raw_scan_hz, startup, version`



HPE Cray EX235n, EX235a – GPU Accelerated Blades

- HPE Cray EX235n :
 - AMD Milan CPU + 4x NVIDIA A100 GPUs
- HPE Cray EX235a :
 - AMD Trento CPU + 4x AMD MI250 GPUs
- These two blades both support the same set of files
 - GPU: accel[0-3]_energy, accel[0-3]_power, accel[0-3]_power_cap
 - CPU: cpu0_temp, cpu_energy, cpu_power,
 - Memory: memory_energy, memory_power
 - Node: energy, power, power_cap
 - Meta-data: freshness, generation, raw_scan_hz, startup, version



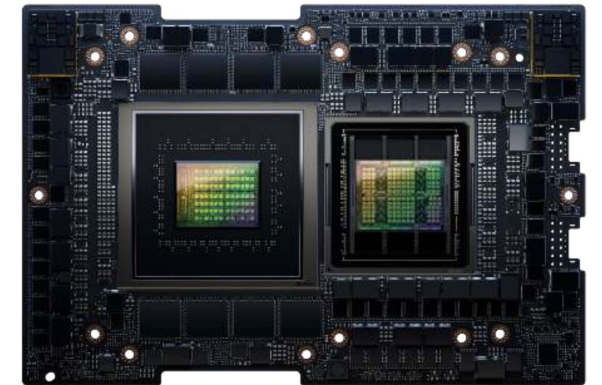
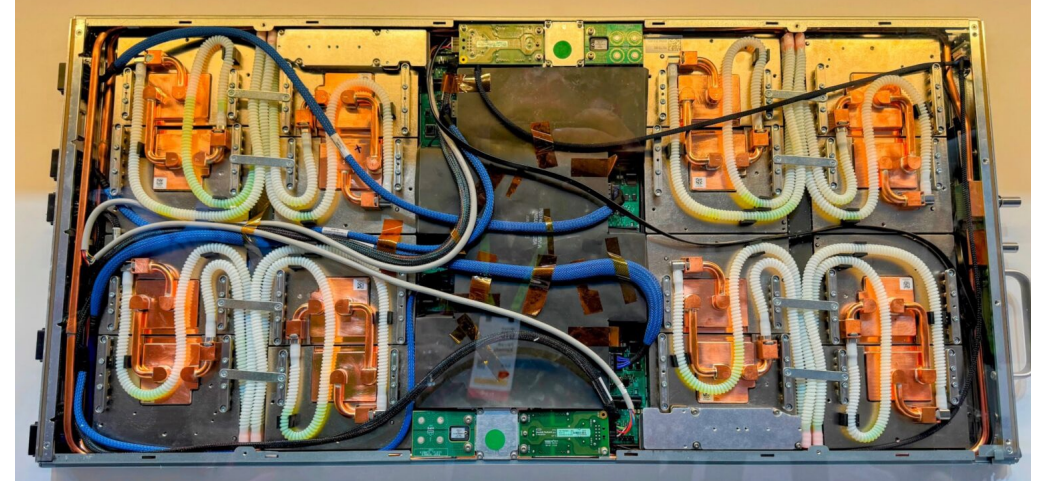
PM Counters on New HPE Cray EX Blades

HPE Cray EX254n and HPE Cray EX255a



HPE Cray EX254n

- Two nodes per blade
 - 4 NVIDIA Grace Hopper GH200 Superchips per node
 - NVIDIA GH200 Grace Hopper Superchip (public specifications)
 - 72 Arm Neoverse V2 cores with up to 480GB LPDDR5X memory
 - NVIDIA H100 Tensor Core GPU with 96GB or 144GB HBM3e
 - 450 to 1000 W (programmable) TDP (CPU + GPU + memory)
 - Quad injection Slingshot 200 per node
- PM Counters data
 - Total node power and energy from SIVOC
 - CPU and Accelerator (GPU) data from NVIDIA modules
 - PLDM for CPU
 - SMBPBI for GPU



HPE Cray EX254n Node -- PM Counter Overview

- Total Files: 34
 - accel[0-3]_energy: J (Joules) # GPU energy
 - accel[0-3]_power: W (Watts) # GPU power
 - accel[0-3]_power_cap: W (Watts) # GPU power cap
 - cpu[0-3]_energy: J (Joules) # CPU energy
 - cpu[0-3]_power: W (Watts) # CPU power
 - cpu[0-3]_temp: C (Celsius) # CPU temperature
 - cpu_energy: J (Joules) # Energy for all 4 CPUs
 - cpu_power: W (Watts) # Power for all 4 CPUs
 - energy: J (Joules) # Energy for the Node
 - power: W (Watts) # Power for the Node
 - power_cap: W (Watts) # Power cap for Node
 - freshness: Counter # Increments at raw_scan_hz (10 Hz)
 - generation: Counter # Increments when a power cap is changed
 - raw_scan_hz: Rate # Update rate for PM Counters
 - startup: Timestamp # Timestamp of counter subsystem
 - version: Revision # Revision of protocol

New in EX254n



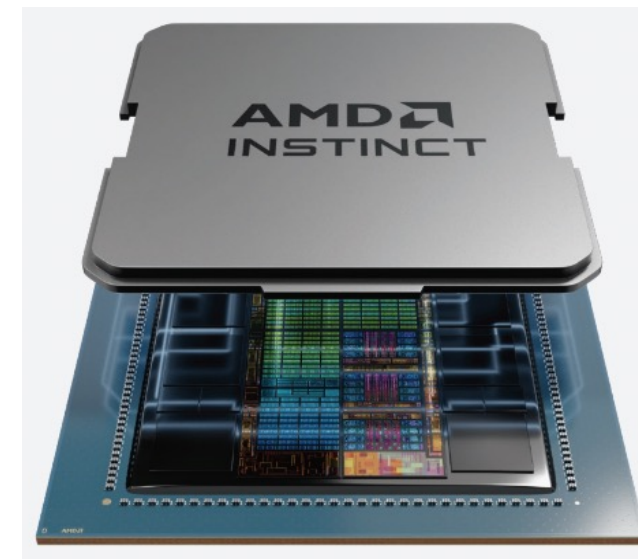
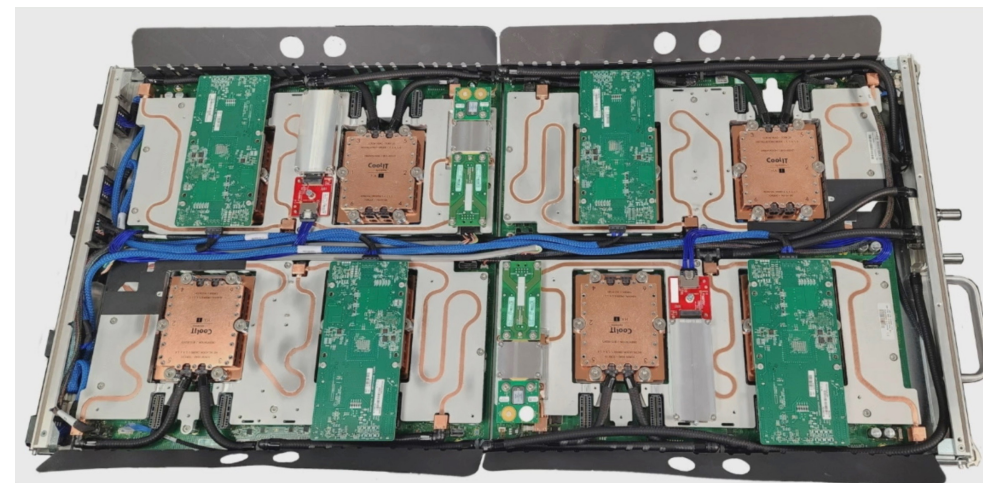
HPE Cray EX254n Node -- PM Counter → Redfish URI Mapping

PM Counters	Redfish URI
accel[0-3]_power	/redfish/v1/Chassis/Node0/Sensors/GPU[0-3]Power
cpu[0-3]_power	/redfish/v1/Chassis/Node0/Sensors/CPU[0-3]Power
cpu[0-3]_temp	/redfish/v1/Chassis/Node0/Sensors/CPU[0-3]Temperature
cpu_power	/redfish/v1/Chassis/Node0/Sensors/ChassisCPUSubsystem0OutputPower
cpu_energy	/redfish/v1/Chassis/Node0/Sensors/ChassisCPUSubsystem0OutputEnergy
power	/redfish/v1/Chassis/Node0/Sensors/ChassisVoltageRegulator0InputPower
energy	/redfish/v1/Chassis/Node0/Sensors/ChassisVoltageRegulator0InputEnergy
power_cap	/redfish/v1/Chassis/Node0/Controls/NodePowerLimit (SetPoint/ControlMode)



HPE Cray EX255a

- Two nodes per blade
 - 4 AMD Instinct™ MI300a APUs per node
 - AMD MI300a (public specifications)
 - 24 AMD ‘Zen 4’ x86 CPU cores
 - 228 GPU compute units
 - 128 GB HBM3
 - 760 W Max TDP when liquid cooled
 - Quad injection Slingshot 200 per node
- PM Counters data
 - Total node power and energy telemetry from Infineon XDP710
 - Capable of Precision input power monitoring and reporting $\leq 2\%$
 - Accelerator (MI300a APU) telemetry data from MPS-MPXXXX



HPE Cray EX255a Node

• Total files: 20

• accel[0-3]_energy:	J (Joules)	# APU energy	→ 1 st Blade with APU Sockets
• accel[0-3]_power:	W (Watts)	# APU power	
• accel[0-3]_power_cap:	W (Watts)	# APU power cap	
• energy:	J (Joules)	# Energy for the Node	
• power:	W (Watts)	# Power for the Node	
• power_cap:	W (Watts)	# Power cap for Node	
• freshness:	Counter	# Increments at raw_scan_hz (10 Hz)	
• generation:	Counter	# Increments when a power cap is changed	
• raw_scan_hz:	Rate	# Update rate for PM Counters	
• startup:	Timestamp	# Timestamp of counter subsystem	
• version:	Revision	# Revision of protocol	



HPE Cray EX255a Node -- PM Counter → Redfish URI Mapping

PM Counters	Redfish URI
accel[0-3]_power	/redfish/v1/Chassis/Node0/Sensors/Accelerator[0-3]VoltageRegulator0InputPower
accel[0-3]_energy	/redfish/v1/Chassis/Node0/Sensors/Accelerator[0-3]VoltageRegulator0InputEnergy
power	/redfish/v1/Chassis/Node0/Sensors/ChassisVoltageRegulator0InputPower
energy	/redfish/v1/Chassis/Node0/Sensors/ChassisVoltageRegulator0InputEnergy
power_cap	/redfish/v1/Chassis/Node0/Controls/NodePowerLimit (SetPoint/ControlMode)





PM Counters on Next Generation Infrastructure

Forward looking subject to change

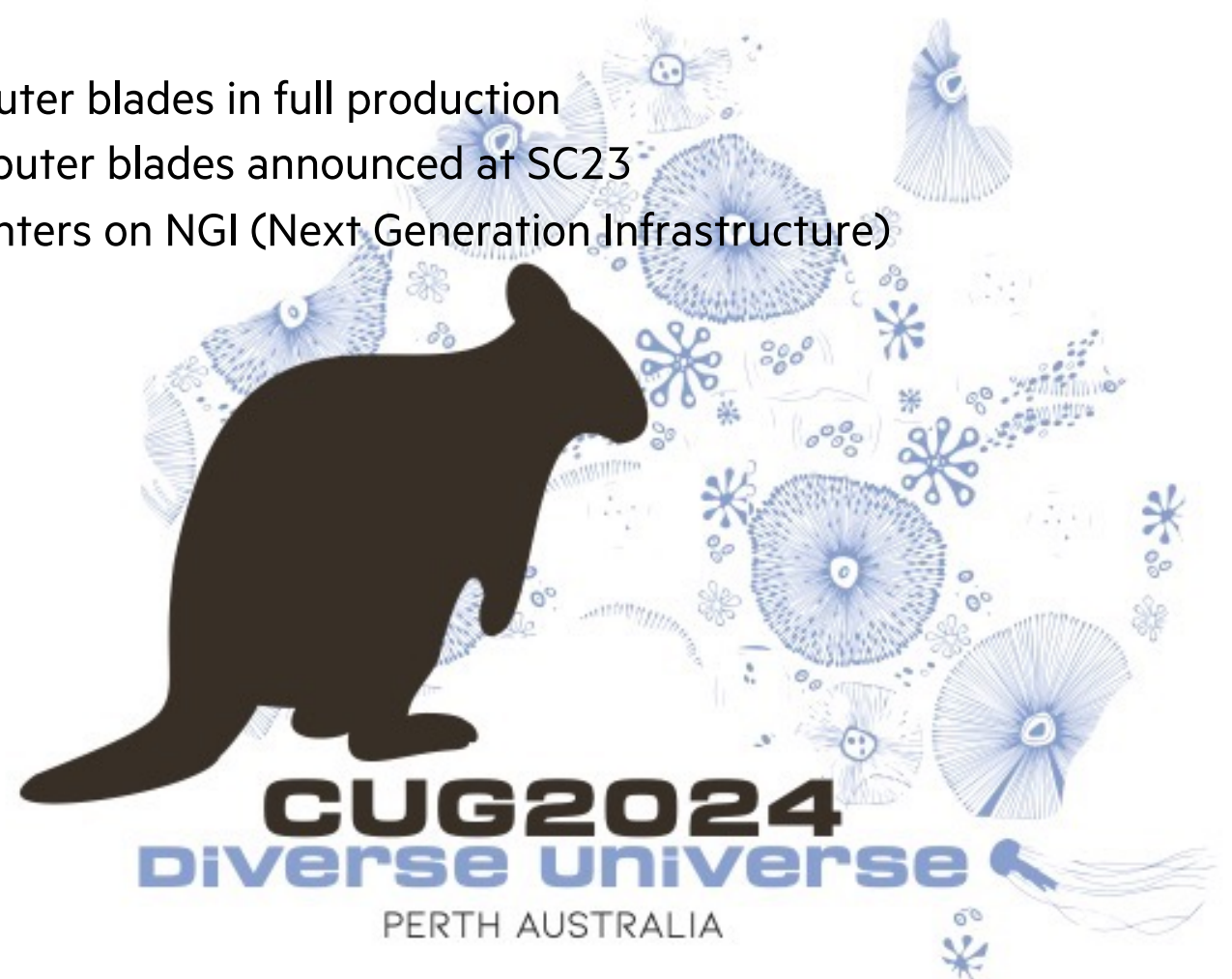


PM Counters on NGI (Next Generation Infrastructure)

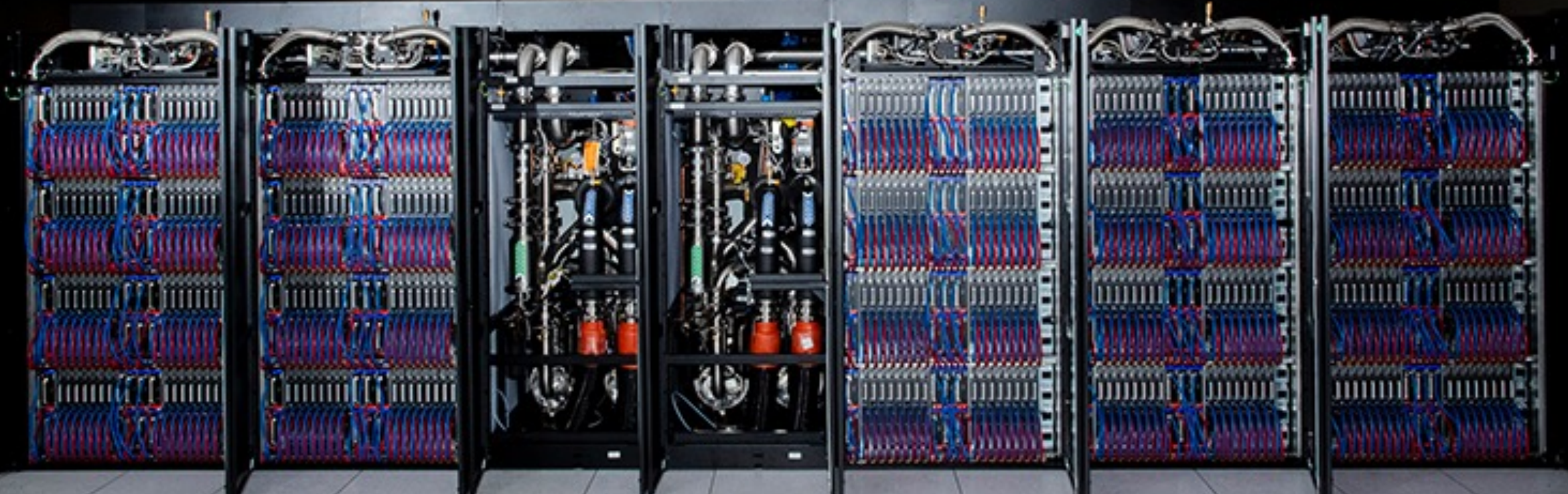
- Motivation
 - Worldwide emphases on sustainability
 - You can't control what you can't measure
- Opportunities:
 - NGI blade architecture will enable collection of high-quality node level power, energy
 - Independent of partner developed components
 - Consistence over multiple blade types and generations
- Challenges:
 - Access to socket-level and component-level power, energy, and thermal data at or above 10Hz
 - May become more incumbered as we leverage partner developed components
 - Greater diversity in partner developed IP
 - Faster development cycles

HPE Cray Power Monitoring (PM) Counters -- Wrap-up

- What was covered
 - Brief history and overview of PM Counters basics
 - Quick look at PM Counters on HPE Cray Supercomputer blades in full production
 - More detailed look at latest HPE Cray EX Supercomputer blades announced at SC23
 - Opportunities and challenges in supporting PM Counters on NGI (Next Generation Infrastructure)
- Questions?
- Contact information and acknowledgements
 - Steven Martin: steven.martin3@hpe.com
 - Brian Collum: brian.collum@hpe.com
 - Sean Byland: sean.byland@hpe.com



Backup

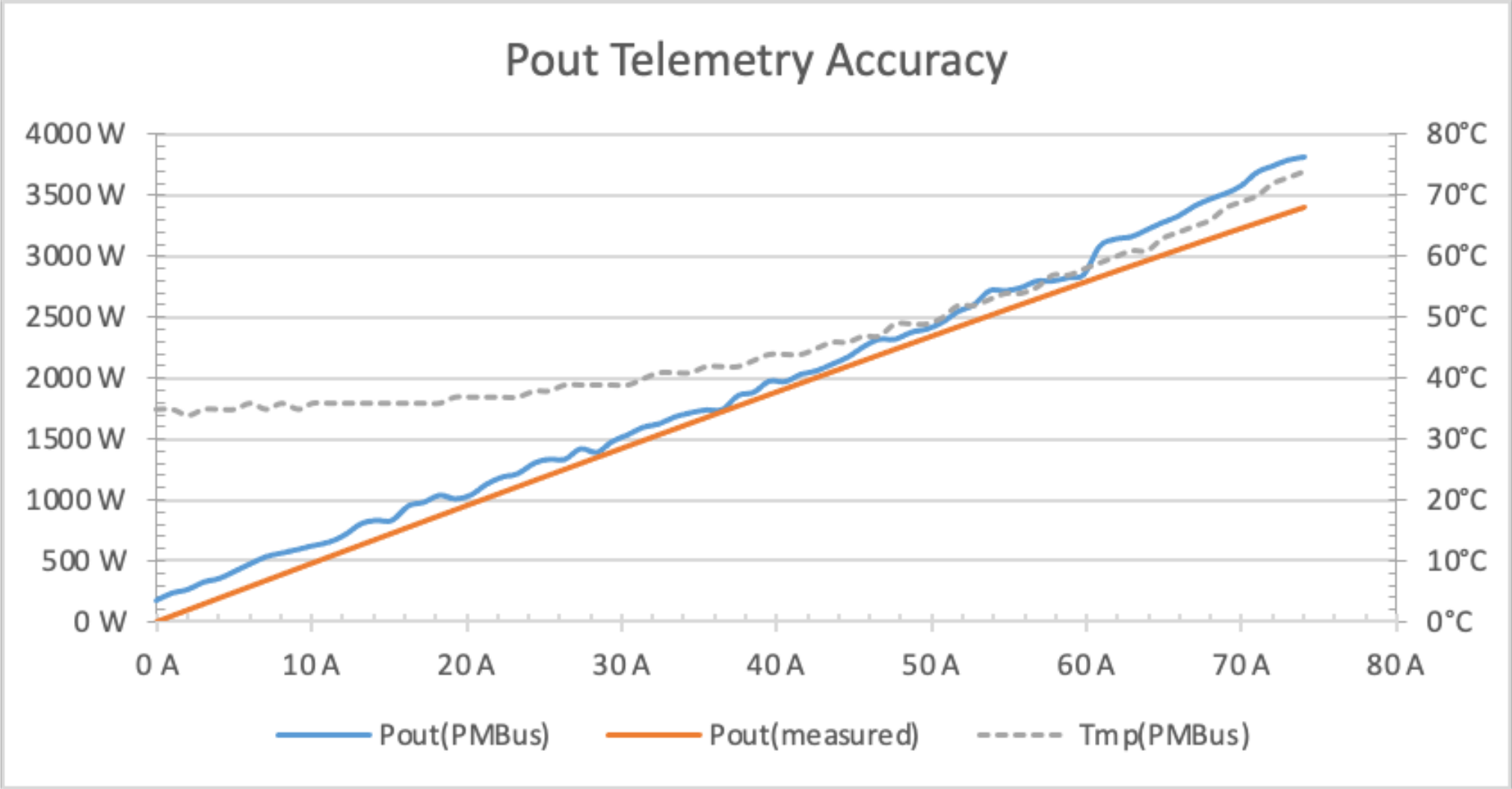


CUG24 Presentation Abstract

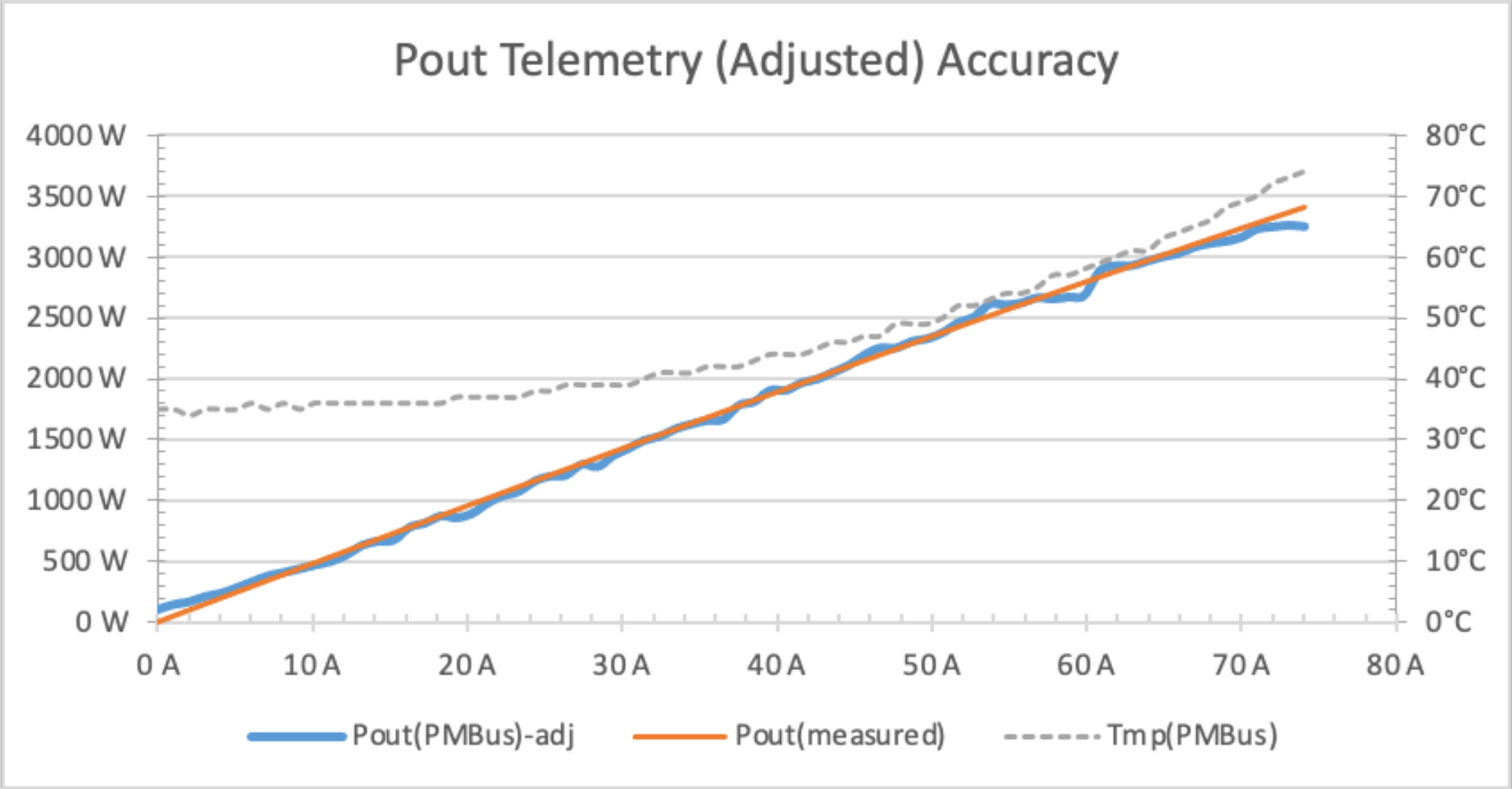
HPE Cray Power Monitoring (PM) Counters were first deployed on Cray XC30 systems, and several papers were presented at CUG in 2014 that described their use. PM Counters expose power energy and related meta-data collected out of band, directly to in-band consumers. Since introduction, PM Counters have been supported on all blades designed for use in Cray XC, and HPE Cray EX Supercomputer systems. PM Counters have continued to be important as system and application power and energy consumption continues to be a top priority for system vendors, application developers, and the wider HPC research community. Over the last decade, the design of PM Counters has remained very stable, with only minor updates to support evolving node architecture changes. This presentation will give a brief history and overview of PM Counters basics, it will then present details of PM Counters on the latest HPE Cray EX supercomputer blades announced at SC23, and then discusses opportunities and challenges in supporting PM Counters on NGI (Next Generation Infrastructure). The presentation will conclude with a reinforcement of the value of PM Counters in supporting research, development, and testing of energy efficient and sustainable HPC systems.



SIVOC Telemetry -- Before



SIVOC Telemetry -- After



Full Chassis: SIVOC Data vs Rectifier Data

