E2 HPC Working Group

SC16 Salt Lake City, Utah | hpc matters.

Sustainably Supporting Science
through committed community action

Energy Efficiency Considerations and HPC Procurement

BoF Session

November 15, 2016

Presentation

Energy Efficiency Considerations and HPC Procurement

ask a question · give feedback

PRIMARY SESSION LEADER: Natalie Bates

SECONDARY SESSION LEADERS: Ladina Gilly, James Laros, James Rogers, Anna Maria Bailey, Daniel Hackenberg, Marek Michalewicz, Bilel Hadri, Thomas Ilsche

EVENT TYPE: Birds of a Feather

EVENT TAGS: Energy | HPC Center Planning and Operations | Intermediate | Power | State of the Practice

TIME: Tuesday, November 15th, 5:15pm - 7pm

E2 HPC Working Group

SC16 Salt Lake City, Utah | hpc matters.

Panel members for today's BoF

Anna Maria Bailey – LLNL
Paul Coteus – IBM
Daniel Hackenberg – TU Dresden
Bilel Hadri - KAUST
Jim Laros - Sandia
Steve Martin – Cray Inc.

E2 HPC Working Group

Sustainably Supporting Science
through committed community action

SC16 Salt Lake City, Utah | hpc matters.

Introduction of the EE HPC WG Document

Jim Laros, Sandia National Laboratories

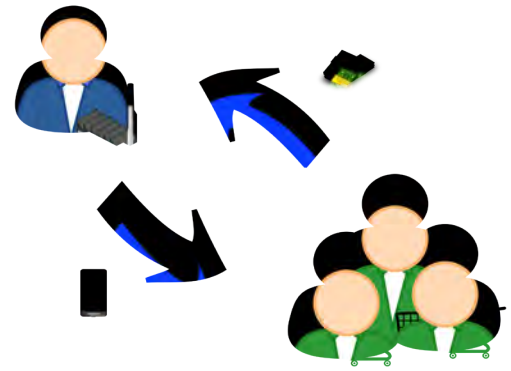
Energy Efficiency Considerations for HPC Procurements

November 15, 2016

Aim of the Document

https://eehpcwg.llnl.gov/pages/compsys_pro.htm

- To put things in perspective – we started in 2012
  - Most efforts were still in the research phase
- Audience:
  - HPC consumers – purchase/manage/use HPC system
  - HPC vendors – provide/sell some aspect of HPC system to consumers
  - HPC community – HPC consumers + HPC vendors
- Document targeted at the entire HPC community
- Living Document – Quickly evolving space
- Document serves as the basis for:
  - Information
    - To HPC "consumers"– what should you consider when writing your next procurement document
      - Not intended to be a cut and paste resource
    - To Vendor community – what to expect as requirements from the HPC community in the short and longer term
  - and Discussion
    - HPC consumers – what we want
      - Not as easy as it might seem
    - HPC community – socialize what we (consumers) want with what can be provided (vendors)
      - Note: what we want typically wins ☺
- Introduction – good source of what document is and is NOT!



Producer - Consumer

APPROACH

- Leverage existing expertise in the area
- Recognized different needs
  - System/Platform/Cabinet
  - Node
  - Component
- Express importance and forecast/predict what we will need
  - **Mandatory** – confident it can be delivered soon
  - **Important** – "think" this is reasonable in the near/mid term
  - **Enhancing** – what we really want even though we likely can't get it today
- Generated lots of lively vendor feedback ☺

System/Platform/Cabinet

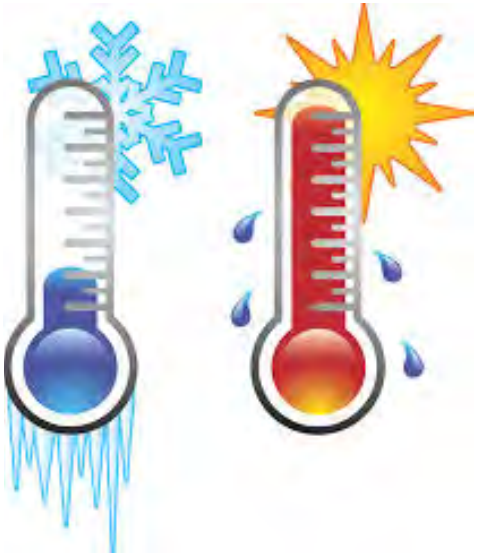
	Internal Sampling Frequency
Mandatory	≥ 10 per second
Important	≥ 100 per second
Enhancing	≥ 1000 per second

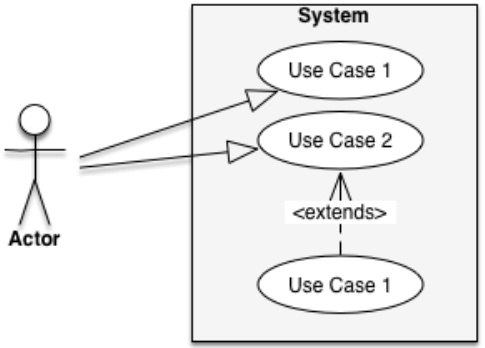
		External Reported Value Frequency
Mandatory	Discrete Power (W)	≥ 1 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 1 per second
Important	Discrete Power (W)	≥ 10 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 1 per second
Enhancing	Discrete Power (W)	≥ 100 per second
	Average Power (W)	≥ 1 per second
	Energy (J)	≥ 10 per second

Also Node and Component

Other Topics

- Timestamps and Clocks
- Temperature Measurement
- Benchmarks
- Cooling
- High Level Objectives
  - TCO, PUE, TUE, ERE
- Use cases
  - Exercising all of these topics





E2 HPC Working Group

Sustainably Supporting Science
through committed community action

SC16 Salt Lake City, Utah | hpc matters.

Considerations from APEX

Jim Laros, Sandia National Laboratories

November 15, 2016

Influenced Trinity RFP

Power measurement and control capabilities (hardware and software tools and application programming interfaces (APIs)) are necessary to meet the needs of future supercomputing energy and power constraints.

1. Describe all power related measurement and control features, capabilities and limitations (hardware and software) of the system including, but not limited to, any tools, system software features and APIs that will be made available at initial acceptance.
2. Describe all power related measurement and control capabilities projected on the Offeror's road map. LANS, UC, and the Subcontractor will work cooperatively to define a set of capabilities that will be delivered beyond initial acceptance.
3. Describe all power related measurement and control capabilities (hardware and software) that would necessitate hardware upgrade or replacement.

Response was used as the basis to develop an Advanced Power Management NRE program to implement the HPC Power API

Influenced Crossroads RFP

Power, energy, and temperature will be critical factors in how the APEX laboratories manage systems in this time frame and must be an **integral part** of overall Systems Operations. The solution must be well integrated into other intersecting areas (e.g., facilities, resource management, runtime systems, and applications). The APEX laboratories expect a growing number of use cases in this area that will require a vertically integrated solution.

- The Offeror shall describe all power, energy, and temperature measurement capabilities (system, rack/cabinet, board, node, component, and sub-component level) for the system, including control and response times, sampling frequency, accuracy of the data, and timestamps of the data for individual points of measurement and control.
- The system should include an integrated API for all levels of measurement and control of power relevant characteristics of the system. It is preferable that the provided API complies with the High Performance Computing Power Application Programming Interface Specification (http://powerapi.sandia.gov).
- And 8 more specific requirements.

E2 HPC Working Group

Sustainably Supporting Science
through committed community action

SC16 Salt Lake City, Utah | hpc matters.

Lessons learned from CORAL procurement

Anna Maria Bailey
Lawrence Livermore National Laboratory
November 15, 2016

CORAL Lessons Learned

CORAL Procurement was a collaborative procurement between Oak Ridge, Argonne and Lawrence Livermore National Laboratories. Energy Efficiency was a core competency of the evaluation process and criteria.

On the facility side:

- Provide as much detail about the facility so that the vendors can specify their solution in detail to make informed energy and sustainable solutions
- Air vs. liquid cooling
- Rear door heat exchangers vs. air cooling room solutions
- AC vs. DC solutions
- 480V vs. 208V solutions

CORAL Lessons Learned

On the system side:

- Evaluate the micro-architectural features that support power efficiency
  - Does the processor support DVFS on a per core basis?
  - What is the response time of power gating and frequency adjustments?

Overall evaluation should include:

- Best value selection process to score and rank each vendor utilizing energy efficiency as a key performance parameter
- Cost benefit analysis tables to evaluate total cost of ownership
- A range of technical expertise involved in the evaluation and for the holders of that range need be cognizant of energy efficiency from the facility to the system

Taurus procurement – lessons learned

Energy Efficiency Considerations and HPC Procurement

November 15th 2016

Daniel Hackenberg (daniel.hackenberg@tu-dresden.de)

High Definition Energy Efficiency Monitoring (HDEEM) on taurus at TU Dresden

- Our requirements were specified quite well in the RFP, with the key targets:
  - Accuracy
  - Temporal granularity
  - Spatial granularity
  - Scalability
- Approach of our vendor Bull/Atos:
  - Setup project for collaborative development
  - Funding of two scientists for five years (2013-2017) at TU Dresden
  - All major goals and production level quality reached after ~3.5 years (mid 2016)

HDEEM Status after 3.5 Years of Development

	Energy Accounting	High Definition Measurement
Domain	Full system	1456 Haswell nodes
Interface	Slurm batch system	HDEEM API
Node level measurement	1 sample/s 2% accuracy (calibrated)	1000 sample/s 2% accuracy (calibrated)
CPU/DRAM measurement	N/A	100 samples/s for 2xCPU and 4x DIMM 5% accuracy
Data access	In-band	In-band or out-of-band
Overhead	Diminishable overhead	Post mortem data access no perturbation during measurement
Timestamps	Timestamping close to the measurement with synchronized clock	
Correctness	Verified power **and** energy measurements	

Key Challenges and Lessons Learned

- Energy correctness
  - Calibrated power measurements
  - Correct timestamps
  - Correct data processing on different HW/SW components
- Low-latency API
  - Turns out that users need it!
  - Post-mortem analysis not always feasible
- Production-readiness
  - Userspace access (non-root)
  - Stability, error handling etc.
- Creating a common understanding about these challenges is another challenge by itself

- Our energy correctness and API requirements
  - apparently differ from most other sites
  - could only be partially integrated into the procurement document
  - were met by no vendor in 2012
  - are met (only?) by Bull/Atos in 2016
- How to avoid this effort for the next system?
  - We need the HDEEM API *functionality* from the next vendor
  - PowerAPI: very welcome, but so far lacking a sufficiently scalability implementation
  - Few HPC vendors have experiences with professional power measurement infrastructures
  - Even RAPL is a contender for the best solution in the next system

Lessons Learned from Shaheen2 Procurement

Bilel Hadri
KAUST Supercomputing Laboratory

SC16 BoF: Energy Efficiency Considerations and HPC Procurement

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
SHAHEEN
SUPERCOMPUTING LABORATORY

Shaheen 2 Procurement

- Constrained by site power and cooling availability
  - During acceptance: 2.9 MW limit
  - After acceptance: 2.3 MW limit before the decommissioning of Shaheen 1 BG/P- 16 racks (~500 kW)

- Shaheen2: 36 cabinet Cray XC40, 197,568 Haswell cores
  - Rmax: 7.2 PF Rpeak=5.53 PF
  - **Power 2.83 MW with** LINPACK, Peaks reached 2.9 MW
  - Technically, it can reach up to 3.5 MW

➔ Critical Power and Cooling constraints. Procurement strategies needed.

What can you recommend as best practices?

- Before the call of proposal:
  - Determine data center limitations, accurate inventory of all systems
  - Get and analyse the power and cooling usage
  - Validate your study by third party and future OEMs (bring them on site)
- During the procurement/evaluation phase:
  - Power and cooling requirement must be clearly stated
  - Perform your own test on early access: most vendors provide such service (both chip manufacturer and OEMs) -- this is a must for new technologies
  - Don’t assume technical specs are correct, measure it, check it with other sites running similar platforms
  - Take into consideration all components (nodes/network/services/PFS…)
- Acceptance: the real test
  - Power and cooling tests should be part of the acceptance site
  - Factory Acceptance Test, detect potential issues before SAT
  - Monitoring real-time power usage

Lessons Learned

- Acceptance
  - Fostering broad collaboration with different key players (chip vendors, OEM, WM, data center, E&PM, campus facilities….)
  - Bridging different field of expertise for successful deployment and acceptance of large HPC system.
  - Identifying specific power metrics and measurement (DC/AC, frequency, average vs peak, node vs cabinet, system overall, w or w/o service node, PFS…)
- Production mode
  - Using real-time power system usage: new approach for improving applications efficiency
    - Power profiling of applications, especially the full scale ones
      - Used when strategizing/optimizing full scale Gordon Bell runs on Shaheen2
      - Detecting issues on applications performance (known compute intensive code drawing less than 200W per node - Found issue in the communication pattern)
  - LINPACK is not the most consuming applications(Memtest, Nekboxtester, MOAO)
  - Power is a scarce resource: power capping brought further awareness to implement energy efficient codes(communication/synchronization reducing).
- Future procurement and upgrade: specify wall plate, peak and nominal power.

E2 HPC Working Group

SC16 Salt Lake City, Utah | hpc matters.

Sustainably Supporting Science
through committed community action

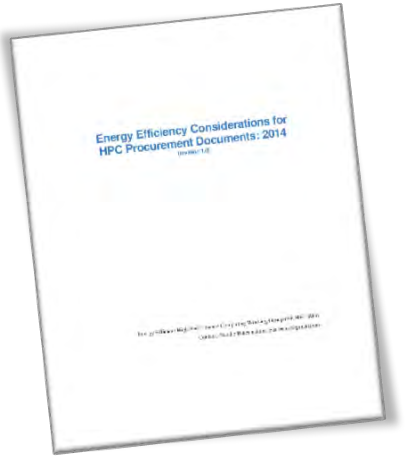
Energy Efficiency Considerations and HPC Procurement

Steven J. Martin (stevem@cray.com)

November 15, 2016

Cray Motivation for Enhanced Monitoring

- **Customer and market demand**
  - Sandia Power API UseCase-powapi.pdf (2013)
  - Energy Efficiency Considerations for HPC Procurement Doc (2014)
  - Trinity Procurement and Trinity APM NRE contracts



- **Research & Development**
  - Enhanced reliability, availability, and serviceability (RAS)
  - Improved performance tuning and analysis opportunities

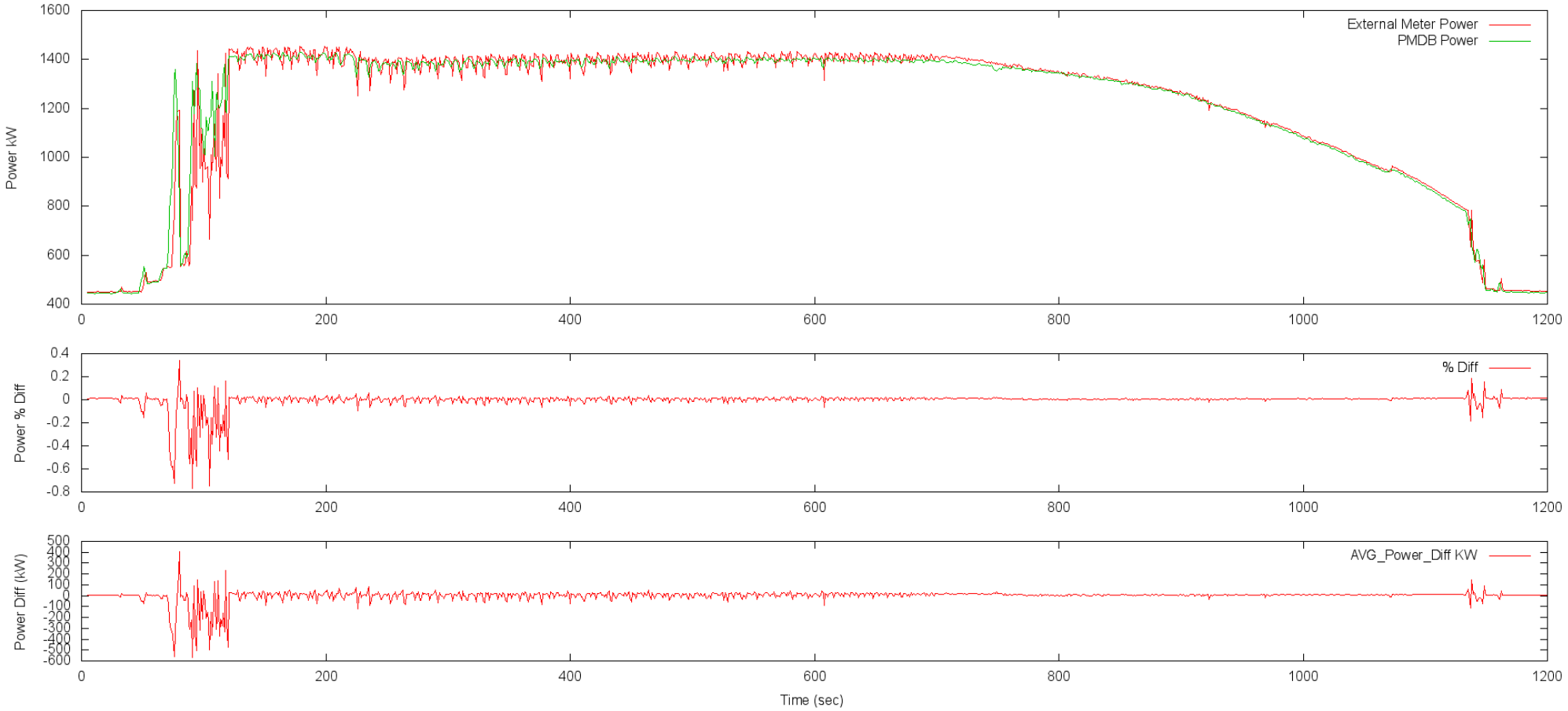
Copyright 2016 Cray Inc.

EEHPC Considerations & Procurement

- **Only use "Mandatory" in an RFP if willing to disqualify a vendor(s)**
  - Words: mandatory, important, or enhancing vs good, better, best…
- **Document your uses case(s)**
  - Clear use case documentation enables vendors ability to deliver
  - Information helps vendors enable what you need
- **Requiring 1% accuracy (for example) may drive up cost where 5% may enable your use case**
  - Balance between enabling features and pricing a system out of the market

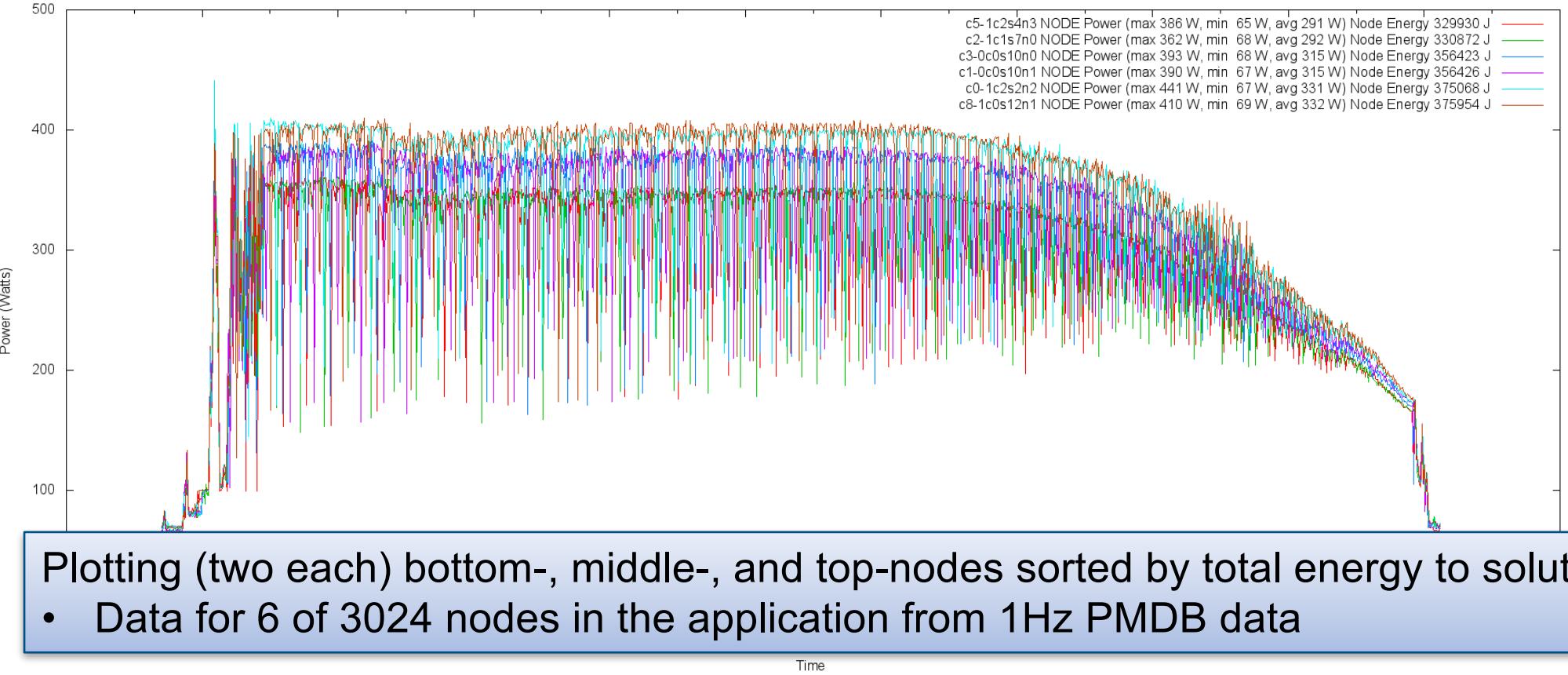
Copyright 2016 Cray Inc.

Cray's Internal Data (PMDB) vs External Meter



Copyright 2016 Cray Inc.

Cray's Internal (PMDB) Node-level Power data



Plotting (two each) bottom-, middle-, and top-nodes sorted by total energy to solution

- Data for 6 of 3024 nodes in the application from 1Hz PMDB data

Copyright 2016 Cray Inc.

Energy Efficiency Considerations for HPC Procurement Documents

IBM Comments on 2014 version, Rev 1.0

For EE HPC WG at SC16

Paul Coteus
IBM Fellow and Chief Engineer IBM Data Centric Systems

Overall Comments

- Motivation is excellent
  - Good instrumentation can allow power aware code optimization, dynamic power management, and energy aware scheduling, and can be used to differentiate suppliers equipment and platform.
  - Meaningful measurements require good time synchronization.
  - A hierarchical approach to data collection and data processing is reasonable and fairly well explained.

- IBM can likely meet if not exceed the spirit of the "mandated" capabilities, in part by establishing energy interfaces with 3rd party components.
  - In some cases development will be needed. However:

- Measurement accuracy and precision is (in places) inadequately specified and/or too stringent for a mandatory requirement.
  - Usually an offerer is denied a contract if all mandatory requirements are not met. For that reason, there should be few mandatory requirements. Those that remain should be essential and clear in how they are determined.

- Measurement frequencies are (in places) too aggressive for a mandatory requirement.
  - There should be a baseline which most if not all suppliers should be able to meet, and which are useful in other environments (i.e., cloud, computing as a service, …

Examples of Specific Concerns

- +/- some amount is too vague.
  - Do we mean +/- one standard deviation, full width, or something else?

- Lets go metric and tailor resolution to the task.
  - +/- 1 °F is too restrictive and wrong unit, +/-0.5C is just too restrictive.

- Mandated, non-impactful external readouts of >100 per second, per node component, is too fast for an exascale system of 10^5-10^6 components!
  - What dedicated processing system will accumulate, analyze, reduce and store this information at full rate? To what end?
  - The issue is one of global measurement capability. We can measure a node at high rate, but to measure all nodes at that high rate is challenging and perhaps not necessary.

- Why are hierarchical measurements (slower rate measurements for a cabinet than for a node) required if all the data MUST be made available?

An Issue of Timeliness …

- We are discussing a “2014” document that was intended to influence a computing system to be delivered and accepted in 2016.
  - A 2year head start is not nearly enough time given recent trends.

- Even if mandated requirements are toned down, it is likely that substantial new energy measurement capability will be required for an Exascale system.

- To ensure compliance, vendors need guidance several years before the RFQ
  - Today would not be too early for DoE labs to state if they intend to make the suggestions of the EE HPC WG into contractual requirements for Exascale, or even options which would influence a purchasing decision

E2 HPC Working Group

Sustainably Supporting Science
through committed community action

SC16 Salt Lake City, Utah | hpc matters.

Thank you for your attention! Questions welcome - Let's discuss!

Please take a moment to provide us with feedback on this BoF at:
https://www.surveymonkey.com/r/3MD5LKT

E2 HPC Working Group

Sustainably Supporting Science
through committed community action

SC16 Salt Lake City, Utah | hpc matters.

Backup slides

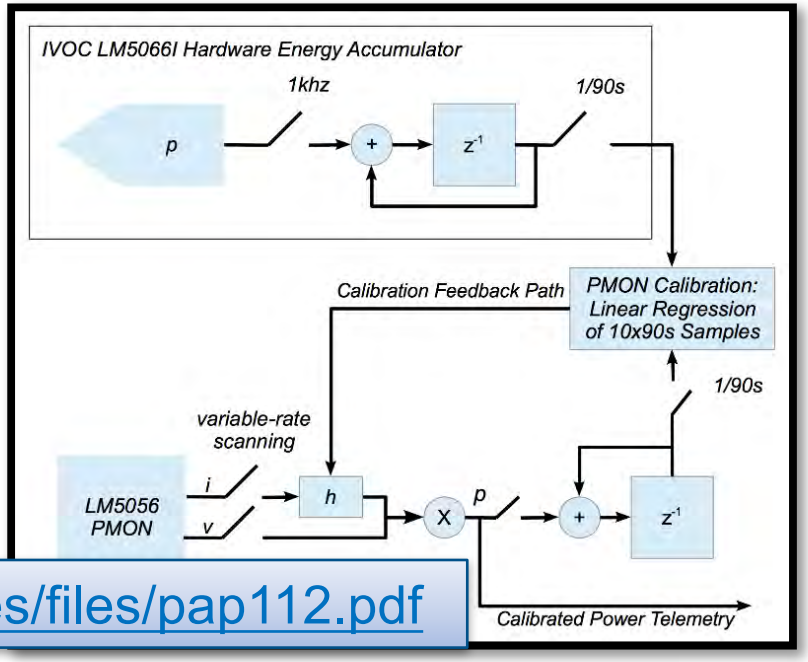
BoF
Energy Efficiency Considerations for HPC Procurements
November 15, 2016

Backup Slides

Copyright 2016 Cray Inc.

Power and Energy Monitoring Enhancements

- **Cray XC PMON Calibration**
  - Leverages factory calibrated IVOC power sensor
  - Higher confidence in data collection
  - Details are in the CUG 2016 paper!

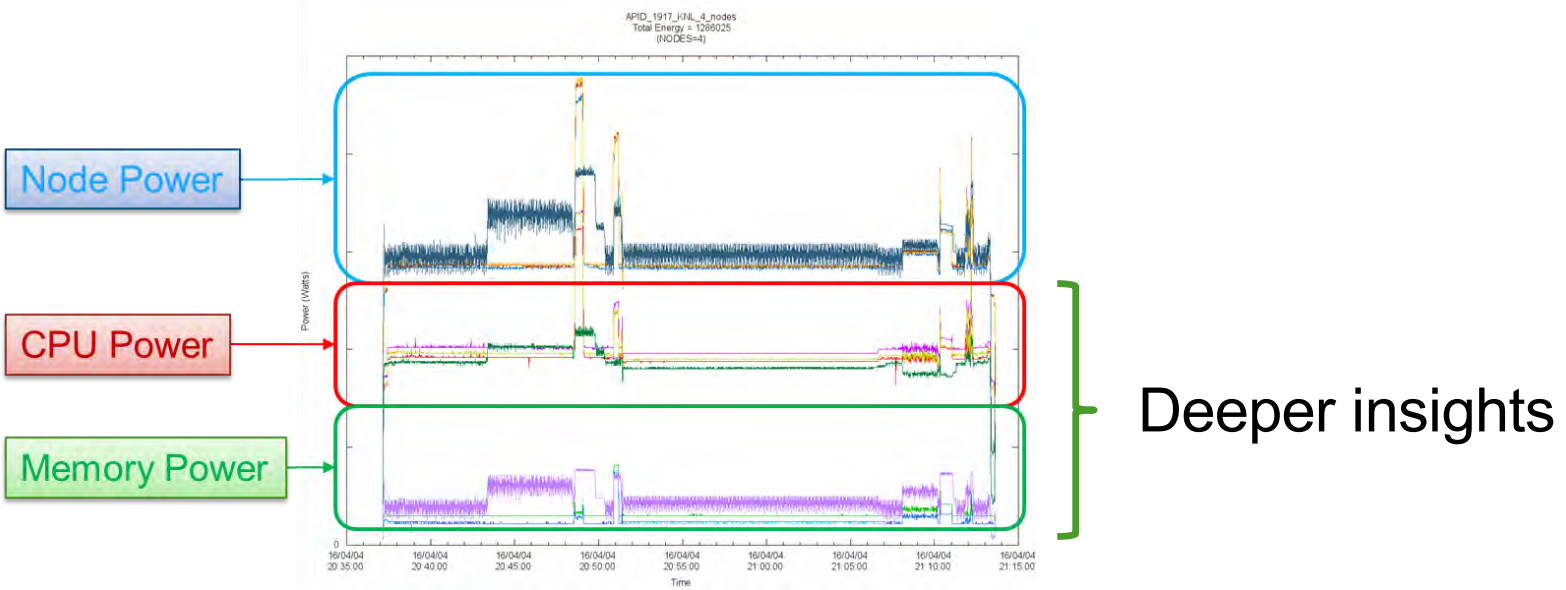


cug.org/proceedings/cug2016_proceedings/includes/files/pap112.pdf

Copyright 2016 Cray Inc.

Power and Energy Monitoring Enhancements

- **Aggregate sensors for CPU and memory telemetry**
  - Cray XC40 Blades with Intel KNL processors, + future blades
  - Enhancement driven by Trinity and EEHPC requirements…



Copyright 2016 Cray Inc.

Cray XC Monitoring and Control Quick List

- Cray Advanced Platform Monitoring and Control
  - RESTful interface for workload manager integration
- Power Management Database (PMDB)
  - System Environmental Data Collection (SEDC)
  - High-speed power/energy data collection
  - Application data (start-,end-time, nodes assigned, User, …)
- In-band access to CLE:/sys/cray/pm_counters
  - In-band access at 10 Hz to node-level data collected out-of-band
  - Resource Utilization Reporting (RUR), PAPI, & CrayPat

Copyright 2016 Cray Inc.