# Energy Efficiency Considerations for HPC Procurement

Cray Inc. Response, Steven J. Martin
stevem@cray.com

5th Annual EEHPC Working Group Workshop
11/17/2014

# Legal Disclaimer

*Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.*

*Cray Inc. may make changes to specifications and product descriptions at any time, without notice.*

*All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.*

*Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.*

*Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.*

*Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.*

*The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.*

*Other names and brands may be claimed as the property of others. Other product and service names mentioned herein are the trademarks of their respective owners.*

# Cray XC System PM Capabilities

- **Blade/Node-level power/energy data collection at 10Hz**
  - + Accelerator data captured on blades with MIC GPU hardware installed
- **Cabinet-level power/energy data collection at 1Hz**
  - Includes data for Blower Cabinets
- **Power management database (PMDB) on the SMW**
  - Cabinet-, blower-, blade-, and node-level power/energy data at 1Hz
  - App data: Job-Id, User-ID, start-time, stop-time, ..., and NID-list
  - PostgreSQL database
- **Node-level power/energy and related data into /sysfs at 10Hz**
  - /sys/cray/pm_counters:
  - RUR energy-plugin, CrayPat, PAPI, and 3rd party tool access

# Cray XC System PM Capabilities (cont)

- **System Environmental Data Collection (SEDC)**
  - Voltage, current, temperature, fan-speed, …
  - Data saved to Database*, or flat-files on Cray SMW
- **Cray Advanced Platform Monitoring and Control (CAPMC) ***
  - Platform monitoring and control API for 3rd party WLM integration
  - Monitoring and control from select service nodes
    - Node power: on | off, system-level and node-level monitoring
    - WLM (Workload Manager) directed system-,node-,job-level power capping
- **Turbo-boost limiting ***
  - Boot time ability to enforce max turbo boost
  - Save energy at large scale due to variation in achieved max turbo boost…

** New Feature released in Oct 2014*

# 2014 Feedback, General impressions

- **A lot of great material**
  - If you have not read it, you should!
- **Supportive use cases for each level of monitoring requested would be helpful**
  - Motivate customers to include details in their procurement documents
  - Justify vendor investment

# 2014 Feedback, Section 2

- Nice write-up with respect to reported vs. internal sampling
- Perhaps guidance with respect to continuous vs. on-demand data collection would be useful
- At the mandatory (lowest recommended) rate, continuous data collection on a large system would create a tremendous amount of telemetry data

| Cabinet Count | Nodes/Cabinet | Components / Node | Bytes/sample | |
|---|---|---|---|---|
| 100 | 192 | 4 | 24 | |
| | ***(Note: Actual Database size / growth rate likely much higher…)*** | | | |
| | Bytes/Second | Bytes/Day | Bytes/Week | Bytes/Year |
| Cabinet Energy | 2,400 | 207,360,000 | 1,451,520,000 | 75,479,040,000 |
| Node Energy | 460,800 | 39,813,120,000 | 278,691,840,000 | 14,491,975,680,000 |
| Component Energy | 1,843,200 | 159,252,480,000 | 1,114,767,360,000 | 57,967,902,720,000 |
| Totals | 2,306,400 | 199,272,960,000 | 1,394,910,720,000 | 72,535,357,440,000 |

- Likely more then four components / nodes.
- Way more if you monitor each power rail!

# 2014 Feedback, Section 4

- **New this year**
- **Not clear that external reporting at 1Hz is well justified**
  - Detailed use case for 1Hz data is needed
  - Perhaps tables with mandatory, important, and enhancing external reporting frequencies like used in section 2
- **Mandatory external reporting at 1 sample per minute is a more justifiable target**

# Backup

- 2013 Presentations from AMD, Cray, IBM, and Intel
  - http://eehpcwg.lbl.gov/sub-groups/equipment-1/procurement-considerations/procurement-considerations-presentations

# Additional Resources

## “Monitoring and managing power consumption on the Cray XC30 system”

- Cray S-0043-72
- http://docs.cray.com/books/S-0043-7202/S-0043-7202.pdf

## “Managing system software for the Cray Linux Environment”

- Cray S-2393-52xx
- http://docs.cray.com/books/S-2393-5202axc/S-2393-5202axc.pdf