Cray XC30 Power Monitoring and Management

CUG 2014
Steven J. Martin
stevem@cray.com

COMPUTE | STORE | ANALYZE

Safe Harbor Statement

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Related Presentations at CUG 2014

- **First Experiences With Validating and Using the Cray Power Management Database Tool**
  - Gilles Fourestey, Benjamin Cumming and Ladina Gilly (Swiss National Supercomputing Centre)
  - Next presentation!

- **User-level Power Monitoring and Application Performance on Cray XC30 supercomputers**
  - Alistair Hart and Harvey Richardson (Cray Inc.), Jens Doleschal, Thomas Ilsche and Mario Bielert (Technische Universität Dresden) and Matthew Kappel (Cray Inc.)
  - Technical Session 18C, Thursday, May 8th

Power Management: Motivation & Philosophy

- **Motivation**
  - System procurements are increasingly constrained
    - Site power & cooling limitations
    - Cost of system power and cooling
  - Customer requirements
    - Power monitoring tools
    - Management of power consumption
    - Better performance per watt
  - Power limitations
    - 20 MW max power target for extreme-scale systems of the future

- **Philosophy**
  - Do not waste energy!
  - Measure power, so you know where it is going
  - Allow customers to affect greater power savings

Power Management: Progression

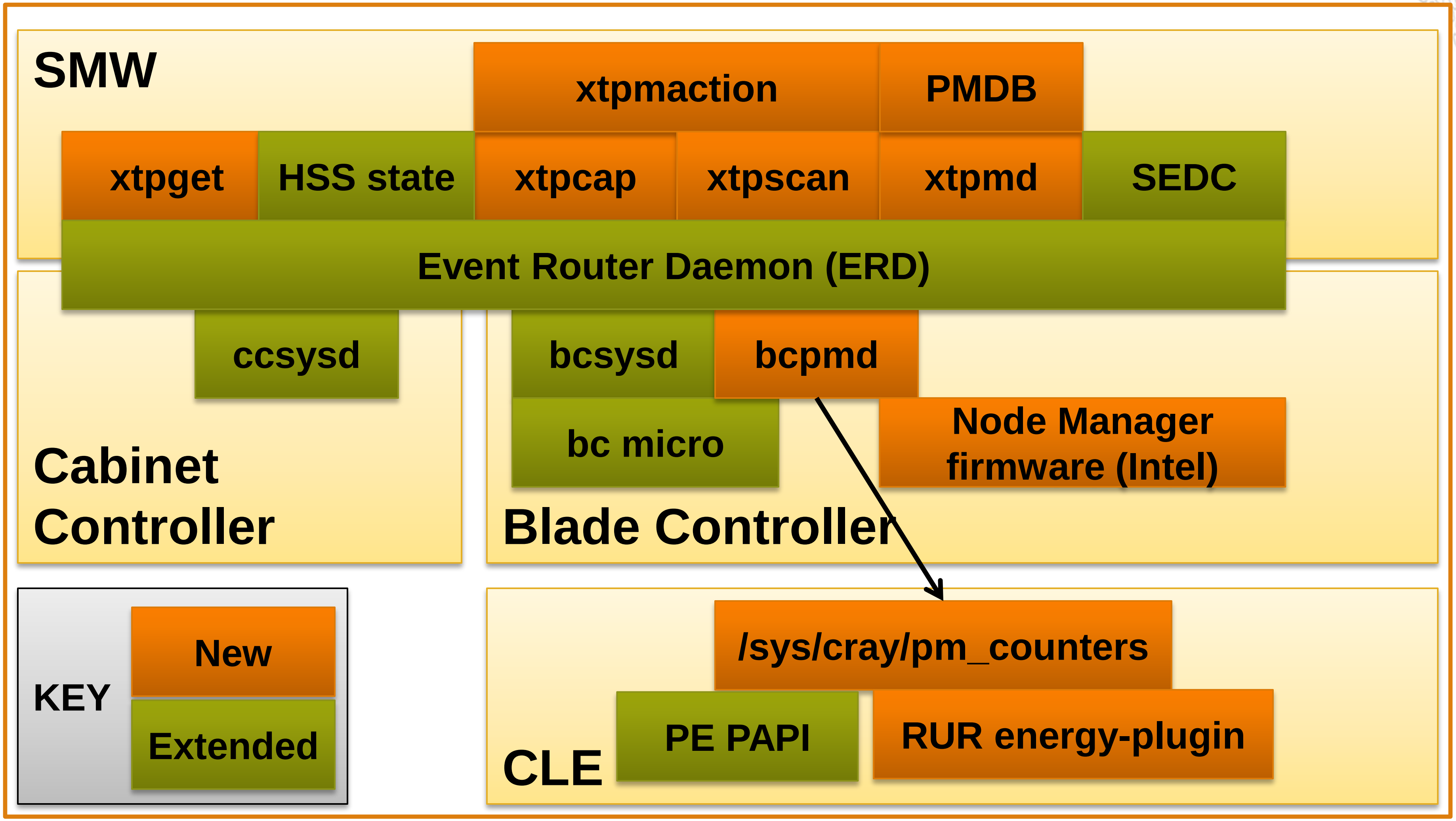
Power monitoring → System level power management → Job level power management → Application level power management

Continuously working to improve monitoring capabilities!

Capabilities Today on XC30

Software stack
Out-of-band monitoring
In-band monitoring
Management

PM Software Stack



XC30 Out-Of-Band Monitoring

- **System Environmental Data Collection (SEDC)**
  - Voltage, current, temperature, pressure, fan-speed, …
  - Readings updated once per minute
  - Data written to flat-files on SMW
- **High-speed power/energy data collection**
  - Cabinet, Blade, Node, and [Accelerator] data
  - Blade level data collection at 10 Hz
- **Power Management Database (PMDB)**
  - Cabinet-level Power (+blowers)
  - Blade-, and Node-level data at 1 Hz

XC30 In-Band Monitoring

- /sys/cray/pm_counters

```
/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
```

- Intel RAPL counters
  - PAPI
  - CrayPat

Out-Of-Band Monitoring Use Cases

- **Real-time system monitoring with *xtpget***
  - Timestamp, Current-,Average-,Peak-Power, and Accumulated Energy
  - User selectable time window for average and peak power
  - Easy command line access, no database access required
- **System-level data from PMDB**
  - System level profiling
  - Cabinet level details
  - Access days or weeks of historic data at 1 Hz
- **Application power/energy profiling from the SMW**
  - Example text report scripts ship with the SMW release
  - Node-level power & accumulated energy data at 1 Hz
  - Application data: job-id, app-id, user, start-time, end-time, and nid-list

In-Band Monitoring Use Cases

- **Cray Resource Utilization Reporting (RUR)**
  - Application energy reporting via energy-plugin
  - Multiple reporting options
    - LLM into system logs on the SMW
    - Direct to user defined locations
    - Extendable by sites and third party workload managers
- **CrayPat**
  - Intel RAPL counters
  - Cray custom counters
- **Direct access to /sys/cray/pm_counters**
  - Unrestricted read-only access

Control Use Cases

- **System power capping**
  - Capping >= max profiled workload
    - Avoid worst case power/cooling costs
    - Prevent budget overruns
  - More aggressive capping:
    - Can it be done while avoiding or mitigate negative performance impacts
    - Tradeoff: lower point-in-time power with increased time-to-solution
  - Capping to ride through a temporary power/cooling event

- **P-State at job launch**
  - Reduce average/peek power by running at lower frequency & voltage
  - May cause total-energy to solution to go up
  - Finding the optimum p-state

Features in Development and Planning Stages

- **Power monitoring and management API**
  - Monitoring and control from select service nodes
  - Node power on, off, and status
  - Flexible system-level and node-level monitoring
  - Node level power capping
- **Unified PMDB + SEDC database**
  - SEDC data into PMDB
  - Improved tools for access and configuration
- **Moving PMDB off-SMW**
  - More capacity for large systems
  - Isolate SMW from database load
  - Allow for more users to access data
  - Small systems will still have option to keep PMDB on the SMW

Conclusion

- **XC30 has PM capabilities available today**
  - Out-of-band monitoring
  - In-band monitoring
  - Power capping & p-state controls
- **Cray is actively developing new PM functionality**
  - Enabling PM for WLM and other 3rd party software
  - Support on new blades as they are developed
  - Ongoing commitment XC30 and beyond
- **We are interested in you feedback!**

Bonus Slides

Additional Resources

"Monitoring and managing power consumption on the Cray XC30 system"

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

"Managing system software for the Cray Linux Environment"

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

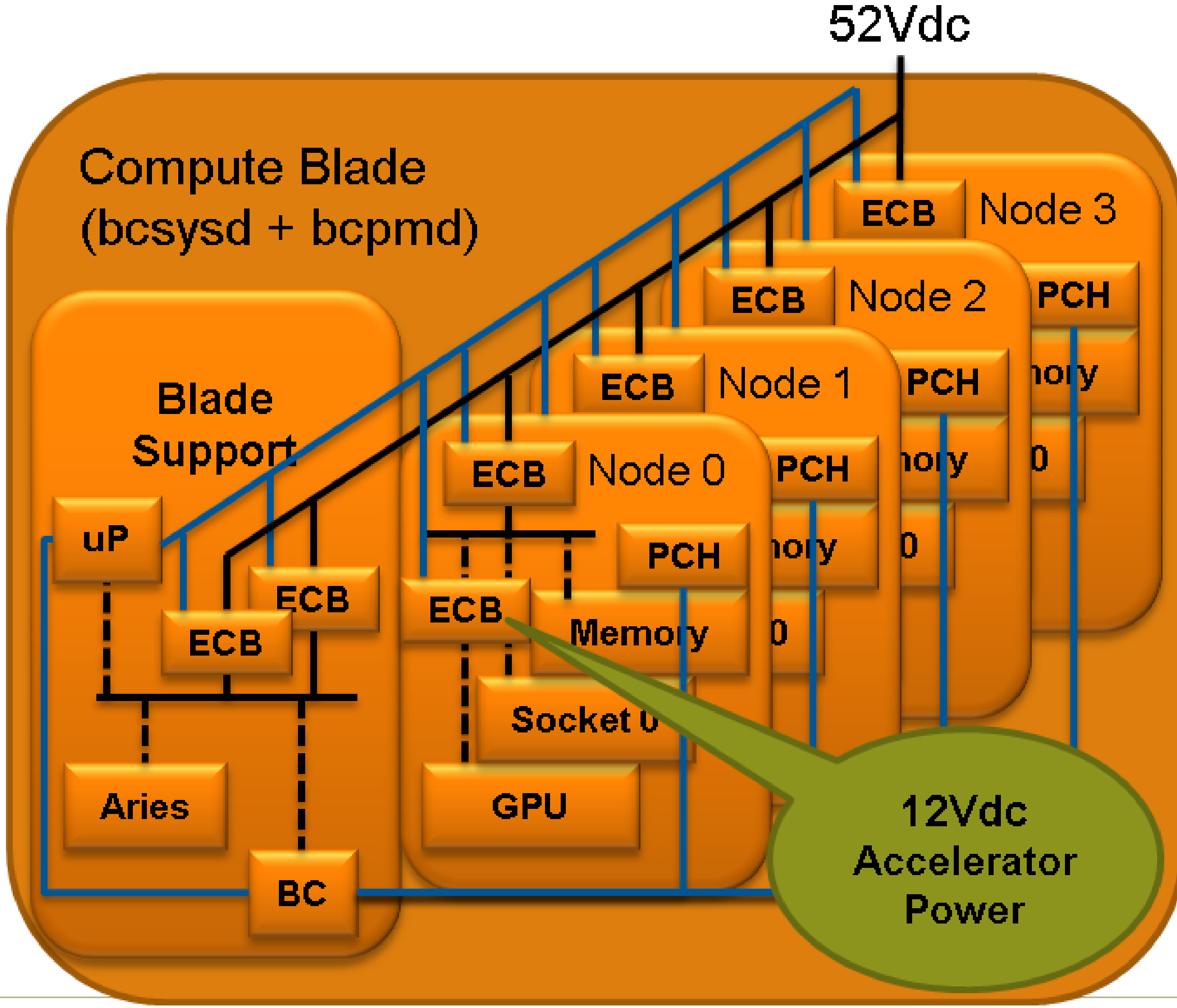
Related PM Papers at CUG 2014

G. Fourestey, B. Cumming, and L. Gilly, **"First experiences with validating and using the Cray power management database tool,"** in *Proc. Cray User Group (CUG) conference*, Lugano, Switzerland, May 2014.

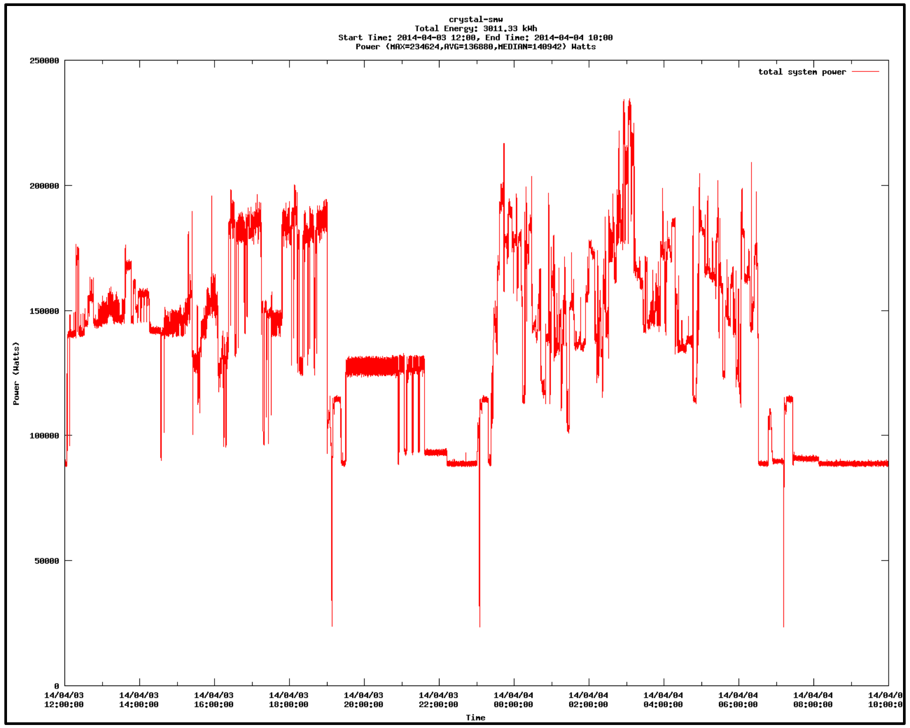
A. Hart *et al.*, **"User-level power monitoring and application performance on Cray XC30 supercomputers,"** in *Proc. Cray User Group (CUG) conference*, Lugano, Switzerland, May2014.

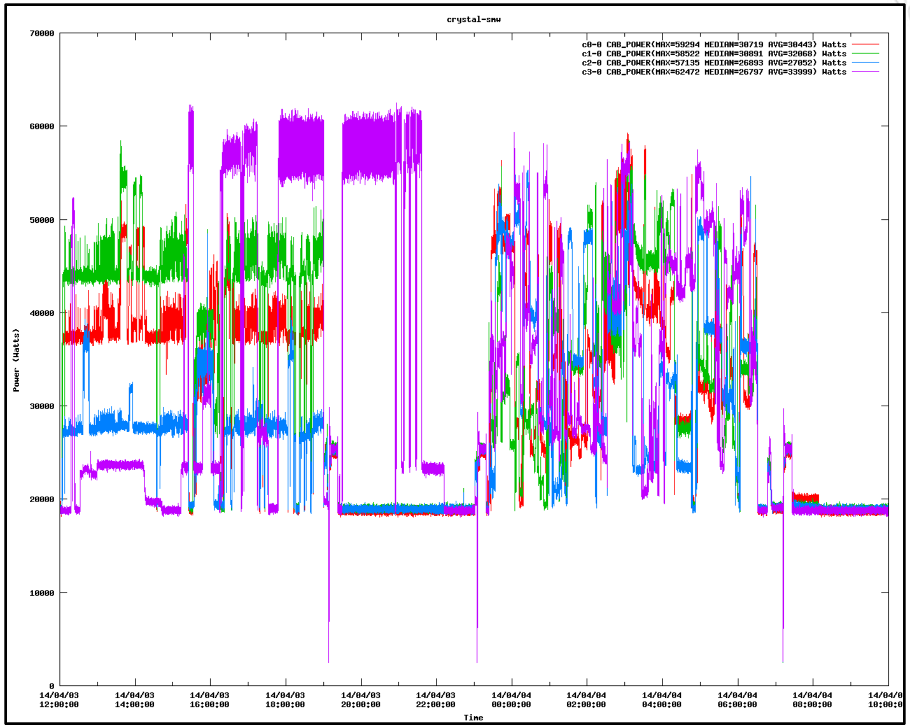
H. Poxon, **"New functionality in the Cray performance analysis and porting tools,"** in *Proc. Cray User Group (CUG) conference*, Lugano, Switzerland, May 2014.

XC30 Out-Of-Band Monitoring



System and Cabinet Level Power Plots



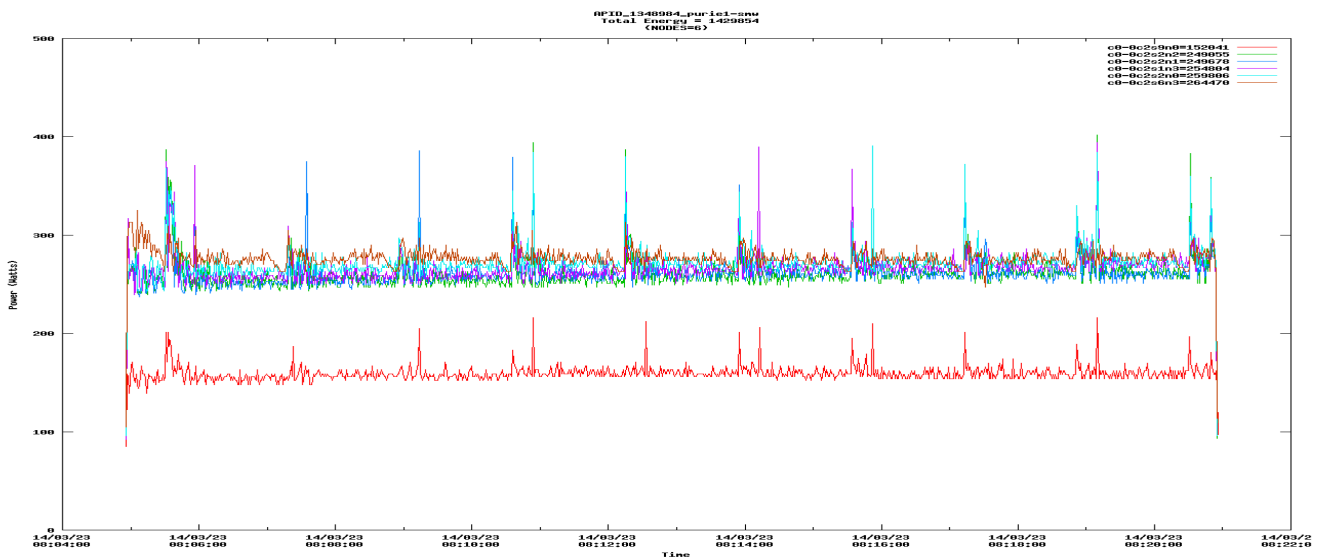


- System power (left), cabinet power (right)
- 22 hours data from PMDB's pmdb.cc data table

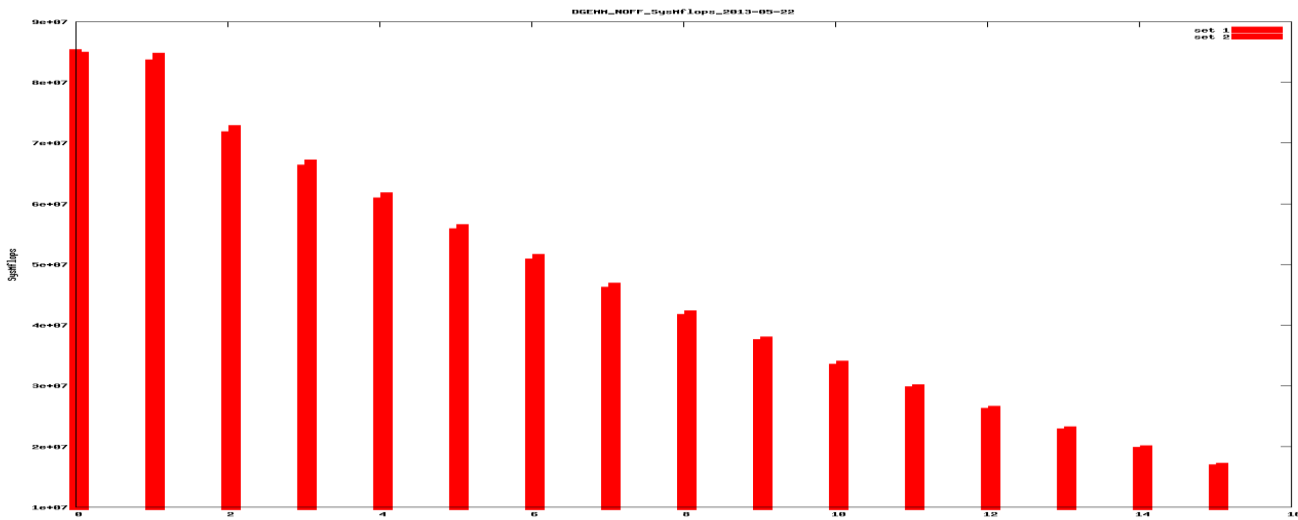
Text & Plot of Data for apid 1348984

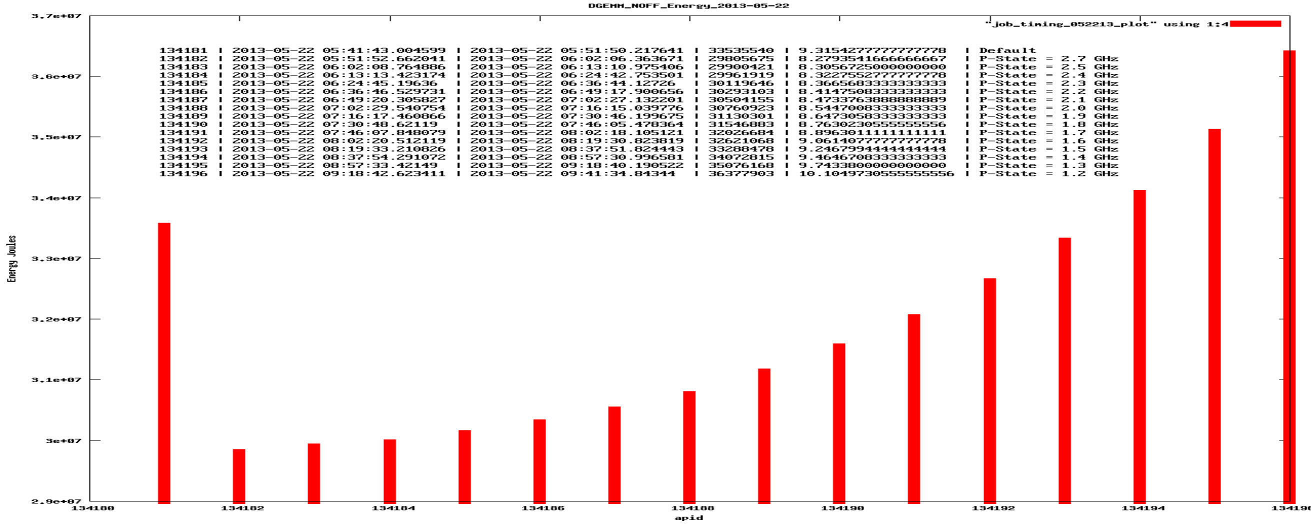
```
crayadm@puriel-smw:> cray_pmdb_report_energy_single_job.sh 1348984
  APID    | Joules  |          KW/h          |     Runtime
---------+---------+------------------------+-----------------
 1348984 | 1429854 | 0.397181666666666667 | 00:16:00.331099
(1 row)

 Component   | NID | Joules
------------+-----+--------
 c0-0c2s1n3 | 135 | 254804
 c0-0c2s2n0 | 136 | 259806
 c0-0c2s2n1 | 137 | 249678
 c0-0c2s2n2 | 138 | 249055
 c0-0c2s6n3 | 155 | 264470
 c0-0c2s9n0 | 164 | 152041
(6 rows)
```



DGEMM: Mflops (top), Energy (bottom) at P-States (Turbo, P1-P15) , 4296 PEs





COMPUTE | STORE | ANALYZE

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 associated model number marks, are 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 trademarks used in this document are the property of their respective owners.

THE SUPERCOMPUTER COMPANY