Cray XC40 Power Monitoring and Control for Intel Knights Landing Processors

Steven J. Martin (stevem@cray.com)

Executive Summary

- **Cray supports Advanced Power Management (APM)**
  - First APM features released for XC system in 2013
  - Cray works with customers, partners, and the broader HPC community to design and deliver new APM features
- **APM updates in R/R and / or blades featuring Intel KNL processors**
  - Highly parallel blade telemetry gathering architecture
  - Node-level power sampling at 1kHz in hardware
  - Factory calibrated node-level power sensors
  - Foundation for higher scan-rates for pm_counters
  - Aggregate sensors for cpu and memory
  - P-State and C-State limiting

Copyright 2016 Cray Inc.

Agenda

- **Introduction to XC power monitoring and control**
  - Short introduction to XC power monitoring and control
- **Cray XC enhanced HSS blade-level monitoring**
  - Motivation for enhanced blade-level monitoring
  - Knights Landing Processor Daughter Card (KPDC)
  - Enhances component level monitoring capabilities
- **Updated Cray XC power monitoring and control interfaces**
  - New sensor data available in PMDB, /sys/cray/pm_counters, and RUR
  - CAPMC updates

Copyright 2016 Cray Inc.

Introduction to XC power monitoring and control

- **Cray PM on XC system**
  - First released in June of 2013
    - System Management Workstation (SMW) 7.0.UP03
    - Cray Linux Environment (CLE) 5.0.UP03
  - Power Management Database (PMDB)
  - System Power Capping
  - PM Counters /sys/cray/pm_counters
  - Resource Utilization Reporting (RUR) (Sept 2013)

- **Online documentation:**
  - http://docs.cray.com/books/S-0043-7204/S-0043-7204.pdf

Copyright 2016 Cray Inc.

Introduction to XC power monitoring and control

- **Cray Advanced Platform Monitoring Control (CAPMC)**
  - Released in the fall of 2014
    - SMW 7.2.UP02 and CLE 5.2.UP02
  - Enabling Workload Managers (WLM)
    - Secure, authenticated, off-smw, monitoring and control interfaces
    - Supported by major WLM partners on XC systems

- **CAPMC online documentation:**
  - http://docs.cray.com/books/S-2553-11/S-2553-11.pdf

Copyright 2016 Cray Inc.

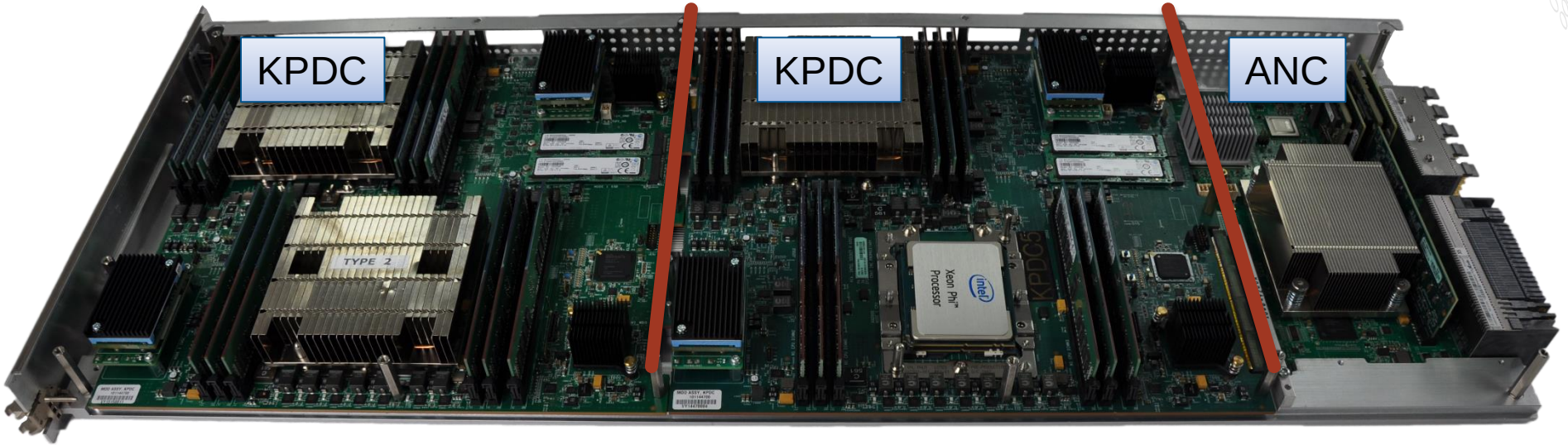
Motivation for Enhanced Blade-Level Monitoring

- **Customer and market demand**
  - Energy Efficiency Considerations for HPC Procurement Documents
    - https://eehpcwg.llnl.gov/documents/compsys/ab_procurement_2013.pdf (EE HPC WG)
    - https://eehpcwg.llnl.gov/documents/compsys/aa_procurement_2014.pdf (EE H
  - Trinity Procurement and Trinity APM NRE contracts

- **Internal use**
  - Enhanced reliability, availability, and serviceability (RAS)
  - Enhanced ability to design, manufacture, and support HPC system
  - Enhanced performance tuning and analysis opportunities

Copyright 2016 Cray Inc.

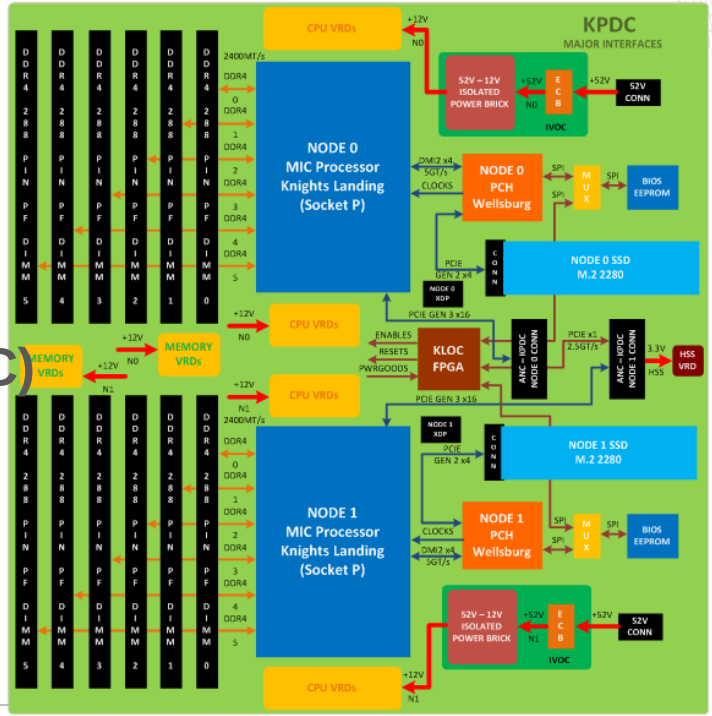
XC40 Blade with Intel KNL Processors



Copyright 2016 Cray Inc.

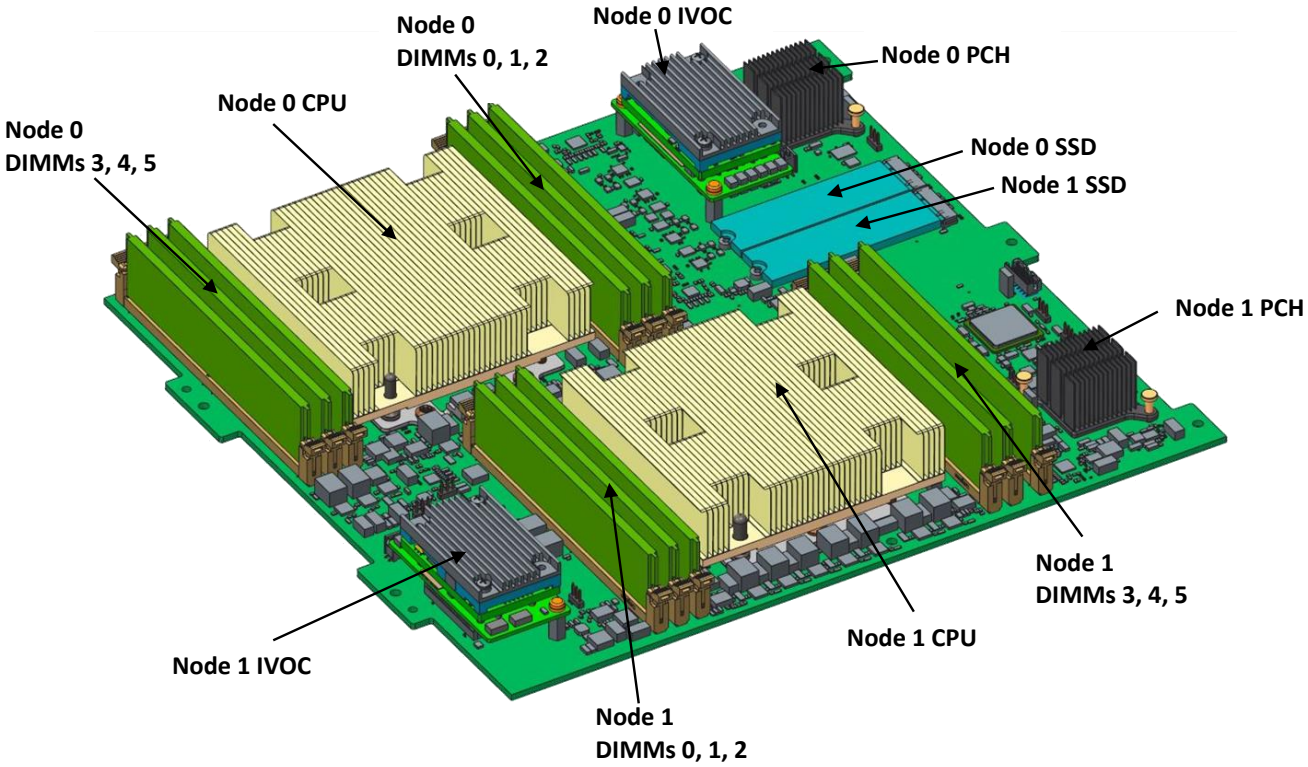
KPDC Logical View

- Two Intel Knights Landing (KNL) sockets
- Twelve DDR4 DIMMs (6 per node)
- Two Platform Controller Hub (PCH) chips
- Two intermediate voltage converters (IVOC)
  - 52V-12V conversion, socketed
- One KLOC FPGA
  - KPDC Level 0 Compute (KLOC)
- Two optional SSD cards (one per node)



Copyright 2016 Cray Inc.

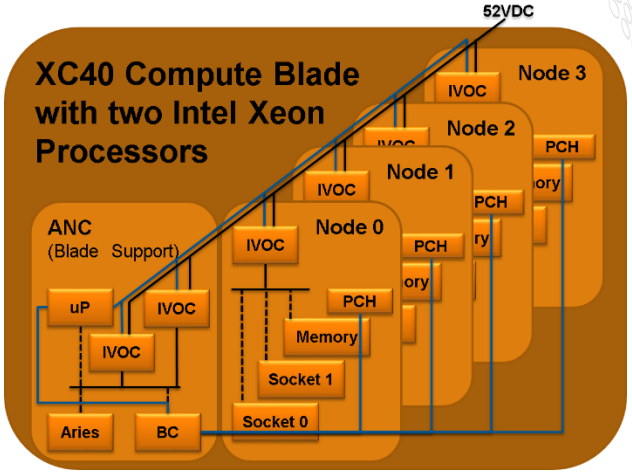
KPDC Isometric View



Copyright 2016 Cray Inc.

Previous XC30 and XC40 blades

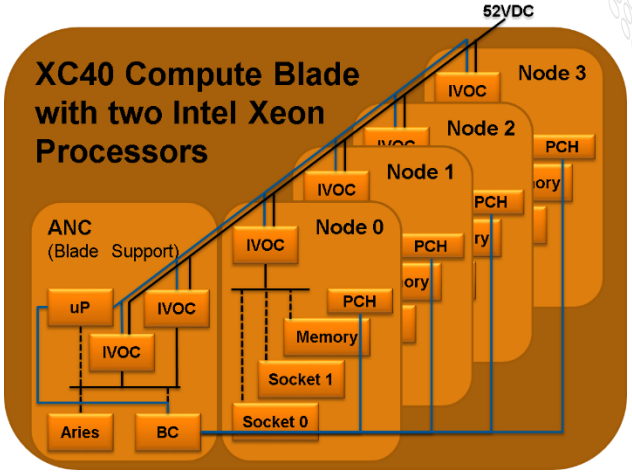
- All I2C devices connected to blade-micro
  - (uP) on the Aries Network Card (ANC)
- One I2C master at 100kHz
- Multiple I2C mux chips
- Cost effective telemetry capability
  - 10Hz max sustained polling rate



Copyright 2016 Cray Inc.

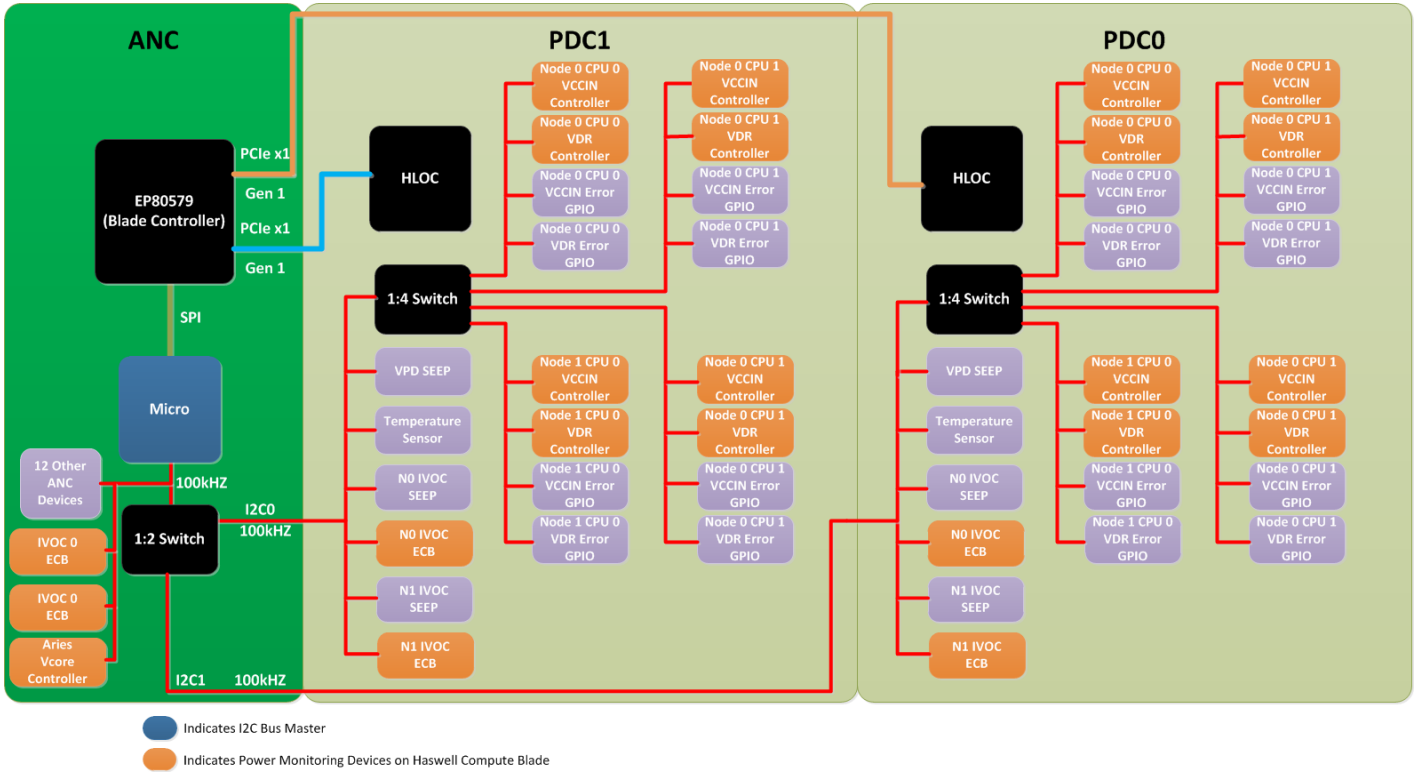
Previous XC30 and XC40 blades

- All I2C devices connected to blade-micro
  - (uP) on the Aries Network Card (ANC)
- One I2C master at 100kHz
- Multiple I2C mux chips
- Cost effective telemetry capability
  - 10Hz max sustained polling rate



Copyright 2016 Cray Inc.

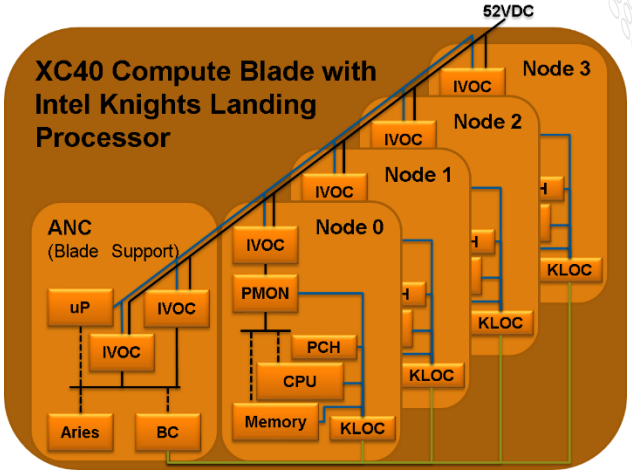
Previous XC30 and XC40 blades



Copyright 2016 Cray Inc.

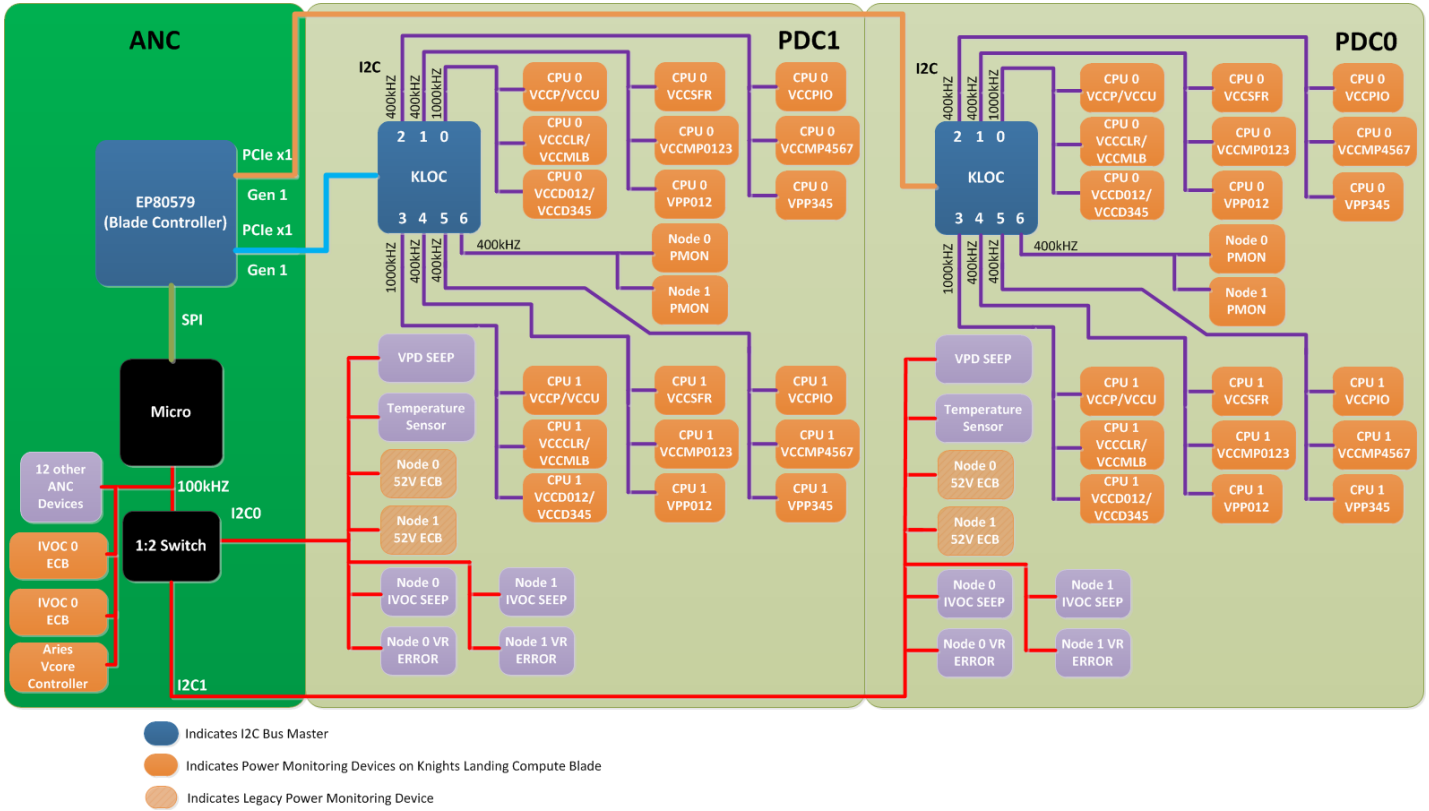
Cray XC enhanced HSS blade-level monitoring

- **I2C devices connected to KLOC**
  - Seven I2C masters / KLOC
  - Fewer devices on each bus
  - Faster I2C clocks (400 kHz or 1MHz)
  - More I2C transactions in-flight (in parallel!)
- **Node-level power sensor (PMON)**
  - 12-bit ADC
  - 1kHz sampling rate
  - Hardware averaging filter, configured to match HSS polling rate
- **Blade Controller connected to KLOC via PCIe**



Copyright 2016 Cray Inc.

Cray XC enhanced HSS blade-level monitoring



Copyright 2016 Cray Inc.

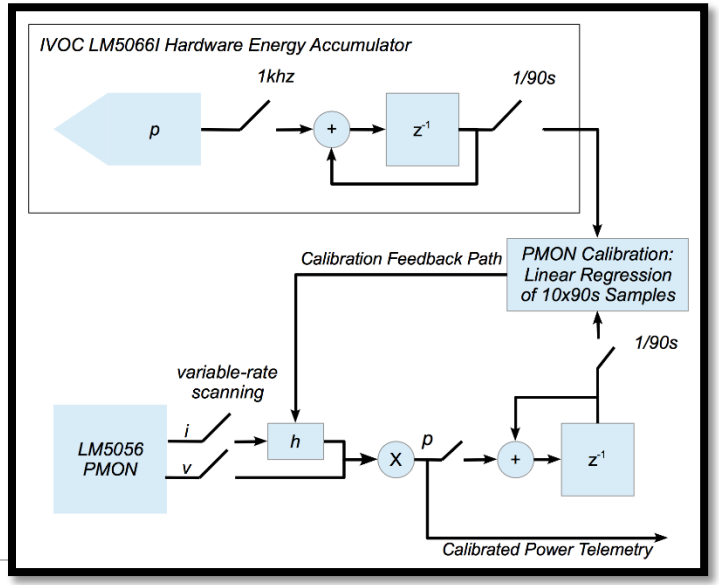
Cray XC enhanced HSS blade-level monitoring

- **PMON (Texas Instruments LM5056)**
  - High Voltage System Power Measurement Device with PMBus
  - http://www.ti.com/lit/ds/symlink/lm5056.pdf
  - Connected to KLOC
- **IVOC (Texas Instruments LM5066I)**
  - High Voltage System Power Management and Protection IC with PMBus
  - http://www.ti.com/lit/ds/symlink/lm5066i.pdf
  - Factory calibrated to better then +- 1% Accuracy

Copyright 2016 Cray Inc.

PMON Calibration

- Takes advantage of the factory calibrated IVOC power sensor
- Compares LM5066I (IVOC) with LM5056 (PMON) readings
- Correcting subsequent PMON readings
- Details are in the paper!

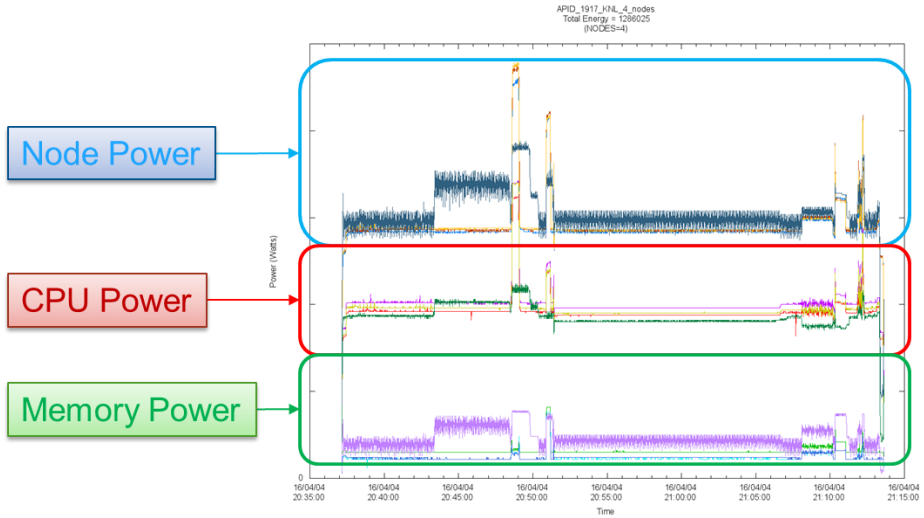


Copyright 2016 Cray Inc.

Power and Energy Monitoring Enhancements

- **Aggregate sensors for cpu and memory telemetry**
  - Abstract interface for this and planned future blades
  - New for XC40 Blades supporting Intel KNL processors

- **New PMDB telemetry**
  - 16 new sensor IDs
  - Data in pmdb.bc_data
  - 1Hz default rate



Copyright 2016 Cray Inc.

Power and Energy Monitoring Enhancements

- **Aggregate sensors for cpu and memory telemetry**
  - Abstract interface for this and planned future blades
  - New for XC40 Blades supporting Intel KNL processors
- **New PMDB telemetry**
  - 16 new sensor IDs
  - Data in pmdb.bc_data
  - 1Hz default rate

ID	Sensor Description	Unit
36	Node 0 CPU Power	W
37	Node 0 CPU Energy	J
44	Node 1 CP	
45	Node 1 CP	
52	Node 2 CP	
53	Node 2 CP	
60	Node 3 CP	
61	Node 3 CP	
68	Node 0 Memory Power	W
69	Node 0 Memory Energy	J
76	Node 1 Memory Power	W
77	Node 1 Memory Energy	J
84	Node 2 Memory Power	W
85	Node 2 Memory Energy	J
92	Node 3 Memory Power	W
93	Node 3 Memory Energy	J

Copyright 2016 Cray Inc.

Power and Energy Monitoring Enhancements

- **Aggregate sensors for cpu and memory telemetry**
  - Abstract interface for this and planned future blades
  - New for XC40 Blades supporting Intel KNL processors

- **New PM Counters (aka: descriptors):**
  - /sys/cray/pm_counters/cpu_energy
  - /sys/cray/pm_counters/cpu_power

  Aggregate CPU Power & Energy

  - /sys/cray/pm_counters/memory_energy
  - /sys/cray/pm_counters/memory_power

  Aggregate Memory Power & Energy

  - /sys/cray/pm_counters/raw_scan_hz
    - Future capability to update at ~= 100Hz?

  Counter update rate, 10Hz default

Copyright 2016 Cray Inc.

Power and Energy Monitoring Enhancements

- **Aggregate sensors for cpu and memory telemetry**
  - Abstract interface for this and planned future blades
  - New for XC40 Blades supporting Intel KNL processors

- **New Resource Utilization Reporting (RUR) telemetry:**
  - **Derived from new CPU and memory PM energy counters**
  - **`cpu_energy_used .....:`** Total CPU energy, joules
  - **`memory_energy_used ..:`** Total memory energy, joules

Copyright 2016 Cray Inc.

Power and Energy Monitoring Enhancements

- **Aggregate sensors for cpu and memory telemetry**
  - Abstract interface for this and planned future blades
  - New for XC40 Blades supporting Intel KNL processors

- **New Resource Utilization Reporting (RUR) telemetry:**
  - **Derived from new CPU and memory PM energy counters**

```
[RUR@34] uid: 12795, apid: 1917,
jobid: 0, cmdname: ./test,
plugin: energy {
  "nodes_throttled":              0, "memory_energy_used":       138156,  "min_accel_power_cap_count": 0,
  "nodes_with_changed_power_cap":0, "max_power_cap_count":      0,       "energy_used":               1285795,
  "max_power_cap":                0, "nodes_memory_throttled":   0,       "accel_energy_used":         0,
  "max_accel_power_cap_count":    0, "nodes_accel_power_capped": 0,       "min_power_cap":             0,
  "max_accel_power_cap":          0, "min_power_cap_count":      0,       "min_accel_power_cap":       0,
  "nodes_power_capped":           0, "nodes":                    4,       "cpu_energy_used":           846865,
  "nodes_cpu_throttled":          0
}
```


Copyright 2016 Cray Inc.

CAPMC Enhancements

Enabling Workload Managers

C-State Limiting

P-State Limiting

CAPMC Enhancements: C-State Limiting

- **`capmc get_sleep_state_limit_capabilities`**
  - Returns all valid C-States for target node(s)
- **`capmc get_sleep_state_limit`**
  - Returns the current C-State limits for target node(s)
- **`capmc set_sleep_state_limit`**
  - Sets the C-State limit for the target node(s)

Use case(s):
- Setting a floor on wakeup latency
- Setting a floor on idle node power consumption

Copyright 2016 Cray Inc.

CAPMC Enhancements: P-State Limiting

- `capmc get_freq_capabilities`
  - Returns all valid P-States for target node(s)
- `capmc get_freq_limit`
  - Returns the current P-State limits for target node(s)
- `capmc set_freq_limit`
  - Sets the P-State limits for the target node(s)

Use case:
- Dynamic control of application frequency from the WLM

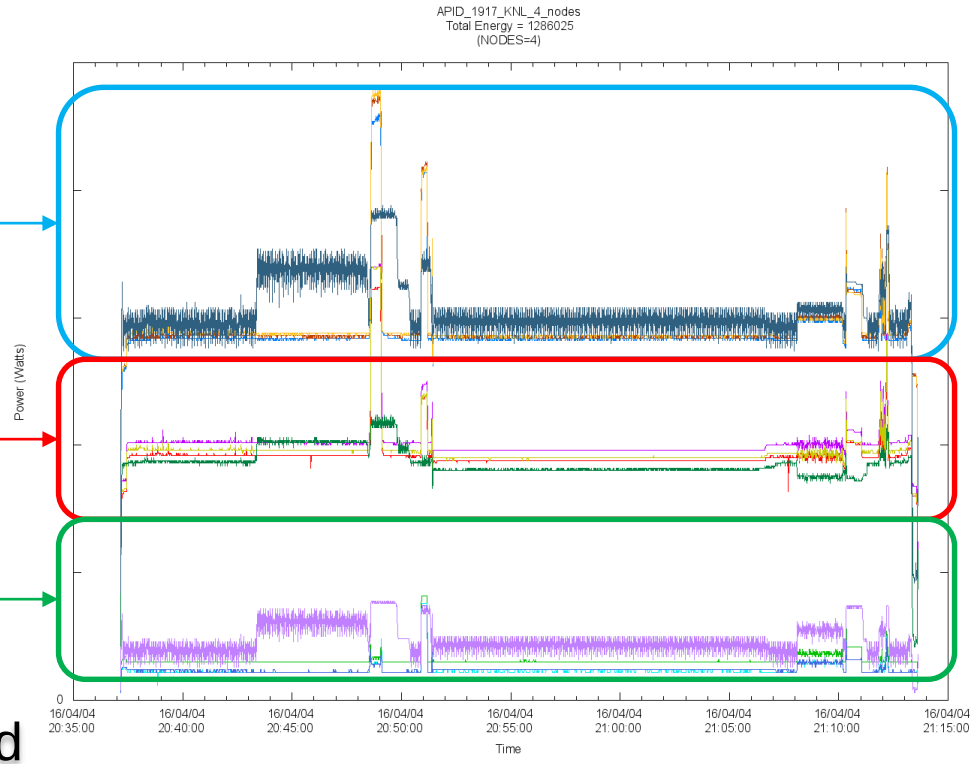
Copyright 2016 Cray Inc.

APID 1917 Test Power Profile (4-Node Test)

Node Power

CPU Power

Memory Power

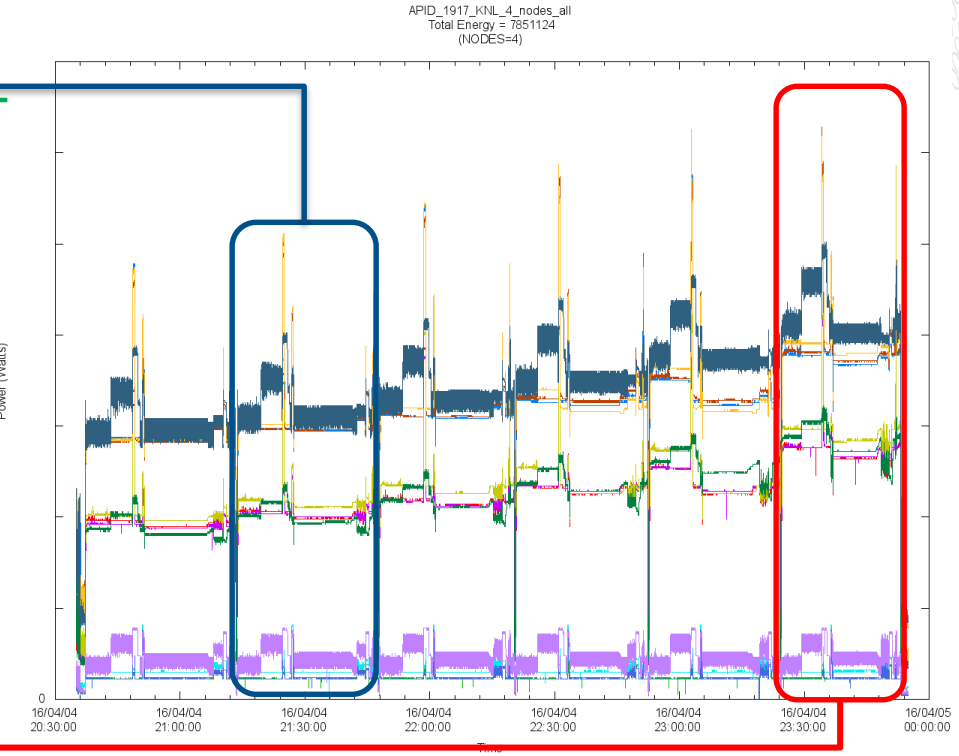


Power Level details removed

Copyright 2016 Cray Inc.

Summary Plot Six 4-Node Runs

APID	Joules	Runtime
1917	1286025	00:36:28.99
1918	1257640	00:34:07.17
1919	1268345	00:32:42.08
1920	1298037	00:31:58.58
1921	1333215	00:31:43.99
1922	1353328	00:28:52.91



Power Level details removed

Copyright 2016 Cray Inc.

Cray XC Blades Featuring Intel KNL Processors

- **Highly parallel blade telemetry gathering architecture**
- **Node-level power sampling at 1kHz in hardware**
- **Factory calibrated node-level power sensors**
- **Foundation for higher scan-rates for pm_counters**
- **Aggregate sensors for cpu and memory available via:**
  - PMDB, /sys/cray/pm_counters, and the RUR energy plugin
- **P-State and C-State limiting**
  - CAPMC controls for workload managers

Copyright 2016 Cray Inc.

Cray XC Monitoring and Control - Wrap-up

- **Cray has supported PM on XC systems since 2013**
  - 4 generations of blades featuring Intel Xeon processors
  - 2 generations of blades featuring Intel Xeon Phi processors
  - Blades featuring NVIDIA GPUs
- **Cray continues to deliver new PM features**
  - Enhanced monitoring capabilities
  - Enhanced control capabilities
  - Close working relationships with customers, partners, and the broader HPC community – Driving new features and innovations!

Copyright 2016 Cray Inc.

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.

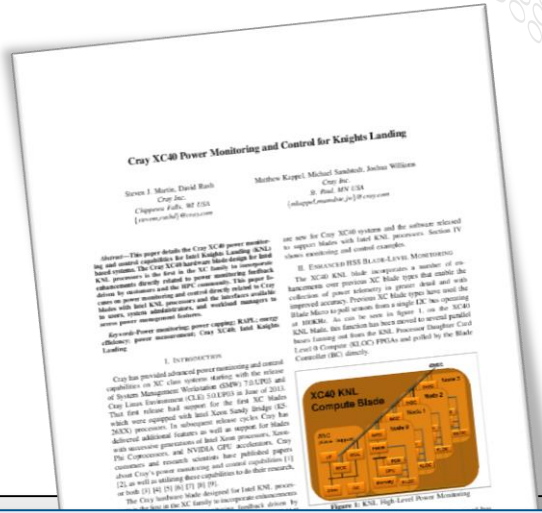
Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, REVEAL,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.

Copyright 2016 Cray Inc.

Q&A

Steven Martin
stevem@cray.com



Steven J. Martin, David Rush
Cray Inc. Chippewa Falls, WI USA
{stevem,rushd}@cray.com

Matthew Kappel, Michael Sandstedt, Joshua Williams
Cray Inc. St. Paul, MN USA
{mkappel,msandste,jw}@cray.com

Additional Resources

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

- **Cray S-0043**
- http://docs.cray.com/books/S-0043-7204/S-0043-7204.pdf

“CLE XC™ System Administration Guide”

- **Cray S-2393**
- http://docs.cray.com/books/S-2393-5204xc/S-2393-5204xc.pdf

“CAPMC API Documentation”

- **Cray S-2553**
- http://docs.cray.com/books/S-2553-10/S-2553-10.pdf

Copyright 2016 Cray Inc.