Cray HPCS Response 10/17/2013

Cray Response to EEHPC Vendor Forum
Slides presented on 12 September, 2013

Steven J. Martin
Cray Inc.

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.

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.

Cray and Sonexion are registered trademarks of Cray Inc. in the United States and other countries, and Cray XC30, Cray CS300, Cray XK7, Cray XE6, Cray Linux Environment, Cray XE6m, Cray XE6m-200, Cray XT6, Cray XT5, Cray XT4, Cray SHMEM, CrayPat, NodeKARE, YarcData and uRiKA are registered trademarks of Cray Inc.

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.

Overview

- **Quick overview of Cray XC30 power monitoring**
  - Features available today in the field
- **Cray XC30 w/respect to EEHPC requirements + comments**
  - Some details about our solution
  - Review slides from the forum on Sept 12
  - Cray comments on the requirements
- **Cray XC30 Power and Cooling**

First Some Acronyms

- **Cray specific acronyms**
  - HSS: Hardware Supervisory System
  - SMW: System Management Workstation
  - CC: Cabinet Controller
    - CC-micro: Cabinet-microcontroller
  - BC: Blade Controller
    - BC-micro: Blade-microcontroller
  - SEDC: System Environmental Data Collection
  - PMDB: Power Management Database
  - ALPS: Application Level Placement Scheduler
  - BASIL: Batch Application Scheduler Interface Layer

- **Other acronyms**
  - PAPI: Performance Application Programming Interface
  - RAPL: Running Average Power Limiting (Intel)

Some Definitions

- **Power capping**
  - Configurable upper limit on power consumption
- **In-band**
  - Communications and/or processing done directly by the application processors and/or done using the high speed network
- **Out-of-band**
  - Communication and control using the Cray HSS network & processors
  - Not disruptive of in-band communications and processing
  - Not a source of OS jitter
- **Sysfs**
  - Sysfs is a virtual file system provided by Linux. Sysfs exports information about devices and drivers from the kernel device model to user space, and is also used for configuration. [http://en.wikipedia.org/wiki/Sysfs]

Cray XC30 Monitoring (available now)

- **Power management database (PMDB)**
  - Cabinet level voltage, current, and power data at 1Hz
  - Blade, and Node level power and accumulated energy data at 1Hz
  - Job/App ID, start time, end time, and node list data
  - Out-of-band job power profiling using data in power database
  - Implemented using PostgreSQL on the SMW
- **In-band node power/energy/cap performance counters**
  - Access node power/energy/cap performance counters via sysfs
  - Enables applications, users, and tools node-level access to power, energy, and capping data
- **In-band energy reporting**
  - Production support via Cray RUR (Resource Utilization Reporting)
  - Accumulated Energy data collected from each node at app start/end
  - Data aggregated to central location

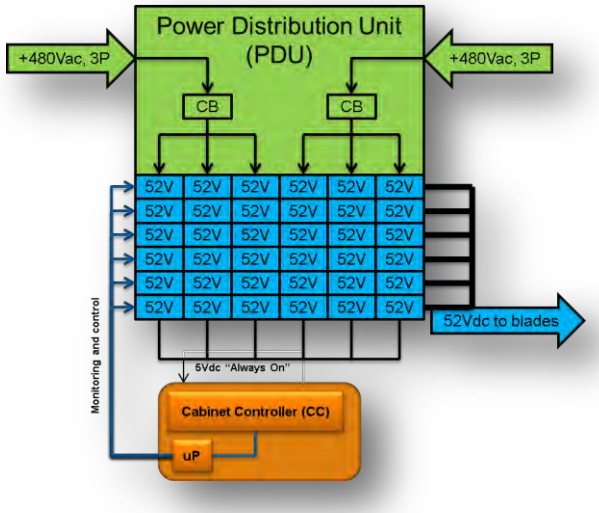
XC30 Power Management (available now)

- **Static system power capping**
  - Power capping configuration management
  - Same power cap setting for all nodes of a given type
    - Service nodes configured with no power cap by default
  - Ability to change power cap without rebooting the system
    - Not intended for rapid changes of the system power cap
- **P-state control at Job/Application Launch**
  - Ability to select P-state or Linux performance governor
    - In production P-state is useful
    - Alternative performance governor may be useful for research
  - Job level control via BASIL interface extensions
    - Enables workload manager integration
  - Application level control via aprun options

System, Platform and Cabinet

Monitoring

Cray XC30LC Cabinet Power Monitoring



- **480 Vac 3-phase input power**
  - Two 3-phase mains into each cabinet
- **36 3000-watt rectifiers**
  - N+1 redundant
  - hot-swap enabled
  - > 95% efficient (load dependent)
- **52 Vdc output to blades**
- **Rectifier monitoring via i2c by the CC-micro**
- **Adjacent blower cabinet power collected by CC-micro**
- **CC (processor) accesses data collected by CC-micro**
- **Reports aggregated cabinet data to SMW at 1Hz**
  - Compute-cabinet: voltage, current, power
  - Blower-cabinet: power
- **Data cached on SMW, and stored into PMDB**
  - PMDB (Power Management Database) PostgreSQL

Slide 6 from EEHPC Presentation

Measurements: System, Platform and Cabinet

(mandatory) Shall be able to measure the current and voltage of the system, platform(s) and cabinet(s).

The current and voltage measurements shall provide a readout capability of
(mandatory) ≥ 1 per second
(important) ≥ 50 per second
(enhancing) ≥ 250 per second

(mandatory) The current and voltage data shall be real electrical measurements, not based on heuristic models

(important) The vendor shall assist in the effort to collect these data in whatever other subsystems are provided (e.g., another vendor's back-end storage system).

(important) Those elements of the system, platform and cabinet that perform infrastructure-type functions (e.g., cooling and power distribution), shall be measured separately with the ability to isolate their contribution to the energy and power measurements.

Cray Cabinet Power Monitoring Comments:

- **"The current and voltage data must be real electrical measurements, not based on heuristic models".**
  - XC30 compute cabinet monitors rectifier output voltage / current
  - AC input can be calculated using rectifier efficiency
  - XC30LC blower cabinet AC power is looked up in table by frequency
  - Adding real AC power meter would require further justification
- **XC30 Cabinet level voltage/current/power data collection**
  - Max frequency is 1Hz (our default)
  - Going above 1 Hz would require additional hardware/software
- **Consideration of optional cabinet-level AC metering**
  - Optional support not free
  - Need to justify with business case
  - This forum is great way to drive requirements

Node

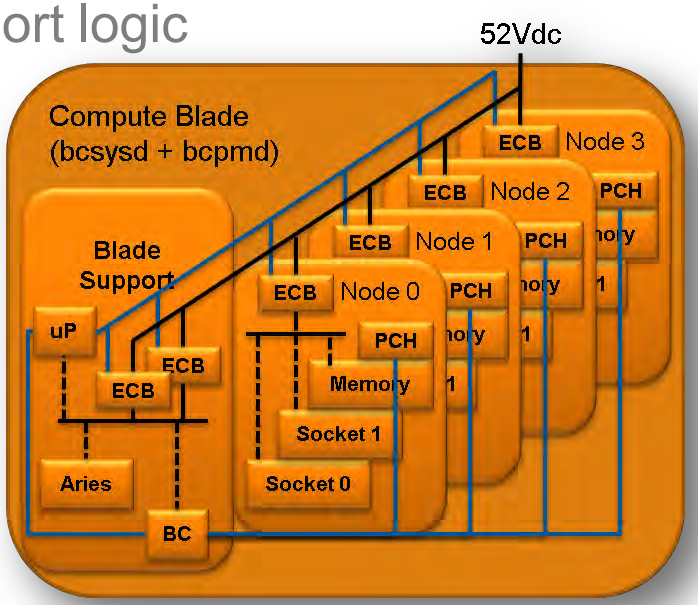
Monitoring

Cray XC30 Node Power / Energy Monitoring

- **Monitoring at 52 Vdc input to node**
  - Voltage and current monitored BC-micro
  - Data collected at target rate of 10Hz
  - Node accumulated energy calculated from 10Hz data
- **Power and accumulated energy saved into PMDB**
  - 1 Hz data saved into PMDB (default rate)
  - Data buffering at blade and SMW levels to enhance performance
- **Power and accumulated energy available in-band**
  - The actual data collection is done out-of-band at ~10Hz
  - Data copied into RAM that can be read by Linux
  - Read only data available to users/apps in /sys/cray/pm_counters
  - Resource Utilization Reporting (RUR) leverages data from /sys/cray/pm_counters to provide energy usage

Cray Marble (Xeon) Blade Power Monitoring

- **Six Electronic Circuit Breakers (ECBs)**
  - Monitor incoming 52Vdc to the blade (and nodes)
  - One ECB to each node
  - Two ECBs support Aries and blade support logic



- **Blade micro-processor**
  - Monitors ECBs via i2c
  - ECBs are polled at ~10Hz
  - ECB data published in sysfs

- **Blade controller software**
  - Reports node power data into Intel chipset on demand
  - Sends power data to SMW for storages and analyses
  - Copies data so it can be accessed in-band via /sys/cray/pm_counters

Slide 7 from EEHPC Presentation

Measurements: Nodes

(Info) A node level measurement shall consist of a combined measurement of all components that make up a node in the architecture. For example, these components may include the CPU, memory and the network interface. If the node contains other components such as spinning or solid state disks they shall also be included in this combined measurement. The utility of the node level measurement is to facilitate measurement of the power or energy profile of a single application. The *node* may be part of the network or storage equipment, such as network switches, disk shelves and disk controllers.

(important) The ability to measure the current and voltage of any and all nodes shall be provided.

The current and voltage measurements shall provide a readout capability of:

(mandatory) ≥ 1 per second
(important) ≥50 per second
(enhancing) ≥ 250 per second

(mandatory) The current and voltage data must be real electrical measurements, not based on heuristic models.

Cray Node level Monitoring Comments

- **XC30 node level collection**
  - Node level voltage/current collection targeted at 10Hz
  - Accumulated energy calculated from 10Hz data
  - 10Hz data available to in-band (Linux) via /sys/cray/pm_counters
  - Data sent to SMW database (PMDB) at 1Hz (default) rate
- **XC30 blade/node level collection at rates > 10Hz**
  - 10Hz is upper limit on current XC30 blades
  - Evaluating paths to increase hardware polling rates on future blades
  - Infrastructure to manage higher data rates & volumes at scale
- **Tradeoff between adding capabilities and COGS**
  - What new use cases are enabled rates > 10Hz?
  - Are sales won / lost / influenced by this capability

Components

Monitoring

Cray XC30 Component Level Monitoring

- **Cray out-of-band component level monitoring**
  - Critical range checking & error handling in HW & BC-micro main loop
    - Independent of higher level monitoring
  - Wide range of data collected via SEDC framework at 1 per minute
    - VRM level voltage & currents, thermal sensors
    - Selected chipset data via PECI interface
  - SEDC data is stored on the SMW in flat files (csv format)

- **In-band access to component level power/energy data**
  - Cray is enabling access via standard interfaces like PAPI
  - Working to insure access is as efficient as possible
  - Working with chipset vendors and customers to meet future needs

- **Future breakout of: CPU, Memory, PCIe, Other**
  - Pushing for fine grained capabilities
  - Some vendor dependencies

Slide 8 from EEHPC Presentation

Measurements: Components

(Info) Components are the physically discrete units that comprise the node. This level of measurement is important to analyze application energy performance trade-offs. This level is analogous to performance counters and carries many of the same motivations. Components may not only be silicon devices. For example, it would be useful to know how much fan energy is being used by the Muffin fans at the back of the rack or by some active rear door cooling methodology. Also, some systems may have a CDU. How much energy is being used by the CDU for motors, fans.

(enhancing) The ability to measure the current and voltage of each individual component must be provided.

The measurement sampling frequency should be:
(mandatory) 10 samples per second
(important) 100 samples per second
(enhancing) 1000 samples per second

(mandatory) The current and voltage data shall be both real electrical measurements and based on heuristic models.

Cray Component Level Monitoring Comments

- **Highest frequency data needs to be available in-band**
  - Enabled by processor vendors
  - Need for low latency / low overhead access
  - Open standards based APIs
- **Out-of-band vs in-band**
  - Cost of data collection
    - COGS, and performance impact
  - Who needs access to the data
- **Need to be realistic**
  - Monitoring frequency should be tuned to the device
  - Fan power at 1KHz likely overkill
    - Even if Fan power is high, it rate of change is likely slow
  - How will data be used
  - Where will data be staged/stored/archived

TCO, PUE, TUE, and Power Distribution

Comments on TCO and PUE

- **TCO**
  - Utilize cooling solutions that use energy for cooling only as needed
    - Do not over cool components
  - Eliminate conditioning of the computer room
    - Make the computers operate in any room environment

- **PUE**
  - Elimination of wet cooling towers or refrigeration assisted cooling
  - Eliminate requirement for chilled water in the facility
  - Higher inlet water temperatures + larger water temperature increase
    - Decrease amount of water required for heat extraction
    - Decrease energy consumption for cooling
  - Best in class power conversion efficiency at all stages
    - Results in lower total power losses
    - Decrease waste heat generation

Comments on TUE, Power Distribution

- **TUE**
  - Reduce system energy associated with the cooling of components
    - More efficient cooling designs
  - Reduce the power losses within the system
    - Higher efficiency converters and distribution
  - Note: We can currently measure power to all of the computational components in Cascade needed to calculate TUE

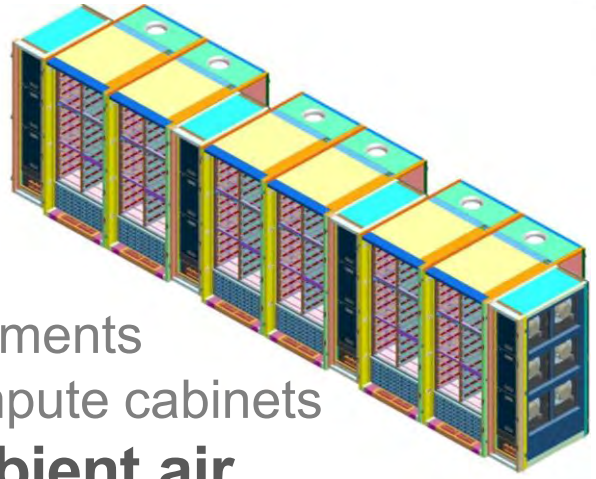
- **Power Distribution**
  - Power design architecture
    - Keep voltage as high as possible as long as possible
      - Keep I^2xR losses to a minimum.
  - Best in class power conversion efficiency at every stage

Cray XC30 Power & Cooling

Cray's Cooling TCO

- **Excellent PUE characteristics**
  - Supportive of facilities with PUE as low as 1.1
- **100% room neutral cooling**
- **Cooling infrastructure maintained across multiple generations of computing technology**
- **Only 3% of cabinet power used for cooling infrastructure**
- **Ambient air re-use minimizes facility air management cost**
- **Full or partial "Free" cooling can be achieved at most data center locations world wide**
- **Excellent system density**
  - 384 sockets/cabinet
  - ~18.5 sockets/sq ft

Cray XC30 Transverse Cooling Advantages



- **Transverse air flow**
  - Allows re-use of air across many cabinets
  - Minimizing facility air management requirements
  - Eliminate facility requirements, for the compute cabinets
- **Accepts up to 32 degree Celsius ambient air**
- **Accepts up to 25 degree Celsius inlet water**
- **Infrastructure cools ALL compute cabinet hardware**
- **Uses facility water**
  - Direct water to air cooling of cabinets
- **Optional pre-conditioner**
  - Allows the system to operate in humid environments
  - Further reduction in facility expense
  - Saves dollars & energy

Cray XC30 Transverse Cooling Advantages

- **Maintainability**
  - Blades can be warm swapped
    - No plumbing connections to disturb
  - Blowers can be hot swapped
    - N+1 cooling redundancy
  - Rectifiers can be hot swapped
    - N+1 rectifiers

XC30LC transverse cooling system

- **Cabinet cooling:**
  - Water coil (radiator) in each cabinet
    - Removing heat from air exiting each cabinet
  - Six horizontally mounted fans per blower cabinet
  - Blower cabinets placement
    - Front and end of each row of cabinet
    - Between pairs of compute cabinets
    - Each blower cabinet supports two compute cabinets
  - Fans are N+1 redundant and support hot swap
  - Optional pre/post condition units

Cray Power Advantages

- **Excellent Power Efficiency**
  - >95% efficient power rectifiers
  - >95% efficiency on 52V to 12V bus converter
  - ~89% efficiency 12V to Logic levels
- **Reliable**
  - High quality power components
  - Cray qualified and tested
  - N+1 power redundancy on rectifiers
- **Flexible**
  - 480V or 400V input power
  - 208V input power (XC30AC)

Questions?