
IBM Response to EEHPC Vendor Forum: Energy Efficiency Considerations for HPC Procurements

November 7, 2013



Agenda

- IBM System x Response
- IBM p775 Response

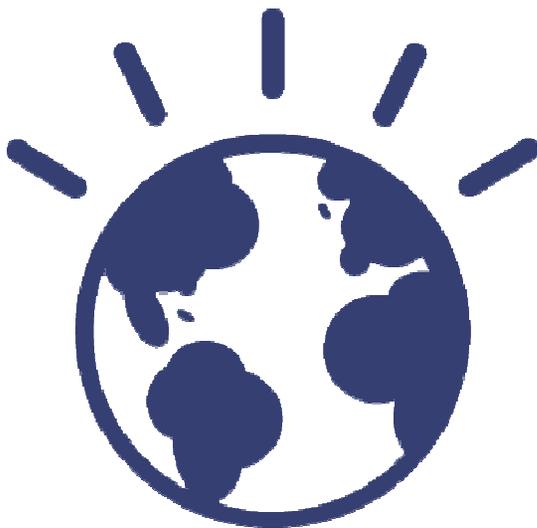
Luigi Brochard

Kevin Covi
Mike Ellsworth



IBM Response to EE HPC WG for System x November 7, 2013

Luigi Brochard, IBM STG Technical Computing
luigi.brochard@fr.ibm.com



**High Performance Computing
For a Smarter Planet**

Agenda for HPC System x

- **Software solutions**
 - xCAT
 - Platform LSF/LoadLeveler

- **Cooling solutions**
 - Direct Water Cooling

- **Answers to EE HPC WG**

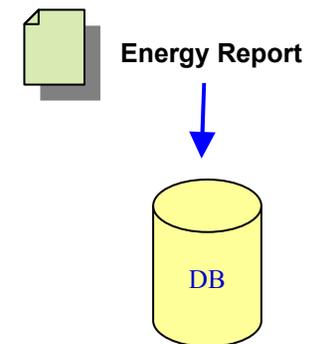
IBM software to monitor and control power

- **Report**

- temperature and power consumption per node
- temperature, power consumption and energy per per job

- **Optimize**

- Reduce power of inactive nodes
- Reduce power of active nodes



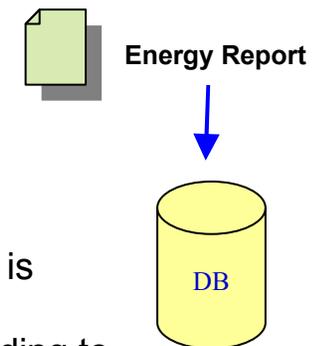
IBM software to monitor and control power

■ xCAT

- Manage power consumption on an ad hoc basis
 - For example, while cluster is being installed, or when there is high power consumption in other parts of the lab for a period of time
 - Query: Power saving mode, power consumed info, CPU usage, fan speed, environment temperature
 - Set: Power saving mode , Power capping value, Deep Sleep (S3 state)

■ Platform LSF / LoadLeveler

- Report power and energy consumption per job
 - Energy report is created and stored in the DB for all jobs submitted
- Optimize power and energy consumption per job
 - Set nodes at lowest power consumption (C6 or S3) when no workload is scheduled on this set of nodes
 - When job is run again, set nodes at optimal processor frequency according to the energy policy selected



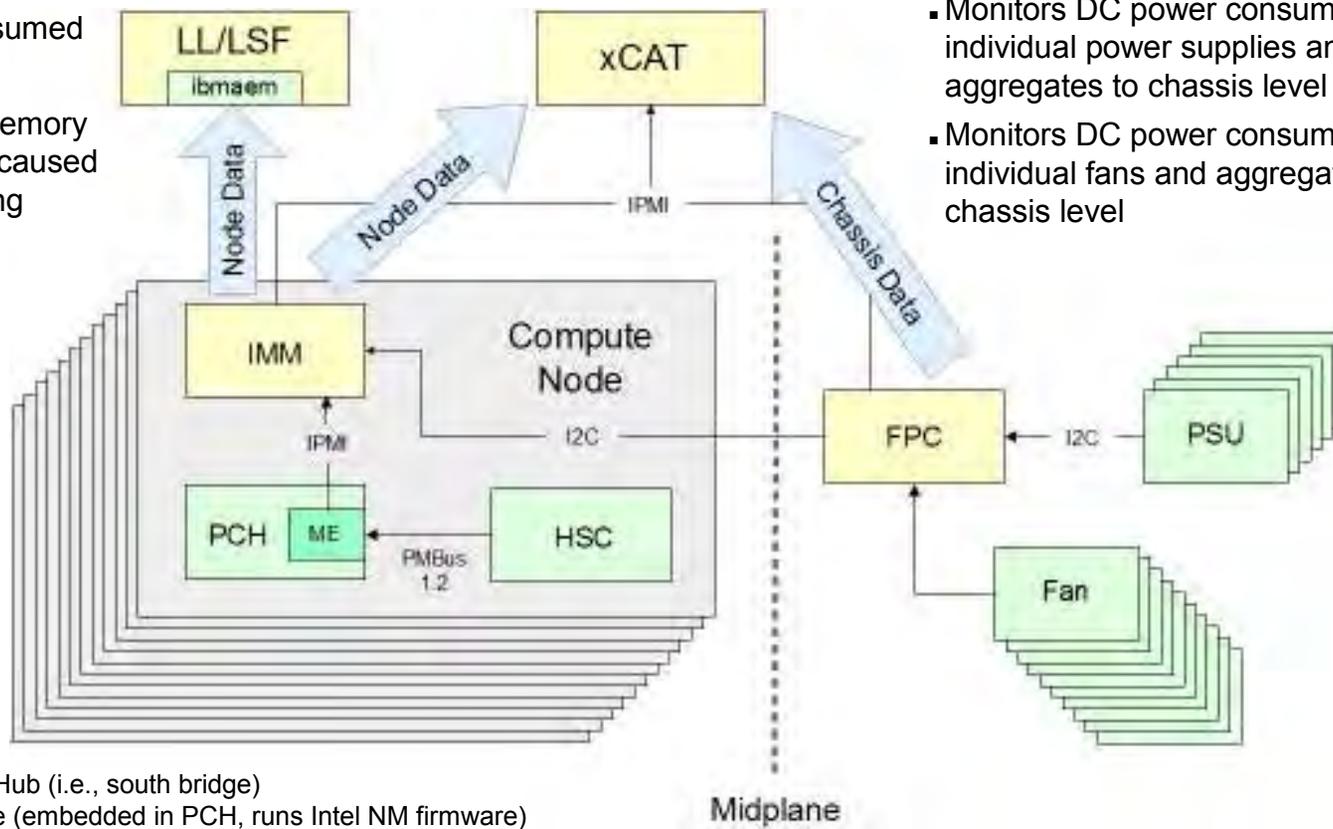
Power Management on NeXtScale

IMM = Integrated Management Module (Node-Level Systems Management)

- Monitors DC power consumed by node as a whole and by CPU and memory subsystems
- Monitors inlet air temperature for node
- Caps DC power consumed by node as a whole
- Monitors CPU and memory subsystem throttling caused by node-level throttling
- Enables or disables power savings for node

FPC = Fan/Power Controller (Chassis-Level Systems Mgmt)

- Monitors AC power consumed by individual power supplies and aggregates to chassis level
- Monitors DC power consumed by individual power supplies and aggregates to chassis level
- Monitors DC power consumed by individual fans and aggregates to chassis level



PCH = Platform Controller Hub (i.e., south bridge)
 ME = Management Engine (embedded in PCH, runs Intel NM firmware)
 HSC = Hot Swap Controller (provides power readings)

xCAT Power/Energy Management

- **Query or set the power capping status – whether or not power capping is currently being enforced (Watt)**
- **Query or set the power capping value – the permitted max wattage per motherboard (Watt)**
- **Query the minimum power cap the system can guarantee (Watt)**
- **Query the cumulative kWh used per motherboard – AC & DC (kWh)**
- **Query from the PSU the recent average AC wattage used (Watt)**
- **Query recent histogram data of wattage usage – how many 1 second intervals it has operated in each wattage range (Watt)**
- **Query current ambient temperature and CPU exhaust temperature (Centigrade)**
- **S3 Suspend and Resume**

How Platform LSF and LL manages idle nodes

- **When a job has completed on a set of nodes, scheduler set those nodes in a state which does let the OS to turn them into C6 state**
- **After nodes have been idle for some time and according to some criteria , scheduler tells xCAT to suspend them into S3 state**
- **When new job is submitted which require nodes to be resumed , scheduler tells xCAT to resume the desired nodes from S3 before it submits the job**

EAS policies available to manage active nodes

▪ Predefined policies

- Minimize Energy within max performance degradation bound of X%
- LL will determine the frequency (lower than default) to match the X% performance degradation while energy savings is still positive
- Minimize Time to Solution
- LL will determine a frequency (higher than default) to match a table of expected performance improvement provided by sysadmin
- This policy is only available when default frequency < nominal frequency
 - Set Frequency
- User provides the frequency he wants host jobs to run
- This policy is available for authorized user only

- Policy thresholds are dynamic, i.e values can be changed any time and will be taken into account when next job is submitted

▪ Site defined policy

- Sysadmin provides an executable which set the frequency based on the information stored in the DB

iDataplex Water Cooling design

- Hot Water Cooled Node with 90% heat recovery.
- Will be capable of cooling all SB-EP Skus { SB-EP 130W 8-core 2.7 GHz & 135W 8-core 2.9GHz}
- Power advantage over air cooled node by 5-7%. {Due to lower CPU temps and absence of fans}
- Water inlet 18°C to 45°C @ 0.5 liters/min per Node {37 liters/min Rack}
- Data Centers can be operated with very low chilled water requirements to manage residual load.
- Thermal design capable of 2DPC memory cooling.
- Typical operating conditions: Ambient air inlet @ 27-35°C and water inlet 18°C to 45°C to the nodes.



iDataplex Rack w/ water cooled nodes

Hybrid Cooling

- Highly energy-efficient **hybrid-cooling** solution:
 - Compute racks with Direct Water Cooling
 - 90% Heat flux to warm water
 - 10% Heat flux to CRAH
 - Switch / Storage racks
 - Rear door heat exchangers
- Compute node **power consumption reduced ~ 10%** due to lower component temperatures and no fans.
- Power Usage Effectiveness $P_{\text{Total}} / P_{\text{IT}}$: **PUE ~ 1.1**
- **Heat recovery** is enabled by the compute node design:
Energy Reuse Effectiveness $(P_{\text{Total}} - P_{\text{Reuse}}) / P_{\text{IT}}$: **ERE ~ 0.3**

3 PFlops SuperMUC system at LRZ

■ **Fastest Computer in Europe on Top 500 June 2012**

- 9324 Nodes with 2 Intel Sandy Bridge EP CPUs
- 3 PetaFLOP/s Peak Performance
- Infiniband FDR10 Interconnect
- Large File Space for multiple purpose
 - 10 PetaByte File Space based on IBM GPFS with 200GigaByte/s aggregated I/O Bandwidth
 - 2 PetaByte NAS Storage with 10GigaByte/s aggregated I/O Bandwidth



■ **Innovative Technology for Energy Effective Computing**

- Hot Water Cooling
- Energy Aware Scheduling

■ **Most Energy Efficient high End HPC System**

- PUE 1.1
- Total Power consumption over 5 years to be reduced by ~ 37% from 27.6 M€ to 17.4 M€

- **Software solutions**
 - xCAT
 - Platform LSF/LoadLeveler

- **Cooling solutions**
 - Direct Water Cooling

- **Answers to EE HPC WG**

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.

Measurements: System, Platform and Cabinet

xCAT reports AC power for chassis

- xCAT is out-of-band through AEM IPMPI for iDataplex and xpm for NeXtScale

xCAT reports AC power ≥ 1 per second

AC power is measured directly

Other tools exist to report AC power per rack though Intelligent PDU

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.

Measurements: Node

EAS reports DC power (energy and performance) per node for the job

- *EAS is in-band through ibmaem linux kernel module (and IMM)*
- *Current/voltage could be reported if required*

EAS reports AC power ≥ 1 per second (≥ 10 in 2015)

DC power is measured, AC power is estimated

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.

Measurements: Component

EAS does not reports DC power per component

- CPU and memory power for CPU and GPU could be reported if required*

EAS reports also CPI , GB/s and GFLOps



IBM Response to EEHPC Vendor Form: Energy & Power Measurement Capabilities for IBM p775

Kevin Covi
kcovi@us.ibm.com

Michael J. Ellsworth, Jr.
mje@us.ibm.com

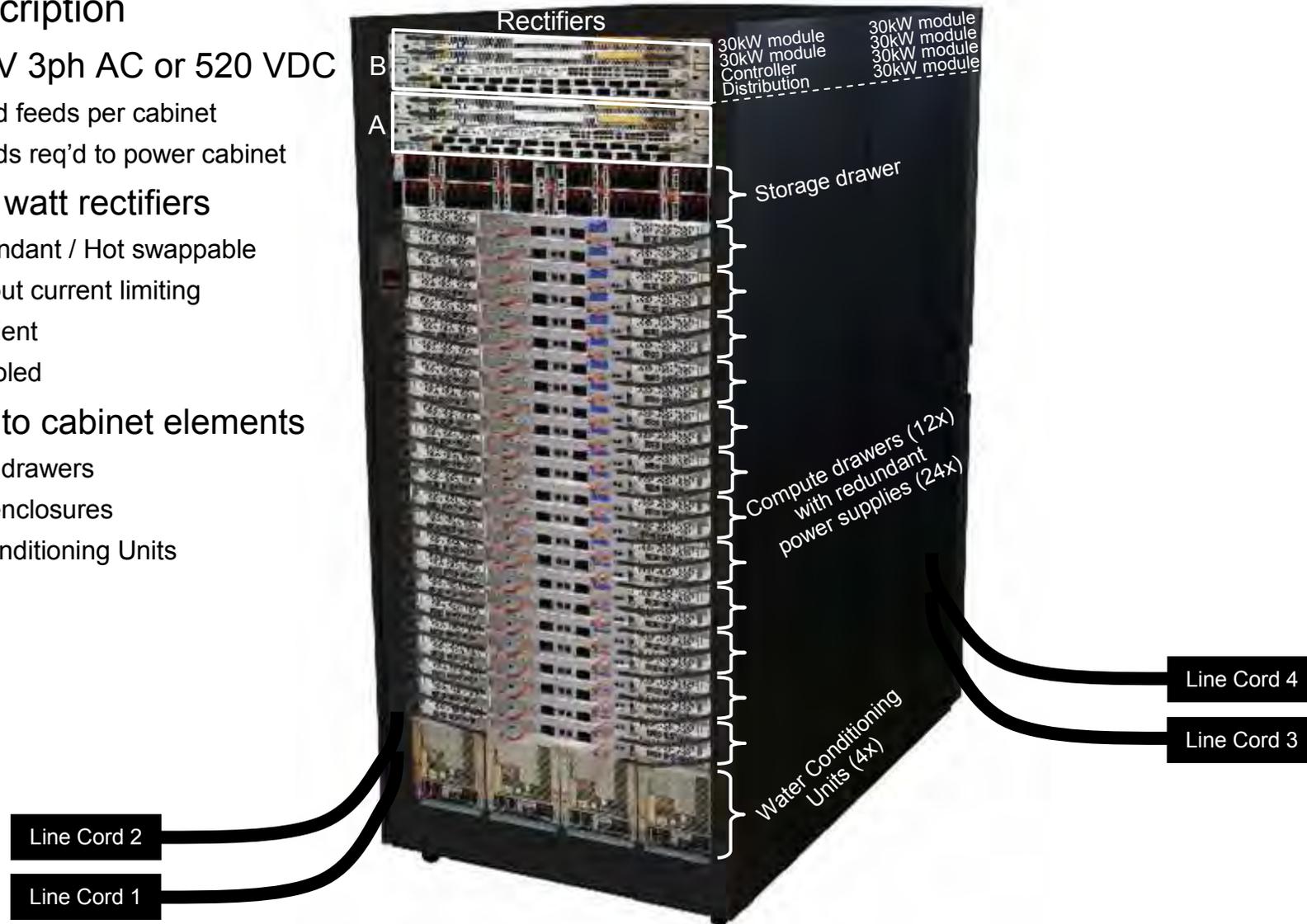
IBM p775 System Development
Poughkeepsie, NY



P775 Cabinet Power Monitoring Capability

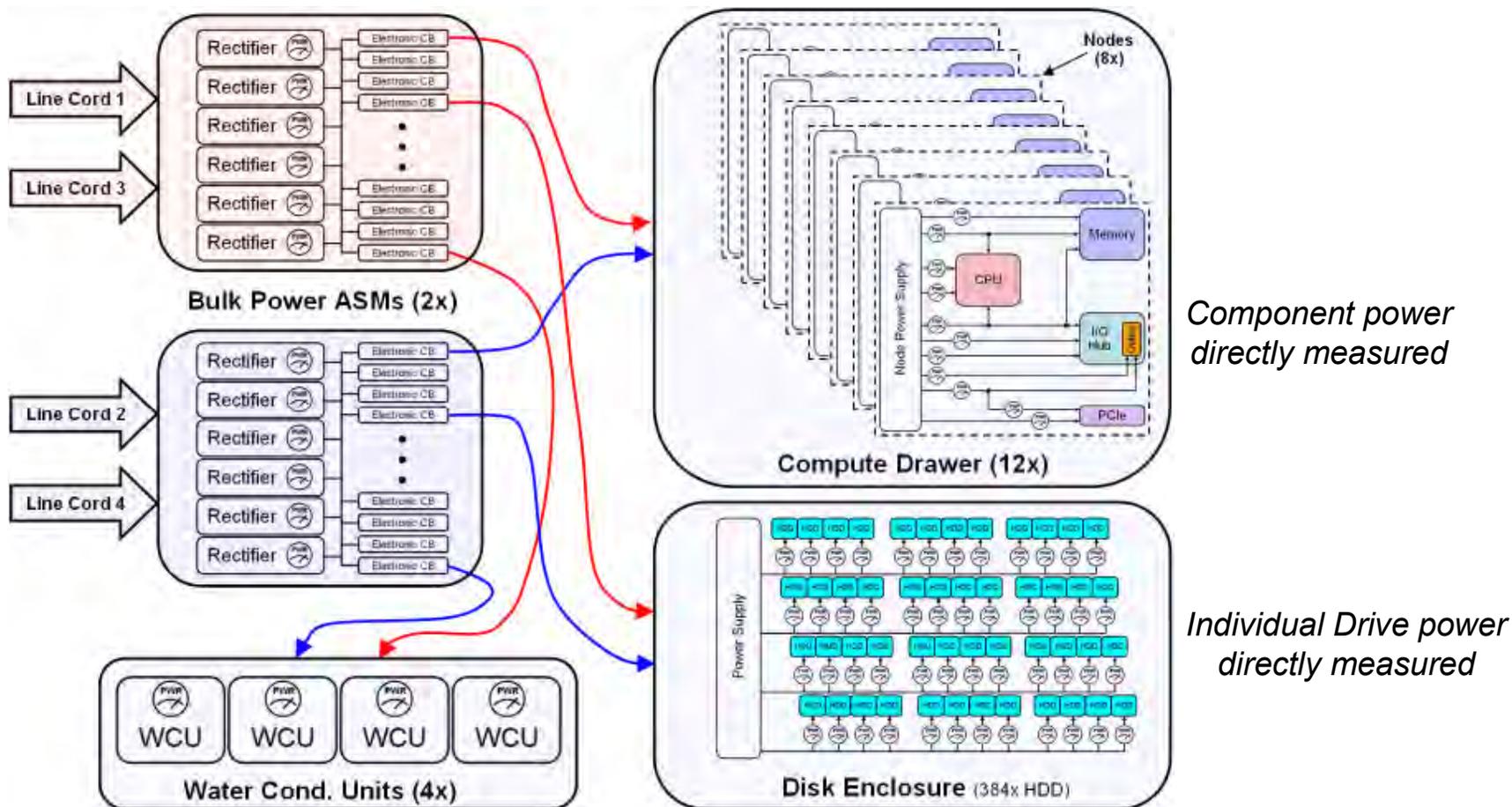
■ Cabinet description

- Input: 480V 3ph AC or 520 VDC
 - 4 line cord feeds per cabinet
 - 3 line cords req'd to power cabinet
- 12 30,000 watt rectifiers
 - N+1 redundant / Hot swappable
 - Active input current limiting
 - 93% efficient
 - Water cooled
- -350 VDC to cabinet elements
 - Compute drawers
 - Storage enclosures
 - Water Conditioning Units



P775 power monitoring diagram

*Power @ ECB
(future enhancement)*



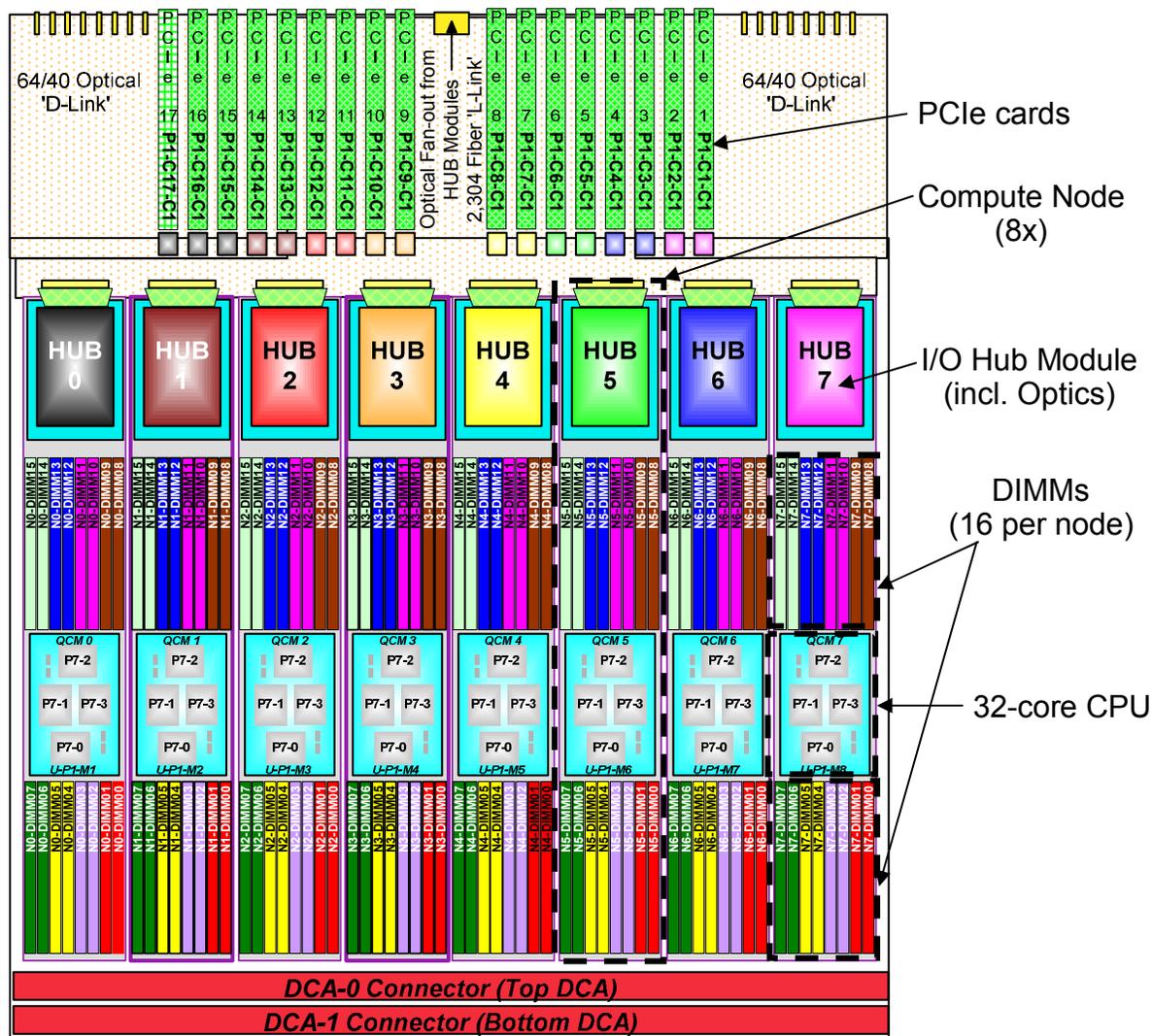
p775 Compute Drawer (8 Compute Nodes)

Drawer description

- Input: -350 VDC
 - 4 feeds per drawer
 - 1 feed req'd to power drawer
- 2 20,000 watt power supplies (not shown)
 - N+1 redundant / Hot swappable
 - 61 voltage levels
 - Water cooled
- 32-55 VDC output to drawer components
 - CPU modules
 - Memory DIMMs
 - I/O Hub
 - Optics
 - PCIe

Power Monitoring

- Direct current & voltage measured
 - 61 voltages & currents
 - Data measured 100x per second
 - Drawer readout updated 1x per second
 - Drawer power estimated from supply efficiency
- Component powers
 - Optics/PCIe directly measured
 - CPU/DIMM/Hub 90% direct 10% estimated due to shared voltage levels on memory & I/O interfaces
 - Direct measurement of all levels at component level would compromise performance & efficiency
 - Flushing fan power estimated
 - Small error since most heat goes to water



IBM comments on Cabinet Power Monitoring

■ Measurements

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

- **Available Now:** Cabinet power estimated from rectifier power using power supply efficiency
- Direct AC power metering will be a cost & reliability detractor
 - What is the value of the incremental increase in accuracy?

❖ Current and voltage readout capability

Mandatory: ≥ 1 per second

Important: ≥ 50 per second

Enhancing: ≥ 250 per second

- **Available now:** 1 Hz. Higher readout rates will be a challenge, especially for large clusters
- Recommend using average power as alternative to increasing readout rate

IBM comments on Node Power Monitoring

■ Measurements

❖ The ability to measure the current and voltage of any and all nodes shall be provided

- **Available Now:** Node power estimated from component power using power conversion efficiency
- **Future Enhancement:** Direct measurement of node power at electronic circuit breaker

❖ Current and voltage readout capability

Mandatory: ≥ 1 per second

Important: ≥ 50 per second

Enhancing: ≥ 250 per second

- **Available now:** 1 Hz. Higher readout rates will be a challenge due to large number of levels (122x) per compute drawer
- Recommend using average power as alternative to increasing readout rate

IBM comments on Component Power Monitoring

■ Measurements

❖ The ability to measure the current and voltage of each individual component must be provided

- **Available Now:** Component power is measured directly *

* When components share a common voltage level, power consumption per component is estimated

❖ Measurement sampling frequency

Mandatory: ≥ 10 per second

Important: ≥ 100 per second

Enhancing: ≥ 1000 per second

- **Available now:** 100 Hz
- Recommend using average power as alternative to increasing sampling rate

Thank You for your Consideration
Questions ?