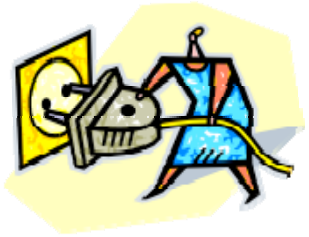
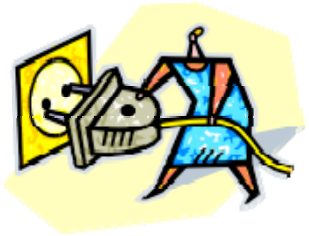


DOE Best Practices Workshop
Power Management
San Francisco, Sept. 28-29, 2010

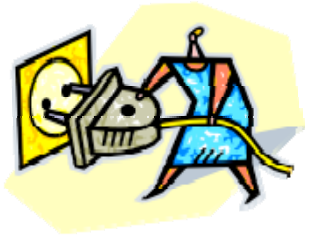
2c: Power-aware system monitoring

Breakout Report



Breakout participants

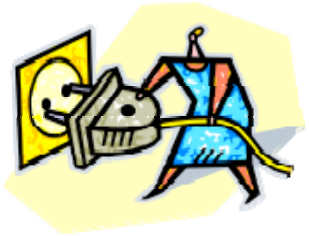
- Susan Coghlan
- Bill Allcock
- Jeff Broughton
- Matthew Campbell
- Kim Cupps
- Thomas Davis
- Marcus Epperson
- Mark Grondona
- Michael Knobloch
- Mike Lang
- Jim Laros
- Josip Loncaric
- Jacques Noe
- Jim Rogers
- Greg Rottman
- David Skinner
- Tisha Stacey
- Ryan Wright
- Mary Zosel



Outline of Breakout Discussion

What is unique about power-aware monitoring vs. other monitoring?

Environmental monitoring or performance monitoring are relatively flat, power monitoring is hierarchical with losses and efficiencies at different levels – you need end-to-end monitoring for power



Experience
Novel / Interesting Approaches

Correlation of applications with hardware events/environmental data

Collection of current draw information at the per second frequency, graphically shows current draw across system as applications are running, using the data to validate system changes made to save power, generating application power utilization signatures. A shorter sampling period would be very useful.

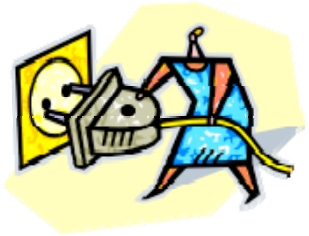
Beginning to correlate data from multiple sources – building mechanical data, system data, sensor data, etc. - proving to be harder then expected.

Looking at having the scheduler feeding data to the chiller plant to warn of an upcoming surge in chilling needs.

Correlation of performance counters with energy draw to increase data gathering frequency. Not clear that this is a valid approach. Is there a possible similar method using temperature data that might be useful for ball parking?

Correlation of changes in temps or other environmentals to potential determine failing equipment

Utilizing building management data to trigger higher system data gathering frequency and ultimately shutting down the system

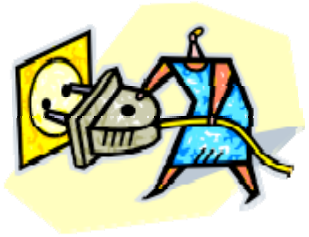


Best Practices in power-aware system monitoring

- Ability to get current draw off individual components (we want to get to where we can pull multiple samples per sec.)
- Integrated facility and system management
- Share thermal histories with vendors
- Analyze your historical data for failure correlation
- Share data from sites increasing their temps
- Run simulations before every major system change
- Do baselines
- Do ROI analysis
- Make sure non-critical monitoring is not in the critical path for operations

Gaps Looking Forward to New Systems

- Want to know whether the key components are within voltage margin or not, also provide events when components fall out of the margin
- Want to get instantaneous measure of any two of voltage, current or power
- Want a software oscilloscope
- Feedback loops for utilizing the data to modify behavior within the facility
- Common standards of measure across all labs
- Will vendors be able to reliably predict onsite power and cooling
- Tools for better power-aware system management
- Tools for applications to monitor their component power usage, particularly high power activities like memory access and data movement.
- Programmable BMC (baseboard management control)

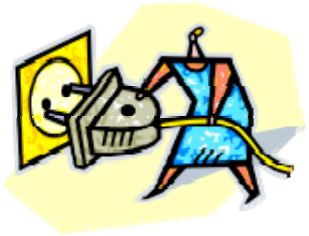


Evolve or start over for future systems?

Evolve (in most areas)

Really large systems need to get data to a centralized location – needs a big tech improvement

May need to start over re the IMPI



Issues shared with large commercial centers

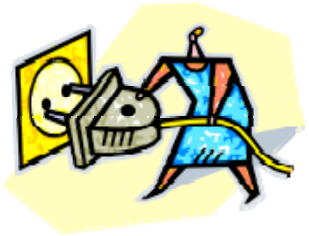
Pretty much everything on the operational side are shared between both HPC and commercial, commercial centers might even drive some areas harder than us

Fine grain data requirements might not overlap (more research related)

Maybe not at the exascale single platform scale

Driven by ROI more than HPC

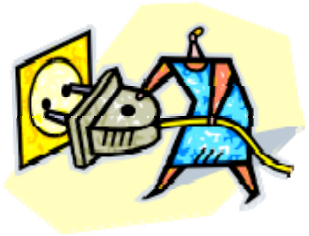
They can look at how they are doing on a single node and multiply



- Hardware/facility/system interfaces to influence

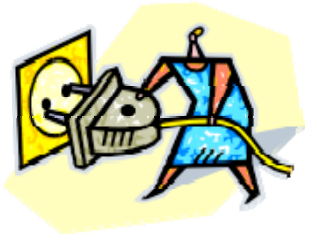
Add watt/hour data output, maybe accumulators, for the hardware

Replace IPMI with something that works consistently and supports the needs of the HPC community



Status of (de facto) standards

- Need public standard for interface for the power data across facility and platform (open source BACNet) both protocols and data format
- IPMI is the current industry standard, but doesn't work well – the protocol is poor, the implementations are poor, ability to not implement portions of the protocols, not reliable for critical tasks
- Monitoring software/methods are all over the place, no dominant player because none of them provide the whole system well
- Need an API that exists on the edge of the RAS system – this is the PAPI for Power (and Environmentals)
- Watch AMQP as a possible exascale solution
- SMASH (system management architecture for server hardware)



Other key findings

For commodity systems, monitoring out-of-band required

High quality, high precision, high accuracy sensors are expensive

The amount of data pulled could be more than the data from the simulations

Modeling could reinforce the monitoring

Good infrastructure design for gathering the data is necessary (Blue Gene approach of polling data serially is too slow)

Different levels of data frequency – operations may just care about coarser grain, researchers may want finer grain – also difference between sampling and error events