BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Power Monitoring At NCSA ISL and Blue Waters

Salishan 2013 Conference

Craig P Steffen
Blue Waters Science and Engineering Applications Suppport Group
csteffen@ncsa.illinois.edu

GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

NCSA: Measuring Power Use and Power Effectiveness of User Applications using Sustained System Performance Metrics

- Use a **real** user application
  - It is important to use a real user application that makes realistic demands on the system
- Measure power of **whole** machine
  - As close as possible to data-center/machine boundary (where the facility charges the owner)
  - Buy (create) and install power monitoring
  - Make the data *available* in a useful form to anyone who can use it
- Measure **whole** application (wall clock) **run time**
  - Scheduler deals with wall clock time (that's the time other applications can't be running)
  - User allocations charged wall clock time

Overview: Power Monitor System Progression

- 32 node Innovative Systems Lab "Accelerator Cluster" system
  - Power monitoring on single node
- 128 node ISL/ECE "EcoG" system
  - Power monitoring on 8 nodes block
- 25,712 compute-node Cray XE/XK "Blue Waters" system
  - Whole-system power monitoring per building transformer *and* per computational cabinet

Stage1: Innovative Systems Lab Accelerator Cluster (AC) (2008?-2012)

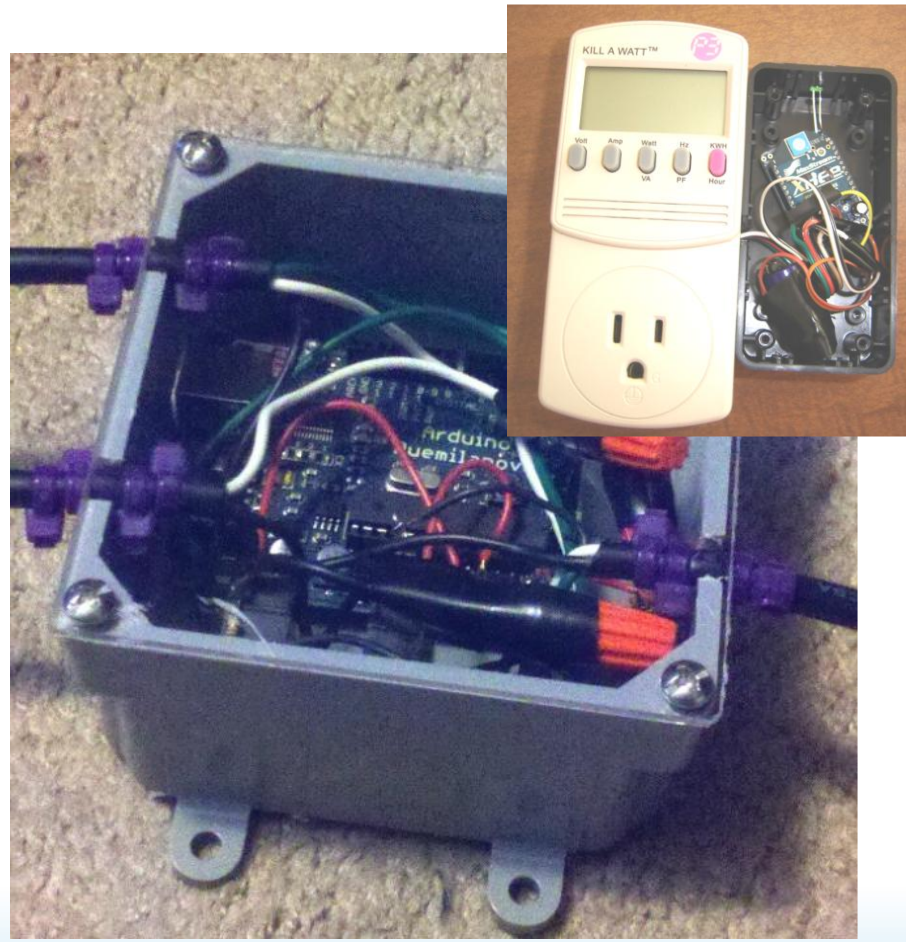
- 32 nodes
- HP xw9400 workstation
  - 2216 AMD Opteron 2.4 GHz dual socket dual-core
  - Infiniband QDR
- Each node: **Tesla S1070 1U GPU Computing Server**
  - 1.3 GHz Tesla T10 processors
  - 4x4 GB GDDR3 SDRAM

Commercial Power Meters (~2010)

	price	readout	monitors
Kill-a-watt	$50	LCD display	120 VAC
PowerSight PS3000	$2450	asynch	100-250 VAC
ElitePro™ Recording Poly-Phase Power Meter	$965	asynch	120 VAC
Watts Up Smart Circuit 20	$194.95	Web-only	120 VAC

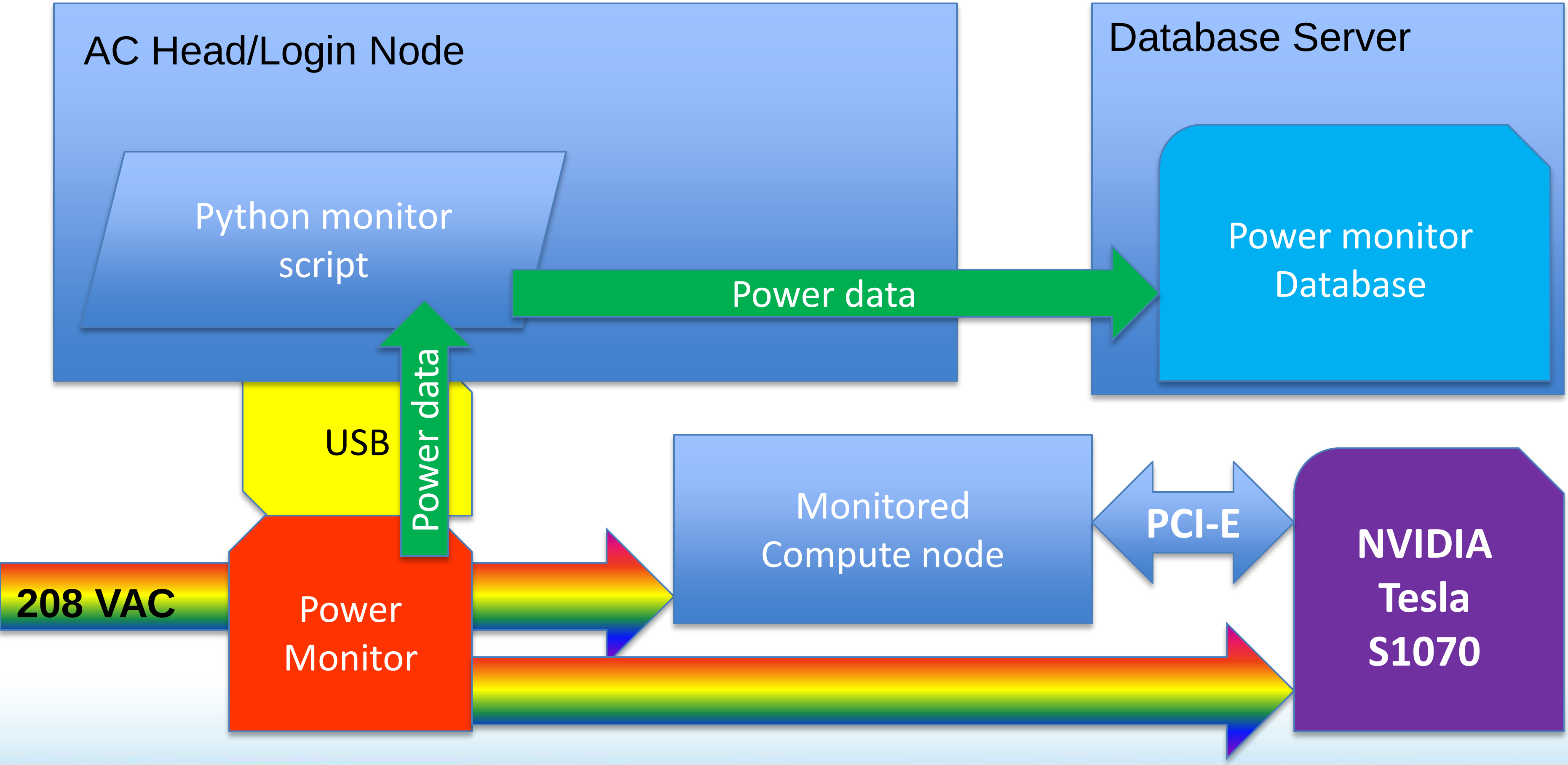
AC Power Hardware Monitor Conclusion: Nothing Suitable Exists; Make One (monitor single node)

- V1: Kill-a-watt based Xbee "Tweet a watt"
  - Wireless transmitter
  - Voltage and current
- V2 and V3: Arduino based power monitors
  - 1 Arduino Duemilanove per chassis
  - 1 Manutech MN-220 20A AC power sensing transformer per measured channel
  - Arduino forms RMS value of AC voltage → current (5 times per second)



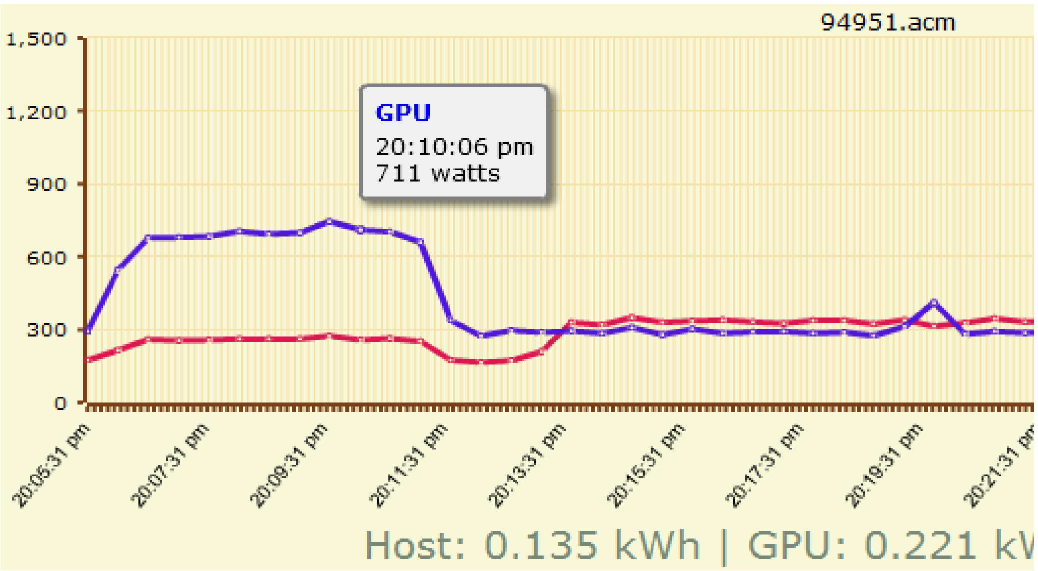
	price	readout	monitors
Kill-a-watt	\$50	LCD display	120 VAC
PowerSight PS3000	\$2450	asynch	100-250 VAC
ElitePro™ Recording Poly-Phase Power Meter	\$965	asynch	120 VAC
Watts Up Smart Circuit 20	\$194.95	Web-only	120 VAC
Our Arduino-based monitor	**~\$100 (2 channels)**	**USB text (python)**	**base config: 120/208/250 VAC**

Power Data Harvesting During Job



AC Power Monitor Software: Tied Into Job Software, Exports Data Automatically

AC Power Utilization

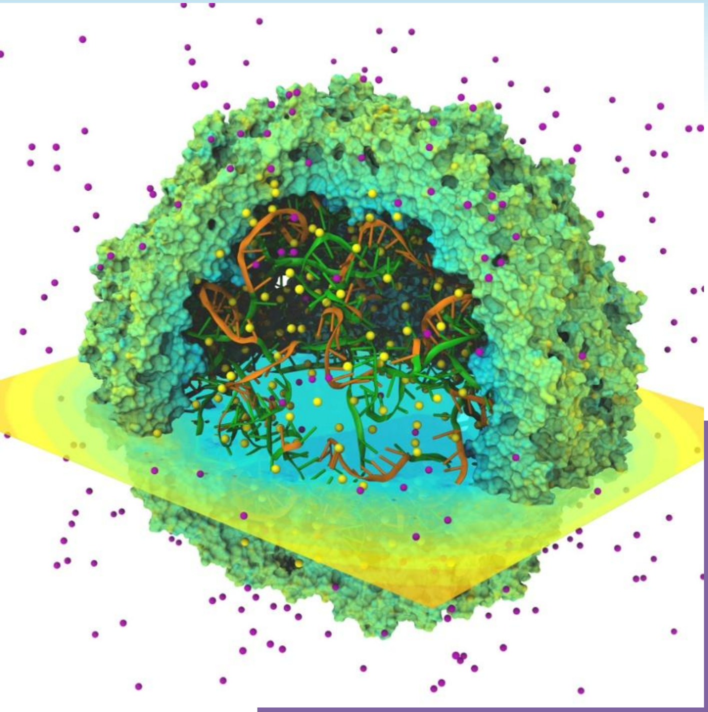


JSON Data

- Blue = GPU power
- Red = CPU node power

- Use qsub to request "powermon" feature
- Prolog script creates link to power graph (left) and raw power data file
- Power graph available during the run; graph and link to data works afterward

(GPU has separate power supply)



Application Speedup Summary
(small or single-node versions of apps)

	Raw GPU speedup (wall clock time)	Speedup scaled by (GPU+node)/node power ratio
NAMD	6	2.8
VMD	26	10.5
QMCPack	62	23
MILC	20	8

Stage 2: EcoG Student-built GPU Cluster (2010-2012)

ECE Independent study course in "cluster building" to create high performance GPU-based architecture

- Maps to GPU math capabilities
- Frequent but not constant node-to-node updates
- Likely target apps:
  - Molecular dynamics
  - Fluid dynamics
  - HPL for testing

- **128 nodes**
- **Tesla 2050 GPUs primary computing element**; single modest CPU per node
- Single-socket motherboard
- Each node:
  - Intel® Core i3 2.93 GHz CPU
  - 4 GB RAM main memory (smal to lower power footprint)
  - 1 two-port QDR infiniband
- High-performance GPUs, lower power CPUs
- NFS root file system (stateless nodes)

EcoG Power Monitoring: modified rack PDU

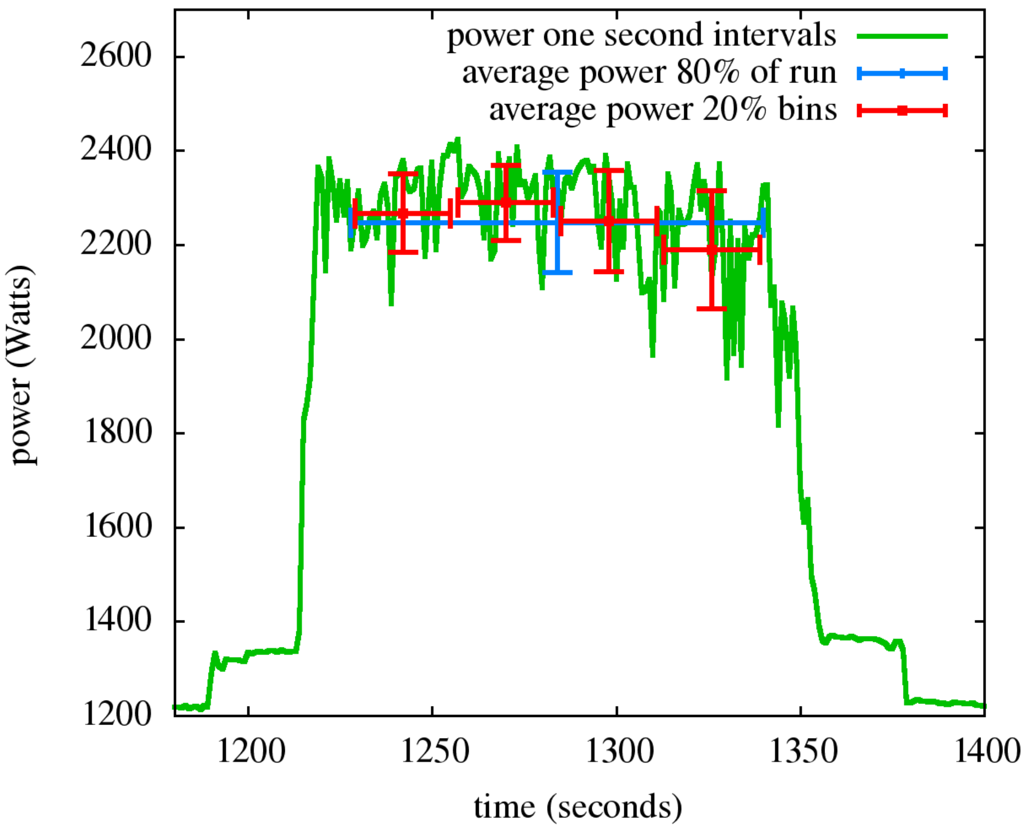
Re-used rack-mounted PDU

- 2 voltage probes for 208V power legs
- 2 clamp-on current probes for current measurement
- Probes secured INSIDE enclosure
- Asynchronous readout via text file

Green 500 Run Rules and EcoG Submission (HPL known quantity for entire system)

- Green 500 run rules as of fall 2011:
  - Fraction: ANY sub-fraction can be measured and scaled up
  - Measurement position: anywhere
  - Time: at least 20% of run in the middle 80%
  - Subsystems: Compute nodes (only) required
- NCSA's EcoG (3 page) Submission + technical report
  - 8 of 128 compute nodes
  - Power measured upstream of node (entire chassis)
  - *We decided* to average power from **whole 80%** of run

HPL is not an application but it is a valid system stress test



Sources of Uncertainty in Green 500 Results

- Subsampling allows small numbers of elements to skew the average (up or down)
- Measuring power inside the system leaves out efficiency of AC/DC or DC/DC conversion efficiency (10 or 20%)
- Not requiring whole run allows lower power section of run to be used (~3% effect)
- Leaving out subsystems (network, head nodes, storage, cooling etc.) artificially lowers power cost of system

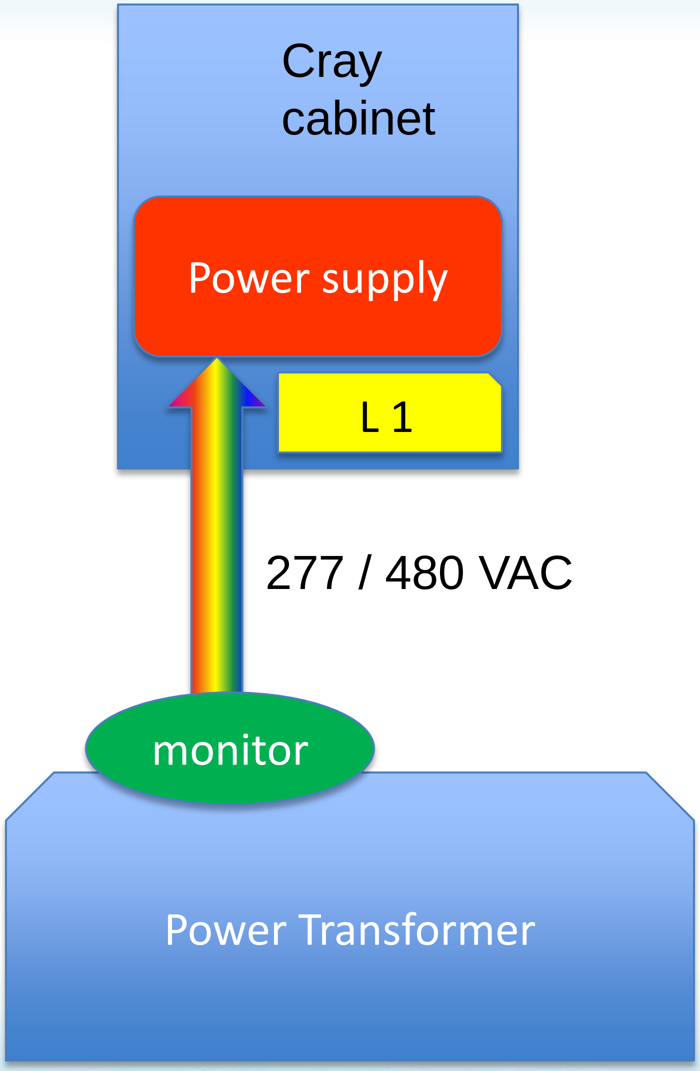
NCSA has been working with EEHPCWG to create a new power monitoring and reporting specification

- Energy Efficient HPC Working group: http://eehpcwg.lbl.gov
- Creating a specification for whole-system power measurement for application efficiency characterization
- Hope to convince Top-500, Green-500, and Green Grid to adopt this standard for "power" part of submissions
- Also to drive specifications for machines and comparisons
- Compare power measurement and performance measurement (general measurement of value)

- Three quality levels
- Level 1 ≈ current Green-500
- Level 3 = current best possible
  - Requires 100% of system
  - Power measured upstream of AC/DC conversion OR loss is accounted for
  - 100% of parallel run used in average power calculation
  - ALL participating subsystems required to be MEASURED
  - Metering devices must be integrating total energy meters

Stage 3: Cray XE6/XK7 Blue Waters (2012--)

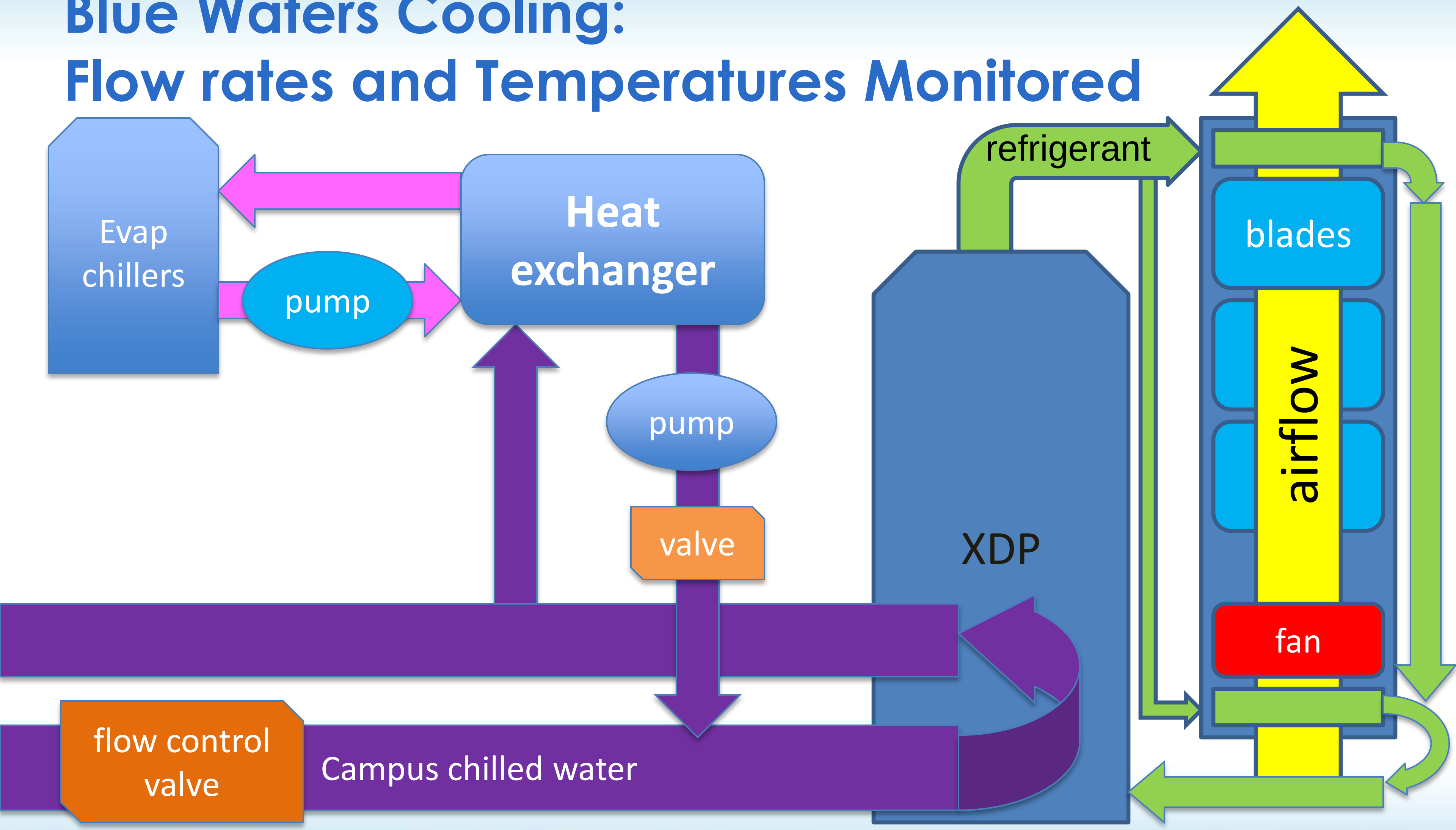
- **3D Torus Gemini network topology 24x23x24**
- **Mix of XE and XK Cray nodes (XK in contiguous block)**
  - **22640 XE compute nodes**
    - **Each node 2-die Interlagos processor 16 Bulldozer cores**
    - **64 GB of RAM**
  - **3072 XK compute nodes**
    - **Each node 1-die Interlagos processor 8 Bulldozer cores**
    - **1 Kepler K20X GPU**
    - **32 GB of RAM**
- **3 large Sonexion Lustre file systems (usable capacities):**
  - **/home (2.2 PB)**
  - **/project (2.2 PB)**
  - **/scratch (21.6 PB)**



Blue Waters Power Delivery

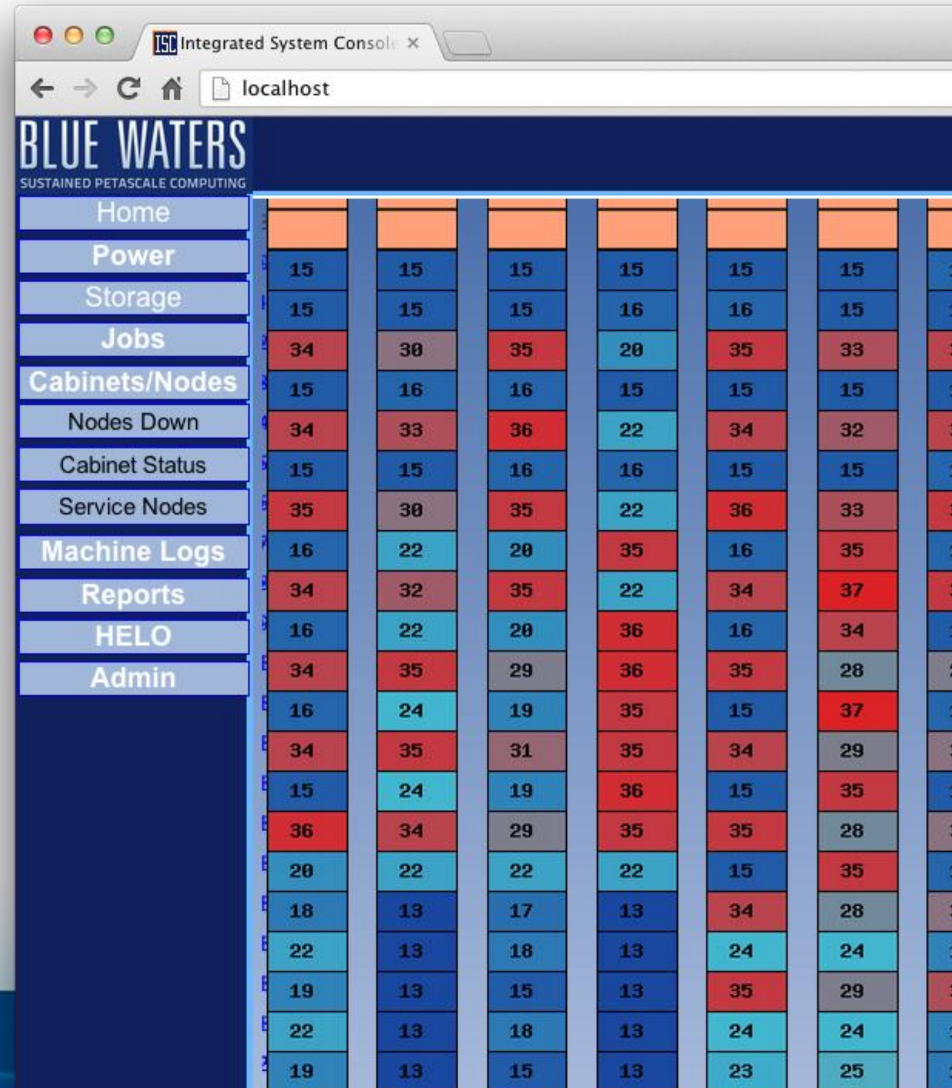
- L1 controller monitors and reports power from within (each) cabinet
- Each transformer output monitored and logged

Blue Waters Cooling:
Flow rates and Temperatures Monitored



Integrated Systems Console

- Tool being developed as system visualization/monitoring/analysis tool for Blue Waters admins
- Selected features are also available on the Blue Waters Portal for users
- All this information in one place allows triage of cooling problems
- All data together



Selected Applications on Blue Waters

During Acceptance four large-scale science applications:
VPIC, PPM, QMCPACK and SPECFEM3DGLOBE
sustained performance **>1PF on Blue Waters**

Weather Research & Forecasting (WRF) run on Blue Waters is the largest WRF simulation ever documented

- Simulating hurricane Sandy
- Submitted as paper to SC 2013

These applications are part of the NCSA Blue Waters Sustained Petascale Performance (SPP) suite and represent valid scientific workloads.

(sizes are in MPI ranks) App	NSF petascale	NSF non-petascale	SPP full system (>1.0 PF/s)	SPP Interlagos	SPP Kepler K20X
turbulence (PSDNS)	360,000				
nwchem				80000 , 8000	
specfem3d			693,600	173400 , 21600	
vpic			180,224	131072, 73728, 8192	
ppm			85,668	33024, 32250, 2112	
milc	316,416				
milc		2048			
milc				65856, 8232	
wrf		512			
wrf				72960, 4920	
qmcpack			90,000	153600, 76800, 38400	700
namd (100M)	25,650				
namd				20000 , 768	768
chroma					768
gamess					1536
paratec		512			
Full system:	XE: 724,480	XK: 49,152		**Composite SPP 1.3 PF/s**	

Friendly User Period Power Usage

- Friendly user Period: users running real production workload
- Average Power February: **9.46 MW**
- Average Power March: **9.71 MW**

Good measurements + effective delivery = ability to understand system behavior

- Real effectiveness is in terms of performance per watt
  - Real performance in terms of wall clock time
  - Real power in terms of full system power
- Get the data to those who can use it

Thanks and References

Thanks to:

- National Science Foundation
- State of Illinois
- Microsoft
- IBM Linux Technology Center

References

Enos, J.; Steffen, C.; Fullop, J.; Showerman, M.; Guochun Shi; Esler, K.; Kindratenko, V.; Stone, J.E.; Phillips, J.C., **"Quantifying the impact of GPUs on performance and energy efficiency in HPC clusters,"** *Green Computing Conference, 2010 International* , vol., no., pp.317,324, 15-18 Aug. 2010
doi: 10.1109/GREENCOMP.2010.5598297

http://www.ncsa.illinois.edu/News/Stories/PFapps/

http://www.nersc.gov/research-and-development/performance-and-monitoring-tools/sustained-system-performance-ssp-benchmark/