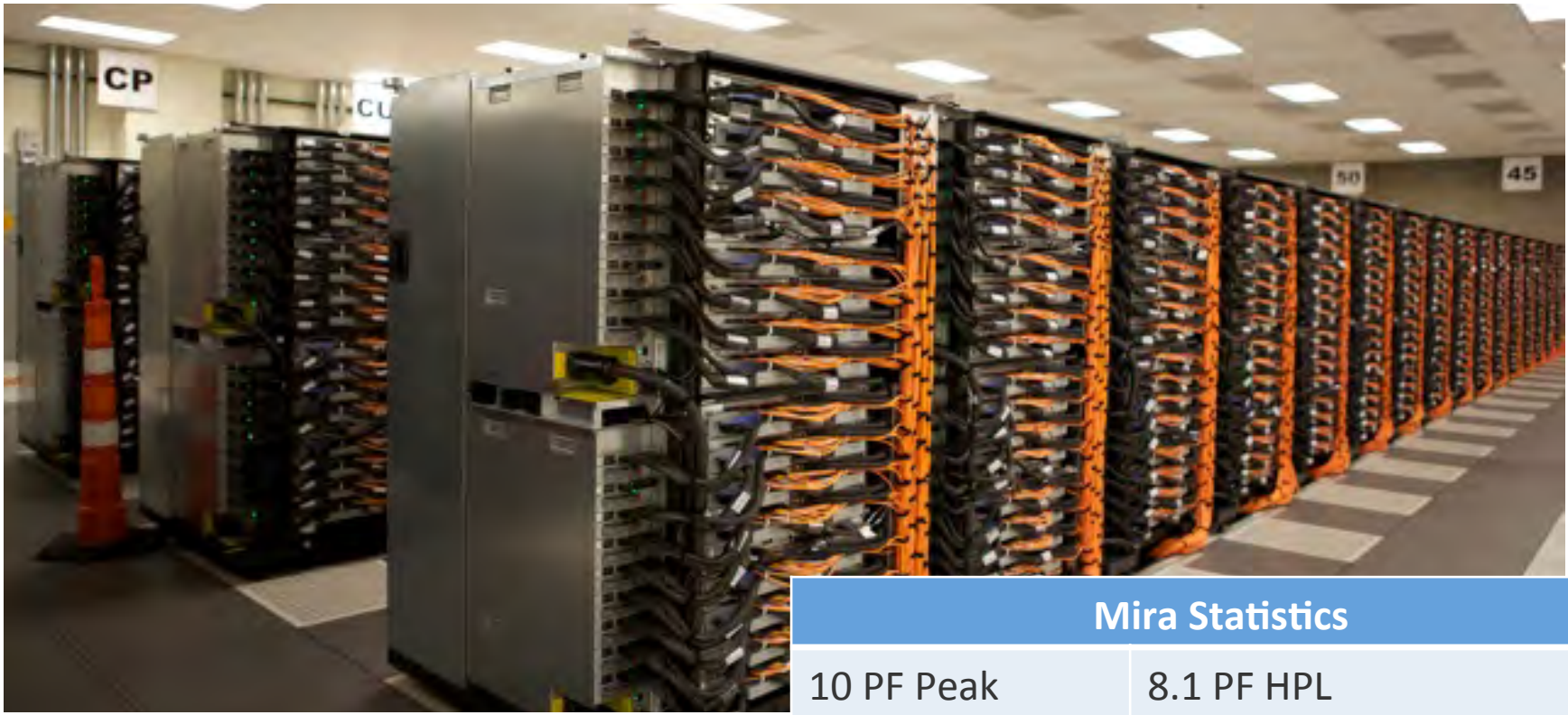
Power Measurements on Mira
Argonne Leadership Computing Facility

Susan Coghlan
Mira Project Manager
Argonne National Laboratory

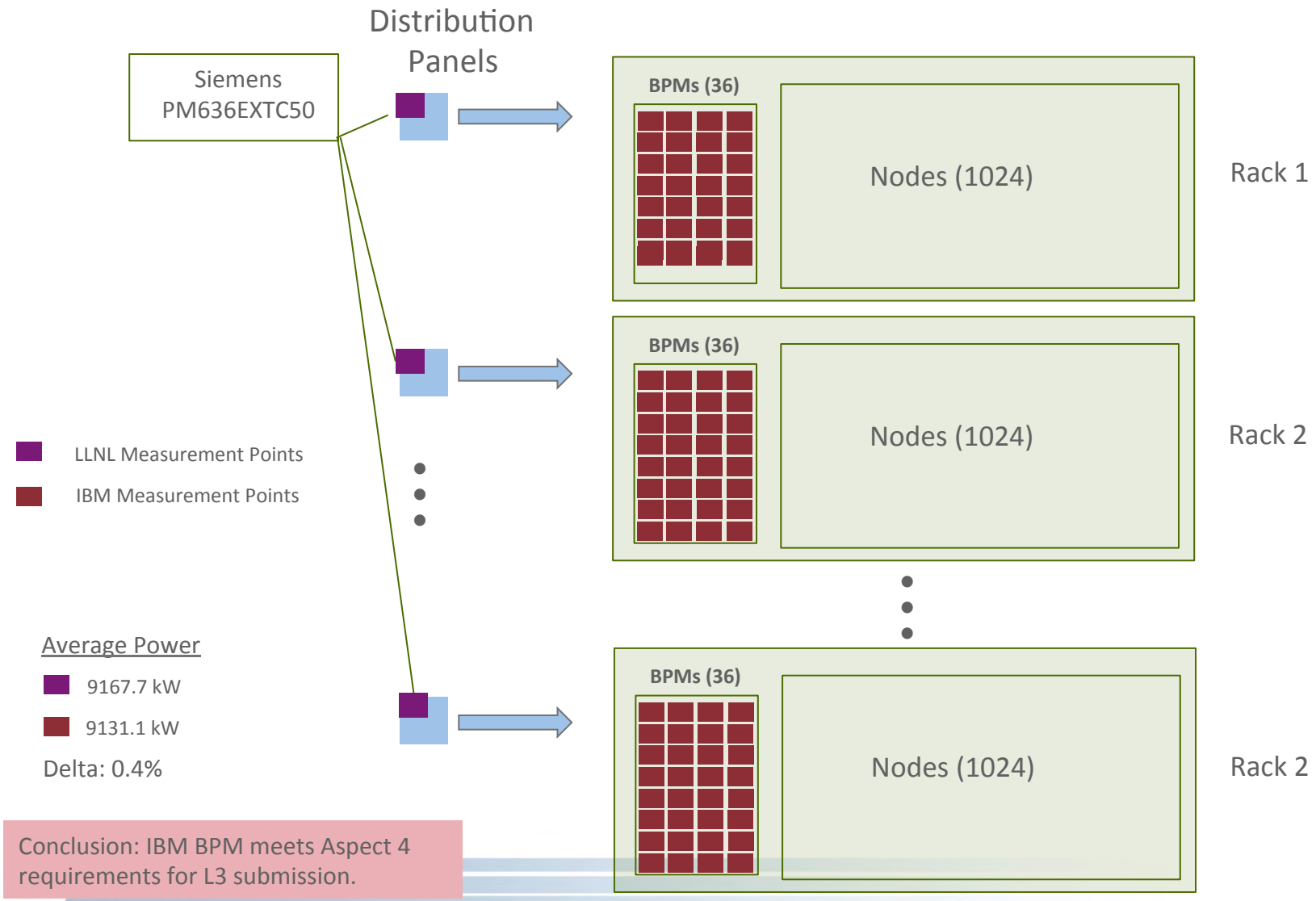
November 14, 2012

The Mira system



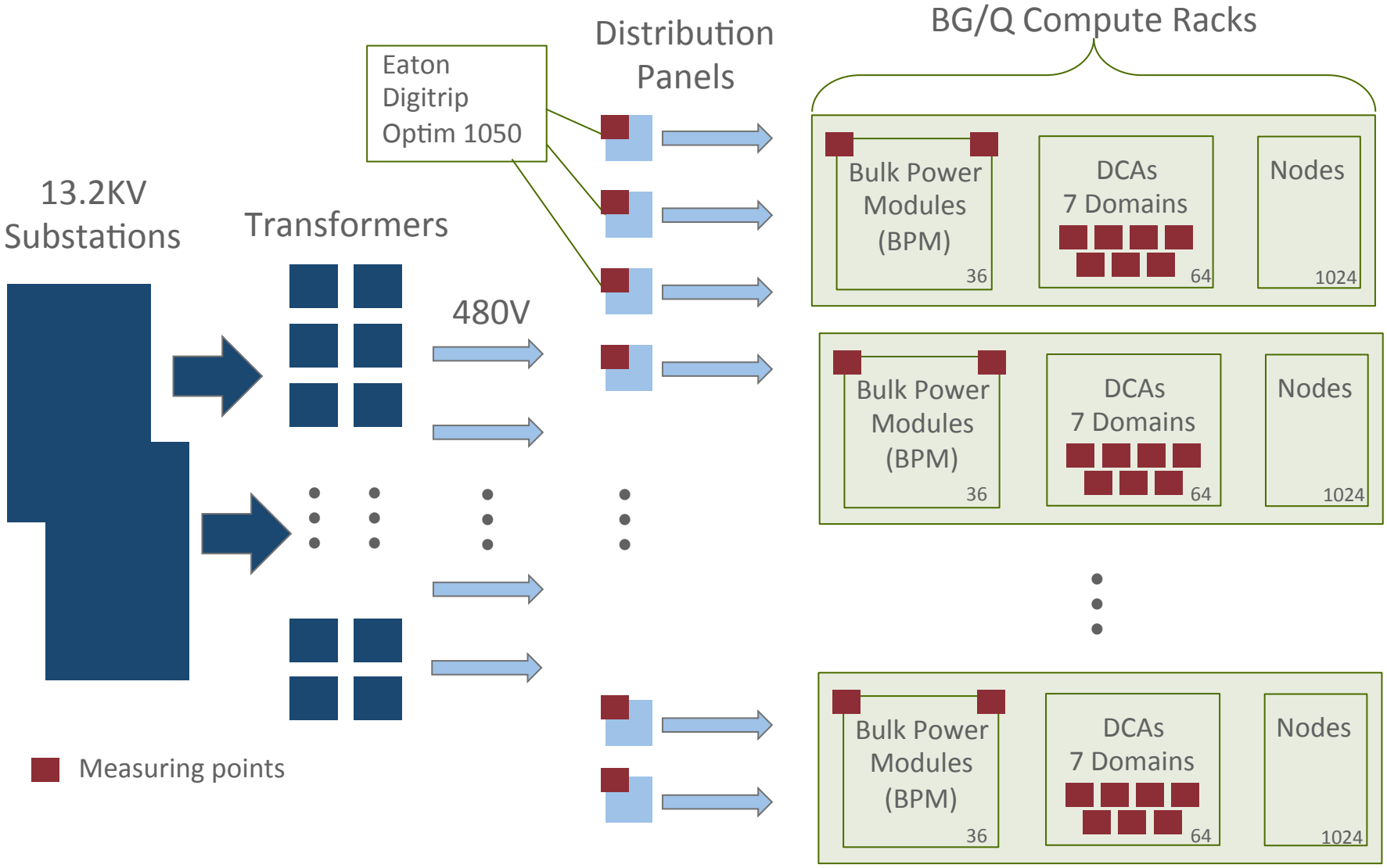
Mira Statistics	
10 PF Peak	8.1 PF HPL
48 Racks	48K compute nodes
768K cores	3.1 B threads
768 TB memory	16 GB memory per node
28 PB disk	240 GB/s
4.8MW peak	48 racks at 100kw each

LLNL power measurement locations



Conclusion: IBM BPM meets Aspect 4 requirements for L3 submission.

Mira 480V compute power distribution system



Data is very different between measuring points Correlating it is non-trivial!

	Panels (10)	BPMs (1728)	DCAs (3072)
Frequency*	5 mins	~5 mins	560 ms
Coverage	5 racks per panel	9 per 256 nodes 36 per rack	2 per 32 nodes 64 per rack
Data per measuring point	phase A, phase B, phase C 3 points	input, output current, voltage 4 points	7 voltage domains current, voltage 14 points
Data points per full system measurement	30 per measurement 360 per hour	6,912 per measurement 82,944 per hour	43,008 per measurement ~276,480,000 per hour
Benefit	Cost accounting	Tracking BPM efficiency, monitoring	App power signature, impact of changes to algo, next generation research

* Current frequency of measurements

DC power information

- Domains:
  - Domain 1 (0.8V): BQC core logic power
  - Domain 2 (1.4V): DDR3 SDRAM, BQC DDR3 I/O
  - Domain 3 (2.5V): Optical module power
  - Domain 4 (3.3V): Optical module power
  - Domain 6 (1.5V): BQC and BGL HSS I/O
  - Domain 7 (0.9V): BQC core array power
  - Domain 8 (1.08V): BQL core power
- **New firmware coming to improve these measurements:**
  - Measures 2000 times per second
  - Provide integrated cumulative energy
- Go see Paul Coteus' talk tonight for the gory details:

  Power and Energy Measurement and Modeling on the Path to Exascale BoF

  5:30pm to 7:00pm, 255-EF

  Last talk

BG/Q Power Monitoring library

Venkatram Vishwanath, Susan Coghlan, Vitali Morozov, Kalyan Kumaran, Michael E. Papka, with much assistance from Paul Coteus and Yutaka Sugawara at IBM

- Captures application power profile
  - Currently only MPI, SPI coming later
- Simple to use
  - Call PMQ_Initialize () after MPI_Init
  - Call PMQ_Finalize() before MPI_Finalize
  - Link with library
- CSV summarized, each nodeboard, each time step
  - Time stamp (NEW), Ticks since job boot using H/w counters,
  - Row, Column, Midplane, Nodeboard,
  - Total of current*voltage for all domains,
  - Current* voltage for each domain
- Background
  - Low overhead (0.2%)
- Compatible with 1 to 64 MPI ranks per node
- Will be available as open source to the community

Data samples from each type of measuring point

Distribution Panel: Excel spreadsheet, tab per panel

Index	Date	Time	Phase A	Phase B	Phase C	Sum	volts	powerfactor	Watts (not phase adjusted)	Watts (phase Adjusted) *
1	10/23/12	17:00:00	337	337	337	1011	480	0.9	485280	252158.8848
2	10/23/12	17:05:00	337	337	337	1011	480	0.9	485280	252158.8848
3	10/23/12	17:10:00	337	337	337	1011	480	0.9	485280	252158.8848
4	10/23/12	17:15:00	337	337	337	1011	480	0.9	485280	252158.8848
5	10/23/12	17:20:00	337	337	337	1011	480	0.9	485280	252158.8848

* $P(W) = \sqrt{3} \times PF \times I(A) \times VL\text{-}L(V)$

BPM: text file, one per time step

LOCATION	TIME	INPUTVOLTAGE	INPUTCURRENT	OUTPUTVOLTAGE	OUTPUTCURRENT
R2E-B1-P8	2012-10-23-21.50.50.929019	+2.81594000000000E+002	+5.39100000000000E+000	+5.08360000000000E+001	+2.76560000000000E+001
R2E-B1-P7	2012-10-23-21.50.50.926648	+2.81156000000000E+002	+5.43800000000000E+000	+5.10310000000000E+001	+2.80160000000000E+001
R2E-B1-P6	2012-10-23-21.50.50.924533	+2.80688000000000E+002	+5.45300000000000E+000	+5.09380000000000E+001	+2.81880000000000E+001
R2E-B1-P5	2012-10-23-21.50.50.921099	+2.80625000000000E+002	+5.39100000000000E+000	+5.09530000000000E+001	+2.77030000000000E+001

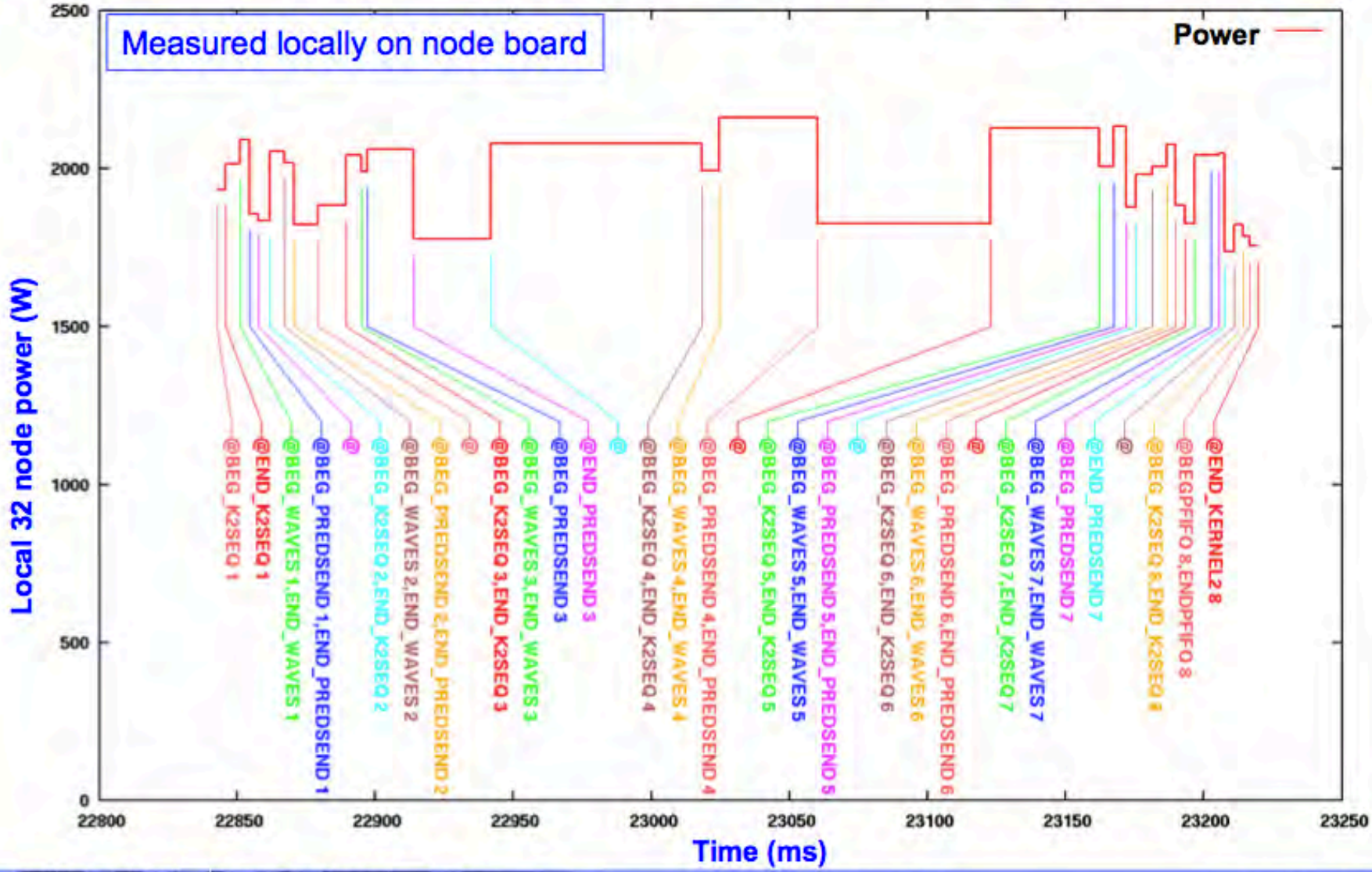
BGQ Power Monitoring Library: csv files, one per nodeboard

```
Mon Nov 12 03:44:02 2012,1352691842,1910783200991,6,1,14,1912.3192,1153.5075,347.7920,57.7530,56.6760,199.9818,49.1508,47.4582
Mon Nov 12 03:44:02 2012,1352691842,1919743211681,6,1,14,2467.5088,1452.2912,603.7532,57.8138,56.2064,200.9434,48.6845,47.8163
Mon Nov 12 03:44:03 2012,1352691843,1928703212291,6,1,14,2397.4751,1377.8339,608.2157,57.2928,56.2057,200.6696,49.5104,47.7469
```

Comparisons

	Panels	BPMs	
Graph500		**Input**	**Output**
Watts (Avg)	2,268,402	2,313,827 (1.02%)	2,170,254 (93.8%)
Watts (Max)	2,576,211	2,912,640 (1.13%)	2,670,385 (91.7%)
Watts (Min)	1,617,708	1,516,627 (93.8%)	1,401,202 (92.4%)
Watt/hrs	4,725,838	4,851,908 (1.03%)	4,552,176 (93.8%)

Correlation of power consumption during Graph 500



	Level 1	Level 2	Level 3
Aspect 1a: granularity of power measurements	1 instantaneous power sampling per second	1 instantaneous power sampling per second	continuously integrated total energy ✔
Aspect 1b: timespan of power measurements	at least one power averaged measurement covering at least 20% of the run	a time series of equally spaced power averaged measurements	a time series of equally spaced integrated total energy values ✔
Aspect 1c: reported analyzed measurements	core phase average power	core phase average power, whole application average power, idle power	core phase average power, whole application average power, idle power ✔
Aspect 2: machine fraction	the greater of 1/64 of the machine or 1 kW	the greater of ⅛ of the machine or 10 kW	whole machine ✔
Aspect 3: subsystems included	all participating subsystems, either measured or estimated	all participating subsystems, either measured or estimated	all participating subsystems must be measured ✔
Aspect 4: power measurement point	upstream of power conversion OR power conversion loss modeled with manufacturer data	upstream of power conversion OR power conversion loss modeled with off-line measurements of single power supply	upstream of power conversion OR power conversion measured simultaneously during the same run ✔