Overview and Power Monitoring of NCSA-UIUC-NVIDIA “EcoG” Computational Cluster

Craig Steffen

NCSA Innovative Systems Lab

at the SC 2010 Green500 BOF

November 18, 2010

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

EcoG Design Goals

- Experiment with low-power, 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 works passably well
- High-performance GPUs, lower power CPUs
- RAM (which also consumes power) just bigger than GPU
- NFS root file system (no hard drive on nodes)

EcoG Final Configuration

- 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
  - 1 two-port QDR infiniband card

EcoG joins "AC" and "Lincoln" GPU-accelerated clusters at NCSA

Will be used soon in scientific development

Donated or Recycled Hardware

- 128 Tesla 2050 GPU cards donated by NVIDIA
- Significant parts of infiniband fabric donated by QLogic

- Ethernet cables, power cables, PDUs, recycled from retired NCSA “Mercury” and “Tungsten” systems

- EcoG cluster sits on food service shelves and occupies 18 square feet

System Assembled and Installed by Students

~13 students from UIUC ECE/CS departments in cluster-building independent study

2 graduate students from the chemistry department

Mike Showerman, Jeremy Enos, Luke Scharf, and Craig Steffen from ISL

Sean Treichler from NVIDA made the scheduling modifications to HPL

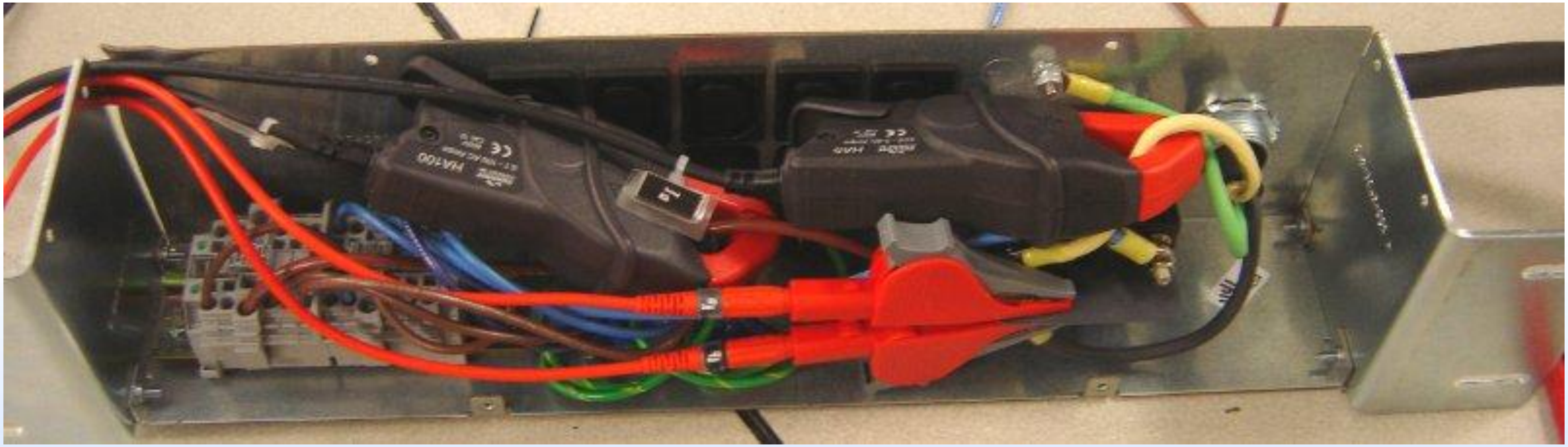
HPL Function Division

- Intel CPU:
  - main application loop
  - panel factorization
  - DTRSM update
  - final triangular solve
  - residual check
- Tesla GPU:
  - Update DGEMM
  - Rowswap scatter/gather

Power Monitoring Setup: Voltage and Current Probes

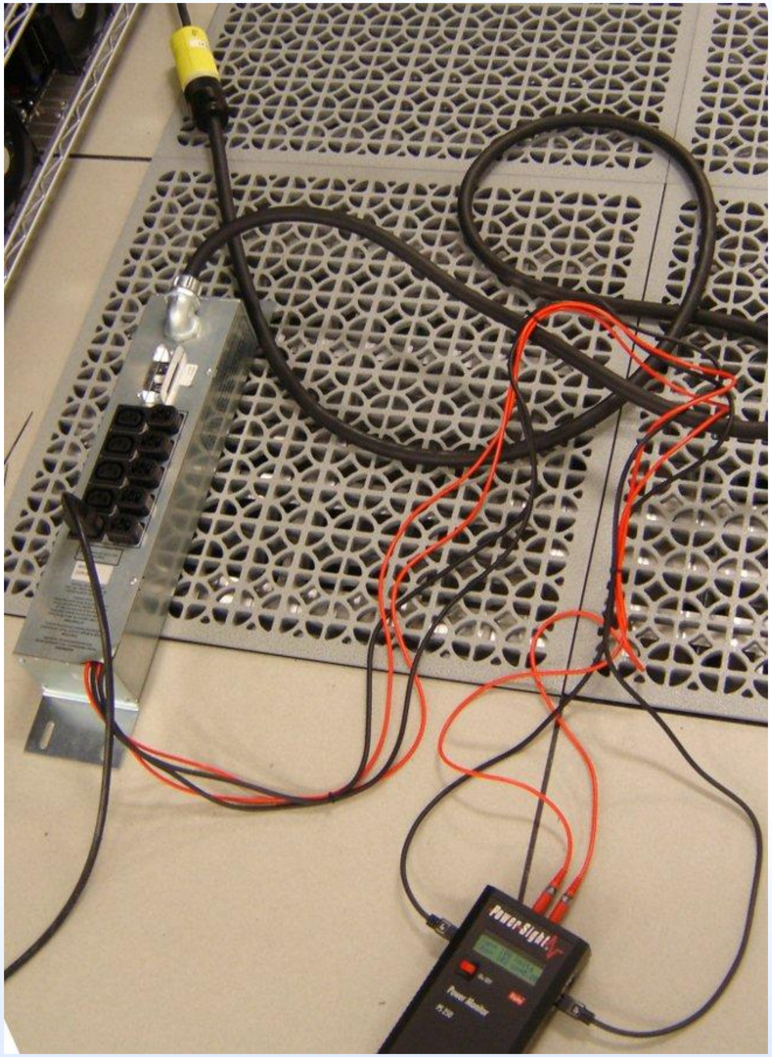
Re-used rack-mounted PDU

- 2 voltage probes for 208V power legs
- 2 clamp-on current probes for current measurement
- Probes secured INSIDE enclosure



Final Power Monitoring Setup: Enclosed for Convenience and Safety

- L6-30 208V 30A input
- Voltage and current instrumented PDU
- 2 outputs each for 4 cluster nodes
- Powersight voltage/current monitor external



PowerSight power monitor

- Records sampled data to internal memory
- Time-stamped data read out later via serial

Power Data File

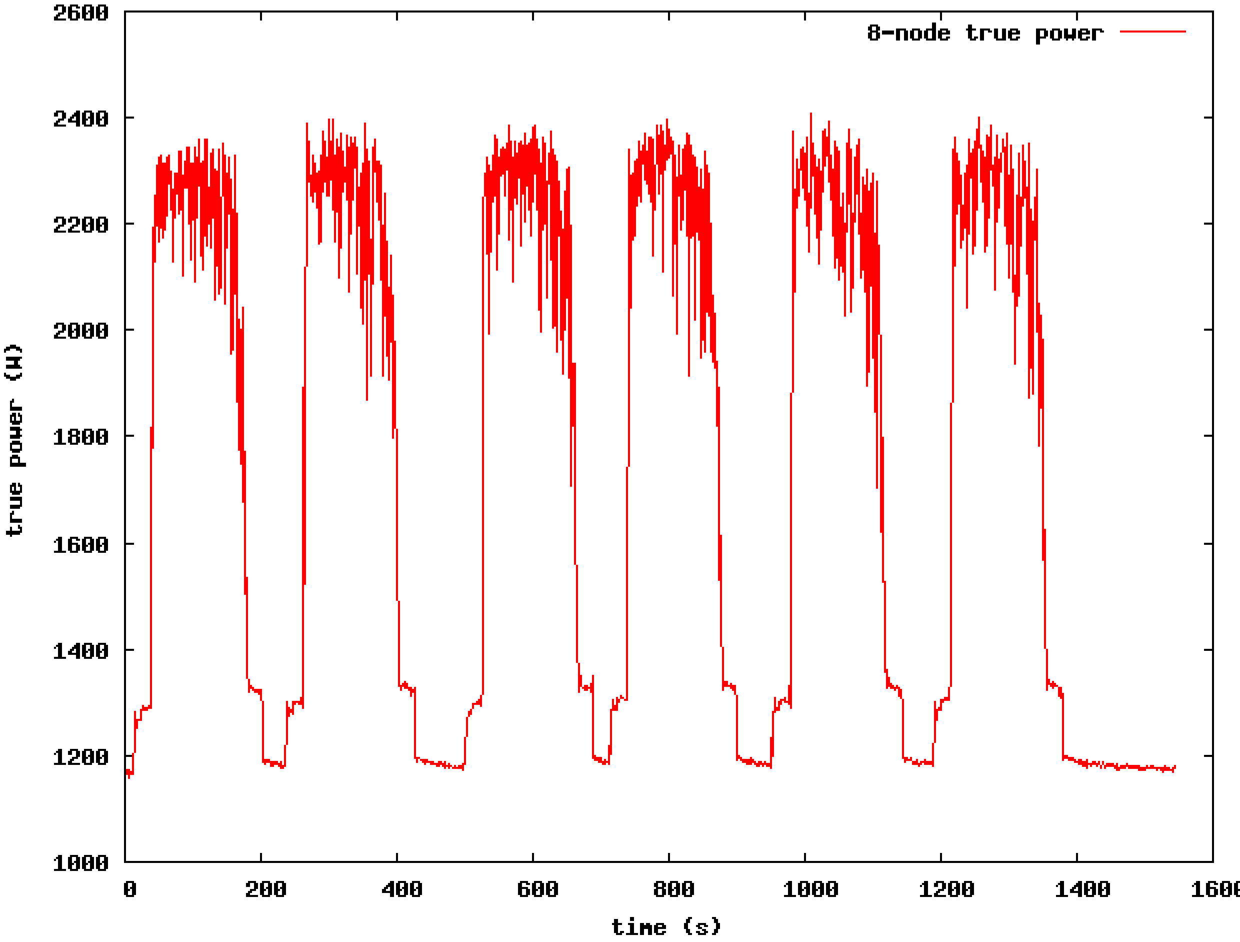
- *
- * Batch Log Began   11/02/10 at 14:16:51
- *
- * Data Type : 0x52 phase-phase
- * Data Period : 62500
- * Data Frames : 1545
- * Mon Period   : 1
- * FreqMode   : 2
- * Date Format : 1
- * Log Type   : 1
- * Software Version : 3.3R
- * Firmware Version : 2.a5
- * Hardware Version : 6.00
- * Serial Number   : 25663

Power Data File

- * Start In VAt Start W1 V12 W2 V23 W3 V31 Wt I1 VA1 I2 VA2 I3 VA3
- * Date Avg Avg Avg Time Avg Avg Avg Avg Avg Avg Avg Avg Avg Avg Avg Avg Avg
- 11/02/10 14:16:51 208.3 100.7 107.2 5.767 5.804 0.000 0.000 603.8 568.2 0.0 1172.0 620.5 584.8 0.0 1204.8
- 11/02/10 14:16:52 208.2 100.9 107.3 5.759 5.819 0.000 0.000 601.0 570.6 0.0 1171.2 617.8 587.5 0.0 1204.8
- 11/02/10 14:16:53 208.5 100.8 107.3 5.767 5.815 0.000 0.000 604.2 569.6 0.0 1173.6 621.0 586.4 0.0 1207.2
- 11/02/10 14:16:54 208.1 100.9 107.3 5.704 5.797 0.000 0.000 596.2 568.5 0.0 1164.0 611.6 585.3 0.0 1196.8

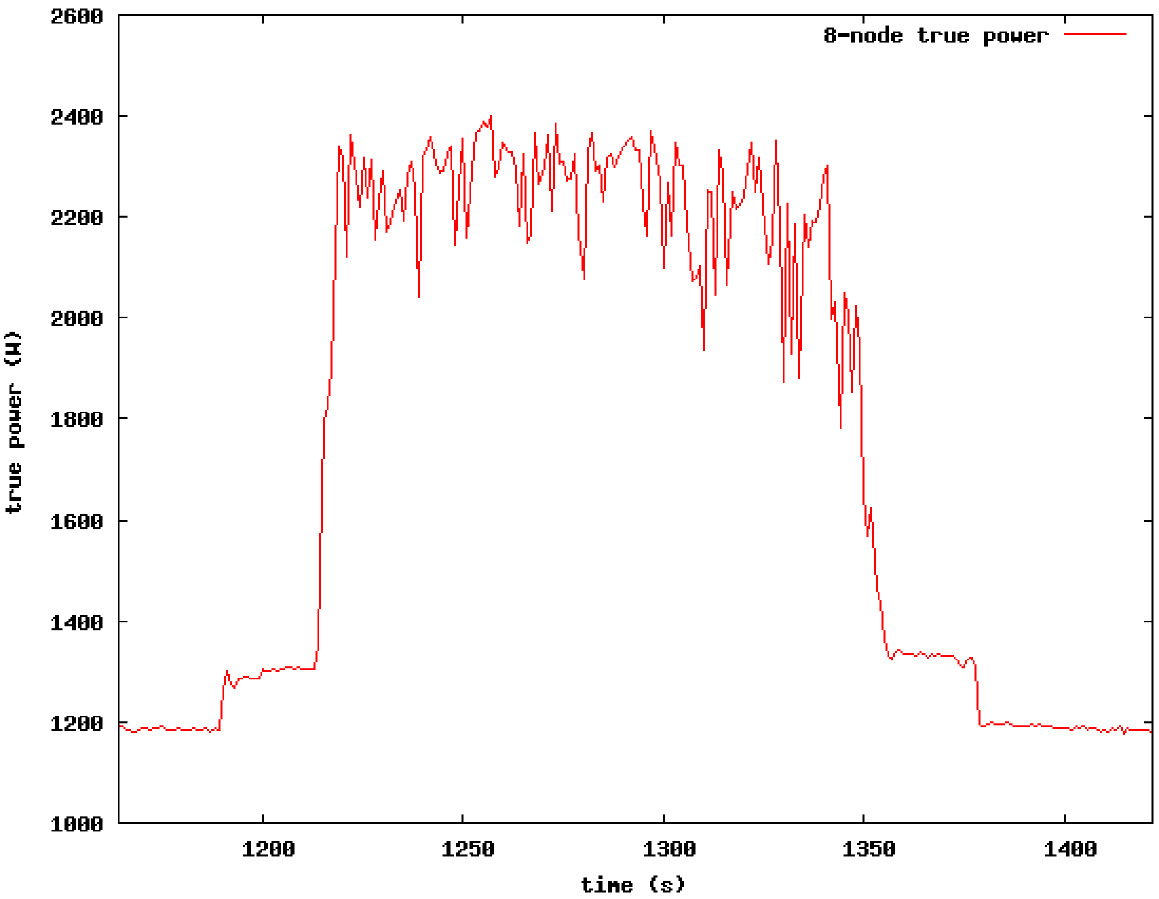
Overall Green500 Entry Test Period (6 HPL Runs)

- 6 HPL runs to get closest match to top500 run and allow for warm-up
- Last (#6) run closest to top500 submission speed



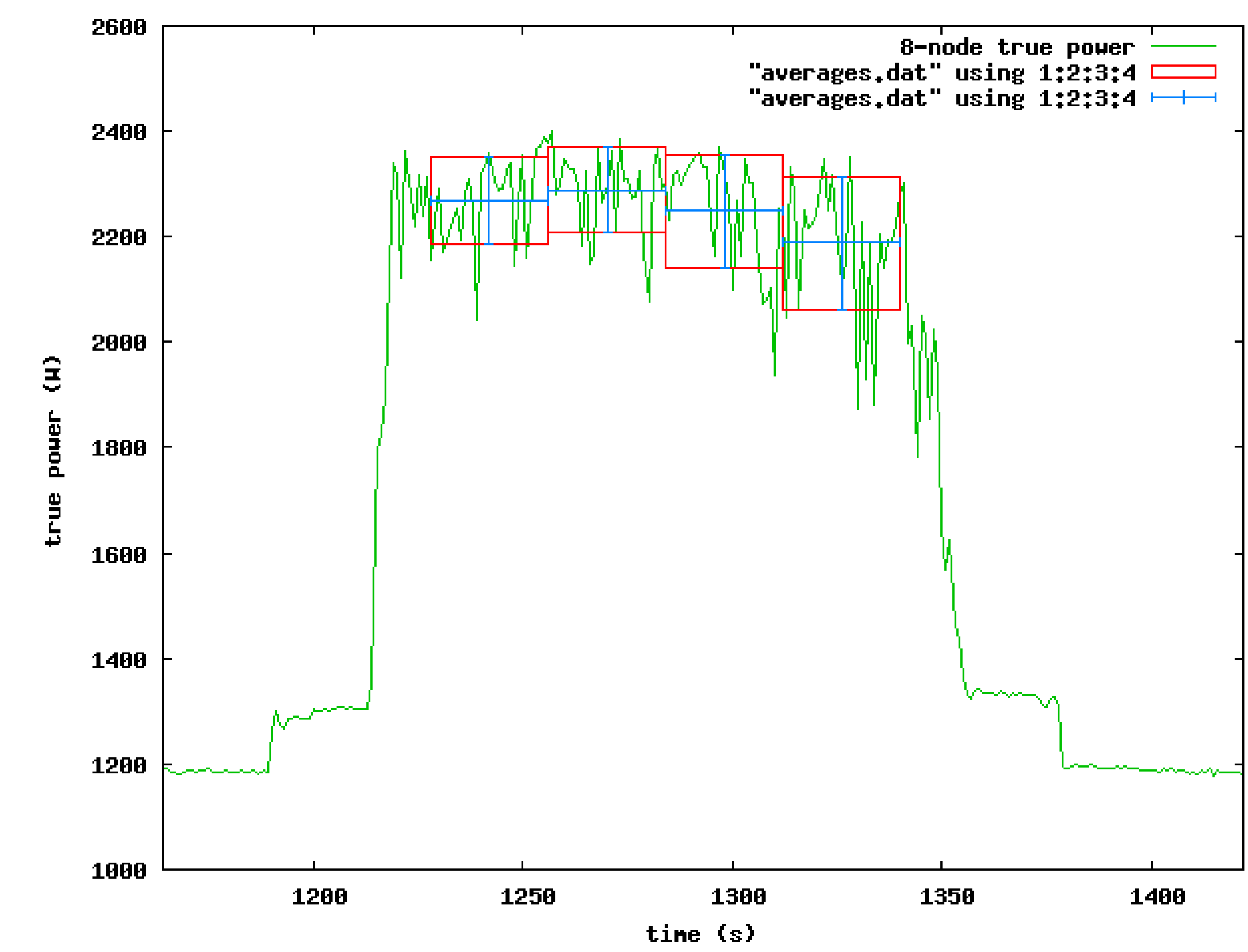
Power Graph for Measured Single HPL Run

- 2 shoulders: front porch for setup, back porch for answer validation
- Features:
  - Negative spikes
  - Power drops slightly over run



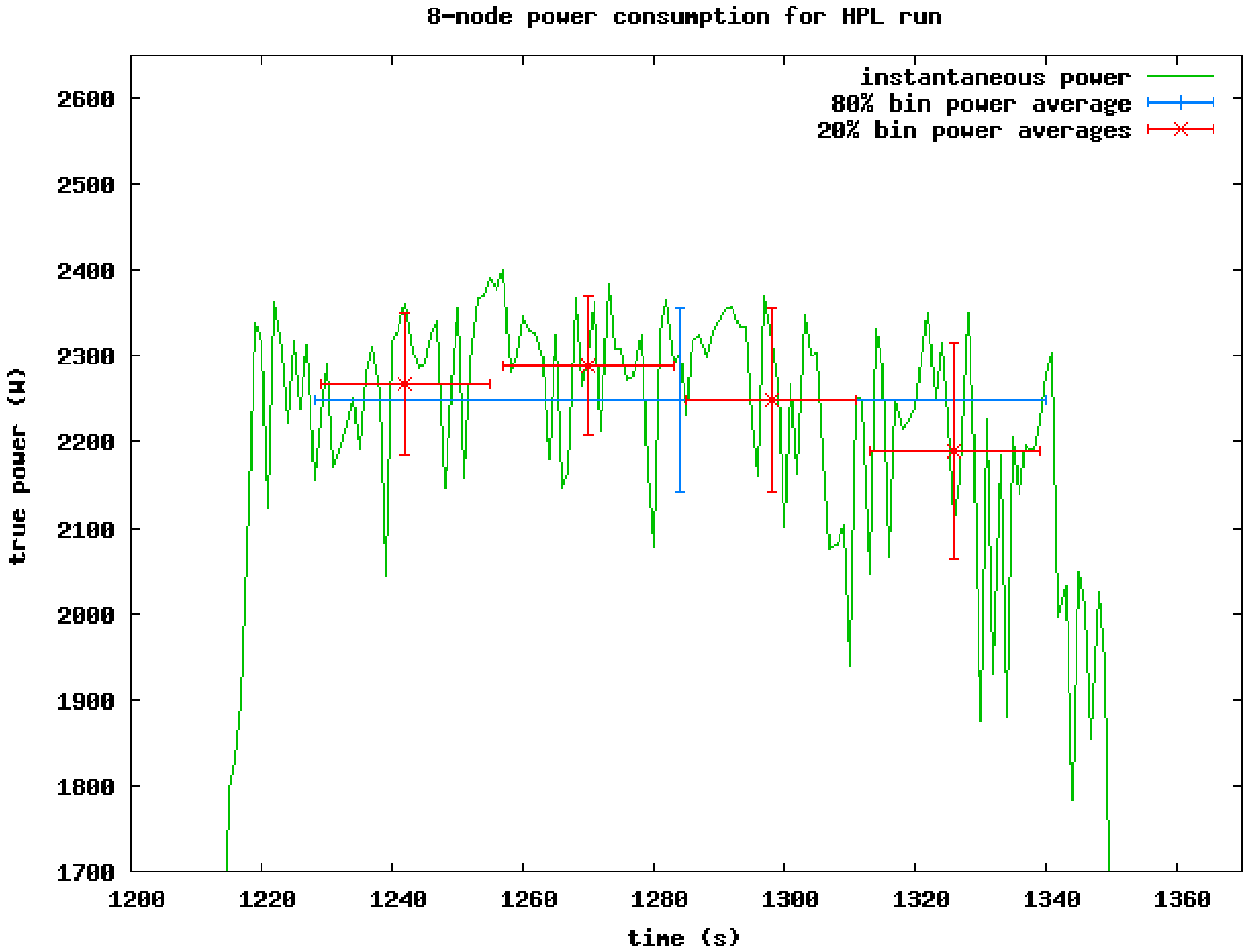
Average 8-node Power Draw In 20% Bins

- Spec for green500 is average power over 20% of run or more
- 4 20% bins in run middle: average 8-node power varies from 2289 W to 2189 W
- Power lowering is real physical effect; GPUS start to run out of computations to do



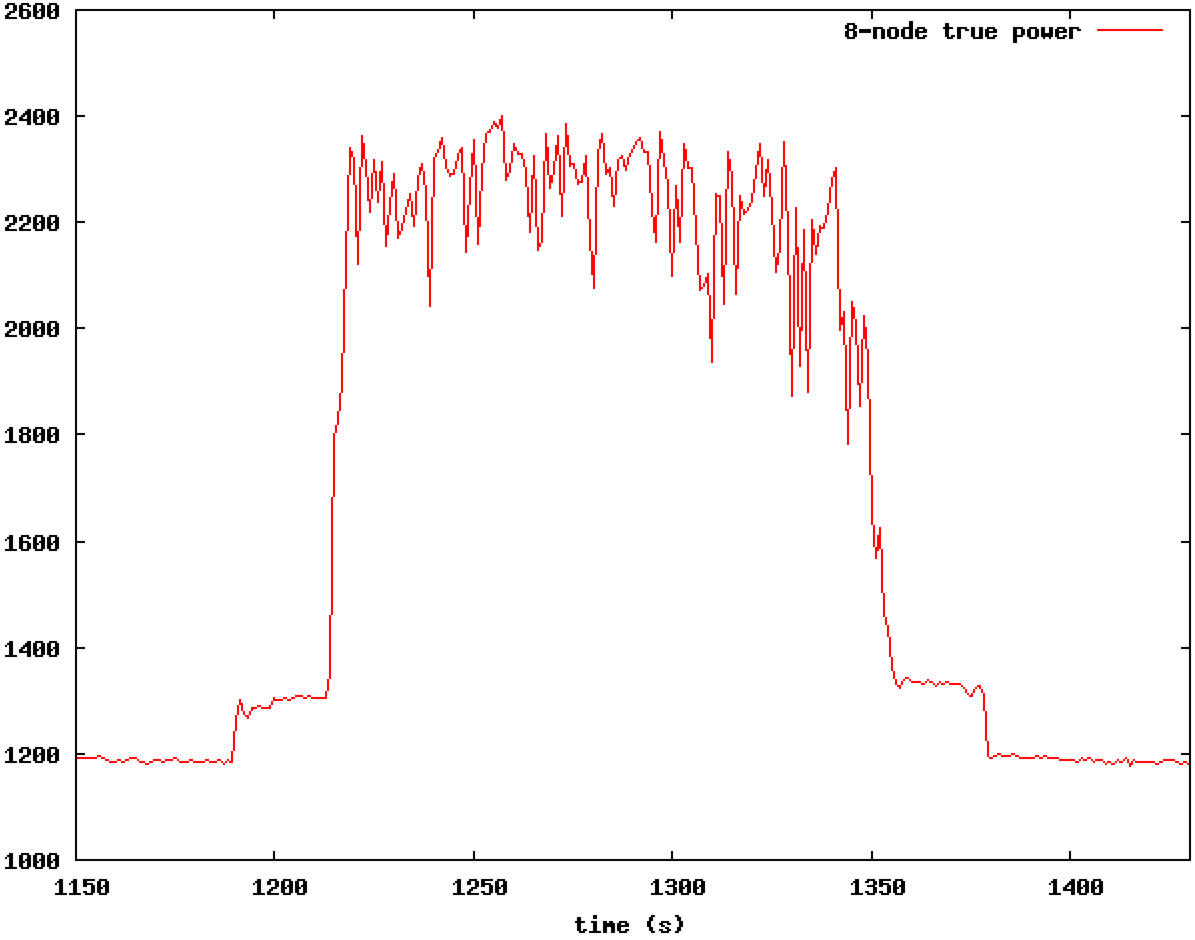
Final Average Power Calculation

- Average power calculated over 10%-90% range
- Calculated to be 2248W (8 nodes) = 35.97 kW for cluster



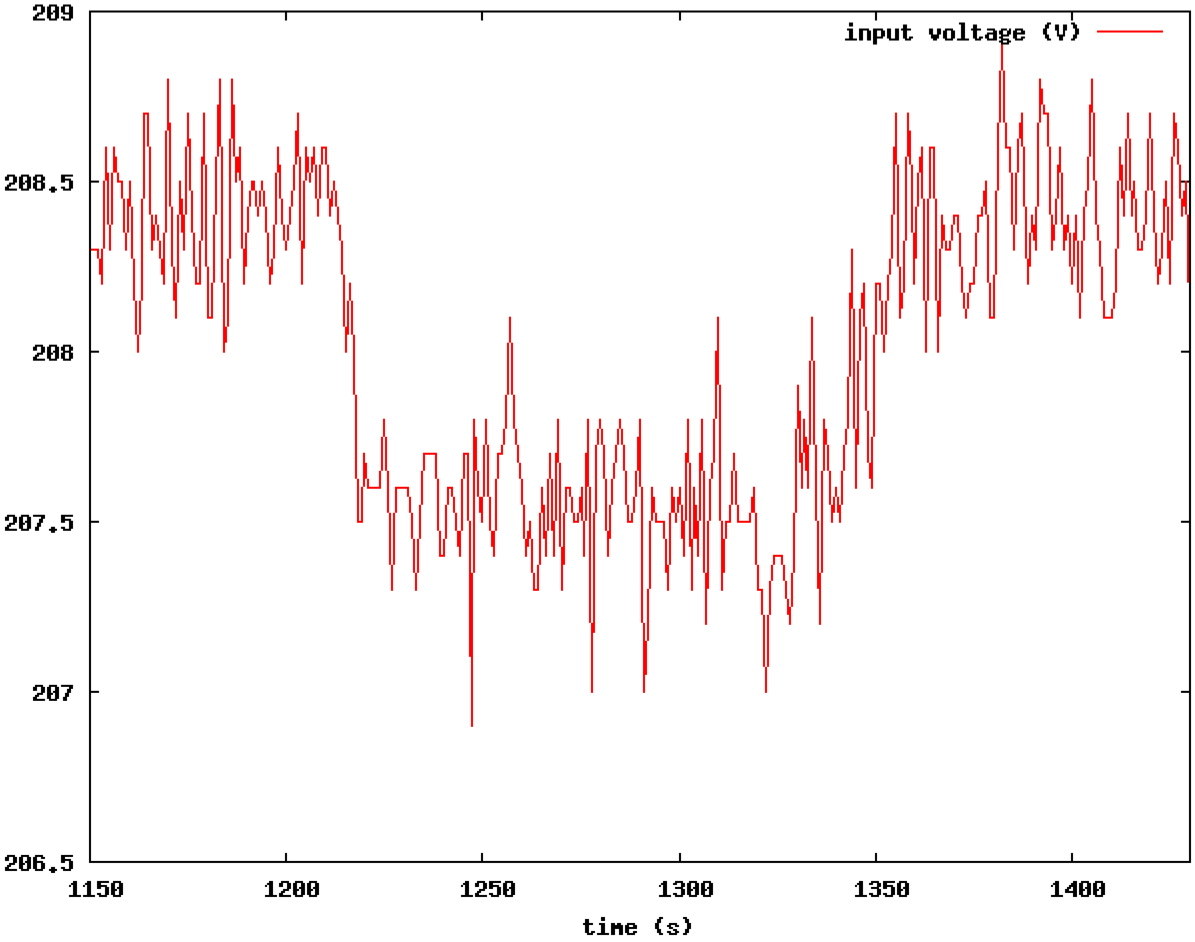
Power Draw for Voltage and Power Factor

- Expanded time range



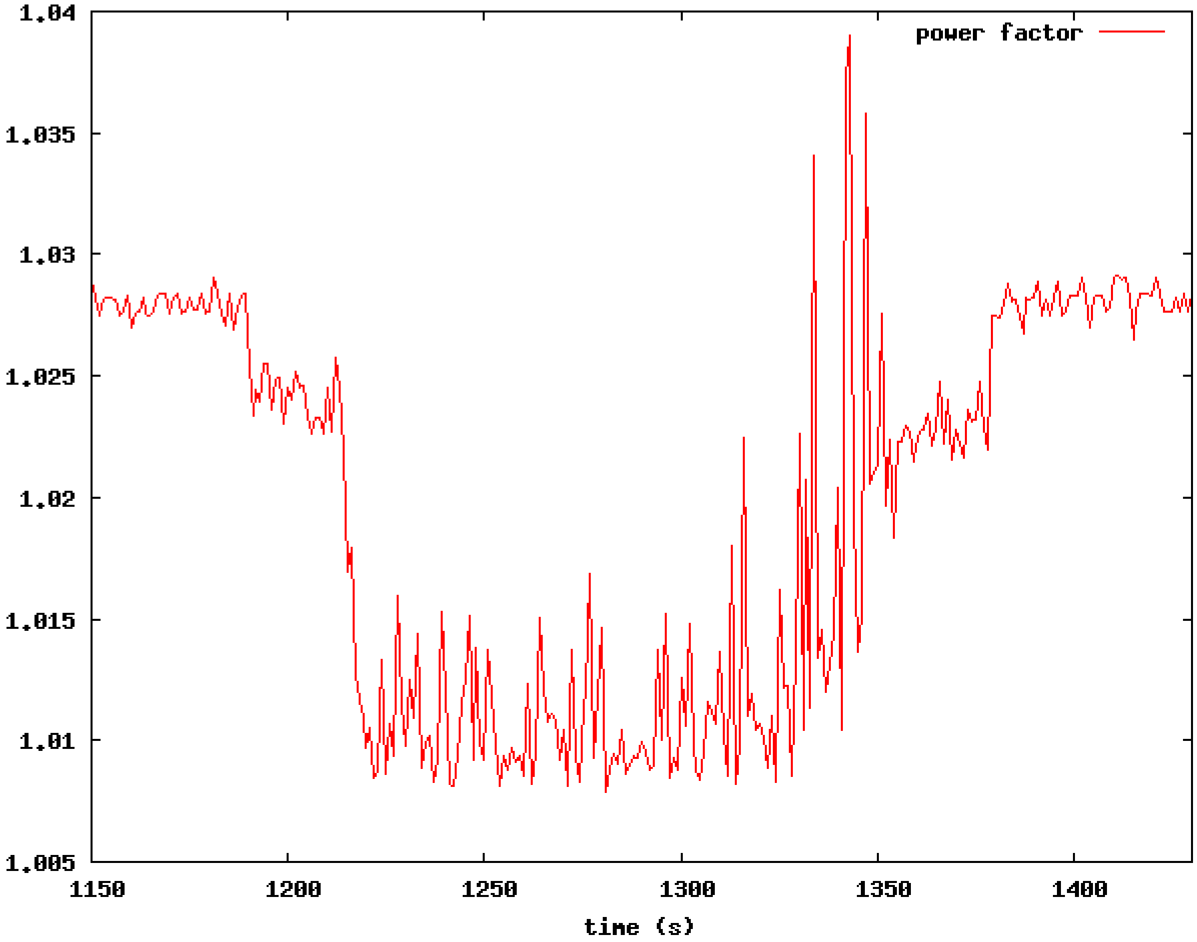
Input Voltage During HPL Runs

- Voltage drops but remains within spec
- Shown here for validation and as a sanity check
- Remains about 207.5 during HPL run



Power Factor

- Power factor remains below 1.035 for whole run including idle time
- Efficient power supplies, not overspecified



Current Questions and Next Steps

- What are the downward power spikes?
  - 1 second resolution *too coarse* to resolve cleanly
  - Need to use .2 second resolution current meter
- What are similar results with 1, 2, 4 nodes?
- How do the high-resolution timing results vary with application phase and input parameters? (Memory saturation tests have smooth graphs.)

More Information

- NCSA front page:

http://www.ncsa.illinois.edu