Cray User Group

May 2009

James H. Laros III
Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Motivation

- Average power consumption of a Top 9 system, 1.33 Mega-Watts (June 2008)
  - 1st time power is reflected on the list
- Average power consumption of a Top 9 system, 2.48 Mega-Watts (Nov 2008)
- 54% Increase in 6 months!
- Jaguar (ORNL) 6.95 Mega-Watts for 1.059 Peta-FLOPS
  - Projecting for 10 Peta-FLOPS 69.5 Mega-Watts
  - Seriously?
- Clearly we will be considering 10's of Mega-Watts for multi Peta-FLOP class systems
  - What about Exe-FLOPS?
  - What about cost (delivery infrastructure etc)?
  - What about cooling (power in power out)

Power Collection Methods

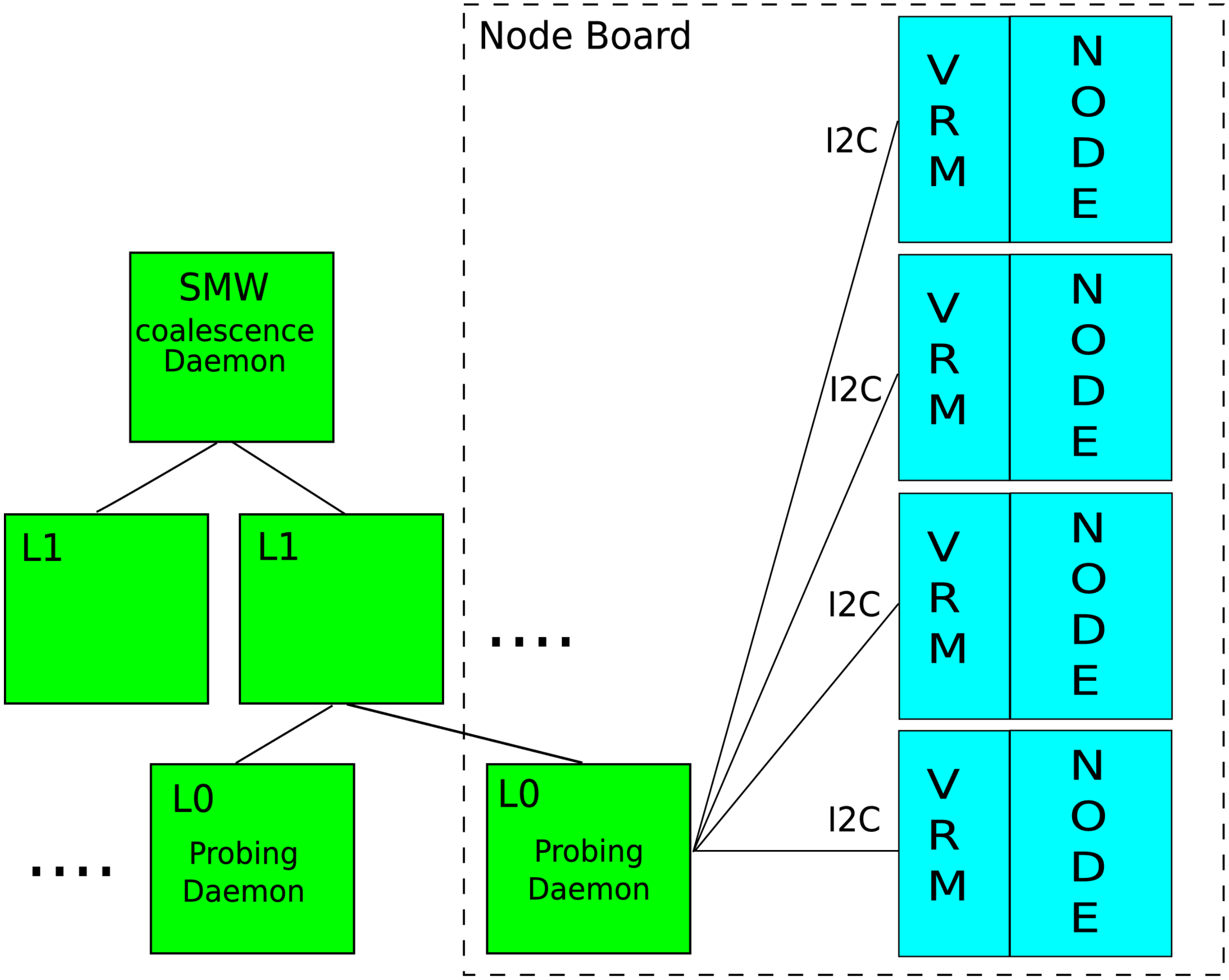
Past and Present

- Measured by Meter
  - Cabinet level
    - Coarse collection
    - Extrapolate to larger system estimate
  - Component level
    - Single components measured
    - Again, extrapolate to larger system estimate
- Performance Counters
  - Typically also used as basis for system level estimates
    - Should be verified
      - Can at an individual node scale but not at system scale

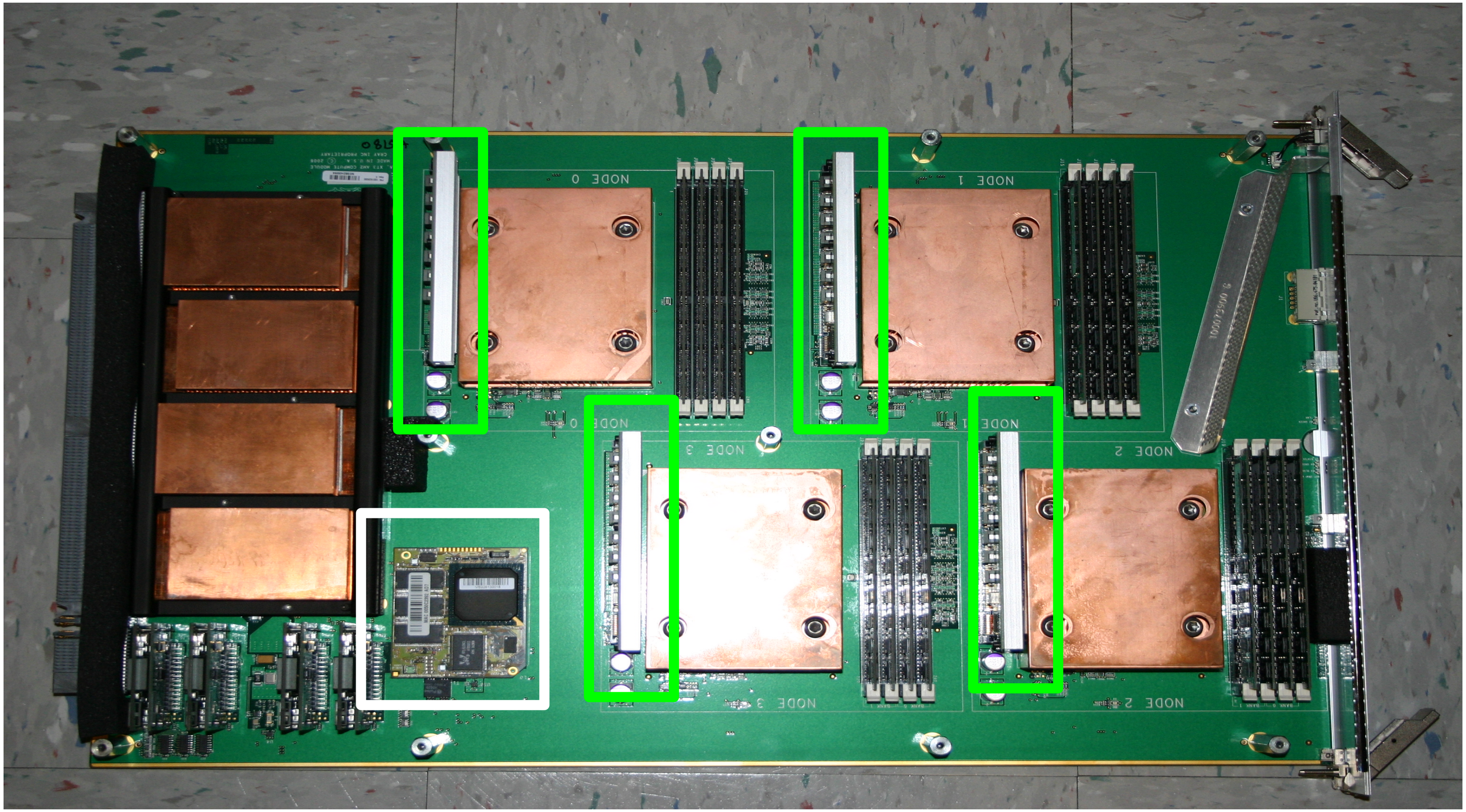
Real Power Collection

- Not currently a feature of CRMS but we can leverage the existing infrastructure (H/W and S/W)
- Additional daemon on each L0 (probing)
  - Registers a call-back in the main event loop
  - Uses event router to get information back up the hierarchy
- Additional daemon on SMW (coalescence)
  - Collects the events and writes them out to flat file
- Results
  - Granular collection (per-node - *socket*)
    - Also Mezzanine (Seastar) but flat line current draw
  - High Frequency (1-100 samples per second)
  - Can collect current and voltage measurements
  - Scalable

CRMS
Cray Reliability Availability and Serviceability Management System



XT4 Board



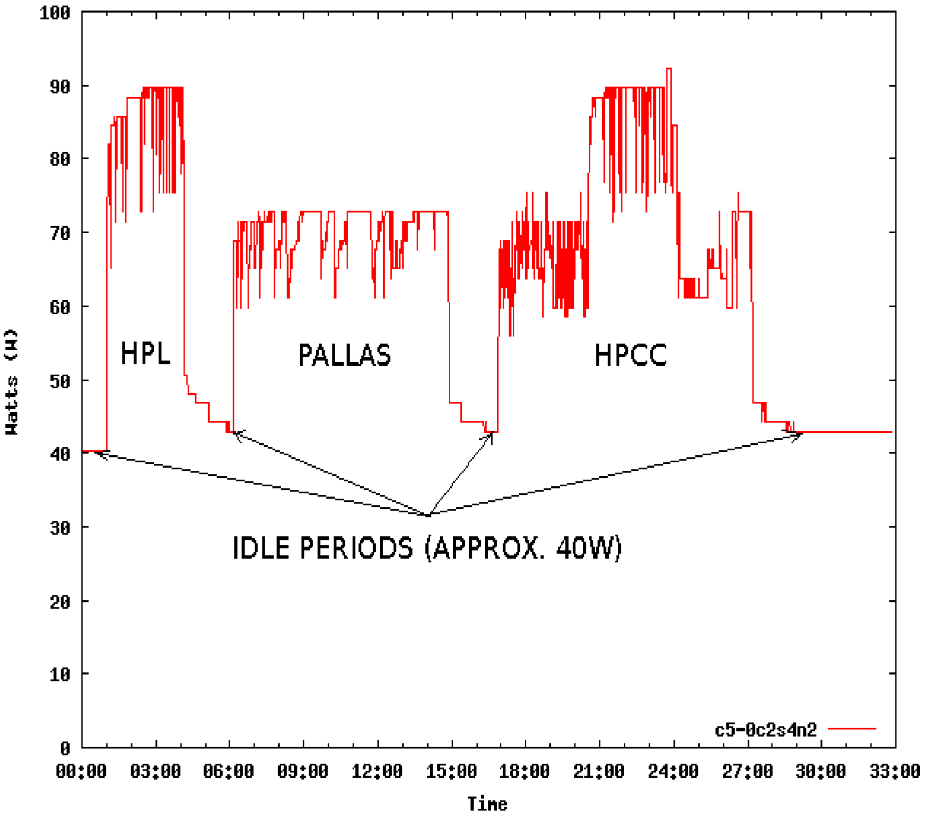
Real Power Collection
(continued)

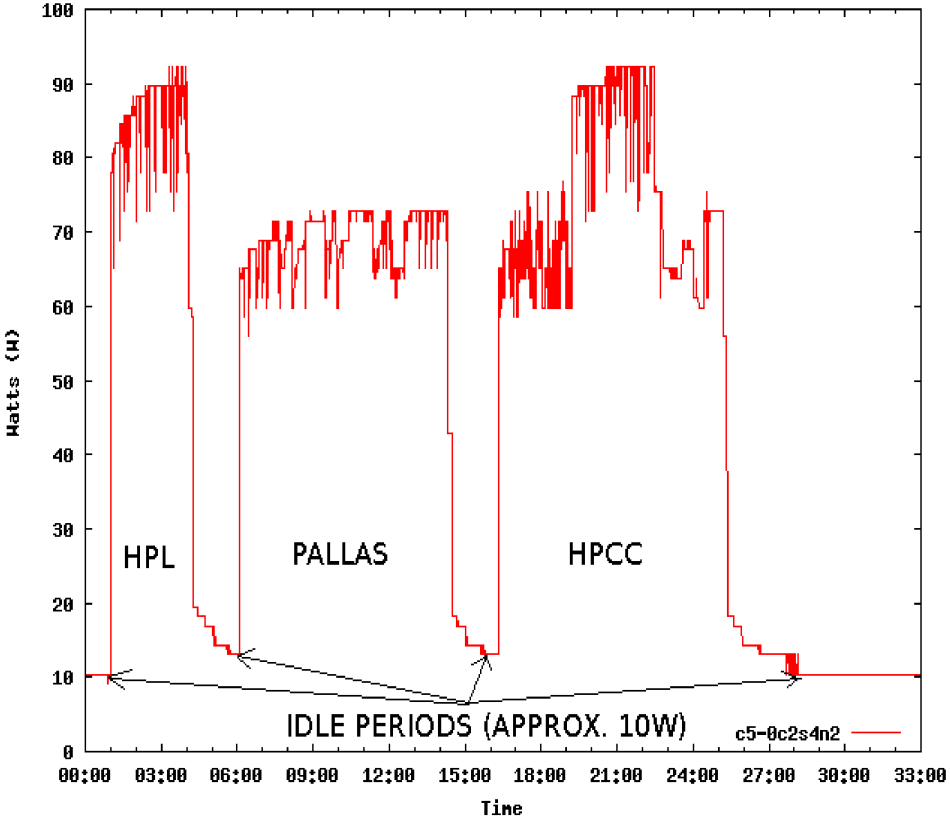
- Output
  - Timestamped Hex values for current
    - and optionally voltage
    - Current in amps +/- 2amp accuracy
- Post process output
  - Graphs (per node, per board)
  - Calculate application energy
    - More later
  - Ultimately, sum energy per job
    - Real time stats?
    - Better integration, output to DB...

Now that we have it what do we do with it?

- Catamount Idle
  - We “thought” it was inefficient
    - Now we know it was
- Linux employs power saving during idle cycles
  - Use for a benchmark to measure our success
- Modified Catamount
  - Relatively straight forward (for OS code :)
  - Only two areas kernel enters during idle
- Contrasted with CNL
  - Discovered our modifications are effective
  - Discovered Linux didn't act as we thought?

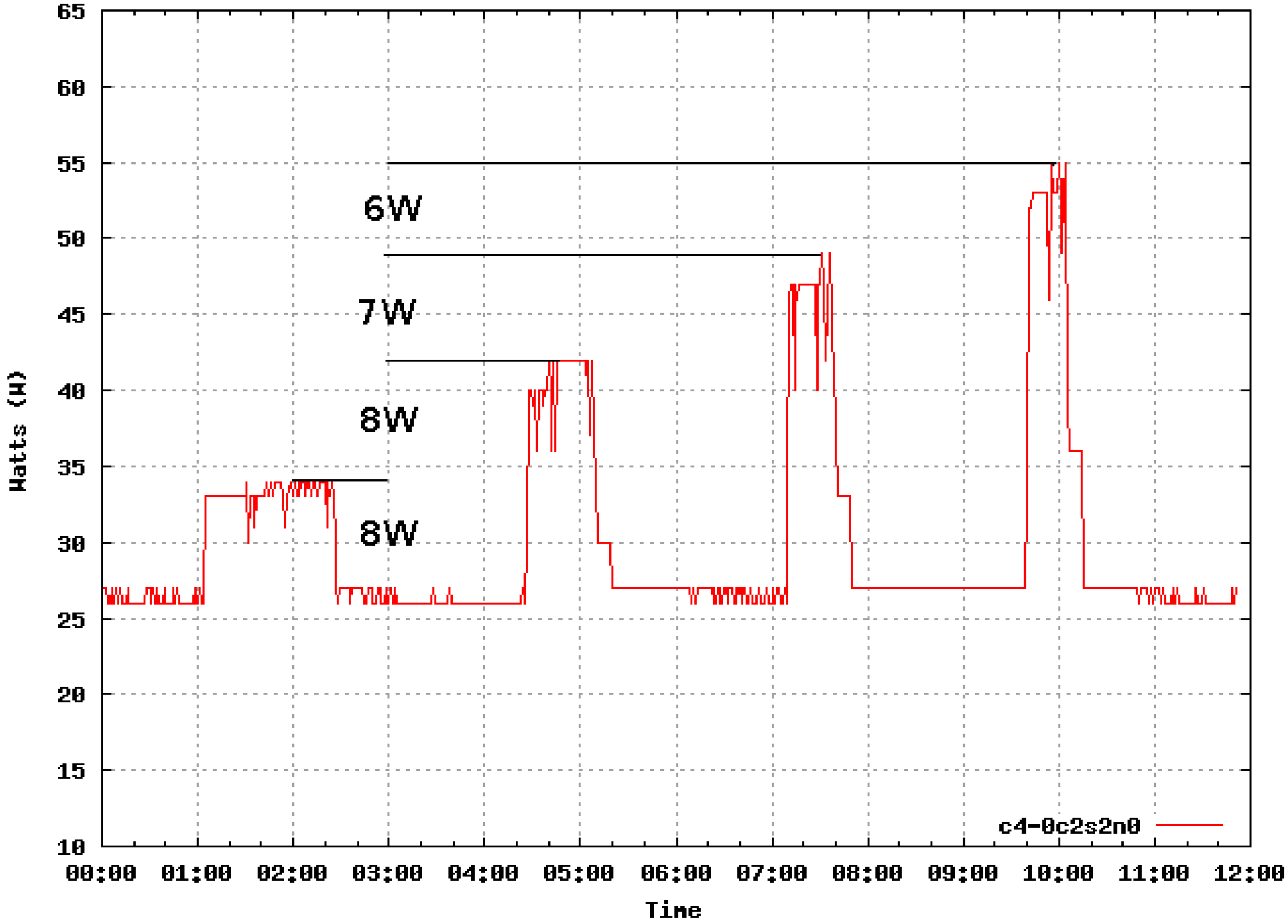
Initial CNL and Catamount IDLE Draw





Sandia National Laboratories

Halt Individual Cores

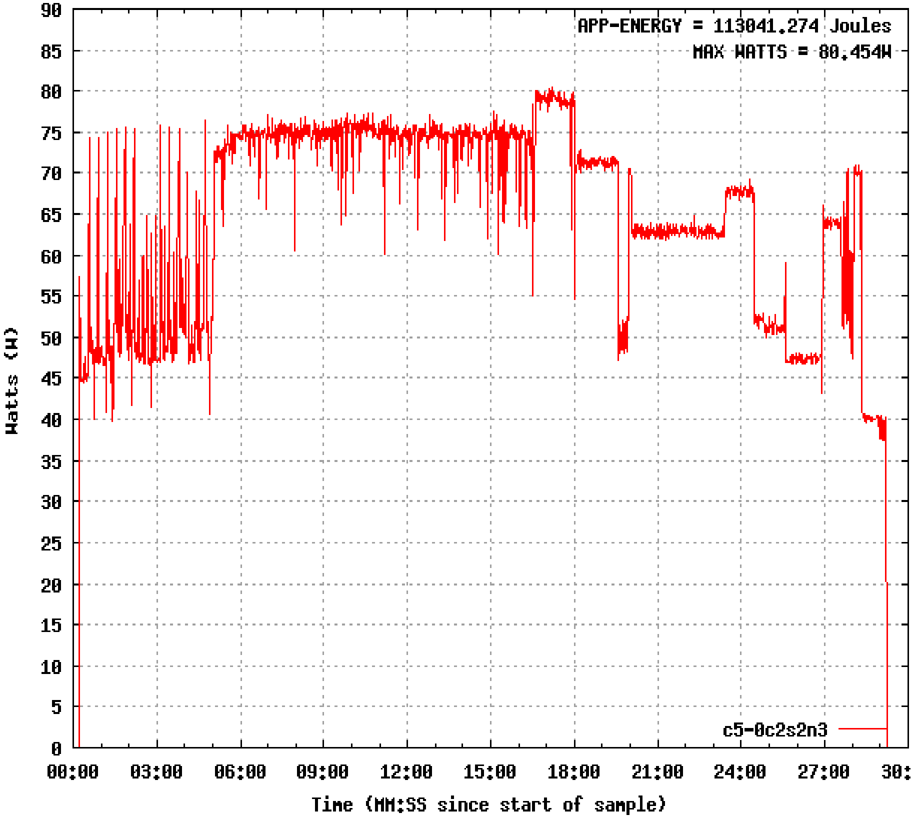


Application Signatures

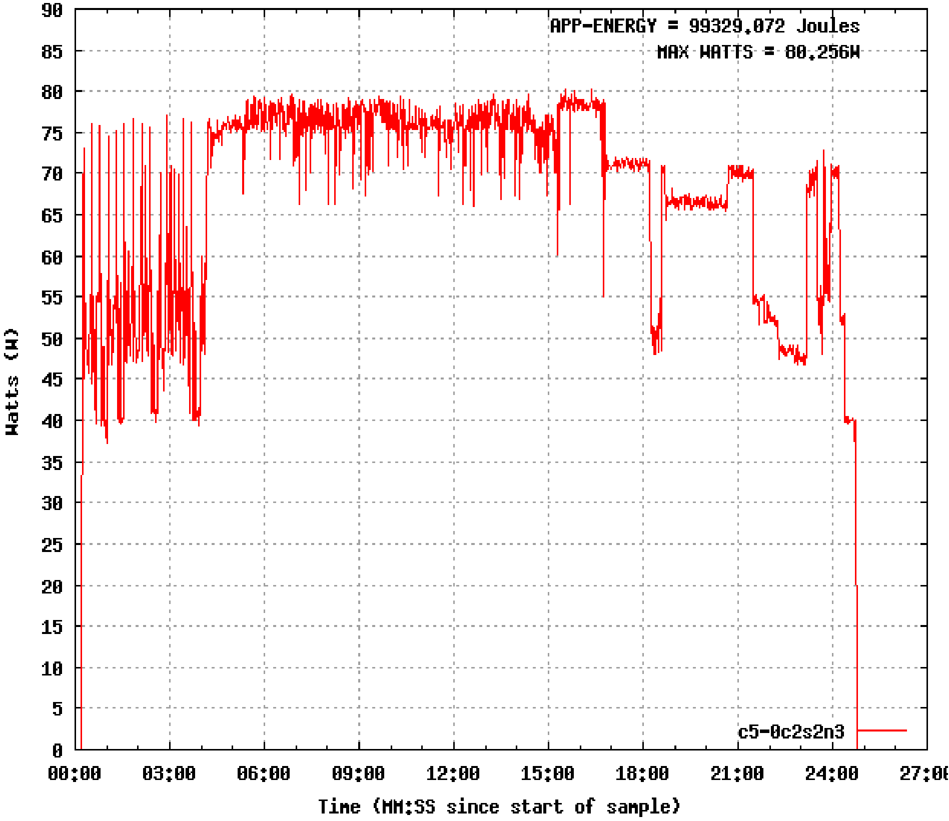
- Noticed graphs of each application has its own, repeatable, recognizable shape
  - Even when run on different OS
- Can we learn anything?
  - Can this be used for debugging?
  - Performance tuning?
- We can calculate application energy
  - Amount of energy used over duration of application
  - Sure, find area under the curve
- We now have “real” power used by applications
  - Use as an additional metric
  - Feed into power aware scheduling

Application Energy

CNL



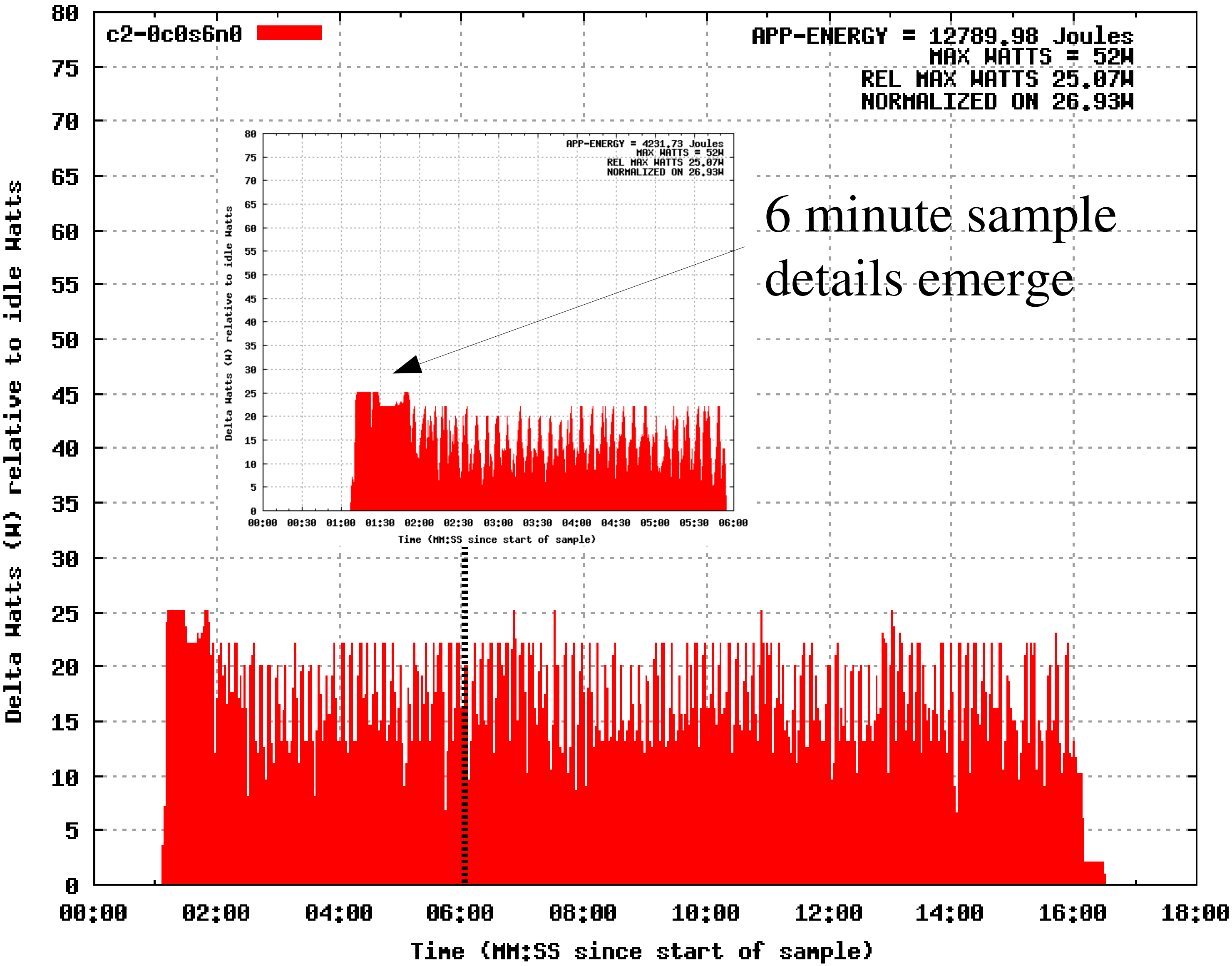
Catamount



Application Energy

- **HPCC**
  - **16% Faster on Catamount**
  - **13% Less energy on Catamount**
- **Obvious but important, longer run time = more energy used**
- **Performance can have other benefits**
- **How do other things that affect performance affect power use?**

Closer examination



6 minute sample
details emerge

Future Work

- Quantify in dollars
- Impact of OS noise on Power
  - We know OS noise can impact performance
  - What is the associated impact on power efficiency?
- Does network imbalance impact Power?
  - Less bandwidth?
  - Higher latency?
- Can we save power when running applications?
  - Go into lower power state while waiting...
- Reduce frequency runs without affecting performance?
  - Little to no impact on run-time, large power savings?

Acknowledgments

- Other Contributors
  - Kevin Pedretti
  - Sue Kelly
  - John Vandyke
  - Courtenay Vaughan
  - Mark Swan (Cray)
- Local Administration Staff

Questions?