SAND2012-1271C

Measuring and Tuning Energy Efficiency on Large Scale High Performance Computing Platforms

James H. Laros III

Sandia National Laboratories

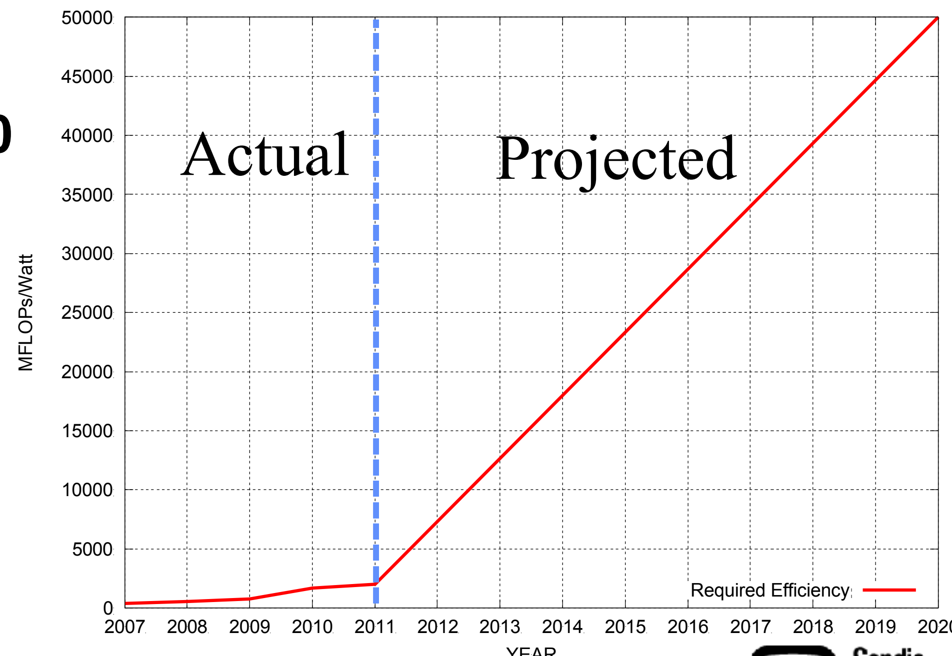
Sandia National Laboratories

Overview

- Motivation – Why?
- Related Research
- Impact at Scale
- Test Platforms
- Measurement and Data Analysis
- Experimental Approach and Results
  - Experiments #1, #2 and #3
- Overall Observations
- Acknowledgments
- Questions

Why?

- **Power efficiency 1^{st} class challenge for Exascale**
- **2011 – Most efficient = 2026 MFLOPs/Watt**
- **Based on this, ExaFLOP requires 494 MW**
- **Target of 20MW requires 50,000 MFLOPs/Watt efficiency**
- **≈25x Increase in efficiency in 9 years**
- **We have seen ≈6x in the past 4**
- **Hardware *might* need help**



Sandia National Laboratories

Related Research

- **Most related**
  - Ge, Feng, Song, Cameron, et al.
  - Virginia Tech – PowerPack – DVS scheduling
  - Component level = Yes
  - Scale = no
- **Simulation and model based research**
  - Microarchitecture level
    - Vendors, labs, academia and various collaborations
  - Structural Simulation Toolkit (SST) at Sandia for example
- **Profile based research**
  - MPI profiles, log profiles, counters
  - Some attempt validation with direct measurement
    - Instrumented single node
- **Coarse level measurements**
  - PDU's
  - External Power meters

Impact at Scale

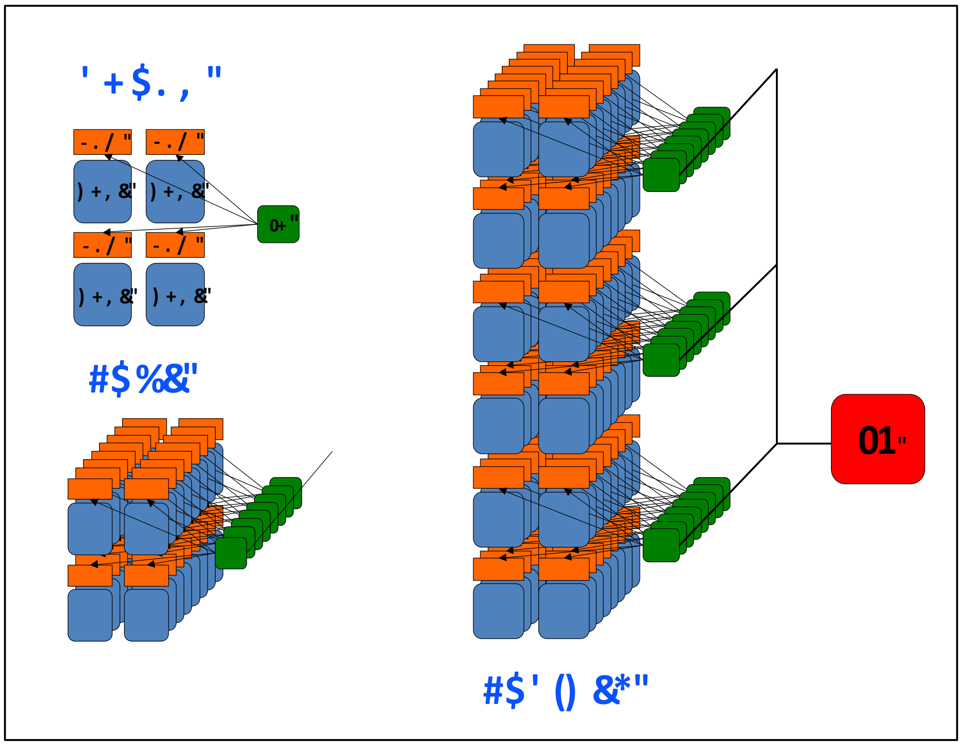
- Unique ability to measure in-situ at large scale
  - Allows application analysis at large scale
- Focus on <u>REAL</u> scientific applications
- Focus on <u>LARGE</u> scale
- Impacting next generation platforms
  - How they will be built
  - How they will be used

Test Platforms

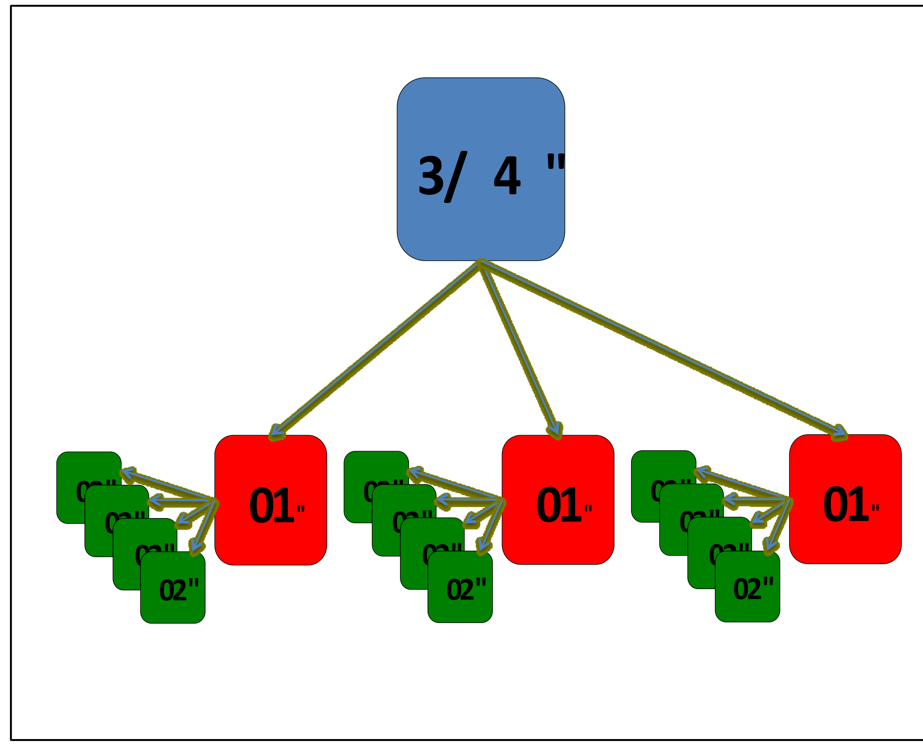
- Sandia National Laboratories
  - Red Storm – 1st Cray XT platform
  - 3,360 Dual Core AMD 64 bit 2.4 GHz nodes – 4GB Memory
  - 6,240 Quad Core AMD 64 bit 2.2 GHz nodes – 8GB Memory
- Oak Ridge National Laboratory
  - Jaguar
  - 7,832 Quad Core AMD 63 bit 2.2 GHz nodes – 8GB Memory
- All
  - Seastar Interconnect
  - 2GB/core
  - Catamount Light-weight Kernel (LWK) Operating System
  - Reliability Availability and Serviceability System (RAS)

In-situ Measurement

- Instrument existing RAS system

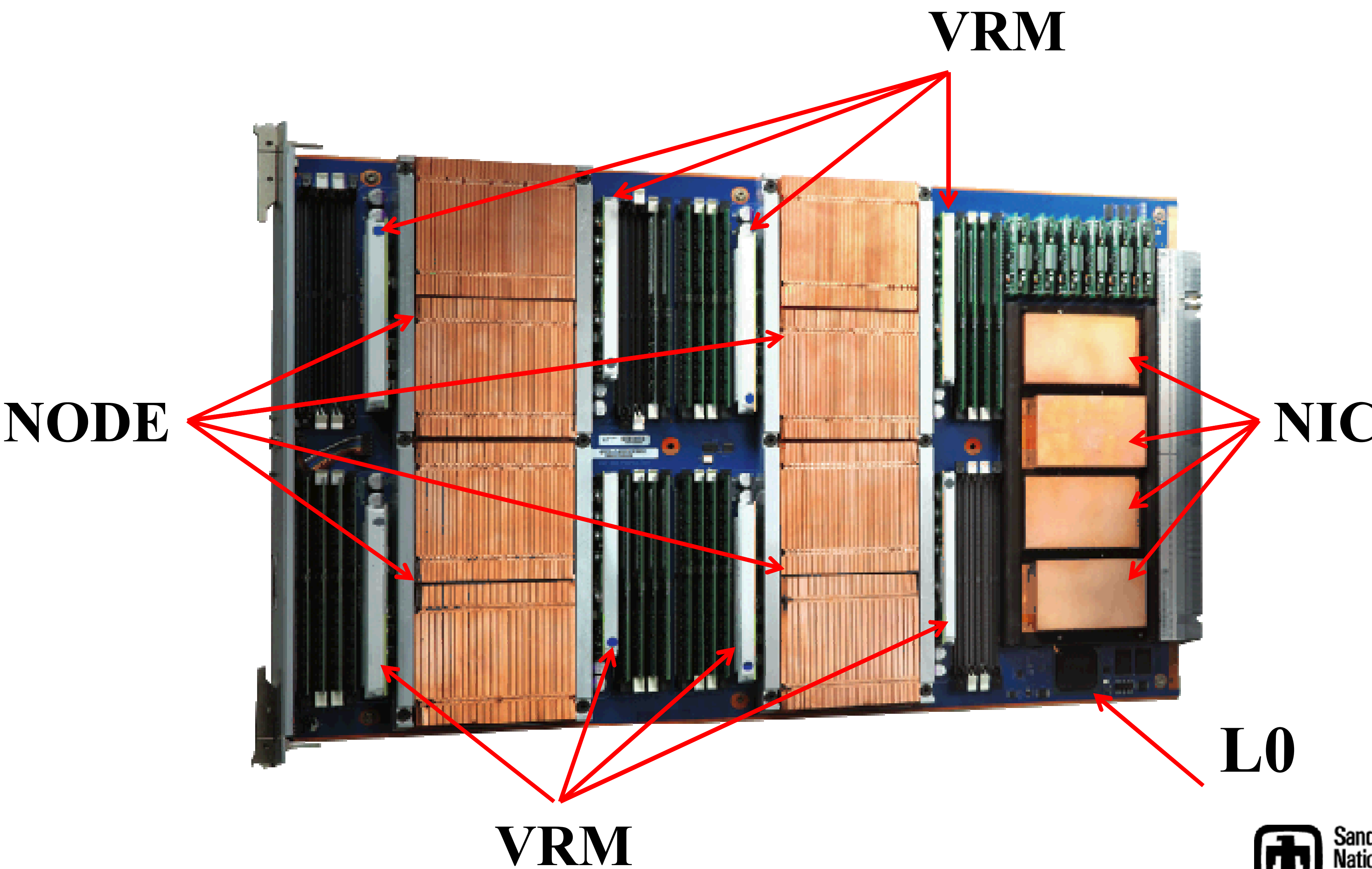


- Leverage existing H/W sensors



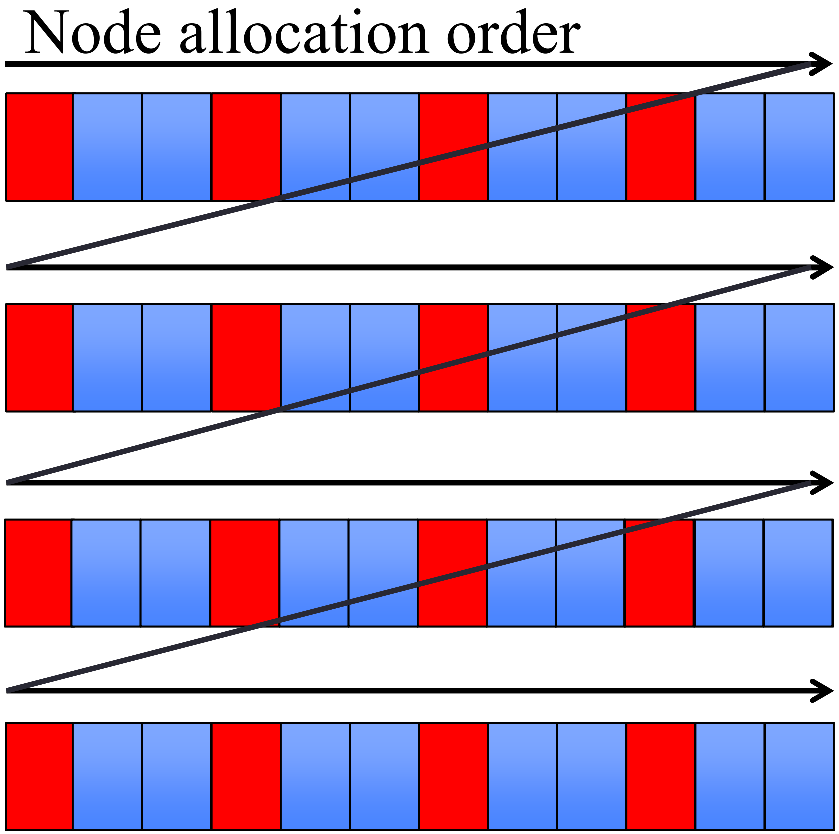
Result: Scalable, in-situ, high-frequency, component level current and voltage measurement

In-situ Measurement



Data Analysis

- 1 sample per second per node
  - Current and Voltage
- Aligned with application execution
- Statistical analysis
  - Median
  - Mean
  - Mode
  - Coefficient of Variation
    - Independent of magnitude
- Per node graphs created
- All done with post processing code



Sandia National Laboratories

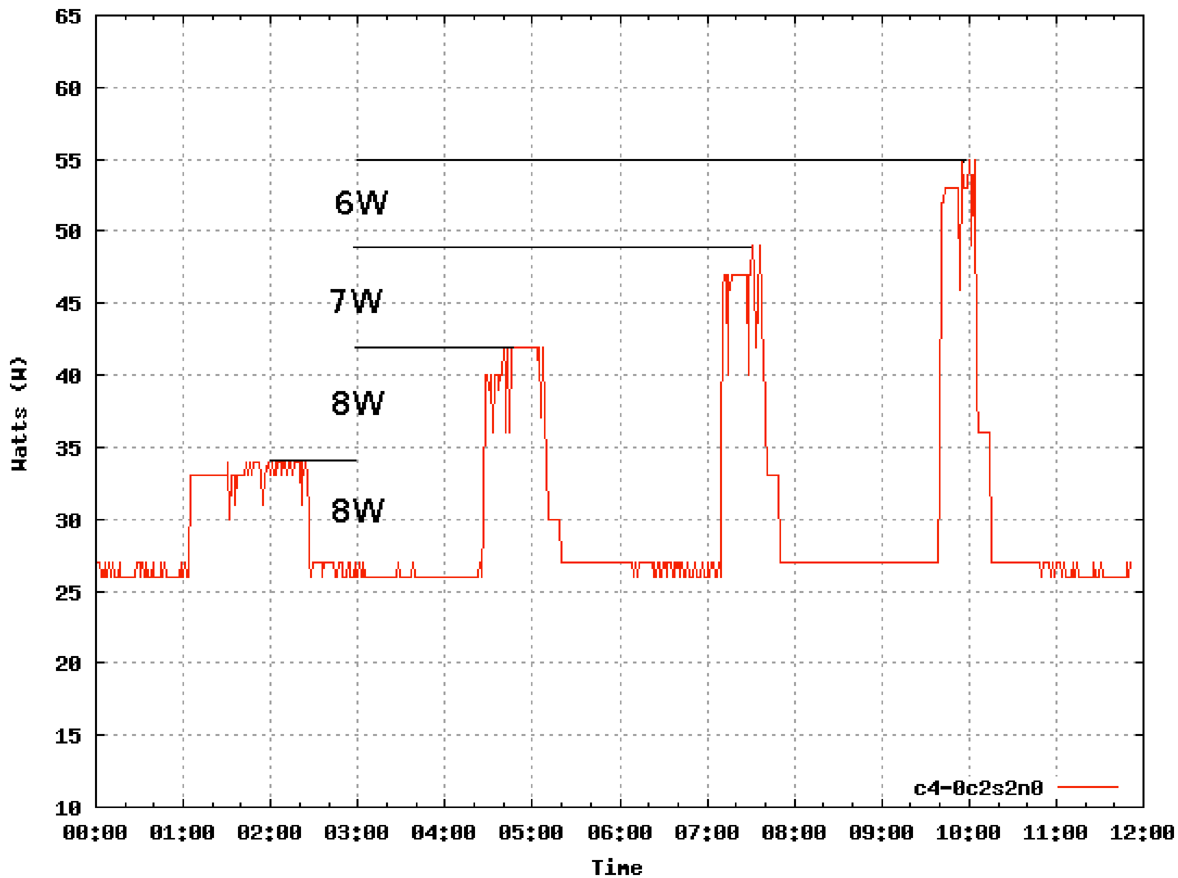
Experiments

- **Experiment #1**
  - Proof of concept
  - Affecting Power During Idle Cycles
- **Experiment #2**
  - Tuning CPU Power During Application Run-time
- **Experiment #3**
  - Network Bandwidth Tuning During Application Run-time

Experiment #1: Affecting Power During Idle Cycles

- Design of Catamount LWK preceded Advanced Power Management features
  - Focus entirely on performance
  - Suspected waste of power during idle cycles
    - Tight busy idle loop
- Targeted modifications
  - Put slave cores in halt when not in use
  - Put master core in halt
  - C and inline assembly
  - Stability sensitive timing considerations
  - Research evolved into production
- Questions:
  - Can we observe the effect of our changes?
  - Can we equal Linux idle characteristics?

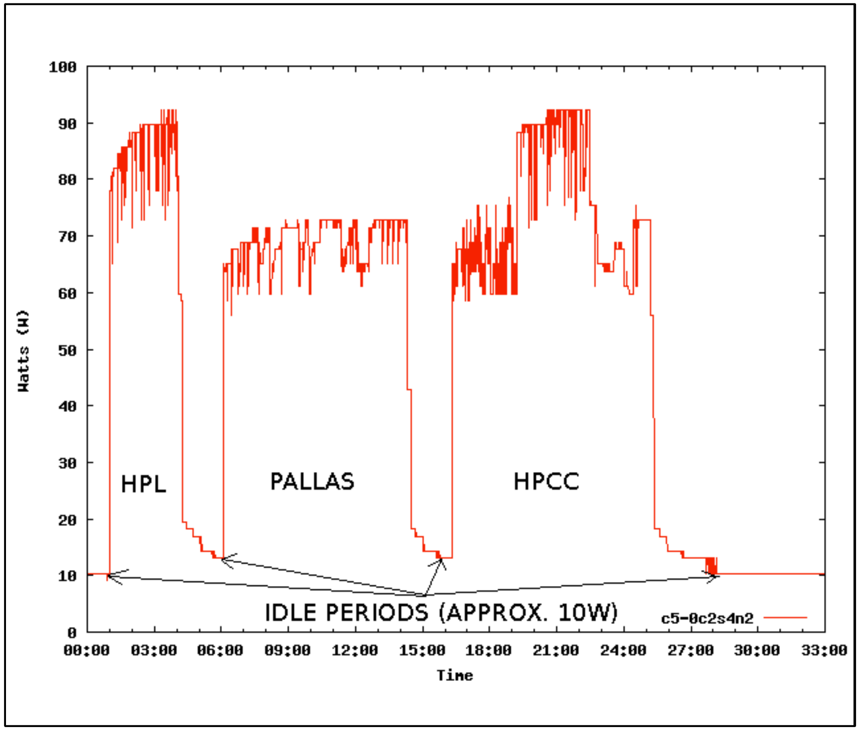
Experiment #1: Results



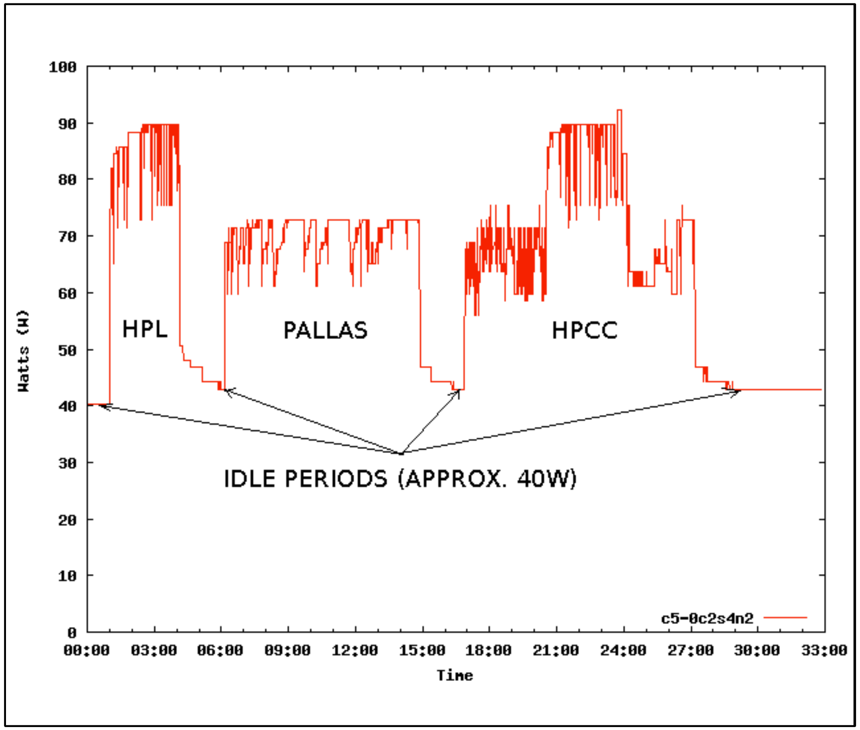
Catamount LWK on QUAD core:
Verification of cores in halt during idle

Sandia National Laboratories

Experiment #1: Results



Catamount LWK after idle modifications



Production Compute Node Linux (CNL)

Sandia National Laboratories

Experiment #1: Results

- Measurement capability characterized
- Initial operating system modifications successful
  - more importantly observed!
- Discovered ability to analyze running applications
- Opened door to further research

Some Accomplishments:

- ≈ 1 million dollars in energy costs since implemented
- DOE/NNSA Environmental Stewardship Award
- DOE/NNSA Defense Programs Award of Excellence
- List Paper?

Experiment #2:
Tuning CPU Power During Application Run-time

- Save energy during application run-time?
- Assumed we would have to *dynamically* tune frequency
- Targeted modifications
  - OS trap to deterministically change P-states
  - User space library to request changes
  - MPI profile layer to intercept potential wait periods
- While testing modifications discovered static tuning had significant impact
  - More stable
  - Easily coordinated
- Experiment #2 based on static tuning
- CPU energy contrasted
  - CPU accounts for 44-57% of total node energy
  - Single largest single component contributor
  - CPU analysis most useful to contrast with other platforms

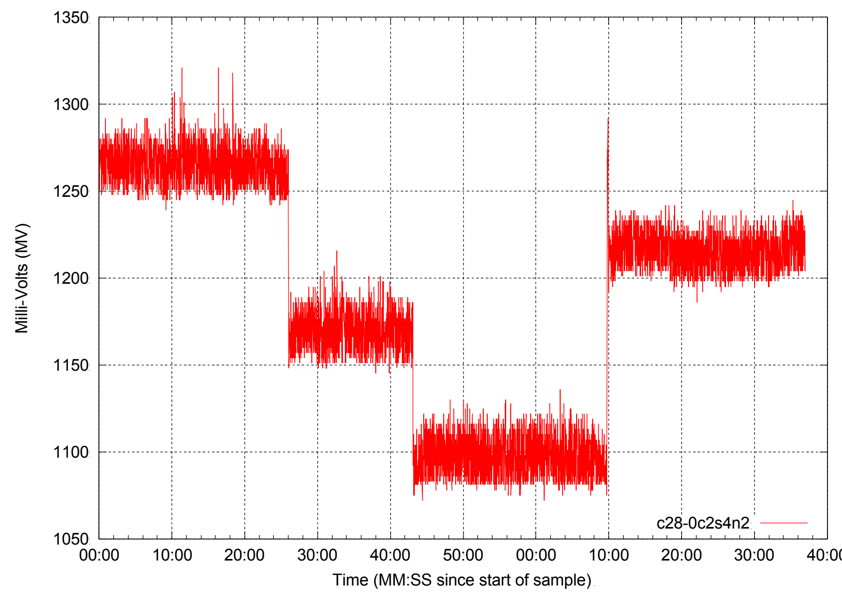
Why Tune CPU Frequency?

- Voltage: quadratically related to Power

$$P = ACV^2 f + AVI_{short} f + Vi_{leak}$$

P-state	CPU FREQ Red Storm	CPU FREQ Jaguar	Input Volt. Red Storm	Input Volt. Jaguar
0	2.2 GHz	2.1 GHz	1.200 V	1.200 V
1	2.0 GHz	2.1 GHz	1.200 V	1.200 V
2	1.7 GHz	1.7 GHz	1.150 V	1.150 V
3	1.4 GHz	1.4 Ghz	1.075 V	1.075 V
4	1.1 GHz	1.1 GHz	1.050 V	1.050 V

P-states, Frequencies and Voltages for Test Platforms



Observed Drop in Voltage During P-state Transitions

Experiment #2: Results

Applications:

- 6 Real scientific High Performance Computing (HPC) applications run at large scale
  - AMG2006, LAMMPS, SAGE, CTH, xNOBEL, UMT, and Charon
  - From 1-24K cores
- Two common benchmark applications
  - Linpack – compute intensive
  - Pallas – communication intensive

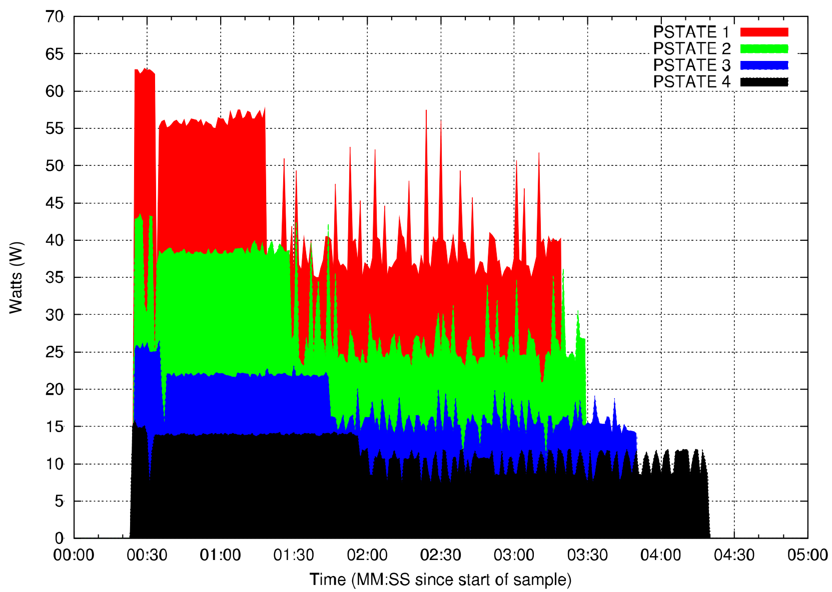
Sandia National Laboratories

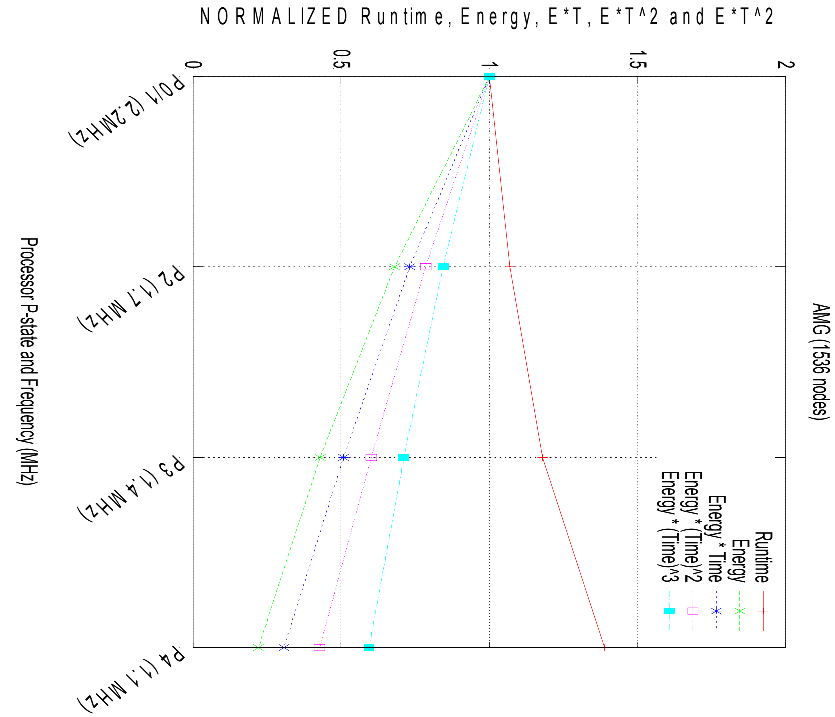
Experiment #2: Results

	Nodes/Cores	P2 Run-time %Diff	P2 Energy %Diff	P2 Run-time %Diff	P3 Energy %Diff	P4 Run-time %Diff	P4 Energy %Diff
HPL	6000/24000	↑ 21.1	↓ 26.4				
Pallas	1024/1024	↑ 2.30	↓ 43.6				
AMG2006	1536/6144	↑ 7.47	↓ 32.0	↑ 18.4	↓ 57.1	↑ 39.1	↓ 78.0
LAMMPS	4096/16384	↑ 16.3	↓ 22.9	↑ 36.0	↓ 48.4	↑ 69.8	↓ 72.2
SAGE	4096/16384	↑ 0.402	↓ 39.5				
SAGE	1024/4096	↑ 3.86	↓ 38.9	↑ 7.72	↓ 49.9		
CTH	4096/16384	↑ 14.4	↓ 28.2	↑ 29.0	↓ 38.9		
xNOBEL	1536/6144	↑ 6.09	↓ 35.5	↑ 11.8	↓ 50.3		
UMT	4096/16384	↑ 18.0	↓ 26.5				
Charon	1024/4096	↑ 19.1	↓ 27.8				

Experiment #2: Additional Analysis

AMG2006: 6144 cores
Pstates 1-4





Analyze Application Signatures

Unified Metric:
Energy Delay Product

Sandia National Laboratories

Experiment #3: Network Bandwidth Tuning During Application Run-time

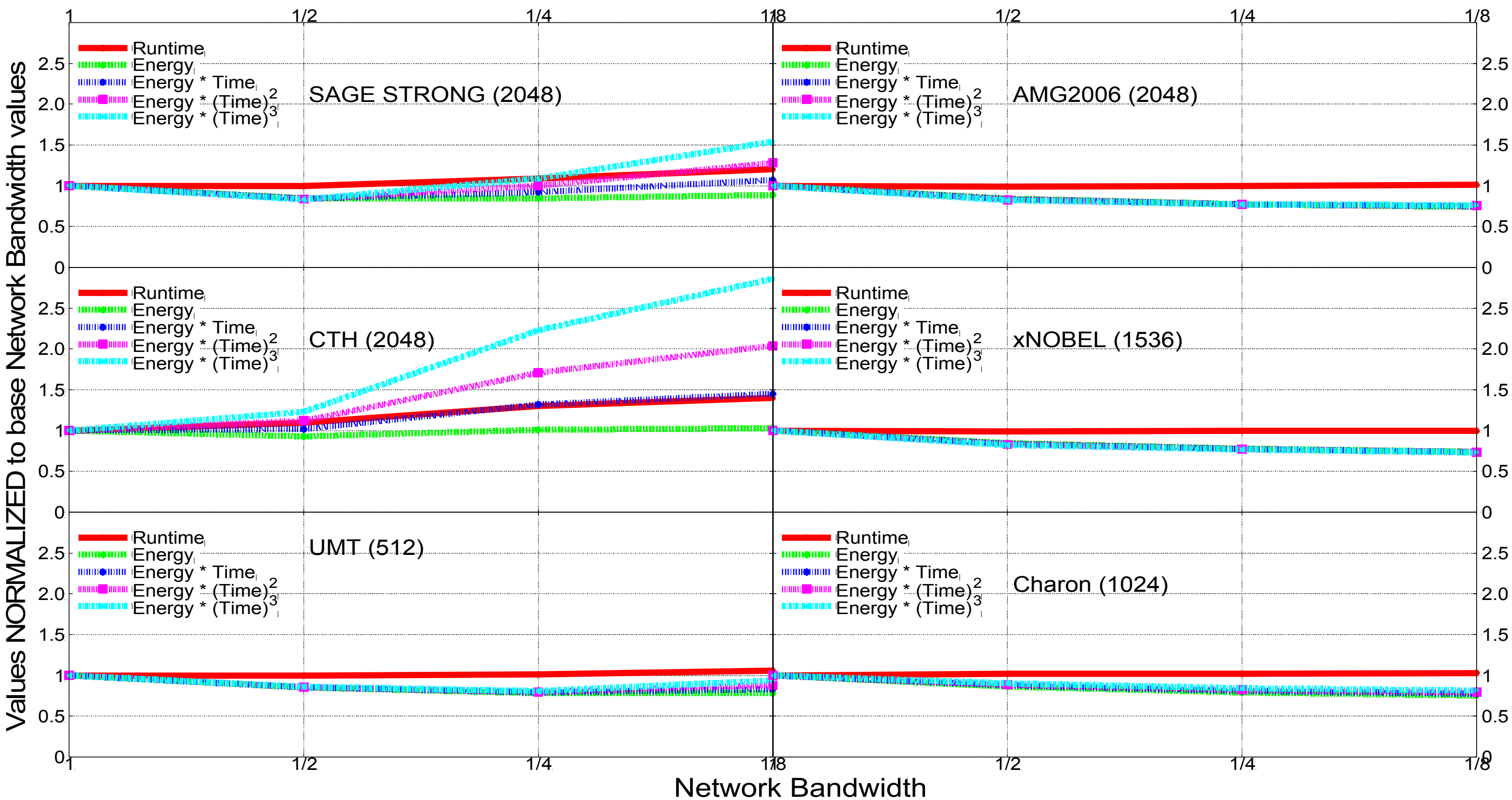
- Same question: Save energy during application run-time?
- Same applications
- Static tuning of network bandwidth
  - Required configuration and BIOS changes
  - Required complete system reboot to alternate settings
- Linear decrease of network energy (assumed)
- Total node energy contrasted
  - Network reduction might have affect on CPU energy
    - CPU energy measured as part of experiment

Experiment #3: Results

	Nodes/Cores	½ BW Run-time	½ BW Energy	1/4th BW Run-time	1/4th BW Energy	1/8th BW Run-time	1/8th BW Energy
SAGE_strong	2048/4096	↓ 0.593	↓ 15.3	↑ 8.90	↓ 15.5	↑ 20.2	↓ 11.4
SAGE_weak	2048/4096	↑ 0.609	↓ 14.3	↑ 8.23	↓ 15.8	↑ 22.6	↓ 9.63
CTH	2048/4096	↑ 9.81	↓ 7.09	↑ 30.2	↑ 1.04	↑ 40.4	↑3.50
AMG2006	2048/4096	↓ 0.815	↓ 15.8	↓ 0.116	↓22.7	↑ 0.931	↓ 25.9
xNOBEL	1536/3072	↓ 0.938	↓ 15.4	↓ 0.375	↓ 22.2	↓ 0.375	↓ 25.9
UMT	512/1024	↑ 0.357	↓ 14.7	↑ 1.07	↓ 21.7	↑ 6.32	↓21.8
Charon	1024/2048	↑ 1.55	↓ 13.7	↑ 2.15	↓ 20.8	↑ 2.67	↓ 24.5

Sandia National Laboratories

Experiment #3: Additional Analysis



Overall Observations

- Large savings can result from relatively simple changes
  - Halting unused cores during idle
- Increased application energy efficiency can result from:
  - Static CPU frequency tuning at large scale
  - Network bandwidth tuning
- Applications exhibit a *sweet spot*
  - Dependent on scale
  - Dependent on platform
  - Dependent on ????
- Tunable platform components
  - Dial in efficiency

Acknowledgments

Committee:

Dr. Wei Shu - Thesis advisor, committee chair and collaborator, University of New Mexico Electrical and Computer Engineering Department

Dr. Howard Pollard - Committee member, University of New Mexico Electrical and Computer Engineering Department

James A. Ang - Committee member and Manager of the Scalable Computer Architectures Department at Sandia National Laboratories

Collaborators:

Kevin Pedretti, Sue Kelly, John Vandyke, Kurt Ferreira and Courtenay Vaughan – Technical Staff Sandia National Laboratories, and Mark Swan - Cray Inc.

Funding:

National Nuclear Security Agency (NNSA) Advanced Simulation and Com- puting (ASC) program and the Department of Energy's (DOE) Innovative and Novel Computational Impact on Theory and Experiment (INSITE) program. Sandia National Laboratories Center 1420 Sudip Dosanjh Senior Manager, Department 1422, James Ang manager and Department 1423 Ronald Brightwell manager.

Sandia National Laboratories

Questions?

.. not everything that can be counted counts,
and not everything that counts can be counted.
- *William Bruce Cameron*