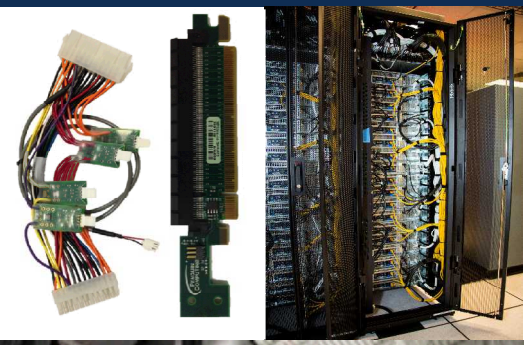
SAND2013-7134P



Power, Energy and High Performance Computing

James H. Laros III

With - Kevin Pedretti, Sue Kelly, Kurt Ferreria, John Van Dyke, Courtney Vaughan, Wei Shu, David DeBonis, Phil Pokorny

Sandia National Laboratories

Exceptional service in the national interest

U.S. DEPARTMENT OF ENERGY

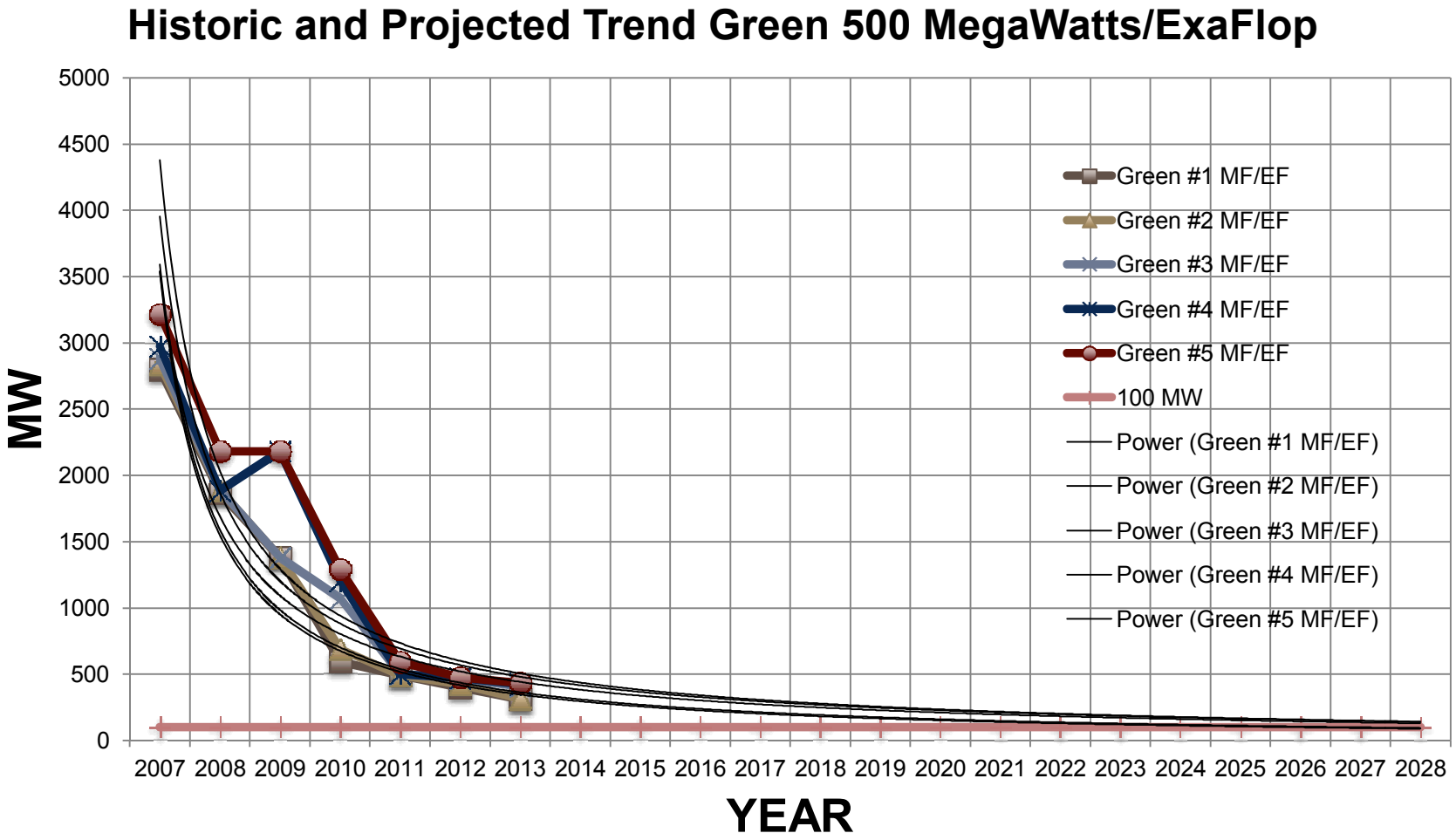
NNSA

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

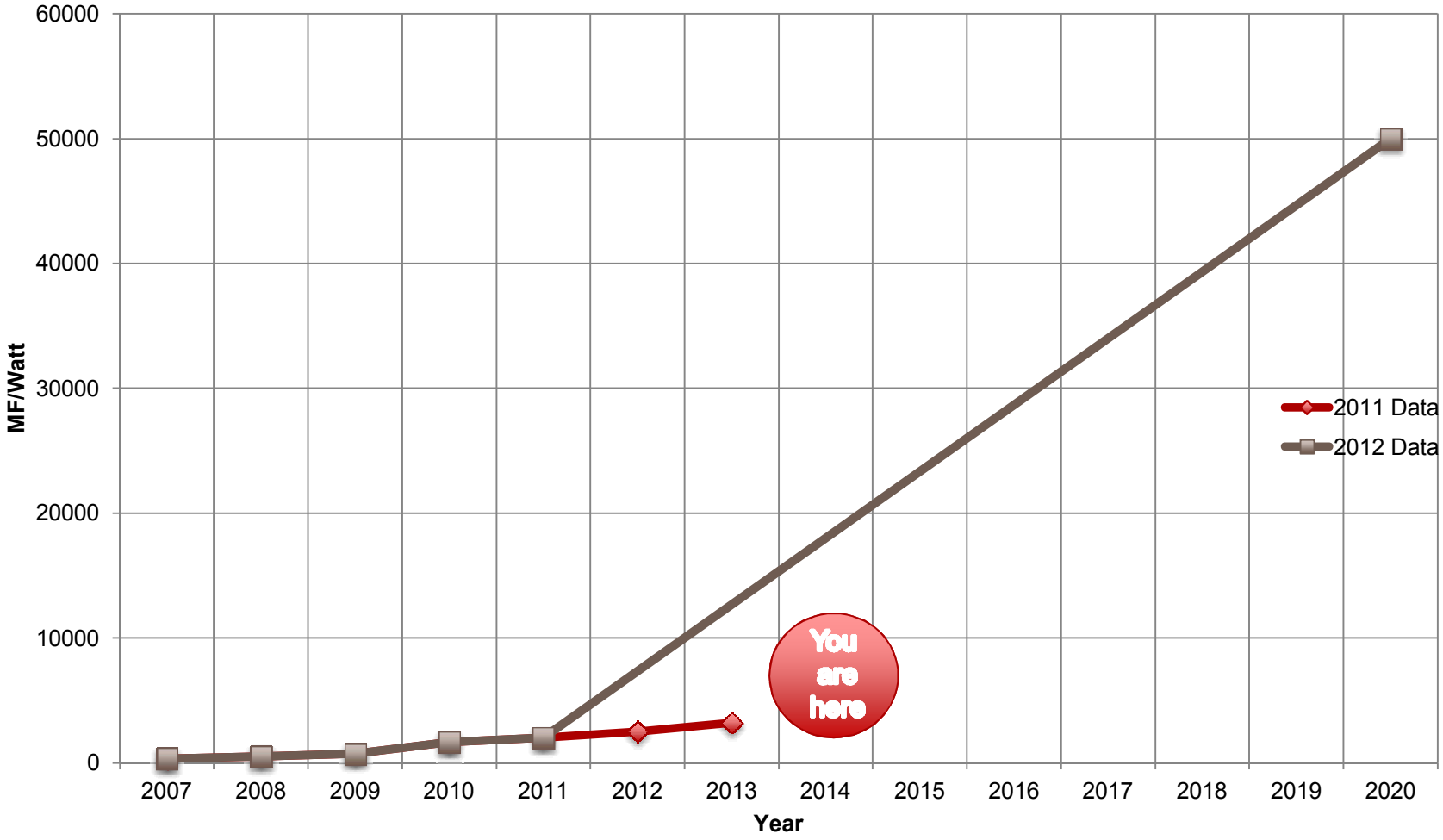
Power Forecast

sunny to mostly cloudy with a 99% chance of storms mixed with afternoon sunshine…

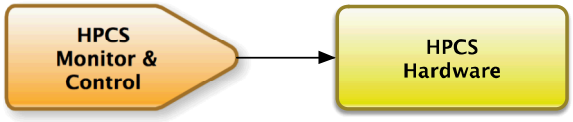
Sandia National Laboratories



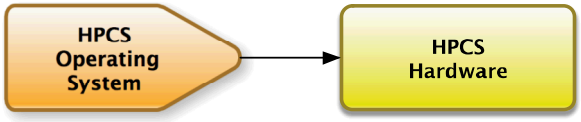
So how are we doing?



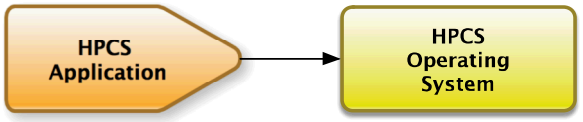
Starting Back in 2007 (ish)



- First research accomplished on Red Storm
- Discovered we could measure voltage and **current** of CPU per node, at scale (few thousand nodes)
  - Could also measure network (Seastar)
- We could also manipulate network bandwidth
  - Measured impact on energy of bandwidth reductions
- Early analysis using REAL applications
  - 6X suite which was later used for Cielo acceptance
  - Important because our priority is impacting Sandia's mission
- Developed measurement infrastructure
  - No small task



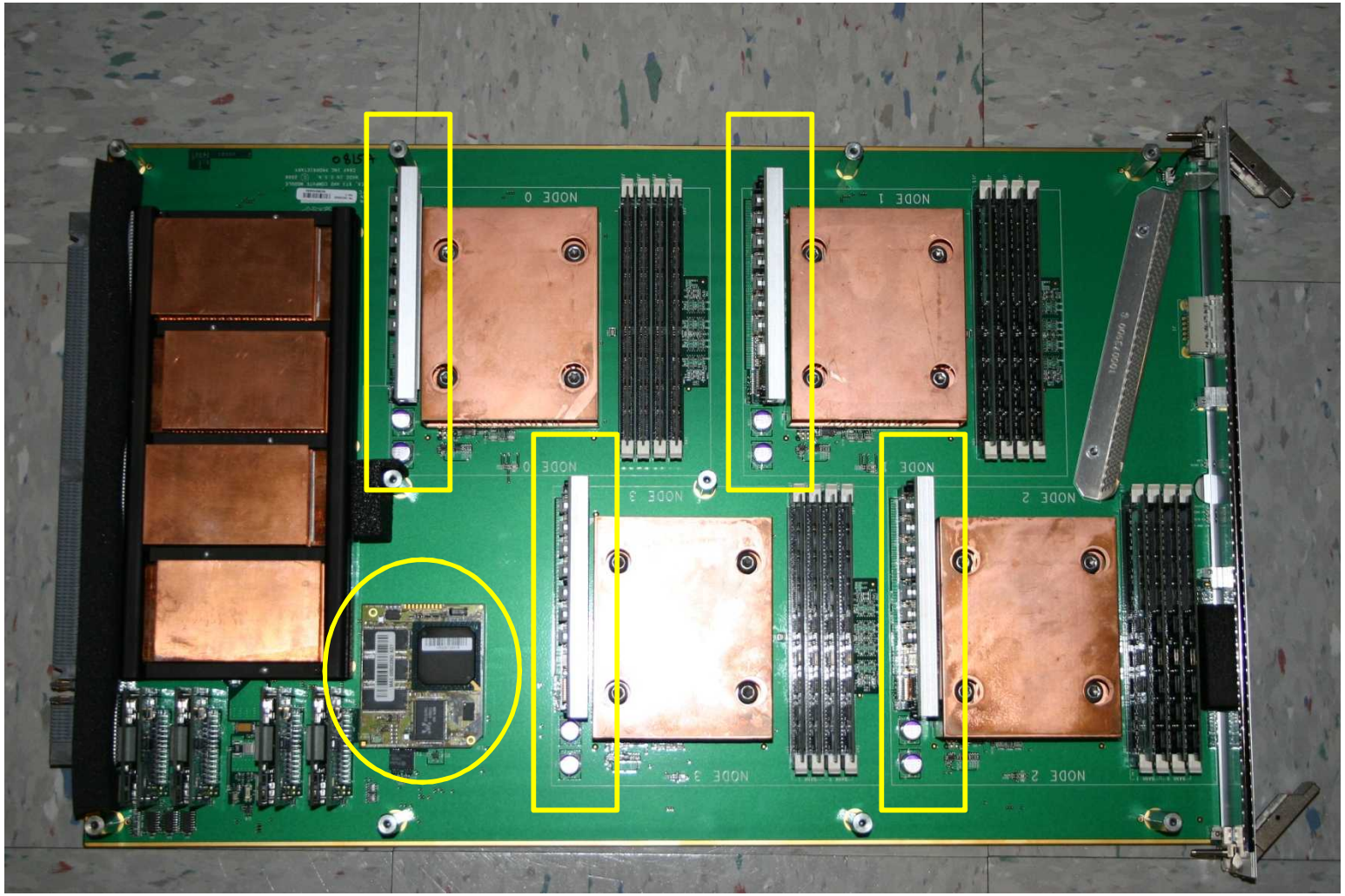
- Instrumented Catamount to deterministically control power states
  - Node centric Linux Governors have proven to be detrimental in practice
- Also implemented MPI-Profiling layer for dynamic tuning
  - Never used in our experiments



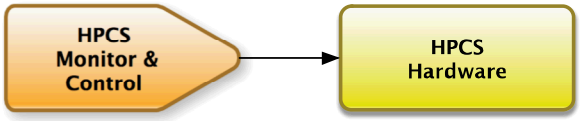
Sandia National Laboratories

MEASUREMENT AND CONTROL

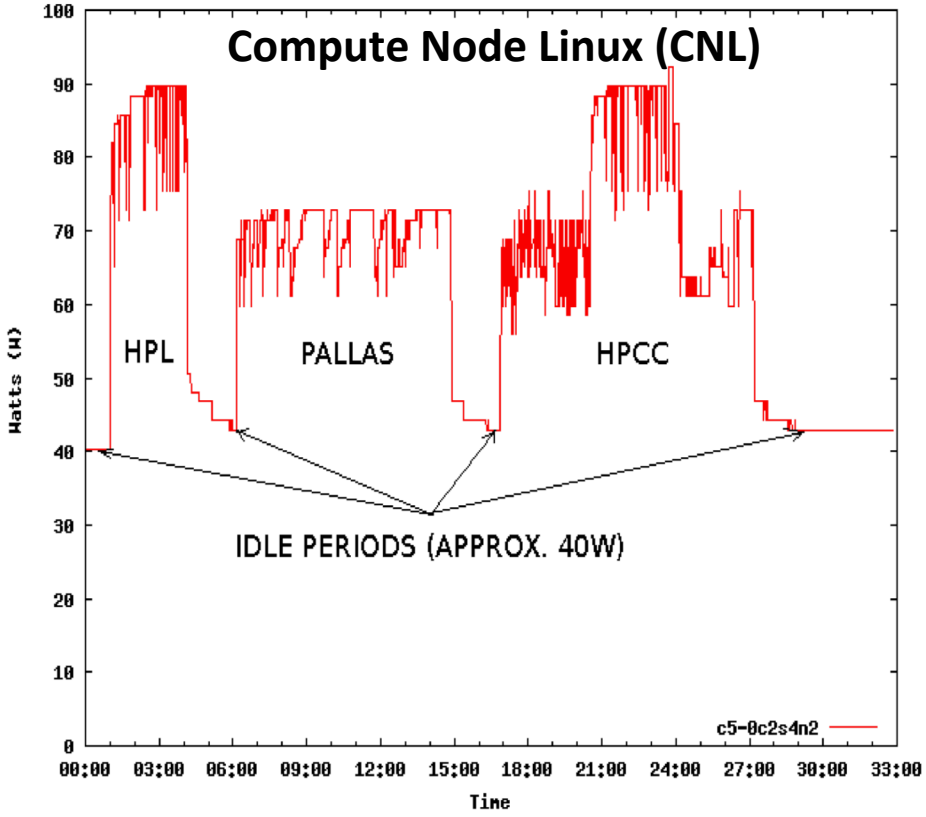
XT4 Board

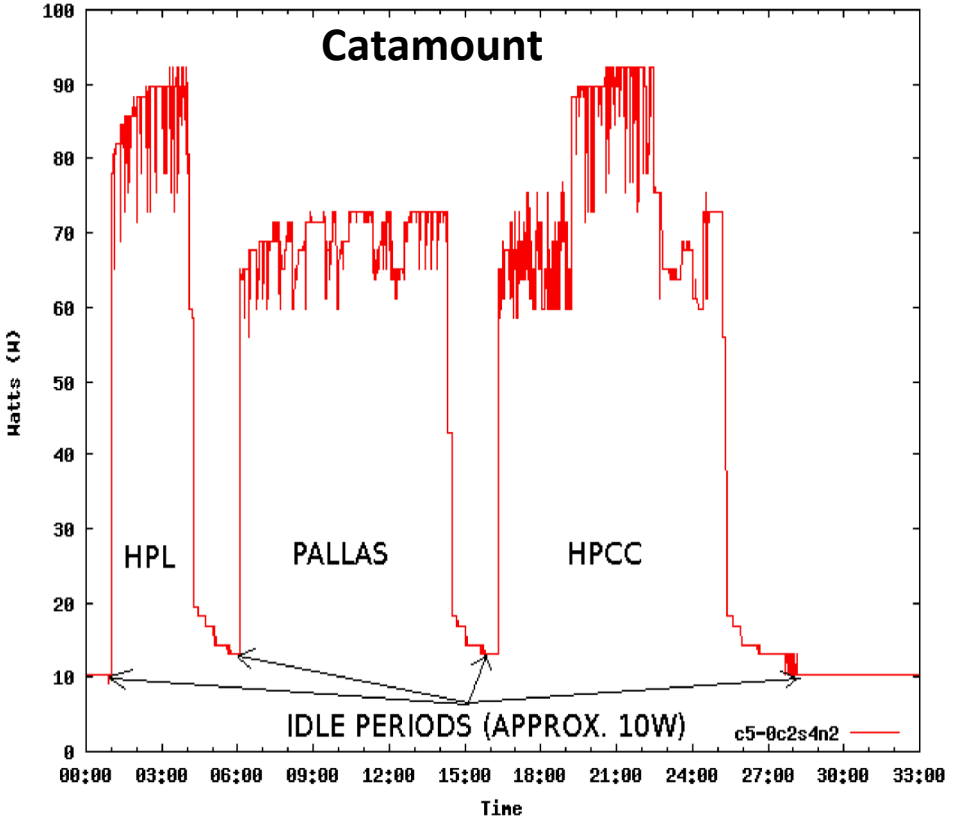


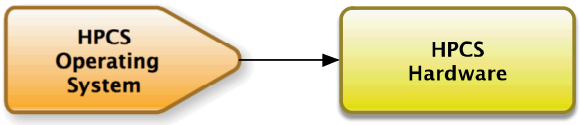
Application Profiles



Sandia National Laboratories

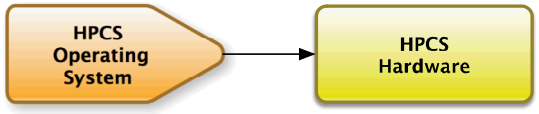




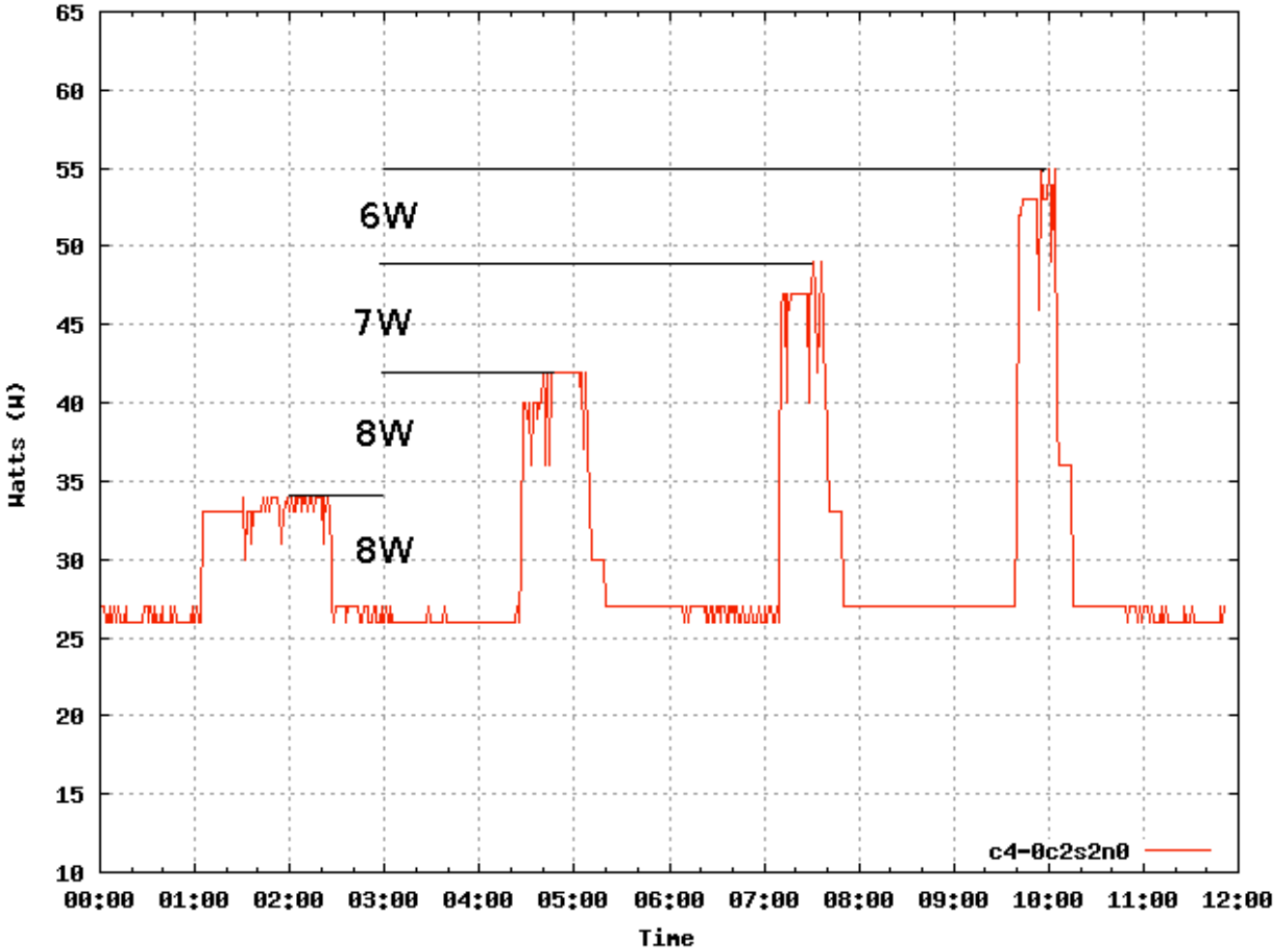


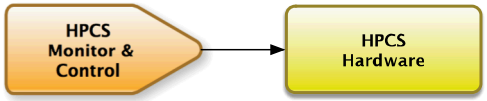
Idle Power Draw

Deterministic Control



Sandia National Laboratories



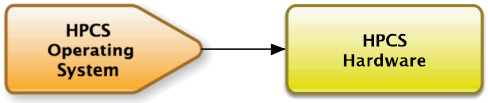


Ability to Observe

Sandia National Laboratories

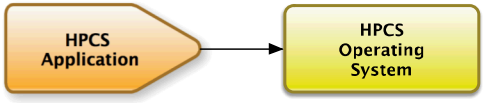
TUNING CPU POWER DURING APPLICATION RUN-TIME

- Save energy during application run-time?
- Targeted modifications
  - OS trap to deterministically change P-states (processor frequency)
  - User space library to request changes
  - MPI profile layer to intercept potential wait periods (for example)

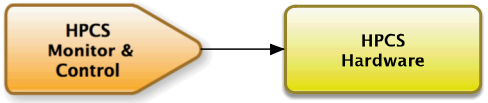


"Science progresses best when observations force us to alter our preconceptions" – Vera Rubin

- Discovered Static Tuning could be highly beneficial
  - More stable
  - Easily coordinated



- CPU energy contrasted
  - CPU accounts for 44-57% of total node energy
  - CPU largest single component consumer of energy
  - CPU analysis most useful to contrast with other platforms



CPU Tuning Results

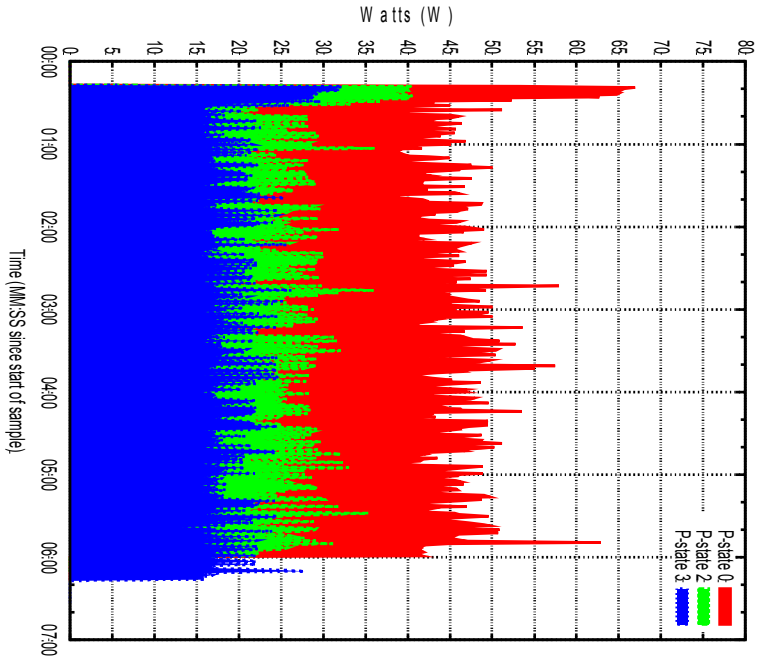
Sandia National Laboratories

	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				

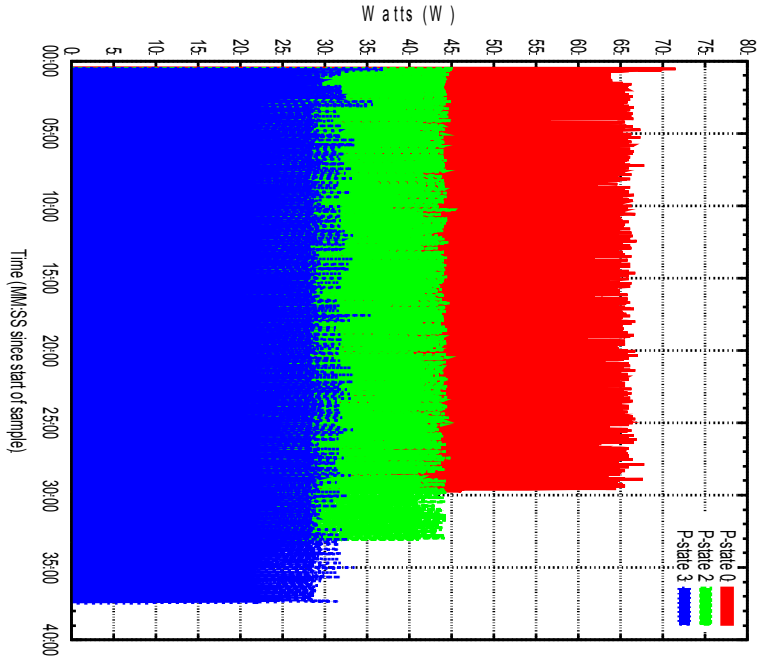
Application Energy Signatures

Sandia National Laboratories

Tuning SAGE = Good



Tuning CTH = Bad

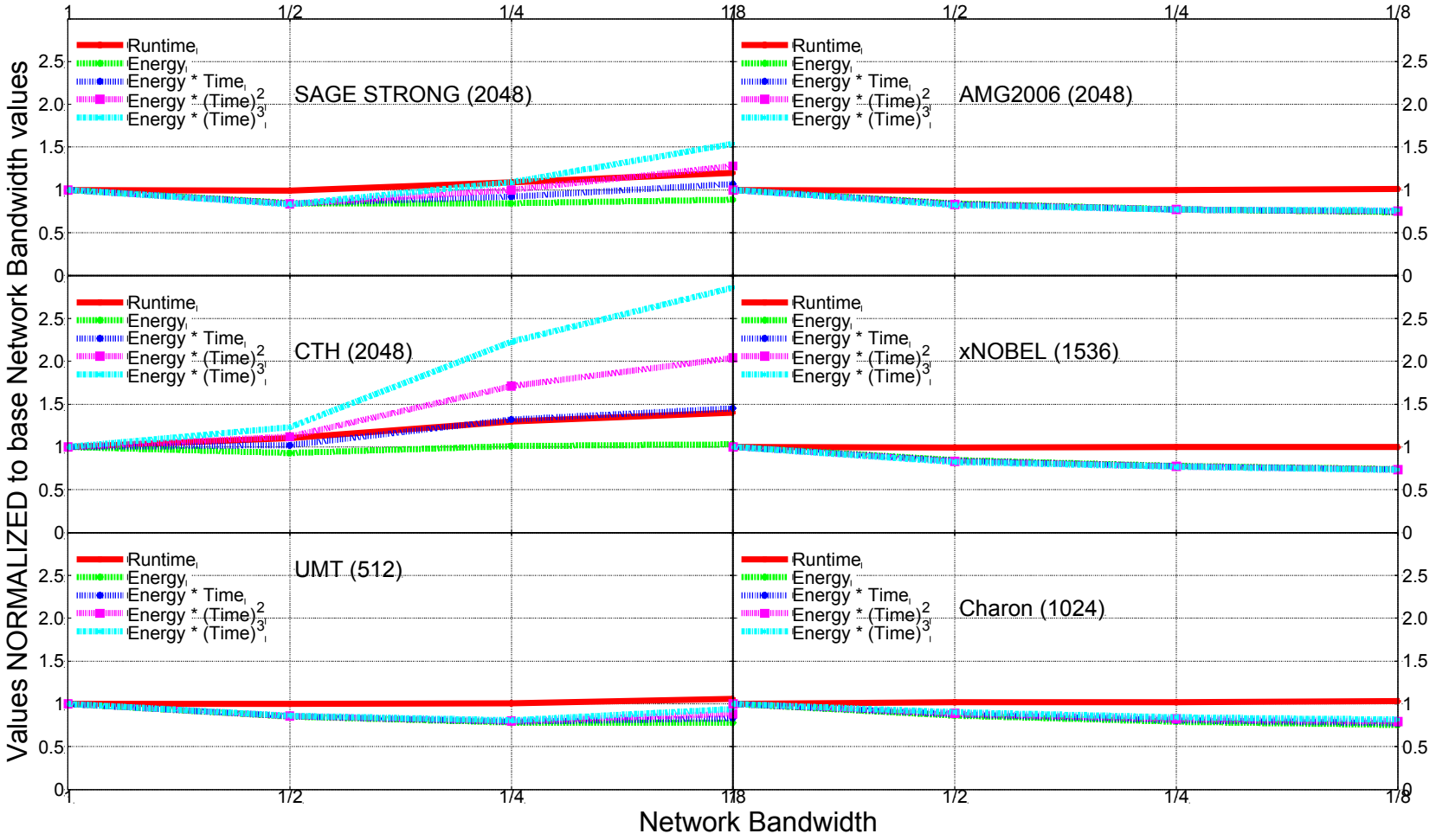


Network Bandwidth Tuning Results

Sandia National Laboratories

	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

EDP: A Fused Metric?



Doesn't consider all of our variables need something better!

So Whats Next?

- Plan A: Collaboration with Cray
  - Not exactly a template for co-design
  - good thing we had a plan B
- Plan B: Commodity path
  - Has the advantage of not being architecture specific
  - Scale is harder
- Enter the Advanced Architecture Test Bed Program
- Partnered with Penguin Computing
- Co-designed Commodity Power Measurement Device

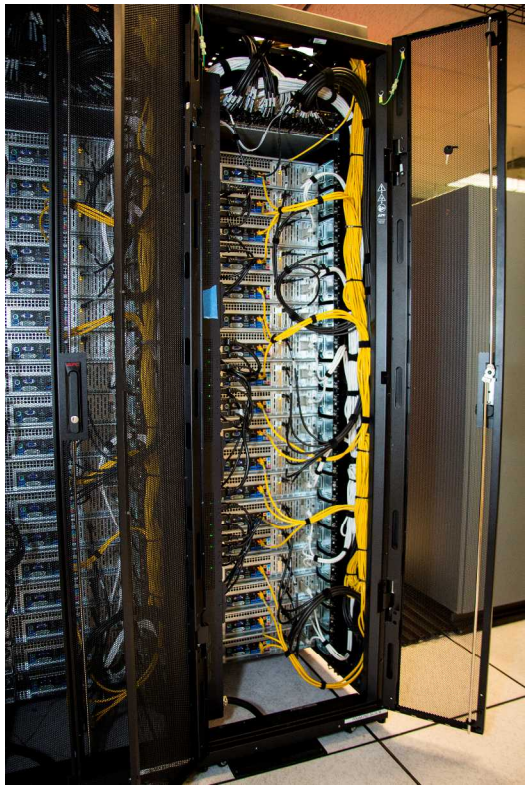
Sandia National Laboratories

POWERINSIGHT

Teller – Platform Architecture

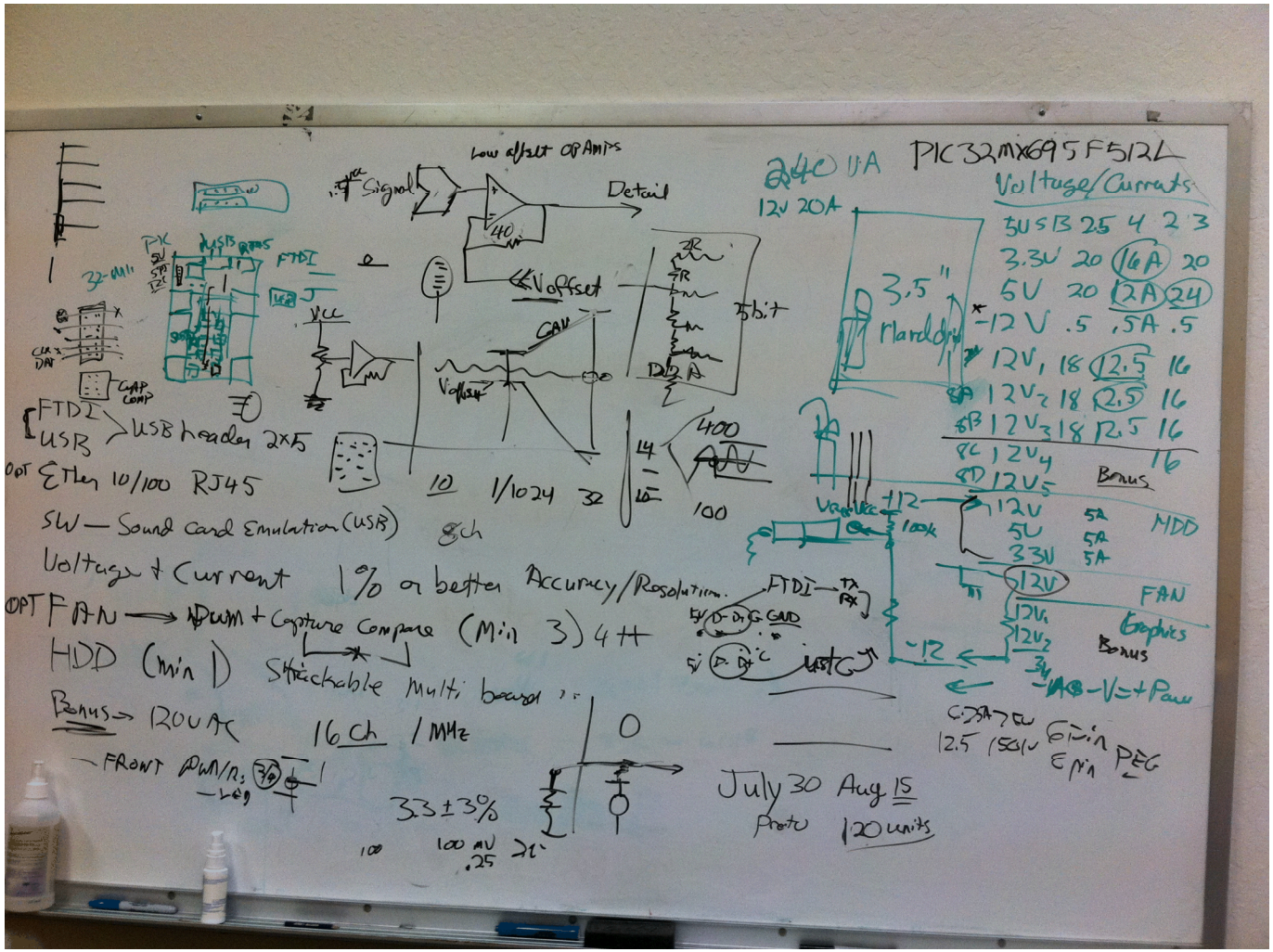
Sandia National Laboratories

- Part of Sandia's **Advanced Architecture Test Bed Program**
- 104 Nodes
  - Single AMD Fusion A10-5800K processor
    - 4 x86 cores @ 3.8GHz (Piledriver)
      - Turbo 4.2GHz
    - 384 Radeon Northern Islands GPUs @ 800MHz
  - Qlogic QDR InfiniBand
  - Ethernet management network
    - 1 node admin
    - 1 PowerInsight out-of-band data
  - 256GB SSD/node
- **PowerInsight**



Initial Brain-Storming

High-Tech Napkin



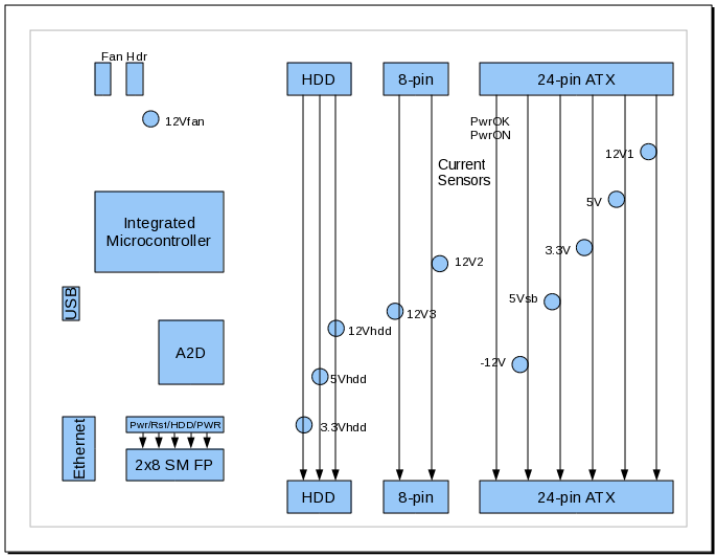
Requirements Development

- Investigated a range of power supplies
  - Only one rail (5V) was rated more than 20A
- Key design goals:
  - 3.5 in hard drive form factor - to hold device
  - **Ethernet (out-of-band data collection)**
  - **USB and Serial (in-band) connectivity to host (OS)**
  - **Support up to 16 channels – Component Level Measurement**
  - System and GPU rails
- Ideas that didn't make the cut
  - Dynamic offset and scaling
    - Too complicated, unnecessary
    - USB sound card emulation
      - Might implement in future version
    - Fan PWM and Tach interface

Evolution of Design Layout

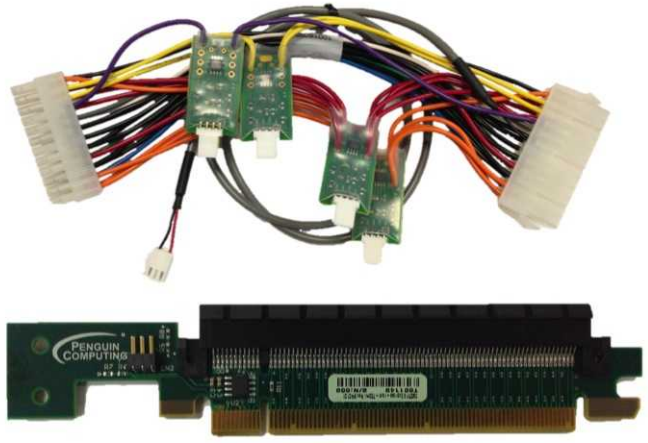
Initial Layout

- Single integrated card with connectors
  - Requires cables to connect to system
  - Would possibly require system harness modifications
  - Fixed CPU configuration



Final Solution

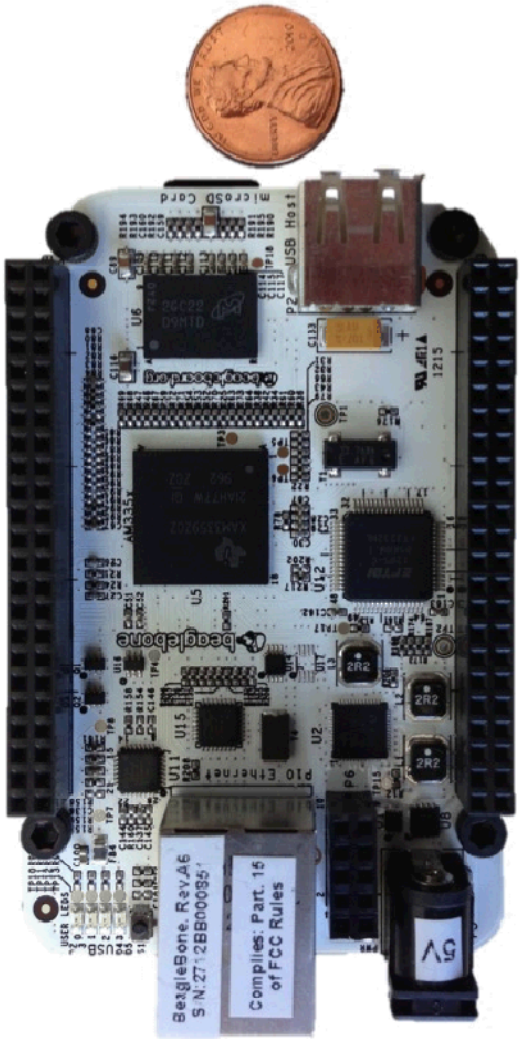
- Distributed sensors built into custom cable harnesses
  - Requires no modification to system harness



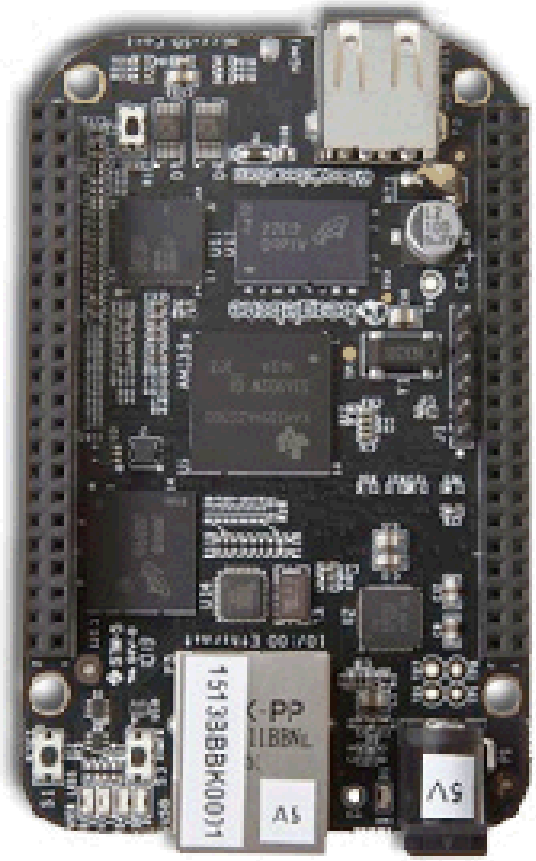
Sandia National Laboratories

BeagleBone (current)

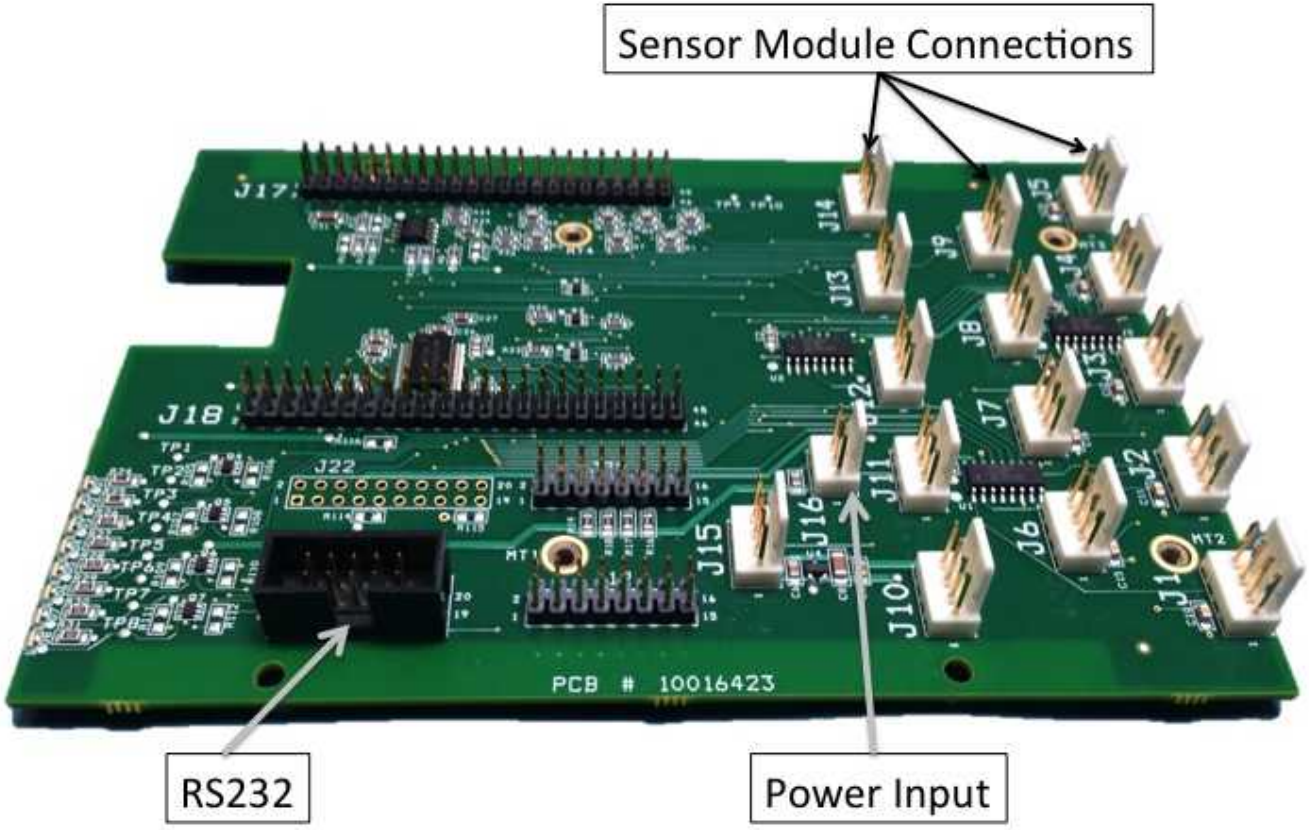
BeagleBone BLACK (future?)

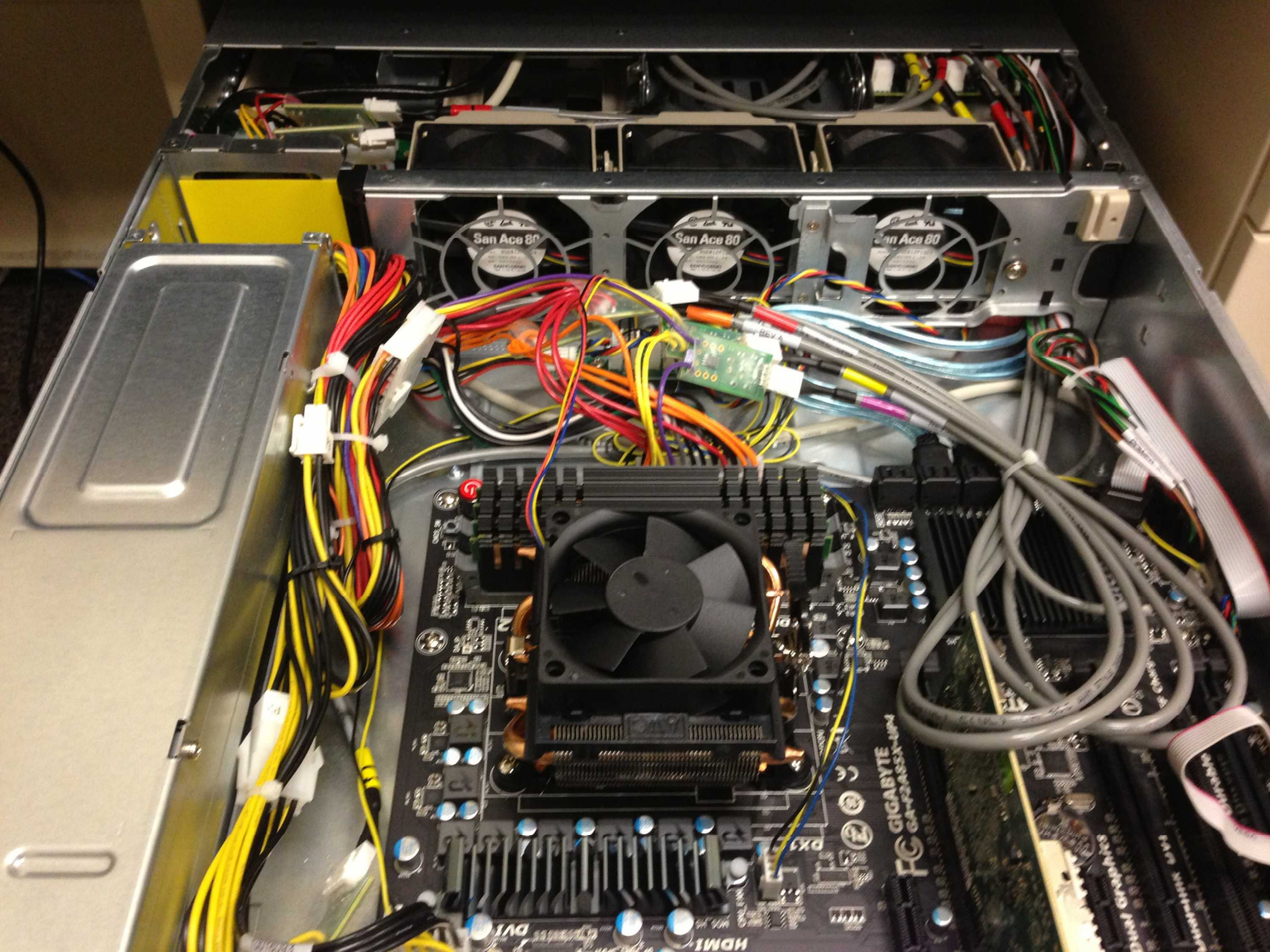


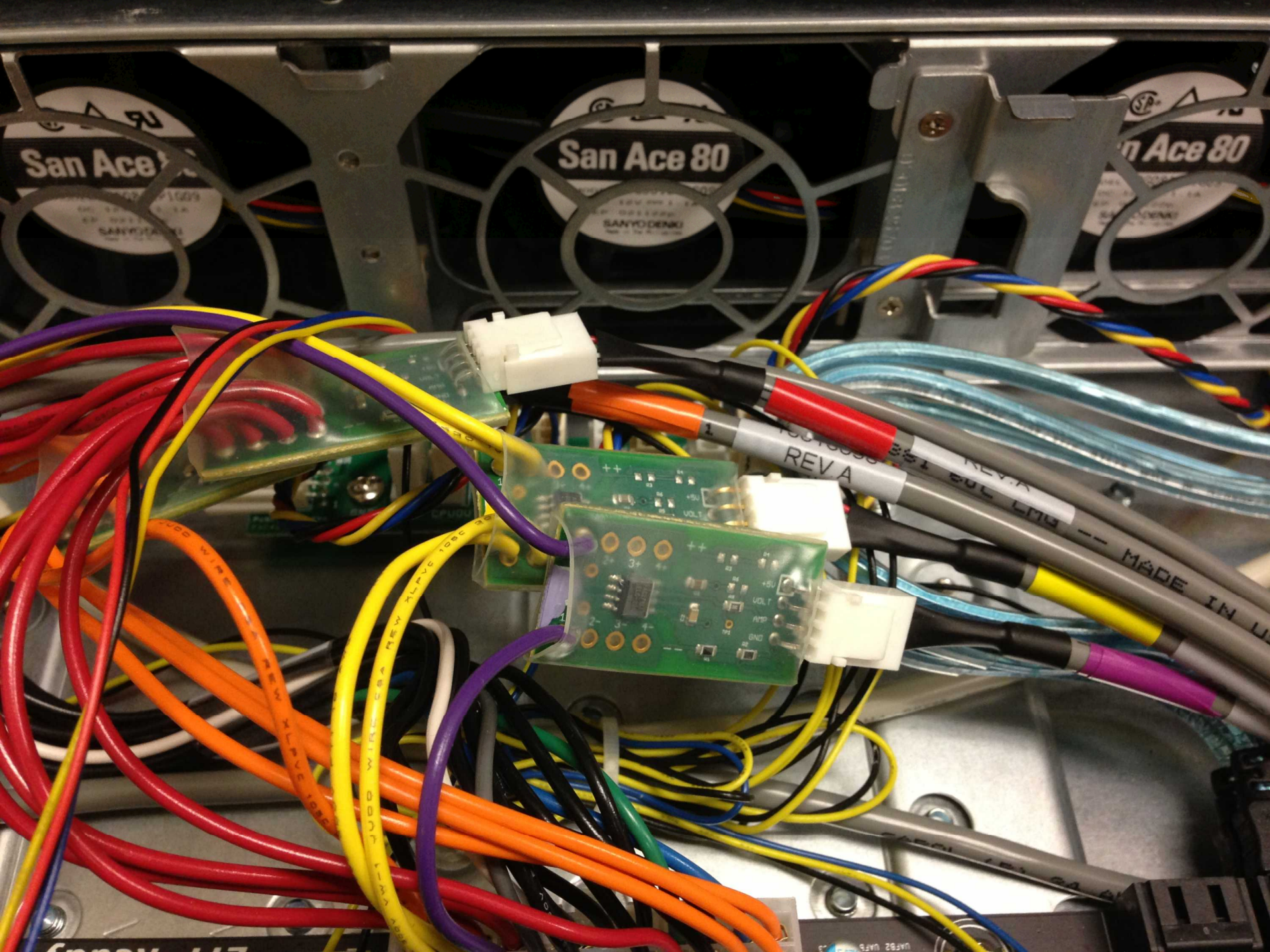
UPGRADABLE



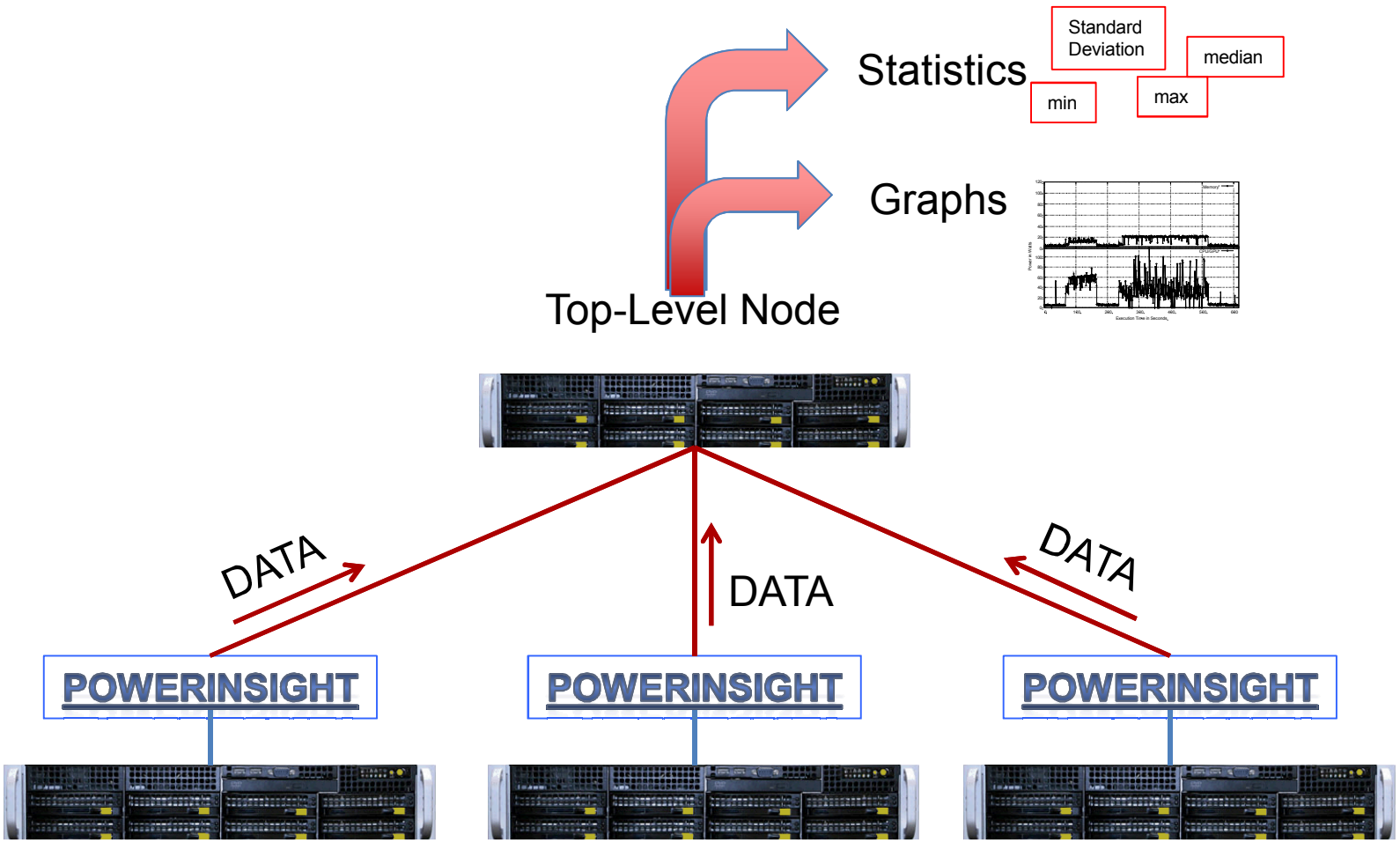
Custom Cape




San Ace 80
San Ace 80
GIGABYTE


San Ace 80
SANYO DENKI
REV.A
VOLT
AMP
GND

Data Collection – System Level



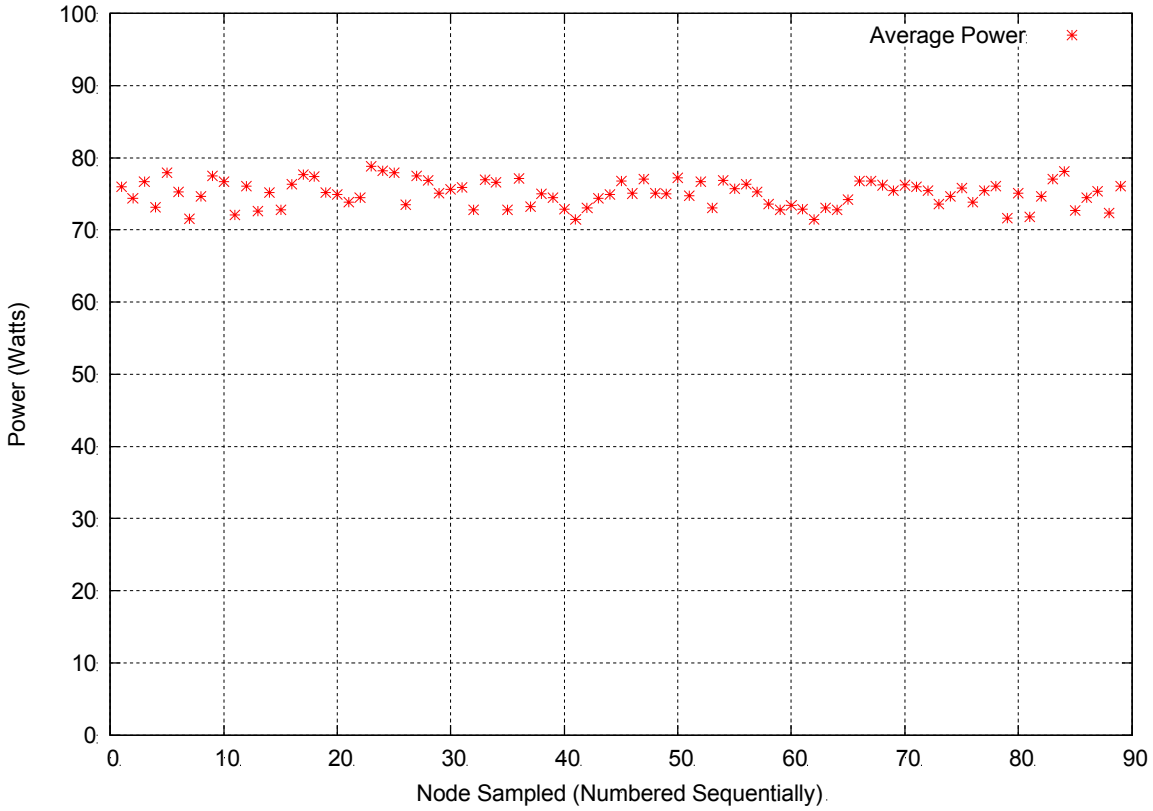
Experimental Validation

Sandia National Laboratories

- Q: Will this device be useful for application power and energy analysis?
- Simple 1/sec sampling (capable of much much more)
- Single Node High Performance Linpack (HPL)
  - All four x86 cores
  - One MPI-task per core
- Compared results of: performance, Power and Energy
  - From node to node
  - Repeatability per node
- Experiments designed to validate PowerInsight's use for application energy analysis

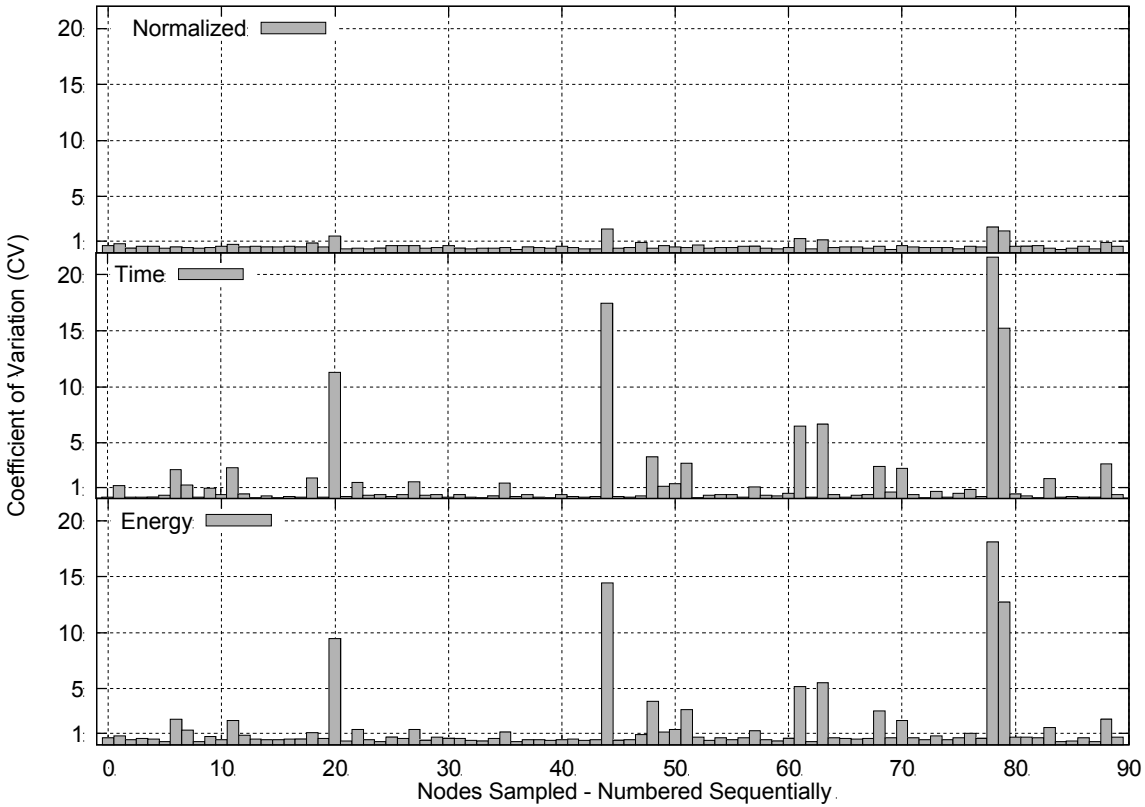
Validation Across Nodes

- Q: Are results consistent from node to node for same workload?
  - Yes
  - As graph shows results for all nodes are between 70 and 80 Watts.
  - Coefficient of Variation (CV) across all nodes is only 2.54%
    - Reasonable considering deviations from die to die
- Is Average Power within expected range?
  - Yes
  - For HPL we expect 70-80% of TDP which is approximately 100W for this chip.
  - Range, 70-80W
  - Confirmed by AMD
- Confirms PowerInsight useful for comparing executions on different nodes



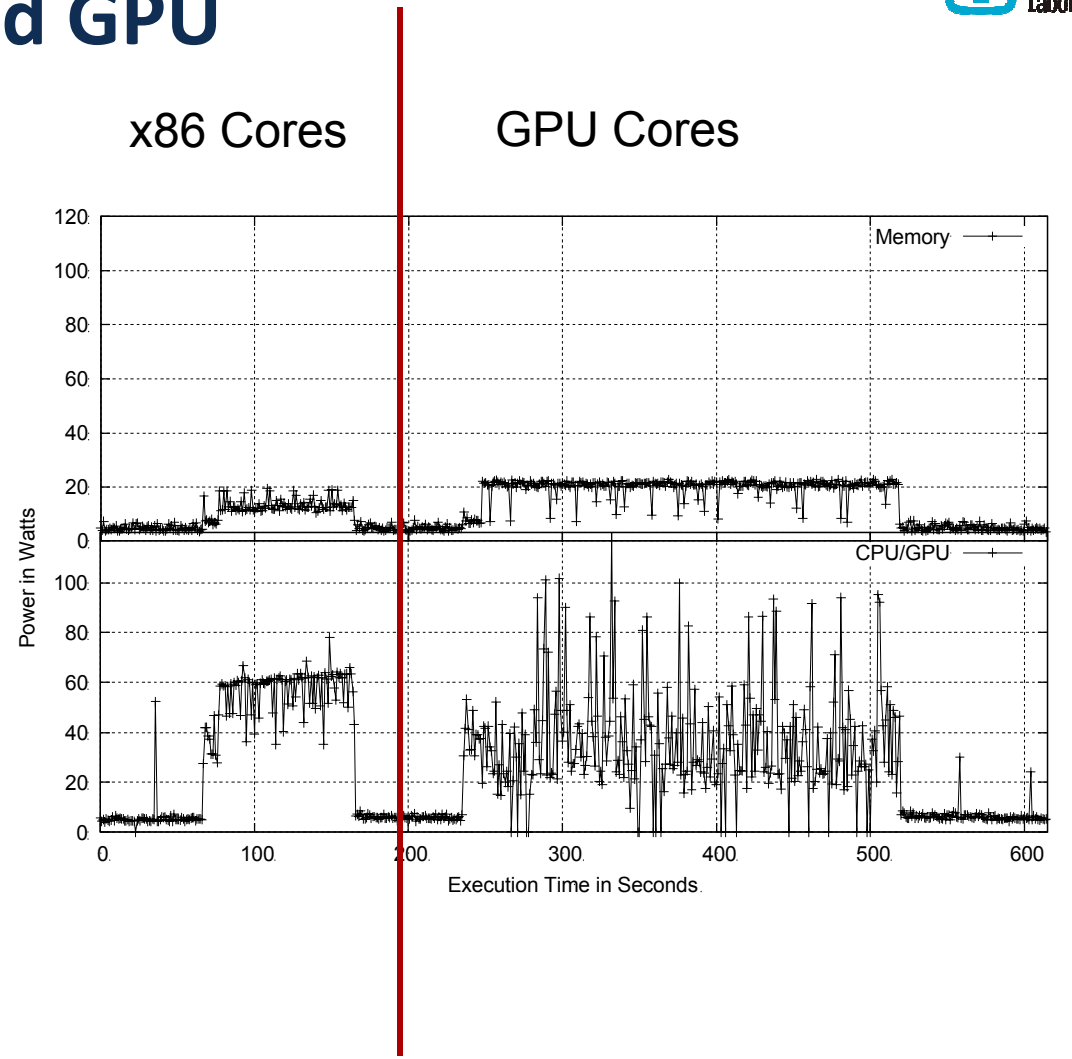
Validation Per Node

- Q: Are results consistent on the same node for the same workload?
  - Yes
  - Included all runs (897 successful out of 900 attempts)
  - When normalized with execution time the variation on 84 out of 90 nodes is less than 1%
  - Including outliers maximum CV per node is 2.28%
  - If we exclude the few anomalies all runs are much less than 1%
- Confirms PowerInsight useful for comparing executions on same node.
- Conducting experiments where differences between baseline and subsequent executions are compared is very common and useful

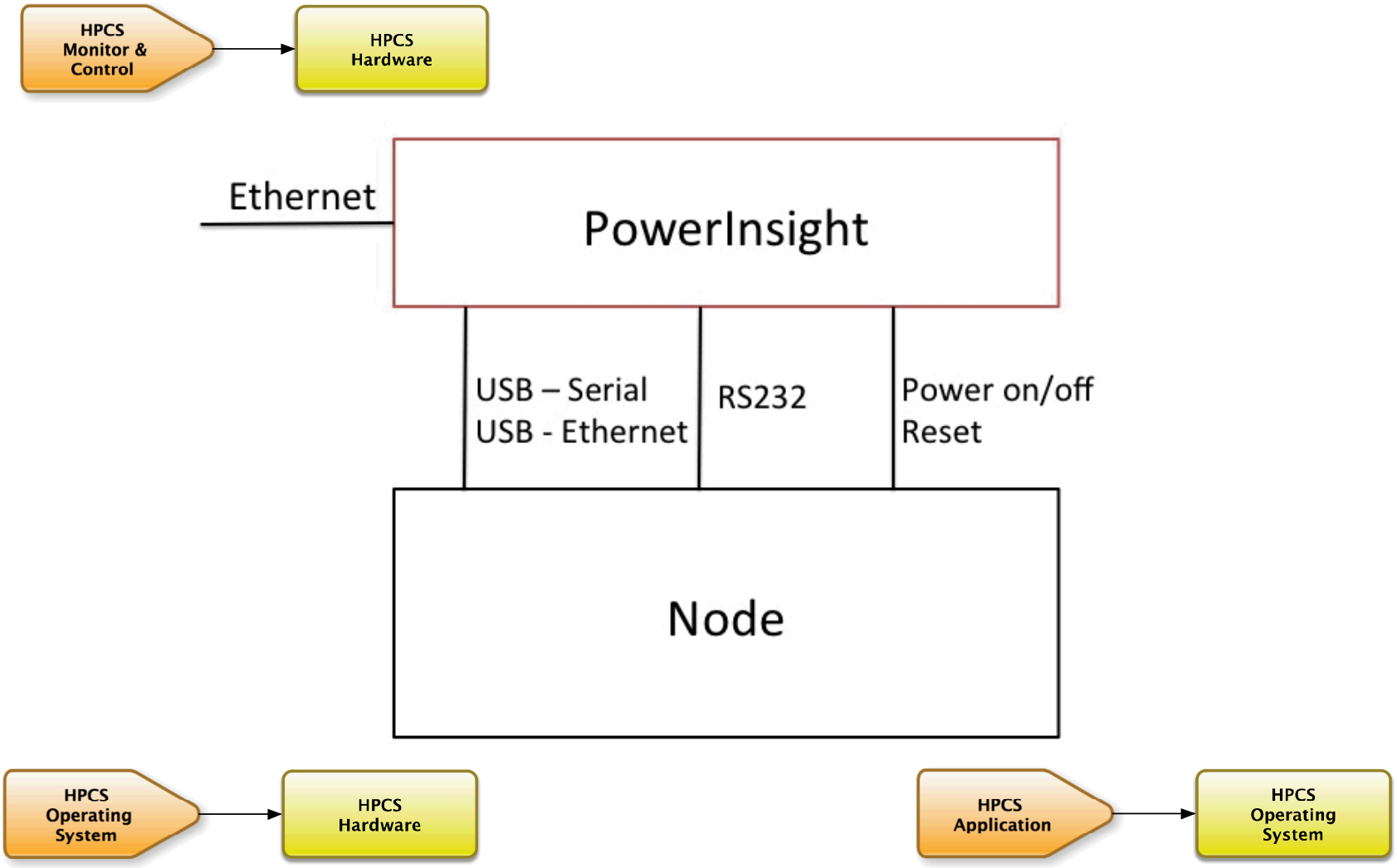


MiniFE on CPU and GPU

- MiniFE is one of Sandia's mini-apps
  - Applications designed to represent core functionality of larger production apps
- First execution only using x86 cores (left part of graph)
- Second execution only using GPU cores (right part of graph)
  - Note, GPU kernel is launched from an x86 core
- Application Energy Profiles
  - Profile of Memory – Top graph
  - Profile of CPU – Bottom graph
- This is an example of the type of application analysis PowerInsight will enable



Designed for Potential



Sandia National Laboratories

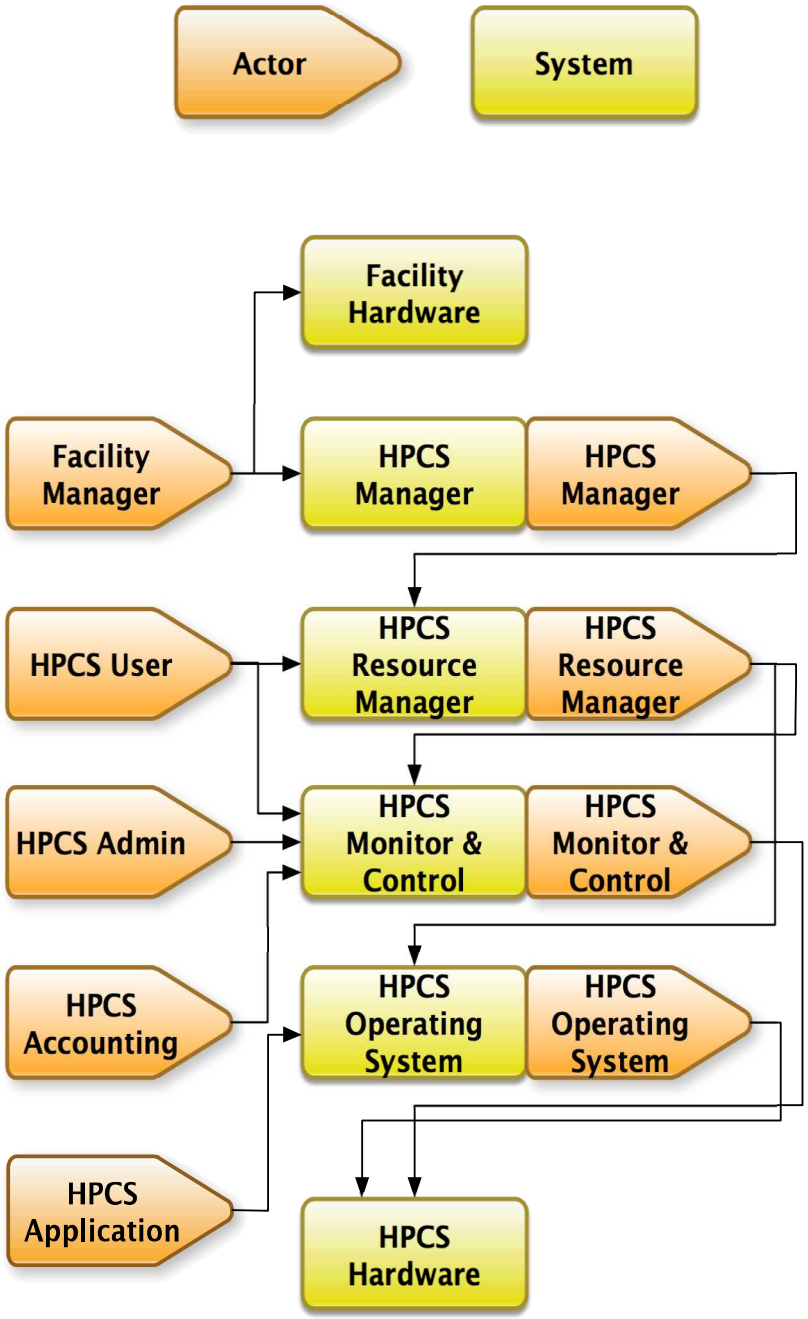
POWER API

Power API – *A Use Case Approach*

Sandia National Laboratories

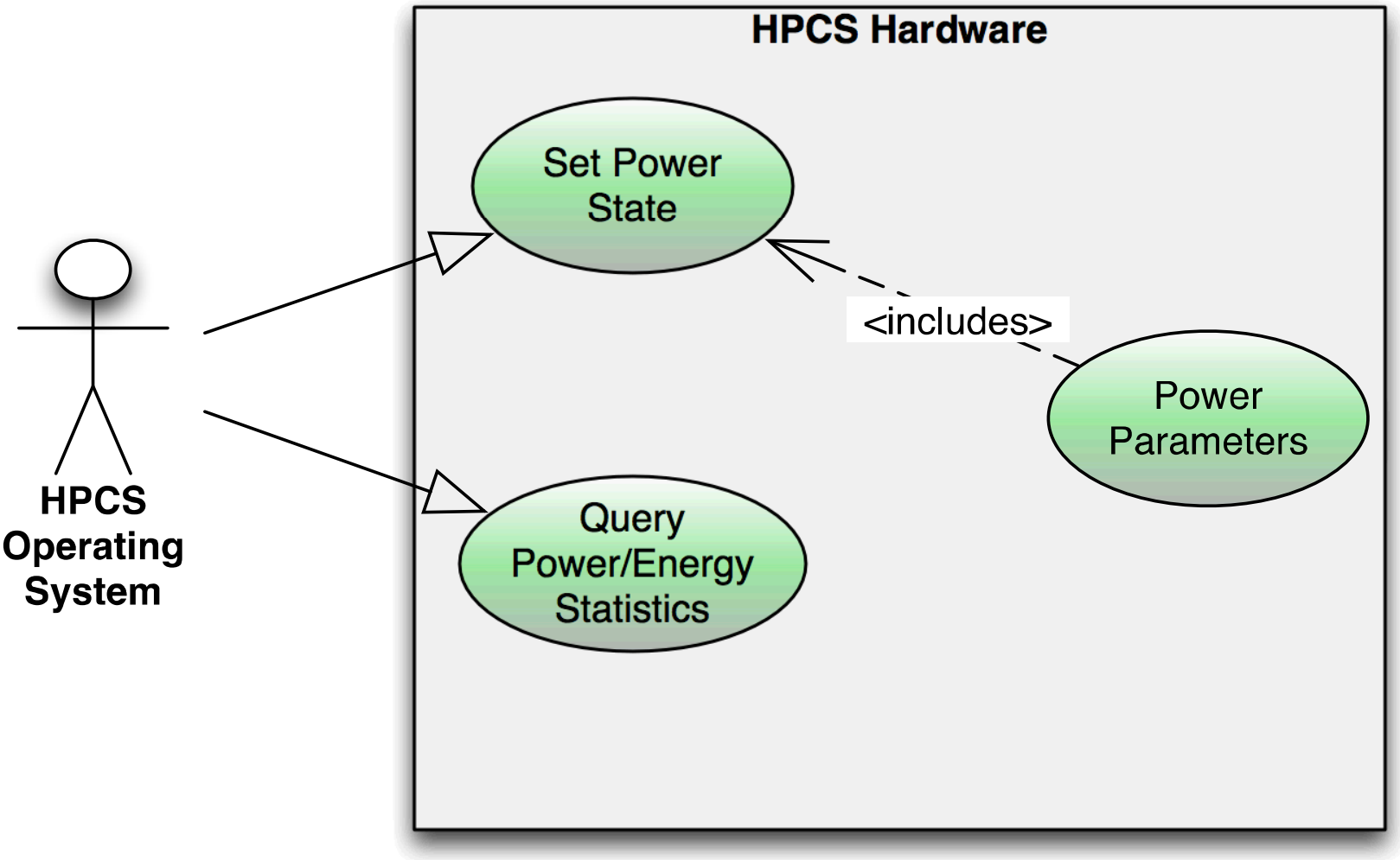
- 2014 L2 Milestone – Power API Definition/Specification
- Use case approach used to define **SCOPE and INTERFACES**
- Reviewed by Labs, Universities and Commercial partners

33

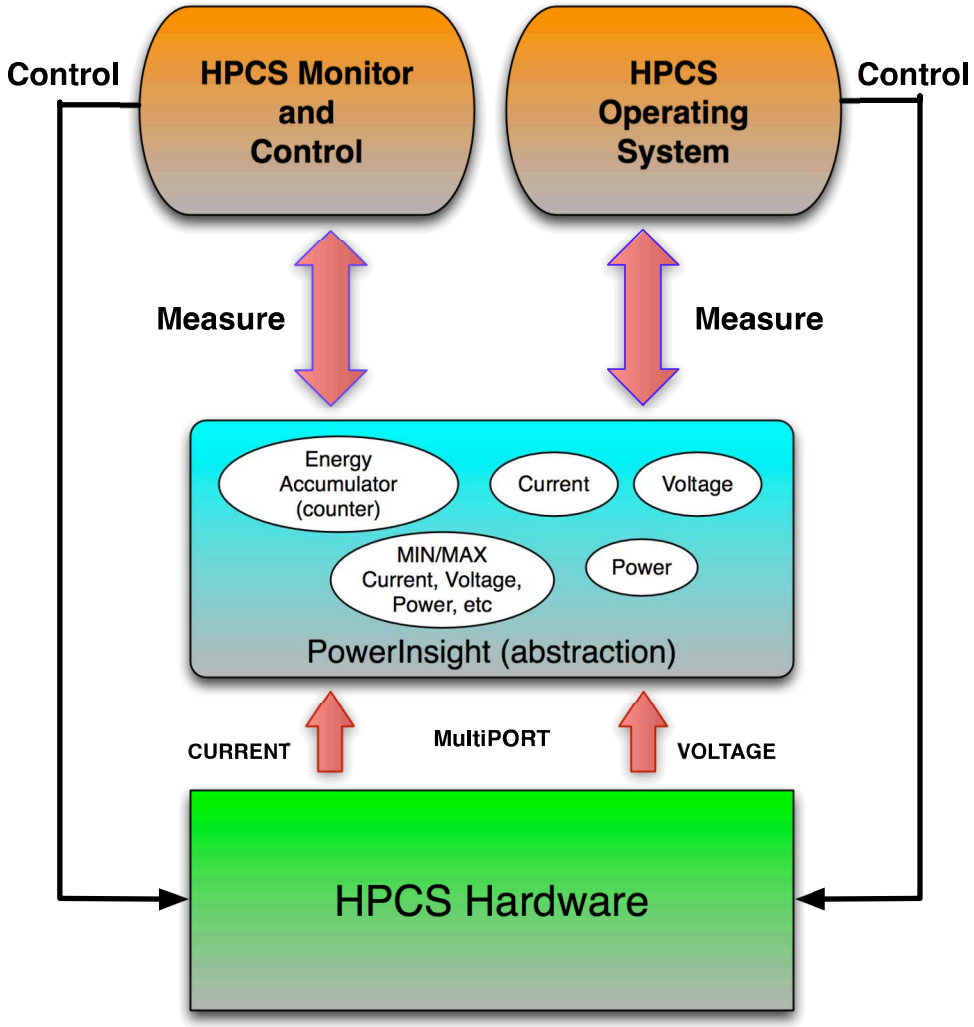


Sandia National Laboratories

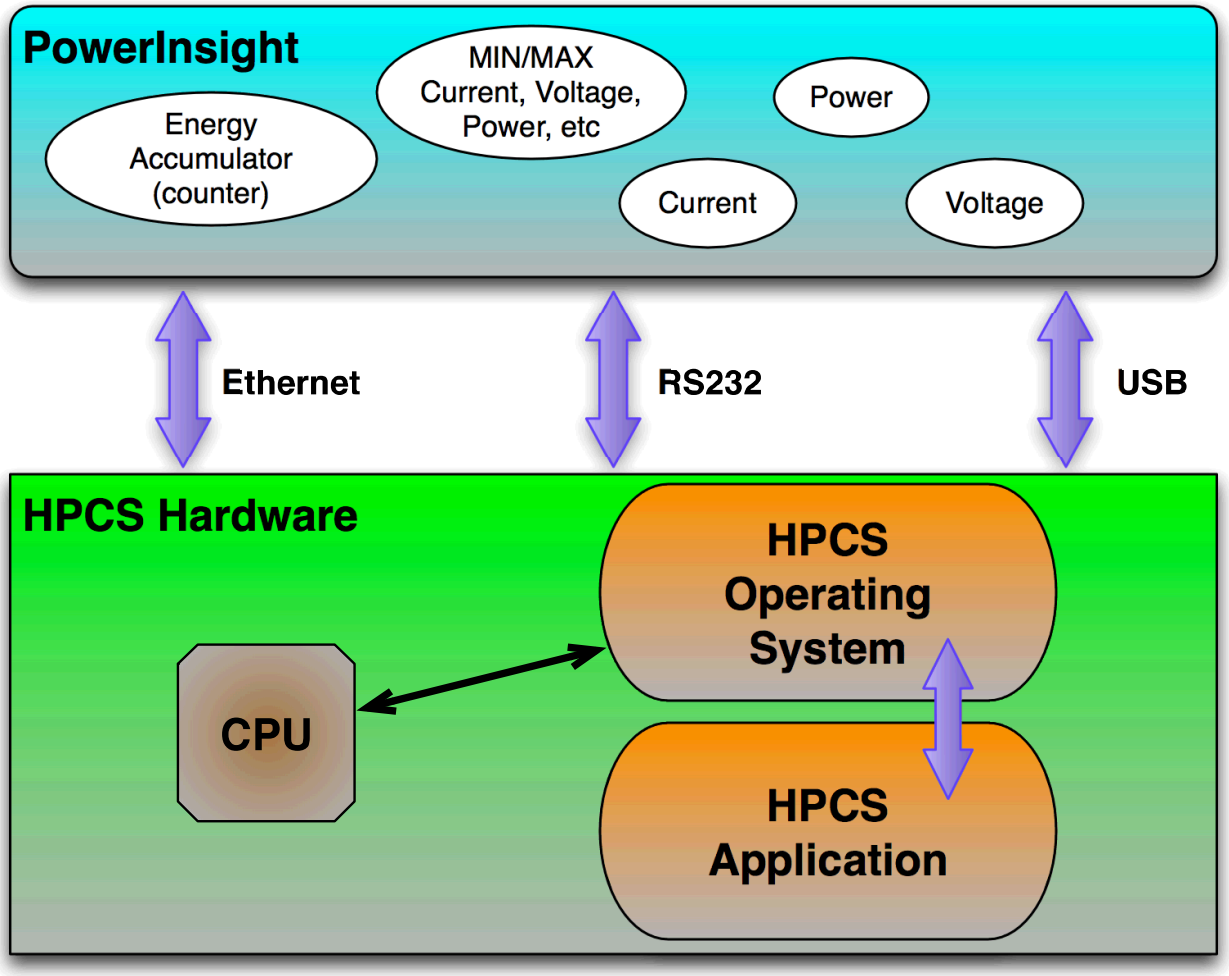
Actor: HPCS Operating System
System: HPCS Hardware



Prototyping with PowerInsight



Prototyping with PowerInsight

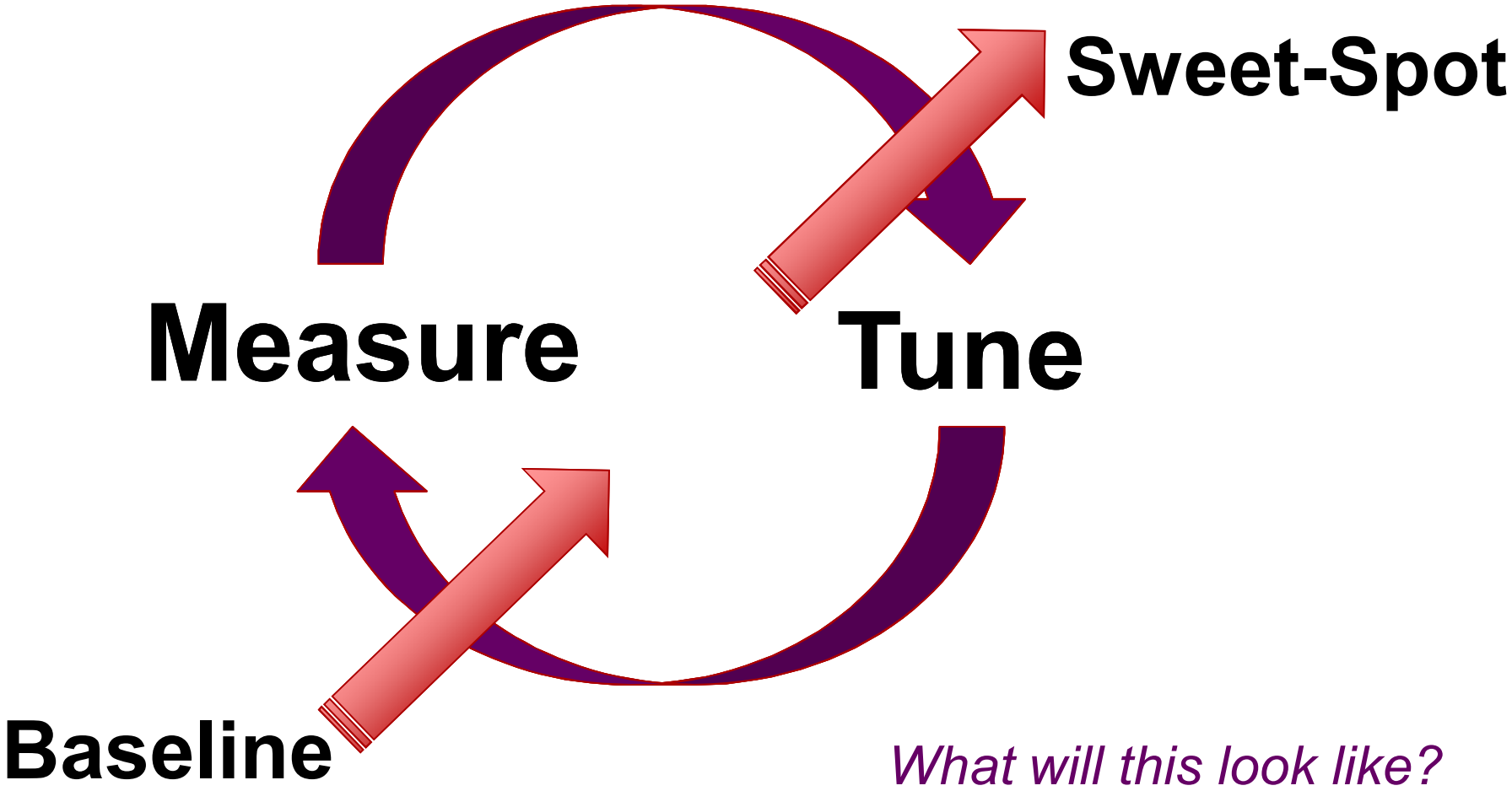


Cray Cascade

Sandia National Laboratories

- Arrived July 19th Accepted August 5th
- Advanced Power measurement and control capabilities
  - Directly impacted by Sandia's early work in this area
- Expands our ability to prototype
- Potential to design experiments at small scale and run at large scale (NERSC)

Going Forward



What will this look like?