Erich Strohmaier, Berkeley Lab & TOP500
Wu Feng, Virginia Tech & Green500
with
Natalie Bates, EE HPC Working Group
Nic Dube, HP & The Green Grid
Craig Steffen, NCSA
And others on the Compute System Metrics Team

SETTING TRENDS FOR ENERGY-EFFICIENT SUPERCOMPUTING

SC11 BoF; November 2011; Seattle
WHY WE ARE HERE

- Context
  - Power consumption and facility costs of HPC are increasing.
- “Can only improve what you can measure”
- What is needed?
  - Converge on a common basis for:
    - METHODOLOGIES
    - WORKLOADS
    - METRICS

for energy-efficient supercomputing, so we can make progress towards solutions.
AGREEMENT IN PRINCIPAL

- Collaboration between Top500, Green500, Green Grid and EE HPC WG
- Evaluate and improve methodology, metrics, and drive towards convergence on workloads
- Form a basis for evaluating energy efficiency of individual systems, product lines, architectures and vendors
- Target architecture design and procurement decision making process
Measure behavior of key system components including compute, memory, interconnect fabric, storage and external I/O

- Workloads and Metrics might address several components at the same time

- Phased implementation planned
PROPOSED WORKLOADS

- Leverage well-established benchmarks
- Must exercise the HPC system to the fullest capability possible
- Use High Performance LINPACK (HPL) for exercising (mostly) compute sub-system
- Use RandomAccess (Giga Updates Per second or GUPs) for exercising memory sub-system (?)
- Need to identify workloads for exercising other sub-systems
UNIFY AND IMPROVE METHODOLOGY

- HPL and RandomAccess measurement methodologies are well established
- Green500 & TOP500 power-measurement methodology
  - Similar, but not identical methodologies
- Issues/concerns with power-measurement methodology
  - Variation in start/stop times as well as sampling rates
  - Node, rack or system level measurements
  - What to include in the measurement (e.g., integrated cooling)
  - Need to increase vendor and/or supercomputing center power-measurement reporting
- June 2010 Green500 List – 56% of the list were submitted/measured numbers, the other 44% were derived by Green500
Current power measurement methodology is very flexible, but compromises consistency between submissions.

Proposal is to keep flexibility, but keep track of rules used and quality of power measurement.

3 Tiers of power measurement quality:
- Sampling rate; more measurements means higher quality
- Completeness of what is being measured; more of the system translates to higher quality
- Common rules for start/stop times
Metric and Methodology

General Definitions:
- **Metric**: a basis for comparison, a reference point against which other things can be evaluated; a measure
- **Methodology**: the system of methods followed in a particular discipline; a way of doing something, especially a systematic way (usually in steps); a measurement procedure

Specific Instantiations:
- **Metric**: Average Power consumed while running workload.
  - Others; peak instantaneous energy, total energy, etc.
- **Methodology**: Time fraction and granularity of run, fraction of the system measured, sub-systems included, how power/energy data is recorded, how average power is computed, and what data is required for reporting
  - Others; environmental conditions, measurement tool requirements, etc.
## Power/Energy Measurement Methodology (Current Proposed)

<table>
<thead>
<tr>
<th>Aspect 1: Time Fraction &amp; Granularity</th>
<th>Aspect 2: Machine Fraction</th>
<th>Aspect 3: Subsystems Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>20% of run: 1 average power measurement</td>
<td>(larger of) 1/64 of machine or 1kW</td>
</tr>
<tr>
<td>Level 2</td>
<td>100% of run: at least 100 average power measurements</td>
<td>(Larger of) 1/8 of machine or 10kW</td>
</tr>
<tr>
<td>Level 3</td>
<td>100% of run: at least 100 running total energy measurements</td>
<td>Whole machine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Why Report 100 Measurements?

Top-500 requires HPL output for performance verification.

Power/Energy Measurement Methodology defines data to be reported for power consumption verification.

– The data is used to verify the computation of average power and matched against total runtime as reported by the workload.
Experimental power data from LLNL Muir

Lawrence Livermore National Lab “Muir” Cluster
Dell Xanadu 3 Cluster, Xeon X5660 2.8 Ghz, QLogic InfiniBand QDR, 1242 nodes
LLNL data center instrument with power and energy meters at the panel board level
Data that follows is from a special HPL run just for power data \textit{(not optimized for performance)}
Data is total machine power every 1 second for the entire run including start and end idle time
HPL Power Profile on Muir (non-tuned)
How Is The Data Used to Compute Average Power?

L1: Average Power is recorded average power

L2: Average power is numerical average of instantaneous power measurements within the run window. Power data is used to verify start and stop times of run window and verify with benchmark time.

L3: Average power is difference of total energy at the end the run window minus the total energy at the beginning, divided by total run time. Energy data is used to determine the beginning and end of run window and verify with benchmark time.
Aspect 1 Time Fraction: Defining Run "Beginning"

![Graph showing power output over time with a highlighted start point labeled HPL timing start. The graph is labeled: Muir Power (test) run October 2011.](image-url)
Aspect 1 Time Fraction: Defining Run End

The runtime length obtained in this way agrees with HPL reported runtime within ~2 seconds.
Power drops from ~390kW to ~360kW over run, a variation of almost 8%
Aspect 1 Granularity: Why Is So Much Data Require For Reporting?

L2: Instantaneous power data is the result of interference between sampling, supply switching and input AC. High resolution maximizes the feature capture.

L3: Integrated total energy measurements have effectively infinite granularity. All power transients taken into account at the device level.
Aspect 1 Granularity Requirement Examples:
Sometimes Wide Variations or Sometimes Not
Aspect 1 Granularity Requirement: Highly Periodic Structure: Sampling Artifact

Muir Power (test) run October 2011
Aspect 1 Granularity Requirement: True Power Variations Over Run
Aspect 2 Machine Fraction: Why Are Larger Machine Fractions Preferred?

Measuring power/energy for a larger machine fractions decreases random power supply and measurement fluctuations.

Larger fractions reduces the effect of unusual (e.g., low/high power) nodes.

A larger machine fraction reduces magnifying instrument error by extrapolation.
Aspect 3: Sub-system List

No particular subsystem is required for any level. Subsystem inclusion is closely tied to power distribution, machine construction, and power instrumentation details.

Reporting of what subsystems are included in the power measurement is required to increase information and to inform future requirements.
Continued focus on improving power measurement methodology

Power/Energy Measurement Methodology Working paper ready for broader review, effective Nov’11

http://eehpcwg.lbl.gov/

“Quality level” implementation targeted for June 2012 Green500 and Top500 Lists
ISSUES TO RESOLVE

- Identify workloads for exercising other sub-systems; e.g., storage, I/O
- Still need to decide upon exact metric
  + Classes of systems (e.g., Top50, Little500)
  + Multiple metrics or a single index
  + Energy and power measurements
    + Average power, total energy, max instantaneous power
- Include non-workload consumption; idle time, job launch and shut down, etc.
- Environmental conditions?
WHY WE ARE HERE

- To gather community momentum and support
- To solicit your feedback
- To ask you to participate by submitting power measurements for Top500 and Green500 Lists
Energy Efficient HPC Working Group

EE HPC WG

- Driving energy conservation measures and energy efficient design in high performance computing.
- Demonstrate leadership in energy efficiency as in computing performance.
- Forum for sharing of information (peer-to-peer exchange) and collective action.

http://eehpcwg.lbl.gov
<table>
<thead>
<tr>
<th>Aspect 1: Time Fraction &amp; Granularity</th>
<th>Aspect 2: Machine Fraction</th>
<th>Aspect 3: Subsystems Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>(larger of) 1/64 of machine or 1kW</td>
<td>[Y] Compute nodes</td>
</tr>
<tr>
<td>20% of run: 1 average power measurement</td>
<td>(Larger of) 1/8 of machine or 10kW</td>
<td>[ ] Interconnect net</td>
</tr>
<tr>
<td>Level 2</td>
<td></td>
<td>[ ] Storage</td>
</tr>
<tr>
<td>Instantaneous power measurements every 1 second over whole run</td>
<td></td>
<td>[ ] Storage Network</td>
</tr>
<tr>
<td>Level 3</td>
<td>Whole machine</td>
<td>[ ] Login/Head nodes</td>
</tr>
<tr>
<td>100% of run: at least 100 running total energy measurements</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Data Center” Measurements
+ Power Usage Effectiveness (PUE)
  - PUE = Total Facility Energy/IT Energy
+ Energy Reuse Effectiveness (ERE)
  - ERE = (Total E.-Reused E.)/IT Energy

AND

“Compute System” Measurements
+ Workload/Productivity Metrics
+ Useful Work / Energy Consumed
STAKEHOLDERS AND PURPOSE?

- Stakeholders
  - **HPC computer system designers and procurement decision makers**, including users, data center and facilities designers/managers

- Purpose
  - **Unite the community behind energy efficiency metric(s) for HPC systems** that form the basis for comparing and evaluating individual systems, product lines, architectures and vendors
WHAT IS A METRIC?

- A basis for comparison
- A reference point against which other things can be evaluated
- A measure
WHAT MAKES A METRIC EFFECTIVE?

- Granular enough
  - Individual components
  - Analyzed in manageable chunks
  - Assigned to specific parties for improvement
- Intuitive, obvious, and clear
- Scientifically accurate and used precisely
- Sufficiently flexible to respond to new technology developments
- Vendor-neutral
- Inexpensive and worthwhile

**WHAT IS A WORKLOAD?**

- The application or benchmark software designed to exercise the HPC system to the fullest capability possible.
Use Workload-based Metrics to Represent HPC Energy Efficiency
+ Use workload-based for numerator and measured power during workload run for denominator

Examples
+ Green500 “FLOPS per Watt”
+ SPEC FP / Measured Watt
+ Green Grid “Productivity Proxies”

Still need to decide upon exact metric
+ Classes of systems (e.g., Top50, Little500)
+ Multiple metrics or a single index
  - Idle as well as full stress
WHAT IS A METHODOLOGY?

- The system of methods followed
- A way of doing something, especially a systematic way; implies an orderly logical arrangement (usually in steps)
- A measurement procedure