Making a Case for a Green500 List

S. Sharma†, C. Hsu†, and W. Feng‡

† Los Alamos National Laboratory
‡ Virginia Tech
Outline

- Introduction
  - What Is Performance?
  - Motivation: The Need for a Green500 List
- Challenges
  - What Metric To Choose?
  - Comparison of Available Metrics
- TOP500 as Green500
- Conclusion
What Is Performance?

TOP500 Supercomputer List

- Benchmark
  - LINPACK: Solves a (random) dense system of linear equations in double-precision (64 bits) arithmetic.
    - Introduced by Prof. Jack Dongarra, U. Tennessee

- Evaluation Metric
  - Performance (i.e., Speed)
    - Floating-Operations Per Second (FLOPS)

- Web Site
  - http://www.top500.org

- Next-Generation Benchmark: HPC Challenge
  - http://icl.cs.utk.edu/hpcc/

Performance, as defined by speed, is an important metric, but...
# Reliability & Availability of HPC

<table>
<thead>
<tr>
<th>Systems</th>
<th>CPUs</th>
<th>Reliability &amp; Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCI Q</td>
<td>8,192</td>
<td>MTBI: 6.5 hrs. 114 unplanned outages/month. HW outage sources: storage, CPU, memory.</td>
</tr>
<tr>
<td>NERSC Seaborg</td>
<td>6,656</td>
<td>MTBI: 14 days. MTTR: 3.3 hrs. SW is the main outage source. Availability: 98.74%</td>
</tr>
<tr>
<td>PSC Lemieux</td>
<td>3,016</td>
<td>MTBI: 9.7 hrs. Availability: 98.33%</td>
</tr>
<tr>
<td>Google (as of 2003)</td>
<td>~15,000</td>
<td>20 reboots/day; 2-3% machines replaced/year. HW outage sources: storage, memory. Availability: ~100%</td>
</tr>
</tbody>
</table>

MTBI: mean time between interrupts; MTBF: mean time between failures; MTTR: mean time to restore

Source: Daniel A. Reed, RENCI, 2004
Costs Associated with HPC

- **Infrastructure**
  - Sizable costs associated with system administration and maintenance. (People resources are $$$$.)
  - Massive construction and operational costs associated with powering and cooling.
    - Google
      - $2M to buy 30 acres of land by The Dalles Dam (Columbia River)
        - Inexpensive power to satisfy their high electrical demand.
        - Water can be used to cool its massive server-filled facility directly rather than relying on more expensive A/C.
    - Lawrence Livermore National Laboratory
      - Building for Terascale Simulation Facility: $55M
      - Electrical Costs: $14M/year to power and cool.

- **Productivity**
  - Downtime means no compute time, i.e., lost productivity.
Recent Trends in HPC

• Low(er)-Power Multi-Core Chipsets
  - **AMD**: Athlon64 X2 (2) and Opteron (2)
  - **ARM**: MPCore (4)
  - **IBM**: PowerPC 970 (2)
  - **Intel**: Smithfield (2) and Montecito (2)
  - **PA Semi**: PWRficent (2)

• Low-Power Supercomputing
  - **Green Destiny** (2002)
  - **Orion Multisystems** (2004)
  - **BlueGene/L** (2004)
  - **MegaProto** (2004)

---

**October 2003**
BG/L half rack prototype
500 Mhz
512 nodes/1024 proc.
2 TFlop/s peak
1.4 Tflop/s sustained
Perspective

● FLOPS Metric of the TOP500
  ❖ Performance = Speed (as measured in FLOPS with Linpack)
  ❖ May not be “fair” metric in light of recent low-power trends to help address reliability, availability, and total cost of ownership.

● The Need for a Different Performance Metric
  ❖ Performance = \( f(\text{speed, “time to answer”, power consumption, “up time”, total cost of ownership, usability, …}) \)
  ❖ Easier said than done …
    ❖ Many of the above dependent variables are difficult, if not impossible, to quantify, e.g., “time to answer”, TCO, usability, etc.

● The Need for a Green500 List
  ❖ Performance = \( f(\text{speed, power consumption}) \) as speed and power consumption can be quantified.
Outline

- Introduction
  - What Is Performance?
  - Motivation: The Need for a Green500 List
- Challenges
  - What Metric To Choose?
  - Comparison of Available Metrics
- TOP500 as Green500
- Conclusion
Challenges for a Green500 List

- **What Metric To Choose?**
  - $ED^n$: Energy-Delay Products, where $n$ is a non-negative int. (borrowed from the circuit-design domain)
  - Variants of $ED^n$
  - Speed / Power Consumed
    - FLOPS / Watt, MIPS / Watt, and so on

- **What To Measure?** Obviously, energy or power ... but
  - Energy (Power) consumed by the computing system?
  - Energy (Power) consumed by the processor?
  - Temperature at specific points on the processor die?

- **How To Measure Chosen Metric?**
  - Power meter? But attached to what? At what time granularity should the measurement be made?
ED$^n$: Energy-Delay Products

- **Original Application**
  - Circuit Design

- **Problem**
  - For $n \geq 1$, the metric is biased towards systems with a larger number of processors as the “delay component” (i.e., aggregate speed) dominates.
  - As $n$ increases, the bias towards aggregate speed, and hence, HPC systems with larger numbers of processors, increases dramatically.
Variants of $ED^n$ : $V_\delta = E^{(1-\delta)} D^{2(1+\delta)}$

- Negative values of $\delta$ (particularly more negative values) marginally offset the bias of the $ED^n$ towards speed.
  - In our benchmarking, they produced identical rankings to the $ED^n$ metric.

- Positive values of $\delta$ place greater emphasis on performance.
  - As $\delta$ increases towards one, the metric approaches the limit $E^0 D^4$ and behaves more like the standard FLOPS metric, which is used for TOP500 List.

**Metric of Choice: FLOPS / Watt (again)**
Outline

- Introduction
  - What Is Performance?
  - Motivation: The Need for a Green500 List
- Challenges
  - What Metric To Choose?
  - Comparison of Available Metrics
- TOP500 as Green500
- Conclusion

W. Feng, feng@cs.vt.edu, (540) 231-1192
## Efficiency of Four-CPU Clusters

<table>
<thead>
<tr>
<th>Name</th>
<th>CPU</th>
<th>LINPACK (Gflops)</th>
<th>Avg Pwr (Watts)</th>
<th>Time (s)</th>
<th>ED (*10^6)</th>
<th>ED2 (*10^9)</th>
<th>Flops/W</th>
<th>V_{\psi=-0.5}</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>P4 3.6G</td>
<td>19.55</td>
<td>713.2</td>
<td>315.8</td>
<td>71.1</td>
<td>22.5</td>
<td>27.4</td>
<td>33.9</td>
</tr>
<tr>
<td>C2</td>
<td>Opt 2.0G</td>
<td>12.37</td>
<td>415.9</td>
<td>499.4</td>
<td>103.7</td>
<td>51.8</td>
<td>29.7</td>
<td>47.2</td>
</tr>
<tr>
<td>C3</td>
<td>Ath64 2.4G</td>
<td>14.31</td>
<td>668.5</td>
<td>431.6</td>
<td>124.5</td>
<td>53.7</td>
<td>21.4</td>
<td>66.9</td>
</tr>
<tr>
<td>C4</td>
<td>Ath64 2.2G</td>
<td>13.40</td>
<td>608.5</td>
<td>460.9</td>
<td>129.3</td>
<td>59.6</td>
<td>22.0</td>
<td>68.5</td>
</tr>
<tr>
<td>C5</td>
<td>Ath64 2.0G</td>
<td>12.35</td>
<td>560.5</td>
<td>499.8</td>
<td>140.0</td>
<td>70.0</td>
<td>22.0</td>
<td>74.1</td>
</tr>
<tr>
<td>C6</td>
<td>Opt 2.0G</td>
<td>12.84</td>
<td>615.3</td>
<td>481.0</td>
<td>142.4</td>
<td>64.5</td>
<td>20.9</td>
<td>77.4</td>
</tr>
<tr>
<td>C7</td>
<td>Ath64 1.8G</td>
<td>11.23</td>
<td>520.9</td>
<td>549.9</td>
<td>157.5</td>
<td>86.6</td>
<td>21.6</td>
<td>84.3</td>
</tr>
</tbody>
</table>
# Green500 Ranking of Four-CPU Clusters

<table>
<thead>
<tr>
<th>Rank</th>
<th>ED</th>
<th>ED²</th>
<th>ED³</th>
<th>V_{\leq-0.5}</th>
<th>V_{\geq0.5}</th>
<th>FLOPS/Watt</th>
<th>TOP 500</th>
<th>Power 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C1</td>
<td>C1</td>
<td>C1</td>
<td>C1</td>
<td>C1</td>
<td>C2</td>
<td>C1</td>
<td>C2</td>
</tr>
<tr>
<td>2</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
<td>C3</td>
<td>C1</td>
<td>C3</td>
<td>C7</td>
</tr>
<tr>
<td>3</td>
<td>C3</td>
<td>C3</td>
<td>C3</td>
<td>C3</td>
<td>C4</td>
<td>C5</td>
<td>C4</td>
<td>C5</td>
</tr>
<tr>
<td>4</td>
<td>C4</td>
<td>C4</td>
<td>C4</td>
<td>C4</td>
<td>C2</td>
<td>C4</td>
<td>C6</td>
<td>C4</td>
</tr>
<tr>
<td>5</td>
<td>C5</td>
<td>C5</td>
<td>C5</td>
<td>C5</td>
<td>C5</td>
<td>C7</td>
<td>C2</td>
<td>C6</td>
</tr>
<tr>
<td>6</td>
<td>C6</td>
<td>C6</td>
<td>C6</td>
<td>C6</td>
<td>C6</td>
<td>C3</td>
<td>C5</td>
<td>C3</td>
</tr>
<tr>
<td>7</td>
<td>C7</td>
<td>C7</td>
<td>C7</td>
<td>C7</td>
<td>C7</td>
<td>C6</td>
<td>C7</td>
<td>C1</td>
</tr>
</tbody>
</table>
Outline

- Introduction
  - What Is Performance?
  - Motivation: The Need for a Green500 List
- Challenges
  - What Metric To Choose?
  - Comparison of Available Metrics
- TOP500 as Green500
- Conclusion
## TOP500 Power Usage

<table>
<thead>
<tr>
<th>Name</th>
<th>Linpack</th>
<th>MFLOPS/W</th>
<th>TOP500 Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlueGene/L</td>
<td>367,000</td>
<td>146.80</td>
<td>1</td>
</tr>
<tr>
<td>ASC Purple</td>
<td>77,824</td>
<td>10.24</td>
<td>3</td>
</tr>
<tr>
<td>Columbia</td>
<td>60,960</td>
<td>17.93</td>
<td>4</td>
</tr>
<tr>
<td>Earth Simulator</td>
<td>40,960</td>
<td>3.44</td>
<td>7</td>
</tr>
<tr>
<td>MareNostrum</td>
<td>42,144</td>
<td>39.35</td>
<td>8</td>
</tr>
<tr>
<td>Jaguar-Cray XT3</td>
<td>24,960</td>
<td>18.75</td>
<td>10</td>
</tr>
<tr>
<td>ASC Q</td>
<td>20,480</td>
<td>2.01</td>
<td>18</td>
</tr>
<tr>
<td>ASC White</td>
<td>12,288</td>
<td>6.02</td>
<td>47</td>
</tr>
</tbody>
</table>
# TOP500 as Green500

<table>
<thead>
<tr>
<th>Relative Rank</th>
<th>TOP500</th>
<th>Green500</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BlueGene/L (IBM)</td>
<td>BlueGene/L (IBM)</td>
</tr>
<tr>
<td>2</td>
<td>ASC Purple (IBM)</td>
<td>MareNostrum (IBM)</td>
</tr>
<tr>
<td>3</td>
<td>Columbia (SGI)</td>
<td>Jaguar-Cray XT3 (Cray)</td>
</tr>
<tr>
<td>4</td>
<td>Earth Simulator (NEC)</td>
<td>Columbia (SGI)</td>
</tr>
<tr>
<td>5</td>
<td>MareNostrum (IBM)</td>
<td>ASC Purple (IBM)</td>
</tr>
<tr>
<td>6</td>
<td>Jaguar-Cray XT3 (Cray)</td>
<td>ASC White (IBM)</td>
</tr>
<tr>
<td>7</td>
<td>ASC Q (HP)</td>
<td>Earth Simulator (NEC)</td>
</tr>
<tr>
<td>8</td>
<td>ASC White (IBM)</td>
<td>ASC Q (HP)</td>
</tr>
</tbody>
</table>

W. Feng, feng@cs.vt.edu, (540) 231-1192
Conclusion

- **Metrics for a Green500 Supercomputer List**
  - Still no definitive metric to use
    - By process of elimination, we converged on FLOPS/watt, which is relatively easy to derive from the TOP500 Supercomputer List.
  - Insight with respect to current metrics
  - Insight with respect to when to use processor energy (or power) versus system energy (or power)

- **TOP500 as Green500**
  - From the data presented, IBM and Cray make the most energy-efficient HPC systems today.
For More Information

- Visit “Supercomputing in Small Spaces” at http://sss.lanl.gov
  - Soon to be re-located to Virginia Tech

- Contact Wu-chun Feng (a.k.a. “Wu”)
  - E-mail: feng@cs.vt.edu
  - Phone: (540) 231-1192
  - Mailing Address
    - 2200 Kraft Drive
    - Department of Computer Science
    - Virginia Tech
    - Blacksburg, VA 24060