Color Me Green!  Color Me Fast!

Prof. Wu FENG, feng@cs.vt.edu
Depts. of Computer Science and Electrical & Computer Engineering
Talk Forecast

• An introspective look at *green supercomputing*
  … where we have been … where we are …
  … and what the future may hold
  *The need for supercomputers to be colored green and fast*

• “Supercomputing in Small Spaces” Project (circa 2001)
  – Autonomic energy and power savings *while* maintaining performance in high-end computing systems.
    • Power-Aware Approach
      – At Integration Time → Green Destiny (2001)
      – At Run Time → β Algorithm (2004)
    • Energy-Efficient Approach: Accelerators?

• The Green500 List
  [http://www.green500.org/](http://www.green500.org/)
The Best Place to Start is …

- Venue: IEEE IPDPS Keynote Talk in 2001
  Speaker: David A. Patterson, UC Berkeley
  Keynote Talk: *New Challenges of the Post-PC Era*
  - We have spent decades focusing on performance, performance, and price/performance.

Hmmm …

*Can’t the same statement be made for high-end computing (HEC)*?
We have spent decades focusing on performance, performance, performance (and price/performance).
The “Cost” of Achieving Performance

- **Electrical power costs $$$$**.

*Worldwide Expense to Power and Cool the Server Installed Base, 1996–2010*

Source: IDC, 2006
The “Cost” of Achieving Performance

• Too much power affects efficiency, reliability, and availability.
  – Anecdotal Evidence from a “Machine Room” in 2001 - 2002
    • **Winter**: “Machine Room” Temperature of **70-75º F**
      – Failure approximately *once* per week.
    • **Summer**: “Machine Room” Temperature of **85-90º F**
      – Failure approximately *twice* per week.
  – **Arrenhius’ Equation** (applied to microelectronics)
    • *For every 10º C (18º F) increase in temperature, … the failure rate of a system doubles.*
Efficiency of HEC Systems

• “Performance” and “Price/Performance” Metrics …
  – Lower efficiency, reliability, and availability.
  – Higher operational costs, e.g., admin, maintenance, etc.

• Examples
  – Computational Efficiency
    • Relative to Peak: Actual Performance/Peak Performance
    • Relative to Space: Performance/Sq. Ft.
    • Relative to Power: Performance/Watt
  – Performance: 10,000-fold increase (since the Cray C90)
    • Performance/Sq. Ft.: Only 65-fold increase.
    • Performance/Watt: Only 300-fold increase.
  – Massive construction and operational costs associated with powering and cooling.
Outline

• Motivation & Background
  – The “Cost” of Achieving Performance
  – The Need for Efficiency, Reliability, and Availability
    • The Need to Color Green and Fast …

• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
    • Architectural: MegaScale, Orion Multisystems, IBM Blue Gene/L
    • Software-Based: $\beta$ adaptation and EcoDaemon™

• The Green500 List

• Conclusion and Future Work
Supercomputing in Small Spaces

• **Goals**
  - Improve efficiency, reliability, and availability in large-scale supercomputing systems (HEC, server farms, distributed systems)
    • Autonomic energy and power savings *while* maintaining performance in a large-scale computing system.
  - Reduce the total cost of ownership (TCO).
    • TCO = Acquisition Cost ("one time") + Operation Cost (recurring)

• **Crude Analogy**
  - Today’s Supercomputer
    • Formula One Race Car: Wins raw performance but reliability is so poor that it requires frequent maintenance. Throughput low.
  - “Supercomputing in Small Spaces” Supercomputer
    • BMW 325i: Loses raw performance but higher reliability and efficiency results in high throughput (i.e., miles driven/month → answers/month).

http://sss.cs.vt.edu/
Green Destiny Supercomputer
(circa December 2001 – February 2002)

• A 240-Node Cluster in Five Sq. Ft.
• Each Node
  – 1-GHz Transmeta TM5800 CPU w/ High-Performance Code-Morphing Software running Linux 2.4.x
  – 640-MB RAM, 20-GB hard disk, 100-Mb/s Ethernet
• Total
  – 240 Gflops peak (Linpack: 101 Gflops in March 2002.)
  – 150 GB of RAM (expandable to 276 GB)
  – 4.8 TB of storage (expandable to 38.4 TB)
  – Power Consumption: Only 3.2 kW (diskless)
• Reliability & Availability
  – No unscheduled downtime in 24-month lifetime.
    • Environment: A dusty 85°-90° F warehouse!

Now in the Computer History Museum.
Transmeta TM5600 CPU: VLIW + CMS

• VLIW Engine
  - Up to four-way issue
    • In-order execution only.
    • 20% reduction on transistor count w.r.t superscalar arch.
  - Two integer units
  - Floating-point unit
  - Memory unit
  - Branch unit

• VLIW Transistor Count (“Anti-Moore’s Law”)
  - 10M transistors \(\rightarrow\) \(\sim 6x-7x\) less power dissipation than CPUs at the time
  - Less power \(\rightarrow\) lower “on-die” temp. \(\rightarrow\) better reliability & availability

• This “colors me green” but what about \textit{fast}?
Transmeta TM5x00 CMS

• Code Morphing Software (CMS)
  – Provides compatibility by dynamically “morphing” x86 instructions into simple VLIW instructions.
  – *Learns and improves with time, i.e., iterative execution.*

• Modules for CMS
  – Interpreter
    • Interprets x86 instructions (*a la* Java).
    • Filters infrequently executed code from being optimized.
    • Collects run-time statistical information.
  – Translator
    • Re-compiles x86 instructions into optimized VLIW instructions (*a la* JIT compiler).
Green Destiny Supercomputer
(circa December 2001 – February 2002)

• A 240-Node Cluster in Five Square Feet
• Each Node
  – 1-GHz Transmeta TM5800 CPU with High-Performance Code
    Morphing Software running Linux 2.4.x
  – 640-MB RAM, 20-GB hard disk, 100-Mb/s Ethernet

• Total
  – 240 Gflops peak (Linpack: 101 Gflops in March 2002.)
  – 150 GB of RAM (expandable to 276 GB)
  – 4.8 TB of storage (expandable to 38.4 TB)
  – Power Consumption: Only 3.2 kW (diskless)

• Reliability & Availability
  – No unscheduled downtime in 24-month lifetime.
    • Environment: A dusty 85°-90° F warehouse!

Goals of the Project
• Improve efficiency, reliability and availability in supercomputers
• Reduce the total cost of ownership

Now in the Computer History Museum.
Green Destiny: Low-Power Supercomputer

Green Destiny “Imitation”: Traditional Supercomputer

Only Difference? The Processors

© J. Gans, Mar. 2007
## Parallel Computing Platforms Running an N-body Gravitational Code

(Apples-to-Oranges Comparison)

<table>
<thead>
<tr>
<th>Machine</th>
<th>Avalon Beowulf</th>
<th>ASCI Red</th>
<th>ASCI White</th>
<th>Green Destiny</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>1996</td>
<td>1996</td>
<td>2000</td>
<td>2002</td>
</tr>
<tr>
<td>Performance (Gflops)</td>
<td>18</td>
<td>600</td>
<td>2500</td>
<td>58</td>
</tr>
<tr>
<td>Area (ft²)</td>
<td>120</td>
<td>1600</td>
<td>9920</td>
<td>5</td>
</tr>
<tr>
<td>Power (kW)</td>
<td>18</td>
<td>1200</td>
<td>2000</td>
<td>5</td>
</tr>
<tr>
<td>DRAM (GB)</td>
<td>36</td>
<td>585</td>
<td>6200</td>
<td>150</td>
</tr>
<tr>
<td>Disk (TB)</td>
<td>0.4</td>
<td>2.0</td>
<td>160.0</td>
<td>4.8</td>
</tr>
<tr>
<td>DRAM density (MB/ft²)</td>
<td>300</td>
<td>366</td>
<td>625</td>
<td>3000</td>
</tr>
<tr>
<td>Disk density (GB/ft²)</td>
<td>3.3</td>
<td>1.3</td>
<td>16.1</td>
<td>960.0</td>
</tr>
<tr>
<td>Perf/Space (Mflops/ft²)</td>
<td>150</td>
<td>375</td>
<td>252</td>
<td>11600</td>
</tr>
<tr>
<td>Perf/Power (Mflops/watt)</td>
<td>1.0</td>
<td>0.5</td>
<td>1.3</td>
<td>11.6</td>
</tr>
</tbody>
</table>
Yet in 2002 …

- “Green Destiny is so low power that it runs just as fast when it is unplugged.”
- “The slew of expletives and exclamations that followed Feng’s description of the system …”
- “In HPC, no one cares about power & cooling, and no one ever will …”
- “Moore’s Law for Power will stimulate the economy by creating a new market in cooling technologies.”

For many of the Los Alamos scientists, the unveiling of Green Destiny was their first introduction to blade servers -- never mind blade servers being used to build a supercomputer. The slew of expletives and exclamations that followed Feng's description of the system made it clear that the blades had captured the audience's attention. Some murmured, "Wow," while others let out multiple shouts of, "Jesus!" as their jaws dropped.

Several scientists here did not share the enthusiasm for Green Destiny, however. Los Alamos, after all, is the home to several massive supercomputers that take up entire floors of buildings and require several cooling systems shaped like mini-nuclear reactors to keep them running. These "real" supercomputers handle serious work, and some of the people running them consider Green Destiny a joke. One scientist walked out of Feng's presentation, making his feelings clear.
Outline

• Motivation & Background
  – The “Cost” of Achieving Performance
  – The Need for Efficiency, Reliability, and Availability
    • The Need to Color Green and Fast …

• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
    • Architectural: MegaScale, Orion Multisystems, IBM Blue Gene/L
    • Software-Based: $\beta$ adaptation and EcoDaemon™

• The Green500 List

• Conclusion and Future Work
Cluster Technology

Low-Power Systems Design

Linux

But in the form factor of a workstation ... a cluster workstation
Color Me Green!  Color Me Fast!

ORION DT-12 DESKTOP CLUSTER WORKSTATION

Imagine a 36 Gflop cluster on your desk!

- LINPACK Performance
  - 14 Gflops
- Footprint
  - 3 sq. ft. \((24" \times 18")\)
  - 1 cu. ft. \((24" \times 4" \times 18")\)
- Power Consumption
  - 170 watts at load
- How did this compare with a traditional desktop?

12 Nodes
in a single computer

36 Gflops
peak processing power

24 GBytes
memory capacity

1 TByte
internal storage

DESIGNED FOR THE INDIVIDUAL
The Orion DT-12 cluster workstation is a fully integrated, completely self-contained, personal workstation based on the best of today’s cluster technologies. Designed to be an affordable individual resource it is capable of 36 Gflops peak performance (18 Gflops sustained) with nodes starting at under $10k.

The Orion DT-12 cluster workstation provides supercomputer performance for the engineering, scientific, financial and creative professionals who need to solve computationally complex problems without waiting in the queue of the back-room cluster.

FASTER SOFTWARE DEVELOPMENT
The Orion DT-12 cluster workstation is the perfect platform for developers writing and deploying cluster software packages. It comes with cluster software development tools pre-installed, including libraries and a parallel compiler that allows you to spread one multiple-file compile to all the nodes in the system. Also included is a suite of system monitoring and management software.

NO ASSEMBLY REQUIRED
Orion workstations are built around industry standards for clustering: x86 processors, Ethernet, the Linux operating system and standard parallel programming libraries, including MPI, PVM and SGE. Existing Linux cluster applications run without modification.

PERFORMANCE AND FEATURES
The Orion DT-12 is a cluster of 12 x86 compatible nodes linked by a switched Gigabit Ethernet fabric. The cluster operates as a single computer with a single on-off switch and a single system image rapid boot sequence, which allows the entire system to boot in less than 90 seconds.

The Orion DT-12 cluster workstation is highly efficient, consuming a maximum of 120 Watts of power under peak load—about the same as an average desktop PC. It operates quietly, plugs into a standard 110V 15A wall socket and fits unobtrusively on a desk or lab bench.
Color Me Green!  Color Me Fast!

© W. Feng, April 2009

2009 Conference on High-End Computing, Salishan, OR.

feng@cs.vt.edu

ORION DS-96 DESKSIDE CLUSTER WORKSTATION

Imagine a 300 Gflop cluster... under your desk.

96 Nodes
in a single computer

300 Gflops
peak processing power

192 GBytes
memory capacity

9.6 TBytes
internal storage

INCREASE YOUR PRODUCTIVITY
The Orion DS-96 cluster workstation is the highest performance general-purpose computing platform that can be plugged into a standard wall outlet and operated in an office or laboratory environment.

PREVENT SOFTWARE INVESTMENTS
Orion workstations are built around industry standards for clustering: x86 processors, the Linux operating system and standard parallel programming libraries, including MPI, PVM and SGE. Your existing Linux cluster software applications can run without modification.

NO ASSEMBLY REQUIRED
Orion workstations are designed from the ground up as a single computer. The entire system boots with the push of a button and has the ergonomics and ease of use of a personal computer. Modular, solid state design allows for flexible configurations and scalability.

PERFORMANCE AND FEATURES
The Orion DS-96 cluster workstation is a fully integrated, completely self-contained personal workstation based on the best of today’s cluster technologies and commodity components. Designed to be an individual or departmental resource, it is capable of 300 Gflops peak performance (150 Gflops sustained). The DS-96 is also highly efficient, consuming a maximum of 1500 Watts of power under peak load. It operates quietly, plugs into a standard 110V 15A wall socket, and fits unobtrusively beneath a desk or lab bench.

The DS-96 is a cluster of 96 x86-compatible nodes linked by an integrated Gigabit Ethernet fabric. The cluster operates as a single computer, with a single on-off switch, and a single-system-image rapid boot sequence which allows the entire system to boot in less than 2 minutes. The DS-96 comes with standard Linux and drivers pre-installed, including an optimized MPI message-passing library. Also included is a suite of cluster software development tools, system monitoring and system management software.

• LINPACK Performance
  – 111 Gflops

• Footprint
  – 3 sq. ft.  (17” x 25”)
  – 6 cu. ft.  (17” x 25” x 25”)

• Power Consumption
  – 1580 watts at load

Recall ....

GD: 101 Gflops
Inter-University Project: MegaScale
http://www.para.tutics.tut.ac.jp/megascale/

Univ. of Tsukuba Booth @ SC2004, Nov. 2004.
IBM Blue Gene/L

- **Chip** (2 processors)
  - 2.8/5.6 GF/s
  - 4 MB
- **Compute Card** (2 chips, 2x1x1)
  - 5.6/11.2 GF/s
  - 0.5 GB DDR
- **Node Card** (32 chips, 4x4x2)
  - 16 Compute Cards
  - 90/180 GF/s
  - 8 GB
- **System** (64 cabinets, 64x32x32)
  - Cabinet
    - (32 Node boards, 8x8x16)
  - 2.9/5.7 TF/s
  - 256 GB DDR

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

© 2004 IBM Corporation
Problem: A Misunderstood Message?

• MegaScale, Orion Multisystems, and IBM “got it” … but what about others?

• “Lost in Translation”
  – Their Takeaway: Blades are the solution.
  – “Let’s take existing motherboards, turn them on their side, and push them together.”
  – Problem
    • High Power Density
    • Lower Reliability and Availability
Reliability & Availability of a Beowulf Cluster

• 18-Node Bladed Beowulf Cluster
  – 80-85° F environment
    • “Silent” failure on Linpack benchmark.
    • 1/3 of nodes became inaccessible.
  – 65° F environment
    • Correct running of Linpack benchmark.

… which led to
Outline

• Motivation & Background
  – The “Cost” of Achieving Performance
  – The Need for Efficiency, Reliability, and Availability
    • The Need to Color Green and Fast …

• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
    • Architectural: MegaScale, Orion Multisystems, IBM Blue Gene/L
    • Software-Based: β adaptation and EcoDaemon™

• The Green500 List

• Conclusion and Future Work
Software-Based Approach:

Power-Aware Computing

\( \beta \) Adaptation Algorithm

- Self-Adapting Software for Energy Efficiency
  Conserve power & energy \textbf{WHILE} maintaining performance.

• Observations
  - Many commodity technologies support \textit{dynamic voltage & frequency scaling (DVFS)}, which allows changes to the processor voltage and frequency at run time.
  - A computing system can trade off processor performance for power reduction.
    • \textit{Power} \( \alpha V^2 f \), where V is the supply voltage of the processor and f is its frequency.
    • \textit{Processor performance} \( \alpha f \).
Power-Aware HPC via DVFS:

Key Observation

- Execution time of many programs is insensitive to CPU speed change.

![Graph showing the relationship between CPU speed and execution time. The graph indicates that as CPU speed is slowed down by 50%, performance is degraded by 4%. The y-axis represents execution time in percentage, ranging from 100 to 110, and the x-axis represents CPU speed in GHz, ranging from 0.6 to 2.2. The graph includes a point marked as 'Clock speed slowed down by 50%' and another marked as 'Performance degraded by 4%'.]
Power-Aware HPC via DVFS:

**Key Idea**

- Applying DVFS to these programs will result in significant power and energy savings at a minimal performance impact.
Why is Power Awareness via DVFS Hard?

- What is cycle time of a processor?
  - Frequency ≈ 2 GHz → Cycle Time ≈ 1 / (2 x 10^9) = 0.5 ns
- How long does the system take to scale voltage and frequency?

O (10,000,000 cycles)
Problem Formulation:

**LP-Based Energy-Optimal DVFS Schedule**

- **Definitions**
  - A DVFS system exports \( n \) \( \{ (f_i, P_i) \} \) settings.
  - \( T_i \): total execution time of a program running at setting \( i \)

- **Given a program with deadline \( D \), find a DVS schedule \((t_1^*, \ldots, t_n^*)\) such that**
  - If the program is executed for \( t_i \) seconds at setting \( i \), the total energy usage \( E \) is minimized, the deadline \( D \) is met, and the required work is completed.

\[
\min E = \sum_i P_i \cdot t_i
\]

subject to

\[
\sum_i t_i \leq D
\]
\[
\sum_i t_i / T_i = 1
\]
\[
t_i \geq 0
\]
Single-Coefficient $\beta$ Performance Model

- **Our Formulation**
  - Define the relative performance slowdown $\delta$ as
    \[
    \frac{T(f)}{T(f_{\text{MAX}})} - 1
    \]
  - Re-formulate two-coefficient model as a single-coefficient model:
    \[
    \frac{T(f)}{T(f_{\text{MAX}})} = \beta \cdot \frac{f_{\text{MAX}}}{f} + (1 - \beta)
    \]
    with
    \[
    \beta = \frac{W_{\text{cpu}}}{W_{\text{cpu}} + T_{\text{mem}} \cdot f_{\text{MAX}}}
    \]
  - The coefficient $\beta$ is computed at run-time using a regression method on the past MIPS rates reported from the built-in PMU.

\[
\beta = \frac{\sum_i ((f_{\text{MAX}} / f_i - 1)(\text{mips}(f_{\text{MAX}}) / \text{mips}(f_i) - 1))}{\sum_i ((f_{\text{MAX}} / f_i - 1)^2}
\]

- C. Hsu and W. Feng.
  “A Power-Aware Run-Time System for High-Performance Computing,”
  *SC'05*, Nov. 2005.
β - Adaptation DVFS Scheduling Algorithm

**Input:** Relative slowdown \( \delta \) and performance model \( T(f) \).

**Output:** Constraint-based DVFS schedule.

For every \( I \) seconds do

- Compute coefficient \( \beta \)
- Compute the desired frequency \( f^* \)
  - If \( f^* \) is not a supported frequency, then
    - Identify \( f_j \) and \( f_{j+1} \).
    - Compute the ratio \( r \).
    - Run \( r \cdot I \) seconds at frequency \( f_j \).
    - Run \((1-r) \cdot I\) seconds at frequency \( f_{j+1} \).
    - Update \( \text{mips}(f_j) \) and \( \text{mips}(f_{j+1}) \).
  - Else run at \( f^* \).
Current DVFS Scheduling Algorithms

• **2step (i.e., CPUSPEED via SpeedStep):**
  - Using a dual-speed CPU, monitor CPU utilization periodically.
  - If utilization > pre-defined upper threshold, set CPU to fastest; if utilization < pre-defined lower threshold, set CPU to slowest.

• **nqPID:** A refinement of the 2step algorithm.
  - Recognize the similarity of DVFS scheduling and a classical control-systems problem → Modify a PID controller (Proportional-Integral-Derivative) to suit DVFS scheduling problem.

• **freq:** Reclaims the slack time between the actual processing time and the worst-case execution time.
  - Track the amount of remaining CPU work $W_{\text{left}}$ and the amount of remaining time before the deadline $T_{\text{left}}$.
  - Set desired CPU frequency at each interval to $f_{\text{new}} = W_{\text{left}} / T_{\text{left}}$.
  - The algorithm assumes that the total amount of work in CPU cycles is known *a priori*, which, in practice, is often unpredictable and not always a constant across frequencies.
### β - Adaptation on Sequential Codes (SPEC)

<table>
<thead>
<tr>
<th>program</th>
<th>β</th>
<th>2step</th>
<th>nqPID</th>
<th>freq</th>
<th>mips</th>
<th>beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>swim</td>
<td>0.02</td>
<td>1.00/1.00</td>
<td>1.04/0.70</td>
<td>1.00/0.96</td>
<td>1.00/1.00</td>
<td>1.04/0.61</td>
</tr>
<tr>
<td>tomcatv</td>
<td>0.24</td>
<td>1.00/1.00</td>
<td>1.03/0.69</td>
<td>1.00/0.97</td>
<td>1.03/0.83</td>
<td>1.00/0.85</td>
</tr>
<tr>
<td>su2cor</td>
<td>0.27</td>
<td>0.99/0.99</td>
<td>1.05/0.70</td>
<td>1.00/0.95</td>
<td>1.01/0.96</td>
<td>1.03/0.85</td>
</tr>
<tr>
<td>compress</td>
<td>0.37</td>
<td>1.02/1.02</td>
<td>1.13/0.75</td>
<td>1.02/0.97</td>
<td>1.05/0.92</td>
<td>1.01/0.95</td>
</tr>
<tr>
<td>mgrid</td>
<td>0.51</td>
<td>1.00/1.00</td>
<td>1.18/0.77</td>
<td>1.01/0.97</td>
<td>1.00/1.00</td>
<td>1.03/0.89</td>
</tr>
<tr>
<td>vortex</td>
<td>0.65</td>
<td>1.01/1.00</td>
<td>1.25/0.81</td>
<td>1.01/0.97</td>
<td>1.07/0.94</td>
<td>1.05/0.90</td>
</tr>
<tr>
<td>turb3d</td>
<td>0.79</td>
<td>1.00/1.00</td>
<td>1.29/0.83</td>
<td>1.03/0.97</td>
<td>1.01/1.00</td>
<td>1.05/0.94</td>
</tr>
<tr>
<td>go</td>
<td>1.00</td>
<td>1.00/1.00</td>
<td>1.37/0.88</td>
<td>1.02/0.99</td>
<td>0.99/0.99</td>
<td>1.06/0.96</td>
</tr>
</tbody>
</table>

*relative time / relative energy*

with respect to total execution time and system energy usage
$\beta$ - Adaptation on Sequential Codes (SPECjbb)

<table>
<thead>
<tr>
<th>Power Management</th>
<th>Watts</th>
<th>% Power Reduction</th>
<th>bops/watt</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>264</td>
<td>0%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Cpuspeed</td>
<td>257</td>
<td>3%</td>
<td>102.56%</td>
</tr>
<tr>
<td>Ondemand</td>
<td>253</td>
<td>4%</td>
<td>104.37%</td>
</tr>
<tr>
<td>$\beta$</td>
<td>196</td>
<td>25%</td>
<td>123.70%</td>
</tr>
</tbody>
</table>

Source: EnergyWare, 2009
NAS Parallel on an Opteron Cluster

NAS/NPB3.2-MPI, C.16

Slowdown  Savings

Energy reduction (15%)  Performance improvement

© W. Feng, April 2009  2009 Conference on High-End Computing, Salishan, OR.
EcoDaemon: Next-Gen Power-Aware Software
– Recent Winner of the SURA IP2M Competition

Energy Savings

Performance Loss
The Green GPGPU?

- Large-scale biomolecular electrostatic potential
  - Long-range interactions.

<table>
<thead>
<tr>
<th>Platform</th>
<th>E * D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial</td>
<td>409.1</td>
</tr>
<tr>
<td>Multi</td>
<td>127.1</td>
</tr>
<tr>
<td>GPU</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Outline

• Motivation & Background
  – The “Cost” of Achieving Performance
  – The Need for Efficiency, Reliability, and Availability
    • The Need to Color Green and Fast …

• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
    • Architectural: MegaScale, Orion Multisystems, IBM Blue Gene/L
    • Software-Based: $\beta$ adaptation and EcoDaemon™

• The Green500 List

• Conclusion and Future Work
The Growth of The **Green500** List

- From 0 to *10,000* page hits per month (on average)

- Number of reported sites increased by 20%

- The **Green500** List BoF
  - SRO
  - Its importance led to filming of the session
Green500 Initial Observations

• Surpassed the 500 Mflops/W milestone

• Over 50% of the top supercomputers are now reporting their energy usage.

• 18% more hardware manufacturers participating!
The Top of the **Green500** List

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
<th>Power (kW)</th>
<th>Mflops / W</th>
<th>TOP500 Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U. Warsaw</td>
<td>BladeCenter QS22</td>
<td>35</td>
<td>536</td>
<td>221</td>
</tr>
<tr>
<td>2</td>
<td>Repsol YPF</td>
<td>BladeCenter QS22</td>
<td>26</td>
<td>530</td>
<td>430</td>
</tr>
<tr>
<td>2</td>
<td>Repsol YPF</td>
<td>BladeCenter QS22</td>
<td>26</td>
<td>530</td>
<td>431</td>
</tr>
<tr>
<td>2</td>
<td>Repsol YPF</td>
<td>BladeCenter QS22</td>
<td>26</td>
<td>530</td>
<td>41</td>
</tr>
<tr>
<td>5</td>
<td>DOE/NNSA/LANL</td>
<td>BladeCenter QS22/LS21</td>
<td>138</td>
<td>458</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>IBM Poughkeepsie</td>
<td>BladeCenter QS22/LS21</td>
<td>138</td>
<td>458</td>
<td>73</td>
</tr>
<tr>
<td>7</td>
<td>DOE/NNSA/LANL</td>
<td>BladeCenter QS22/LS21</td>
<td>2483</td>
<td>445</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>U. Groningen</td>
<td>Blue Gene/P</td>
<td>95</td>
<td>372</td>
<td>76</td>
</tr>
</tbody>
</table>
Closing Thoughts

• The Green500 List is a work in progress … with the ultimate goal of raising awareness in energy efficiency in supercomputing

• The Green500 Team
  – Encourages fair use of the list rankings to promote energy efficiency in high-performance systems
  – Discourages use of the list to disparage a particular vendor, system, or user
For More Information

- The Green500 List
  http://www.green500.org/

The Green500 List: Encouraging Sustainable Supercomputing

Wen-chao Feng and Kirk W. Cameron
Virginia Tech

The performance-at-any-cost design mentality ignores supercomputers' excessive power consumption and need for heat dissipation and will ultimately limit their performance. Without fundamental change in the design of supercomputing systems, the performance advances common over the past two decades won't continue.
Selected References:  Available at http://sss.cs.vt.edu/ Green Supercomputing

• “Green Supercomputing Comes of Age,” IT Professional, Jan.-Feb. 2008.
• “Green Supercomputing in a Desktop Box,” IEEE Workshop on High-Performance, Power-Aware Computing at IEEE IPDPS, Mar. 2007.
• “Green Destiny + mpiBLAST = Bioinfomagic,” 10th Int’l Conf. on Parallel Computing (ParCo’03), Sept. 2003.
• “Honey, I Shrunk the Beowulf!,” Int’l Conf. on Parallel Processing, Aug. 2002.
Recap

- An introspective look at **green supercomputing**
  ... where we have been ... where we are ....
  ... and what the future may hold

*The need for supercomputers to be colored green and fast*

- “Supercomputing in Small Spaces” Project (circa 2001)
  - Autonomic energy and power savings *while* maintaining performance in high-end computing systems.
    - Power-Aware Approach
      - At Integration Time → **Green Destiny** (2001)
      - At Run Time → $\beta$ Algorithm (2004)
    - Energy-Efficient Approach: Accelerators?

- **The Green500 List**
  [http://www.green500.org/](http://www.green500.org/)
Conclusion

• Recall that …
  – *Computational horsepower* requires electrical *power* consumption.

• Consequences
  – Electrical power costs $$$$.
  – “Too much” power affects efficiency, reliability, availability.

• Solution
  – Autonomic energy and power savings while maintaining performance in the supercomputer and datacenter.
    • Low-Power Architectural Approach
    • Power-Aware Commodity Approach
    • Energy-Aware Accelerator Approach
Acknowledgments

• Technical Contributions
  – Jeremy S. Archuleta (LANL & VT)  Former LANL Intern, now PhD
  – Chung-hsing Hsu (LANL)           Former LANL Postdoc
  – Michael S. Warren (LANL)         Former LANL Colleague
  – Eric H. Weigle (UCSD)            Former LANL Colleague

• Sponsorship & Funding
  – AMD
  – DOE Los Alamos Computer Science Institute
  – LANL Information Architecture Project
  – National Science Foundation
  – Virginia Tech
Wu FENG  

feng@cs.vt.edu

SyNeRGy Laboratory

http://synergy.cs.vt.edu/

http://www.chrec.org/

http://www.mpiblast.org/

http://www.green500.org/

http://sss.cs.vt.edu/