Green Destiny: A 240-Node Compute Cluster in One Cubic Meter

Wu-chun (Wu) Feng

Research & Development in Advanced Network Technology (RADIANT)
Computer & Computational Sciences Division
Los Alamos National Laboratory
Outline

- Where is Supercomputing?
  - Architectures from the Top 500.
- Evaluating Supercomputers
  - Metrics: Performance & Price/Performance
- An Alternative Flavor of Supercomputing
  - Supercomputing in Small Spaces → Bladed Beowulf
- Architecture of a Bladed Beowulf
- Performance Metrics
- Benchmark Results
- Discussion & Status
- Conclusion
- Acknowledgements & Media Coverage
Flavors of Supercomputing
(Picture Source: Thomas Sterling, Caltech & NASA JPL)
Architectures from the Top 500 Supercomputer List

Wu-chun Feng
feng@lanl.gov
http://www.lanl.gov/radiant
http://sss.lanl.gov
Metrics for Evaluating Supercomputers

- **Performance**
  - Metric: Floating-Operations Per Second (FLOPS)
  - Example: Japanese Earth Simulator

- **Price/Performance → Cost Efficiency**
  - Metric: Cost / FLOPS
  - Examples: SuperMike, GRAPE-5, Avalon.
Performance (At Any Cost)

- Japanese Earth Simulator ($400M)

<table>
<thead>
<tr>
<th></th>
<th>Performance</th>
<th>Price/Perf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak</td>
<td>40.00 Tflop</td>
<td>$10.00/Mflop</td>
</tr>
<tr>
<td>Linpack</td>
<td>35.86 Tflop</td>
<td>$11.15/Mflop</td>
</tr>
<tr>
<td>n-Body</td>
<td>29.50 Tflop</td>
<td>$13.56/Mflop</td>
</tr>
<tr>
<td>Climate</td>
<td>26.58 Tflop</td>
<td>$15.05/Mflop</td>
</tr>
<tr>
<td>Turbulence</td>
<td>16.40 Tflop</td>
<td>$24.39/Mflop</td>
</tr>
<tr>
<td>Fusion</td>
<td>14.90 Tflop</td>
<td>$26.85/Mflop</td>
</tr>
</tbody>
</table>

Wu-chun Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov
## Price/Performance

### Cost Efficiency

- **LSU’s SuperMike**  
  (2002: $2.8M)
  
<table>
<thead>
<tr>
<th>Performance</th>
<th>Price/Perf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linpack</td>
<td>2210 Gflops</td>
</tr>
</tbody>
</table>

- **U. Tokyo’s GRAPE-5**  
  (1999: $40.9K)
  
<table>
<thead>
<tr>
<th>Performance</th>
<th>Price/Perf</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-body</td>
<td>5.92 Gflops</td>
</tr>
</tbody>
</table>

- **LANL’s Avalon**  
  (1998: $152K)
  
<table>
<thead>
<tr>
<th>Performance</th>
<th>Price/Perf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak</td>
<td>149.40 Gflops</td>
</tr>
<tr>
<td>Linpack</td>
<td>19.33 Gflops</td>
</tr>
</tbody>
</table>
The Need for New Supercomputing Metrics

- Analogy: Buying a car. Which metric to use?
  - Raw performance, price/performance, fuel efficiency, reliability, size, etc.

- Issues with today’s supercomputing metrics
  - Focus: Performance & price/performance
    - Important metrics, but ...
Flavors of Supercomputing
(Picture Source: Thomas Sterling, Caltech & NASA JPL)

Bigger and faster machines simply aren’t good enough anymore.
Efficiency, reliability, and availability (ERA) will be the key issues of this decade.
Why ERA Metrics?

- Observations
  - Strong hints of the tradeoffs that come with "performance" and "price/performance" metrics ...
    - Lower efficiency, reliability, and availability.
    - Higher operational costs, e.g., admin, maintenance, etc.
  - Institutional consumers that use clusters as a tool ...
    - Pharmaceutical, financial, actuarial, retail, aerospace, data centers for web-server farms.
  - A couple of informational data points:
      - Reliability, transparency, and ease of use.
      - Low power, NOT speed.
      - DRAM density, NOT speed.
An Alternative Flavor of Supercomputing

- Supercomputing in Small Spaces (http://sss.lanl.gov)
  - First instantiation: *Bladed Beowulf*
    - MetaBlade (24), MetaBlade2 (24), and Green Destiny (240).
- Goal
  - Improve *efficiency, reliability, and availability* (ERA) in large-scale computing systems.
    - Sacrifice a little bit of raw performance.
    - Improve overall system throughput as the system will “always” be available, i.e., effectively no downtime, no hardware failures, etc.
  - Reduce the *total cost of ownership* (TCO).
- Analogy
  - Ferrari 550: Wins raw performance but reliability is poor so it spends its time in the shop. Throughput low.
  - Toyota Camry: Loses raw performance but high reliability results in high throughput (i.e., miles driven).
Architecture of a Bladed Beowulf

A Fundamentally Different Approach to High-Performance Computing
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")**
  - $\frac{1}{4}$ of Intel PIII $\rightarrow$ ~ 6x-7x less power dissipation
  - Less power $\rightarrow$ lower “on-die” temp. $\rightarrow$ better reliability & availability

BIOS, OS, Applications

Code Morphing Software

VLIW engine

x86

Wu-chun Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov

Los Alamos National Laboratory
Transmeta TM5x00 Comparison

<table>
<thead>
<tr>
<th>Intel P4</th>
<th>MEM</th>
<th>MEM</th>
<th>2xALU</th>
<th>2xALU</th>
<th>FPU</th>
<th>SSE</th>
<th>SSE</th>
<th>Br</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmeta TM5x00</td>
<td>MEM</td>
<td>2xALU</td>
<td>FPU</td>
<td>Br</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Current-generation Transmeta TM5800 performs comparably to an Intel PIII over iterative scientific codes on a clock-for-clock-cycle basis.
- Next-generation Transmeta CPU rectifies the above mismatch in functional units.
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).
RLX ServerBlade™ 633 (circa 2000)

Transmeta™
TM5600 633 MHz

Crusoe™

Public NIC
33 MHz PCI

Private NIC
33 MHz PCI

Management NIC
33 MHz PCI

Code Morphing Software
(CMS), 1 MB

Status LEDs
Serial RJ-45
debug port
Reset Switch

128MB, 256MB, 512MB
DIMM SDRAM
PC-133

512KB Flash ROM

ATA 66
0 or 1 or 2 - 2.5” HDD
10 or 30 GB each

128KB L1 cache, 512KB L2 cache
LongRun, Northbridge, x86 compatible

RLX ServerBlade™ 667 @ $960.
933 @ TBD.
1066 @ alpha.

Wu-chun Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov

Los Alamos
National Laboratory
RLX System™ 324
3U chassis that houses 24 blades

- 3U vertical space
  - 5.25” x 17.25” x 25.2”
- Two hot-pluggable 450W power supplies
  - Load balancing
  - Auto-sensing fault tolerance
- System midplane
  - Integration of system power, management, and network signals.
  - Elimination of internal system cables.
  - Enabling efficient hot-pluggable blades.
- Network cards
  - Hub-based management.
  - Two 24-port interfaces.

RLX System™ 300ex
- Interchangeable blades
  - Intel, Transmeta, or both.
- Switched-based management

Wu-chun Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov
"Green Destiny" Bladed Beowulf

- A 240-Node Beowulf in One Cubic Meter
- Each Node
  - 667-MHz Transmeta TM5600 CPU
    - Upgrade to 933-MHz Transmeta TM5800 CPUs
  - 640-MB RAM
  - 20-GB hard disk
  - 100-Mb/s Ethernet (up to 3 interfaces)
- Total
  - 160 Gflops peak (224 Gflops with upgrade)
  - 240 nodes
  - 150 GB of RAM (expandable to 276 GB)
  - 4.8 TB of storage (expandable to 38.4 TB)
Who Cares? So What? It’s a Smaller Beowulf ...

- **Goal**
  - Improve *efficiency*, *reliability*, and *availability* (ERA) in large-scale computing systems.
  - Reduce the *total cost of ownership* (TCO).

- **How to quantify ERA?**

- **What exactly is TCO?**
  - Can it be concretely quantified?
  - Or is it a “foofy” metric?
What is TCO?

- **Cost of Acquisition**
  - $$$ to buy the supercomputer.

- **Cost of Operation**
  - **Administration**
    - $$$ to build, integrate, configure, maintain, and upgrade the supercomputer over its lifetime.
  - **Power & Cooling**
    - $$$ in electrical power and cooling that is needed to maintain the operation of the supercomputer.
  - **Downtime**
    - $$$ lost due to the downtime (unreliability) of the system.
  - **Space**
    - $$$ spent to house the system.

Fixed, one-time cost
Variable, recurring cost
Total Price-Performance Ratio

- Price-Performance Ratio
  - Price = Cost of Acquisition
  - Performance = Floating-Point Operations Per Second

- Total Price-Performance Ratio (ToPPeR)
  - Total Price = Total Cost of Ownership (TCO)
  - Performance = Floating-Point Operations Per Second
Quantifying TCO?

- Why is TCO hard to quantify?
  - Components
    - Acquisition + Administration + Power + Downtime + Space
Why is TCO hard to quantify?

- Components
  - Acquisition + Administration + Power + Downtime + Space

Too Many Hidden Costs
Institution-specific
Quantifying TCO?

- Why is TCO hard to quantify?
  - Components
    - Acquisition + Administration + Power + Downtime + Space
  - Too Many Hidden Costs
  - Institution-Specific

- Traditional Focus: Acquisition (i.e., equipment cost)
  - Cost Efficiency: Price/Performance Ratio
Why is TCO hard to quantify?

- **Components**
  - Acquisition + Administration + **Power** + Downtime + **Space**

  Institution-Specific
  Too Many Hidden Costs

- Traditional Focus: Acquisition (i.e., equipment cost)
  - Cost Efficiency: Price/Performance Ratio

- **New Quantifiable Efficiency Metrics**
  - Power Efficiency: Performance/Power Ratio
  - Space Efficiency: Performance/Space Ratio

Wu-chun Feng
feng@lanl.gov
http://www.lanl.gov/radiant
http://sss.lanl.gov
Moore's Law for Power Dissipation

Chip Maximum Power in watts/cm²

1000
100
10
1

Year
1985 1995 2001

Pentium 4 – 75 watts
Pentium III – 35 watts
Pentium II – 35 watts
Pentium Pro – 30 watts
Pentium – 14 watts
I386 – 1 watt
I486 – 2 watts

Not too long to reach Nuclear Reactor

Surpassed Heating Plate

Itanium – 130 watts

Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta

Wu-chun Feng
feng@lanl.gov
http://www.lanl.gov/radiant
http://sss.lanl.gov
What’s wrong with high power?
- Costs $$$ to power such a system; costs $$$ to cool it.
- Causes reliability problems. Why?
  - Higher power implies higher temperatures.

Arrhenius’ Equation (circa 1980s)
- As temperature increases by 10° C ...
  - The failure rate of a system doubles.
  - The reliability of a system is cut in half.
- Twenty years of unpublished empirical data.
Empirical Data on Temperature

- From off to system boot-up, after 25 seconds:

<table>
<thead>
<tr>
<th>Processor</th>
<th>Clock Freq.</th>
<th>Voltage</th>
<th>Peak Temp.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Pentium III-M</td>
<td>500 MHz</td>
<td>1.6 V</td>
<td>252° F (122° C)</td>
</tr>
<tr>
<td>Transmeta Crusoe TM5600</td>
<td>600 MHz</td>
<td>1.6 V</td>
<td>147° F (64° C)</td>
</tr>
</tbody>
</table>

*Peak temperature measured with no cooling.

- Arrehenius’ Equation
  - Every 10° C increase, doubles the failure rate.

Implication: Without cooling facilities, PIII-M is 32 times more likely to fail!
Summary of Performance Metrics

- **Total Price/Performance Ratio (ToPPeR)**
  - Price is more than the *cost of acquisition*.
  - Operational costs: sys admin, power & cooling, space, downtime.

- **Performance/Power Ratio → “Power Efficiency”**
  - How efficiently does a computing system use energy?
  - How does this affect reliability and availability?
    - Higher Power Dissipation $\alpha$ Higher Temperature $\alpha$ Higher Failure Rate

- **Performance/Space Ratio → “Space Efficiency”**
  - How efficiently does a computing system use space?
  - Performance has increased by 2000 since the Cray C90; performance/sq. ft. has only increased by 65.
Benchmark Results
## Gravitational Microkernel Benchmark (circa June 2002)

### Processor Performance

<table>
<thead>
<tr>
<th>Processor</th>
<th>Math sqrt</th>
<th>Karp sqrt</th>
</tr>
</thead>
<tbody>
<tr>
<td>500-MHz Intel PIII</td>
<td>87.6</td>
<td>137.5</td>
</tr>
<tr>
<td>533-MHz Compaq Alpha EV56</td>
<td>76.2</td>
<td>178.5</td>
</tr>
<tr>
<td>633-MHz Transmeta TM5600</td>
<td>115.0</td>
<td>144.6</td>
</tr>
<tr>
<td>800-MHz Transmeta TM5800</td>
<td>174.1</td>
<td>296.6</td>
</tr>
<tr>
<td>375-MHz IBM Power3</td>
<td>298.5</td>
<td>379.1</td>
</tr>
<tr>
<td>1200-MHz AMD Athlon MP</td>
<td>350.7</td>
<td>452.5</td>
</tr>
</tbody>
</table>

Units are in Mflops.

**Memory Bandwidth for Transmetas (via STREAMS): 350 MB/s**
SSS Demo at SC 2001

- MetaBlade: 24 ServerBlade 633s
- MetaBlade2: 24 ServerBlade 800s

- MetaBlade Bladed Beowulf: 2.1 Gflops (unoptimized)
- MetaBlade2 Bladed Beowulf: 3.3 Gflops (unoptimized)

No failures since September 2001 despite no cooling facilities.
## Treecode Benchmark for n-Body

<table>
<thead>
<tr>
<th>Site</th>
<th>Machine</th>
<th>CPUs</th>
<th>Gflops</th>
<th>Mflops/CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NERSC</td>
<td>IBM SP-3</td>
<td>256</td>
<td>57.70</td>
<td>225.0</td>
</tr>
<tr>
<td>LANL</td>
<td>SGI O2K</td>
<td>64</td>
<td>13.10</td>
<td>205.0</td>
</tr>
<tr>
<td>LANL</td>
<td>Green Destiny</td>
<td>212</td>
<td>38.90</td>
<td>183.5</td>
</tr>
<tr>
<td>SC'01</td>
<td>MetaBlade2</td>
<td>24</td>
<td>3.30</td>
<td>138.0</td>
</tr>
<tr>
<td>LANL</td>
<td>Avalon</td>
<td>128</td>
<td>16.16</td>
<td>126.0</td>
</tr>
<tr>
<td>LANL</td>
<td>Loki</td>
<td>16</td>
<td>1.28</td>
<td>80.0</td>
</tr>
<tr>
<td>NASA</td>
<td>IBM SP-2</td>
<td>128</td>
<td>9.52</td>
<td>74.4</td>
</tr>
<tr>
<td>SC'96</td>
<td>Loki+Hyglac</td>
<td>32</td>
<td>2.19</td>
<td>68.4</td>
</tr>
<tr>
<td>Sandia</td>
<td>ASCI Red</td>
<td>6800</td>
<td>464.90</td>
<td>68.4</td>
</tr>
<tr>
<td>CalTech</td>
<td>Naegling</td>
<td>96</td>
<td>5.67</td>
<td>59.1</td>
</tr>
<tr>
<td>NRL</td>
<td>TMC CM-5E</td>
<td>256</td>
<td>11.57</td>
<td>45.2</td>
</tr>
</tbody>
</table>
“Cost Efficiency” Metrics

- **Price-Performance Ratio**
  - *Price* = Cost of Acquisition
  - ✓ *Performance* = Floating-Point Operations Per Second

- **Total Price-Performance Ratio (ToPPeR)**
  - *Total Price* = Total Cost of Ownership (TCO)
  - ✓ *Performance* = Floating-Point Operations Per Second
ToPPeR Metric

- **ToPPeR**: Total Price-Performance Ratio (over the lifetime of a 24-node cluster in a 80° F environment)

<table>
<thead>
<tr>
<th>Cost Parameter</th>
<th>Alpha</th>
<th>Athlon</th>
<th>PIII</th>
<th>P4</th>
<th>TM5600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquisition</td>
<td>$17K</td>
<td>$15K</td>
<td>$16K</td>
<td>$17K</td>
<td>$26K</td>
</tr>
<tr>
<td>System Admin</td>
<td>$60K</td>
<td>$60K</td>
<td>$60K</td>
<td>$60K</td>
<td>$5K</td>
</tr>
<tr>
<td>Power &amp; Cooling</td>
<td>$11K</td>
<td>$6K</td>
<td>$6K</td>
<td>$11K</td>
<td>$2K</td>
</tr>
<tr>
<td>Space</td>
<td>$8K</td>
<td>$8K</td>
<td>$8K</td>
<td>$8K</td>
<td>$2K</td>
</tr>
<tr>
<td>Downtime</td>
<td>$12K</td>
<td>$12K</td>
<td>$12K</td>
<td>$12K</td>
<td>$1K</td>
</tr>
<tr>
<td>TCO (four yrs)</td>
<td>$108K</td>
<td>$101K</td>
<td>$102K</td>
<td>$108K</td>
<td>$36K</td>
</tr>
</tbody>
</table>

- **Problem**: Too many hidden costs & institution-specific
- **ToPPeR metric is approximately 2x better ...**
Price/Performance vs. ToPPeR

- **Green Destiny**
  - **Price/Performance Ratio**
    - $26K / 38.9 Gflops = $0.67 / Mflop
  - **Total Price/Performance Ratio (ToPPeR)**
    - $36K / 38.9 Gflops = $0.92 / Mflop

- But ToPPeR is a “foofy” metric ...
Parallel Computing Platforms

- **Avalon (1996)**
  - 140-Node *Traditional Beowulf Cluster*

- **ASCI Red (1996)**
  - 9632-CPU *MPP*

- **ASCI White (2000)**
  - 512-Node (8192-CPU) *Cluster of SMPs*

- **Green Destiny (2002)**
  - 240-Node *Bladed Beowulf Cluster*
### Parallel Computing Platforms Running the N-body Code

<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>39</td>
</tr>
<tr>
<td>Area (ft²)</td>
<td>120</td>
<td>1600</td>
<td>9920</td>
<td>6</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>25000</td>
</tr>
<tr>
<td>Disk density (GB/ft²)</td>
<td>3.3</td>
<td>1.3</td>
<td>16.1</td>
<td>800.0</td>
</tr>
<tr>
<td>Power density (watts/ft²)</td>
<td>150</td>
<td>750</td>
<td>202</td>
<td>833</td>
</tr>
<tr>
<td>Space efficiency (Mflops/ft²)</td>
<td>150</td>
<td>375</td>
<td>252</td>
<td>6500</td>
</tr>
<tr>
<td>Power efficiency (Mflops/watt)</td>
<td>1.0</td>
<td>0.5</td>
<td>1.3</td>
<td>7.5</td>
</tr>
</tbody>
</table>
## Parallel Computing Platforms Running the N-body Code

<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>39</td>
</tr>
<tr>
<td>Area (ft²)</td>
<td>120</td>
<td>1600</td>
<td>9920</td>
<td>6</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>25000</td>
</tr>
<tr>
<td>Disk density (GB/ft²)</td>
<td>3.3</td>
<td>1.3</td>
<td>16.1</td>
<td>800.0</td>
</tr>
<tr>
<td>Power density (watts/ft²)</td>
<td>150</td>
<td>750</td>
<td>202</td>
<td>833</td>
</tr>
<tr>
<td>Space efficiency (Mflops/ft²)</td>
<td>150</td>
<td>375</td>
<td>252</td>
<td>6500</td>
</tr>
<tr>
<td>Power efficiency (Mflops/watt)</td>
<td>1.0</td>
<td>0.5</td>
<td>1.3</td>
<td>7.5</td>
</tr>
</tbody>
</table>
**Green Destiny vs. Japanese Earth Simulator**

<table>
<thead>
<tr>
<th>Machine</th>
<th>Green Destiny+</th>
<th>Earth Simulator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>2002</td>
<td>2002</td>
</tr>
<tr>
<td>LINPACK Performance (Gflops)</td>
<td>144 (ext.)</td>
<td>35,860</td>
</tr>
<tr>
<td>Area (ft²)</td>
<td>6</td>
<td>17,222</td>
</tr>
<tr>
<td>Power (kW)</td>
<td>5</td>
<td>7,000</td>
</tr>
<tr>
<td>Cost efficiency ($/Mflop)</td>
<td>2.33</td>
<td>11.15</td>
</tr>
<tr>
<td>Space efficiency (Mflops/ft²)</td>
<td>24,000</td>
<td>2,085</td>
</tr>
<tr>
<td>Power efficiency (Mflops/watt)</td>
<td>28.8</td>
<td>5.13</td>
</tr>
</tbody>
</table>

Disclaimer: This is not exactly a fair comparison. Why?  
(1) LINPACK performance is extrapolated for Green Destiny+.  
(2) Use of area and power does not scale linearly.
Discussion: Interesting Tidbits

- DARPA Contributes $2M to IBM’s Low Power Center in Aug. 2001
  - [http://www.computerworld.com/industrytopics/defense/story/0,10801,73289,00.html](http://www.computerworld.com/industrytopics/defense/story/0,10801,73289,00.html)

- Transmeta performance on N-body code can match Intel performance on a clock-for-clock-cycle basis.
  - Problem: Fastest Transmeta? Fastest Intel?

- Low component count on blade server enhances reliability.
  - 100 parts per RLX node vs. 800-1000 parts per typical node.

- Intel-based Bladed Beowulf: 18 nodes in 3U
  - 80° F environment: “Silent” failure on LINPACK.
    - 1/3 of nodes inaccessible.
  - 65 ° F environment: ~20% better performance vs. 933-MHz Transmeta.

- Why 10/100? GigE has been available for two years now.
  - In 2000-01, GigE ~12-15 W. Now, GigE ~6-8W?

- Systems community vs. applications community.
Recent Work

- April 2002: Assembled and integrated a 240-node Beowulf in one cubic meter called *Green Destiny*.
- July 2002: Worked with Transmeta to demonstrate comparable performance to similarly-clocked Intels.
- July 2002: Worked with DOE SciDAC-funded *3-D Supernova* project to demo “base code” on *Green Destiny*. (A vertically-integrated solution from hardware on up to the application.)

Future Work

- Demo first 3-D supernova on a Linux-based cluster at SC.
- Work with additional code teams, e.g., climate modeling, computational fluid dynamics, large-scale molecular dynamics.
- Upgrade *Green Destiny* processors from 667 MHz to 933 MHz.
Conclusion

- **New Performance Metrics**
  - Overall Efficiency
    - ToPPeR: Total Price-Performance Ratio
  - Power Efficiency
    - Performance-Power Ratio
  - Space Efficiency
    - Performance-Space Ratio

- **Predictions**
  - Traditional clustering and supercomputing as we know it will *NOT* scale to petaflop computing due to issues of efficiency, reliability, and availability.
  - “Supercomputing in Small Spaces” is a single step in the right direction ...
Conclusion

- Keeping It In Perspective
  - The “Supercomputing in Small Spaces” project [(http://sss.lanl.gov)](http://sss.lanl.gov) is *not* meant to replace today’s large supercomputers.
    - Focus on metrics related to efficiency, reliability, and availability (ERA) rather than raw performance.
      - i.e., SSS = “Toyota Camry” of supercomputing.
    - Works particularly well as a departmental cluster (or even institutional cluster if there exists power and space constraints).
Acknowledgments

- **Technical Co-Leads**
  - Mike Warren and Eric Weigle

- **Contributions**
  - Mark Gardner, Adam Engelhart, Gus Hurwitz

- **Enablers**
  - J. Thorp, A. White, R. Oldehoeft, and D. Lora (LACSI)
  - W. Feiereisen and S. Lee (CCS Division Office)

- **Funding Agencies**
  - LACSI
  - IA-Linux

- **Encouragement & Support**
  - Gordon Bell, Chris Hipp, Linus Torvalds
The “Hype”: A Sampling of Press Coverage

- “Not Your Average Supercomputer,” *Communications of the ACM*, 8/02.
- “Supercomputing Coming to a Closet Near You?” *HPCwire*, 5/31/02.
- “Bell, Torvalds Usher Next Wave of Supercomputing,” *HPCwire*, 5/24/02.
- “Bell, Torvalds Usher Next Wave of Supercomputing,” *CNN.com*, 5/21/02.
SUPERCOMPUTING
in SMALL SPACES

http://sss.lanl.gov

Wu-chun (Wu) Feng
Research and Development in Advanced Network Technology

RADIANT

http://www.lanl.gov/radiant