The Road to a Linux-Based Personal Desktop Cluster (PDC)

Wu-chun Feng
feng@lanl.gov

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

Outline

- Motivation & Background
  - Where is High-Performance Computing (HPC)?
  - The Need for Efficiency, Reliability, and Availability
- Supercomputing in Small Spaces (http://sss.lanl.gov)
    - Architecture & Experimental Results
    - MegaScale Project: MegaProto
    - Orion Multisystems: DT-12 and DS-96
- Publications & Recognition
- Conclusion
We have spent decades focusing on performance, performance, performance (and price/performance).

**Where is High-Performance Computing?**

Top 500 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](http://www.top500.org)
Metrics for Evaluating Supercomputers (or HPC)

- **Performance (i.e., Speed)**
  - Metric: Floating-Operations Per Second (FLOPS)
  - Example: Japanese Earth Simulator, ASCI Thunder & Q.

- **Price/Performance \(\rightarrow\) Cost Efficiency**
  - Metric: Acquisition Cost / FLOPS
  - Examples: LANL Space Simulator, VT System X cluster.
  
  (In general, Beowulf clusters.)

Performance & price/performance are important metrics, but what about ...

---

Reliability & Availability of Leading-Edge Clusters & Constellations

<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</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, UNC
Efficiency of Leading-Edge Clusters & Constellations

- “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**: 2000-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.

Where is Cluster Computing?

(Pictures: Thomas Sterling, Caltech & NASA JPL and Wu Feng, LANL)

Efficiency, reliability, and availability will be the key issues of this decade.
The Need for Efficiency, Reliability, and Availability

- Requirement: Near-100% availability with efficient and reliable resource usage.
  - E-commerce, enterprise apps, online services, ISPs, data and HPC centers supporting R&D.

- Problems
  - Frequency of Service Outages
    - 65% of IT managers report that their websites were unavailable to customers over a 6-month period.
  - Cost of Service Outages
    - NYC stockbroker: $6,500,000/hour
    - Ebay (22 hours): $225,000/hour
    - Amazon.com: $180,000/hour
  - Social Effects: negative press, loss of customers who “click over” to competitor (e.g., Google vs. Ask Jeeves)

Source: David Patterson, UC-Berkeley

Peter Bradley, Pratt & Whitney
  - Business: Aerospace Engineering (CFD, composite modeling)
  - HPC Requirements

Eric Schmidt, Google
  - Business: Instantaneous Search
  - HPC Requirements
    - Low Power, Availability and Reliability, and DRAM Density
    - NOT speed. Speed \(\rightarrow\) High Power & Temps \(\rightarrow\) Unreliability

Myself, LANL
  - Business: Research in High-Performance Networking
  - Problem: Traditional cluster fails weekly (or more often)
  - HPC Requirements
Operating Environment
♦ 85-90°F warehouse at 7,400 feet above sea level.
♦ No air conditioning, no air filtration, no raised floor, and no humidifier/dehumidifier.

Computing Requirement
♦ Parallel computer to enable high-performance network research in simulation and implementation.

Old Solution: Traditional Supercomputing Cluster
♦ 100-processor cluster computer that failed weekly in the above operating environment.

New Solution
Power-Aware, Reliable Supercomputing Cluster
♦ Green Destiny: A 240-Node Supercomputer in Five Sq. Ft.

Outline
Motivation & Background
♦ Where is High-Performance Computing (HPC)?
♦ The Need for Efficiency, Reliability, and Availability

Supercomputing in Small Spaces (http://sss.lanl.gov)
  ♦ Architecture & Experimental Results
  ♦ MegaScale Project: MegaProto
  ♦ Orion Multisystems: DT-12 and DS-96

Publications & Recognition

Conclusion
Supercomputing in Small Spaces: Efficiency, Reliability, and Availability via Power Awareness

- **Goal**
  - Improve efficiency, reliability, and availability (ERA) in large-scale computing systems.
  - Sacrifice a bit of raw performance (potentially).
  - Improve overall system throughput as the system will "always" be available, i.e., effectively no downtime, no HW failures, etc.
  - Reduce the total cost of ownership (TCO). Another talk ...

- **Crude Analogy**
  - Formula One Race Car: Wins raw performance but reliability is so poor that it requires frequent maintenance. Throughput low.
  - Honda S2000: Loses raw performance but high reliability results in high throughput (i.e., miles driven $\rightarrow$ answers/month).

How to Improve Efficiency, Reliability & Availability?

- **Observation**
  - High power density $\alpha$ high temperature $\alpha$ low reliability
  - **Arrhenius’ Equation**
    (circa 1890s in chemistry $\rightarrow$ circa 1980s in computer & defense industries)
    - As temperature increases by 10° C ...
      - The failure rate of a system doubles.
    - Twenty years of unpublished empirical data.

* The time to failure is a function of $e^{-E_a/kT}$ where $E_a$ = activation energy of the failure mechanism being accelerated, $k$ = Boltzmann’s constant, and $T$ = absolute temperature
Moore’s Law for Power

Can we build a low-power supercomputer that is efficient, reliable, and highly available but is still considered high performance?

Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta
“Green Destiny” Bladed Beowulf  
(circa April 2002)

- A 240-Node Beowulf Cluster in Five Sq. Ft.
- Each Node (with High-Performance CMS)
  - 667-MHz Transmeta TM5600 CPU w/ Linux 2.4.x
    - Upgraded to 1-GHz Transmeta TM5800 CPUs
  - 640-MB RAM, 20-GB hard disk, 100-Mb/s Ethernet (up to 3 interfaces)
- Total
  - 160 Gflops peak (240 Gflops with upgrade)
  - 150 GB of RAM (expandable to 276 GB)
  - 4.8 TB of storage (expandable to 38.4 TB)
  - Power Consumption: Only 3.2 kW.
- Reliability & Availability
  - No unscheduled failures in 24 months.

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

- Motivation & Background
  - Where is High-Performance Computing (HPC)?
  - The Need for Efficiency, Reliability, and Availability
- Supercomputing in Small Spaces (http://sss.lanl.gov)
    - Architecture & Experimental Results
    - MegaScale Project: MegaProto
    - Orion Multisystems: DT-12 and DS-96
- Publications & Recognition
- Conclusion
The Road from Green Destiny to Orion Multisystems

- Trends in High-Performance Computing
  - Rise of cluster-based high-performance computers.
    - Price/performance advantage of using “commodity PCs” as cluster nodes (Beowulf: 1993-1994.)
    - Different flavors: “homebrew” vs. “custom”
The Road from Green Destiny to Orion Multisystems

- Trends in High-Performance Computing
  - Rise of cluster-based high-performance computers.
  - Price/performance advantage of using “commodity PCs” as cluster nodes (Beowulf: 1993-1994.)
  - Different flavors: “homebrew” vs. “custom”
  - Maturity of open-source cluster software.
    - Emergence of Linux and MPI as parallel programming APIs.
  - Rapid decline of the traditional workstation.
    - Replacement of workstation with a PC.
    - 1000-fold (and increasing) performance gap with respect to the supercomputer.
    - Still a desperate need for HPC in workstation form.
Evolution of Workstations: Performance Trends

- PC performance caught up with workstations
  - PC OSes: NT and Linux

- A large gap has opened between PCs and supercomputers
  - 3 Gflops vs. 3 Tflops

Source: Orion Multisystems, Inc.

Cluster Technology
Low-Power Systems Design
Linux
But in the form factor of a workstation ... a cluster workstation
Imagine a 36 Gflop cluster **on your desk!**

- **LINPACK Performance**: 13.80 Gflops
- **Footprint**: 3 sq. ft. (24” x 18”)
- **Power Consumption**: 170 watts at load
- **How does this compare with a traditional desktop?**

**ORION DT-12 DESKTOP CLUSTER WORKSTATION**

**12 Nodes** in a single computer
**36 Gflops** peak processing power

**Designed for the Individual**
The Orion DT-12 cluster workstation is fully integrated, completely self-contained, and personal workstation based on the best of today’s cluster technologies. Designed to be an affordable individual workstation, it is capable of 36 Gflops peak performance (18 Gflops sustained) with no add-on options or add-on kit. The Orion DT-12 cluster workstation provides superior computing performance for the engineering, scientific, financial, and creative professional who need to solve computationally complex problems without any additions to the power of the desktop or server cluster.

**TASER SOFTWARE DEVELOPMENT**
The Orion DT-12 cluster workstation is the perfect platform for developers writing high-performance cluster software packages. It comes with cluster software development tools pre-installed, including compilers and a parallel compiler that allows you to develop your own parallel code. It is a node in a cluster of up to 12 nodes, with a single user interface and a single user account. It also includes a suite of system monitoring and management software.

**24 GBytes** memory capacity
**1 TByte** internal storage

**NO ASSEMBLY REQUIRED**
These configurations are designed for the desktop user who needs fast, powerful computing performance. The entire system comes with the push of a button and has the expandability and ease of use of a personal computer. These solid-state designs allow for flexible configurations and scalability.

---

**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 parallel-processing computing platform that can be plugged into a standard wall socket and operated in an office or laboratory environment.

**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 components. Designed to be an individual desktop computer, it is capable of 300 Gflops peak performance (150 Gflops sustained) with no add-ons or add-on kit. It runs on standard 32-bit Windows operating systems and Windows-based applications with no add-ons or add-on kit. It also includes a suite of system monitoring and management software.

**Recall ...**

**GDGD**: 101 Gflops

**Road to Tflop?**

- **10 DS-96s \( \Rightarrow \) ~ 1 Tflop LINPACK**
## Parallel Computing Platforms

### Running LINPACK

<table>
<thead>
<tr>
<th>Machine</th>
<th>ASCI Red</th>
<th>ASCI White</th>
<th>Green Destiny+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>1996</td>
<td>2000</td>
<td>2002</td>
</tr>
<tr>
<td>Performance (Gflops)</td>
<td>2379</td>
<td>7226</td>
<td>101.0</td>
</tr>
<tr>
<td>Area (ft²)</td>
<td>1600</td>
<td>9920</td>
<td>5</td>
</tr>
<tr>
<td>Power (kW)</td>
<td>1200</td>
<td>2000</td>
<td>5</td>
</tr>
<tr>
<td>DRAM (GB)</td>
<td>585</td>
<td>6200</td>
<td>150</td>
</tr>
<tr>
<td>Disk (TB)</td>
<td>2.0</td>
<td>160.0</td>
<td>4.8</td>
</tr>
<tr>
<td>DRAM density (MB/ft²)</td>
<td>366</td>
<td>625</td>
<td>30000</td>
</tr>
<tr>
<td>Disk density (GB/ft²)</td>
<td>1</td>
<td>16</td>
<td>960</td>
</tr>
<tr>
<td>Perf/Space (Mflops/ft²)</td>
<td>1487</td>
<td>728</td>
<td>20202</td>
</tr>
<tr>
<td>Perf/Power (Mflops/watt)</td>
<td>2</td>
<td>4</td>
<td>20</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Machine</th>
<th>ASCI Red</th>
<th>ASCI White</th>
<th>Green Destiny+</th>
<th>Orion DS-96</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>1996</td>
<td>2000</td>
<td>2002</td>
<td>2005</td>
</tr>
<tr>
<td>Performance (Gflops)</td>
<td>2379</td>
<td>7226</td>
<td>101.0</td>
<td>109.4</td>
</tr>
<tr>
<td>Area (ft²)</td>
<td>1600</td>
<td>9920</td>
<td>5</td>
<td>2.95</td>
</tr>
<tr>
<td>Power (kW)</td>
<td>1200</td>
<td>2000</td>
<td>5</td>
<td>1.58</td>
</tr>
<tr>
<td>DRAM (GB)</td>
<td>585</td>
<td>6200</td>
<td>150</td>
<td>96</td>
</tr>
<tr>
<td>Disk (TB)</td>
<td>2.0</td>
<td>160.0</td>
<td>4.8</td>
<td>7.68</td>
</tr>
<tr>
<td>DRAM density (MB/ft²)</td>
<td>366</td>
<td>625</td>
<td>30000</td>
<td>32542</td>
</tr>
<tr>
<td>Disk density (GB/ft²)</td>
<td>1</td>
<td>16</td>
<td>960</td>
<td>2603</td>
</tr>
<tr>
<td>Perf/Space (Mflops/ft²)</td>
<td>1487</td>
<td>728</td>
<td>20202</td>
<td>37119</td>
</tr>
<tr>
<td>Perf/Power (Mflops/watt)</td>
<td>2</td>
<td>4</td>
<td>20</td>
<td>69</td>
</tr>
</tbody>
</table>
Selected Publications
http://sss.lanl.gov (… about three years out of date …)


Sampling of Media Overexposure

Adding to the Media Hype …

Conclusion

- Efficiency, reliability, and availability will be the key issues of this decade.
- Approach: Reduce power consumption via HW or SW.
- Performance Metrics for Green Destiny (circa 2002)
  - Performance: 2x worse than fastest AMD/Intel.
  - Price/Performance: 2x worse.
  - Overall Efficiency (Total Price-Performance Ratio)
    - 1.5x to 2.0x better. See ACM Queue, Oct. 2003.
  - Power Efficiency (Perf/Power): 10x to 20x better.
  - Space Efficiency (Perf/Space): 20x to 60x better.
  - Reliability: “Infinite”
  - Availability: Nearly 100%.