Introduction to Clusters

Rocks-A-Palooza II
Ground Rules

◆ Interrupt me!
  ✍ If you have a question and need more information
  ✍ Would like me to go into more detail, or skip over some material
  ✍ I already know this stuff

◆ Tell me to slow down
  ✍ I tend to talk very fast
  ✍ We have about 200 slides to go through (in six hours)
    • But we will skip some, and other are very short
    • We have plenty of time
    • Last session will be unstructured (you’ve been warned)

◆ I don’t have to use my slides
  ✍ This workshop is for you
  ✍ Other topics are welcome (but also see track2)

◆ Tomorrow we will go over some of Track2
Introduction

A brief introduction to clustering and Rocks
Brief History of Clustering
(very brief)

- NOW pioneered the vision for clusters of commodity processors.
  - David Culler (UC Berkeley) started early 90’s
  - SunOS / SPARC
  - First generation of Myrinet, active messages
  - Glunix (Global Unix) execution environment
- Beowulf popularized the notion and made it very affordable.
  - Tomas Sterling, Donald Becker (NASA)
  - Linux
Definition: Beowulf

- Collection of *commodity PCs* running an *opensource* operating system with a *commodity network*
- Network is usually Ethernet, although non-commodity networks are sometimes called Beowulfs
- Come to mean any Linux cluster
- [www.beowulf.org](http://www.beowulf.org)
Types of Clusters

- Highly Available (HA)
  - Generally small, less than 8 nodes
  - Redundant components
  - Multiple communication paths
  - This is not Rocks

- Visualization Clusters
  - Each node drives a display
  - OpenGL machines
  - This is not core Rocks
  - But, there is a Viz Roll

- Computing (HPC Clusters)
  - AKA Beowulf
  - This is the core of Rocks
Definition: HPC Cluster Architecture

Frontend Node

Public Ethernet

Private Ethernet Network

Application Network (Optional)

Node

Node

Node

Node

Node

Node

Node

Node

Node

Node

© 2006 UC Regents
Minimum Components

- Local Hard Drive
- Power
- Ethernet

- i386 (Pentium/Athlon)
- x86_64 (Opteron/EM64T)
- ia64 (Itanium) server
Optional Components

- High-performance network
  - Myrinet
  - Infiniband (Infinicon or Voltaire)

- Network-addressable power distribution unit

- Keyboard/video/mouse network not required
  - Non-commodity
  - How do you manage your management network?
Cluster Pioneers

- In the mid-1990s, Network of Workstations project (UC Berkeley) and the Beowulf Project (NASA) asked the question:

Can You Build a High Performance Machine From Commodity Components?
The Answer is: Clusters now Dominate High-End Computing

Case Scenario

What does 128-node cluster look like?
128 Node cluster [In 6 months]

- **Frontend**
  - Dual-Processor (e.g. Xeon/Opteron 3.x Ghz) [4P= Dual-Socket Dual-Core @ 2.x GHz]
  - 2GB RAM [4GB/8GB]
  - Dual On board Gigabit Ethernet
  - 500 GB Storage (2 x 250GB SATA Drives) [1TB Storage 2x500GB]
  - CDROM
  - On board video

- **Compute Nodes**
  - Dual-Processor [4P= Dual-Socket Dual-Core @ 2.x GHz]
  - 2GB RAM [4GB/8GB]
  - Dual On board Gigabit Ethernet
  - 250 GB Storage
  - CDROM [No CDROM]
  - On board video
Additional Components

- Machine Racks
- Power
  - Network addressable power units
  - Power cords
- Network
  - 48 Port gigabit Ethernet switches
  - CAT5e cables
- VGA monitor, PC101 keyboard, mouse
## SPEC Benchmark

<table>
<thead>
<tr>
<th>Processor</th>
<th>GHz</th>
<th>SPECfp</th>
<th>SPECfp Rate</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athlon 64 X2 (1S/2C)</td>
<td>2.4</td>
<td>1634</td>
<td>33.1</td>
<td>649</td>
</tr>
<tr>
<td>Pentium 4 EE (1S/2C)</td>
<td>3.7</td>
<td>2236</td>
<td>37.7</td>
<td>1059</td>
</tr>
<tr>
<td>Opteron 285 (2S/4C)</td>
<td>2.6</td>
<td>2095</td>
<td>82.4</td>
<td>1049</td>
</tr>
<tr>
<td>Opteron 254 (2S/2C)</td>
<td>2.8</td>
<td>2223</td>
<td>53.8</td>
<td>674</td>
</tr>
<tr>
<td>Pentium 4 Xeon (2S/2C)</td>
<td>3.6</td>
<td>1868</td>
<td>33.0</td>
<td>640</td>
</tr>
<tr>
<td>Itanium 2 (2S/2C)</td>
<td>1.6</td>
<td>2712</td>
<td>51.5</td>
<td>1199</td>
</tr>
<tr>
<td>Power5+ (4C)</td>
<td>1.9</td>
<td>3007</td>
<td>133</td>
<td>???</td>
</tr>
</tbody>
</table>
## Processors

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
<th>Processors</th>
<th>Year</th>
<th>R&lt;sub&gt;max&lt;/sub&gt;</th>
<th>R&lt;sub&gt;peak&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NNSA/LLNL, United States</td>
<td>BlueGene/L - eServer Blue Gene Solution IBM</td>
<td>131072</td>
<td>2005</td>
<td>280600</td>
<td>367000</td>
</tr>
<tr>
<td>2</td>
<td>IBM Thomas J. Watson Research Center, United States</td>
<td>BGW - eServer Blue Gene Solution IBM</td>
<td>40960</td>
<td>2005</td>
<td>91290</td>
<td>114688</td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA/LLNL, United States</td>
<td>ASC Purple - eServer pSeries p5 575 1.9 GHz IBM</td>
<td>10240</td>
<td>2005</td>
<td>63390</td>
<td>77824</td>
</tr>
<tr>
<td>4</td>
<td>NASA/Ames Research Center/NAS, United States</td>
<td>Columbia - SGI Altix 1.5 GHz, Voltaire Infiniband SGI</td>
<td>10160</td>
<td>2004</td>
<td>51870</td>
<td>60960</td>
</tr>
<tr>
<td>5</td>
<td>Sandia National Laboratories, United States</td>
<td>Thunderbird - PowerEdge 1850, 3.6 GHz, Infiniband Dell</td>
<td>8000</td>
<td>2005</td>
<td>38270</td>
<td>64512</td>
</tr>
<tr>
<td>6</td>
<td>Sandia National Laboratories, United States</td>
<td>Red Storm Cray XT3, 2.0 GHz Cray Inc.</td>
<td>10880</td>
<td>2005</td>
<td>36190</td>
<td>43520</td>
</tr>
<tr>
<td>7</td>
<td>The Earth Simulator Center, Japan</td>
<td>Earth-Simulator NEC</td>
<td>5120</td>
<td>2002</td>
<td>35860</td>
<td>40960</td>
</tr>
<tr>
<td>8</td>
<td>Barcelona Supercomputer Center, Spain</td>
<td>MareNostrum - JS20 Cluster, PPC 970, 2.2 GHz, Myrinet IBM</td>
<td>4800</td>
<td>2005</td>
<td>27910</td>
<td>42144</td>
</tr>
<tr>
<td>9</td>
<td>ASTRON/University Groningen, Netherlands</td>
<td>Stella - eServer Blue Gene Solution IBM</td>
<td>12288</td>
<td>2005</td>
<td>27450</td>
<td>34406.4</td>
</tr>
<tr>
<td>10</td>
<td>Oak Ridge National Laboratory, United States</td>
<td>Jaguar - Cray XT3, 2.4 GHz Cray Inc.</td>
<td>5200</td>
<td>2005</td>
<td>20527</td>
<td>24960</td>
</tr>
</tbody>
</table>
Interconnects

- **Weak interconnect**
  - Gigabit Ethernet

- **Strong interconnect**
  - Myrinet ($800-$1000/ port)
  - Infiniband ($800-$1000 / port)

- **Dual Xeon compute node**
  - Node cost $2000
  - All of the above interconnects = $2500

- One of the surprising, but often essential, costs of a cluster
Myrinet

- Long-time interconnect vendor
  - Delivering products since 1995
- Deliver single 128-port full bisection bandwidth switch
- Performance (Myrinet MX):
  - Latency: 2.7 us
  - Bandwidth: 245 MB/s
  - Cost/port (based on 64-port configuration): $1000
    - Switch + NIC + cable
- Newer Myrinet 10G is Dual Protocol
  - 10GigE or 10G Myrinet
Myrinet

System sustains 64% of peak performance

- But smaller systems hit 70-75% of peak
Quadrics

- QsNetII E-series
  - Released at the end of May 2004
- Deliver 128-port standalone switches
- Performance:
  - Latency: 3 us
  - Bandwidth: 900 MB/s
  - Cost/port (based on 64-port configuration): $1800
    - Switch + NIC + cable
Quadrics

Sustains 74% of peak

- Other systems on Top500 list sustain 70-75% of peak
Infiniband

- Newest interconnect
- Currently shipping 32-port and 96-port switches
  - Requires 32-port switches requires 12 switches (and 256 Cables) to support a full bisection bandwidth network for 128 nodes
- Performance:
  - Latency: 6.8 us (New Adapter from Pathscale takes this to 1.3us without a switch)
  - Bandwidth: 840 MB/s
  - Estimated cost/port (based on 64-port configuration): $1000-$1200
    - Switch + NIC + cable
    - http://www.techonline.com/community/related_content/24364
## Infiniband

- Sustained 58% of peak
- Other Infiniband machines on Top500 list have achieved 64% and 68%

<table>
<thead>
<tr>
<th>Rank</th>
<th>Institute</th>
<th>Country</th>
<th>Model Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Virginia Tech</td>
<td>United States</td>
<td>1100 Dual 2.0 GHz Apple G5/Mellanox Infiniband 4X/Cisco GigE / 2200 Self-made</td>
</tr>
</tbody>
</table>

© 2006 UC Regents
Ethernet

- Latency: 30-80 us (very dependent on NIC, Switch, and OS Stack)

- Bandwidth: 100 MB/s

- Top500 list has ethernet-based systems sustaining between 35-59% of peak
## Ethernet

- What we did 3 years ago with 128 nodes and a $13,000 ethernet network
  - $101 / port
  - Sustained 48% of peak

<table>
<thead>
<tr>
<th>201</th>
<th>UCSD/Cal-IT²/SDSC United States/2003</th>
<th>Rocks V60x Cluster 2.8 GHz, Gig Ethernet / 256 Sun</th>
<th>699 1433.6</th>
</tr>
</thead>
</table>

- With Myrinet, would have sustained 1 Tflop
  - At a cost of ~$130,000
    - Roughly 1/3 the cost of the system
Rockstar Topology (Bisection BW made a difference)

- 24-port switches
- Not a symmetric network
  - Best case - 4:1 bisection bandwidth
  - Worst case - 8:1
  - Average - 5.3:1
Low Latency Ethernet?

- Bring os-bypass to Ethernet
- Projected performance:
  - Latency: less than 20 us
  - Bandwidth: 100 MB/s
- Potentially could merge management and high-performance networks
- Pioneering Vendor
  “Ammasso” is out of business

- At 10GigE Force 10 just introduced a 200ns switch (down from ~10us)
Sample Application Benefits

![Graph showing parallel efficiency scaling for different applications](graph.png)
Interconnect Observations

- If your application can tolerate latency, then Ethernet will deliver the best bang for the buck.
- Myrinet, Quadrics and Infiniband all have excellent low latency properties
- Myrinet delivers 2x bandwidth over Ethernet
- Quadrics and Infiniband deliver 2x bandwidth over Myrinet

- Observation: codes are often sensitive first to messaging overhead, then latency, then bandwidth
Details

<table>
<thead>
<tr>
<th>Size</th>
<th>Unit Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Nodes</td>
<td>128</td>
<td>2000</td>
</tr>
<tr>
<td>Frontend Nodes</td>
<td>1</td>
<td>3000</td>
</tr>
<tr>
<td>Total Node Count</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>Racks</td>
<td>5</td>
<td>800</td>
</tr>
<tr>
<td>Ethernet Switches</td>
<td>5</td>
<td>1400</td>
</tr>
<tr>
<td>Power Cords</td>
<td>135</td>
<td>0</td>
</tr>
<tr>
<td>Network Cables</td>
<td>130</td>
<td>5</td>
</tr>
<tr>
<td>Power Strips</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>Crash Cart</td>
<td>1</td>
<td>300</td>
</tr>
<tr>
<td>Total Hardware Cost</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• System Cost at **fixed size** is **relatively constant**
  – It is performance that changes

• Memory footprint can change pricing dramatically

• If your application needs low-latency buy a good interconnect
Add KVM

<table>
<thead>
<tr>
<th>Size</th>
<th>Unit Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Nodes</td>
<td>128</td>
<td>2000</td>
</tr>
<tr>
<td>Frontend Nodes</td>
<td>1</td>
<td>3000</td>
</tr>
<tr>
<td>Total Node Count</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>Racks</td>
<td>5</td>
<td>800</td>
</tr>
<tr>
<td>Ethernet Switches</td>
<td>5</td>
<td>1400</td>
</tr>
<tr>
<td>Power Cords</td>
<td>135</td>
<td>0</td>
</tr>
<tr>
<td>Network Cables</td>
<td>130</td>
<td>5</td>
</tr>
<tr>
<td>Power Strips</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>Crash Cart</td>
<td>1</td>
<td>300</td>
</tr>
<tr>
<td>KVM Cables</td>
<td>129</td>
<td>50</td>
</tr>
<tr>
<td>KVM Switch</td>
<td>9</td>
<td>1000</td>
</tr>
</tbody>
</table>

• $15K USD additional cost (~ 5%)

• KVM’s are low volume networks that will require management. Are they worth it?
Add Myrinet

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Unit Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Nodes</td>
<td>128</td>
<td>2000</td>
<td>$256,000</td>
</tr>
<tr>
<td>Frontend Nodes</td>
<td>1</td>
<td>3000</td>
<td>$3,000</td>
</tr>
<tr>
<td>Total Node Count</td>
<td>129</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Racks</td>
<td>5</td>
<td>800</td>
<td>$4,000</td>
</tr>
<tr>
<td>Ethernet Switches</td>
<td>5</td>
<td>1400</td>
<td>$7,000</td>
</tr>
<tr>
<td>Power Cords</td>
<td>135</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Network Cables</td>
<td>130</td>
<td>5</td>
<td>$650</td>
</tr>
<tr>
<td>Power Strips</td>
<td>17</td>
<td>30</td>
<td>$510</td>
</tr>
<tr>
<td>Crash Cart</td>
<td>1</td>
<td>300</td>
<td>$300</td>
</tr>
<tr>
<td>Myrinet NIC</td>
<td>128</td>
<td>500</td>
<td>$64,000</td>
</tr>
<tr>
<td>Myrinet Cables</td>
<td>128</td>
<td>100</td>
<td>$12,800</td>
</tr>
<tr>
<td>Myrinet Switch</td>
<td>1</td>
<td>30000</td>
<td>$30,000</td>
</tr>
<tr>
<td>Total Hardware Cost</td>
<td></td>
<td></td>
<td>$378,260</td>
</tr>
</tbody>
</table>

- Added $100K USD. ~ 33% of complete system
- Often essential to get codes to scale
1U Servers (Rack of 32 + Frontend)

- 64 Sockets (64-128 Cores)
- 5 electrical circuits (20A, 208V)
- Cable count
  - 65 = power & network
  - 97 with Myrinet
  - 193 with KVM
  - 225 with Serial Port management
Cluster Software Space

Rocks is not alone
Other efforts
Where Rocks fits
The Dark Side of Clusters

- Clusters are phenomenal price/performance computational engines …
  - Can be hard to manage without experience
  - High-performance I/O is still unsolved
  - Finding out where something has failed increases at least linearly as cluster size increases

- Not cost-effective if every cluster “burns” a person just for care and feeding

- Programming environment could be vastly improved

- Technology is changing very rapidly. Scaling up is becoming commonplace (128-256 nodes)
The Top 2 Most Critical Problems

- The largest problem in clusters is *software skew*
  - When software configuration on some nodes is different than on others
  - Small differences (minor version numbers on libraries) can cripple a parallel program

- The second most important problem is adequate job control of the parallel process
  - Signal propagation
  - Cleanup
Rocks (open source clustering distribution)  
www.rocksclusters.org

- Technology transfer of commodity clustering to application scientists
  - “make clusters easy”
  - Scientists can build their own supercomputers and migrate up to national centers as needed.
- Rocks is a cluster on a CD
  - Red Enterprise Hat Linux (opensource and free)
  - Clustering software (PBS, SGE, Ganglia, NMI)
  - Highly programmatic software configuration management
- Core software technology for several campus projects
  - BIRN
  - Center for Theoretical Biological Physics
  - EOL
  - GEON
  - NBCR
  - OptIPuter
- First Software release Nov, 2000
- Supports x86, Opteron/EM64T, and Itanium
- RedHat/CentOS 4.x
Philosophy

- Caring and feeding for a system is not fun
- System Administrators cost more than clusters
  - 1 TFLOP cluster is less than $200,000 (US)
  - Close to actual cost of a fulltime administrator
- The system administrator is the weakest link in the cluster
  - Bad ones like to tinker
  - Good ones still make mistakes
Philosophy continued

- All nodes are 100% automatically configured
  - Zero “hand” configuration
  - This includes site-specific configuration
- Run on heterogeneous standard high volume components
  - Use components that offer the best price/performance
  - Software installation and configuration must support different hardware
  - Homogeneous clusters do not exist
  - Disk imaging requires homogeneous cluster
Philosophy continued

- Optimize for installation
  - Get the system up quickly
  - In a consistent state
  - Build supercomputers in hours not months
- Manage through re-installation
  - Can re-install 128 nodes in under 20 minutes
  - No support for on-the-fly system patching
- Do not spend time trying to issue system consistency
  - Just re-install
  - Can be batch driven
- Uptime in HPC is a myth
  - Supercomputing sites have monthly downtime
  - HPC is not HA
OpenMosix

- **Overview**
  - Single system image - all nodes look like one large multiprocessor
  - Jobs migrate from machine to machine (based on machine load)
  - No changes required for apps to use system

- **Interconnects supported**
  - All IP-based networks

- **Custom Linux Kernel**
  - Download a new kernel
  - Or patch and compile
  - Install kernel on all nodes

- **Supports**
  - Diskfull
  - Diskless
Warewulf

- **Overview**
  - Install frontend first
    - Recommend using RPM-based distribution
  - Imaged based installation
    - “Virtual node filesystem”
  - Attacks problem of generic slave node management

- **Standard cluster software not included**
  - Added separately
  - Use ‘chroot’ commands to add in extra software

- **Supports**
  - Diskfull
  - Diskless
Scyld Beowulf

- Single System Image
  - Global process ID
  - Not a global file system

- Heavy OS modifications to support BProc
  - Patches kernel
  - Patches libraries (libc)

- Job start on the frontend and are pushed to compute nodes
  - Hooks remain on the frontend
  - Does this scale to 1000 nodes?

- Easy to install
  - Full distribution
  - Often compared to Rocks
SCore

◆ Research group started in 1992, and based in Tokyo.
◆ Score software
  ✜ Semi-automated node integration using RedHat
  ✜ Job launcher similar to UCB’s REXEC
  ✜ MPC++, multi-threaded C++ using templates
  ✜ PM, wire protocol for Myrinet
◆ Development has started on SCore Roll
Scalable Cluster Environment (SCE)

- Developed at Kasetsart University in Thailand
- SCE is a software suite that includes
  - Tools to install, manage, and monitor compute nodes
    - Diskless (SSI)
    - Diskfull (RedHat)
  - A batch scheduler to address the difficulties in deploying and maintaining clusters
  - Monitoring tools (SCMSWeb)
- User installs frontend with RedHat and adds SCE packages.
- Rocks and SCE are working together
  - Rocks is good at low level cluster software
  - SCE is good at high level cluster software
  - SCE Roll is now available for Rocks
Open Cluster Group
(OSCAR)

- OSCAR is a collection of clustering best practices (software packages)
  - PBS/Maui
  - OpenSSH
  - LAM/MPI
- Image based installation
  - Install frontend machine manually
  - Add OSCAR packages to frontend
  - Construct a “golden image” for compute nodes
  - Install with system imager
  - “Multi-OS” – Currently only supports RPM-based Distros
    - Dropping “Mandriva”..
- Started as a consortium of industry and government labs
  - NCSA, ORNL, Intel, IBM, Dell, others
  - Dell now does Rocks. NCSA no longer a contributor. IBM?
System Imager

- Originally VA/Linux (used to sell clusters) (now “bald guy software)
- System imaging installation tools
  - Manages the files on a compute node
  - Better than managing the disk blocks
- Use
  - Install a system manually
  - Appoint the node as the golden master
  - Clone the “golden master” onto other nodes
- Problems
  - Doesn’t support heterogeneous
  - Not method for managing the software on the “golden master”
  - Need “Magic Hands” of cluster-expert admin for every new hardware build
Cfengine

- Policy-based configuration management tool for UNIX or NT hosts
  - Flat ASCII (looks like a Makefile)
  - Supports macros and conditionals
- Popular to manage desktops
  - Patching services
  - Verifying the files on the OS
  - Auditing user changes to the OS
- Nodes pull their Cfengine file and run every night
  - System changes on the fly
  - One bad change kills everyone (in the middle of the night)
- Can help you make changes to a running cluster
Kickstart

- RedHat
  - Automates installation
  - Used to install desktops
  - Foundation of Rocks

- Description based installation
  - Flat ASCII file
  - No conditionals or macros
  - Set of packages and shell scripts that run to install a node
LCFG

- Edinburgh University
  - Anderson and Scobie

- Description based installation
  - Flat ASCII file
  - Conditionals, macros, and statements
    - Full blown (proprietary) language to describe a node

- Compose description file out of components
  - Using file inclusion
  - Not a graph as in Rocks

- Do not use kickstart
  - Must replicate the work of RedHat

- Very interesting group
  - Design goals very close to Rocks
  - Implementation is also similar
Rocks Basic Approach

- Install a frontend
  1. Insert Rocks Base CD
  2. Insert Roll CDs (optional components)
  3. Answer 7 screens of configuration data
  4. Drink coffee (takes about 30 minutes to install)

- Install compute nodes:
  1. Login to frontend
  2. Execute insert-ethers
  3. Boot compute node with Rocks Base CD (or PXE)
  4. Insert-ethers discovers nodes
  5. Goto step 3

- Add user accounts
- Start computing

Optional Rolls
- Condor
- Grid (based on NMI R4)
- Intel (compilers)
- Java
- SCE (developed in Thailand)
- Sun Grid Engine
- PBS (developed in Norway)
- Area51 (security monitoring tools)
- Many Others …
Minimum Requirements

- **Frontend**
  - 2 Ethernet Ports
  - CDROM
  - 18 GB Disk Drive
  - 512 MB RAM

- **Compute Nodes**
  - 1 Ethernet Port
  - 18 GB Disk Drive
  - 512 MB RAM

- Complete OS Installation on all Nodes
- No support for Diskless (yet)
- Not a Single System Image
- All Hardware must be supported by RHEL
ROCKS

HPCwire Reader’s Choice Awards for 2004/2005

◆ Rocks won in Several categories:
  ➤ Most Important Software Innovation (Reader’s Choice)
  ➤ Most Important Software Innovation (Editor’s Choice)
  ➤ Most Innovative - Software (Reader’s Choice)

Five Years, Five Awards

In its fifth year of development, Rocks has won its fifth HPCwire Award.

© 2006 UC Regents
Commercial Interest

Makes Beowulf Clusters child's play!

Scalable Rocks Web Console

- simple, web-based configuration
- improved maintenance
- simpler deployment, reduced network and decreased business risk

With Platform Rocks, you can:
- Rapidly deploy massive Linux-based computing infrastructure
- Reduce the hourly cost of ownership without compromise
- Reduce the headaches and business risk associated with deploying and managing Linux clusters

MX-2G Roll for Rocks v4.1

<table>
<thead>
<tr>
<th>Processor</th>
<th>Type of NIC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCIXD (Linux X86)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>or PCIXE (Linux X86P)</td>
<td></td>
</tr>
<tr>
<td>Myminet Roll for i86</td>
<td>MX-2G-1.11 roll for i86</td>
<td></td>
</tr>
<tr>
<td>Myminet Roll for x86</td>
<td>MX-2G-1.11 roll for x86</td>
<td></td>
</tr>
<tr>
<td>Myminet Roll for x86_64</td>
<td>MX-2G-1.11 roll for x86_64</td>
<td></td>
</tr>
</tbody>
</table>

Note: Each Myminet roll contains MX-2G 1.11, MXP2-MX 1.26.6.94, OpensUSE 1.0, and JPL. Installation instructions are available on the Rocks homepage.

© 2006 UC Regents
### Rocks Cluster Register

Add your cluster to the Register. Click on a column header to sort by that field. Click on an id for details and to edit your cluster.

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Org</th>
<th>CPU Type</th>
<th>CPUs</th>
<th>CPU Clock (GHz)</th>
<th>FLOPS (GFLOPS)</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>673</td>
<td>Total CPUs, Aver CPU Clock, Total FLOPS:</td>
<td>39982</td>
<td>2.14</td>
<td>150797.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>497</td>
<td>Tungsten-2</td>
<td>NCSA</td>
<td>EM64T</td>
<td>1040</td>
<td>3.60</td>
<td>7488</td>
<td>Urbana, IL</td>
</tr>
<tr>
<td>51</td>
<td>GridKa</td>
<td>Forschungszentrum</td>
<td>Radboud</td>
<td>1538</td>
<td>2.37</td>
<td>7384.92</td>
<td>Karlsruhe, Germany</td>
</tr>
<tr>
<td>71</td>
<td>EM6S-rocks</td>
<td>EMGS</td>
<td>EM64T</td>
<td>1860</td>
<td>3.40</td>
<td>7208</td>
<td>Tromsø, Norway</td>
</tr>
<tr>
<td>652</td>
<td>Athena_59</td>
<td>ACME</td>
<td>EM64T</td>
<td>949</td>
<td>3.40</td>
<td>6589.2</td>
<td>Brazil</td>
</tr>
<tr>
<td>180</td>
<td>Lionstar</td>
<td>TACC</td>
<td>Pentium 4</td>
<td>1024</td>
<td>3.06</td>
<td>6006.88</td>
<td>Austin, Texas</td>
</tr>
<tr>
<td>665</td>
<td>Tatanka</td>
<td>University Of Calgary</td>
<td>EM64T</td>
<td>624</td>
<td>3.40</td>
<td>4243.2</td>
<td>Calgary, Alberta</td>
</tr>
<tr>
<td>599</td>
<td>USCMDS Fermilab Tier1</td>
<td>Fermi National Accelerator Lab</td>
<td>Pentium 4</td>
<td>794</td>
<td>2.80</td>
<td>4294.4</td>
<td>Batavia, IL</td>
</tr>
</tbody>
</table>
User Base

- > 1300 Users on the Discussion List
- 5 Continents
- University, Commercial, Hobbyist
Beowulf Mentality

Why DIY is wrong
A Tale of a Cluster Tuner
(288 AthlonMP Hand Built Machine)
07.2002: The Idea
08.2002 - 11.2002: Construction
12.2002: Build Complete & Celebration

- Machine only 50% operational
- But, they are getting results
- Machine is fully operational 3 months later
Summary

- 07.2002
  - Design system
- 08.2002 - 11.2002
  - Build system
- 03.2003
  - System in Production
- 7 months (maybe 8)
  - Concept to Cluster
  - Still just a Beowulf
  - Moore-cycle is 18 months
    - Half life for performance
    - Half life for cost
  - Useful life is 36-48 months
- What did they optimize for?
Rocks Cluster Timeline

- Day 1 - Idea
- Day 30 - Production
- Not just us, world wide user base has done the same
Lost Time = Lost Computation

- Assumption
  - Rocks
    - 256 2.2 GHz Pentium IV
    - 1,126 GFlops
    - Available at same time as tuner build
    - 1 month to build
  - Tuner
    - 144 - 264 Athlon-MP 2200+
    - 512 - 950 Gflops
    - 5 - 7 months to build

- Baseline of 50% CPU efficiency for Rocks
- Tuner improvement beyond baseline
  - 10% (55% efficiency)
  - 20% (60% efficiency)
  - 50% (75% efficiency)
- Tuner must have 50% gain to catch baseline after 1.5 years
Invest in Hardware not People

- Assumptions
  - Two salaried tuners
  - “Full burden” (salary, grant overhead, office space, etc) is $180k / year.

- Invest
  - 5 months salary into baseline
  - $150k (5 months)
  - Just buy more nodes
    - $2500k / node

- Month 7
  - Baseline cluster grows
  - 54 2.2 GHz servers
  - Ignoring Moore’s Law!

- Baseline wins
Other Tuners

- Kernel Tuning
  - “My handcrafted kernel is X times faster.”

- Distribution Tuning
  - “Distribution Y is X times faster.”
  - RFP: “Vendor will be penalized for a Red Hat only solution”

- White-box Tuning
  - “White-box vendor Y has a node that is X times cheaper.”
Rocks

Making Clusters Easy
When You Need Power Today

Young Frankenstein - Gene Wilder, Peter Boyle
Two Examples

Rockstar - SDSC
Tungsten2 - NCSA
Rockstar Cluster

- 129 Sun Fire V60x servers
  - 1 Frontend Node
  - 128 Compute Nodes
- Gigabit Ethernet
  - $13,000 (US)
  - 9 24-port switches
  - 8 4-gigabit trunk uplinks
- Built live at SC’03
  - In under two hours
  - Running applications
- Top500 Ranking
  - 11.2003: 201
  - 06.2004: 433
  - 49% of peak
Rockstar Topology

- 24-port switches
- Not a symmetric network
  - Best case - 4:1 bisection bandwidth
  - Worst case - 8:1
  - Average - 5.3:1
Super Computing 2003 Demo

- We wanted to build a Top500 machine live at SC’03
  - From the ground up (hardware and software)
  - In under two hours
- Show that anyone can build a super computer with:
  - Rocks (and other toolkits)
  - Money
  - No army of system administrators required
- HPC Wire Interview
  - **HPCwire**: What was the most impressive thing you’ve seen at SC2003?
  - **Larry Smarr**: I think, without question, the most impressive thing I’ve seen was Phil Papadopoulos’ demo with Sun Microsystems.
Building Rockstar
NCSA
National Center for Supercomputing Applications

- Tungsten2
  - 520 Node Cluster
  - Dell Hardware
  - Topspin Infiniband
- Deployed 11.2004
- Easily in top 100 of the 06.2005 top500 list
- “We went from PO to crunching code in 2 weeks. It only took another 1 week to shake out some math library conflicts, and we have been in production ever since.” -- Greg Keller, NCSA (Dell On-site Support Engineer)

source: topspin (via google)
© 2006 UC Regents