NASA SuperComputing
A Ground Based Instrument for Exploration and Discovery

LUG 2015
Bob Ciotti
Chief Architect/Supercomputing Systems Lead
LUG 2015 - Denver
Discussion

- What is Pleiades
- The NASA Workload
- System Build Strategy
- Operational Strategy
- Tools and Analysis Software
- Issues Do We See
- Whats Lustre Does
- What We Want
Supercomputing Support for NASA Missions

- Agency wide resource
- Production Supercomputing
  - Focus on availability
- Machines mostly run large ensembles
- Some very large calculations (50k)
  - Typically 6500 jobs running

- Example applications
  - ARMD
    - LaRC: Jet wake vortex simulations, to increase airport capacity and safety
    - GRC: Understanding jet noise simulations, to decrease airport noise
  - ESMD
    - ARC: Launch pad flame trench simulations for Ares vehicle safety analysis
    - MSFC: Correlating wind tunnel tests and simulations of Ares I-X test vehicle
    - ARC/LaRC: High-fidelity CLV flight simulation with detailed protuberances
  - SMD
    - Michigan State: Ultra-high-resolution solar surface convection simulation
    - GSFC: Gravity waves from the merger of orbiting, spinning black holes
  - SOMD
    - JSC/ARC: Ultra-high-resolution Shuttle ascent analysis
  - NESC
    - KSC/ARC: Initial analysis of SRB burn risk in Vehicle Assembly Building

Jet aircraft wake vortices
Detailed CLV Geometry
Jet engine noise emissions
Orbiting, Spinning Black Holes
Solar surface convection
2-SRB Burn in VAB
Shuttle Ascent Configuration
KOI-157 :: Teff = 5685  logg = 4.38  Rs = 1.06
ECCO – Ocean Modeling
Planetary Defense
LBUG!
NASA’s Computational Landscape

Embarrassingly Parallel

Highly Complex and Evolving Computations

Simple Well Understood Computations

Compute Bound

Data/Storage Intensive

Tightly Coupled
SGI ICE Dual Plane – Topology

- **ib0**: 2x 11d hypercube
  - full: 2048 vertices
  - Pleiades: 1336/11d (2672 across both cubes)

- **ib1**:


LUG - Denver

Apr 2015
Infiniband Subnet LAN

LAN Implemented with out board IB switches

- Archive Servers
- NFS File Servers
- Hyperwall Graphics System
- Data Transfer Nodes
- Bridge Nodes
- Front End Nodes
- Data Analysis Nodes
- Lustre Filesystems
I/O Network

480 GB/sec
382 GB/sec
428 GB/sec ib0+ib1
857 GB/sec

Lustre Server

Hyperwall 128-Display Graphics Array

105 OSS+MDS
64 racks – 2008
393 teraflops
NASA (Pleiades) Rack Layout

92 racks – 2008
565 teraflops (#3 t500)
112 racks – 2009
683 teraflops
144 racks – 2010
969 teraflops
156 racks – 2010
1.08 petaflops
168 racks – 2011
1.18 petaflops
170 racks – 2011
1.20 petaflops
182 racks – 2011
1.31 petaflops
186 racks – 2011
1.33 petaflops
Pleiades - Sustained SpecFP rate base

- **SpecFP rate base estimates** (eliminates cell/GPU/blue-gene/SX vec)

<table>
<thead>
<tr>
<th>Spec Top500</th>
<th>Machine</th>
<th>CPU</th>
<th>#Sockets</th>
<th>FPR/Socket</th>
<th>TSpec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jaguar</td>
<td>AMD-2435</td>
<td>37,360</td>
<td>65.2</td>
<td>2,436,246</td>
</tr>
<tr>
<td>2</td>
<td>Tera-100</td>
<td>Intel-7560</td>
<td>17,296</td>
<td>133.4</td>
<td>2,307,805</td>
</tr>
<tr>
<td>3</td>
<td>Hopper</td>
<td>AMD-6176</td>
<td>12,784</td>
<td>149.8</td>
<td>1,800,115</td>
</tr>
<tr>
<td>4</td>
<td>Tianhe-1a</td>
<td>Intel-x5670</td>
<td>14,336</td>
<td>119.5</td>
<td>1,713,868</td>
</tr>
<tr>
<td>5</td>
<td>Pleiades</td>
<td>Intel-x</td>
<td>21,632</td>
<td>72.2</td>
<td>1,562,510</td>
</tr>
<tr>
<td>6</td>
<td>Cielo</td>
<td>AMD-6136</td>
<td>13,394</td>
<td>115.5</td>
<td>1,547,408</td>
</tr>
<tr>
<td>7</td>
<td>Kraken</td>
<td>AMD-2435</td>
<td>16,448</td>
<td>65.2</td>
<td>1,075,182</td>
</tr>
<tr>
<td>8</td>
<td>RedSky</td>
<td>Intel-x5570</td>
<td>10,610</td>
<td>90.3</td>
<td>958,401</td>
</tr>
<tr>
<td>9</td>
<td>Lomonosov</td>
<td>Intel-x5570</td>
<td>8,840</td>
<td>90.3</td>
<td>798,517</td>
</tr>
<tr>
<td>10</td>
<td>Ranger</td>
<td>AMD-2356</td>
<td>15,744</td>
<td>37.3</td>
<td>588,196</td>
</tr>
</tbody>
</table>

- Tspec == number of 2-core 296mhz UltraSPARC II
182 racks – 2012
1.7 petaflops

* Install – 3/30/2012 Note: RK 299 and RK 300 are RLC racks, Racks 301-312 and Racks 317-328 are Intel E5 Processors

Note: Rack 221 will cable to on 11D to rack 92. There is no 11D for Rack 222, this is a problem. If we remove rack 92 then we have issue with rack 222.
158 racks – 2012
deinstall
158 racks – 2012 deinstall
158 racks – 2012
deinstall
64 rack deinstall
2013

* Install – 3/30/2012 Note: RK 299 and RK 300 are RLC racks, Racks 301-312 and Racks 317-328 are Intel E5 Processors

Note: Rack 221 will cable to on 110 to rack 92. There is no 11d for Rack 222, this is a problem. If we remove rack 92 then we have issue with rack 222.

National Aeronautics and Space Administration LUG - Denver Apr 2015
167 racks – 2013
2.9 petaflops
160 racks – 2013
3.1 petaflops

NASA (Pleiades) Rack Layout as of 12/30/2013

A (ICE DORJ)
R4401.410 SG1 ICE X
(bybridge) Prev5 SW

B (ICE DORJ)
R4407.432 SG1 ICE X
(bybridge) Prev5 SW

C (ICE DORJ)
R453.448 SG1 ICE X
(bybridge) Prev5 SW

D (ICE DORJ)
R448 464 SG1 ICE X
(bybridge) Prev5 SW

E (ICE DORJ)
R461-161
1/1 Altix ICE 8400
Netstorage + ICE DORJ
1/1/161 Altix ICE 8400
EX (Westmore)

F (ICE DORJ)
R447
1/1/161 Altix ICE 8400
Netstorage + ICE DORJ
1/1/161 Altix ICE 8400 EX (Westmore)

G (ICE DORJ)
R430-317 SG1 ICE X
Sandbridge Prev5 SW
R4313-316 128 node
Pyramid in hypercube topology

H (ICE DORJ)
R4317-330 SG1 ICE X 5N13
Prev5 SW

I (ICE DORJ)
R420-144 Altix ICE 8400 EX

J (ICE DORJ)
R435-145
1/1/161 Altix ICE 8400 EX

K (ICE DORJ)
R431-193
200 Altix ICE 8400 EX

L (ICE DORJ)
R431-218
Altix ICE 8400 EX
R431-230 Cospair board
Westmore with GPGPU
R42013 in hypercube EX
R42013 Altix ICE 8400 EX

National Aeronautics and Space Administration
LUG - Denver
Apr 2015
168 racks – 2013
3.2 petaflops
168 racks – 2014
3.3 petaflops
170 racks – 2014
3.5 petaflops
168 racks – 2014
4.5 petaflops
168 racks – 2015
5.4 petaflops
Pleiades 2015 – Based on MemoryBW (ignore GPU/PHI)

<table>
<thead>
<tr>
<th>Machine</th>
<th>Type</th>
<th>11/14</th>
<th>Sockets</th>
<th>Type</th>
<th>Mem BW</th>
<th>Mem BW</th>
<th>Mega</th>
<th>Rmax</th>
<th>Rpeak</th>
<th>PctPeak</th>
</tr>
</thead>
<tbody>
<tr>
<td>K computer</td>
<td>Sparc64</td>
<td>4</td>
<td>88,128</td>
<td>VIII fx</td>
<td>64.0</td>
<td>373.2</td>
<td>5,640</td>
<td>10,510</td>
<td>11,280</td>
<td>93.2%</td>
</tr>
<tr>
<td>Sequoia</td>
<td>BGQ/Power</td>
<td>3</td>
<td>98,304</td>
<td>BGQ-A2</td>
<td>42.7</td>
<td>144.3</td>
<td>4,198</td>
<td>14.2</td>
<td>17,173</td>
<td>20,132</td>
</tr>
<tr>
<td>BlueWater</td>
<td>XK6/XK7</td>
<td>4</td>
<td>49,200</td>
<td>6276</td>
<td>51.2</td>
<td>176.0</td>
<td>2,519</td>
<td>8.7</td>
<td>71,378</td>
<td>85.3%</td>
</tr>
<tr>
<td>Mira</td>
<td>BGQ/Power</td>
<td>5</td>
<td>49,152</td>
<td>BGQ-A2</td>
<td>42.7</td>
<td>144.3</td>
<td>2,099</td>
<td>7.1</td>
<td>8,586</td>
<td>10,066</td>
</tr>
<tr>
<td>Tianhe-2</td>
<td>Xeon/Xeon Phi</td>
<td>1</td>
<td>32,000</td>
<td>E5-2692v2</td>
<td>59.7</td>
<td>321.5</td>
<td>1,910</td>
<td>10.3</td>
<td>33,862</td>
<td>54,902</td>
</tr>
<tr>
<td><strong>Pleiades</strong></td>
<td><strong>SGI/Xeon Mix</strong></td>
<td><strong>11</strong></td>
<td><strong>22,896</strong></td>
<td><strong>XeonMix</strong></td>
<td><strong>54.8</strong></td>
<td><strong>283.7</strong></td>
<td><strong>1,255</strong></td>
<td><strong>6.5</strong></td>
<td><strong>3,375</strong></td>
<td><strong>3,987</strong></td>
</tr>
<tr>
<td>Juqueen</td>
<td>BGQ/Power</td>
<td>8</td>
<td>28,672</td>
<td>BGQ-A2</td>
<td>42.7</td>
<td>144.3</td>
<td>1,224</td>
<td>4.1</td>
<td>5,008</td>
<td>5,872</td>
</tr>
<tr>
<td>Secret2</td>
<td>XC30/Xeon</td>
<td>13</td>
<td>18,832</td>
<td>E5-2697v2</td>
<td>59.7</td>
<td>341.0</td>
<td>1,124</td>
<td>6.4</td>
<td>3,143</td>
<td>4,881</td>
</tr>
<tr>
<td>Vulcan</td>
<td>BGQ/Power</td>
<td>9</td>
<td>24,576</td>
<td>BGQ-A2</td>
<td>42.7</td>
<td>144.3</td>
<td>1,049</td>
<td>3.5</td>
<td>4,293</td>
<td>5,033</td>
</tr>
<tr>
<td>Titan</td>
<td>XK7/Opteron/K20x</td>
<td>2</td>
<td>18,688</td>
<td>6274</td>
<td>51.2</td>
<td>173.0</td>
<td>957</td>
<td>3.2</td>
<td>17,590</td>
<td>27,112</td>
</tr>
<tr>
<td>SuperMUC</td>
<td>iData/Xeon</td>
<td>14</td>
<td>18,432</td>
<td>E5-2680</td>
<td>51.2</td>
<td>244.5</td>
<td>944</td>
<td>4.5</td>
<td>2,897</td>
<td>3,185</td>
</tr>
<tr>
<td>Pangea</td>
<td>SGI/Xeon</td>
<td>20</td>
<td>13,800</td>
<td>E5-2670</td>
<td>51.2</td>
<td>240.5</td>
<td>707</td>
<td>3.3</td>
<td>2,098</td>
<td>2,296</td>
</tr>
<tr>
<td>Stampede</td>
<td>Dell/Xeon/Phi</td>
<td>7</td>
<td>12,800</td>
<td>E5-2680</td>
<td>51.2</td>
<td>244.5</td>
<td>655</td>
<td>3.1</td>
<td>5,168</td>
<td>8,520</td>
</tr>
<tr>
<td>Hornet</td>
<td>XC40/Xeon</td>
<td>16</td>
<td>7,884</td>
<td>E5-2680v3</td>
<td>68.0</td>
<td>396.5</td>
<td>536</td>
<td>3.1</td>
<td>2,763</td>
<td>3,784</td>
</tr>
<tr>
<td>Tianhe-1A</td>
<td>Xeon/Nvidia2050</td>
<td>17</td>
<td>14,336</td>
<td>X5670</td>
<td>32.0</td>
<td>132.0</td>
<td>459</td>
<td>1.9</td>
<td>2,566</td>
<td>4,701</td>
</tr>
<tr>
<td>Secret1</td>
<td>CS/Xeon/K40</td>
<td>10</td>
<td>7,280</td>
<td>E5-2660v2</td>
<td>59.7</td>
<td>287.5</td>
<td>435</td>
<td>2.1</td>
<td>3,577</td>
<td>6,131</td>
</tr>
<tr>
<td>HPC2</td>
<td>iData/Xeon/K20x</td>
<td>12</td>
<td>7,200</td>
<td>E5-2680v2</td>
<td>59.7</td>
<td>313.0</td>
<td>430</td>
<td>2.3</td>
<td>3,188</td>
<td>4,605</td>
</tr>
<tr>
<td>Excalibur</td>
<td>XC40/Xeon</td>
<td>19</td>
<td>6,254</td>
<td>E5-2698v3</td>
<td>68.0</td>
<td>434.0</td>
<td>425</td>
<td>2.7</td>
<td>2,485</td>
<td>3,682</td>
</tr>
<tr>
<td>Piz Daint</td>
<td>XC30/Xeon/K20x</td>
<td>6</td>
<td>5,272</td>
<td>E5-2670 snb</td>
<td>51.2</td>
<td>240.5</td>
<td>270</td>
<td>1.3</td>
<td>6,271</td>
<td>7,788</td>
</tr>
<tr>
<td>Cascade</td>
<td>Xeon/Xeon Phi</td>
<td>18</td>
<td>1,880</td>
<td>E5-2670</td>
<td>51.2</td>
<td>240.5</td>
<td>96</td>
<td>0.5</td>
<td>2,539</td>
<td>3,388</td>
</tr>
<tr>
<td>Tsubame</td>
<td>Nec/Xeon/K20x</td>
<td>15</td>
<td>2,816</td>
<td>X5670</td>
<td>32.0</td>
<td>132.0</td>
<td>90</td>
<td>0.4</td>
<td>2,785</td>
<td>5,735</td>
</tr>
</tbody>
</table>

Numbers in Red are sWAG
Pleiades Environment

• 11,280 compute nodes – 22,560 sockets - 211,360 x86 cores
• 128 visualization nodes
• 192 GPU Nodes
• 192 Xeon Phi Nodes
• 10 Front End Nodes
• 4 “Bridge Nodes”
• 4 Archive Front Ends
• 8 Data Analysis Nodes
• 8 Archive Nodes
• 2 large memory nodes 2 TB + 4 TB

• Everything cross mounted. NFS Home, Modules, Nobackup (NFS, lustre)

• + a couple hundred administration/management nodes of various types.
## Pleiades /nobackup Filesystems (production)

### Lustre

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>#OST</th>
<th>#OSS</th>
<th>Size (PB)</th>
<th>Write BW</th>
<th>Read BW</th>
<th>controller</th>
</tr>
</thead>
<tbody>
<tr>
<td>p5</td>
<td>180</td>
<td>12</td>
<td>2.2</td>
<td>19</td>
<td>15</td>
<td>DDN SFA10k</td>
</tr>
<tr>
<td>p6</td>
<td>180</td>
<td>12</td>
<td>1.3</td>
<td>17</td>
<td>14</td>
<td>DDN SFA10k</td>
</tr>
<tr>
<td>p7</td>
<td>84</td>
<td>18</td>
<td>1.9</td>
<td>18</td>
<td>18</td>
<td>NetApp 5400</td>
</tr>
<tr>
<td>p8</td>
<td>312</td>
<td>26</td>
<td>7.1</td>
<td>65</td>
<td>52</td>
<td>NetApp 5400</td>
</tr>
<tr>
<td>p9</td>
<td>240</td>
<td>18</td>
<td>3.5</td>
<td>22</td>
<td>21</td>
<td>DDN SFA10k</td>
</tr>
</tbody>
</table>

- Total: 996 OSTs, 86 OSSs, 16 PB, 141 Write BW, 120 Read BW

### NFS
Incremental Expansion – Driving Factors

- Annual Funding/Budget Uncertainty
- Synthetic Leases/Sarbanes-Oxley cost
- Risk Mitigation for Fast moving technology
- Supports Short Lead/Opportunistic Strategy
- Timed adoption based on technology readiness
- Decouples technologies on different development cycles
- Dynamic project support

Maintains leading edge components throughout our “Ground Based Instrument”
Production Software Environment

- 4 different production selectable operating systems
  - AOE: 3 sles, centos
  - Additional test images

- 251 different loadable modules
  - 58 different compilers (32 intel, 8 PGI, 4 gcc, 3 cuda, 3 matlab… )
  - 26 different MPIs (10 SGI MPT, 12 Intel MPI, 8 MVAPICH)
  - 23 libraries (13 hdf, 6 netcdf, 4 mkl)

- Various debuggers, performance analyzers, plotting/graphing, editors

- Driven by user requests/requirements

This is an HPC Cloud
What is Today's General Purpose Supercomputer

- 1980s/1990s – a monolithic system with limited access
  - Typically served smaller communities
  - Local dedicated disk with limited network connectivity

- Today – it’s a collection of heterogeneous elements both SW & HW
  - Supports a wide variety and types of computation
  - Tuned for user productivity

- General Purpose - a compromise in some ways
  - MAY not be the #1 top 500 machine
  - But should be the most productive for highly varied requirements in multiple science and engineering domains.
Continuous Availability
24/7 Operations

- Goal – never take the whole system down
  - Outages are very disruptive
  - Dedicated time very costly
  - Not even possible to update entire system in one dedicated session.
  - Things go wrong

- Components
  - Lustre, NFS, CXFS, OFED, OpenSM, Linux Distro patches, cluster management software,
  - Firmware
    - its in everything – including the cables.
Continuous Availability
24/7 Operations

• Rolling updates of various components
  – Lustre/NFS clients/compute node images
    – Easy – simply done at end of user job
  – NFS, Lustre servers
    – Hot swap
      – Nfs hard mounts
      – Lustre recovery
    – Suspend/Resume
  – Schedule filesystems as a resource in addition to nodes
    – Allow us to use all compute nodes and figure out share later
    – Various admin, front ends, bridge nodes are easier or less urgent.
Continuous Availability
24/7 Operations

• Hot Plug - Grow system while in operation
  • Cable up new components powered off
    • Check cabling
    • Signal OpenSM to turn off sweep
  • Power on equipment
  • Run ibnetdiscover to verify cabling
  • Signal OpenSM to sweep
  • Mount file systems and go
• Cable Maintenance
### PBS Lustre Stats

Exiting at: Sat Mar 14 14:34:55 2015

```
==============================================================================
LUSTRE Filesystem Statistics
==============================================================================
nbp8 Metadata Operations
open close stat statfs read (GB) write (GB)
1056469 1056469 1058349 0 274 312

Read 4KB 8KB 16KB 32KB 64KB 128KB 256KB 512KB 1024KB
114 147 1 16 9 29 144 748 48185

Write 4KB 8KB 16KB 32KB 64KB 128KB 256KB 512KB 1024KB
5091 51 51 353 36 48 2120 49 297141
```

Job Resource Usage Summary for 3075801.pbspl1.nas.nasa.gov

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Time Used</td>
<td>259:36:54</td>
</tr>
<tr>
<td>Real Memory Used</td>
<td>37024436kb</td>
</tr>
<tr>
<td>Walltime Used</td>
<td>10:52:49</td>
</tr>
<tr>
<td>Exit Status</td>
<td>0</td>
</tr>
<tr>
<td>Number of CPUs Requested</td>
<td>816</td>
</tr>
<tr>
<td>Walltime Requested</td>
<td>24:00:00</td>
</tr>
<tr>
<td>Execution Queue</td>
<td>sls_aero1</td>
</tr>
<tr>
<td>Charged To</td>
<td>e0847</td>
</tr>
<tr>
<td>Job Stopped</td>
<td>Sat Mar 14 14:35:36 2015</td>
</tr>
</tbody>
</table>
File Transfer - Shiftc

- File transfers have become quite complex:
  - Best source/destination
    - Systems have multiple interfaces – want to pick best one
  - Threading
    - Big performance wins by parallelizing within a node
    - Big performance wins by parallelizing across nodes
  - Error checking
    - Checksum
      - Partial resend for hash mismatches
      - Ability to save partial hash to detect location of corruptions
  - Restart/Completion
    - Systems fail or reboot
      - Will restart transfer and notify upon completion
- Alternative to lustre-hsm, but some potential application…
- Multi GB/sec transfer within a filesystem
- Working on similar capability to DMF Archive

- Credit: Paul Kolano
Log File Analysis

• Lumber - Tool written to go through all the log file data (GBs/day)
  • Lustre logs
    • Server and Clients
  • PBS Logs
  • Console Logs
  • System Logs

• Absolutely necessary to track system issues

• Can specify a job ID and get all the log information across all systems during that timeframe.

• Can do arbitrary searches – across all logs

• Credit: Dave Barker
Daily Failure Logs – Past 24 hours

Daily Report for 04/10/2015 on pbspl1
Job Failure Summary from Fri Apr 10 00:00:00 2015 to Fri Apr 10 23:59:59 2015
There were 3197 jobs in the time region, of which 22 indicate as failed.
The total SBUs of these jobs was 500795.64, of which 6.70 (%0.00) belonged to the failed jobs.

Job Failure Summary Sorted by Frequency of Failure Types:

<table>
<thead>
<tr>
<th>Count</th>
<th>UID/GID</th>
<th>SBUs</th>
<th>Failure type</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6/6</td>
<td>0.00 (% 0.00)</td>
<td>head node lost connection with a sister node</td>
</tr>
<tr>
<td>6</td>
<td>5/5</td>
<td>6.38 (% 0.00)</td>
<td>job experienced out of memory (oom)</td>
</tr>
<tr>
<td>5</td>
<td>3/3</td>
<td>0.00 (% 0.00)</td>
<td>the PBS Server discarded the job because it appeared a node was down</td>
</tr>
<tr>
<td>1</td>
<td>1/1</td>
<td>0.05 (% 0.00)</td>
<td>job produced too much spool output (stdout/stderr)</td>
</tr>
<tr>
<td>1</td>
<td>1/1</td>
<td>0.28 (% 0.00)</td>
<td>PBS unable to start job</td>
</tr>
<tr>
<td>1</td>
<td>1/1</td>
<td>0.00 (% 0.00)</td>
<td>PBS server lost connection with head node</td>
</tr>
</tbody>
</table>
Weekly Failure Logs – Past 24 hours

Daily Report for last 7 days to 04/10/2015 on pbspl1
Job Failure Summary from Sat Apr  4 00:00:00 2015 to Fri Apr 10 23:59:59 2015
There were 14650 jobs in the time region, of which 148 indicate as failed.
The total SBUs of these jobs was 3598210.40, of which 239289.38 (%6.65) belonged to the failed jobs.

Job Failure Summary Sorted by Frequency of Failure Types:

<table>
<thead>
<tr>
<th>Count</th>
<th>UID/GID</th>
<th>SBUs</th>
<th>Failure type</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>19/17</td>
<td>480.60 (% 0.01)</td>
<td>job experienced out of memory (oom)</td>
</tr>
<tr>
<td>24</td>
<td>3/3</td>
<td>0.00 (% 0.00)</td>
<td>job start error 15010, node could not JOIN_JOB successfully</td>
</tr>
<tr>
<td>8</td>
<td>5/5</td>
<td>1361.84 (% 0.04)</td>
<td>job produced too much spool output (stdout/stderr)</td>
</tr>
<tr>
<td>8</td>
<td>6/6</td>
<td>0.00 (% 0.00)</td>
<td>the PBS Server discarded the job because it appeared a node was down</td>
</tr>
<tr>
<td>8</td>
<td>6/6</td>
<td>0.00 (% 0.00)</td>
<td>head node lost connection with a sister node</td>
</tr>
<tr>
<td>7</td>
<td>5/5</td>
<td>0.00 (% 0.00)</td>
<td>the PBS Server discarded the job for unknown reasons</td>
</tr>
<tr>
<td>6</td>
<td>4/2</td>
<td>145034.72 (% 4.03)</td>
<td>MPT error - receive completion flushed</td>
</tr>
<tr>
<td>6</td>
<td>2/2</td>
<td>210.25 (% 0.01)</td>
<td>node had RCU sched stalls</td>
</tr>
<tr>
<td>5</td>
<td>3/2</td>
<td>46686.32 (% 1.30)</td>
<td>MPT error - MPI_SGI_ctrl_recv failure</td>
</tr>
<tr>
<td>5</td>
<td>5/5</td>
<td>1553.60 (% 0.04)</td>
<td>node dropped into kdb</td>
</tr>
<tr>
<td>4</td>
<td>3/3</td>
<td>6074.78 (% 0.17)</td>
<td>MPT error - xmpi_net_send failure</td>
</tr>
<tr>
<td>3</td>
<td>3/3</td>
<td>3584.49 (% 0.10)</td>
<td>job experienced uncorrectable ecc memory error</td>
</tr>
<tr>
<td>2</td>
<td>2/2</td>
<td>90.62 (% 0.00)</td>
<td>at least one node associated with the job booted for unknown reasons</td>
</tr>
<tr>
<td>2</td>
<td>2/2</td>
<td>0.00 (% 0.00)</td>
<td>mlx4 internal error</td>
</tr>
<tr>
<td>2</td>
<td>2/2</td>
<td>0.26 (% 0.00)</td>
<td>PBS server lost connection with head node</td>
</tr>
<tr>
<td>1</td>
<td>1/1</td>
<td>34110.72 (% 0.95)</td>
<td>MPT error - continuous IB fabric problems</td>
</tr>
<tr>
<td>1</td>
<td>1/1</td>
<td>47.64 (% 0.00)</td>
<td>MPT error - network error in starting shepherd</td>
</tr>
<tr>
<td>1</td>
<td>1/1</td>
<td>53.27 (% 0.00)</td>
<td>MPT error - shepherd terminated</td>
</tr>
<tr>
<td>1</td>
<td>1/1</td>
<td>0.28 (% 0.00)</td>
<td>PBS unable to start job</td>
</tr>
</tbody>
</table>
### Real Time I/O Monitor

**Every 1.0s:** `abracadabra -i 1`

**Mar 26 00:31:37 2012**

<table>
<thead>
<tr>
<th>io_swx</th>
<th>nbp1 read</th>
<th>write</th>
<th>nbp2 read</th>
<th>write</th>
<th>nbp3/4 read</th>
<th>write</th>
<th>nbp5 read</th>
<th>write</th>
<th>nbp6 read</th>
<th>write</th>
<th>tot read</th>
<th>write</th>
</tr>
</thead>
<tbody>
<tr>
<td>r999i_mds</td>
<td>0.7</td>
<td>0.4</td>
<td>2.4</td>
<td>1.4</td>
<td>16.7</td>
<td>11.5</td>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>0.7</td>
<td>20.7</td>
<td>13.9</td>
</tr>
<tr>
<td>r999i_oss1</td>
<td>2.3</td>
<td>6.5</td>
<td>18.4</td>
<td>208.5</td>
<td>4.1</td>
<td>11.6</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>2.3</td>
<td>11.0</td>
<td>22.6</td>
</tr>
<tr>
<td>r999i_oss2</td>
<td>3.5</td>
<td>122.1</td>
<td>2.8</td>
<td>51.3</td>
<td>2.5</td>
<td>7.0</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>2.3</td>
<td>13.4</td>
<td>184.9</td>
</tr>
<tr>
<td>r999i_oss3</td>
<td>2.3</td>
<td>9.7</td>
<td>16.0</td>
<td>39.7</td>
<td>2.5</td>
<td>4.8</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>3.2</td>
<td>25.3</td>
<td>59.6</td>
</tr>
<tr>
<td>r999i_oss4</td>
<td>2.3</td>
<td>8.1</td>
<td>79.9</td>
<td>34.1</td>
<td>2.4</td>
<td>4.0</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>2.2</td>
<td>89.2</td>
<td>50.7</td>
</tr>
<tr>
<td>r999i_oss5</td>
<td>2.4</td>
<td>9.0</td>
<td>2.7</td>
<td>42.5</td>
<td>2.2</td>
<td>10.4</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>11.7</td>
<td>66.4</td>
</tr>
<tr>
<td>r999i_oss6</td>
<td>2.3</td>
<td>10.6</td>
<td>6.4</td>
<td>38.7</td>
<td>2.2</td>
<td>5.6</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>15.5</td>
<td>59.4</td>
</tr>
<tr>
<td>r999i_oss7</td>
<td>2.3</td>
<td>10.6</td>
<td>6.3</td>
<td>23.5</td>
<td>2.2</td>
<td>12.3</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>15.3</td>
<td>50.8</td>
</tr>
<tr>
<td>r999i_oss8</td>
<td>2.3</td>
<td>10.2</td>
<td>270.5</td>
<td>35.7</td>
<td>2.2</td>
<td>7.1</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>3.2</td>
<td>279.3</td>
<td>58.4</td>
</tr>
<tr>
<td>Total</td>
<td>20.4</td>
<td>187.2</td>
<td>405.4</td>
<td>475.4</td>
<td>37.0</td>
<td>74.3</td>
<td>18.0</td>
<td>19.3</td>
<td>20.6</td>
<td>481.4</td>
<td>566.7</td>
<td></td>
</tr>
</tbody>
</table>

**Max**

- RcvData: 2809.2 16138.9 5943.9 14297.8 3963.5 10441.4 8181.3 6729.7 5473.9 7.1 3.6 9.2 1.9 8.8 1.7 11.1 1.2 3.6 16847.1
- XmitData: 14.1 1393.7 6645.3 3405.3 1478.8 5506.1 13417.8 1675.2 2846.6 2498.5 1365.8 8.8 2.0 6.9 3.8 10.4 1.2 8.9 2.1 15130.8

**Total**

- RcvData: 0.1 62.4 4.1 6.0 5.7 14.4 52.2 22.8 128.4 18.4 171.4 288.3 0.3 0.3 0.3 0.0 0.2 0.1 0.2 15.1
- XmitData: 0.1 17.7 11.2 6.4 105.0 15.0 15.0 8.9 9.8 2.8 301.8 0.3 0.1 0.3 0.0 0.3 0.2 0.4 22.2

**Max**

- RcvData: 2809.2 16138.9 5943.9 14297.8 3963.5 10441.4 8181.3 6729.7 5473.9 7.1 3.6 9.2 1.9 8.8 1.7 11.1 1.2 3.6 16847.1
- XmitData: 14.1 1393.7 6645.3 3405.3 1478.8 5506.1 13417.8 1675.2 2846.6 2498.5 1365.8 8.8 2.0 6.9 3.8 10.4 1.2 8.9 2.1 15130.8

---

**National Aeronautics and Space Administration**

**LUG - Denver**

**Apr 2015**
Lustre Metadata Caching

- Implemented a methodology to keep metadata cached
  - Identify sections of OST where metadata is stored.
    - Inodes, bitmaps, etc.
  - Open the raw block device and read those blocks every 5 minutes.
  - Read Caching Turned off on OSS

- Helps to limit the impacts of certain types or user behavior.
  - Vast improvement on certain operations.

- Thought we could turn off in 2.4, but returned to this after meltdown.
What Do We Want from a Filesystem?

- Reliable
- Easy to Use
- Performance
- Free
What Do We Want from a Filesystem?

- Reliable
  - Some things are surprisingly reliable
    - Suspend/lflush/reboot
    - LBUG in OSS doesn’t kill everyone
      - Limited evictions
      - Recovery Works (sometimes)
  - Some things not
    - Cascading failures
      - LBUG or KDB across all servers
      - 1000’s of client evictions
    - *Always* hit already patched bugs
What Do We Want from a Filesystem?

- Easy to Use
  - Generally – Very easy to use (POSIX compliant)
  - Maybe a few odd end cases
    - E.g. partial read or write
What Do We Want from a Filesystem?

- Performance
  - Can get very good performance

- Things you need to do to get performance doesn’t always map easily to many applications.
  - ECCO

- Large system
  - I/Os look random once they get to the back end
What Do We Want from a Filesystem?

- Free
  - Yes – In the Stallman sense.

- Still require high levels of support
  - Bug tracking/patching - steep curve here
Issues

- Intel kept two maintenance releases 2.4 and 2.5, then dropped 2.4
- Got on 2.4 early, and then had problems moving to 2.5
- Hit many bugs that were already patched

- Bug tracking jira and Bug patching gerrit system need to talk. Missed some updated patch sets, resulted in more crashes.
Issues

- Resilience
  - Cascading failures.
  - Rebooting all 110 lustre servers
  - Commit on Share (help recovery?)

- Quiesce Filesystem for administrative work/upgrades

- Performance
  - Single user can drag down performance
  - Network Request Scheduler (LU-398) is on out list to test

- Single client performance
Issues

– Quotas seldom work. Moving to the OSTs made them more fragile

– We seem to always hit bugs that are already patched.
  – Over and over again. Since the beginning of time.
What Does NASA Want from Lustre

• Increased Stability
  – Better Patch Management
• Better Workload Performance (500+ jobs).
• QoS – Limiting Damage of Creative Users
• Administrative Shutdown