HPCG/HPL on Pleiades

Bob.Ciotti@nasa.gov
Chief Systems Architect
NASA Ames Research Center
SGI ICE Dual Plane – Topology

2x 11d hypercube
full 11d == 2048 vertices
Pleiades – partial 11d - 1336 vertices (2672 across both cubes)

64 racks – 2008
393 teraflops
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
1.33 petaflops

186 racks – 2011

Gpup racks 219 and 220
but configured as rack
219. Note switches on
gpup are in rear of rack
so cable lengths needs to
be adjusted to reflect this.

Note: Racks 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 racks 221 &
222.

Gpup racks 219 and 220
but configured as rack
219. Note switches on
gpup are in rear of rack
so cable lengths needs to
be adjusted to reflect this.

Note: Racks 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 racks 221 &
222.
Pleiades - Sustained SpecFP rate base  
(2011 timeframe)

- 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>2</td>
<td>Jaguar</td>
<td>AMD-2435</td>
<td>37,360</td>
<td>65.2</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>Tera-100</td>
<td>Intel-7560</td>
<td>17,296</td>
<td>133.4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>Hopper</td>
<td>AMD-6176</td>
<td>12,784</td>
<td>149.8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Tianhe-1a</td>
<td>Intel-x5670</td>
<td>14,336</td>
<td>119.5</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>Pleiades</td>
<td>Intel-x</td>
<td>21,632</td>
<td>72.2</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>Cielo</td>
<td>AMD-6136</td>
<td>13,394</td>
<td>115.5</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>Kraken</td>
<td>AMD-2435</td>
<td>16,448</td>
<td>65.2</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>RedSky</td>
<td>Intel-x5570</td>
<td>10,610</td>
<td>90.3</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
<td>Lomonosov</td>
<td>Intel-x5570</td>
<td>8,840</td>
<td>90.3</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>Ranger</td>
<td>AMD-2356</td>
<td>15,744</td>
<td>37.3</td>
</tr>
</tbody>
</table>

- Tspec == number of 2-core 296mhz UltraSPARC II
158 racks – 2012
1.15 petaflops
deinstall
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

Gpgpu racks 219 and 220 but configured as rack 219. Not switches on gpgpu are in rear of rack so cable lengths needs to be adjusted to reflect this.

Note: Racks 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 racks 221 & 222.
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

Gpgpu racks 219 and 220 but configured as rack 219. Note switches on gpgpu are in rear of rack so cable lengths needs to be adjusted to reflect this.

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 racks 221 & 222.
167 racks – 2013
2.9 petaflops

Note: 06/21/2013 - Rack 001-004 are RD racks for RLC and switches. Row A,B,C,D – 46 racks are proposed NPB. They will connect via 10G to Row D and P. This will be partial 9D and partial 10D.

Note: 1st delivery: 8 racks in Row C
2nd delivery: 8 racks in Row D
3rd delivery: add 8 racks to row D
C
Rack 001-004 are the admin racks that house the RLC and ethernet switches. There is one being added for Row A.

* Install – Rk 313-316
are the pyramid with MIC racks and are configured as two racks of SGI ICE X except there are only 64 nodes per rack. They are virtually racks 313 and 314. They will not be delivered till Nov 2012 to NASA

This is the switch rack

Gigaswitch racks 219 and 220 but configured as rack 219, note switches on gigaswitch are in rear of rack, so cable lengths needs to be adjusted to reflect this.

Note: Rack 221 will cable to on 11D to rack 52. There is no 1 fit for Rack 222, this is a problem. If we remove rack 52 then we have issue with racks 221 & 222.
3.1 petaflops

160 racks – 2013
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>32.9</td>
<td>10,510</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>85.3%</td>
</tr>
<tr>
<td>BlueWater</td>
<td>XK6/XK7</td>
<td></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></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>85.3%</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>61.7%</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>XeonMix</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>85.3%</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>64.4%</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>85.3%</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>64.9%</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</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
  - Westmere, Sandybridge, Ivybridge, Haswell
- 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
- + a couple hundred administration/management nodes of various types.
Load balancing with 1 MPI task/core

- SNB E5-2670 8c 2.6 GHz: STREAM Triad 75.9 GB/s/node, **4.74** GB/s/core
- IVB E5-2680v2 10c 2.8 GHz: STREAM Triad 95.7 GB/s/node, **4.79** GB/s/core
- HSW E5-2680v3 12c 2.5 GHz: STREAM Triad 117.2 GB/s/node, **4.89** GB/s/core

Performance measurements

- SNB: 99.3% scaling efficiency from 1 to 1868 nodes
- IVB: 97.4% scaling efficiency from 1 to 5347 nodes
- HSW: 95.9% scaling efficiency from 1 to 2073 nodes

Credit - Cheng Laio - SGI
The SGI code is optimized using common techniques such as contiguous memory, storage format tuning, multi-color reordering and combined computations.

The code is pure MPI.

Improvements and Extensions are being planned.
Performance on Various Systems

- e5-2690v3-ES
- e5-2697v3-COD
- e5-4667v3 (UV)
- e7-8890v3 (UV)

Credit - Cheng Laio - SGI
**HPCG June 2015**

<table>
<thead>
<tr>
<th>HPCG Rank (Jun 15)</th>
<th>#5</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top500 Rank (Nov 15)</td>
<td>#13</td>
<td>#1</td>
<td>#4</td>
<td>#2</td>
<td>#5</td>
</tr>
<tr>
<td></td>
<td>Pleiades</td>
<td>Tianhe</td>
<td>K Computer</td>
<td>Titan</td>
<td>Mira</td>
</tr>
<tr>
<td>Cores</td>
<td>186,288</td>
<td>3,120,000</td>
<td>705,024</td>
<td>560,640</td>
<td>786,432</td>
</tr>
<tr>
<td>HPCG PF</td>
<td>0.131</td>
<td>0.580</td>
<td>0.461</td>
<td>0.322</td>
<td>0.167</td>
</tr>
<tr>
<td>HPL PF</td>
<td>4.089</td>
<td>33.863</td>
<td>10.51</td>
<td>17.59</td>
<td>8.567</td>
</tr>
<tr>
<td>Peak PF</td>
<td>4.970</td>
<td>54.902</td>
<td>11.280</td>
<td>27.112</td>
<td>10.066</td>
</tr>
<tr>
<td>HPCG MF/Core</td>
<td>703.21</td>
<td>185.90</td>
<td>653.59</td>
<td>574.88</td>
<td>212.35</td>
</tr>
<tr>
<td>HPL GF/Core</td>
<td>21.95</td>
<td>10.85</td>
<td>14.91</td>
<td>31.37</td>
<td>10.89</td>
</tr>
<tr>
<td>Peak GF/Core</td>
<td>26.68</td>
<td>17.60</td>
<td>16.00</td>
<td>48.36</td>
<td>12.80</td>
</tr>
<tr>
<td>HPCG %of HPL</td>
<td>3.20%</td>
<td>1.71%</td>
<td>4.38%</td>
<td>1.83%</td>
<td>1.95%</td>
</tr>
<tr>
<td>HPCG %of Peak</td>
<td>2.64%</td>
<td>1.06%</td>
<td>4.09%</td>
<td>1.19%</td>
<td>1.66%</td>
</tr>
</tbody>
</table>

No one has built a 1 PetaFlop machine yet 😞
HPL Runtime Performance

Memory error burst or partial DIMM failure result in ~5TF performance drops or run failure on green try30.

<3TF lost on some memory errors or a single network transmission error in red try26.

Nice to modify HPCG to give similar real time performance metrics.
Fail
Node crash on DIMM issue in last seconds.

<table>
<thead>
<tr>
<th>T/V</th>
<th>N</th>
<th>NB</th>
<th>P</th>
<th>Q</th>
<th>Time</th>
<th>Gflops</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHC01L2L4</td>
<td>5459520</td>
<td>192</td>
<td>128</td>
<td>72</td>
<td>26568.92</td>
<td>4.08318e+06</td>
</tr>
</tbody>
</table>

HPL_pdgesv() start time Thu Jun 4 23:07:09 2015
HPL_pdgesv() end time Fri Jun 5 06:29:58 2015

||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)= 77939.0629011 ...... FAILED
||Ax-b||_oo = 224.874403
||A||_oo = 1366608.843671
||A||_1 = 1366651.513851
||x||_oo = 3.483180
||x||_1 = 2887333.826247
||b||_oo = 0.500000
Success
After many years of HPL, we observe that a successful run always begin late evening.

<table>
<thead>
<tr>
<th>T/V</th>
<th>N</th>
<th>NB</th>
<th>P</th>
<th>Q</th>
<th>Time</th>
<th>Gflops</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHC01L2L4</td>
<td>5455920</td>
<td>192</td>
<td>128</td>
<td>72</td>
<td>26528.30</td>
<td>4.08943e+06</td>
</tr>
</tbody>
</table>

HPL_pdgesv() start time Fri Jun 5 20:14:31 2015

HPL_pdgesv() end time Sat Jun 6 03:36:40 2015

||Ax-b||_oo/ (||A||_oo*||x||_oo+||b||_oo)*N) = 0.0026823 ...... PASSED

Finished 1 tests with the following results:
1 tests completed and passed residual checks,
0 tests completed and failed residual checks,
0 tests skipped because of illegal input values.
HPL vs HPCG Runs

Memory DIMM errors dominate failures on HPL runs. These far exceed those seen in normal production likely due to larger memory footprint and higher CPU load (temperature). HPL then exposes these errors that were latent (dormant).

HPCG retries were mostly related to debugging a network layer/MPI issue that had been occurring in production without a reproducer – until this HPCG run. This allowed us to identify and correct the issues. Some HW/memory fallout did occur when bringing up system after two days powered off while system returns to nominal operating temperature.
Summary

HPL Strengths:
• Good for burn-in, clean-up.
• Useful in finding problems.
  • SW, Processors, memory, network, building power distribution, cooling.
HPL too time consuming, skipped runs on several major upgrades.

HPCG Strengths:
• Easy to map to system
• Configurable runtime
• Useful performance information on short (<1 hr) runs.
• Also found problems – SW, HW
• More Representative of performance seen on NASA codes

Most significant issue by far: Memory DIMMS
Credits to the Team

John Baron SGI
Cheng Laio SGI
Michael Raymond SGI
Jay Lan SGI
Scott Emery SGI
Jennifer Fung SGI
Jose Rodriguez SGI
Matt Lepp SGI
Jason Inoue SGI
Rich Davila SGI
John Dugan SGI

Davin Chan CSC
Dale Talcott CSC
Jim Karella CSC
Greg Matthews CSC
Herbert Yeung CSC
Mahmoud Hanafi CSC
Mike Hartman CSC
Jeff Becker CSC
Bill Thigpen NASA
Mark Tangney NASA
Bob Ciotti NASA