The Blue Water’s File/Archive System

Data Management Challenges
Michelle Butler
mbutler@ncsa.illinois.edu
NCSA is a...

- World leader in deploying supercomputers and providing scientists with the software and expertise needed to fuel discoveries in science and engineering
- Unique partnership among the University of Illinois, state of Illinois, and federal government
- Home to more than 250 computing experts and students
- Key partner in the National Science Foundation’s TeraGrid project
- Home to Blue Waters, expected to be the most powerful computer for open scientific research when it comes online in 2011
NCSA’s current computing power

• 4 production systems
• More than 155 teraflops (155 TRILLION calculations every second)
• About 1,500 users nationwide
• Researchers receive time at no cost through peer review
• Archive environment at 6PB and growing at 75%/year
Let’s get Blue Waters specific!
### Diverse Large Scale Computational Science

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanoscience</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Chemistry</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Fusion</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Climate</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Combustion</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Astrophysics</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Nuclear</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>System Balance Implications</td>
<td>General Purpose balanced System</td>
<td>High Speed CPU, High Flop/s rate</td>
<td>High Performance Memory</td>
<td>High Interconnect Bisection bandwidth</td>
<td>High Performance Memory</td>
<td>High Speed CPU, High Flop/s rate</td>
<td>Irregular Data and Control Flow</td>
<td></td>
</tr>
</tbody>
</table>

*Use, reproduction, or disclosure is subject to the restrictions as stated in the "IBM Agreement for Exchange of Confidential Information – Blue Waters" as documented in the "IBM Statement of Work for Blue Waters System Prepared for the National Center for Supercomputing Applications (NCSA)"*
# Blue Waters Petascale Computing System

## Blue Waters Computing System

<table>
<thead>
<tr>
<th>System Attribute</th>
<th>Typical Cluster (NCSA Abe)</th>
<th>Track 2 (TACC)</th>
<th>Blue Waters*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vendor</strong></td>
<td>Dell</td>
<td>Sun</td>
<td>IBM</td>
</tr>
<tr>
<td><strong>Processor</strong></td>
<td>Intel Xeon 5300</td>
<td>AMD</td>
<td>Power 7</td>
</tr>
<tr>
<td><strong>Peak Perf. (PF)</strong></td>
<td>0.090</td>
<td>0.58</td>
<td>~10</td>
</tr>
<tr>
<td><strong>Sustained Perf. (PF)</strong></td>
<td>~0.005</td>
<td>~0.06</td>
<td>~1.0</td>
</tr>
<tr>
<td><strong>Number of cores</strong></td>
<td>9,600</td>
<td>62,976</td>
<td>&gt;300,000</td>
</tr>
<tr>
<td><strong>Amount of Memory (PB)</strong></td>
<td>0.0144</td>
<td>0.12</td>
<td>&gt;1.0</td>
</tr>
<tr>
<td><strong>Amount of Disk Storage (PB)</strong></td>
<td>0.1</td>
<td>1.73</td>
<td>&gt;18</td>
</tr>
<tr>
<td><strong>File system Performance (GB/s)</strong></td>
<td>11</td>
<td>30</td>
<td>&gt;1500</td>
</tr>
<tr>
<td><strong>Amount of Archival Storage (PB)</strong></td>
<td>11</td>
<td>30</td>
<td>~500</td>
</tr>
<tr>
<td><strong>External Bandwidth (Gbps)</strong></td>
<td>40</td>
<td>10</td>
<td>100-400</td>
</tr>
</tbody>
</table>

*Reference petascale computing system (no accelerators).
Storage Management – BW Approach

*Have the right data at the right place at the right time*

- Blue Waters will proactively use the new storage functions to implement a new state of the practice in HPC storage management (hours to fill diskcache, day to write to tape)
- Goal – no pain (for users anyway 😊)
  - To have one extremely large storage space – with on-line and near-line limits
  - Approach storage as with virtual memory
    - Large virtual storage
    - Limited work sets of data
    - Try to keep the data with the most temporal locality in the highest (fasted) levels of storage when it is needed
- Goal vs reality needs to be explored
- Fall back is to implement a more standard
## Possible Layout

<table>
<thead>
<tr>
<th>Usage</th>
<th>File System</th>
<th>On-line Usable Capacity</th>
<th>Near-Line Capacity</th>
<th>Managed</th>
<th>Quota</th>
<th>Backup</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Home Directories</td>
<td>Midperf</td>
<td>4PB</td>
<td>20 PB?</td>
<td>Yes</td>
<td></td>
<td>Yes - via GHI – relatively rapid backup (&gt; 24 hours? residency) All files &gt; 1MB Metadata backed up weekly</td>
</tr>
<tr>
<td>High Performance Large Files</td>
<td>Highperf</td>
<td>14 PB</td>
<td>480PB?</td>
<td>Yes</td>
<td></td>
<td>Yes – via GHI Longer (&gt; 7 day residency?) Metadata backed up before upgrades. Alternative is to subdivide with and without GHI</td>
</tr>
<tr>
<td>Large Scale Test</td>
<td>Test</td>
<td>.2PB</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>For new system testing</td>
</tr>
</tbody>
</table>

- Midperf: Midperformance
- Highperf: High performance
- Test: Test
- PB: Petabyte
- GHI: Grid High-Performance Internet
On going research

• Batch jobs –
  • users tell us ahead of time
    • what objects need to be online before job can be started
    • how much storage space is needed for the job
  • NCSA behind the scenes will move up the data from near-line(on-demand stage) or from across country (gridftp)
  • Using attributes in GPFS to “lock” files on disk so that they don’t get “punched or purged” before all the data is on-line.
On going research -

- What files need to stay on disk for further analysis? (post analysis)
  - what can go to archive immediately (safe keeping),
  - what can be deleted? Checkpoints?
- Post job data management step
On going research -

• For retrieval: how will the files need to be associated together
  • Using GPFS filesets for the PRAC projects
  • Researching the filesets environments
    • so policy scans can be run in parallel over filesets
    • quotas implemented at fileset level
    • use HPSS family of files for project from GPFS filesets
National Petascale Computing Facility at a Glance

- 88,000 GSF over two stories—45’ tall
  - 30,000+ GSF of raised floor
  - 20,000+ unobstructed net for computers
  - 6’ clearance of raised floor
- 24 MW initial power feeds + backup
  - Three 8 MW feeds + One 8 MW for backup
  - 13,800 volt power to the each
- 5,400 Tons of cooling
  - Full water side economization for 50%+ of the year
  - Automatic Mixing of mechanical and ambient chilled water for optimal efficiency
  - Adjacent to (new) 6.5M gallon thermal storage tank
- 480 Volt distribution to computers
- Energy Efficiency
  - PUE - ~1.02 to <1.2 (projected)
  - USGBC LEED Silver-Gold (Platinum?) classification target

www.ncsa.illinois.edu/BlueWaters
Questions? See me

Michelle Butler
NCSA/University of Illinois
Technical Program Manager
mbutler@ncsa.illinois.edu- http://www.ncsa.uiuc.edu/BlueWaters