Lustre at the Australian National Computational Infrastructure (NCI)

Daniel Rodwell  (NCI)
daniel.rodwell@anu.edu.au

Shuichi Ihara  (DDN)
sihara@ddn.com
Agenda

• What is NCI?
• Petascale Machine at NCI (Raijin)
• Root over Lustre
• Lustre Storage on the Petascale Machine
• Other Lustre Storage at NCI
• Future Plans & Collaboration Possibilities
• Lustre patch - source contrib
WHAT IS NCI?
In case you’re wondering where are we located?

- In the Nation’s capital, at its National University ...
• NCI is Australia’s national high-performance computing service
  – comprehensive, vertically-integrated research service
  – providing national access on priority and merit
  – driven by research objectives

• Operates as a formal collaboration of ANU, CSIRO, the Australian Bureau of Meteorology and Geoscience Australia

• As a partnership with a number of research-intensive universities, supported by the Australian Research Council.
Our mission is to foster ambitious and aspirational research objectives and to enable their realisation, in the Australian context, through world-class, high-end computing services.
**Specialised Support**
- Climate Science and Earth System Science
- Astronomy (optical and theoretical)
- Geosciences: Geophysics, Earth Observation
- Biosciences: Bioinformatics
- Social Sciences

- Growing emphasis on data-intensive computation
  - Cloud Services
  - Earth System Grid
Specialised Support: Scientific Visualisation

- NCI VizLab in existence since early-1990s
- Innovative software development (Drishti and Voluminous)
- Skilled visualisation programmers who deal with multi-terabyte datasets
- Lustre use-case: access from visualization desktops, driving video walls, on-demand GPU clusters, on-demand volume visualization

http://youtu.be/1JxUYUKSnLs
Engagement: Government Infrastructure Initiatives

- **Engagement with RDSI and NeCTAR**
  - Approximately $100M in funding from the Australian Federal Government
  - RDSI – National Storage Initiative
    - NCI High-Performance Data Node
      - Hosting data collections of national importance, seeding storage initiatives across the country
  - NeCTAR – National Research Cloud Initiative
    - High-Performance node of NeCTAR Cloud
  - Major Participant in Virtual Labs (VLs)
    - Weather and Climate VL
    - All-Sky Virtual Observatory VL
    - Contributing to Characterisation VL, Virtual Exploration Geophysics Laboratory (VEGL)
  - Tools—volume visualisation in the cloud
PRIORITY SCIENCE AREAS
Engaging with Priority Research: Climate

Case Study: Building a National Climate Modelling Capability

Partners: CAWCR (Bureau of Met, CSIRO), ARC Centre for Climate Systems Science, NCI, Fujitsu

Goals:

• Enhance the value of investment in ACCESS model development
• Harness and develop Australia’s international value in Climate Research (CAWCR + AU Universities)
• Build research infrastructure in harmony with operational environment

Requirements:

• High Performance Computing at NCI available at competitive level to support Climate
• Provide integrated environment for supporting:
  – Simulations
  – Data repository: Online and Deep Archive
  – Cloud capability for data processing, analysis and visualization
INFRASTRUCTURE
Background: Evolution of Peak Facilities at NCI/APAC

<table>
<thead>
<tr>
<th>System</th>
<th>Procs/ Cores</th>
<th>Memory</th>
<th>Disk</th>
<th>Peak Perf. (Tflops)</th>
<th>Sustained Perf. (SPEC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001–04 Compaq Alphaserver (31)</td>
<td>512</td>
<td>0.5 Tbyte</td>
<td>12 Tbytes</td>
<td>1 TFlop</td>
<td>2,000</td>
</tr>
<tr>
<td>2005–09 SGI Altix 3700 (26)</td>
<td>1920</td>
<td>5.5 Tbytes</td>
<td>30 (+70) Tbytes</td>
<td>14 Tflops</td>
<td>21,000</td>
</tr>
<tr>
<td>2008–12 SGI Altix XE (-)</td>
<td>1248</td>
<td>2.5 Tbytes</td>
<td>90 Tbytes</td>
<td>14 TFlops</td>
<td>12,000</td>
</tr>
<tr>
<td>2009–13 Sun Constellation (35)</td>
<td>11,936</td>
<td>37 Tbytes</td>
<td>800 Tbytes</td>
<td>140 TFlops</td>
<td>240,000</td>
</tr>
<tr>
<td>2013–Fujitsu Primergy (24)</td>
<td>57,500</td>
<td>160 Tbytes</td>
<td>12.5 Pbytes</td>
<td>1200 Tflops</td>
<td>1,400,000+</td>
</tr>
</tbody>
</table>
Fujitsu Primergy Petascale System (2013–)
Current Infrastructure - Compute

- **Raijin**—Fujitsu Primergy cluster—June 2013
- Approx. 57,500 Intel Sandy Bridge (2.6 GHz)
- 157 TBytes memory, 8 PBytes short term storage
- FDR Infiniband
- 150 GB/s bandwidth to filesystem
- Centos 6.4 Linux; PBS Pro scheduler
- Good Performance — well balanced, appreciated
  - 1195 Tflops, 1,400,000 SPECFPrate
- Significant growth in highly scaling application codes
  - Largest: 40,000 cores; many 1,000 core tasks
Data Storage

- HSM (massdata) – DMF based: 8PB as at September 2012 [2 copies]

- /projects: SGI CXFS (Interactive f/s space) HSM (shared with massdata), achieves 2.5 GB/sec from tape

- Global Lustre Filesystem
  - 6 PB by end of 2013 and growing
  - Global bandwidth: 25 GB/sec
Current Infrastructure - Cloud

- VMware ESX cluster—providing mission-critical hosting of essential services in a high availability environment
  - DCC: Specialised cluster for data-intensive applications
    - Climate, earth-system observation and bioinformatics
    - Part virtualized, part bare-metal

- Cloud computing
  - NeCTAR Research Cloud node at NCI
    - Australia’s highest performance cloud
    - Architected for strong computational and I/O performance needed for “big data” research
    - Intel Sandy Bridge (3200 cores)
    - 160 TB of SSDs; 56GigE + RoCE for compute and I/O performance
    - Planning to use RoCE for LNET
  - Private cloud: RedHat OpenStack
    - SLA centric, on-demand scientific computation
How does all of the pieces link together?
ROOT OVER LUSTRE
What is root over Lustre?
- The root filesystem is provided by Lustre
- We use oneSIS for provisioning with minor patches

Why?
- Simplicity: Ease of management
  - Diskless compute nodes
  - One golden image for multiple clusters
  - ‘yum update’ the entire cluster
- Synchronous: Rolling out updates
  - Once an update is made, all nodes see it
- Security: Better/Coherent patching

We have been using root over Lustre since 2008
• Key feature: oneSIS loads Lustre kernel modules and parses the location of the root filesystem from its boot command line:

```
lustreroot=10.9.103.1@o2ib3:10.9.103.2@o2ib3:/images/NCI/centos-6.4-compute-03
```

• NCI implements root-over-lustre by modifying oneSIS. Work done by Robin Humble

• IB Flexboot provides boot over IB
• Initial bugs ironed out
• Planning to roll into next scheduled downtime window

<table>
<thead>
<tr>
<th># of Nodes</th>
<th>Time to boot (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Node</td>
<td>6 min.</td>
</tr>
<tr>
<td>4 Nodes (1 chassis)</td>
<td>6 min.</td>
</tr>
<tr>
<td>72 Nodes (1 rack)</td>
<td>7 min. (±11 seconds)</td>
</tr>
</tbody>
</table>
LUSTRE ON RAIJIN
Storage Architecture of Raijin

- Storage for the Petascale machine provided by DDN SFA block appliances

- 5 storage building blocks of SFA12K40-IB with 10 x SS8460, 84 bay disk enclosures

- Each building block:
  70 x RAID6 (8+2) 3TB 7.2k SAS pools
  20 x RAID1 (1+1) 3TB 7.2k SAS pools
  40 x RAID1 (1+1) 900GB 10k SAS pools
  12 x 3TB 7.2k SAS hot spares
  8 x 900GB 10k SAS hot spares

- Building blocks scale diagonally with both capacity & performance
Fully redundant SAS enclosures, FDR IB between storage and OSSes and uplinks to Raijin
Metadata storage is based on the DDN EF3015 storage platform.

Each metadata storage block has 12 RAID1 (1+1) 300GB 15kSAS pools. There are 2/4 storage blocks for each MDS.

Fully redundant Direct Attached FC8 fabric.

Fully redundant FDR IB uplinks to main cluster IB fabric.
• Lustre servers are Fujitsu Primergy RX300 S7
  Dual 2.6GHz Xeon (Sandy Bridge) 8-core CPUs
  128/256GB DDR3 RAM

  6 MDS (3 HA pairs)
  50 OSS (25 HA pairs)

• All Lustre servers are diskless
  Current image is CentOS 6.3, Mellanox OFED 2.0, Lustre v 2.1.6, corosync/pacemaker
  (image was updated 8 September – simply required a reboot into new image)
  HA configuration needs to be regenerated whenever a HA pair is rebooted

• 5 Lustre file systems:
  /short – scratch file system (rw)
  /images – images for root over Lustre used by compute nodes (ro)
  /apps – user application software (ro)
  /home – home directories (rw)
  /system – critical backups, benchmarking, rw-templates (rw)
• NCI MDS requirements:

*MDT Storage on LVM on top of software RAID1 configuration of hardware RAID1 LUNs*
- 4-way mirror (1+1) + (1+1).

• NCI acceptance testing requirements for the scratch file system, **/short**

*Demonstrate IOR exceeds 120GB/s for sustained streaming write performance:*
*Achieved 143 GB/s (Updated after reconfiguration 152 GB/s)*

*Demonstrate IOR exceeds 7.5GB/s for random 1MB write performance:*
*Achieved 75.5 GB/s*

*Demonstrate mdtest test can create, stat and delete 65536 files in a shared directory within 53 seconds:*

**Achieved**
- File Creation 3.57s
- File Stat 2.88s
- File Delete 6.20s
- **Total 12.65s**
Lustre on Raijin: Sequential IO Performance

- Graph showing Aggregate Sequential IO performance with read and write bandwidths.
- Y-axis: Bandwidth (MB/s)
- X-axis: No. of Processes
- Graph displays steady performance as the number of processes increases.
### Lustre on Raijin: Filesystem composition

<table>
<thead>
<tr>
<th>File System</th>
<th>RAID</th>
<th>OST/OSS</th>
<th>Total OST</th>
<th>Total Size</th>
<th>Performance*</th>
</tr>
</thead>
<tbody>
<tr>
<td>/short</td>
<td>RAID6 (8+2) 7.2k SAS</td>
<td>7</td>
<td>350</td>
<td>7.5PB</td>
<td>152 GB/s</td>
</tr>
<tr>
<td>/images</td>
<td>RAID1 (1+1) 10k SAS</td>
<td>2</td>
<td>100</td>
<td>80TB</td>
<td>17.8 GB/s**</td>
</tr>
<tr>
<td>/apps</td>
<td>RAID1 (1+1) 10k SAS</td>
<td>2</td>
<td>100</td>
<td>80TB</td>
<td>17.9 GB/s**</td>
</tr>
<tr>
<td>/home</td>
<td>RAID1 (1+1) 7.2k SAS</td>
<td>1</td>
<td>50</td>
<td>135TB</td>
<td>6.9 GB/s**</td>
</tr>
<tr>
<td>/system</td>
<td>RAID1 (1+1) 7.2k SAS</td>
<td>1</td>
<td>50</td>
<td>135TB</td>
<td>8.1 GB/s</td>
</tr>
</tbody>
</table>

* Aggregate Sequential write bandwidth with IOR (Aug 2013)
** File system was not idle

- Currently investigating a Lustre read performance issue:
  During acceptance testing in Dec 2012 /short read performance was 160 GB/s.
  From later benchmarking (May 2013) /short read performance is 88 GB/s
SITE-WIDE LUSTRE
• In order to avoid moving data between clusters and storage, the NCI has implemented a site-wide Lustre F/S, visible both to compute clusters and virtual machine hosts

• We have decided to use islands of storage to create multiple Lustre F/S which are vendor/technology specific
2 x Management Servers: oneSIS, DNS, DHCP etc

4 x NFS servers

21 x LNET routers connected to Raijin

9 Mellanox FDR IB switches, in non-blocking fat tree fabric

72 x OSSes

13 x SGI arrays hosting OSTs

3 x DDN SFA12K hosting OSTs

2 x MDSs

Legend
- 8 Gb MM FC
- 10 GigE
- 56Gb FDR IB

NCI Data Centre

/g/data1 ~8.3PB
/g/data2 ~6.4PB

2 x SGI arrays hosting 4 MDTs

NCI Data Centre

www.nci.org.au
**Site-wide Lustre – Functional Composition**

**Array Performance Expectations for new /gdata1, without LNET routers**
- 48GB/sec read
- 43GB/sec write

**LNET Performance Expectation**
- 25GB+/sec using 21 routers to Raijin

Legend
- 8 Gb MM FC
- 10 GigE
- 56Gb FDR IB
- 2 x SGI arrays hosting 4 MDTs
- 2 x MDSs

NCI Data Centre

/g/data1 ~8.3PB
/g/data2 ~6.4PB
Future Plans & Collaboration Possibilities

• Site-wide Lustre to tie together HPC, Cloud and Visualization

• Complex workflows, post-simulation, will use the NCI’s NeCTAR OpenStack node, and requires access to Lustre

• We are keen to implement in the future: Lustre HSM, WAN and Kerberos feature sets
A recent bug fix on Raijin

WORKING WITH LUSTRE UPSTREAM
Bug fix by Dongyang Li (NCI)

- `open(2)` an existing file with `O_RDONLY | O_CREAT` fails with `-EROFS`, ro mounted client

- With thanks to Dale Roberts and Dr Andrey Bliznyuk (NCI, User Support Services)
/apps is read only mounted
10.9.103.2@o2ib3:10.9.103.1@o2ib3:/apps on /apps type lustre
(ro,nosuid,localflock)

Got -EROFSc when trying to run Fortran apps.

open("/apps/foo", O_RDONLY|O_CREAT, 0600) = -1 EROFS
(Read-only file system)

GCC fortran runtime tends add O_CREAT flag
when opening files.

Remove the flag will work. However...
On a local ro mounted ext4:

- open("/apps/bar", O_RDWR|O_CREAT, 0600) = -1 EROFS (Read-only file system)
- open("/apps/bar", O_WRONLY|O_CREAT, 0600) = -1 EROFS (Read-only file system)
- open("/apps/bar", O_RDONLY|O_CREAT, 0600) = 3
Open is handled on MDS.

- if (client is ro mount && the open has O_CREAT)
  -EROFS;

At least check if the file is there first!
Don't bother the O_CREAT if we can do the open.

Patch landed for 2.5 release.
• A JIRA ticket has been opened to track the issue
  – [https://jira.hpdd.intel.com/browse/LU-3557](https://jira.hpdd.intel.com/browse/LU-3557)
  – Affected versions: 2.1.X to 2.4.0
• Test and commit your patch locally using acceptance-small.sh
• The patch follows the Requirements for patch submission:
  – The Commit Comments are well formatted and useful
  – Verify patch follows Lustre Coding Guidelines (at least "git show | contrib/scripts/checkpatch.pl -" passes)
  – A regression test has been created that fails without the patch and passes with the patch
  – The patch has the appropriate signed off by line
• The patch has been uploaded to Gerrit
  – git push ssh://USER@review.whamcloud.com:29418/fs/
    lustre-release HEAD:refs/for/master
  – http://review.whamcloud.com/#/c/6893/
• Request at least two Patch Inspection approvals (preferably ones with experience in this area of code) on the Gerrit change
• The branch gatekeeper will review the patch, confirm the test results, and submit it when everything goes well
  – Automatically Build and Test
    • Triggered once you refresh the patch.
    • After a successful build, automatic test system maloo will spin up VMs, carry out the test using Auster, finally push the result back to review board.

Find more at
https://wiki.hpdd.intel.com/display/PUB/Submitting+Changes

Finally backport the patch to Raijin :-}
Thank you

Visit us at:

www.nci.org.au

Daniel Rodwell
daniel.rodwell@anu.edu.au

Shuichi Ihara
sihara@ddn.com