Lustre at the Australian National Computational Infrastructure (NCI)

Joseph Antony  (NCI)
joseph.antony@anu.edu.au

Patrick Fitzhenry  (DDN)
pfitzhenry@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
WHAT IS NCI?
NCI – an overview

Mission:
• to foster ambitious and aspirational research objectives and to enable their realisation, in the Australian context, through world-class, high-end computing services

NCI is:
• being driven by research objectives
• a comprehensive, vertically-integrated research service
• providing national access on priority and merit, and
• being built on, and sustained by, a collaboration of national organisations and research-intensive universities
In case you’re wondering where are we located?

- In the nation’s capital, at its national university ...
Engagement: Research Communities

• **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
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, VEGL

- Tools—volume visualisation in the cloud
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
PRIORITY SCIENCE AREAS
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
VIDEO: ANDY HOGG @ ANU

http://youtu.be/zUF2rsq7ej8
CURRENT INFRASTRUCTURE
## Background: Evolution of Peak Facilities at NCI/APAC

<table>
<thead>
<tr>
<th>System (Top500 rank)</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, 10 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
  – 4.4 PB by end Sept 2012 and growing
  – Global bandwidth: 25 GB/sec
  – Migrating /projects off of CXFS

• Object Storage: Ceph
  – Initially object store for NeCTAR cloud
  – Considering use in long term on-disk copy when erasure coding backend is mature
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
      – RoCE for LNET

  – Private cloud: RedHat OpenStack
    • SLA centric, on-demand scientific computation
How does all of the pieces link together?
How does all of the pieces link together?

Today's talk is about these Lustre F/S
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

Boot chart for r1 (Thu Sep 12 17:37:46 EST 2013)
uname: Linux 2.6.32-358.14.1.el6.x86_64 #1 SMP Tue Jul 16 23:51:20 UTC 2013 x86_64
release: CentOS release 6.4 (Final)
CPU: Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz (16)
kernel options: selinux=0 exec-shield=0 audit=0 console=tty0 console=ttyS0,115200n8 ro initrd=initramfs
time: 3:22

<table>
<thead>
<tr>
<th></th>
<th>5s</th>
<th>10s</th>
<th>15s</th>
<th>20s</th>
</tr>
</thead>
<tbody>
<tr>
<td>init</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Boot Chart for Root over Lustre

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>5s</td>
<td>kthreadd</td>
</tr>
<tr>
<td>10s</td>
<td>infiniband/17</td>
</tr>
<tr>
<td>15s</td>
<td>ib_mcast</td>
</tr>
<tr>
<td>20s</td>
<td>mlx4_ib</td>
</tr>
<tr>
<td>25s</td>
<td>mlx4_ib_mcq</td>
</tr>
<tr>
<td>30s</td>
<td>ib_mad1</td>
</tr>
</tbody>
</table>

### Additional Events
- to_fifo
- ib_cm/0
- ipoib
- ipoib_auto_mode
- iw_cm_wq
- ib_addr
- rdma_cm
- cfs_wi_sd000
- wi_serial_sd
- obd_zombid
- ptrpc_hr_0
- kibind_sd_00
- kibind_connd
- router_checker
## Boot Chart for Root over Lustre

<table>
<thead>
<tr>
<th></th>
<th>80s</th>
<th>85s</th>
<th>90s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ptlrpcd-brw</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ptlrpcd-brw-rcv</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ptlrpcd</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ptlrpcd-rcv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ll</td>
<td>ping</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sptlrpc_qc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ll</td>
<td>capa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>bl 00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>bl 01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>cn 00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>watchdobjd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>cn 01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>cb 00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>cb 01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>elt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ldlm</td>
<td>poold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ll</td>
<td>close</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scsi</td>
<td>eh 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scsi</td>
<td>wg 0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Boot Chart for Root over Lustre

- 80s
  - sh
  - rc.preinit
- 85s
  - bootchartd
  - sleep
  - rc.sysinit
  - start_udev
  - udevadm
  - mount
  - mount.lustre
- 90s
  - udevd
  - modprobe
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)
Lustre on Raijin: Acceptance Requirements

• 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

**Aggregate Sequential IO (/short)**

- **Write (MB/s)**
- **Read (MB/s)**

![Graph showing the performance of Lustre on Raijin for sequential IO operations.](image-url)
**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) 7.2kSAS</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

13 x LNET routers connected to Raijin

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

2 x SGI arrays
13 x SGI arrays hosting OSTs
1 x SFA12K hosting OSTs

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

2 x MDSs

60x OSSes

2 x MDSs

NCI Data Centre
2 x Management Servers: oneSIS, DNS, DHCP etc

4 x NFS servers

13 x LNET routers connected to Raijin

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

60x OSSes

This FS is called /gdata1 and runs Intel Enterprise Lustre

Legend

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

NCI Data Centre

www.nci.org.au
Array Performance Expectations for new /gdata1, without LNET routers
- 48GB/sec read
- 43GB/sec write

LNET Performance Expectation
- 25GB/sec using 13 routers to Raijin
Site-wide Lustre – Functional Composition

NCI Data Centre

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

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

- 2 x Management Servers: oneSIS, DNS, DHCP etc
- 4 x NFS servers
- 13 x LNET routers connected to Raijin
- 13 x SGI arrays hosting OSTs
- 1 x SFA12K hosting OSTs

2 x SGI arrays hosting 4 MDTs

60x OSSes

This F/S is called /gdata2, runs IEL and uses same MDS as /gdata1
• 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 Lustre HSM, WAN and Kerberos feature sets