Lustre Update

- Brent Gorda
  CEO
  Whamcloud, Inc.
  bgorda@whamcloud.com
Agenda

- Lustre History / Whamcloud introduction

- Current status
  - Community
  - Roadmap
  - Current development

- Looking forward
  - Growing the Lustre ecosystem
  - What to expect in 2012
  - Exascale thoughts
Lustre Major Releases

- 1999 – Lustre project startup
- 2001 – ASCI Pathforward
- 2003 V1.0 – CFS
- 2007 V1.6 – Sun
- 2009 V1.8 – Sun
- 2010 V2.0 – Oracle
- 2011 V2.1 – Whamcloud
- 2012 V2.2 soon - Whamcloud
Whamcloud introduction

- VC-backed California Corp. Formed July 2010
- ~50 people worldwide
  - Unique advantage: critical mass for Lustre technology
- ~150 supported sites worldwide
- Today’s offerings:
  - Worldwide L3 Lustre support
  - Lustre development
  - Training
  - Chroma - management

Whamcloud is widely recognized as the source for Lustre
We have the only HW vendor-neutral offering
Whamcloud management team

CEO Brent Gorda (> 25 yrs HPC)
- DOE program manager for Lustre, BlueGene, TLCC
- Entrepreneur: Myrias Research, Bonsai Software, MetaExchange

CTO Eric Barton (> 25 yrs in HPC)
- LNet developer since 2002, Lustre architect since 2008
- Entrepreneur: Meiko Scientific founder, Quadrics & consulting

Daniel Ferber – Cray/SGI/Sun/Oracle, bizdev
Peter Jones – Lustre support manager
Bryon Neitzel – Lustre development manager
Jessica Popp – Lustre project management
Robert Read – Lustre 2.0 lead engineer, Chroma
**Lustre community**

- Whamcloud community membership
  
  ![EOFS](http://www.eofs.eu) ![OpenSFS](http://www.opensfs.org)


- Whamcloud maintains the community assets
  - Jira bug tracker: [http://bugs.{OpenSFS.org,whamcloud.com](http://bugs.{OpenSFS.org,whamcloud.com}
  - Git repositories: [http://git.whamcloud.com](http://git.whamcloud.com)
  - Gerrit code review: [http://review.{OpenSFS.org,whamcloud.com](http://review.{OpenSFS.org,whamcloud.com}
  - Build: [http://build.whamcloud.com](http://build.whamcloud.com)

- No copyright assignment on source contributions
  - Ensures no single entity can own whole copyright on Lustre
  - Has support of OpenSFS and EOFS
Lustre Development Activity

- **Lustre development**
  - Majority development at Whamcloud
  - Majority of the same engineers who worked on Lustre through its history
  - Long range architectural vision for a reliable roadmap

- **Lustre tree maintenance**
  - Stability a priority
  - Lustre test automation & QA
  - Testing at scale
  - Test results database

- **Lustre support**
  - Direct access to developers
  - Experienced Lustre engineers worldwide
  - Minimum time to incident resolution
  - Solutions avoid technical debt
On Lustre quality

"The Spider parallel file systems at Oak Ridge National Laboratory, three of the largest parallel file systems in the world, powered by Data Direct Networks, Dell, Mellanox, and Lustre supported by Whamcloud, have provided unprecedented reliability in supporting the I/O needs of all major compute systems at our Leadership Computing Facility. On our largest file system we have **achieved 100% availability** in 8 of the 12 months of 2011 with scheduled availability of **99.26% over the entire year**. The stability of this system is unprecedented given its scale."

-- Galen M. Shipman,
Technology Integration Group Leader,
National Center for Computational Sciences (NCCS) at ORNL
Quality achieved by planning/ testing

<table>
<thead>
<tr>
<th>Test Case</th>
<th>2.2.50 368b67d 2012-03-06</th>
<th>2.1.56 e41a9f0 2012-02-18</th>
<th>2.1.55 1255aa5 2012-01-25</th>
<th>2.1.54 107a010 2012-01-11</th>
<th>2.1.52 31569d7 2011-11-12</th>
<th>2.1.0 9d71fe8 2011-09-16</th>
<th>2.1.0-RC1 1f1c672 2011-08-23</th>
</tr>
</thead>
<tbody>
<tr>
<td>conf-sanity</td>
<td>☉ 3/3</td>
<td>☉ 6/6</td>
<td>☉ 6/13</td>
<td>☉ 2/3</td>
<td>☉ 7/7</td>
<td>☉ 1/20</td>
<td>☉ 1/2</td>
</tr>
<tr>
<td>insanity</td>
<td>☉ 3/3</td>
<td>☉ 5/5</td>
<td>☉ 8/8</td>
<td>☉ 3/3</td>
<td>☉ 7/7</td>
<td>☉ 20/20</td>
<td>☉ 3/3</td>
</tr>
<tr>
<td>large-scale</td>
<td>☉ 2/2</td>
<td>☉ 5/5</td>
<td>☉ 7/7</td>
<td>☉ 3/3</td>
<td>☉ 7/7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lfsck</td>
<td>☉ 2/3</td>
<td>☉ 7/8</td>
<td>☉ 1/10</td>
<td>☉ 1/3</td>
<td>☉ 5/7</td>
<td>☉ 23/25</td>
<td>☉ 2/2</td>
</tr>
<tr>
<td>liblustre</td>
<td>☉ 3/5</td>
<td>☉ 10/16</td>
<td>☉ 2/3</td>
<td>☉ 1/7</td>
<td>☉ 7/25</td>
<td>☉ 2/2</td>
<td></td>
</tr>
<tr>
<td>inet-selftest</td>
<td>☉ 2/2</td>
<td>☉ 5/5</td>
<td>☉ 7/7</td>
<td>☉ 3/3</td>
<td>☉ 7/7</td>
<td>☉ 16/16</td>
<td>☉ 3/4</td>
</tr>
<tr>
<td>lustre-rsync-test</td>
<td>☉ 2/2</td>
<td>☉ 5/5</td>
<td>☉ 9/9</td>
<td>☉ 2/2</td>
<td>☉ 7/7</td>
<td>☉ 13/13</td>
<td></td>
</tr>
<tr>
<td>metadata-updates</td>
<td>☉ 2/2</td>
<td>☉ 5/5</td>
<td>☉ 10/10</td>
<td>☉ 3/3</td>
<td>☉ 7/7</td>
<td>☉ 20/20</td>
<td>☉ 3/3</td>
</tr>
<tr>
<td>mmp</td>
<td>☉ 2/2</td>
<td>☉ 5/5</td>
<td>☉ 8/8</td>
<td>☉ 2/3</td>
<td>☉ 7/7</td>
<td>☉ 18/19</td>
<td>☉ 4/4</td>
</tr>
<tr>
<td>obdfilter-survey</td>
<td>☉ 2/2</td>
<td>☉ 5/5</td>
<td>☉ 6/8</td>
<td>☉ 3/3</td>
<td>☉ 0/7</td>
<td>☉ 16/16</td>
<td>☉ 4/4</td>
</tr>
</tbody>
</table>
Community Lustre Roadmap

Maintenance Releases every quarter
- 1.6.6
- 1.6.7
- 2.1.1
- 2.1.2
- RHEL6 client support
- 24TB ext4 LUNs

Feature Releases
- 2.1
- 2.2
- 2.3
- 2.4
- Full RHEL6 support
- Async journal commits
- 128TB ext4 LUNs
- Stability enhancements
- Imperative Recovery
- Dirop SMP Scaling
- Wide Striping
- Statahead performance
- Server Stack SMP Scaling
- Online check/scrub
- OSD restructuring
- Distributed namespace
- HSM

Q2 2011
Q3 2011
Q4 2011
Q1 2012
Q2 2012
Q3 2012
Q4 2012
Q1 2013
Q2 2013

Sponsor for Whamcloud Development: ORNL, OpenSFS, LLNL, Whamcloud
Third Party Development: CEA
Lustre development at Whamcloud

http://wiki.whamcloud.com/display/PUB/Community+Lustre+Roadmap

- Imperative recovery
  - Reduced server recovery latency
- Dirop SMP scaling
  - Improve MDS throughput
- Server stack SMP scaling
  - Improve SMP node affinity for LNET and ptlrpc
- Statahead performance
  - Fast `ls -l`, `du` etc
- Wide striping
  - Allow files to stripe over ~1500 OSTs
- OSD restructuring
  - Single backend storage abstraction (ext4, ZFS, btrfs?, others?)
- Online check/scrub
  - Filesystem consistency checking and repair
- Distributed Namespace
  - Horizontal scale for metadata
Lustre 2.3 and Beyond

- Lustre 2.3 (Q3 2012)
  - Server Stack SMP Scaling
  - Online check/scrub

- Lustre 2.4 (Q1 2013)
  - OSD Restructuring
  - Distributed Namespace
  - HSM

- A great deal of other development is underway but not tied to a specific release yet

- Roadmap refreshes now happening regularly
Chroma Manager (Appliance)

- Administration
  - Provisioning
  - Maintenance
  - HA Setup
  - Fault diagnosis
- Management information
  - Performance
  - Utilization
  - Alerts
- Intuitive interfaces
  - GUI (single pane of glass)
  - Scriptable CLI (automation)
- System integration
  - Multi-vendor storage management
  - Multi-vendor cluster/site management
  - Partners can build their own appliance
Exascale filesystem

- Conventional namespace
  - Works at human scale
  - Administration, security, accounting
  - Supports legacy data and applications

- DAOS Container
  - Works at exascale
  - Embedded in conventional namespace
  - Provides storage for application domain specific object schemas (data + metadata)
Take away from this talk

Lustre is stable today

- Technically: v1.8.x is rock solid, 2.x features
- Politically: the community has stepped up
  - Whamcloud is the safe hands on the tree
  - All development expected to be done in the open for review

The Ecosystem is growing:

- EOFS + OpenSFS + Whamcloud
- Single tree from which to pull
- Active dev community: Whamcloud + others
- More storage vendors shipping Lustre today
"The maturation of the Lustre community, anchored by OpenSFS and Whamcloud, was the most significant trend in HPC in 2011."

– Addison Snell, CEO of Intersect360 Research