Lustre Development Update

- Dan Ferber
  Whamcloud, Inc.
  dferber@whamcloud.com
Agenda

- **Lustre current status**
  - Community & Roadmap
  - Releases
  - Current development processes

- **Looking forward**
  - Chroma™
  - Exascale
Whamcloud Today

• ~55 people worldwide
  - Unique advantage:
    • Critical mass for Lustre technology
    • Time to delivery or resolution (engineers, project managers)

• ~175 supported sites worldwide

• Our offerings:
  - Worldwide Lustre support
  - Lustre development
  - Training
  - Community releases (OpenSFS)
  - Chroma

Whamcloud is widely recognized as the source for Lustre
We have the only HW vendor-neutral offering
Broad Community Success

Lustre is stable today

- Technically: v1.8.x is very solid, 2.x features
- Politically: the community has stepped up
  - The tree is safe, stable, and reliable
  - 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
Lustre Community - Resources

- Whamcloud community membership
  http://www.opensfs.org/release-planning-group/technical-working-group

- Whamcloud maintains the community assets
  - All Lustre releases:  http://www.whamcloud.com/downloads
  - Jira bug tracker:  http://bugs.{OpenSFS.org,whamcloud.com}
  - Git repositories:  http://git.whamcloud.com
  - Gerrit code review:  http://review.{OpenSFS.org,whamcloud.com}
  - Build:  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
LUG 2012 in Austin, Texas
http://insidehpc.com/category/events/lug-2012/
Whamcloud Community Lustre Releases

- Single community-wide source tree
  - Hosted at Whamcloud, tested and available via RPM
  - Formally recognized by community

- Whamcloud defines two release streams
  - The designated maintenance release stream is targeted for conservative users wishing to use a well-proven release
  - The feature release stream is targeted for those requiring first access to new features

- Bugfix releases every quarter for maintenance release stream
  - Currently 2.1.x is the designated maintenance release stream
  - Use combination of support metrics and customer demand to determine when to switch

- Feature releases every six months
  - Not all feature releases will have maintenance releases
Lustre 1.8.x

- Whamcloud continues to support Lustre 1.8.x
  - We will offer Lustre 1.8.x support as long as there is sufficient demand
- Whamcloud to release 1.8.8-wc1
  - A supplemental release to provide RHEL6.2 support
Lustre 2.1.x

- Lustre 2.1.0 released Q3 2011
- Lustre 2.1.1 released in Q1
  - Provided bugfixes and RHEL6.2 server and client support
- Lustre 2.1.2 scheduled for release in Q2
  - Will incorporate bugfixes from large sites running 2.1.x in production
- Intention is to match Lustre 1.8.x for stability
  - Still early days in this process
Lustre 2.2

- Released March 30th as per the schedule
- Changelog is available at: [http://wiki.whamcloud.com/display/PUB/Changelog+2.2](http://wiki.whamcloud.com/display/PUB/Changelog+2.2)
- **First release to benefit from OpenSFS funding**
  - Covers some of the costs of maintaining the community tree
  - The remaining costs are covered by Whamcloud support contracts
- **Regular updates throughout release cycle**
  - JIRA filters provide dynamic list of blockers
  - Open access to issues (JIRA); patches (git/gerrit) and test results (maloo)
  - Biweekly email update to wc-discuss and CDWG mailing lists
  - Quarterly reports posted on OpenSFS CDWG wiki
Lustre 2.2 Changelog

version 2.2.0

Support for networks:

- oiblind - OFED 1.5.4

Server support for kernels:

- 2.6.32-220.4.2.e16 (RHEL6)

Client support for unpatched kernels:

- 2.6.18-274.16.1.e16 (RHEL5)
- 2.6.32-220.4.2.e16 (RHEL6)
- 2.6.32-35-6.5 (SLES11)

Recommended e2fsprogs version:

- 1.41.90.wc4

Known Issues in 2.2.0:

- LU-1185: Parallels can occur running iscsi test while running with kernel debug options turned on
- LU-1131: Due to a known NFS-related bug in the kernel - https://bugzilla.kernel.org/show_bug.cgi?id=22572 - users wanting to re-export Lustre via NFS should apply the kernel patch for this issue.
Lustre 2.2 – Accomplishments

- Amount of test automation greatly increased
  - Reduces the overhead of making releases and increases the likelihood of catching regression earlier in the release cycle
- Test reports facility added
  - https://maloo.whamcloud.com/reports
  - Highlights areas where tests need improving
- Testing resources at IU and FZJ utilized during test cycle
- System wide Hyperion testing conducted
  - 1100+ clients
- Community Development wiki established
  - http://wiki.whamcloud.com/display/PUB/Lustre+Community+Development+in+Progress
  - Aim is to foster greater collaboration and prevent duplication of effort with development across organizations
Lustre 2.2 – Features

- **Asynchronous Glimpse Lock/Statahead** (LU925/LU389)
  - Improved performance for `ls -l`/find and accessing object attributes (file sizes/xtime etc)
  - Development funded by ORNL

- **Client Parallel Checksums** (LU884)
  - Improved support for `mmap` and better performance using checksums
  - Development funded by ORNL

- **Imperative Recovery** (LU580)
  - Faster recovery
  - Development funded by ORNL

- **Large Xattrs** (aka Wide Striping) (LU80)
  - Maximum stripe size raised from 160 to 2000; max file size increased from 320 TB to 64PB
  - Completing work by Sun funded by ORNL
  - ORNL/Xyratex helped with this initiative

- **Mds-survey** (LU593/LU633)
  - Tool for MDS performance benchmarking
  - Development funded by LLNL

- **Parallel Directory Operations** (LU50)
  - Improved performance when multiple processes access the same directory in parallel
  - Development funded by OpenSFS
Lustre Community Work

Lustre 2.2 Landings by Company

- Whamcloud: 258
- Xyratex: 19
- CEA: 3
- Cray: 5
- DDN: 3
- LLNL: 10
- NICS: 1
- ORNL: 15
- TACC: 3

Lustre Landings by Contract

- OpenSFS: 6
- Whamcloud: 50
- LLNL: 4
- ORNL: 9

OpenSFS and Whamcloud: 92
Lustre 2.3 and Beyond

• Lustre 2.3 (September 2012)
  – Server Stack SMP Scaling (LU-56)
  – LFSCK Online OSD check/ OI scrub (Ph.1, LU-957)
  – OSD restructuring (ZFS OST capability, LU-1305)

• Lustre 2.4 (March 2013)
  – OSD Restructuring (ZFS on MDTs, LU-1305)
  – Distributed Namespace (Ph.1 Remote Directories, LU-1187)
  – HSM (CEA implementation, LU-941, 169, 827, 1338, 1333)
  – LFSCK Online check/scrub – Distributed Repair (Ph.2, LU-957)

• Lustre 2.5 (September 2013)
  – LFSCK MDT-MDT Consistency (Ph. 3, LU-957)
  – Distributed Namespace (Ph.2 Distributed Directories, LU-1187)
Other Areas of Interest

- **Networking**
  - LNET Dynamic Configurations,
  - Channel Bonding,
  - Health Networks,
  - IPv6

- **Storage Management**
  - Tiered Storage: Policy-driven object storage placement
  - Migration: OST rebalancing, Async mirroring
  - Small File Performance: Unified Targets

- **Other**
  - Administrative Shutdown
  - Test frameworks
  - JobStats
  - Hadoop Integration
Community Lustre Roadmap

- Maintenance Releases on an ad-hoc basis
  - 1.8.7-wc1
  - 1.8.8-wc1

- Maintenance Releases every quarter
  - 2.1.1
  - 2.1.2

- Feature Releases
  - 2.2
  - 2.3
  - 2.4
  - 2.5

- Feature Releases:
  - Imperative Recovery
  - Dirop SMP Scaling
  - Wide Striping
  - Statahead Speedup
  - Server Stack SMP Scaling
  - Online check/scrub
  - OSD restructuring
  - DNE Phase 1
  - HSM
  - LFSCK MDT-OST Consistency
  - Network Request Scheduler
  - DNE Phase 2
  - LFSCK MDT-MDT Consistency

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

Sponsor for Whamcloud Development and Releases:
- ORNL
- OpenSFS
- LLNL
- Whamcloud

Third Party Development:
- CEA
- Xyratex
Lustre Feature Lookup

- Go to:  
  http://wiki.whamcloud.com/display/PUB/Lustre+Community+Development+in+Progress

- Find your feature of interest
- Find the feature on the roadmap (schedule)
- Look at the bug in Jira (details on feature)
The Tools We Use Today

- **JIRA** is Whamcloud’s Issue and Agile management tool
- **JENKINS** is the build tool that continuously builds mainstream branches and all patches submitted by the community
- **GIT** is source code tool used for managing the Lustre canonical tree
- **GERRIT** is code review tool that allows the whole community to be part of the code review process

Jira, Jenkins
Git and Gerrit

Tools Live Today

jira.whamcloud.com
build.whamcloud.com
review.whamcloud.com
Work Flow

JIRA Change Request → Developer Carries Out The Work → Developer Tests With PAT - Results To Maloo

Developer Lands Using Gerrit → Jenkins Builds Patched Lustre → PAT Tests The Jenkins Build

Tests Results Go To Maloo → Peer Review Using Gerrit → Patch landed

http://wiki.whamcloud.com/display/PUB/Submitting+Changes
Gerrit

- When one developer writes code, another developer is asked to review that code
- A careful line-by-line critique
- Happens in a non-threatening context
- Goal is cooperation, not fault-finding
- An integral part of the Lustre coding process
Maloo

- Maloo is the authoritative test results database
  - Autotest and Developer results are stored in Maloo

- Testing results from development
  - Results from development provide landing collateral
  - Failures are as important as passes
    - Good to see the transition from failure to pass

- Landing requires passing results in Maloo
  - Maloo / Jenkins / Gerrit work in unison to ensure Reviews, Build and Test have all occurred.
## Test sessions

Sessions for user:  All users  

<table>
<thead>
<tr>
<th>Host</th>
<th>Group</th>
<th>User</th>
<th>Run at</th>
<th>Imported at</th>
<th>Sets passed</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>client-20-ib</td>
<td>review</td>
<td>Whamcloud Autotest</td>
<td>2011-05-20 00:43:45 UTC</td>
<td>2011-05-20 02:40:59 UTC</td>
<td>1/2</td>
<td>gerrit:4f1aa57ed</td>
</tr>
</tbody>
</table>

This is a link to the test suite detail.
Looking Forward

- Chroma™
- Exascale
Chroma Terminology
http://insidehpc.com/category/events/lug-2012/

• Chroma Manager
  – The User Interface for FS Management

• Chroma Storage
  – Software appliance (includes Lustre)
  – Installed on bare-metal servers
  – Attached to Chroma-supported storage

• Chroma Enterprise
  – Complete Scalable Storage Solution
  – Consists of Chroma Manager and Chroma Storage
Drive Chroma the Way You Want to

• Management console
  – Simplify provisioning, monitoring and maintenance of Lustre storage
  – Monitor and analyze filesystem performance
  – Deep integration with storage vendor partner products
    • Appliance toolkit

• Scriptable user interface
  – Command line
  – Web service

• REST API for Lustre management
  – Architecture for using standard web protocols
    • Separate front-end and server logic
    • Used internally by GUI and CLI
  – Enables deep integration with 3rd party site/cluster management products
Chroma Manager
Chroma Enterprise 1.0
Available from DDN Now

- Chroma Manager
  - Provides graphical and command-line interfaces for filesystem management
  - Central repository of filesystem configuration, state, and measurements

- Chroma Storage
  - Software appliance (includes Lustre)
  - Installed on bare-metal servers
  - Attached to Chroma-supported storage

- Other partners working on Chroma
- Chroma is targeted for partners
Exascale Challenges
http://insidehpc.com/category/events/lug-2012/

Application data + metadata

• Explosive growth
  - Large, sophisticated models
  - Uncertainty Qualification
  - Billions – trillions of “Leaf” data objects
  - Complex analysis

• Filesystem namespace pollution
  - Keep filesystem namespace for storage management / administration
  - Separate namespace for application data + metadata
    - Distributed Application Object Storage (DAOS) containers

• Preserve model integrity in the face of all possible failures
  - Very large atomic, durable transactions
  - Integrity APIs at all levels of the I/O stack

• Search / query / analysis
  - Non-resident index maintenance & traversal / non-sequential data traversal
  - Move query processing to global storage
    - Same programming model as apps?