Cray’s Lustre Model and Roadmap

Cory Spitz*, Cray Inc. and* **Derek Robb***, Cray Inc.*

ABSTRACT: *Since 2003, Cray, our customers, and the wider HPC community have developed the Lustre file system as a key technology component for our success. In order to ensure that Lustre will continue to grow and develop Cray has played a founding role, with other leaders in the HPC community, in launching OpenSFS and have joined EOFS. Cray plans to incorporate new Lustre features, produced through the efforts of these consortia and their member companies, into its products. This paper will lay out the support model and new software release details for Cray’s use of Lustre in CLE and esFS in 2011, 2012, and beyond.*

KEYWORDS: Lustre, OpenSFS, esFS, CLE, roadmap, file systems

1. Introduction

The Lustre file system is a key component of Cray systems. Cray has historically provided value to its Lustre users by performing development, integration, performance scaling improvements, stringent testing, and support. To accommodate all of this work, the release of Lustre software as a part of larger, integrated Cray release always lags the general availability of Lustre software.

This paper discusses Cray’s supported Lustre offerings and our roadmap for the future, which continues the model outlined above. However, our roadmap isn’t a natural extension of our current offerings. A trend in the industry is to remove “islands” of data and move file system servers out of the mainframe where they are accessible by multiple systems simultaneously. Certain conditions in the Lustre ecosystem are forcing Cray to continue to move into that direction. I’ll detail the changes. In addition, Lustre is moving to a very open community development model with OpenSFS and I’ll discuss Cray’s role thereof.

This paper assumes that the reader is familiar with Cray’s existing software release models, specifically for CLE. In short, Cray uses a “train” model with quarterly updates. The updates can contain new features, but the base kernel version (not necessarily service pack level) remains constant. The final update to a release may only contain critical and urgent bug fixes.

2. Current Lustre Offerings

Cray currently supports two major Lustre releases in two separate CLE releases. CLE 2.2 includes Lustre version 1.6.5. CLE 3.1 supports Lustre version 1.8.x.

CLE 2.2

CLE 2.2 supports Cray XT3, XT4, and XT5 platforms only. Its kernel is based on SLES 10 and its Lustre is based on version 1.6.5, although Cray’s Lustre incorporates hundreds of patches many of which are back-ported from later Lustre releases.

The latest CLE update is CLE 2.2 UP03 and there are no more updates planned for CLE 2.2. That is, CLE 2.2 is in maintenance mode and UP04 is not planned. We will only support CLE 2.2 with patches and will do so for as long as we have customers in the field with support contracts.

CLE 2.2 includes Cray Lustre failover capability and scales well. And because it has been deployed by our customers and supported for years it has matured and is stable.

Although systems running this release are actively supported, there are a number of compelling reasons to upgrade to CLE 3.1.

CLE 3.1

CLE 3.1 is the next main release and this supports both XT (XT6-SeaStar) and XE (XE5, XE6, and XE6m-Gemini) customers. Its kernel is based on SLES 11[1] and its Lustre is based on version 1.8.x. Although Oracle supports 1.8.x on SLES 10, Cray does not and customers must upgrade to CLE 3.1 based on SLES 11 to deploy Lustre 1.8.x

The initial CLE 3.1 release included Lustre 1.8.1 and newer Lustre versions were included in UP02 and UP03. Lustre 1.8.2 was integrated into UP02 and version 1.8.4 into UP03.

The latest CLE update to CLE 3.1 is UP03, but it is not yet available to earlier XT customers. Support for XT4 and XT5 will be announced soon. CLE 3.1 UP04, the final update for CLE 3.1, will be created after the earlier XT systems are supported and upgraded, so that new issues found there can be addressed and included in the last update. The ship date will depend on general availability of UP03, but is tentatively set for early July 2011.

The general availability of CLE 3.1 will give our earlier XT-SeaStar customers using Lustre 1.6.5 an upgrade path to Lustre 1.8 for the first time. Lustre 1.8 includes a lot of nice features (1) including Adaptive Timeouts, OSS Read cache, OST pools, Version Based Recovery, and a Cray exclusive, Imperative Recovery (2).

CLE 3.1 will be the last CLE release to support SeaStar customers. That means that Lustre 1.8.4 is last supported version for SeaStar systems.

esFS

Cray Custom Engineering Data Management Practice (CE DMP) has deployed Lustre external service file systems (esFS) and those systems are accessed from XT and XE systems running CLE 2.2 and CLE 3.1. The installations are custom and run various versions of Lustre, but all of the external Lustre servers run some version of 1.8, most frequently with CentOS. The mainframe access to esFS is made by using LNET routers that bridge the mainframe's HSN to the external fabric. The routers exist on the mainframe service nodes so that they have an HCA to make the connection. In most cases, the external fabric is Infiniband.

Support for these systems will continue via support contracts with CE DMP. From a maintainability perspective, it would be best to upgrade XT systems to CLE 3.1 UP03 in order to receive the best support.

3. Future Lustre Offerings

Cray is developing a Lustre roadmap for 1.8.x and 2.x. Lustre 2.x represents a disruptive technology insertion and so we will discuss the two roadmaps separately.

Lustre 1.8.x

First, it will be helpful to discuss the upstream development of Lustre 1.8. At this time, Lustre 1.8.x is in maintenance mode and no new features (aside from SLES 11 SP1 support discussed below) will be developed.

Oracle is continuing to maintain so-called b1_8, the head of 1.8.x development, and create quarterly maintenance releases from it. There is no long term roadmap from Oracle, but they have stated that fixes and quarterly releases will continue for as long as there is demand. Cray's goal is to release the latest generally available Lustre version in each CLE update, aside from any UP04, which is reserved for urgent and critical fixes only and never contains "new features".

CLE 4.0

CLE 4.0 (code name Ganges) will be released shortly. General availability is planned for June 2011. CLE 4.0 is based on SLES 11 SP1 and Lustre 1.8.4. Although, Oracle did not include SLES 11 SP1 support in Lustre until 1.8.5, Cray ported the kernel support backwards to our 1.8.4. Our development schedule could not accommodate waiting for version 1.8.5.

The next quarterly release of CLE, CLE 4.0 UP01, which is scheduled to ship in September, will include support for Lustre version 1.8.6. Lustre 1.8.6 general availability was announced by Oracle in May 2011.

Beyond UP01, Cray has no firm commitments to integrate future Lustre 1.8 releases since there is no 1.8 roadmap from Oracle. However, we expect that 1.8.7 will align with our schedules for CLE 4.0 UP02. Similarly, the expectation is that we would include 1.8.8 into UP03.

b1_8 will eventually reach an end with CLE. The next major CLE release after 4.0 (code named Nile) will be the last CLE release to support 1.8.x. Nile will also include Lustre 2.1. However, support for 1.8 will not end anytime soon as Nile is only scheduled to ship in June 2012. As usual, at least three SW updates will be included and those will continue to include quarterly Lustre releases of b1_8 from Oracle. Therefore, the Nile 1.8.x roadmap will extend out to March 2013 for the final release of CLE with support of 1.8. That would mean that 1.8.x support would exist through 2014.

esFS

Historically, CE DMP provided Lustre 1.8 solutions that were far ahead of the current 1.6.5 offerings from R&D. However, moving forward it is Cray's intentions to standardize the esFS offerings and base them on the tested CLE stack. Since the esFS installations are asynchronous from CLE, this can occur over time and as always, at the customer's pace.

[1] SLES 11 is often referred to as SP0, or service pack 0.

Lustre 2.x.

Nile will be the first CLE release to support Lustre 2.x with Lustre 2.1. Lustre 2.1 doesn't bring many new additional features, but major components were rewritten as a basis for future technology insertions, such as Clustered MetaData (CMD) (3). However, it severely breaks the traditional Cray Lustre model. That is because SLES server support has been deprecated in Lustre 2.x. However, SLES clients are still supported. In addition, Lustre 2.1 clients are not backwards compatible with Lustre 1.8 servers. This means that it will not be possible to use Lustre 2.1 at all with traditional direct-attached Lustre!

Therefore, CLE will only include support for Lustre 2.1 clients (and LNET routers). External file systems will be the only deployable Lustre 2 solution. This is another reason why 1.8 will have a long life and continue into Nile. This will give our XE customers with Danube and Ganges with direct-attached file systems opportunity to upgrade their CLE systems without requiring them to install a new external file system.

Nile GA with Lustre 2.1 is scheduled for June 2012 and that is well behind the expected "summer" 2011 general availability date for the "community" Lustre 2.1 release. I'll discuss the community release and Cray's role with OpenSFS in the next section.

Even though Lustre 2.1 clients are not backward compatible with Lustre 1.8.x servers, the opposite is possible. That is, Lustre 2.1 servers are backwards compatible with Lustre 1.8.x clients. This means that it will be possible to deploy external Lustre 2.1 based file systems and mount them with 1.8.x clients from a Cray mainframe. CE DMP has not made a commitment for deploying Lustre 2.1 versioned esFS servers since they have no roadmap and do custom installations. Tentative plans call for 2.1 deployments that align with the CLE schedule for Nile (June 2012). In this timeframe, CE DMP may also deploy third-party Lustre appliances. But again, there are no firm commitments.

CE DMP is currently testing early versions of Lustre 2.1 on esFS servers that would interoperate with Lustre 1.8.x clients. In addition, ORNL, as a part of OpenSFS is executing an interoperability test plan that tests Lustre 2.1 servers with Cray CLE clients. (4) Therefore, it is expected that these systems could be easily supported because ORNL is bringing considerable test resources to bear.

CE DMP has no migration plans developed for migrating existing direct attached systems to esFS or Lustre appliances, but this would be possible. Lustre 2.1 servers can understand the 1.8.x on-disk format. It would even be possible to then downgrade servers and serve the same file system with 1.8.x servers. Even if the file system was a new 2.1 installation, it can be created in a way that can be safely downgraded as to be backward compatible with version 1.8.

CLE releases beyond Nile are not yet planned, but will include further updates to the Lustre 2 base from OpenSFS. Cray plans to carry it's traditional model forward where we integrate an upstream release and feature set into CLE, it is just the case that our upstream provider will now be OpenSFS. CLE will still be based on main releases with quarterly updates.

4. OpenSFS

Cray Inc., Data Direct Networks (DDN) Inc., Lawrence Livermore National Laboratory (LLNL) and Oak Ridge National Laboratory (ORNL) co-founded OpenSFS in-part because of concerns for the future development of Lustre for Linux. Cray was also a participant in HPCFS for the same reason.

OpenSFS's original goals were to provide an OpenSFS branded release of Lustre that was not a fork of the canonical source base maintained by Oracle. The release would contain new features that were funded with member dues and ongoing bug fixes for maintenance. All of these changes would be pushed "upstream" to the canonical source base maintained by Oracle. However, recently it has become clear that Oracle will no longer maintain the 2.x version sources of Linux.

With Oracle's absence, it was unclear how the community would then move forward. Amid the confusion, Whamcloud Inc. announced that they would lead the effort to finish the next planned Lustre release, version 2.1. It was hoped that in the meantime that the community could reach consensus about how to move forward.

Fortunately, there has been much progress on reaching Lustre community consensus in recent weeks. At the Lustre User's Group meeting in April HPCFS and OpenSFS announced that HPCFS would merge into OpenSFS. The legal details are still being worked out, but it is clear that the Lustre community will not fracture. The European Open File System (EOFS) group will remain separate, mostly for legal and funding reasons, but there will be a Memorandum of Understanding between EOFS and OpenSFS and they will work closely together to ensure that efforts are not duplicated and that new features are not in conflict.

Also at LUG, the two main Lustre development houses, Whamcloud and Xyratex, announced that they would bring their lead developers together for a summit to chart out the technology needed for Lustre in the future. Whamcloud published a roadmap (5) and Xyratex published a white paper (6) discussing their position. The two directions are quite complimentary, which bodes well for being able to maintain a single canonical version.

OpenSFS Working Groups

Cray is taking a leadership role with OpenSFS. As part of co-founding OpenSFS at the "promoter" level, Cray pays $500K in annual dues and has a seat on the board. In addition we are involved in the working groups.

When OpenSFS started, it commissioned working groups to define and drive operations. The Technical Working Group (TWG) was to engage the community to define requirements for functionality and performance and then craft RFPs to commission work to address the most popular or needed enhancements based on the collected requirements. The TWG has a whitepaper discussing their role at http://ww.opensfs.org. (5) John Carrier from Cray co-chairs the TWG along with David Dillow from ORNL. The author is a contributor to that group as well.

A Release Planning Working Group (RPWG) and a Support Working Group (SWG) attempted to define how the OpenSFS SW model would work, but after the disruption caused by Oracle's abandonment of the 2.x community, they agreed to hibernate to focus entirely on working with Whamcloud to get version 2.1 tested and released. The author is also a member of the joint RPWG+SWG.

OpenSFS is truly open. There are a number of open email lists that support the working groups and general discussion. In addition, meeting minutes are recorded and posted to these lists. A Communications Working Group was created to aid these efforts.

There is also a working group focused on Lustre for wide area networks.

Cray is not directly involved with either the WAN working group or the communications working group.

Cray's board member is David Wallace, who is Software Product Manager for the Cray Product Division. Derek Robb was Cray's envoy to HPCFS and now will participate with OpenSFS.

Cray Involvement in Working Groups and Updates

After gathering requirements from the community, the TWG generated two RFPs and gained board approval to publish them. The TWG wanted one RFP to focus on performance improvements and the other that to address foundational enhancements. The first RFP primarily addresses metadata performance. The RFP specifies the community requirements that need to be met for metadata performance in 2012 and 2014, but does not dictate what solutions should be deployed. Respondents have a chance to propose whatever feature enhancements they deem necessary. The second RFP calls for further definition of quotas support for inclusion into the new OSD back-end API. Currently in 1.8.x, Lustre quotas support depends on the quotas implementation in ldiskfs. The new OSD layer could support other back-end file systems, but the quotas support will need to be extracted up a layer.

Both RFPs ask that respondents address not only the proposed solutions, but also how they would pay-off technical debt. Technical debt is the undesirable side effect that results when design decisions and implementation trade-offs create code that is difficult to maintain. Technical debt accrues to an unacceptable level over time if not addressed. The Lustre community feels that it is important to begin paying down Lustre technical debt in order to keep Lustre viable in the future.

There is a sub-team of five individuals that are reviewing the TWG RFP responses and both John Carrier from Cray and the author are on the team.

The TWG will ensure proper design and code quality. In the future the joint RPWG+SWG will ensure that the code is landed safely to an OpenSFS release. Cray fully intends to adopt OpenSFS sponsored enhancements and will volunteer test resources and contribute to test plans as necessary in addition to maintaining our roles in the working groups. Cray will encourage appliance vendors to do the same.

OpenSFS with guidance from senior Lustre engineers determined that it would take approximately 12 full-time FTEs to perform the necessary maintenance, testing, gate keeping required to productize a quality SW release and maintain it. OpenSFS has requested member groups to volunteer these resources. This work is currently being defined. Current plans call for OpenSFS to "fund the gaps" after the release of Lustre 2.1. It is likely that the Linux foundation model will be followed where contributors continue to work for their employers but that their work would be for the community and compensated or reimbursed by OpenSFS. Pam Hamilton from LLNL now chairs the joint RPWG+SWG and is working to renew the discussions about what would be the canonical release or canonical sources. This is perhaps the most important work that OpenSFS will do.

Conclusion

Lustre file systems are a very important resource for Cray systems and Cray's customers. In order to provide the best Lustre experience Cray executes substantial testing and software stabilization as a part of distributing CLE. These efforts will continue for both 1.8.x and 2.x versions going forward.

1.8.x will continue to be included in quarterly updates through the Nile CLE release and through 2014. Lustre version 2.x will be new beginning with the Nile CLE release in June 2012, but will only be deployed with external file systems due to the lack of SLES server support, which is required for Cray service nodes. Therefore, CE DMP will deploy traditional esFS systems and upcoming Lustre appliances, but there are no firm delivery dates at this time.

Cray is very involved with OpenSFS to ensure its success, the success of Lustre 2.x, and future Lustre feature development.

In short, Cray has a well defined plan for supporting Lustre 1.8 versions through 2014. Cray is developing a plan for upcoming Lustre 2.x releases on external servers. Cray believes that the Lustre Community has become organized, and is funded, to allow for a new, more effective level of development and support for Lustre that will allow it to remain as the premier scalable, parallel file system for HPC.

Acknowledgements

We would like to thank OpenSFS founders for having the vision to create a truly open development ecosystem for Lustre. The success of OpenSFS will have a direct impact upon Cray's future Lustre offerings and in turn the Cray system experience.

We would also like to thank John Carrier and Dave Wallace for helping to define Cray's Lustre requirement and thus a roadmap.

References

1. **Oracle Corporation.** Lustre 1.8. *http://www.lustre.org.* [Online] 2010. http://wiki.lustre.org/index.php/Lustre_1.8.
2. *Imperative Recovery for Lustre Failover.* **Cory Spitz, Chris Horn, Nicholas Henke.** Edinburgh : s.n., 2010. CUG 2010 Proceedings.
3. **Oracle Corporation.** Lustre 2.0 Features. *http://www.lustre.org.* [Online] 2009. http://wiki.lustre.org/index.php/Lustre_2.0_Features.
4. **Peter Jones, Sarp Oral, et. al.** Lustre 2.1 Community Release Test Plan. *http://groups.google.com/group/lustre-21.* [Online] 2011. http://groups.google.com/group/lustre-21/attach/e03c897c6a568282/Lustre+2.1+Community+Release+Test+Plan+-DRAFT+v3.doc?part=2&view=1.
5. **Whamcloud Inc.** Whamcloud Lustre Roadmap. *http://wiki.whamcloud.com.* [Online] 2011. http://wiki.whamcloud.com/display/PUB/Whamcloud+Lustre+Roadmap.
6. **Xyratex.** Xyratex Lustre Architecture Priorities Overview. *http://www.xyratex.com.* [Online] 2011. http://www.xyratex.com/pdfs/whitepapers/Xyratex_white_paper_Lustre_Architecture_Priorities_Overview_1-0.pdf.
7. **OpenSFS TWG.** OpenSFS Technical Working Group. *http://www.opensfs.org.* [Online] 2011. http://lists.opensfs.org/pipermail/twg-opensfs.org/attachments/20101202/46e7fccb/attachment-0001.pdf.

About the authors

Cory Spitz is the team lead for Lustre integration in Cray's OSIO division. He can be reached via email at spitzcor@cray.com. Derek Robb is the Storage and I/O Product Manager for CE DMP and can be reached at derekr@cray.com. Both authors work at 380 Jackson Street, St. Paul, MN, 55101.