lustre®

Lustre Center of Excellence Scalability Workshop

Oak Ridge National Laboratory, February 10 - February 11, 2009 Meeting Notes

John K. Dawson **T** «Phone» John.Dawson@Sun.com

Meeting Summary

The Lustre Center of Excellence (LCE) at Oak Ridge National Laboratory hosted a workshop on scalability of the Lustre file system on February 10 and 11 of 2009. The purpose of the workshop was to identify key requirements for supporting the next generation of large scale parallel systems with the Lustre file system. The meeting was focused on supporting systems requiring Terabytes per second of bandwidth with hundreds of Petabytes to be deployed in 2012 or after.

Invitees were asked to identify their top three requirements or issues before the meeting. After introductions, the meeting began with Galen Shipman presenting an overview of Oak Ridge's current compute and storage capability and high level description of their expected requirements for IO and storage in 2012 as well as a discussion of trends in performance and capacity of IO devices.

Next, Eric Barton presented the Lustre roadmap in detail. The attendees had a lot of discussion around the capabilities and interactions of planned features. As the discussion progressed, issues that arose in addition to those submitted by the attendees in advance were captured.

We then walked through all of the requirements that were submitted in advance and identified which issues or requirements were either not met at all by the current Lustre roadmap or would not be addressed in the required time frame.

Our discussion identified fifteen gaps in the Lustre roadmap. We discussed and voted on the fifteen items to identify and order the top 10 by priority. Lustre and LCE personnel did not vote on priorities. The top 10 gaps are listed and described at the end of these notes.

John K. Dawson **T** «Phone» John.Dawson@Sun.com

Note that many of the requirements or issues the attendees brought *were* covered by the existing Lustre roadmap. These no-gaps, so to speak, are not highlighted nor listed in this paper.

John K. Dawson **T** «Phone» John.Dawson@Sun.com

Follow up actions from the meeting will be:

1. A meeting summary (this document and attachments). - March, 1, 2009
2. A whitepaper on Lustre Scalability. - March 1, 2009
3. A response from Sun to all of the identified gaps. – May 2009
4. A second workshop to present the initial response to the gaps and to discuss IO performance and scalability requirements for 2015 and beyond. - May 19-20, 2009
5. A revised Lustre roadmap that incorporates the Sun responses. - July 31, 2009

John K. Dawson **T** «Phone» John.Dawson@Sun.com

Scalability Workshop Agenda

Date: February 10-11, 2009

Location: Oak Ridge National Laboratory

Day 1

Time	Topic	Presenter
9:00 AM	Introductions & Logistics	John Dawson
9:15 AM	Presentation of ORNL's 2012 requirements and draft configuration.	Galen Shipman
10:45 AM	Break	
11:00 AM	Presentation of existing Lustre plans and roadmap, focused on the capabilities planned for 2011-2012.	Eric Barton
12:30 PM	Lunch	Catered
1:30 PM	Gap analysis - Group discussion - Group walk through of requirements flagging any requirements that can not be met with Lustre today. Group walk through of requirements from other sites that are not captured in ORNL requirements and the exiting Lustre roadmap.	John Dawson
3:30 PM	Break	
3:45 PM	Group discussion to Identify any requirements either not covered by current Lustre roadmap or not planned to be ready when needed.	John Dawson
5:00 PM	Adjourn	
6:00	Dinner at Flatwater Grille	

John K. Dawson **T** «Phone» John.Dawson@Sun.com

Day 2

Time	Topic	Presenter
9:00 AM	Continue gap analysis	John Dawson
10:45 AM	Break	
11:00 AM	Continue gap analysis	Eric Barton
12:30 PM	Lunch	Catered
12:30 PM	Prioritize gaps - Identify and vote to rank order top ten gaps by priority.	John Dawson
13:45 PM	Wrapup	John Dawson
2:00 PM	Adjourn	

John K. Dawson **T** «Phone» John.Dawson@Sun.com

Workshop Attendees

John – are these people that actually attended, versus were invited but did not attend?

Name	Email address	Affiliation
Sadaf Alam	alamsr@ornl.gov	ORNL
Eric Barton	eeb@sun.com	Sun Microsystems
Michael Booth	michael.booth@sun.com	Sun Microsystems
Shane Canon	canon@nersc.gov	LBNL/NERSC
John Carrier	carrier@cray.com	Cray, Inc.
John Dawson	John.Dawson@Sun.Com	Sun/LCE
Andreas Dilger	adilger@sun.com	Sun Microsystems, Inc
David Dillow	dillowda@ornl.gov	ORNL
Oleg Drokin	oleg.drokin@sun.com	SUN
Bryan Embry	zerotrace000@gmail.com	DoD
Rob Farber	rob.farber@pnl.gov	PNNL
Evan Felix	e@pnl.gov	PNNL
Jeff Flick	Jeff.Flick@noaa.gov	Geophysical Fluid Dynamics Laboratory
Mark Gary	mgary@llnl.gov	LLNL
Sherry Hempfling	hgv@ornl.gov	ORNL
Nicholas Henke	nic@cray.com	Cray
Isaac Huang	he.huang@sun.com	Sun Microsystems
Frank Indiviglio	frank.indiviglio@noaa.gov	NOAA

Name	Email address	Affiliation
Heike Jagode	jagode@eecs.utk.edu	UTK / ORNL
Rainer Keller	keller@hlrs.de	ORNL
Anton Kozhevnikov	khv@ornl.gov	JICS
jacques-charles lafoucriere	jc.lafoucriere@cea.fr	CEA
Keith Miller	kmiller@datadirectnet.com	NCCS/DDN
Steve Monk	smonk@sandia.gov	Sandia National Laboratories
Colin Morgan	colin.morgan@noaa.gov	NOAA
Christopher Morrone	morrone2@llnl.gov	LLNL
Sarp Oral	oralhs@ornl.gov	ORNL NCCS
Robert Read	rread@sun.com	Sun Microsystems
james rosinski	rosinskijm@ornl.gov	ORNL
Nathaniel Rutman	nathan.rutman@sun.com	Sun Microsystems
Galen Shipman	gshipman@ornl.gov	ORNL
DI WANG	di.wang@sun.com	Sun Microsystems
Feiyi Wang	fwang2@ornl.gov	ORNL
Kit Westneat	kwestneat@datadirectnet.com	Data Direct Networks

lustre®

Top Ten Gaps

At the conclusion of the workshop we identified and prioritized the top ten requirements or issues that were seen as gaps in the Lustre roadmap. These ten gaps are:

1. Asymmetric impact of failures – availability and reliability
   Hardware or software failures on a subset of file system resources should only impact those resources residing on the failed/failing equipment. For instance, a failed OSS node should only preclude access to files controlled by that OSS – it shouldn't snowball into MDS hangs because of exhausted threads, router backups. A problem on one file system should never impact another file system in any way.
2. Known milestones for order of magnitude incremental MDS performance improvements on servers of sufficient configuration over next two years. CMD performance will then become a significant/gating need after two years.
   The performance of Metadata operations is a current and growing issue. While Clustered Metadata Servers (CMD) are going to address this, there is a need for improved MDS performance before CMD is available. The workshop participants would like to know what interim performance improvements to expect and when.

John K. Dawson **T** «Phone» John.Dawson@Sun.com

3. Lustre ZFS Licensing: - There was concern over the license used with ZFS (CDDL). While CDDL is an Open Source license people were concerned that it is not compatible with GPL and that this may impact ZFS in Lustre the following ways:
   - Community involvement - The Lustre community may be unwilling to invest in enhancing ZFS if it not licensed with GPL.
   - Protect investment - Concern that the CDDL does not sufficiently protect any investments made in ZFS. Sun could drop ZFS and the user investment in the code could be lost or encumbered.
   - Contaminating – Concern that programmers working on the ZFS code will be exposed to Sun owned Intellectual Property and then not be free to work on other file systems. (The SCO UNIX/LINUX legal issues came up here.)
   - Link level – compatibility of licenses – Will the kernel implementation of the ZFS DMU be compatible with both letter and spirit of the GPL and the Linux kernel's license checks?

4. Quality of service - Network Request Scheduler
   The ability to assign quality of service levels to individual clients is required to handle:

   - Individual jobs competing for bandwidth on a single system (direct attached)
     - Example: an aggressive reader to a single OST can slow a competing simulation job during a checkpoint.
   - Individual machines competing for bandwidth in a center wide configuration
     - Example: A visualization cluster may consume a disproportionate amount of bandwidth in a center-wide file system reducing the I/O performance of other simulation platforms.
   - Metadata processing rates to provide a more real-time responsive system.

- Example: Users on login nodes receive poor file system responsiveness when other metadata "hungry" applications are running

- Open issues:
  - How do we provide quality of service in routed configuration when router QOS would be based on available bandwidth while MDS and OSS QOS may be based on RPC rates?
  - What NRS policies will be supported?
  - How are timeouts affected as we may slow some reader/writers for aggregation purposes?
  - How do we assign QOS values?

5. Performance Variability
   Users should see consistent performance across a continuum of workloads assuming some the individual I/O operations are reasonably large. For example, users should achieve a large fraction of the ideal I/O rates for large block I/O regardless of alignment, exact size, and number of clients.

6. Policy engine to set allocation, migration, tier classes, locality to client OST performance stats: load avg, etc. Such a policy engine would use the following information to help implement such policies:
   OST fill information: percent full, file count, etc
   OST group/pool info
   Uid/gid info
   File size/type/mime-info

7. Manageable at scale: This is the ability to quickly and precisely identify failures and potential failures. Examples are:

   • Offering more information in proc

   • Improved syslog information

   • RAS interface that will give 3rd parties something to develop tools against?

   • Other mechanisms of aggregating this information in Lustre

   • Manage normal operations

8. Failover duration - Failover is not widely used today because of difficulty in configuring it and because the time required for the system to detect a failure and complete the failover can be longer than the time required to reboot the system. A mechanism is needed to significantly speed up the process. Eric discussed the work being done to look at adding a virtual health network to address this.

9. Small file performance and efficiency. Improve the efficiency and performance of small file management by aggregating files on OST's and placing small files on MDS.

10. Wide stripe performance

    Users would prefer to not have to worry about setting striping for their files. A system wide default should achieve a large fraction of the best case I/O without requiring the user to manually set the stripe count. One likely consequence of this is a small file (on the order of the stripe size) that is widely striped should perform as well as narrowly striped file.