Cray Lustre Model Roadmap

Cory Spitz and Derek Robb

Cray Inc.

5/24/2011

CRAY
THE SUPERCOMPUTER COMPANY

Introduction and Agenda

- Current Cray Lustre model
- Current Lustre offerings
- Cray and OpenSFS
- The Lustre 2.1 community
- Future Cray Lustre model
- Cray Lustre roadmap

Cray Lustre Model

- Lustre is incorporated into the Cray Linux Environment (CLE)
  - Complete SW stack is tested and stabilized as a whole system
  - CLE releases are close to a “train” model with quarterly updates
    - There are defining features that will hold the train
    - UP01, UP02, UP03 may contain new features
    - UP04 bug fixes only
  - CLE is the delivery vehicle for any new direct-attached Lustre version
- esFS by Cray CE Data Management Practice
  - Removes the “island” of data on the mainframe
  - LNET routers within the mainframe bridge networks
  - All esFS servers run some form of 1.8.x, CentOS preferred
  - Interoperates with all CLE clients
  - Deployments and updates are not tied to CLE schedules

Current Lustre Offerings – CLE 2.2

- Supports Cray XT3, XT4, and XT5
- SLES 10 based
- Lustre 1.6.5
  - With hundreds of patches(!)
- Lustre 1.6.5 features
  - Failover
  - LNET routers
  - Scales with portals
  - Stable
- UP03 is the latest release
- UP04 not planned
- There are compelling reasons to upgrade

Current Lustre Offerings – CLE 3.1

- Supports Cray XT6, XE5, XE6, and XE6m
- SLES 11 (service pack 0) based
- Lustre 1.8.x
- UP03 is the latest release
  - Includes Lustre 1.8.4
  - XT4 and XT5 support coming soon
- UP04 planned for July 2011
- Cray Lustre 1.8.x features
  - Failover
  - Scales with Portals/Gemini
  - Adaptive Timeouts
  - OSS Read Cache
  - OST Pools
  - Version Based Recovery
  - Imperative Recovery (Cray exclusive)

Cray and OpenSFS

- Cray has co-founded OpenSFS to:
  - Ensure the development of Lustre for Linux and HPC
  - Foster the Lustre community
  - Be successful by funding these efforts with capital
- Cray is taking a leadership role with OpenSFS
  - Cray is an OpenSFS “Promoter” with $500K annual dues
    - Cray has a seat on the board
      - David Wallace, Software Product Manager, CPD
  - Cray has an active role in OpenSFS working groups
    - John Carrier co-chairs the Technical Working Group (TWG)
    - Cory Spitz is a contributor as well
    - Both John and Cory are on the TWG RFP sub-team
    - Cory Spitz is a contributor to the joint Release Planning Working Group and Support Working Group (RPWG+SWG)

Lustre 2.1 Community Release

- Oracle has abdicated the support and development of Lustre 2.x
- Whamcloud has graciously offered to release Lustre 2.1
  - Release planning
  - Development
  - Gatekeeping
  - Release testing
- OpenSFS estimates 12 full-time FTEs for releases
  - OpenSFS would like members to volunteer FTEs
  - Plan is to fund the gaps

Future Cray Lustre Model

- Continue 1.8.x quarterly updates from Oracle
- Carry the current model forward for 2.x
  - Upstream provider becomes OpenSFS
- Cray leadership in OpenSFS guides 2.x development
- Cray will encourage appliance vendors to adopt the OpenSFS stack

Cray Lustre 1.8.x roadmap

- CLE 4.0 (code name Ganges)
  - GA planned for June 2011
  - SLES 11 SP1 based
  - Lustre 1.8.4
    - SLES 11 SP1 support backported from Lustre version 1.8.5
- CLE 4.0 UP01
  - Planned for September release
  - Lustre 1.8.6
- So-called “b1_8” in maintenance mode by Oracle
  - No stated 1.8.x roadmap from Oracle
  - Oracle has stated they will continue quarterly Lustre releases
  - Therefore future 1.8.7 and 1.8.8 are targets for 4.0 UP2 and UP03
- esFS moves to CLE SW stack
- CLE “Nile” will be the last CLE release with 1.8.x

Cray Lustre 2.1 roadmap

- Lustre 2.x lacks SLES server support
  - External file systems required for 2.x (!)
    - esFS or Lustre appliances only
  - Direct attached file systems must remain on 1.8.x
- CLE “Nile” first release with Lustre 2.1
  - GA June 2012
- Lustre 2.1 features
  - Rsync, commit on share, changelogs, rewrite for CMD and OSD
  - Servers compatible with 1.8.x clients
    - Clients not compatible with 1.8.x servers
  - Can upgrade and fallback

Lustre 2.0/2.1 Features

Feature	Benefit
Restructured Server and Client	Enforce strict layering with OS to establish stable foundation for portability (OSD) and performance optimizations (CMD)
Server Change Logs	Record changes to namespace, which create audit trail that can be used to interoperate with HSMs
Lustre_rsync	Replicate namespace and data to an external backup system (change logs avoid the need to scan the file system for inode changes and modification times)
Commit on Share	Remove dependent replays from clients by committing metadata updates immediately
Imperative Recovery (preview)	Reduce recovery delays with explicit client notification on server restart
Clustered MetaData (preview)	Provide client and server foundations for CMD, which will support file systems with multiple MDSs (preview has no recovery mechanisms)
Size on MDS (preview)	Cache object attributes from OSTs on MDS to speed up directory searches (ls –l)
Kerberos (preview)	Enable Lustre authentication in Kerberized environments

Cray Lustre Roadmap

Cray Lustre Roadmap based on CLE Releases

5/24/2011

Linux Version	CLE Version	Code Name	Cray Systems Supported	Direct Attached Client/Lnet	Direct Attached Lustre OSS/MDS	esFS/Appliance Client/Lnet	esFS/Appliance Lustre OSS/MDS	2011	2012	2013
SLES10	2.2.x	Congo	XT Only	1.6.5*	1.6.5*	1.6.5*	1.6.x, 1.8.x†			
SLES11, SP0	3.1 UP01 & UP02	Danube	XT and XE	1.8.2	1.8.2	1.8.2	1.8.4†			
SLES11, SP0	3.1 UP03 & UP04	Danube	XT and XE	1.8.4	1.8.4	1.8.4	1.8.4†			
SLES11, SP1	4.0 UP00	Ganges	XE Only	1.8.4	1.8.4	1.8.4	1.8.6†			
SLES11, SP1	4.0 UP01/02/03	Ganges	XE Only	1.8.6+	1.8.6+	1.8.6+	1.8.6+, 2.1.x†			
SLES11, SP1	TBD	Nile	XE Only	1.8.6+	1.8.6+	2.x	2.2†			
SLES11, SP1	TBD	Nile	Cascade Only	-	-	2.x	2.x†			

* Heavily patched

Means actively being worked and patched

Means critical problems patches ONLY

† RHEL based

Means fixed in next release - upgrade encouraged