Lustre User Group 2009

Lustre 2.0 Features

Andreas Dilger
Sun Microsystems

2.0 Features List

- Metadata ChangeLogs
- Replication
- Commit on Share
- Simplified Interoperability
- Clustered Metadata Preview
- Kerberos Preview

Lustre 2.0 Architecture

New platform for future development

- MDS rewrite: CMD, ZFS, portable
- Client IO rewrite: SMP, simpler, portable
- ChangeLogs: replication, backup, HSM

Target scale: HPCS 2012+

- 100PB+ data, 1T files
- 100k nodes, 1M processor cores
- End-to-end data integrity

MDS Rewrite

- New MDS stack independent of VFS
- FIDs abstracted from backing inodes
- Allows CMD functionality
- Portable to Solaris kernel
- Allows layering on top of ZFS
- Future client MD writeback cache

Client IO Rewrite

- Client IO stack separated from VFS
- Portable to Windows, Solaris, OS/X
- Designed to allow future features
  - Server Network Striping (SNS==RAID)
  - SMP scalability, multi-core optimization

Metadata ChangeLogs

- Provides stream of server changes
- Persistent and atomic
- Filtered in kernel
- Can register multiple consumers
- Single log for all consumers
- Configured on server (for security)
- Can be consumed on client

Replication

- First Metadata ChangeLog consumer
- Keep a remote filesystem in sync
  - Incremental Backup
  - Read-only mirror
- Target not required to be Lustre

Commit On Share

- Improved Recoverability
- Remove inter-node dependency
- With VBR no dependent failures
- Performance impact only as needed

Simplified Interoperability

- Important for live upgrade process
- Reduces interoperability matrix
  - Removes recovery from interoperation
- Shutdown notification
  - Server notifies clients of impending shutdown
  - Clients flush buffers and block ops
- No recovery during normal unmount

Clustered Metadata Preview

- Scale metadata performance like IO
- Complements single-MDS speedups
- Create, unlink, lookup, getattr scaling
- 2.0 preview missing some features
  - Full power-off recovery of cross-MDS ops
  - Some usability features (pools, striping)
  - May still optimize on-disk format
- Production debut with simple recovery
  - Ordered synchronous cross-MDS ops
  - Rename, hard-link, mkdir may be slower

Kerberos Security Preview

- Kerberos user authentication
- Authentication, encryption of RPCs
- Authentication, encryption of data

- Undergoing security review
- Capabilities still unfinished

Lustre 2.0 Alpha

- Early availability to 2.0 features
  - For test filesystems only
  - Has undergone reasonable testing
- Regular builds will be made available
  - Current build v1_9_167 available today
  - Next build scheduled for May 9

Lustre 2.x and 3.0

Focus: Performance & reliability for HPC

- Performance features
  - Ongoing IO and MD performance gains
  - Size on MDS, CMD produciton
  - OSS write cache
  - Network Request Scheduler (NRS)
- Reliability features
  - ZFS for checksums, redundancy, scale
  - Kerberos Security production
- Information Management
  - HSM support (HPSS, SAM)
  - ZFS for backups, snapshots

Thank you

THANK YOU

<adilger@sun.com>