From lab to enterprise - growing the Lustre* ecosystem

Malcolm Cowe, Eric Barton, Andreas Dilger

High Performance Data Division

Drivers for change

Lustre has always supported high performance computing

- Extreme performance at extreme scale

New challenges for Lustre as HPC expands into new IT domains and markets

- Performance requirements are changing
  - Not just about massive streaming IO performance and huge files
  - Small random IO to large files, massive collections of tiny files
  - Diverse and unstructured
- Reliability, Availability, and Serviceability (RAS)
  - Resilience, service level agreements (many 9’s uptime)
  - Disaster recovery across sites
- Security of data in flight and at rest

Requirements of key market segments

Life sciences

- Small file workloads – very large file populations, millions of files
- Security and privacy – personal data, protected health information

Weather and climate

- Reliability – mission-critical workloads for forecasts and emergency modelling
- Small files – mixed workloads, but small file workloads are prevalent

Media, Manufacturing and EDA

- Small files, Reliability

Financial services

- Small files, Reliability, Security

Scaling metadata performance

Increasing single client metadata performance

- Lustre currently limits each client to 1 in-flight metadata modifying RPC
  - Single last_rcvd slot on MDT for each client to reconstruct RPC reply
- Change to dynamic log removes in-flight limit
  - Improved client multi-threading

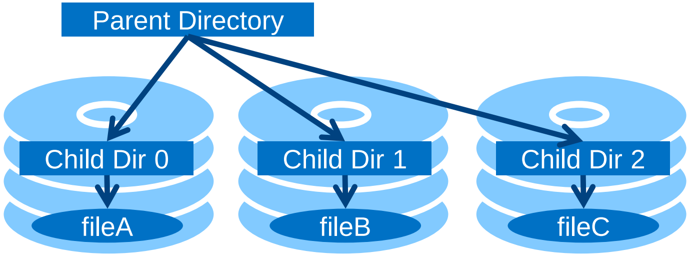
Scaling metadata performance

Horizontally scaling metadata performance

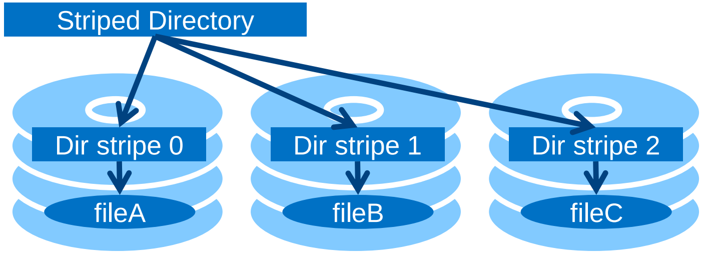
- Phase 1: Remote directories distribute a directory tree onto a separate MDT
- Phase 2: Striped directories distribute a single directory across multiple MDTs

Efficient general purpose distributed transaction protocol

- Remove disk sync latency from critical RPC path
- Assured recovery on client and/or server failure



DNE Phase 1

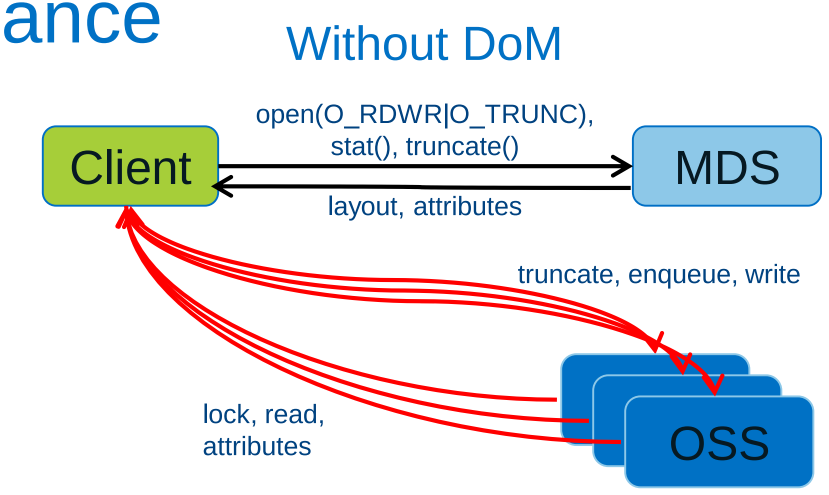


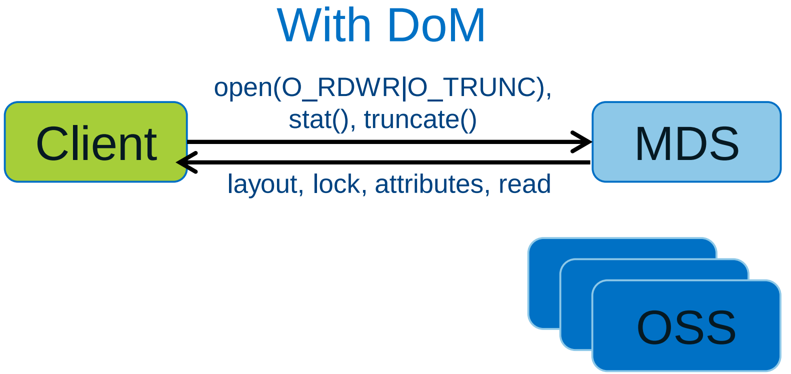
DNE Phase 2

Scaling small file performance

Data on MDT

- Co-locate data and metadata for small files
- Large streaming IO on OSTs not disturbed
- Further optimize IO rates with flash storage
- Scale out performance with striped directories





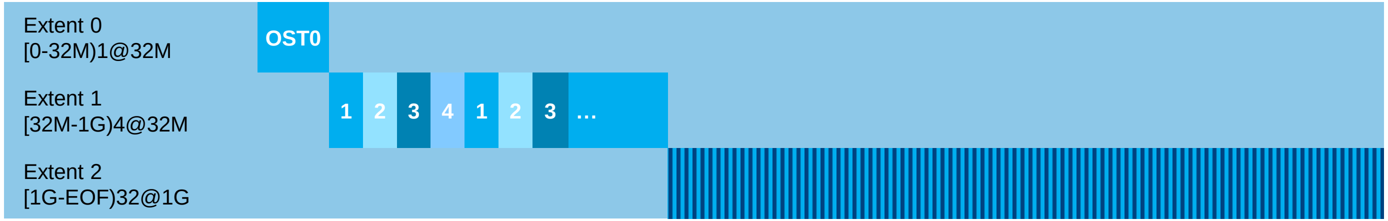
Layout enhancement

Allow file layouts beyond simple striping

- Different layouts for different ranges of each file
- Layouts can overlap (mirror) and be on different types of storage

Progressive File Layout

- Increase stripe count as file size increases
- Automatic layout for optimal performance of small and large files
- Layout extents can be disjoint or overlapping
  - RAID-1 mirroring → overlapping [0, EOF), [0, EOF)



Fault tolerance

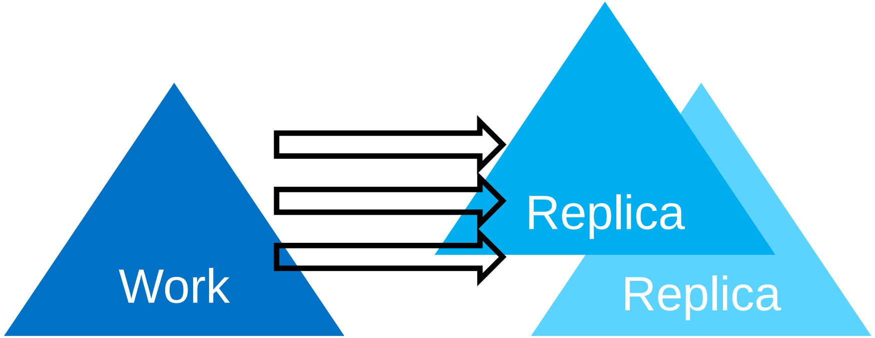
Replication within the filesystem

4 stripes 3 mirrors	0	1	2	3	0	1	2	…
	0’	1’	2’	3’	0’	1’	2’	…
	0’’	1’’	2’’	3’’	0’’	1’’	2’’	…

- Improve reliability of commodity storage hardware
- Increased data availability
  - No need to wait for failover
- Delayed or immediate mirroring of writes to replicas (overhead vs. availability)
- Improved read performance from multiple replicas

Replication to external storage

- Off-site disaster recovery
- Multi-version backups
- Requires…
  - Incremental update
  - Safe, reliable, efficient data migration



Snapshot

Data protection mechanism for checkpointing a file system

Several purposes

- Quick undo / undelete / roll-back in case of user/administrator error
- Prepare a consistent, read-only view of data for backup
- Prepare for software upgrade

ZFS* Snapshot

- Leverage the native snapshot in ZFS
- Create a coordinated snapshot across all storage targets

Security – market drivers

Demand for control of restricted information

- Life sciences, including health care (HIPAA regulation)
- Government, e.g. defense (ICD 503 directive)
- Aerospace, shipbuilding

Increased regulation of personally identifiable information

Movement of workloads to cloud – access must be constrained, data secured

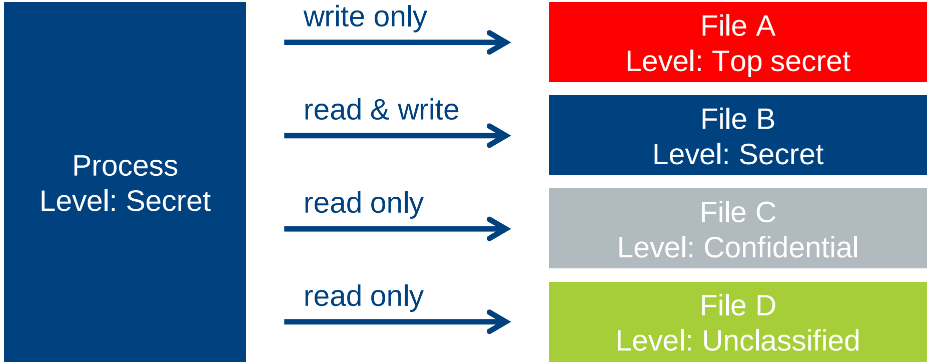
Financial impact of data theft is significant

- Healthcare average cost per breach $3.5M in 2013, some cases significantly larger
- Loss of credibility, loss of revenue as people move to other providers

Access control

SELinux provides fine-grained, mandatory and role-based access control

- MAC – administrative control of policy definitions
  - Mandatory means enforcement by the OS – users cannot bypass
- RBAC – access controls are assigned to roles, not users
  - Users are then assigned to one or more roles
- MLS – multi-level security:



Encryption

Encryption of data in flight

- Native implementation in Lustre
  - IU Shared-Key Crypto
  - Kerberos

Encryption of data at rest

- Block device encryption with DM-Crypt / LUKS – no change to Lustre required
- Potential for client-side encryption / decryption integrated into Lustre client

Summary

The Lustre community must continue to drive innovation in HPC storage

Increase Lustre’s versatility for an ever-widening spectrum of applications

- Deliver performance across a wide range of workloads

Enterprise data management

- Fault tolerance for critical production data
- HSM
- Replication for disaster recovery
- Snapshot

Security and encryption for sensitive data

Legal Information

- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.
- Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.
- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.
- For more complete information about performance and benchmark results, visithttp://www.intel.com/performance.
- Intel and the Intel logo, are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others

© 2015 Intel Corporation.

intel®

Software