Look Inside™

# From lab to enterprise - growing the Lustre* ecosystem

**Eric Barton**

**High Performance Data Division**

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

# Legal Disclaimers

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.

FTC Optimization Notice: Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

# 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

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

# Performance – Market Drivers

Increasingly diverse data workloads requiring large scale storage systems

- [Very] Large files
- Millions of small files per directory
- Millions of files in complex directory hierarchies, mixture of sizes
- Sequential, streaming IO
- Random IO

Contained in a single addressable name space

Requires a versatile, scalable file system platform

# 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

# Increasing versatility

Flexible layouts to accommodate diverse requirements in a single name space

- Decisions can be made per file, per directory, per filesystem
- Data on MDT for small file optimisation
- Replication for fault tolerance
- Progressive file layout
- As always, striping for throughput

HSM for long-term archival of permanent production data

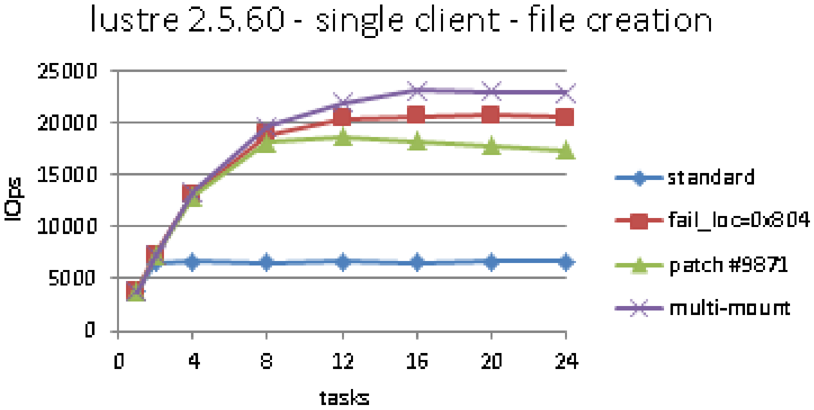
High performance parallel data movers for replication, disaster recovery

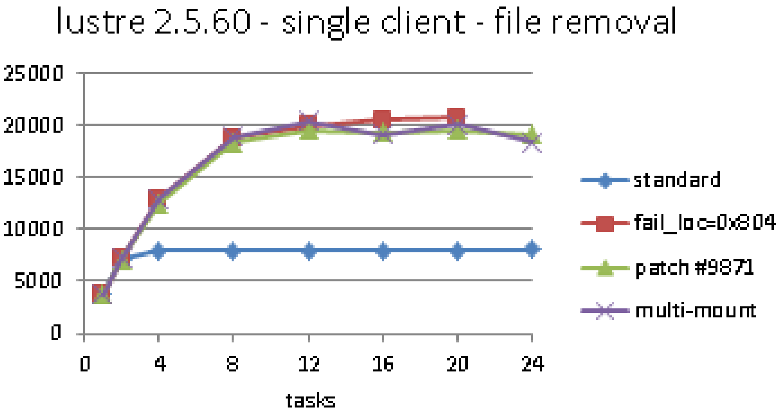
Securing data: access control and encryption

# 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





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

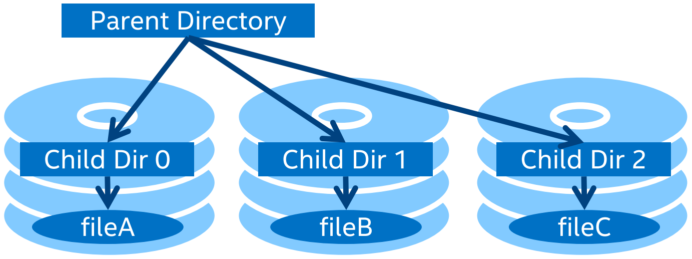
# Scaling Metadata Performance

## Horizontally scaling metadata performance

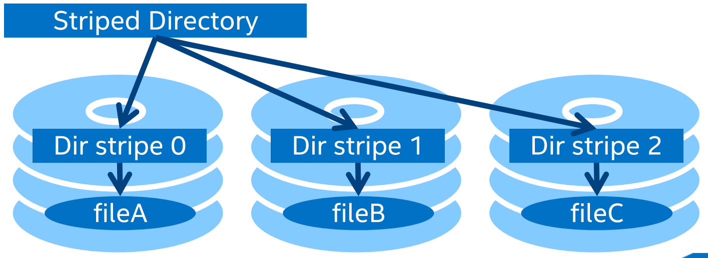
- Phase 1: Remote directories distribute a directory tree onto a separate MDT
- Phase 2: Striped directories distribute a single directory across multiple MDTs

## 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

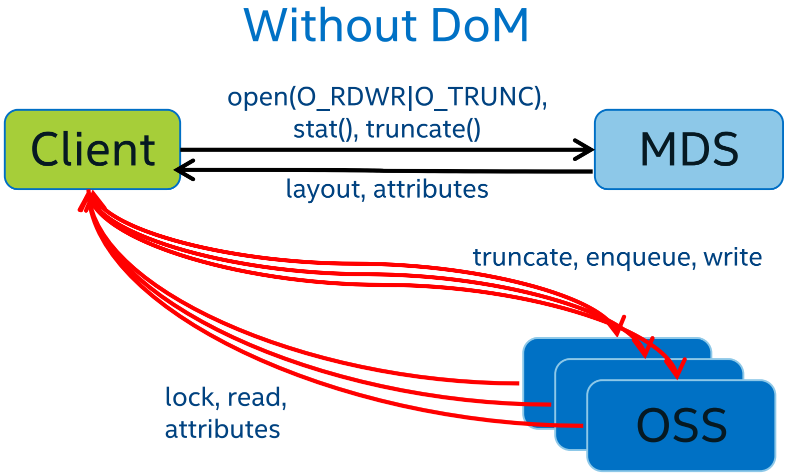
# Scaling Small File Performance

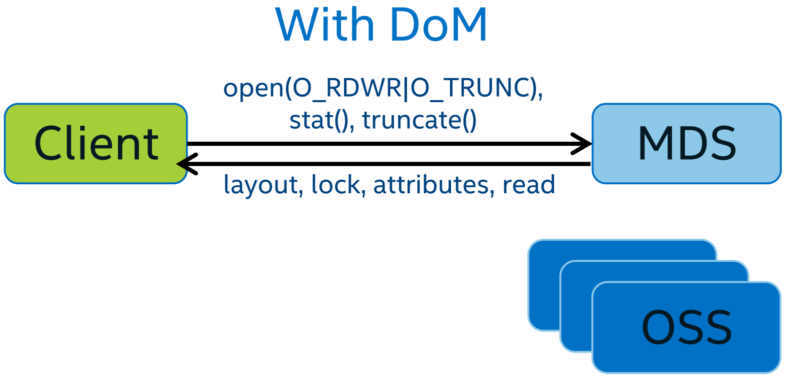
## Data on MDT

- 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

## Differentiated Storage Services (DSS)

- All stack levels classify I/O
  - OSD: ext4 extent metadata
  - OST/MDT: object index
  - Application: Frequently accessed directory/file
- Classifications drive caching policies
  - SSD tier integrated into OSD and/or block storage
  - Intelligently prioritize cache utilization





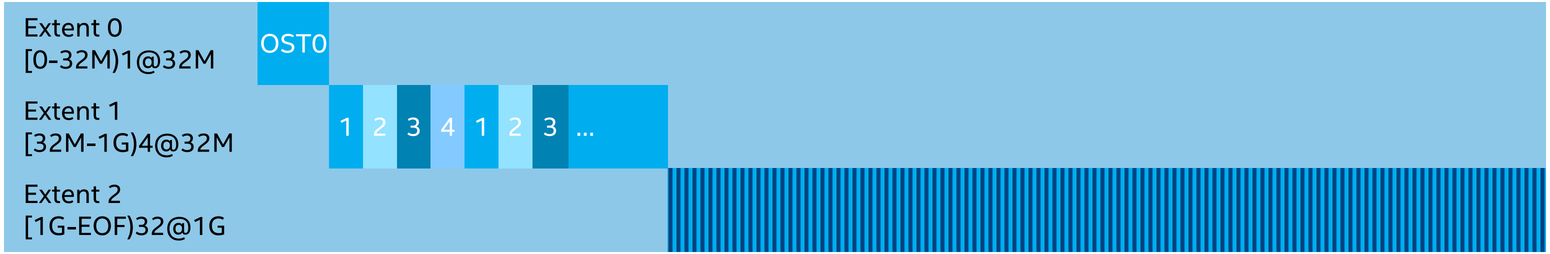
# Layout Enhancement

## 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)
  - Dynamic stripes → disjoint [0, 32M), [32M, 1G), [1G, EOF)



# Fault Tolerance

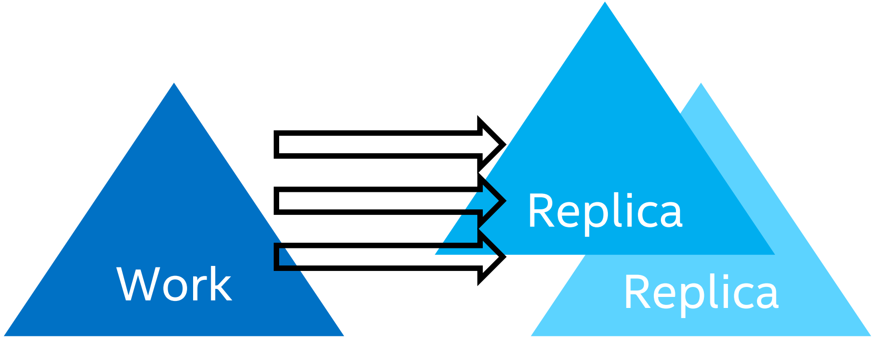
## Replication within the filesystem

| 4 stripes<br>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



# Scaling Capacity and Performance with HSM

## Hierarchical Storage Management

- Tiered storage provides an online library of permanent production data
- Massive performance in the Lustre* tier(s)
- Massive capacity in the archive tier
- Framework in place since Lustre 2.5
- Allows multiple storage tiers within the filesystem itself

## Ongoing investment to provide complete platform

- Parallel data mover – high performance interface to multiple archives
- Policy engine – data management automation for billions of files

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

# Parallel Data Mover

Highly scalable parallel copy tool

- General-purpose “engine”
- Extensions to support diverse range of HSM archives
- Extensions to support multi-site replication
  - Disaster recovery
  - Online backup

When data has to be transferred, it should be transferred as fast as possible

# Policy Engine

Policy engine provides data management automation for digital assets

Defines rules for managing capacity, archival, replication, migration, etc.

- Archive and purge inactive files
- Migrate files between storage tiers within filesystem
- Manage file replicas in case of OST failures
- Copy critical data to DR site every 2 hours

A policy engine for Lustre* must support very large scale

- Billions of inodes
- Multiple metadata servers
- High transaction rates

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

# 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

# Features of a Secure System

Authentication - proper identification of systems and users

- Node or user based authentication

Authorization - permission based access control

- Allow only specific authorized users access to resources

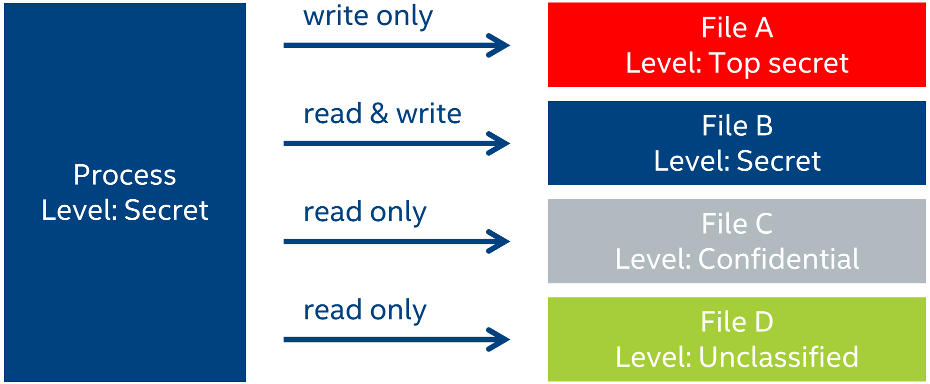
Encryption - protect data in flight and at rest

- RPC traffic encryption
- Disk or filesystem data encryption

# 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

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

# Summary

Intel and the Lustre* community continue to drive innovation

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

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