The Evolving Lustre* Landscape

Jessica Popp
Director of Engineering, High Performance Data Division

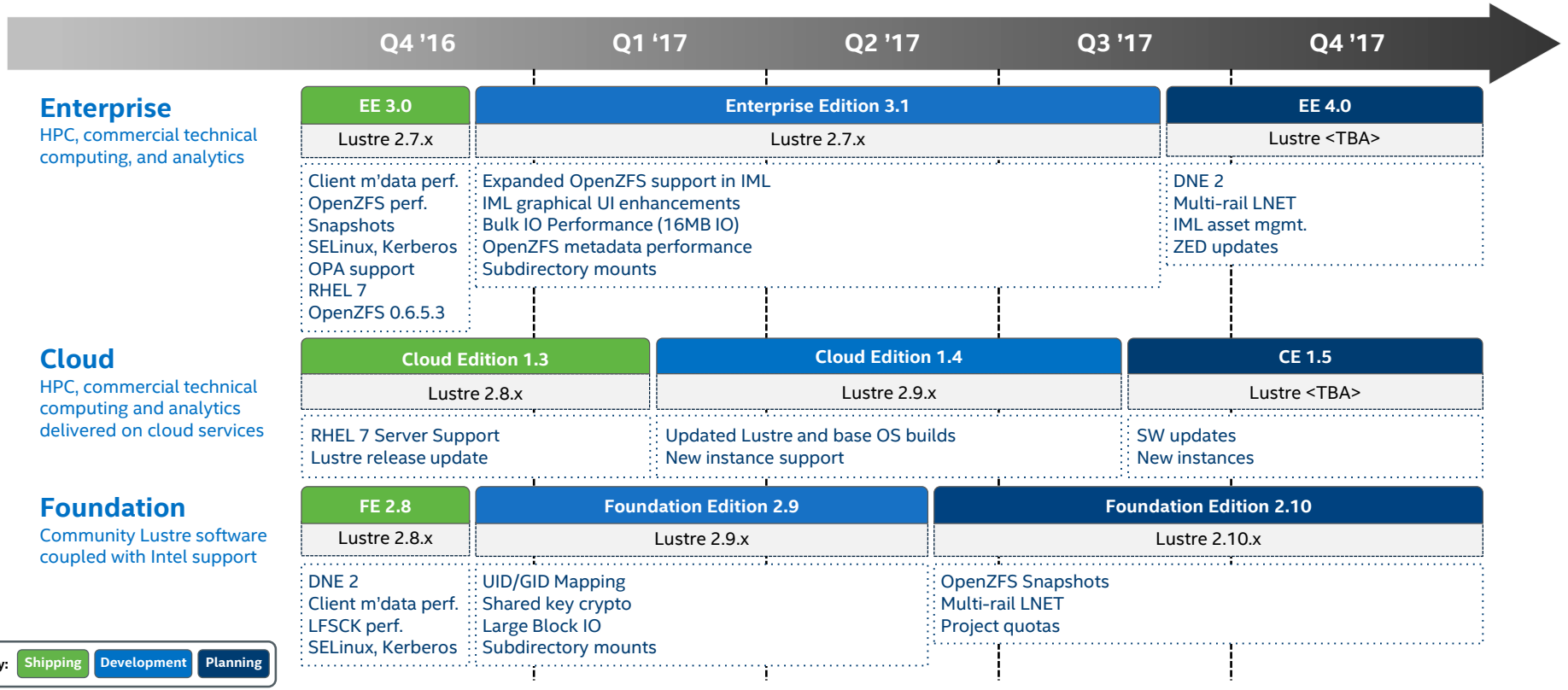
Legal Disclaimers

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.

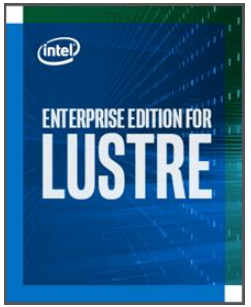
Intel® Solutions for Lustre* Roadmap



Product placement not representative of final launch date within the specified quarter

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

Intel® Enterprise Edition for Lustre* software 3.1



Target GA Q4 2016

- Based on community 2.7 release
- Latest RHEL 7.x server/client support
- IML managed mode for OpenZFS*
- GUI enhancements
- Bulk IO Performance for ldiskfs (16MB IO)
- Metadata performance improvements for Lustre on OpenZFS
- Subdirectory Mounts

Intel® Cloud Edition for Lustre* software 1.3

Software defined storage cluster available via AWS

Provides a performant, scalable filesystem on demand

Powered by Lustre 2.8

Provides secure clients, IPSec and EBS encryption

Available through AWS Cloud, AWS GovCloud, Microsoft Azure

Microsoft Azure

amazon web services™

*Features may vary by platform

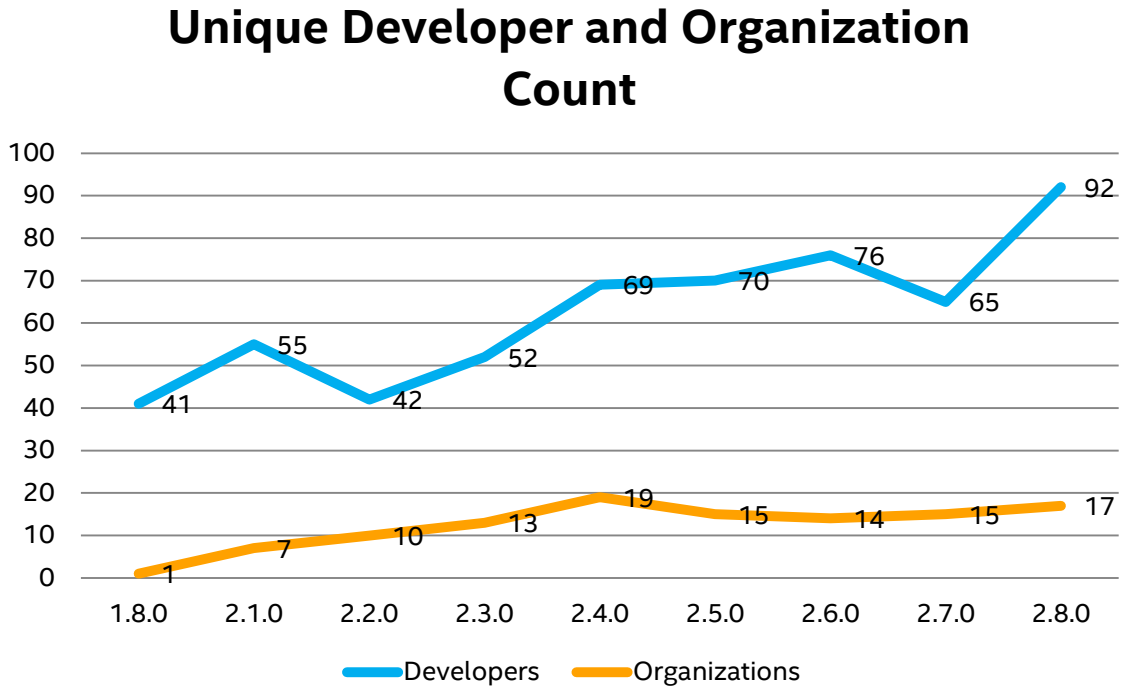
Intel® Foundation Edition for Lustre* software 2.9

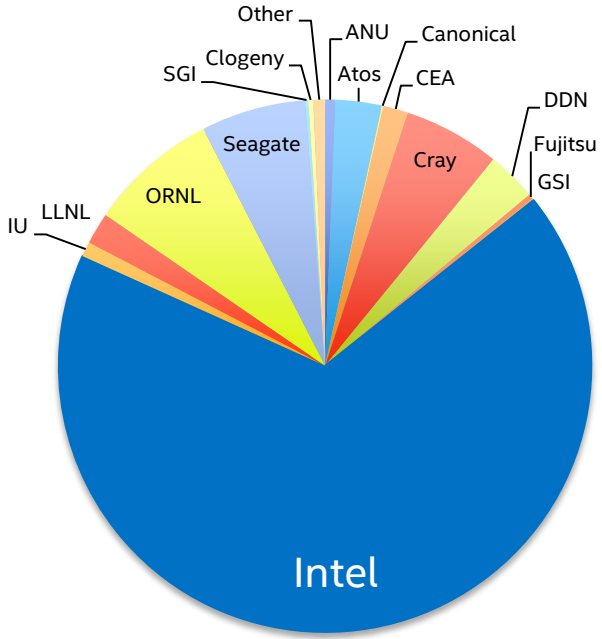
GA November 2016

Features

- RHEL* 7.2 servers/clients; SLES12* SP1 client support
- ZFS* Enhancements (Intel, LLNL)
- UID/GID mapping (IU, OpenSFS*)
- Shared Secret Key Encryption (IU, OpenSFS)
- Subdirectory mounts (DDN*)
- Server IO advice and hinting (DDN)
- Large Block IO

Development Community Growth



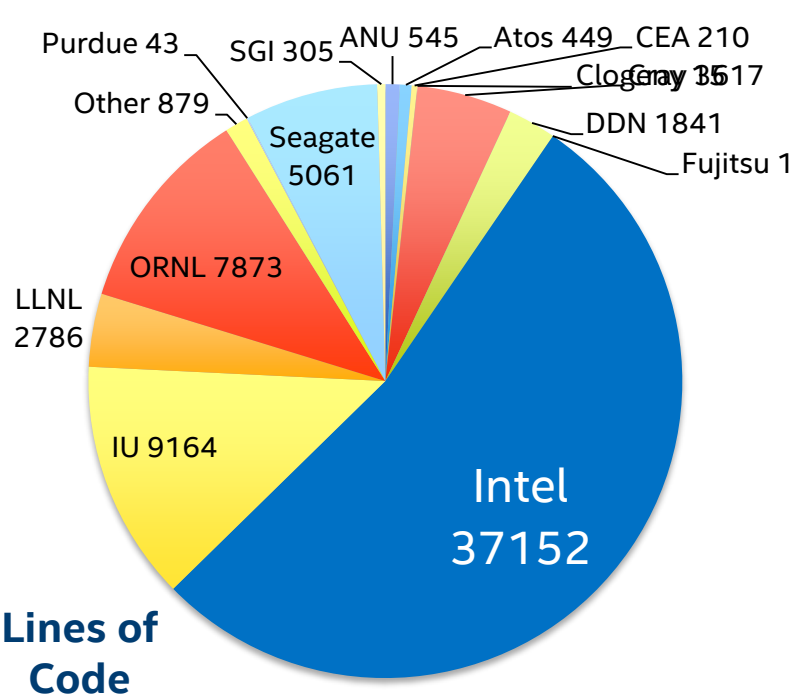


Lustre 2.8 Commits

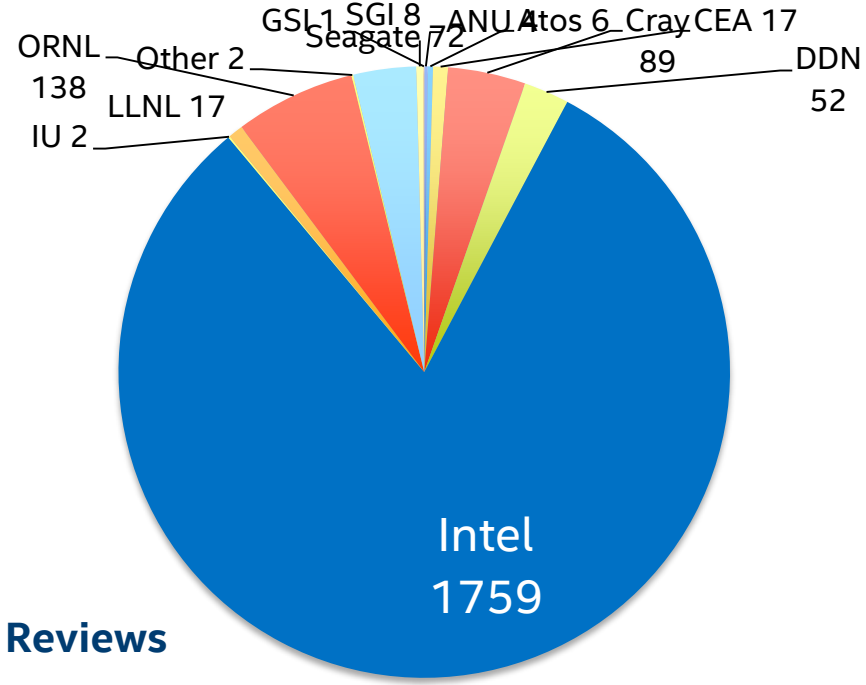
Statistics courtesy of Lawrence Livermore National Laboratories (Chris Morrone)

Source:
http://git.whamcloud.com/fs/lustre_release.git/shortlog/refs/heads/b2_8

Lustre 2.9 – Contributions So Far



Lines of Code



Reviews

Statistics courtest of Dustin Leverman (ORNL)
Source: http://git.whamcloud.com/fs/lustre-release.git
Aggregated data by organization between 2.8.50 and 2.8.60

Advanced Lustre Research
Intel Parallel Computing Centers

Uni Hamburg + German Client Research Centre (DKRZ)

- Client-side data compression
- Adaptive optimized ZFS data compression

GSI Helmholtz Centre for Heavy Ion Research

- TSM* HSM copytool

University of California Santa Cruz

- Automated client-side load balancing

Johannes Gutenberg University Mainz

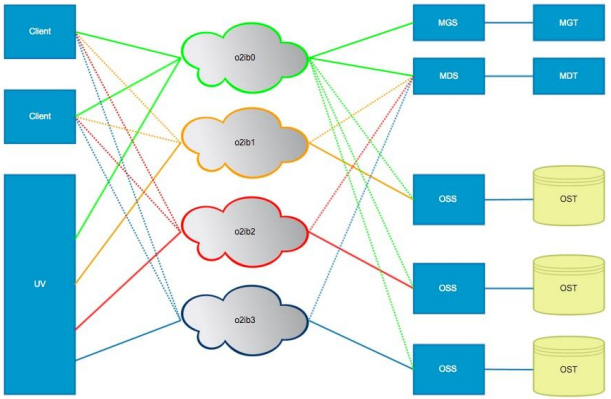
- Global adaptive IO scheduler

Lawrence Berkeley National Laboratory

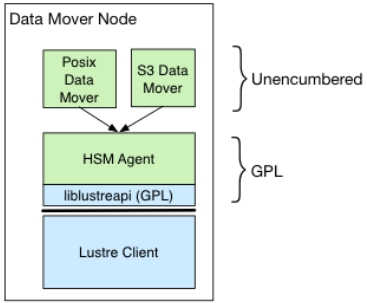
Lustre Feature Development

Multi-Rail LNet (SGI*, Intel)

- Improve performance by aggregating bandwidth
- Improve resiliency by trying all interfaces before declaring msg undeliverable



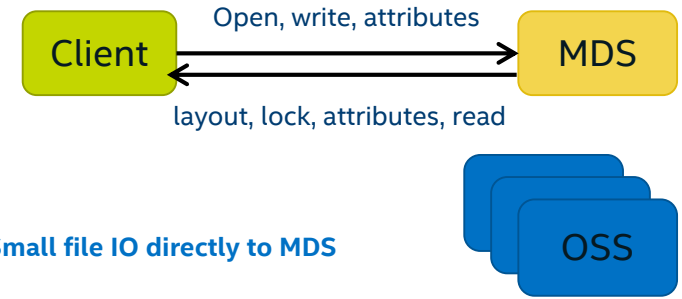
- HSM Data Mover
  - Copy tool to interface between Lustre and 3rd party storage
  - Early Access availability of agent with POSIX and S3 data movers via GitHub
  - Additional data movers can be licensed as desired by developers



Lustre Feature Development

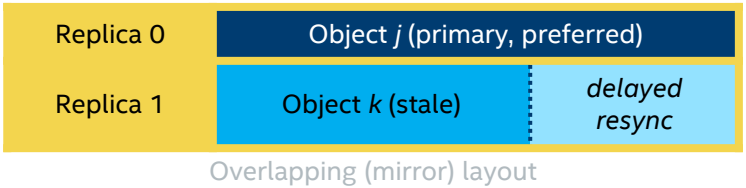
Data on MDT (Intel)

- Optimize performance of small file IO
- Small files (as defined by administrator) are stored on MDT

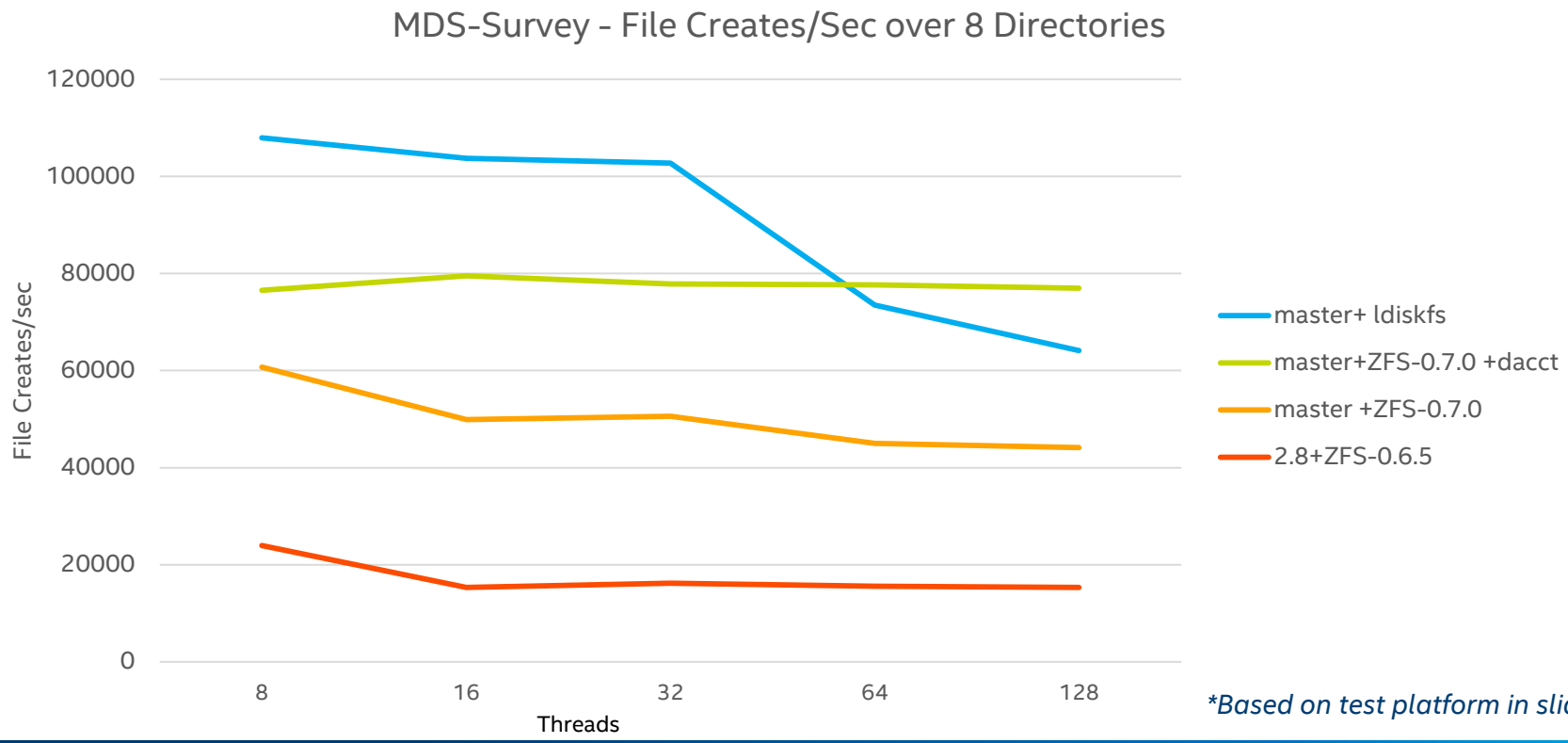


File Level Replication

- Robustness against data loss/corruption
- Provides higher availability for server/network failures
- Redundant layouts can be defined on per-file or directory basis
- Replicate/migrate between storage classes



ZFS Metadata Performance Improvements



*Based on test platform in slide 20

Data Security

Available in Intel® EE for Lustre* software 3.0

- **Secure Clients** – SELinux8 support to enforce access control policies, including Multi-Level Security (MLS) on Lustre clients
- **Secure Authentication and Encryption** – Kerberos* functionality can be applied to Intel® EE for Lustre* software to establish trust between Lustre servers and clients and to support encrypted network communications

In Development (2.9+)

- **Shared Secret Key Crypto** – data encryption for networks including RDMA
- **Node Maps** – UID/GID mapping for WAN clients
- **Data isolation via filesystem containers** – subdirectory mounts with client authentication

Upstream Linux Kernel Client

Refocused efforts on in-kernel Lustre client

- LNET virtually up-to-date with master
- Bug fixes to 2.5.51 landed, 2.5.58 pending

Challenges

- Features currently at 2.3.54 (now DNE, HSM)
- Significant code cleanups still needed

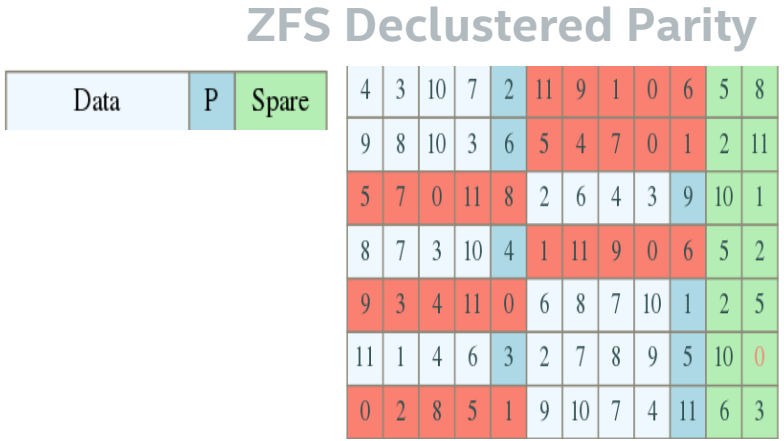
Lustre community developers cited among most active for Linux 4.6 kernel

- Oleg Droking (Intel): #5 by lines of code
- James Simmons (ORNL): #16 by lines of code

Lustre Development for CORAL

dRAID

- Optimizations for large streaming IO
- Distributes data, parity, and spare blocks evenly across all disks
- Optimized and scalable RAID rebuild limits exposure to cascading disk failures and minimally impacts application IO



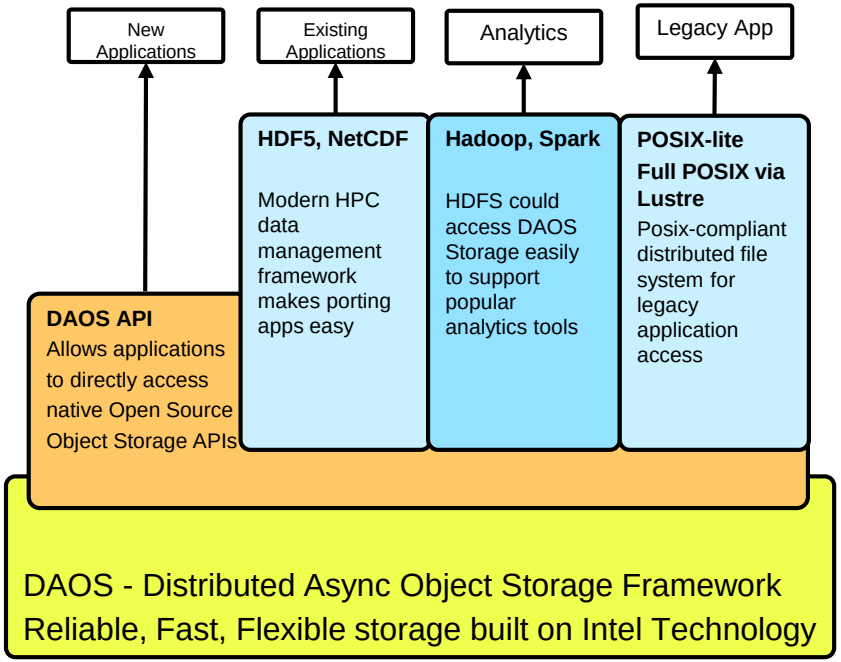
Lustre Streaming

- End-to-end large block support
- Block allocation improvements

All work will be landed in appropriate upstream trees (Lustre and OpenZFS)

DAOS: The Future of Storage

- Open source, scalable, software-defined userspace storage system providing object file system storage on Intel hardware
- Applications can realize immediate benefit w/ little to no change via middleware modifications
- Fine-grained IO and Versioning provide greater speed and control over data
- DAOS exploits 3D-Xpoint™ DIMM & NVMe and Intel® OmniPath Architecture
  - Intel technology improves performance by reducing latency & keeping data closer to computing resources



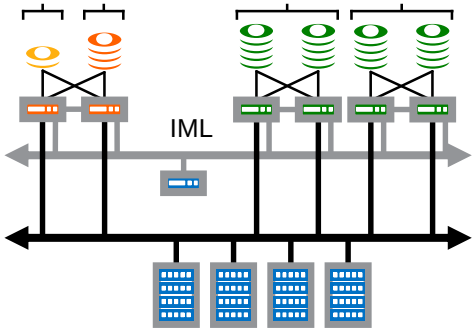
The Future is both Evolutionary & Revolutionary

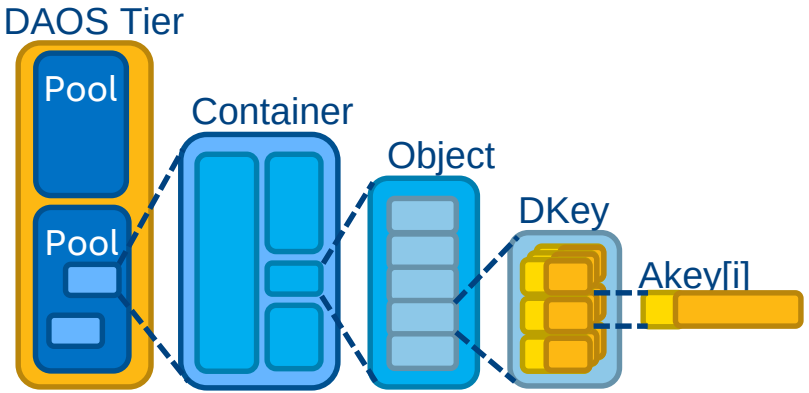
Lustre evolving in response to:

- A Growing Customer Base
- Changing use cases
- Emerging HW capabilities

DAOS exploring new territory:

- What may lay beyond POSIX
- Use new HW capabilities as storage
- Object storage model exposes new capability for scalable consistency





Thank You.

Backup

Test Configuration

- 2 x Intel(R) Xeon(R) CPU E5-2660 v2 @ 2.20GHz – 20 cores
- 64GB RAM
- 3 x 500GB local SATA HDD 7200 RPM
- CentOS 7.2.1511
- 3.10.0-327.28.2.el7 kernel (RHEL7)
- No remote clients, just local MDS testing (mds-survey script)

HPE Scalable Storage with Intel Enterprise Edition for Lustre*

High Performance Storage Solution

Meets Demanding I/O requirements

Performance measured for an Apollo 4520 building block

- **Up to 17 GB/s Read/15 GB/s Writes with EDR[1]**
- **Up to 16 GB/s Reads and Writes with OPA[1]**
- **Up to 21GB/s Reads and 15GB/s Writes with all SSD's[2]**

Designed for PB-Scale Data Sets

Density Optimized Design For Scale

- Dense Storage Design Translates to Lower $/GB
- Limitless Scale Capability – Solution Grows Linearly in

"Appliance Like" Delivery Model

Pre-Configured Flexible Solution

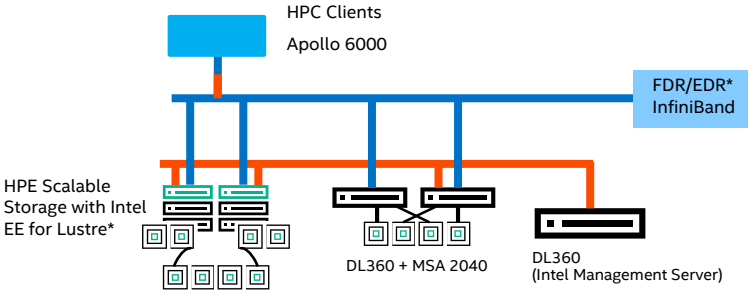
- Deployment Services for Installation
- Simplified Wizard Driven Management thru Intel Manager for Lustre

Innovative Software Features

Leading Edge Yet Enterprise Ready Solution

- ZFS software RAID provides Snapshot, Compression & Error Correction
- ZFS reduces hardware costs with uncompromised performance
- Rigorously Tested for Stability & Efficiency

Built on Apollo 4520



1: Different Conditions & Workloads can affect the Performance

2: 24 x1.6TB MU SSD's in A4520 no JBOD

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


intel®
experience
what's inside™