Hewlett Packard Enterprise

DMF for Lustre
Apollo 4520 Lustre Solution
Multi-Rail Lustre LNET
HPC Data Management

Feb 2017

HPC (Compute/Storage) Roadmap

Forward-Looking Statements

This document contains forward looking statements regarding future operations, product development, product capabilities and availability dates. This information is subject to substantial uncertainties and is subject to change at any time without prior notification. Statements contained in this document concerning these matters only reflect Hewlett-Packard Enterprise's predictions and / or expectations as of the date of this document and actual results and future plans of Hewlett-Packard Enterprise may differ significantly as a result of, among other things, changes in product strategy resulting from technological, internal corporate, market and other changes. This is not a commitment to deliver any material, code or functionality and should not be relied upon in making purchasing decisions.

Agenda

– DMF for Lustre Summary Overview

– Data Flow and Architecture

– Status and Best Practices

– RobinHood Status and Roadmap

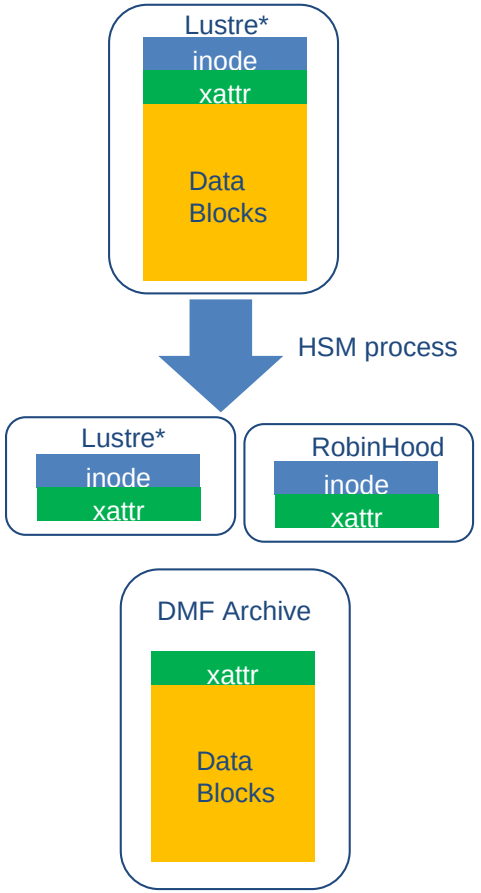
– DMF Policy Engine Roadmap

– HPE 4520 Lustre Platform

– Discussion

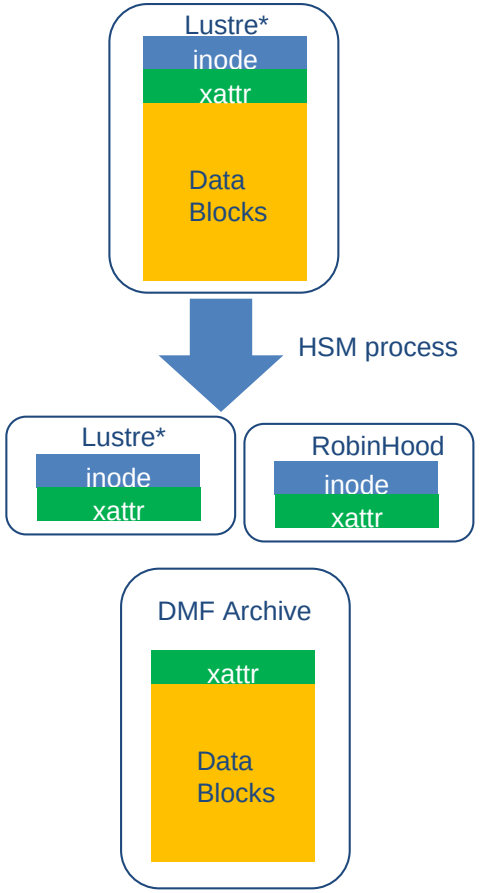
Data Management | **DMF for Lustre** Summary

- HSM support included as a feature of Lustre since Lustre 2.5
  - Three basic components to a Lustre* file:
    - inode (metadata – permissions, times)
    - xattr (extended attributes – striping layout (lov))
    - Data blocks (data)
  - Lustre* DMF HSM archives the data blocks and the xattr information
    - Inodes are asynchronously updated in Robinhood Policy Agent

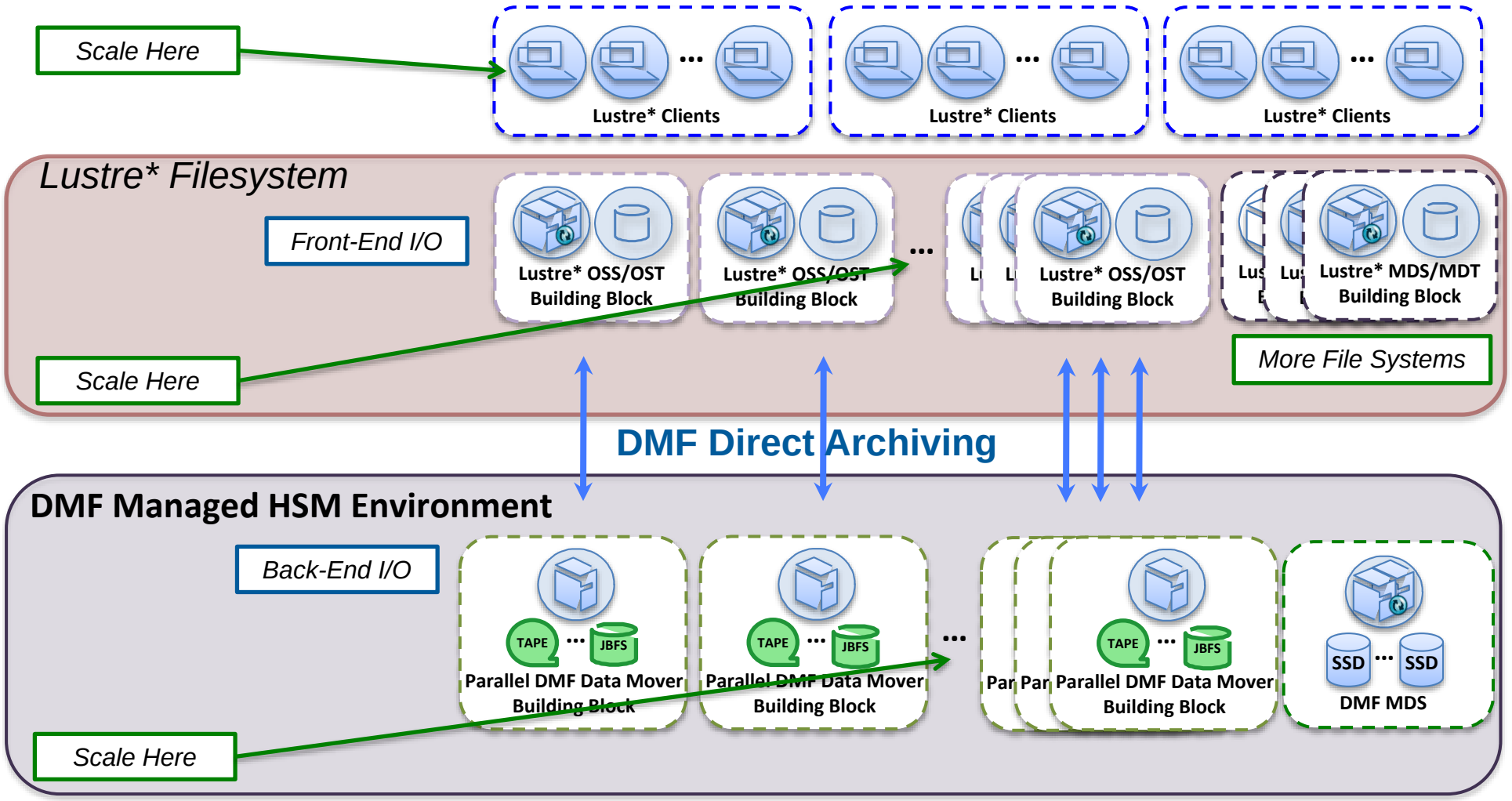


Data Management | **DMF for Lustre** Summary

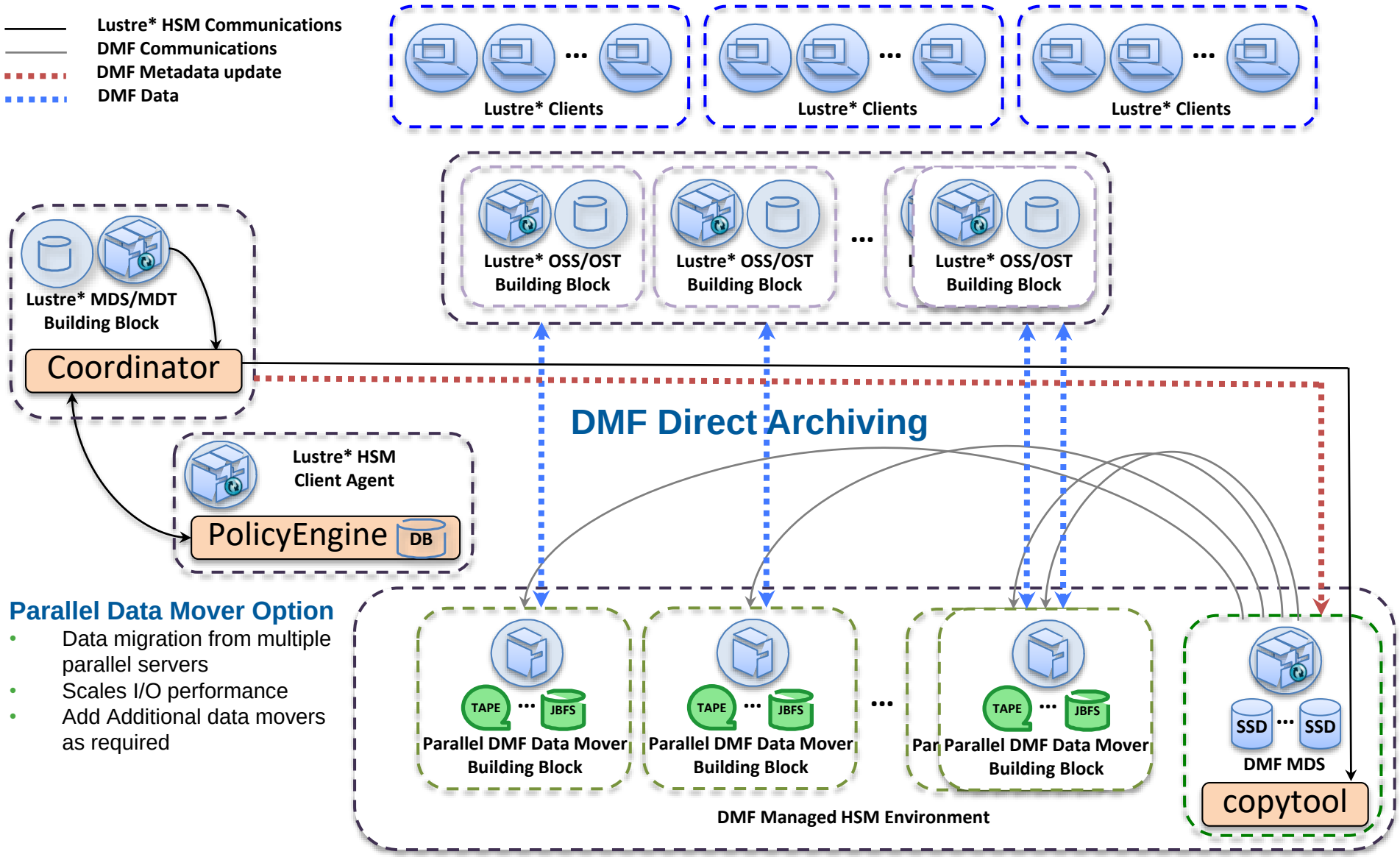
– A file stub and the xattr information stays in place on Lustre*
  – inode stays on primary storage (about 2K in size)
  – Typically this is called a stub
  – After a period of time, the data blocks on primary storage are released
– Files can be “restored” when needed (automatic)
– Lustre* has a fixed stub size (inode size ~2K)
– The Robinhood Policy Agent keeps a copy of the Inode
  – Asynchronously updated
  – Adds data protection and a recovery mechanism for Lustre
  – RobinHood is included as part of Intel EE for Lustre and supported as part of EE for Lustre support agreements



Data Management | **DMF for Lustre** Scalability & Performance



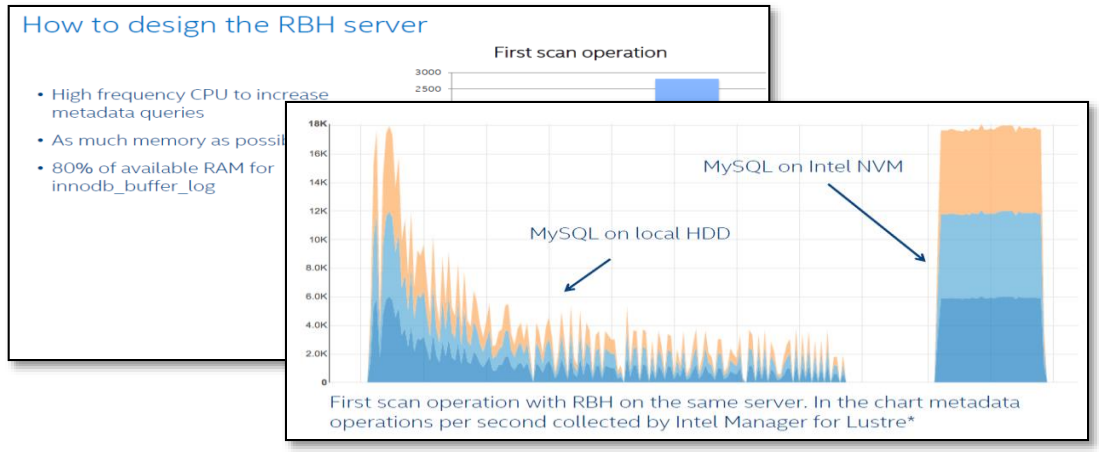
Data Management | **DMF for Lustre** Communication & Data Flow

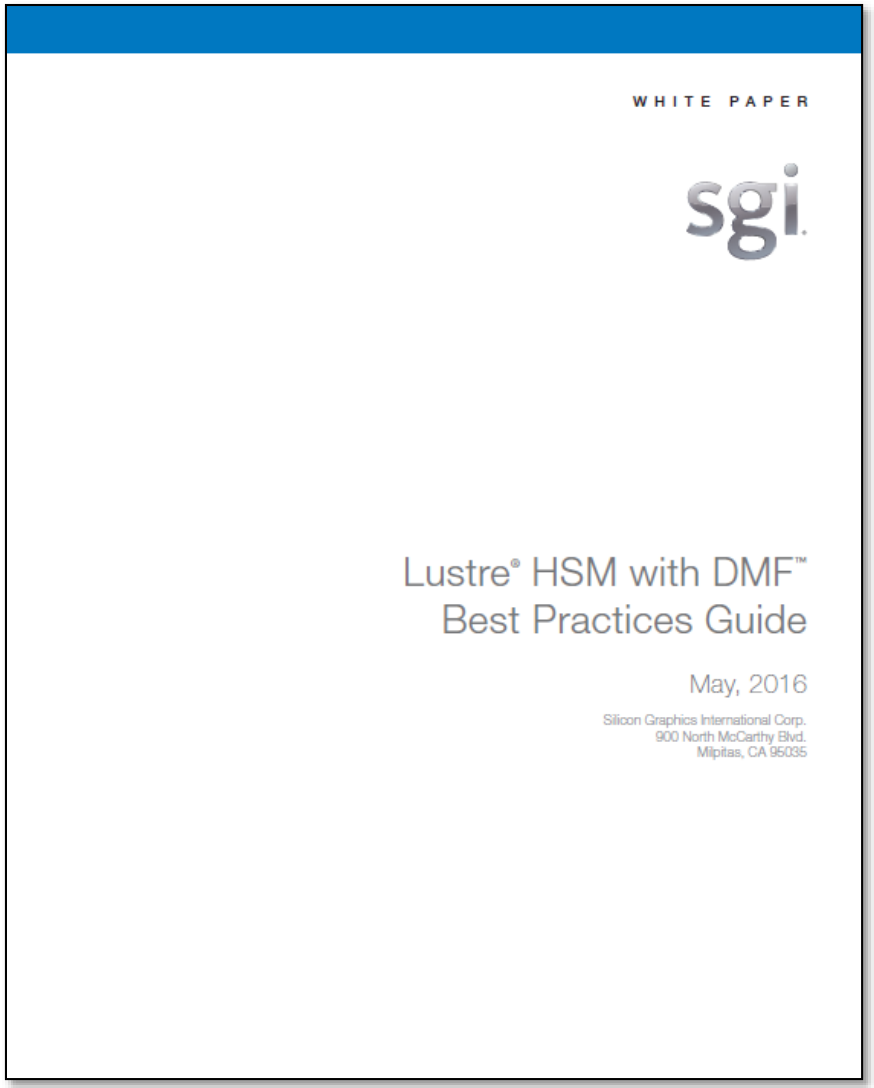


Hewlett Packard Enterprise

Data Management | **DMF for Lustre** Best Practice Guide

- Numerous customers in production
- Field deployments have occurred with both IEEL Lustre on SGI servers and with DDN ExaScaler (Intel IEEL).
- Sites have been extremely stable
- SGI Best Practice Guide available
  - Install, tuning, RH file recovery commands and process





Hewlett Packard Enterprise

Data Management | **DMF for Lustre** Evolution

- Lustre HSM integration using DMF v6 will continue to leverage Robinhood
  - Robinhood v3 has been successfully tested with DMF and will be validated with HPE Lustre solutions
  - DMF Best Practice guides will be updated to incorporate changes in Robinhood v3
- DMF for Lustre has been validated for use with Lustre DNE (multiple metadata servers)
- DMF v7.x (details in DMF v7 session) will incorporate native capabilities for log processing, HSM and file system recovery tools (no need for Robinhood)


SGI DMF for Lustre Data Sheet

SGI InfiniteStorage
DMF for Lustre

Extending Lustre Capacity
With Policy-based Tiered Data Management

Key Features
Policy Management & Active Data Migration
Storage Virtualization & Tiered Data Management
Cold Storage Integration with SGI ZeroWatt™ Power Control

Expanding the Capacity and Value of Lustre

Building on SGI's multi-decade strength as a leader in tiered data management solutions, SGI DMF for Lustre allows organizations to expand their use of Lustre file systems to include long term data storage in addition to being the file system of choice for high performance data access within cluster compute environments.

With DMF for Lustre, active data is always available on the fastest tier of Lustre storage (OST) while tunable policies allow the SGI solution to automatically migrate data to and from additional tiers of storage that can include lower cost disk, public and private cloud and even robotic tape libraries. The result is a system where all data is online and available for access at all times - but where storage costs are managed and minimized based on the organization's objectives and data usage patterns.

Broad Support for Storage Technologies

SGI's DMF supports storage tiers that can include Storage Area Network (SAN) arrays and Network Attached Storage (NAS) devices from a broad range of open systems suppliers. Object-based storage and cloud-based storage that leverage Amazon S3 interfaces are also supported in both private cloud and public cloud models using adapters that are created, validated and supported by SGI.

Additionally, SGI is pioneering the delivery of integrated hardware-and-software capabilities for intelligently managed cold storage with extremely low storage cost characteristics while also delivering operational savings through software-managed ZeroWatt™ power control that can significantly reduce the power and cooling costs for disks containing less active data. And for more static "dark data", DMF supports a wide array of library-based tape storage platforms that can deliver very appealing cost per terabyte levels while maintaining the data in a ready-for-access state.

SGI DMF for Lustre: Direct Integration with Lustre Storage Environments

sgi

Hewlett Packard Enterprise

HPE Lustre Solution

High-Performance and Cost-Optimized Lustre Solution with ZFS

HPE 4520 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 GiB/s Read/15 GiB/s Writes with EDR[1]
- Up to 16 GiB/s Reads and Writes with OPA[1]
- Up to 21GiB/s Reads and 15GiB/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
- Linear performance and capacity scaling

Innovative Software Features

Leading Edge Yet Enterprise Ready Solution

- ZFS RAID provides Snapshot, Compression & Error Correction
- HPE integration with Intel Manager for Lustre

Services and support

Installation and support services

- Factory tested and validated, deployment services for installation
- 24/7 Support services

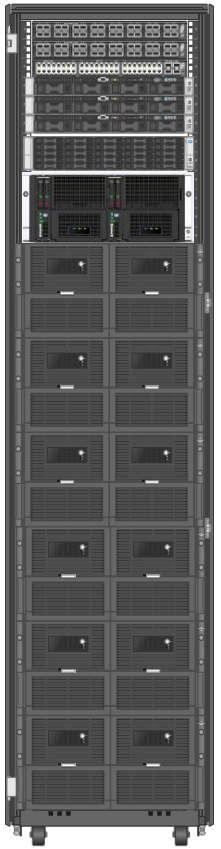
Hewlett Packard Enterprise

HPE 4520 with Intel Enterprise Edition For Lustre*

Easily optimized for a variety of needs

Capacity Optimized

- Up to 5.5 PB per rack
- Maximize $/GB
  - Up to 6 JBODs per Apollo 4520
  - Minimal software licensing
- Efficient JBOD chaining
  - HA configured to handle cable failures
- Licensing not based on capacity



Bandwidth Optimized

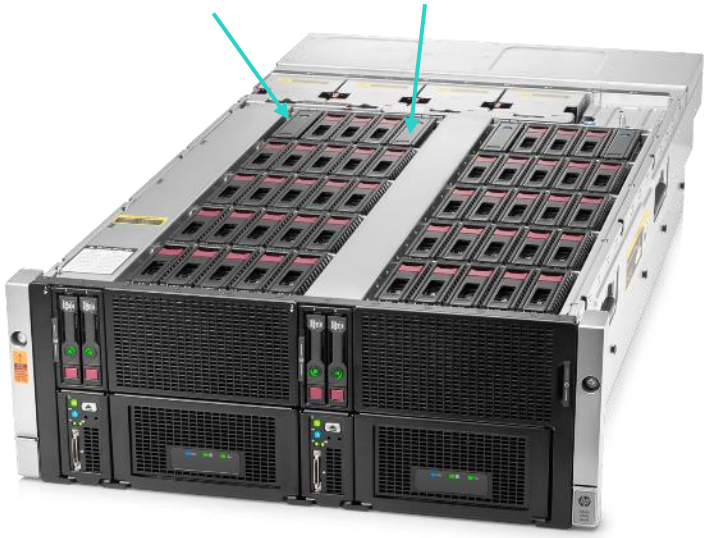
- Up to 21 GiB/s per Apollo 4520 using SSDs
  - ~200 GiB/s per rack when filled with Apollo 4520's
- Up to 17 GiB/s per Apollo 4520 using HDDs and two JBODs

Cost Optimized

- Lustre file system all in a 4U space with over 340TB usable capacity
- MDS/MGS all integrated in single 4520
- IML in separate 1U server
- Up to 5 GB/s writes and 4 GB/s reads
- Supports 140 million files with average size of 1.4MB

HPE Apollo 4520: System Details

SAS expanders in LFF drive bays



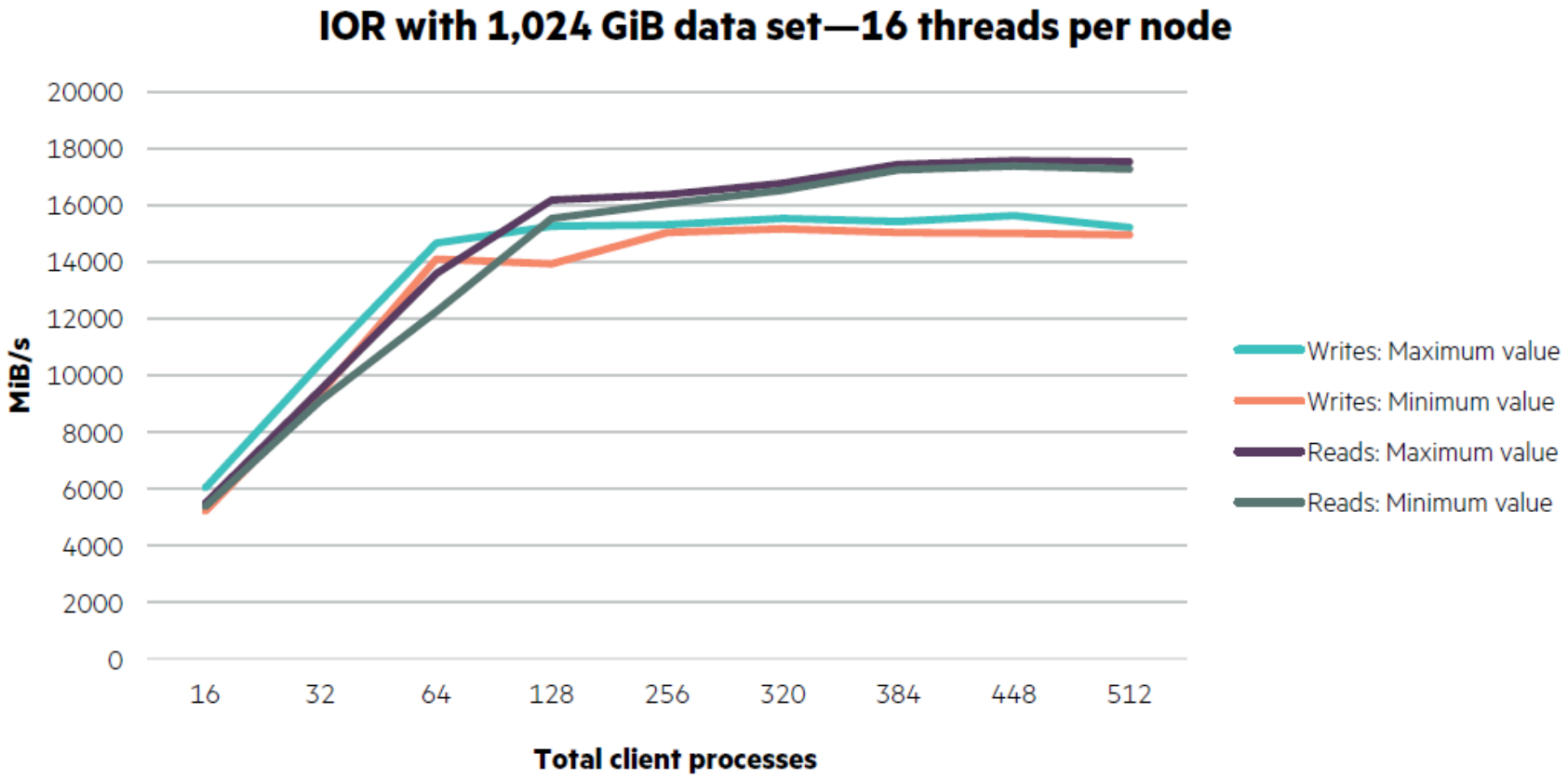
Feature	Server tray details
Processors	Up to 2 **Intel® Broadwell Xeon E5-2600 v4** processors
Memory	**16** DIMMs (8 per processor), registered, DDR4(2400/2133) w/ ECC
Drive Support	− 46 LFF SAS(12 Gb) drives in 2 server nodes − Includes support of SFF drives in converter − Hot pluggable SAS expanders − Support for Dual M.2 SSDs
Network	Dual-Port 1GbE with FlexibleLOM support **Support for EDR InfiniBand and Omni-Path**
Expansion	Up to **4 Low Profile PCIe Gen3 Slots** **Support for x16 HPC I/O module**
Display	SUV port, Video, Power/Health/UID Buttons and LEDs
Management	iLO 4 + one optional dedicated iLO NIC port
Other Features	**4U chassis height**, hot-plug redundant fans, HPE Gen9 Flex Slot power supplies (AC and DC versions), Support for dual M.2 SSDs
JBOD Expansion	Expand up to 6 D6020 Series JBODs

Hewlett Packard Enterprise

HPE Scalable Storage with Intel Enterprise Edition For Lustre*

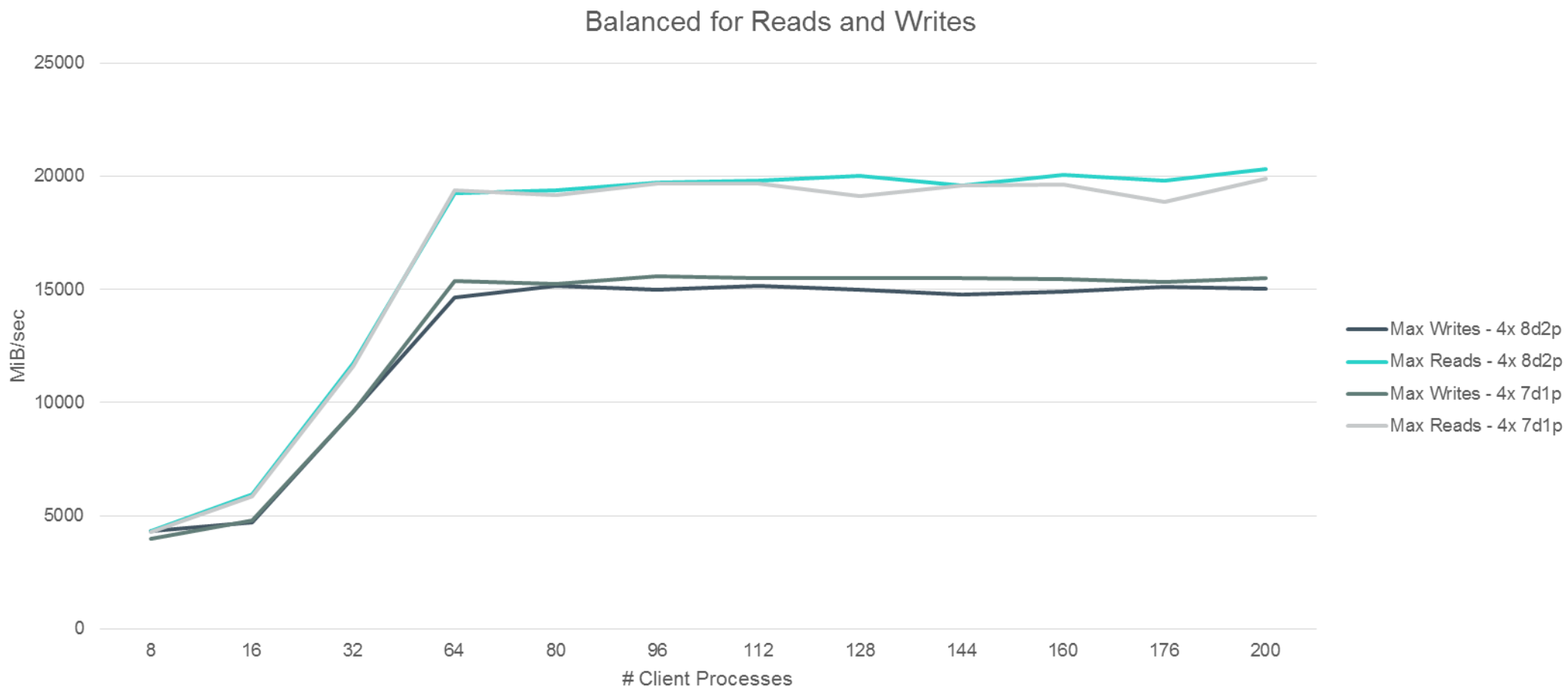
Performance with single Apollo 4520 and two D6020 JBODs (all HDDs)

Peak Writes: 15,630 MiB/s
Peak Reads: 17,568 MiB/s



HPE Scalable Storage with Intel Enterprise Edition For Lustre*

Performance with Single Apollo 4520 and SSDs



Hewlett Packard Enterprise

Flexible I/O Choices to Meet Your $/GB and $/Bandwidth Targets

x16 I/O module supports for EDR InfiniBand and Intel® Omni-Path

Choice of: Three x8 and one x16 PCIe 3.0 slots for highest bandwidth

OR

Four x8 PCIe 3.0 slots and one FlexibleLOM for greatest flexibility

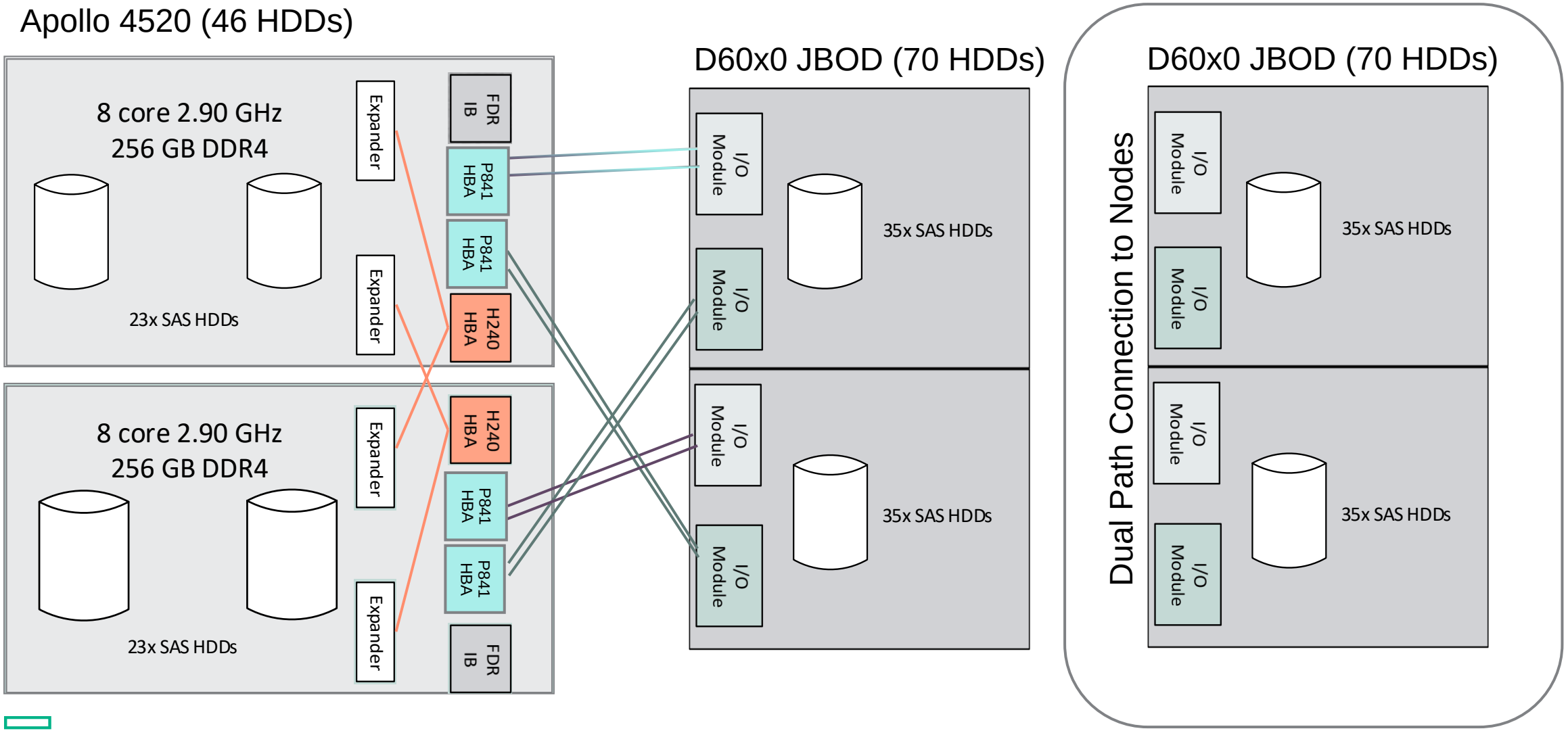
HPE Advanced Power Manager 2.0

(4) Flex Slot, hot-swap power supplies

- Shared across both server nodes for power efficiency
- 800W and 1400W options

Hewlett Packard Enterprise

Apollo 4520 Cabling



Hewlett Packard Enterprise

Multi-Rail Lustre LNET Capability

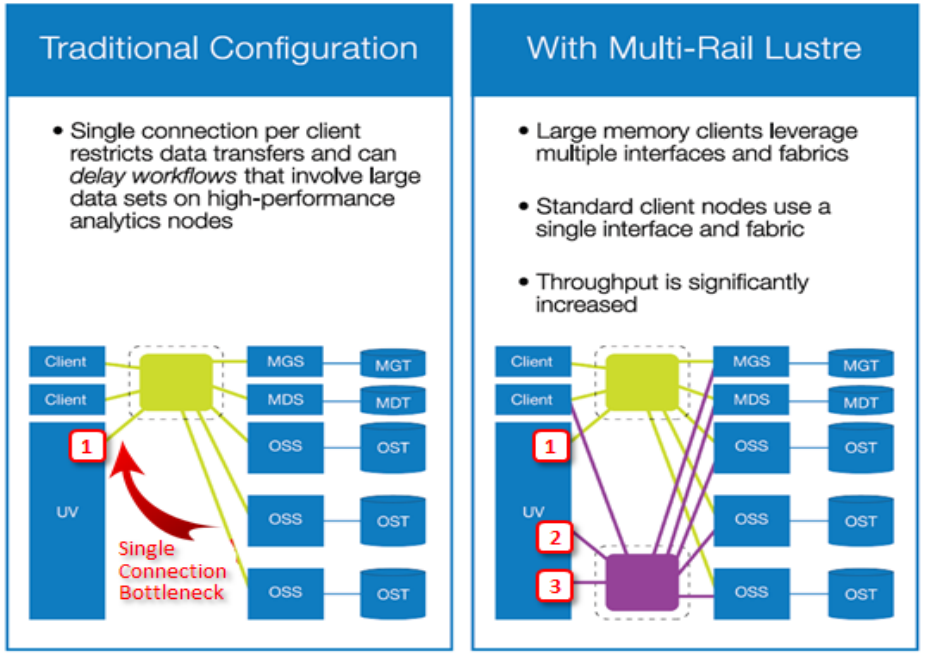
Scalable LNET-based Data Pipelines Leveraging Multiple Intel® Omni-Path or InfiniBand* Connections

Key Features :

- **Expands the IO throughput** of Lustre client or server nodes – and allows mixed us of multi-rail and single-rail nodes.
- **Supports both InfiniBand and Omni-Path** – and works with SGI UV and standard systems
- Enables multiple paths for **added resiliency**

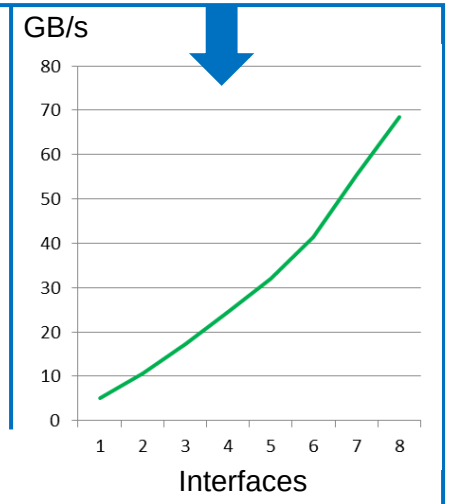
Key Benefits :

- **Accelerates data-intensive HPC and HPDA workflows** by significantly reducing time spent on data movement operations
- **Enhances the ROI** of existing Lustre storage infrastructures by enabling faster performance from in-place systems
- **Simple and consistent support model** as a standard feature in an upcoming release of Intel® Enterprise Edition for Lustre* Software



Performance Results:

- The plot gives I/O throughput in GB/s against number of IB interfaces.
- **Data throughput increases with the number of interfaces installed on a system**

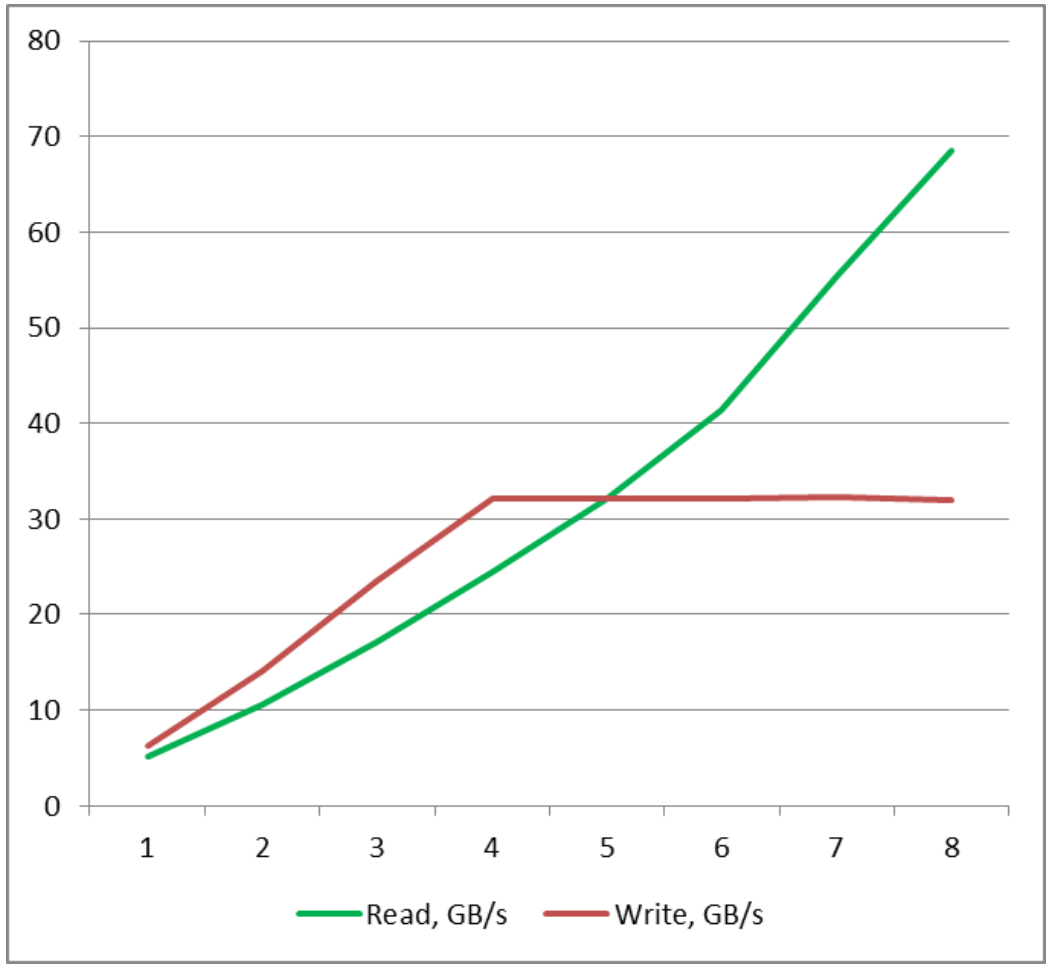


- SGI® UV™ In-Memory Supercomputers with Intel Xeon processors
- Intel Storage Cards & SSDs
- Intel Enterprise Edition Lustre
-

Hewlett Packard Enterprise

Multi-Rail Lustre LNET Capability | **Performance** as shown at SC16

- Actual test scenario demonstrated live at SC16
  - 4MB Direct I/O transactions using 512 threads distributed evenly among 32 sockets
- MC990 X (UV 300) with 32 Intel Xeon sockets
  - UV has eight Intel 100Gb/s OmniPath Adapter (OPA) cards
  - Line rate of all adapters is 100GB/s
- 8 OSS
  - Each OSS with OST based on ZFS and 4 Intel DC P3700 NVMe SSD flash cards
  - Each OSS has one Intel 100Gb/s OmniPath Adapter (OPA) cards
- Read performance of each card is around 2.4GB/s
  - Theoretical storage bandwidth on reads is 76GB/s
  - IOR shows 68GB/s (close to 90%)
- Write performance is around 1GB/s
  - Theoretical storage bandwidth on writes is 32GB/s
  - IOR show 32GB/s, (close to 100%)



Industry Adoption of ZFS based Lustre

– Path to Exascale
  – Coral and future follow-on architectures are scoped with ZFS.
– LLNL Sequoia1 (55PB File System)
  – Cheaper, less complex, higher performance file system for Sequoia.
– Heavy investment in advancing Lustre with ZFS
  – ZFS event daemon validation on HPE Apollo 4520.
  – ZFS small file performance testing.
  – Collaboration with OpenZFS community on new features.
  – Breakthrough metadata performance: LAD'16

[1] http://computation.llnl.gov/projects/zfs-lustre

Data Management | **DMF for Lustre** Summary

- DMF for Lustre is a stable and validated element of the DMF portfolio
- Solution has gone through significant testing and tempering
- Near term DMF v6 validation on HPE platforms and with Robinhood v3
- DMF v7 native log file processing and file system integration will simplify the deployment architecture (no need for Robinhood)

Hewlett Packard
Enterprise

Thank you