INTEL STORAGE software

Bryon Neitzel
Director, High Performance Data Division

THE NEW CENTER OF POSSIBILITY

Intel High Performance Data Division

- Mission:
  - Develop high performance IO software solutions for the worlds most challenging data movement and storage problems
- Scope:
  - Lustre* Feature and Maintenance Releases, Lustre L3 Support
  - Complete IO stack for large scale Deployments
  - Future Storage Software Technology (DAOS)

Lustre Business Model Changes

How we deliver products to the marketplace has changed:

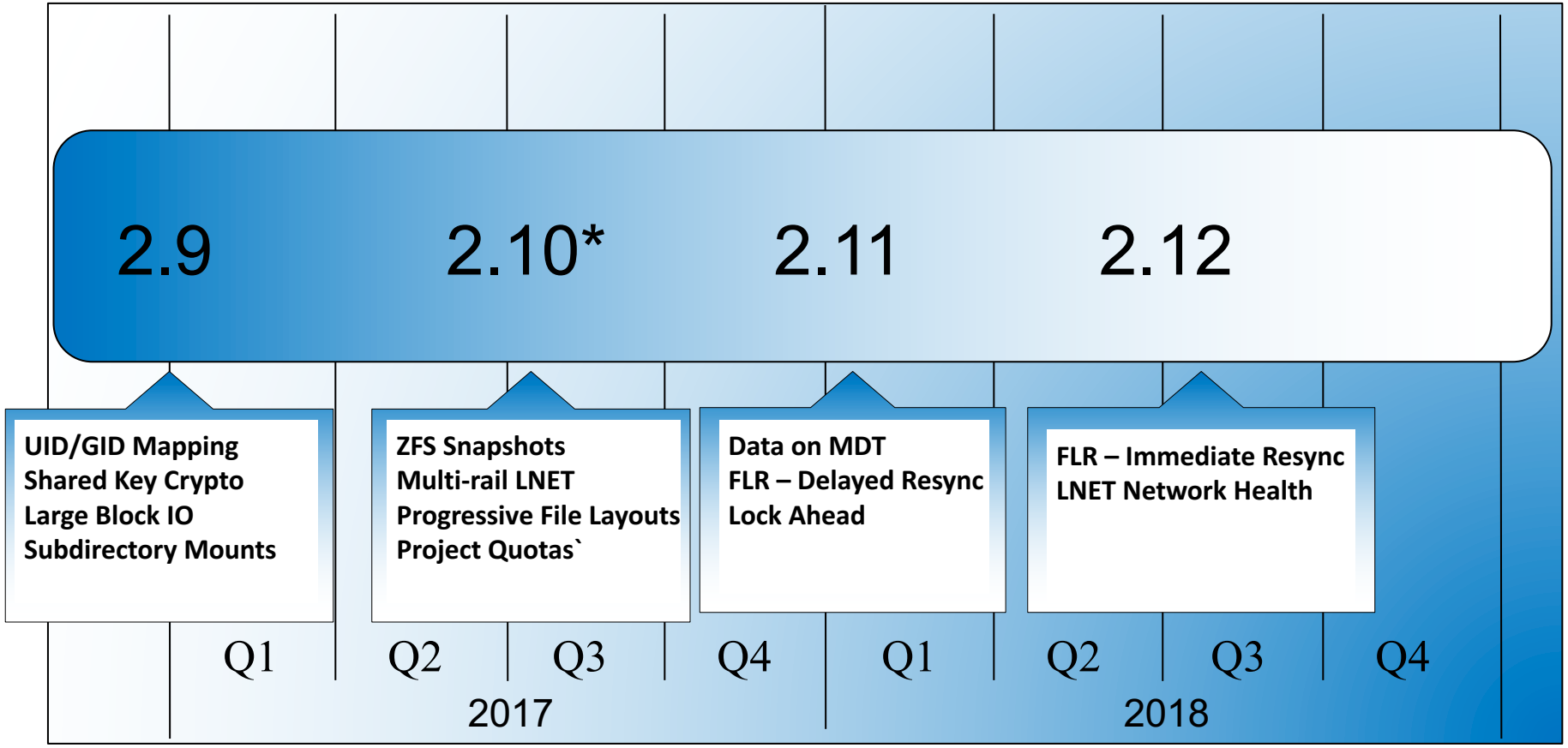
- All Intel contributions go **directly to Open Source** projects
  - Moving away from Intel-branded releases
  - All formerly proprietary components from Intel-branded releases have been open sourced (HAL, HAM, IML)
    - https://github.com/intel-hpdd/
- **Consolidating** Sales Functions with other Intel organizations
  - Focus on Level 3 support for future customers
  - Continued support for existing customers
- **Enhanced testing and stability** for Lustre Community Edition
  - One release means more focus on LTS Lustre stabilization and hardening, plus free maintenance releases
  - (Community, Foundation, Enterprise) -> Community

Lustre Business Model

What we build did **not** change:

- Ongoing delivery of **feature releases**
- **Support** for existing and ongoing installations for Intel branded releases.
- Lustre **development, test, release, support, and R&D** teams
- **Intel-funded hardware** for development, build, and testing of Lustre
- Involvement in **large scale machine** deployments
- Involvement in Lustre **community events** and groups like OpenSFS, EOFS, LUG, LAD, etc

Community Release Roadmap

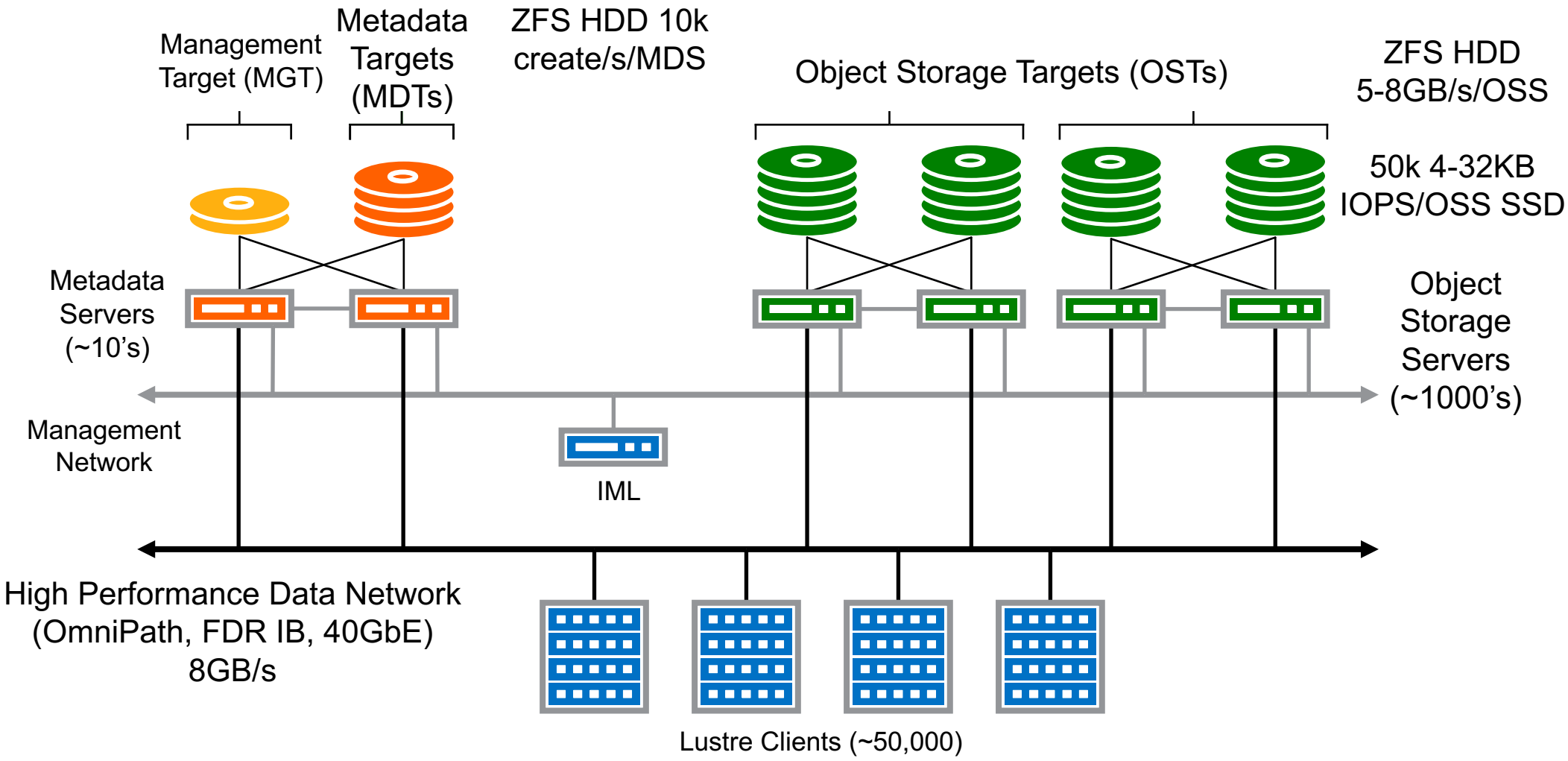


*LTS Release with maintenance releases provided

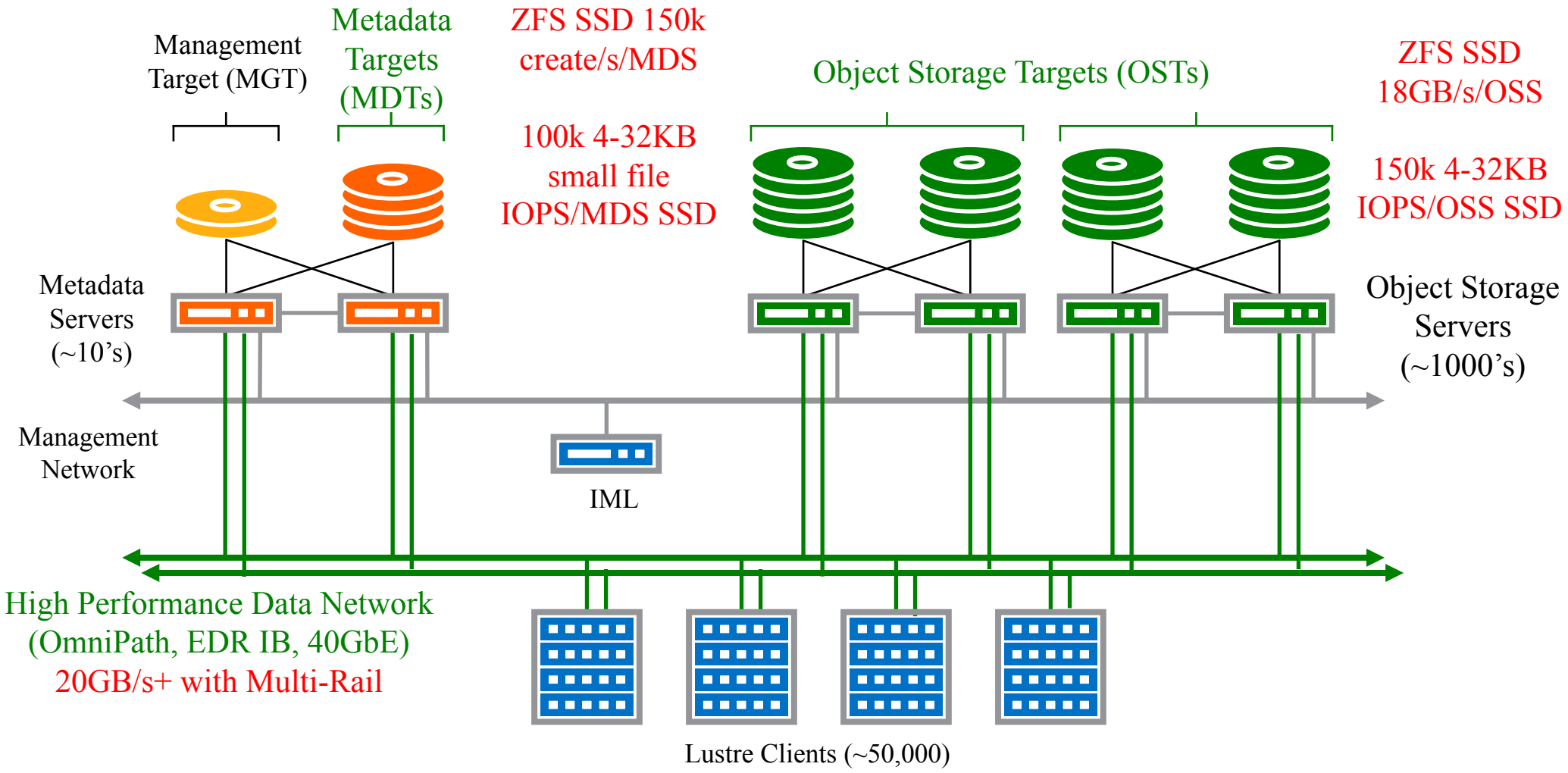
Estimates are not commitments and are provided for informational purposes only

Fuller details of features in development are available at http://wiki.lustre.org/Projects
Last updated: April 20th 2017

Improvements in Lustre Performance - Today

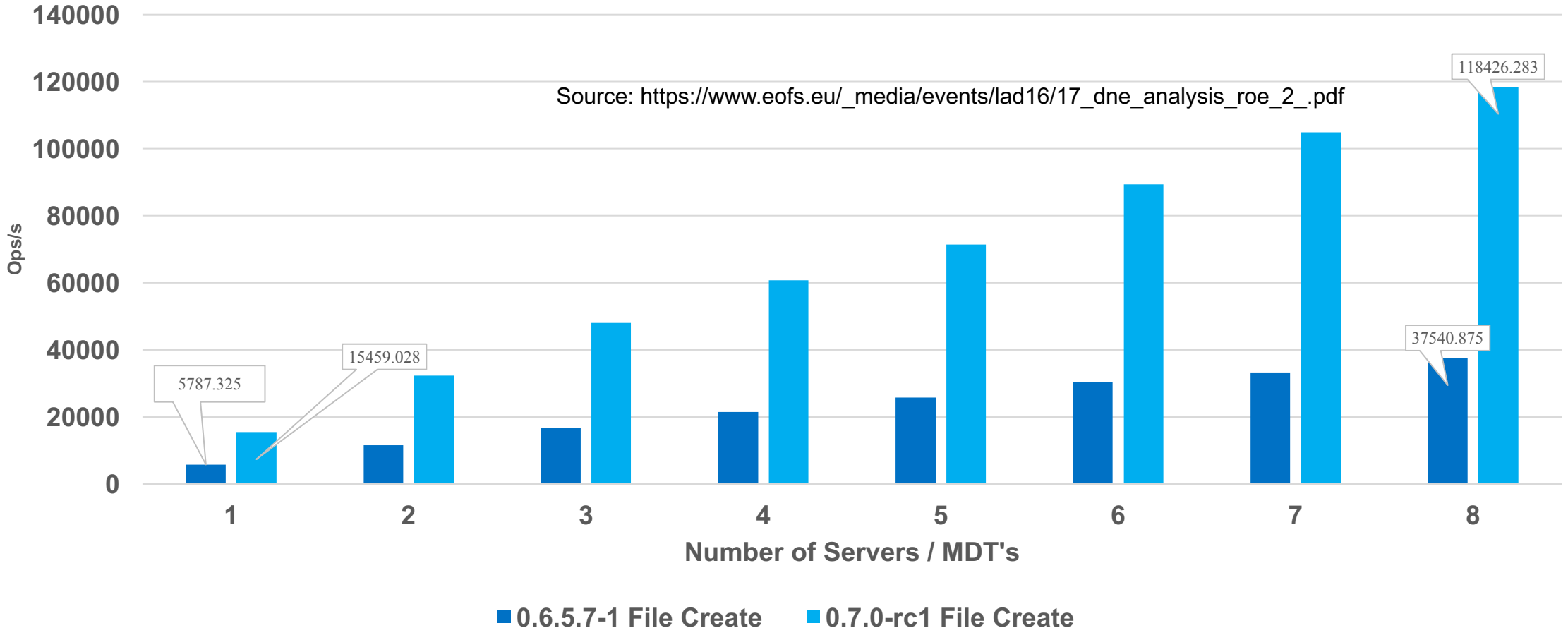


Improvements in Lustre Performance – 2.12 targets



Intel Focus on Scalability and Performance

DNE Phase I - File Create: ZFS 0.6.5.7-1 vs. 0.7.0-rc1



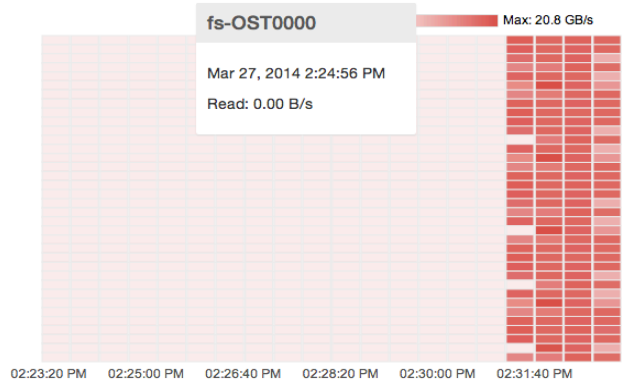
* Intel measured or estimated as of September 2016. Please see configuration details at end of deck

$MPICMD ./mdtest -i 3 -I 10000 -F -C -T -r -u -d /mnt_point/@/mnt/point2/@etc.

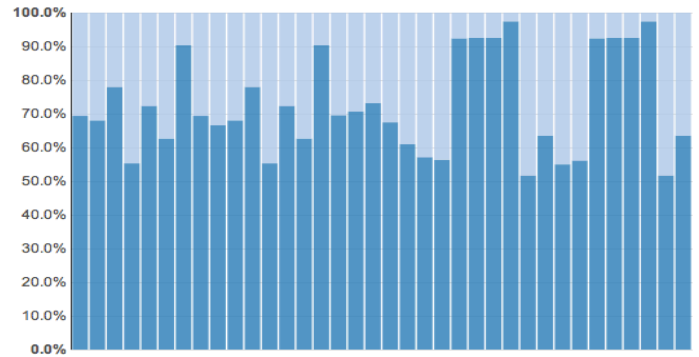
IML: Community-based Lustre Manager

Management and Monitoring Tool

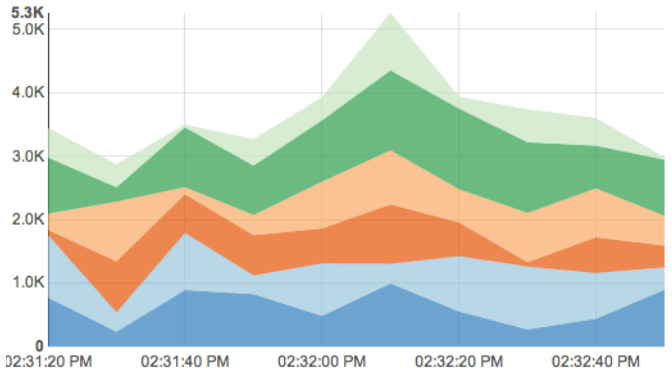
Read/Write Heat Map



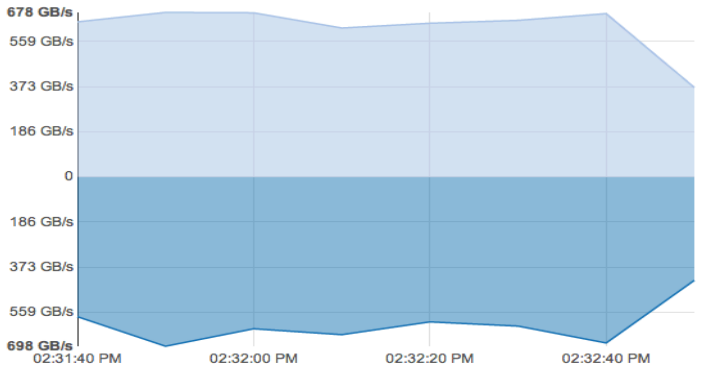
OST Capacity



Metadata Operations



Read/Write Bandwidth



- Intuitive, browser-based administration
- Lustre installation and configuration
- Real-time system monitoring
- Extensible through open, documented APIs

IML: Community-based Lustre Manager

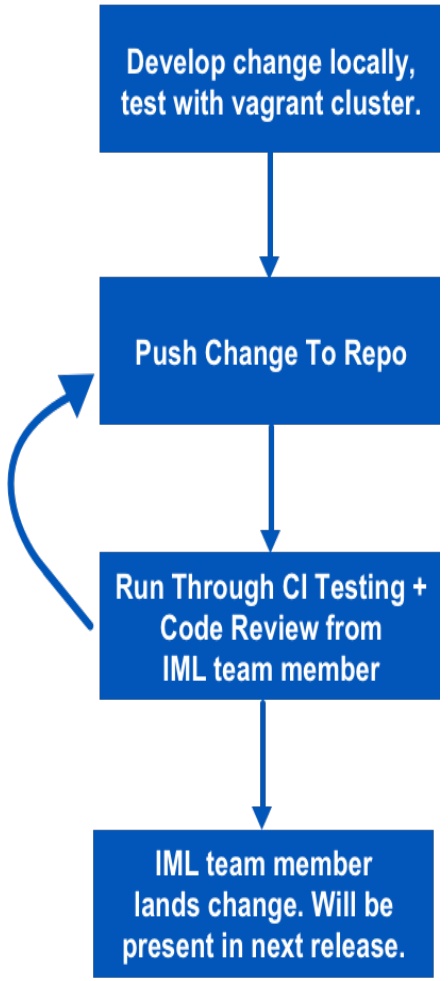
Details of Open Source Project

IML now available under MIT license at https://github.com/intel-hpdd/

- IML is a monorepo with a series of collaborator repos
- Each has CI mechanism running tests over changes: https://travis-ci.org/intel-hpdd/
- Providing convenient way to demo tool and test proposed change using Vagrant
  - https://atlas.hashicorp.com/boxes/search?utf8=%E2%9C%93&sort=&provider=&q=manager-for-lustre
- Following typical GitHub workflow; issues and pull-requests can be opened against specific repos. Use these to communicate / propose changes to IML team
  - Examples: https://github.com/intel-hpdd/intel-manager-for-lustre/issues, https://github.com/intel-hpdd/intel-manager-for-lustre/pulls.

Developing open-source roadmap; input from community greatly appreciated

IML 4.0 will be compatible with Lustre 2.10.x LTS releases; targeted for early Q3 release

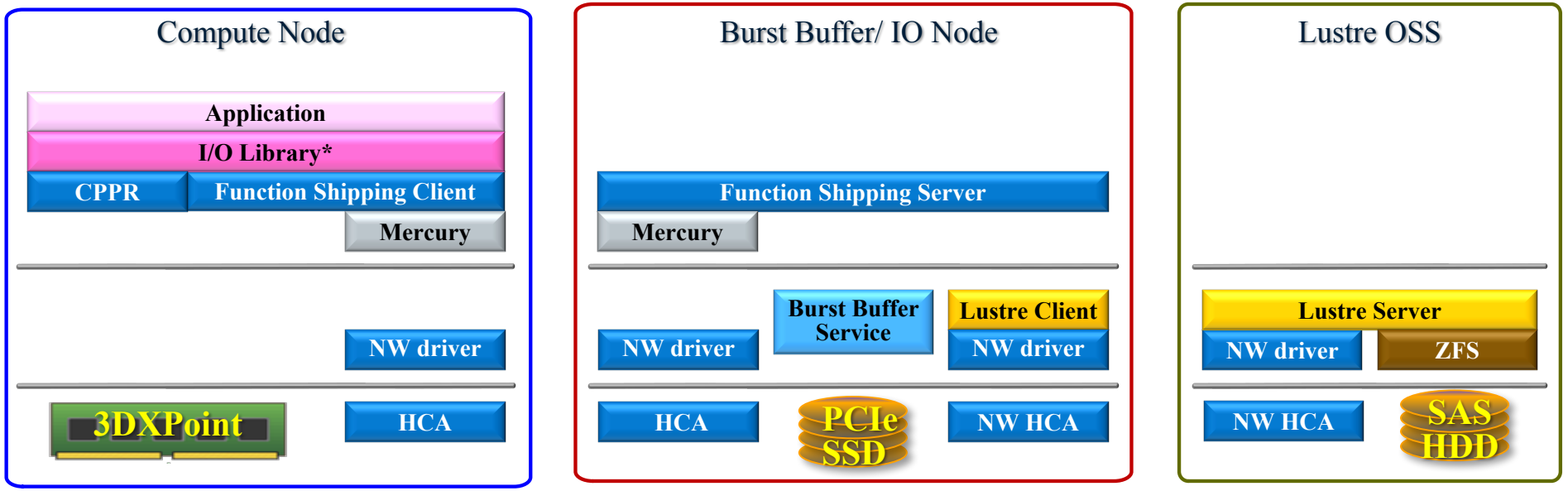


HPDD Storage Software

Open Source Landing Zones

HPDD Aurora I/O Programs

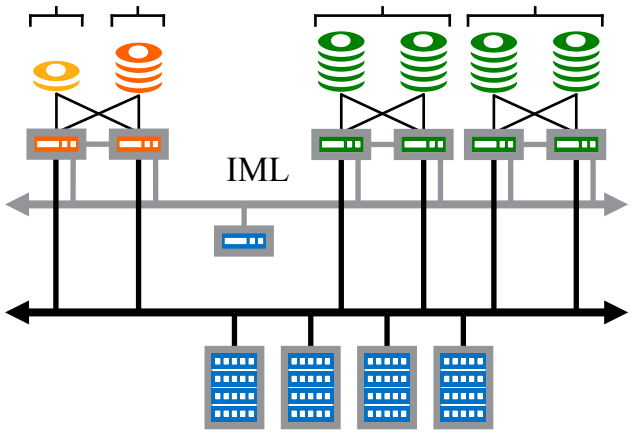
- ZFS — OpenZFS, ZFSonLinux
- Lustre — Lustre
- IO Forwarding — Mercury, OpenHPC
- CPPR — OpenHPC
- I/O LibTuning — HDF5, Various MPI



The Future is both Evolutionary & Revolutionary

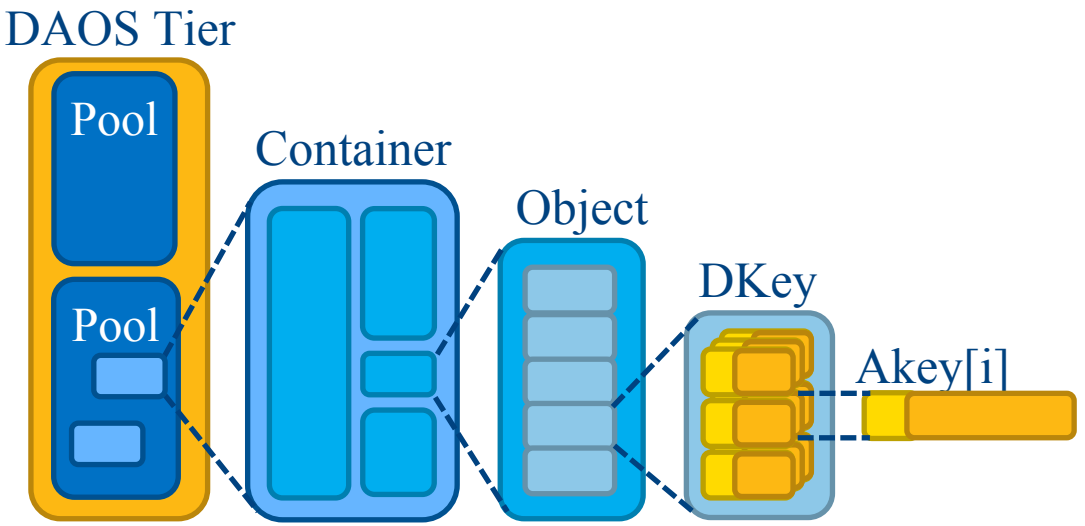
Lustre evolving in response to:

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

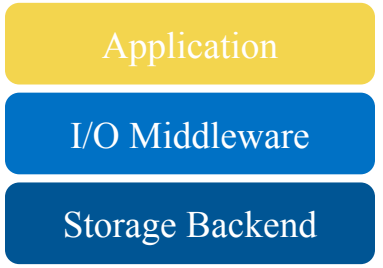


DAOS exploring new territory:

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



Extreme Scale Storage IO (ESSIO)



Joint Project with HDF Group to explore:

- New architectural directions
  - Massively distributed storage
  - Hot tier close to compute nodes
- Future programming models, runtimes and workflows
  - Legion
  - Asynchronous producer/consumer
- Analytics
  - Capture & index metadata
  - Help to derive value from data being produced as volumes explode

Pre-Productization version capability

- Data Model and KV interface
- Data replication / Online Rebuild
- Large & Small record support
- Metadata replication
- Snapshots (aka Epochs)
- Libfabric support
- HDF5 Support

Summary

- Mission:
  - Develop a rich portfolio of high performance storage products to solve the worlds most challenging data storage and IO problems
- Scope:
  - **Lustre** is the future of scalable **POSIX** storage
    - Advancing the Roadmap, Feature and Maintenance Releases, Commercial Level 3 Support
  - **DAOS** is the future of scalable **Object** storage
    - Complete IO stack for pre-Exascale Deployments, Rich High-Level Object Interfaces
  - Next Generation Storage R&D Projects **(including both Lustre and DAOS)**
- See Peter, Micah or Bryon during the break to ask questions – Thanks!

Notices and Disclaimers

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

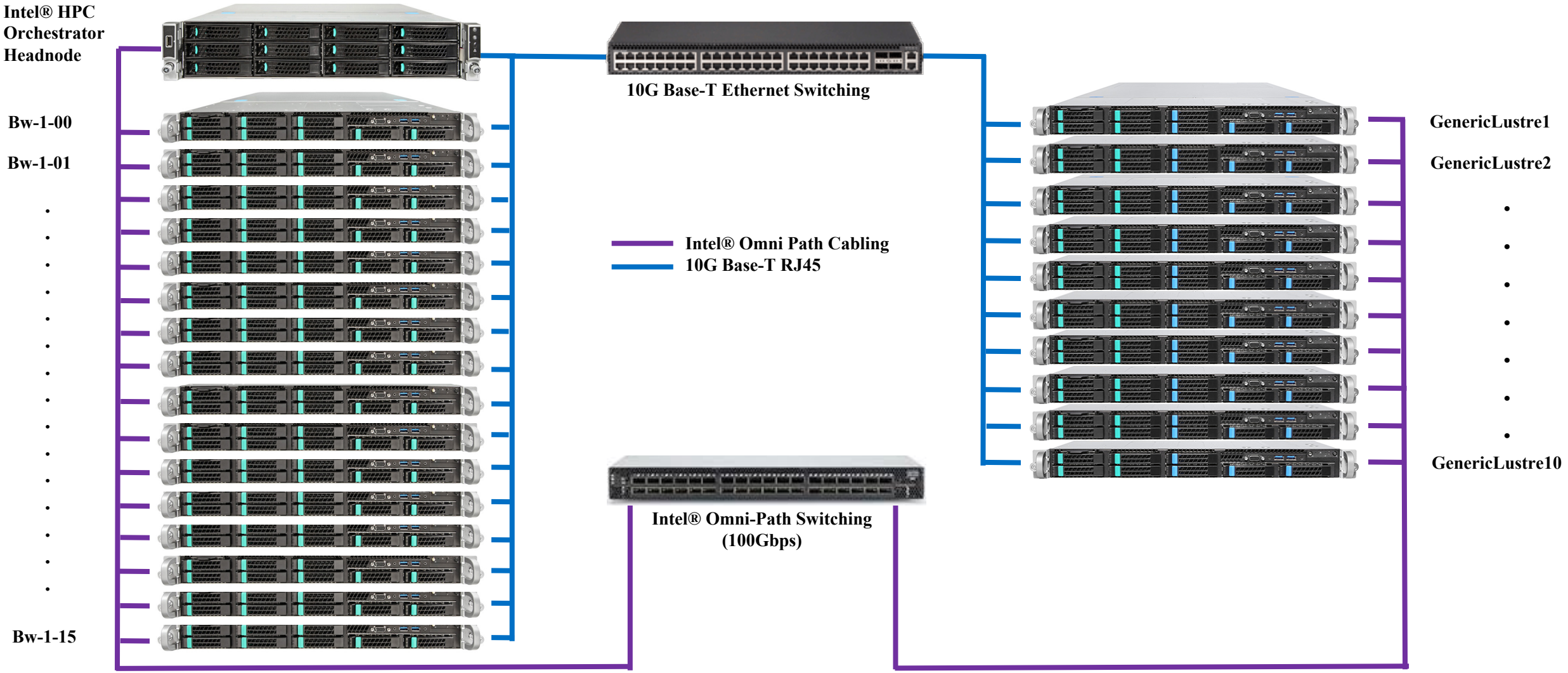
Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

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.

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, visit **www.intel.com/benchmarks**

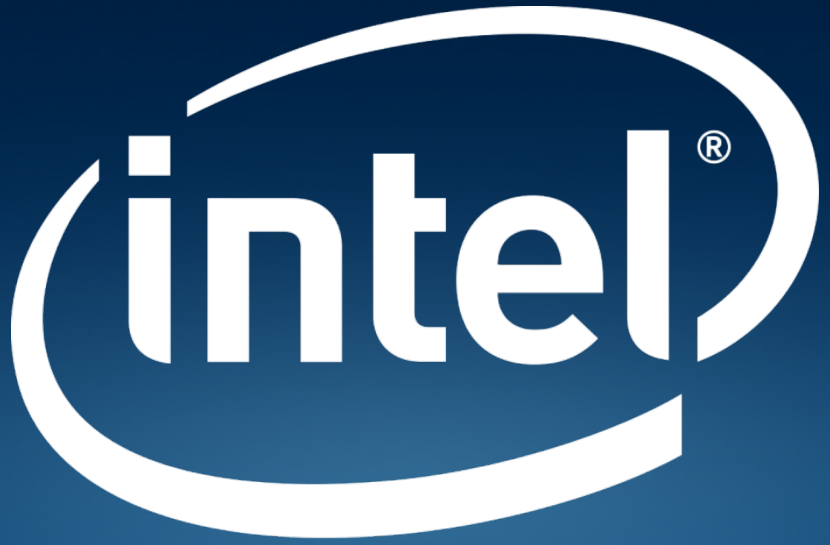
Testbed Architecture



Testbed Architecture (Cont.)

Server

- 10x Generic Lustre servers with two slightly different configurations
  - Each System comprises of:
    - 2x Intel® Xeon E5-2697v3 (Haswell) CPU's
    - 1x Intel® Omni-Path x16 HFI
    - 128GB DDR4 2133MHz Memory
    - Eight of the nodes contain - 4x Intel P3600 2.0TB 2.5" (U.2) NVMe devices, while the other two have 4x Intel® P3700 800GB 2.5" (U.2) NVMe devices
    - One node equipped with 2x Intel® S3700 400GB's for MGT
- 16x 2S Intel® Xeon E5v4 (Broadwell) Compute nodes
  - 1x Intel® HPC Orchestrator (Beta 2) Headnode
  - Hardware Components:
    - 2x Intel® Xeon E5-2697v4 (Broadwell) CPU's
    - 1x Intel® Omni-Path x16 HFI
    - 128GB DDR4 2400MHz Memory
    - Local boot SSD
- 100Gbps Intel® Omni-Path Fabric
  - None-blocking fabric with single switch design.
  - Server side optimisations: "options hfi1 sge_copy_mode=2 krcvqs=4 wss_threshold=70"
    - Improve generic RDMA performance on Lustre server side, generally you can be more aggressive with krcvqs on the server side



experience
what's inside™