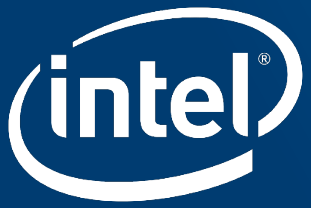


ISC'17: Lustre BoF

Trish Damkroger, VP Technical Computing
Adam Roe, HPC Solutions Architect

Legal Information

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps

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 http://www.intel.com/performance.

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.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

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

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

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.

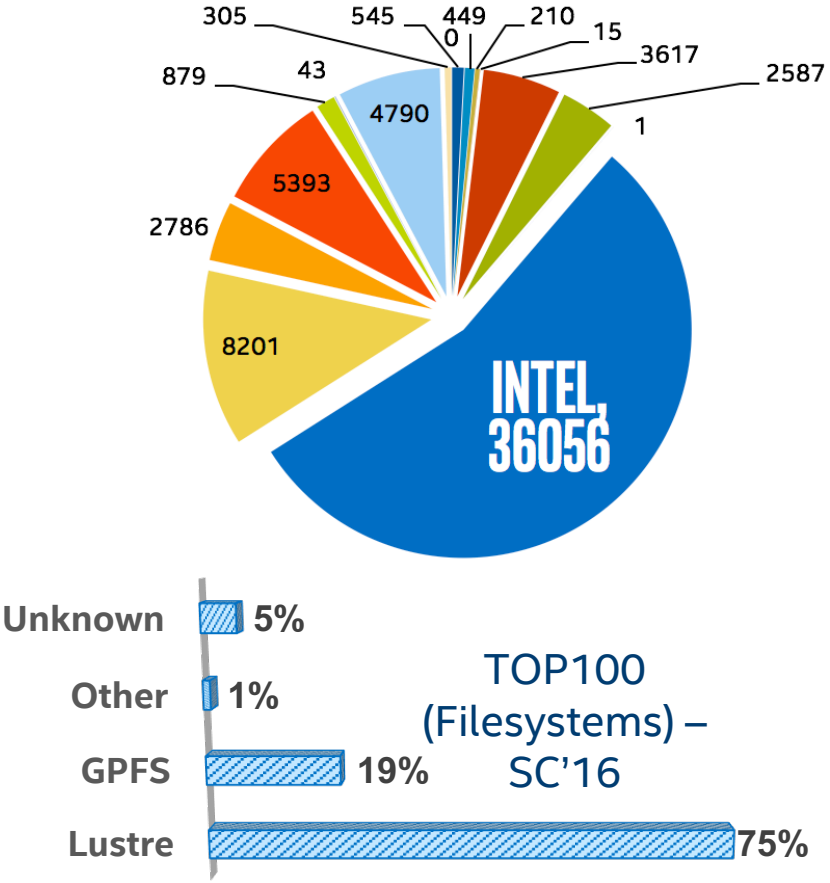
This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

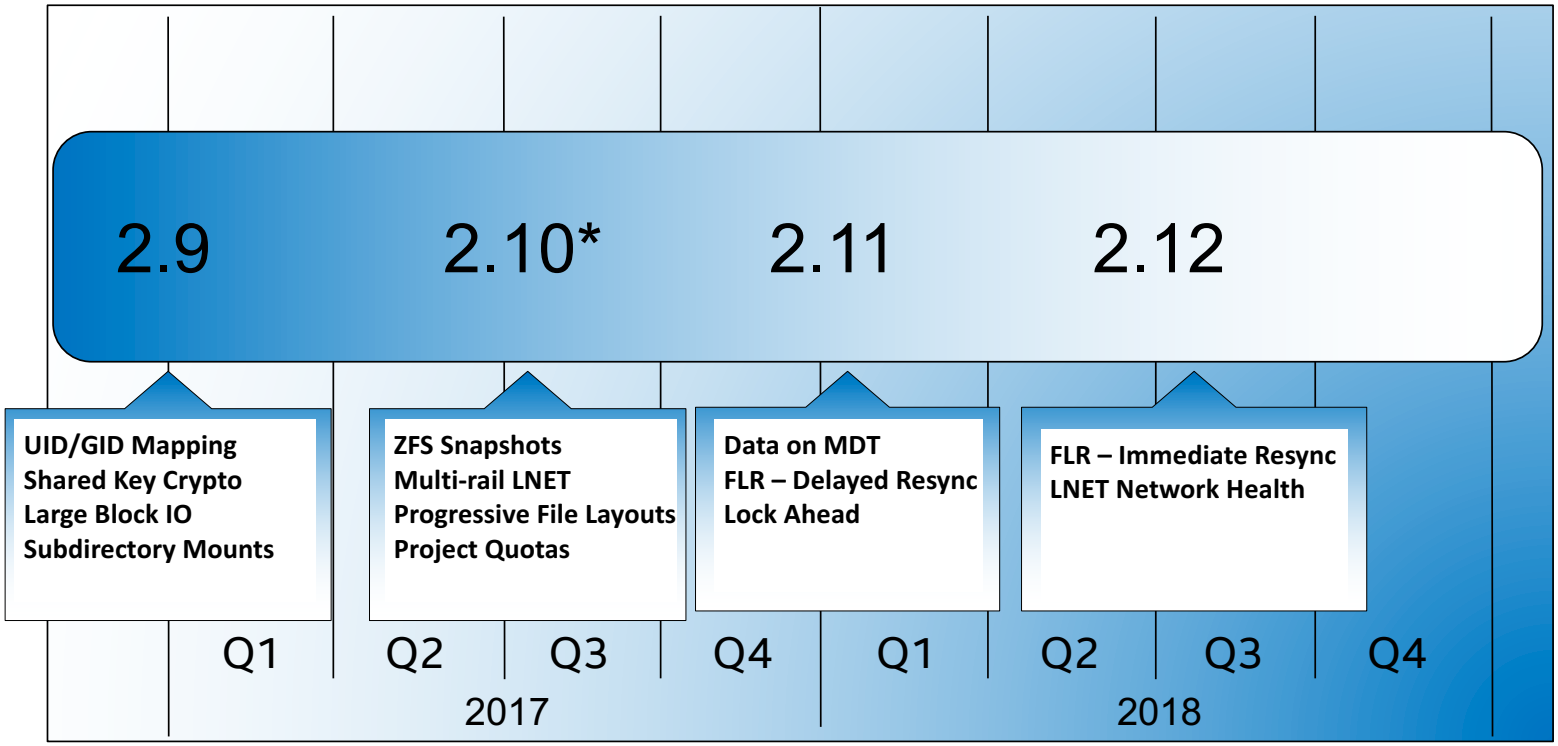
Lustre Today

LUSTRE* IS AN OBJECT-BASED, OPEN SOURCE (GPLV2), DISTRIBUTED, PARALLEL, CLUSTERED FILE SYSTEM

- Designed for maximum performance (2TB/s in production) and scalable to Exabyte's
- Number 1 HPC Filesystem (TOP500 List)
  - Performance, Scale & Adoption
- Open Source (GPLv2) & Strong Developer community
- L3 Technical Support from Intel®



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

Lustre 2.10

Targeted GA June 2017

Will support RHEL 7.3 servers/clients and SLES12 SP2 clients

Interop/upgrades from Lustre 2.9 servers/clients

Will be designated an LTS Release and have freely available maintenance releases

- Lustre 2.10.1 targeted for Q3 release

http://wiki.lustre.org/Release_2.10.0

Lustre 2.10.x – Additional Content

Confirmed in 2.10.0

- ZFS Metadata Improvements (LU-7895)
- Single thread performance improvements (LU-8964)
- OPA Performance improvements (LU-8943)
- Pacemaker scripts (LU-8457/8458)
- Upgrade possible from EE 3.x Lustre releases

Coming in 2.10.x maintenance release

- Patchless servers (LU-20)
- Support for 4.9 kernel Lustre clients (LU-9183)
- SLES12 SP2 server support
- Ubuntu 16.04 LTS Lustre client support
- MOFED 4.x support

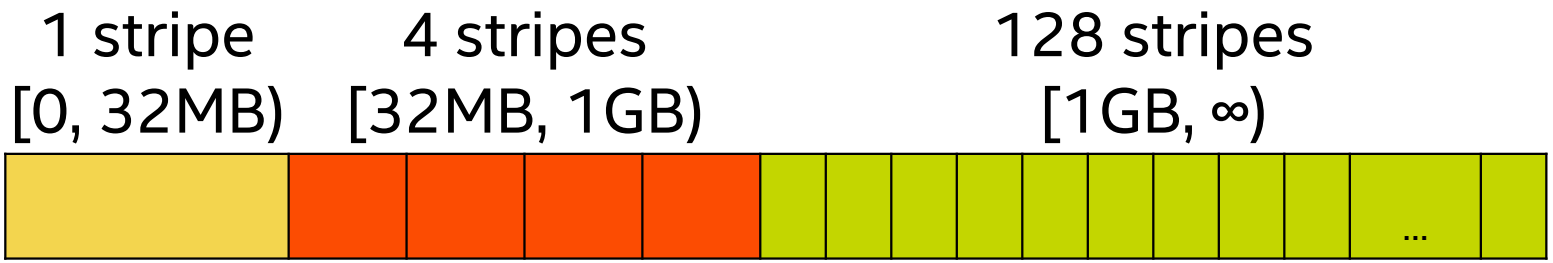
Lustre 2.10 – Progressive File Layouts

Progressive File Layout (PFL) simplifies usage for users and admins

- Optimize performance for diverse users/applications
- One PFL layout could be used for all files
- Low stat overhead for small files
- High IO bandwidth for large files

Collaboration between Intel and ORNL

Example progressive file layout with 3 components



Lustre 2.10 – Multi-Rail LNet

Allow LNet across multiple network interfaces

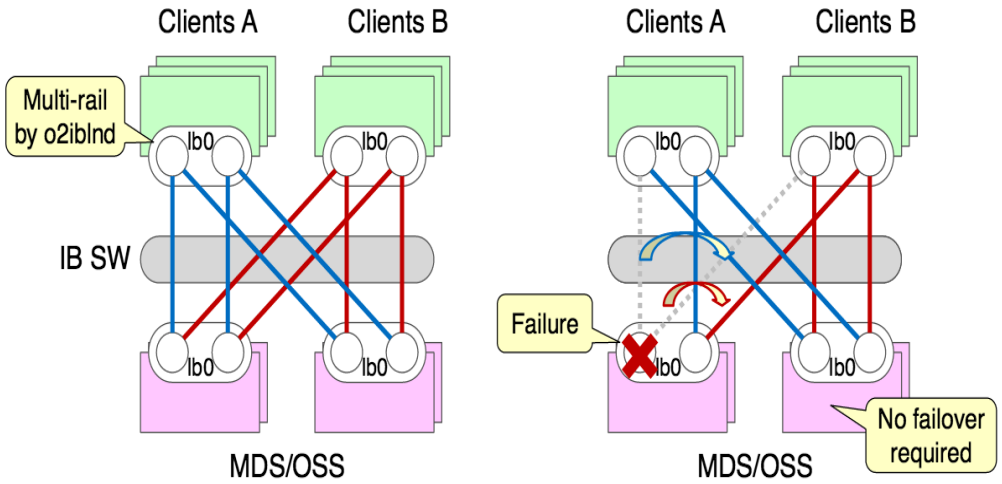
- Supports all LNet networks – LNet layer instead of LND layer
- Allows concurrent use of different LNDs (e.g. both TCP & IB at once)

Scales performance significantly

Improves reliability

- Active-active network links between peers

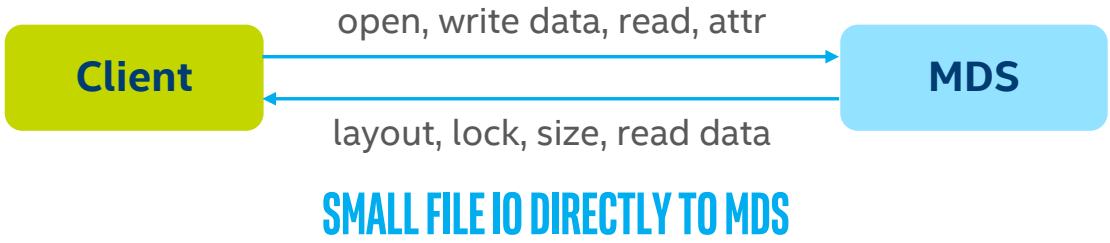
Collaboration between Intel and HPE/SGI



Improved Small File Performance (2.11)

Data-on-MDT optimizes small file IO

- Avoid OST overhead (data, lock RPCs)
- High-IOPS MDTs (mirrored SSD vs. RAID-6 HDD)
- Avoid contention with streaming IO to OSTs
- Prefetch file data with metadata
- Size on MDT for files
- Manage MDT usage by quota



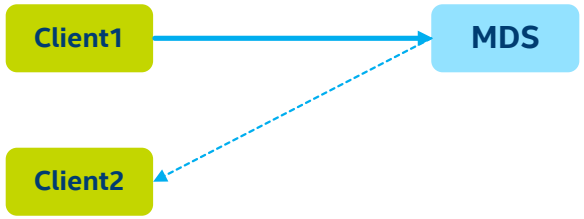
Complementary with DNE 2 striped directories

- Scale small file IOPS with multiple MDTs

Feature Optimisation: Data-on MDT

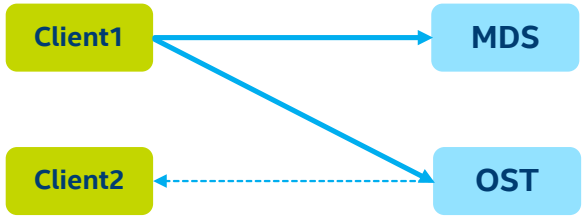
GLIMPSE-AHEAD

DoM File



1 RPC (2 with GLIMPSE)

Traditional File



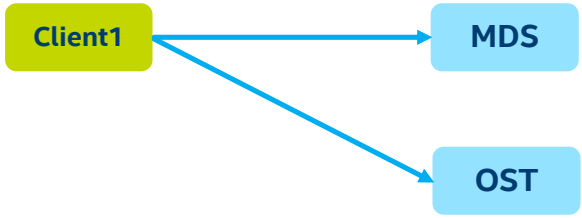
2 RPCs (3 with GLIMPSE)

LOCK ON OPEN

DoM File

1 RPC

Traditional File



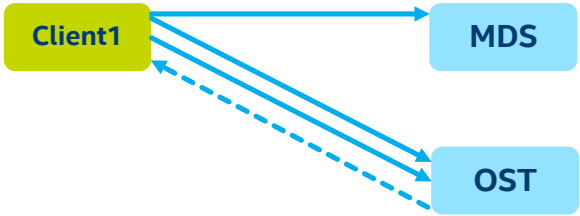
2 RPCs

READ ON OPEN

DoM File

1 RPC + BULK if size >128k

Traditional File



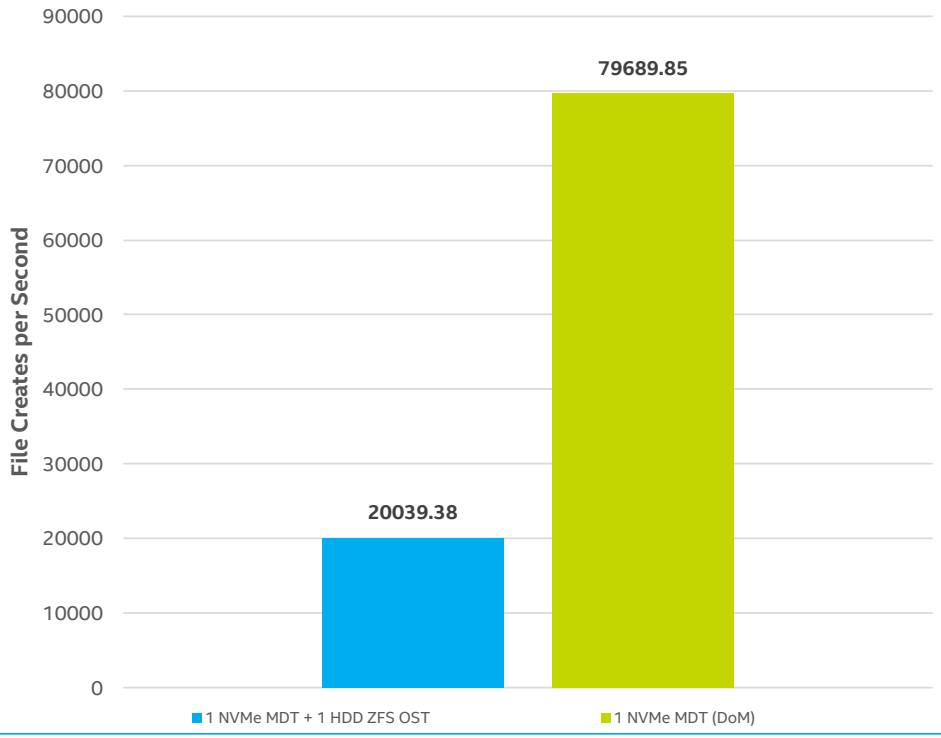
3 RPCs + BULK

Feature Optimisation: Data-on MDT (Cont.)

SMALL FILE CREATES DIRECTLY ON THE LUSTRE MDT

File Create (4KiB): HDD vs. NVMe OST vs. DoM

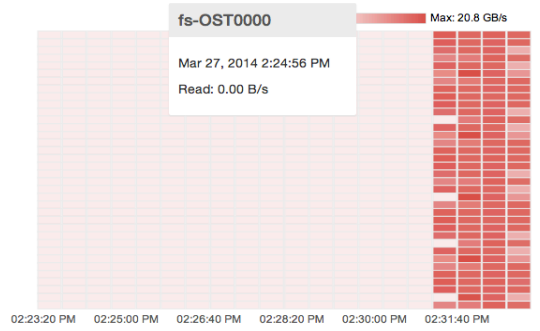
- Architecturally very different, both from a hardware and software perspective
  - Space used and load on the MDT is considerably higher
- 3x Speed up when using DoM for small files on an NVMe Lustre MDT (~4-32KiB tested)
- 1.9x of that is just from efficiency improvements in the network, i.e. less/better use of RPC's



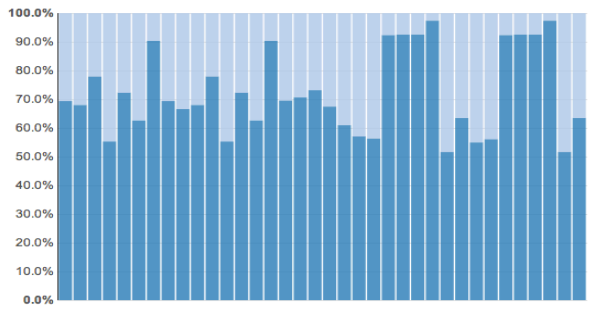
IML: Community-based Lustre Manager

Management and Monitoring Tool

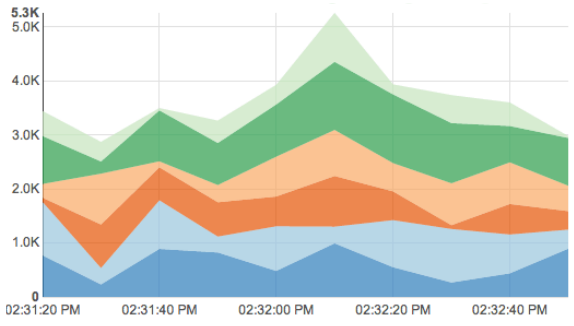
Read/Write Heat Map



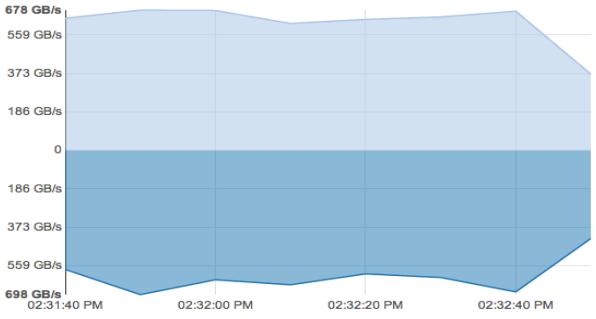
OST Capacity



Metadata Operations



Read/Write Bandwidth

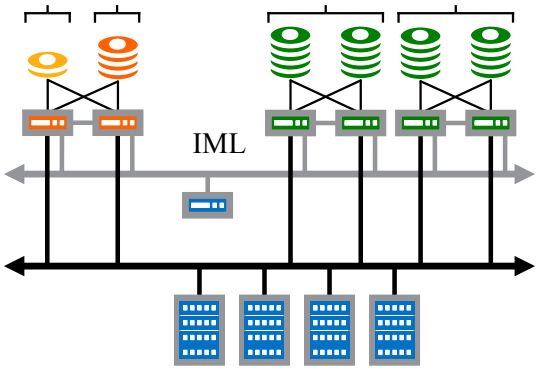


- Intuitive, browser-based administration
- Lustre installation and configuration
- Real-time system monitoring
- Extensible through open, documented APIs
- Now available under MIT license at https://github.com/intel-hpdd/

The Future is both Evolutionary & Revolutionary

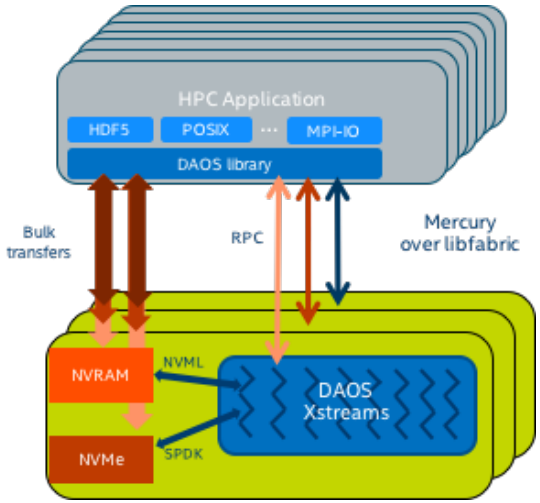
Continued Lustre Evolution

- Scaling for performance & capacity
- Stable functional improvements



Exascale DAOS Revolution

- Leverage NVRAM and NVMe storage technologies
- Userspace I/O architecture for lowest latency
- Direct integration into userspace libraries/apps



Summary & Resources

Mission

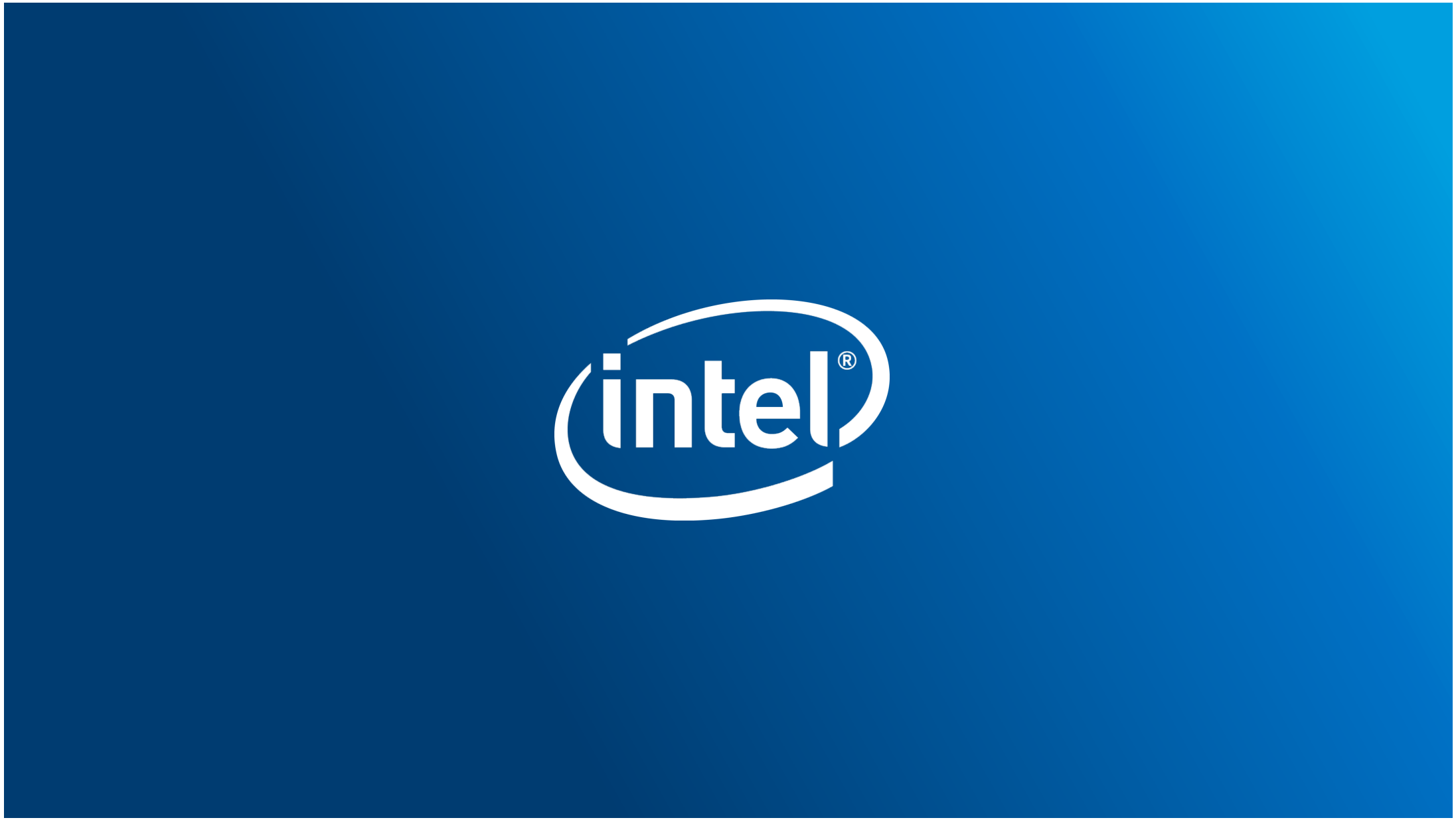
- Develop a rich portfolio of high performance storage products to solve the worlds most challenging data storage and IO problems

Scope

- Lustre is the future of scalable POSIX-compliant storage
- DAOS is the future of scale-out object storage
- Next-generation storage R&D Projects combining both Lustre & DAOS

Resources

- Lustre*
  - GPLv2 License
  - https://git.hpdd.intel.com/fs/lustre-release.git
- IML & HAL
  - MIT License
  - https://github.com/intel-hpdd
- DAOS
  - Apache 2.0 License
  - https://github.com/daos-stack/daos
- Support
  - https://jira.hpdd.intel.com


intel®