HPC STORAGE FUTURES
A 5-YEAR OUTLOOK

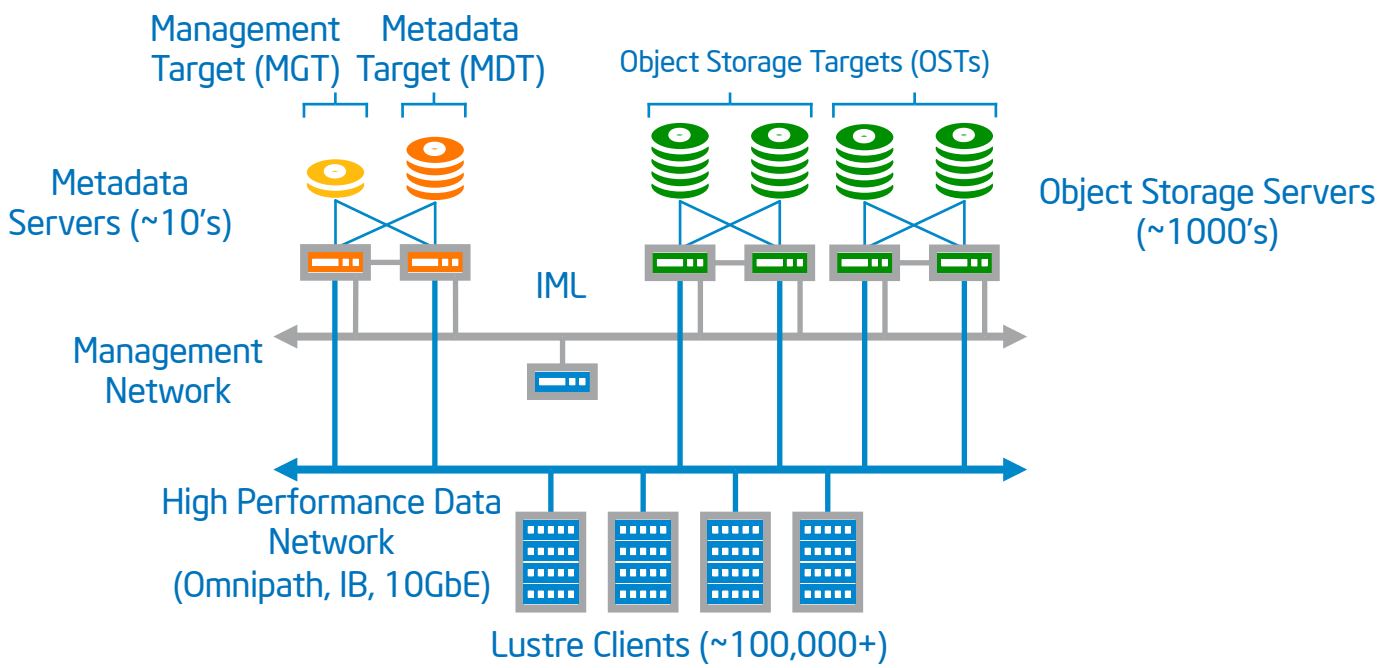
Brent Gorda
GM, HPC Storage
Intel® HPC Platform Group
March 2016

THE NEW CENTER OF POSSIBILITY

2nd International Workshop on The Lustre Ecosystem: Enhancing Lustre Support for Diverse Workloads

Baltimore, Maryland

HPC I/O TODAY – THE PARALLEL FS

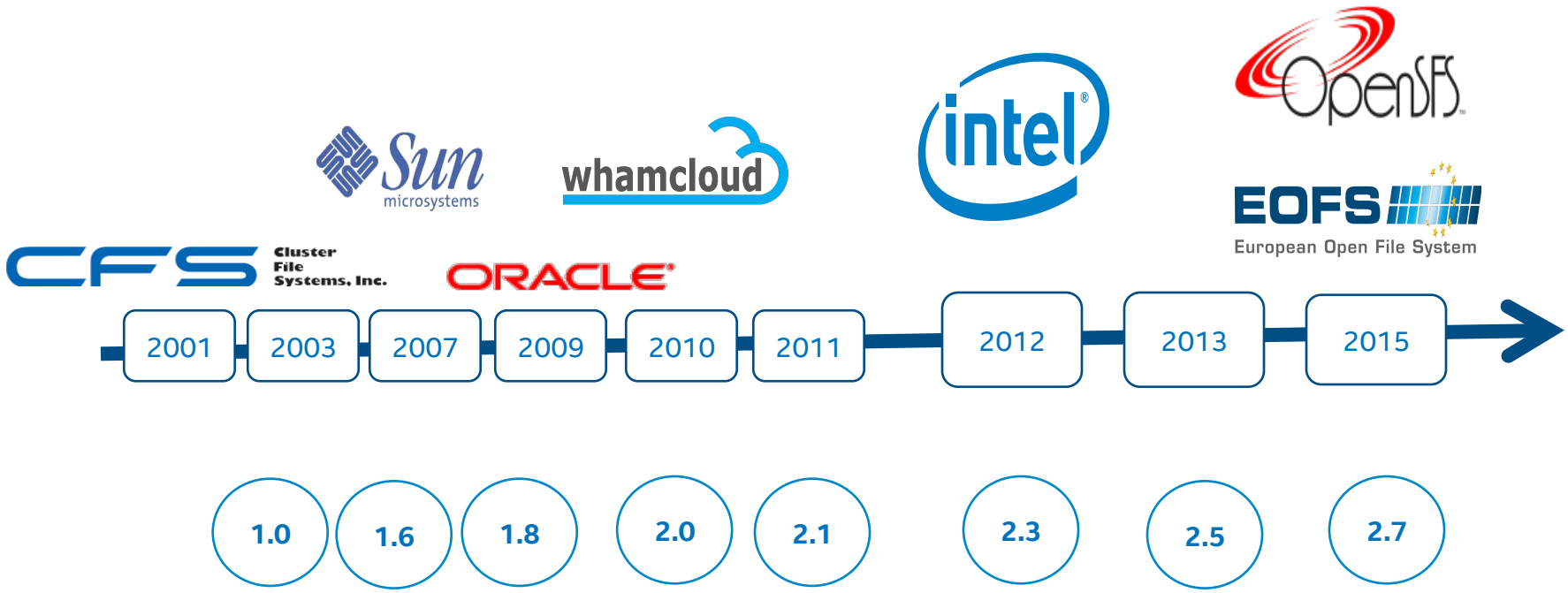


Lustre* is the leading scale-out parallel posix/file solution for High Performance Computing.

A healthy Lustre* ecosystem

A BRIEF HISTORY OF LUSTRE*

A JOURNEY INTO INNOVATION AND FREEDOM



LUSTRE* COMMUNITY

Two community efforts

- European Open File Systems
- OpenSFS

A shift toward building the open source community:

- OpenSFS & Community Releases:: OpenSFS Lustre* Working Group (Sarp/Peter)
- Community Release of Lustre* – 2.8 in final stages of release
- Lustre* User Group growth (India, China, Japan, Australia)
- Intel® a major contributor/donor to the community efforts

Advanced Lustre* Research

New initiatives starting at Intel® Parallel Computing Centers

Lawrence Berkeley National Laboratory

Johannes Gutenberg University

University of California Santa Cruz

GSI Helmholtz Centre for Heavy Ion Research

University of Hamburg

Expanding Collaborations

*Other brands and names are the property of their respective owners

INTEL® SOLUTIONS FOR LUSTRE* ROADMAP

Q4 '15 | Q1 '16 | Q2 '16 | Q3 '16 | Q4 '16

Enterprise

HPC, commercial technical computing, and analytics

Enterprise Edition 2.4.1

EE Lustre 2.5.x

RHEL 6.7 server/client
RHEL 7.2 client
SLES 11sp4 server/client
Robinhood 2.5.5
ZFS 0.6.4.2

Enterprise Edition 3.0

EE Lustre 2.7.x

MDT directory striping (preview)
Client metadata RPC scaling
OpenZFS performance
OpenZFS Snapshots
SELinux on client + Kerberos
Intel Omni-Path (OPA) support
RHEL 7 server/client, SLES 12

Enterprise Edition 3.1

EE Lustre 2.7.x

Expanded OpenZFS support, Intel Manager for Lustre
Enhanced UI, Intel Manager for Lustre

Cloud

HPC, commercial technical computing and analytics delivered on cloud services

Cloud Edition 1.2

FE Lustre 2.7.1

IPSec over-the-wire encryption
Disaster recovery
Supportability enhancements

Cloud Edition 2.0

FE Lustre 2.7.x

Disaster Recovery Automation
Lustre patches and enhancements
Support for the latest instance types

Foundation

Community Lustre software coupled with Intel support

Foundation Edition 2.7.1

Lustre 2.7.1

RHEL 7.1 server / client
ZFS 0.6.4.2 + performance
Intel Omni-Path (OPA) Gen 1 support

Foundation Edition 2.8

Lustre 2.8.x

DNE 2 – MDT directory striping
Client MD RPC scaling
LFSCK performance tuning
SELinux on client + Kerberos

Foundation Edition 2.9

Lustre 2.9.x

UID/GID Mapping
Shared key network encryption
<TBA>

Key: Shipping | Development | Planning

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

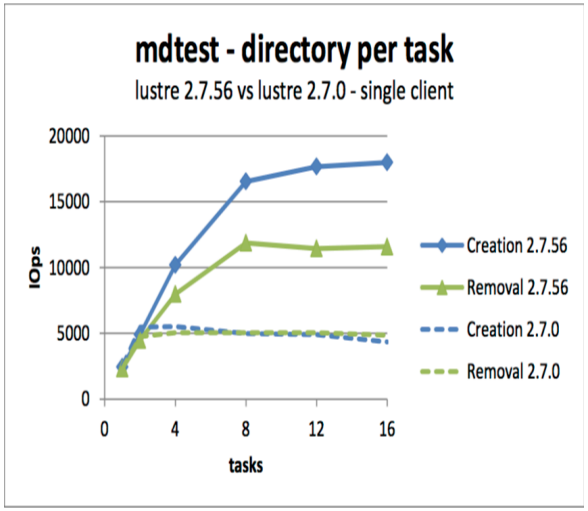
EVOLVING CAPABILITY AND SCALABILITY

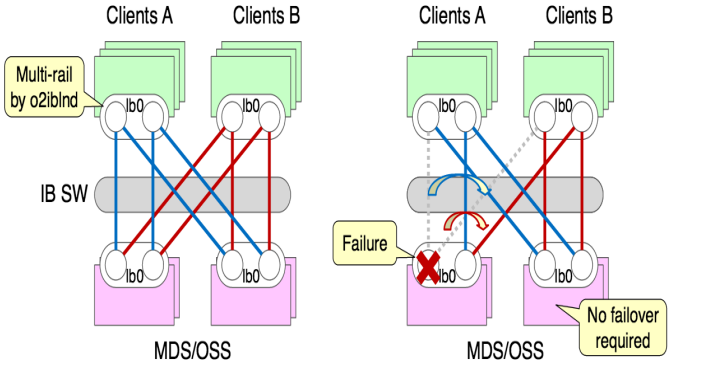
Improved performance for back-end storage

- Significant investment in ZFS performance
- Improved robustness for very large servers
- Quality of Service for users, jobs, nodes

Improved networking capabilities

- Multi-Rail support for all network types
- New Intel® Omni-Path network support
- Support for EDR and FDR InfiniBand™
- Crypto for RDMA networks like OPA/IB

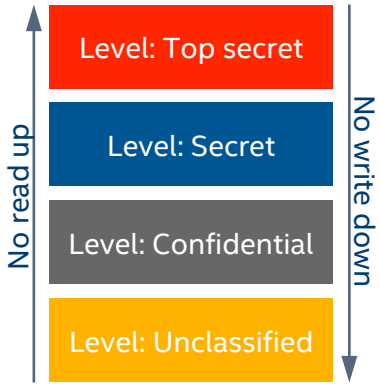




DATA SECURITY AND EASE OF USE CRITICAL IN ALL ENVIRONMENTS

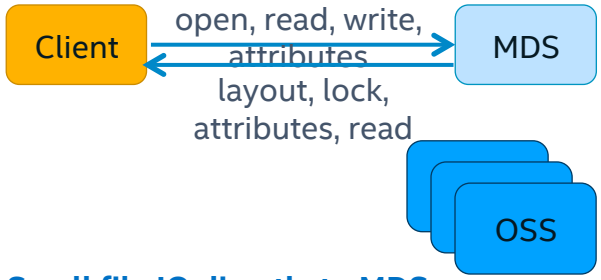
Security enhancements

- Multi-level access control
- Data isolation via filesystem containers
- Client node identification and authorization
- Strong crypto for network communication



New file layout options for users & apps

- Progressive file layouts simplify usage
- File level replication enables multiple areas
- Data-on-MDT improves small file IO
- HSM policy engine and fast data movers



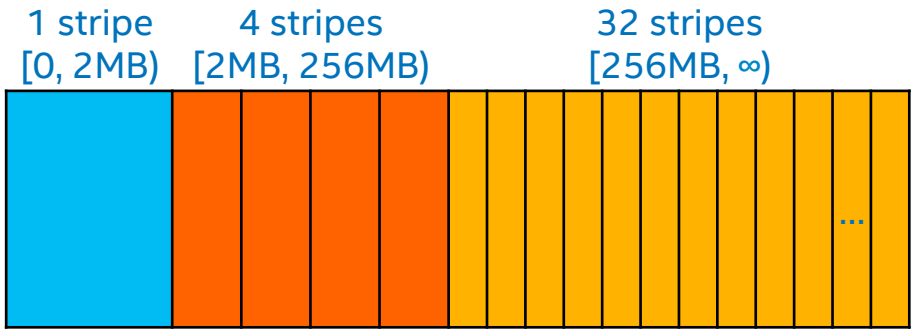
Small file IO directly to MDS

COMPLEX FILE LAYOUTS DRIVING INNOVATION IN LUSTRE* USAGE

Progressive file layouts simplify usage

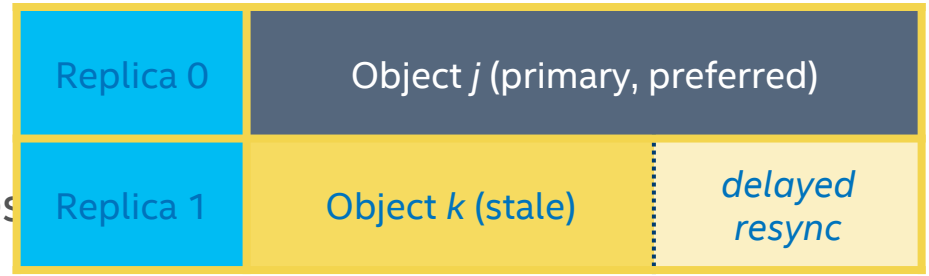
- Optimize perf for diverse users
- Low user and administrative burden
- Multiple storage classes in a single file

Example progressive file layout with 3 components



File level replication = significant value

- Can be selected on a per-file basis
- HA for server/network failure
- Robustness vs data loss/corruption
- Increased read speed for common files
- Migrating files between storage classes

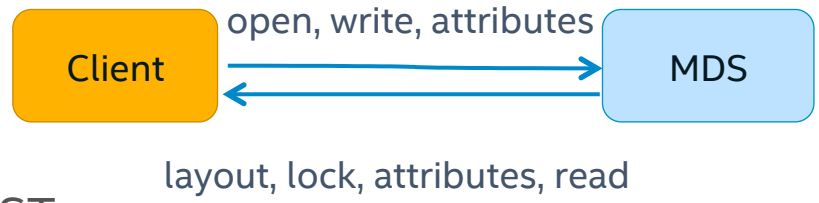


Overlapping (mirror) layout

LEVERAGE LOW-LATENCY STORAGE AT THE SERVER AND CLIENT

Data-on-MDT optimizes small file IO

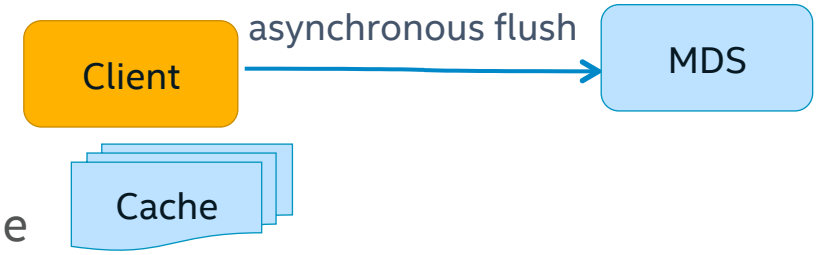
- Reduced RPC overhead for data access
- Leverage high-IOPS storage on MDTs
- Avoid contention with streaming IO to OSTs
- Prefetch data and metadata concurrently
- Optimize data placement decisions



Small file IO directly to MDS

Client-side IO and metadata optimizations

- Horizontal namespace scaling
- Single- and multi-threaded IO efficiency
- Client side persistent data read/write cache
- Client metadata operations writebackcache



Client Metadata Operations Cache

IMPROVED PERFORMANCE FOR BACK-END ZFS STORAGE

Lustre+ZFS lowers the cost of storage

- Integrated data checksums
- Online data integrity checking and repair
- Snapshots, data compression
- Hybrid HDD/SSD storage

CORAL driving significant ZFS improvement

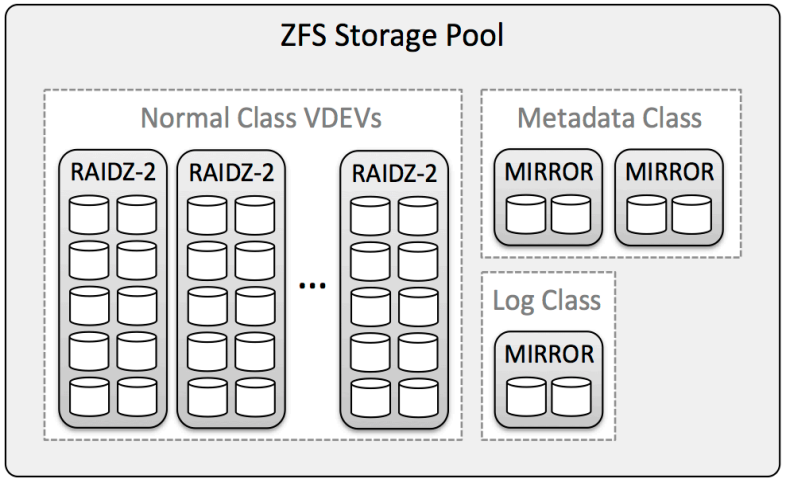
- Optimizations for large streaming IO
- Segregation of internal data+metadata
- Optimized and scalable RAID rebuild
- Improved large-scale management

Optimizations for all-flash storage

ZFS Declustered Parity

Data	P	Spare

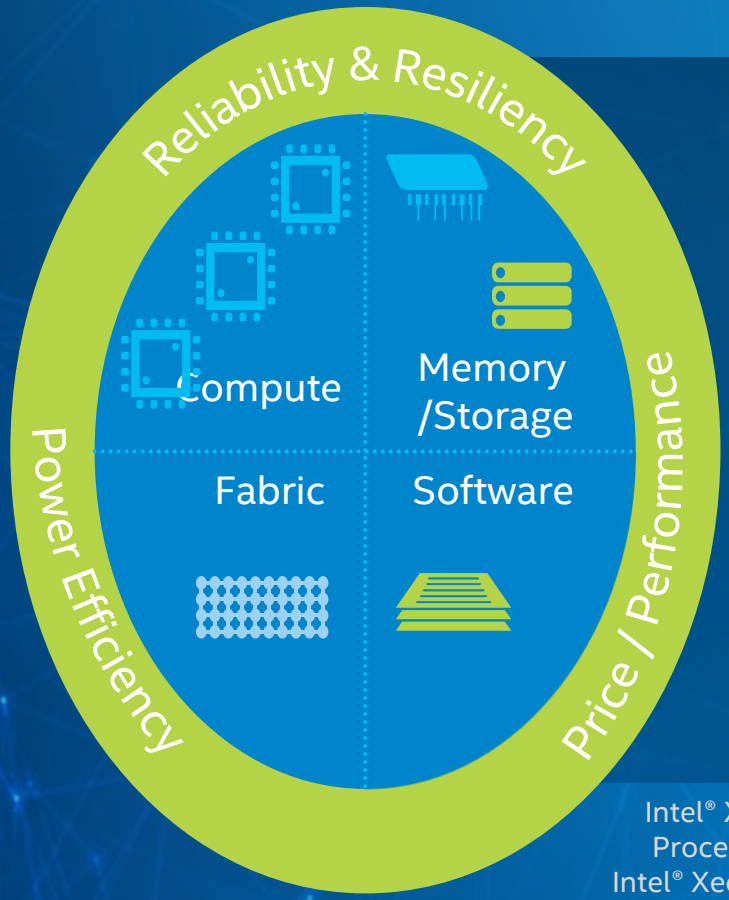
4	3	10	7	2	11	9	1	0	6	5	8
9	8	10	3	6	5	4	7	0	1	2	11
5	7	0	11	8	2	6	4	3	9	10	1
8	7	3	10	4	1	11	9	0	6	5	2
9	3	4	11	0	6	8	7	10	1	2	5
11	1	4	6	3	2	7	8	9	5	10	0
0	2	8	5	1	9	10	7	4	11	6	3



A healthy HPC ecosystem

Intel's HPC Scalable System Framework

A design foundation enabling wide range of highly workload-optimized solutions



Small clusters through Supercomputers

Compute and Data-Centric Computing

Standards-Based Programmability

Intel® Xeon® Processors Intel® Xeon Phi™ Coprocessors Intel® Xeon Phi™ Processors	Intel® True Scale Fabric Intel® Omni-Path Architecture Intel® Ethernet	Intel® SSDs Intel® Lustre-based Solutions Intel® Silicon Photonics Technology	Intel® Software Tools

Intel® Scalable System Framework

Intel's HPC Scalable System Framework

Public statements of adoption since April '15

Intel® SSF Design Guidance

Simplifies...

System Design and Build

Software Development

Procurement, Deployment, Management

Coming Q1'16

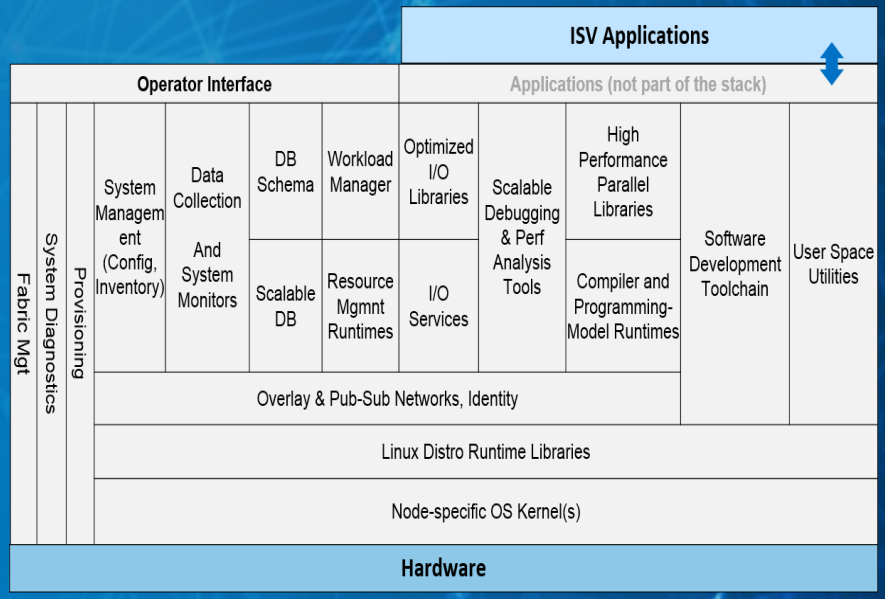
Reference Architectures
designs for compatibility

Reference Designs
system build recipes

Validation Tools
streamlined testing

Community-driven building blocks hpc

Intel®: Founding Member of OpenHPC



- Modular & comprehensive
- Established integration conventions
- Pre-packaged and pre-tested
- Flexible development environment
- Centralized package repository
- Bring your own license model

Initial Release Nov. 12, 2015

Intel® Supported Versions

An element of the Intel® Scalable System Framework

Coming in 2016

Intel's Commitment to the HPC Community

Intel® Modern Code Developer Community

An Online Community
to Reach **400,000** Developers and Partners with Tools, Trainings and Support

Hands on Training
for 10,000 Developers and Partners

Remote Access
to Intel® Xeon® Processor and Intel® Xeon Phi™ Coprocessor-based Clusters

Intel® Supporting the Software Community

Leading Contributor
to Multiple **Open Source Projects**, Including Linux*, Luster*, OpenHPC, Embree

Working with ISVs and the Community to Help Modernize Codes Across the Ecosystem

Intel® Parallel Compute Centers

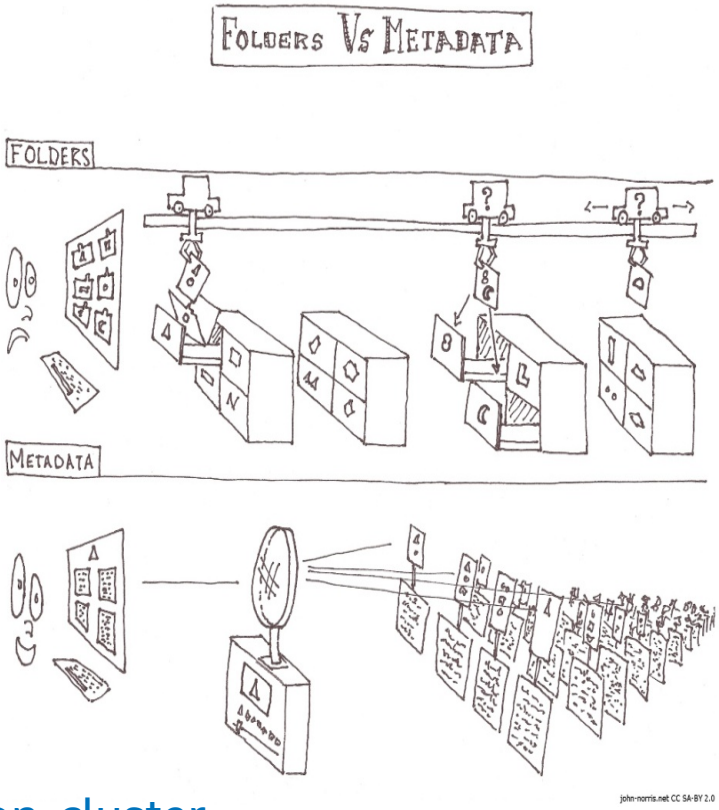
Focused on Modernizing Community Code
50+ Intel® PCC Modernizing More Than 90 HPC Computing Codes Across 16 Domains

A Global Effort
in 15 Countries and Four Continents

Join the Online Community Today!

Changing demands and opportunities

EMERGING TRENDS



Increased computational power…

- Huge expansion of simulation data volume & metadata complexity
- Complex to manage and analyze

…achieved through parallelism

- 100,000s nodes with 10s millions cores
- More frequent hardware & software failures

Tiered storage architectures

- High performance fabric & solid state storage on-cluster
- Lower performance, high capacity disk-based storage off-cluster

EMERGING TRENDS… CONTINUED

Small file I/O dominating workload

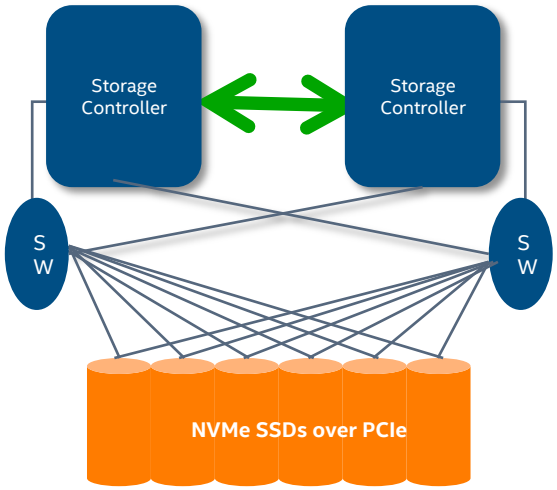
- Used to be: C/R + streaming workload
- Today stories of 90% of files < 1 MB

Lustre evolving / adapting in response

- Distributed Namespace (DNE) (v2.4, v2.8)
- Data on MDT, file replication, snapshots

Other features: ZFS, progressive file layouts,

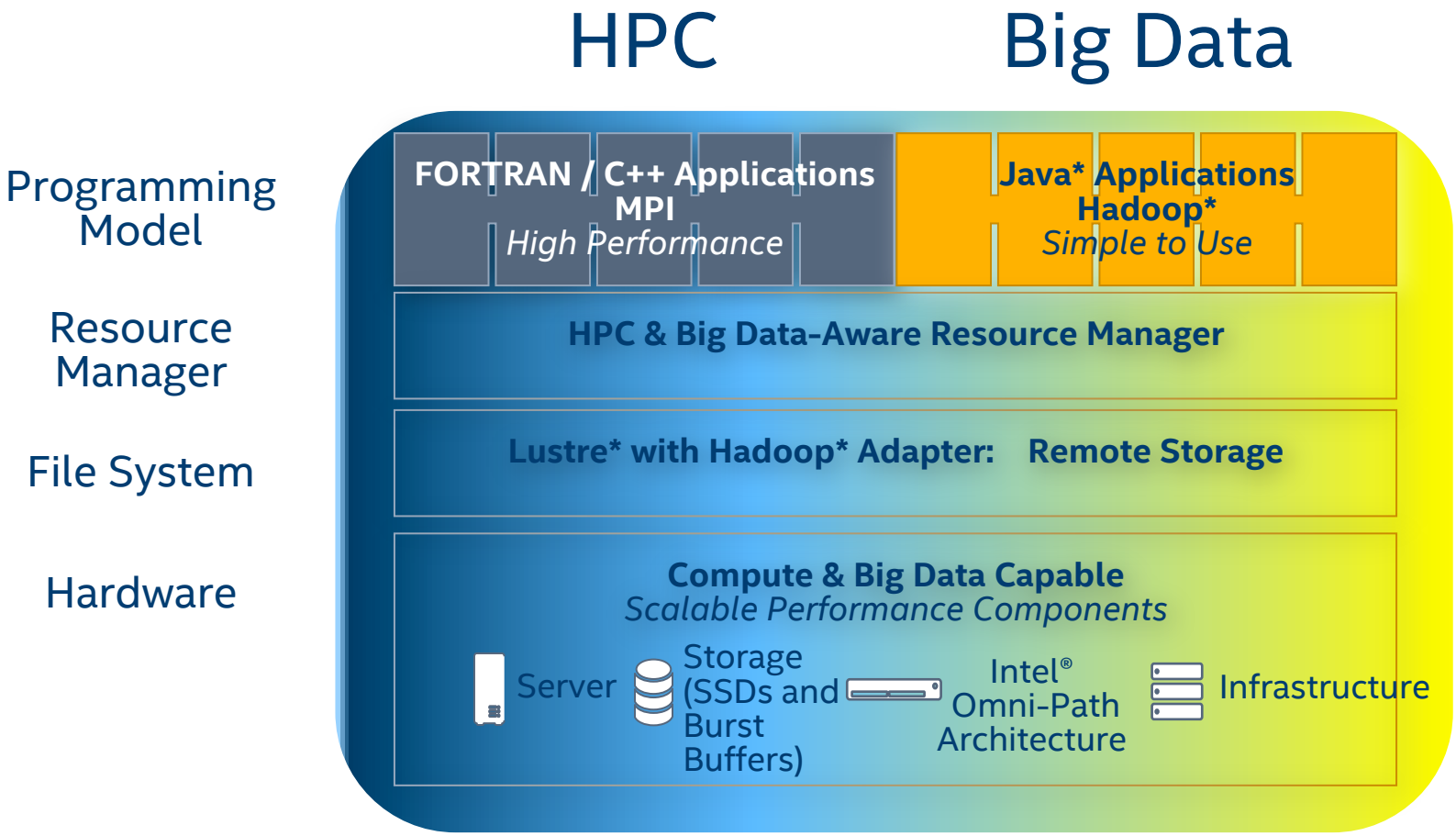
single threaded performance, etc….



There is further talk of converging architectures: HPC, Cloud, BigData

- How are we going to feed data to these systems as they grow?
- Are HPC Achilles heels going to drive architecture?
- **Some interesting storage hardware on the horizon**

Converged Architecture for HPC and Big Data

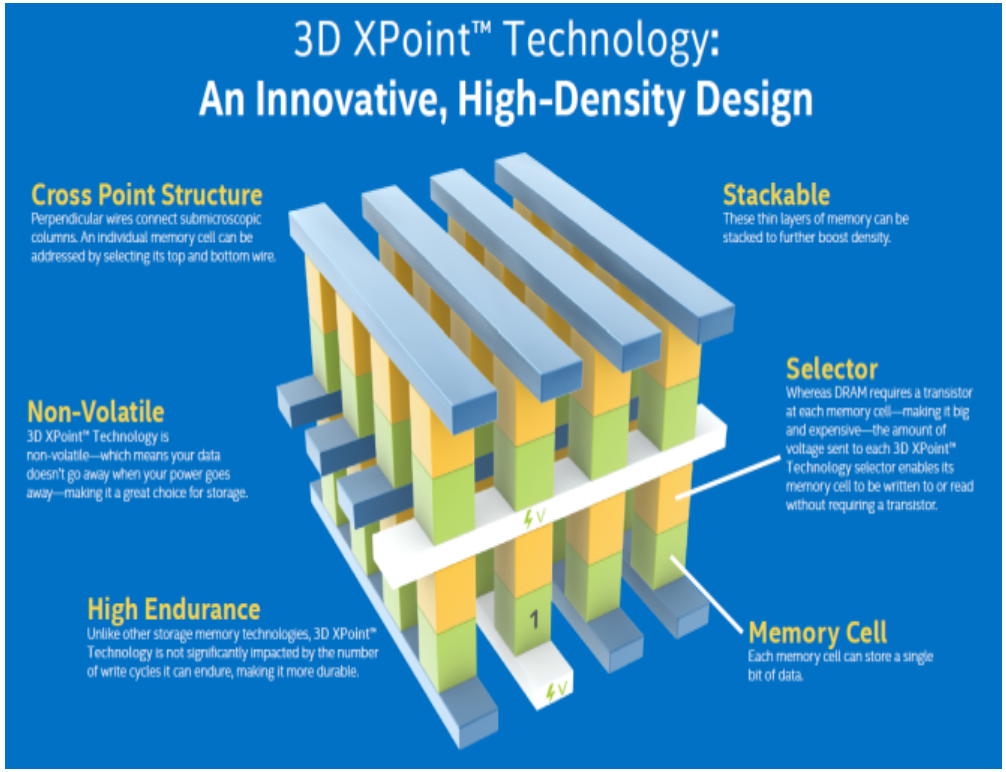


New hardware capabilities

BRIDGING THE MEMORY-STORAGE GAP

INTEL® OPTANE™ TECHNOLOGY BASED ON 3D XPOINT™

Intel® Scalable System Framework



SSD

Intel® Optane™ SSDs 5-7x Current Flagship NAND-Based SSDs (IOPS)[1]

DRAM-like performance

Intel® DIMMs Based on 3D-XPoint™

1,000x Faster than NAND[1]

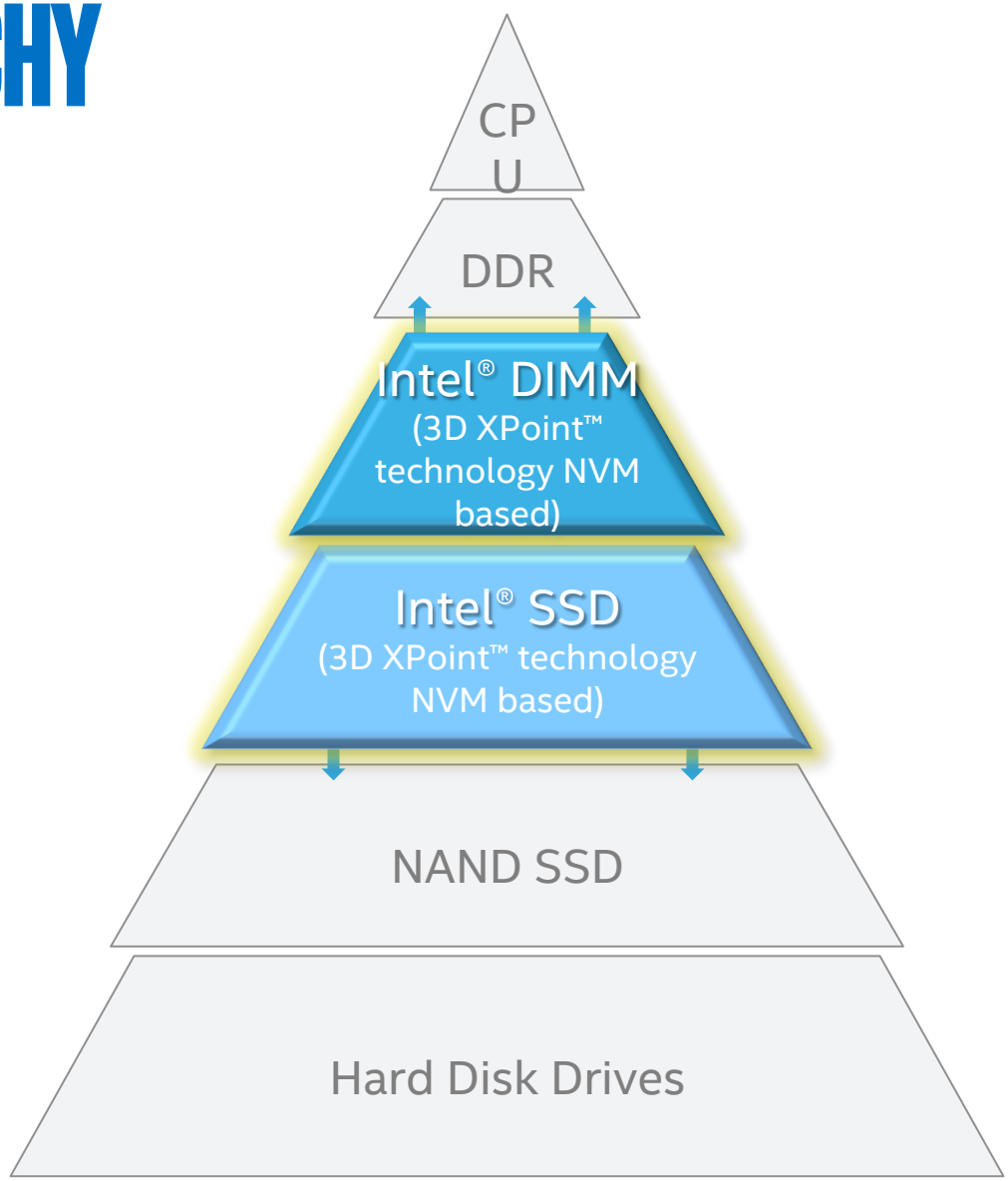
1,000x the Endurance of NAND[2]

Hard drive capacities

10x More Dense than Conventional Memory[3]

[1] Performance difference based on comparison between 3D XPoint™ Technology and other industry NAND
[2] Density difference based on comparison between 3D XPoint™ Technology and other industry DRAM
[2] Endurance difference based on comparison between 3D XPoint™ Technology and other industry NAND

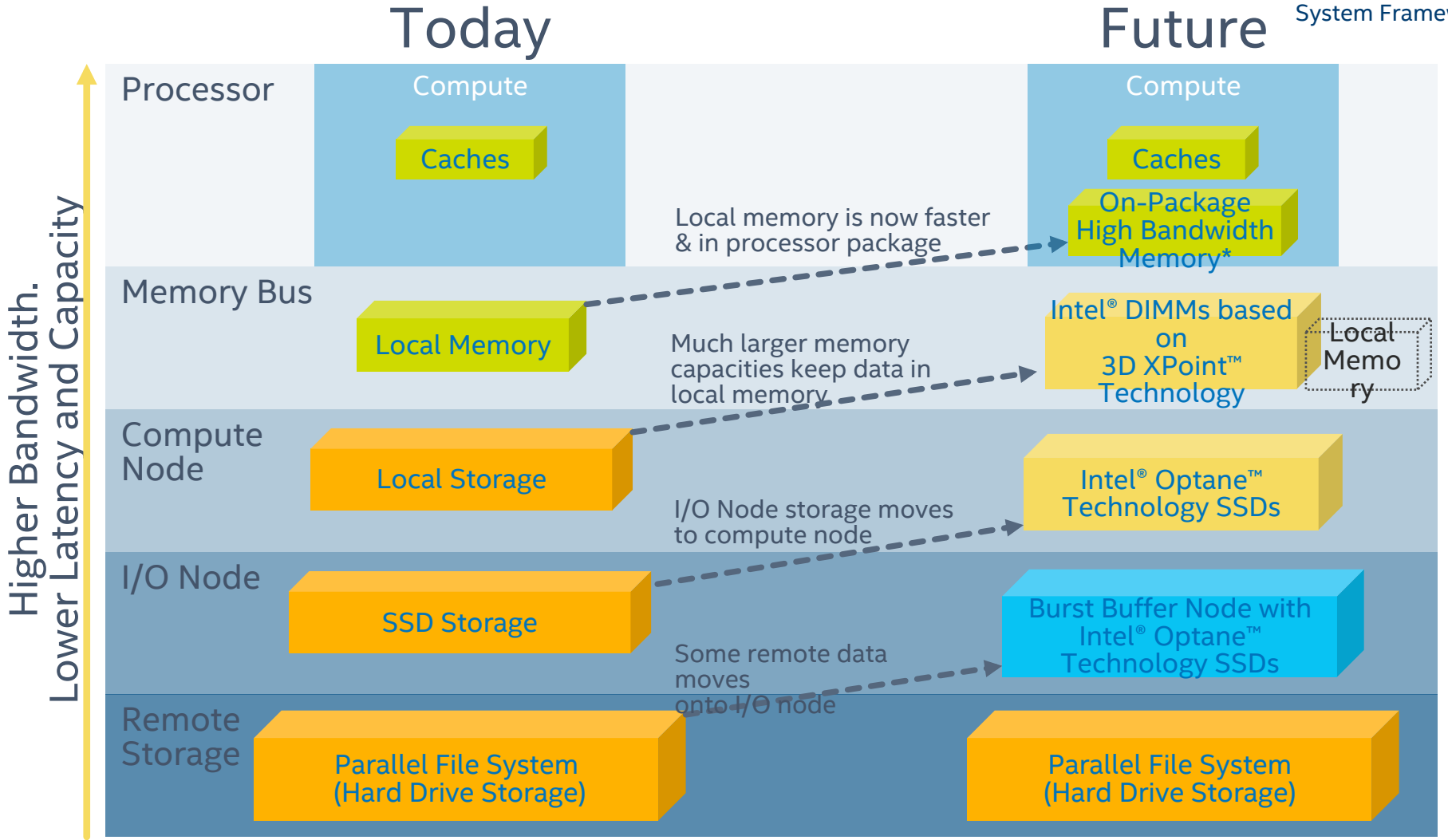
A NEW HIERARCHY



TIGHTER SYSTEM-LEVEL INTEGRATION

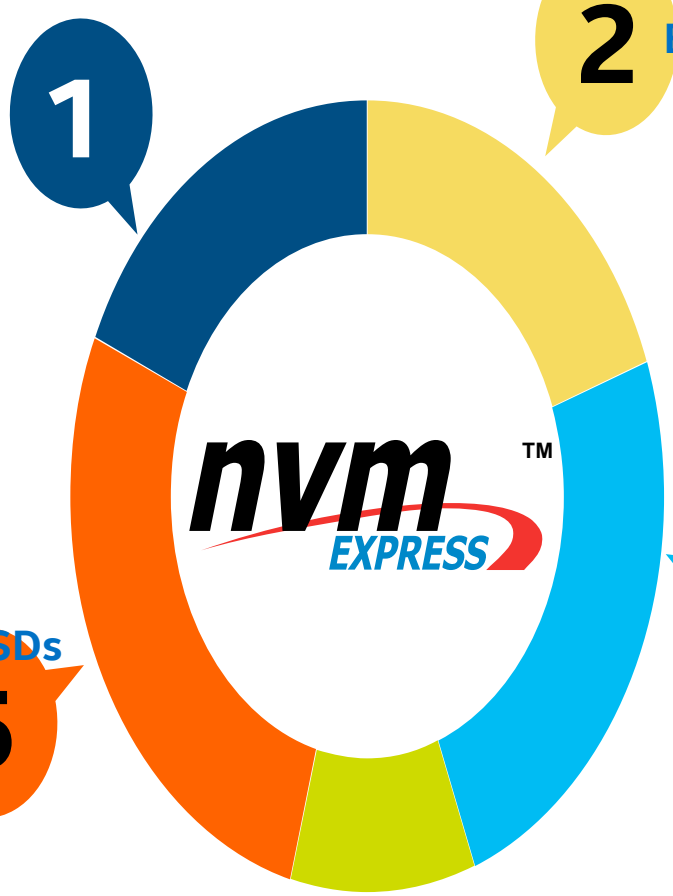
INNOVATIVE MEMORY-STORAGE HIERARCHY

Intel® Scalable System Framework



*cache, memory or hybrid mode

NVM EXPRESS™



1

What is NVMe?

NVM Express is a standardized high performance software interface for PCI Express® Solid-State Drives

2

Built for SSDs

Architected from the ground up for SSDs to be efficient, scalable, and manageable

3

Developed to be lean

Streamlined protocol with new efficient queuing mechanism to scale for multi-core CPUs, low clock cycles per IO

4

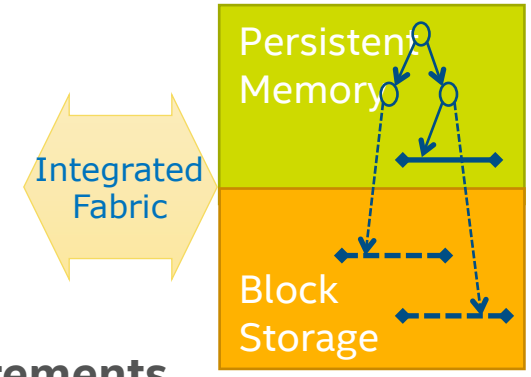
Industry standard software and drivers that work out of the box

5

Ready for next generation SSDs

New storage stack with low latency and small overhead to take full advantage of next generation NVM

DISRUPTIVE CHANGE



NVRAM + Integrated fabric

- Byte-granular storage access
- Sub-µS storage access latency
- µS network latency

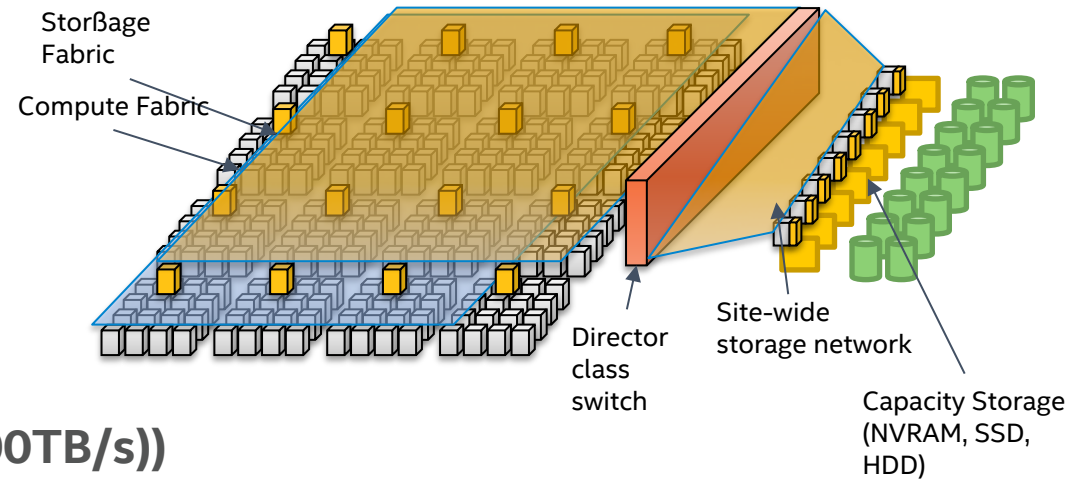
Conventional storage software

- Block granular access limits scaling
- High overhead
  - 10s µS lost to comms S/W
  - 100s µS lost to F/S & block I/O stack

I/O stack requirements

- Minimal software overhead
  - OS bypass
    - Communications
    - Latency sensitive I/O
- Fail-out resilience
- Persistent Memory storage
  - Fine-grain data & metadata
- Block storage
  - Bulk data

I/O ARCHITECTURE



Performance storage tier (O(100TB/s))

- NVRAM: ultra fine-grain I/O
  - Accessible as memory by local CNs
- NAND: additional bulk storage capacity
  - Eliminate unnecessary staging
- Fully distributed across compute cluster
  - Global I/O object address spaces
  - Jobs can burst at full storage bandwidth
  - Storage BW scales with system size
- Dedicated storage fabric
  - Applications unaffected by staging I/O and I/O from unrelated jobs

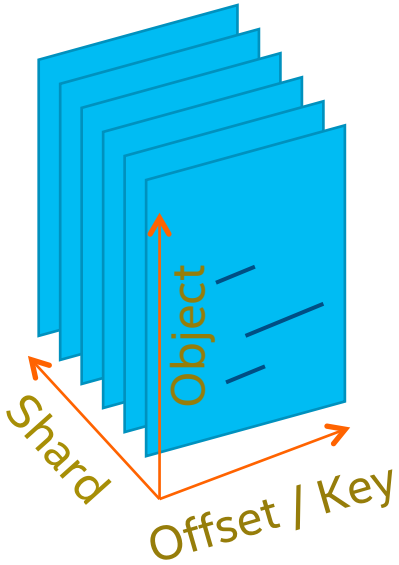
Capacity storage tier (O(TB/s))

- NVRAM
  - System metadata
  - Hot application metadata
- NAND
  - Application metadata
  - Fine-grain application data
- HDD
  - Bulk application data
  - Aggregated application data

DAOS

Distributed Application Object Storage

- Object PGAS supporting multiple higher level storage models
  - Container = set of container shards
  - Shard = set of objects
  - Object = {KV store, byte array}
- Fine-grained versioning
  - Writes eagerly accepted in arbitrary order
  - Reads sample given version snapshot
  - Abort discards uncommitted versions
- Container Metadata
  - Simple list of container shards + commit state
  - Resiliently replicated over container shards



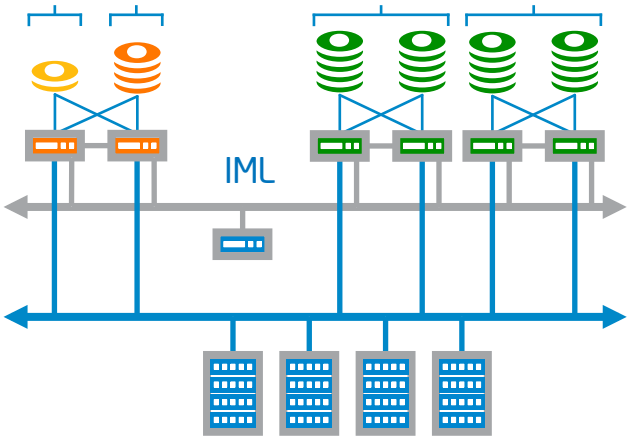
DAOS on Persistent Memory (DAOS-M)

- Deliver fabric/NVRAM latency benefit
  - End-to-end OS bypass
  - Persistent Memory server
- Connectionless networking & security
  - Peer-to-peer connectivity ~1000x conventional client/server
  - Heavyweight security checks only on container open
  - Communicator == capability

THE FUTURE IS BOTH EVOLUTIONARY & REVOLUTIONARY

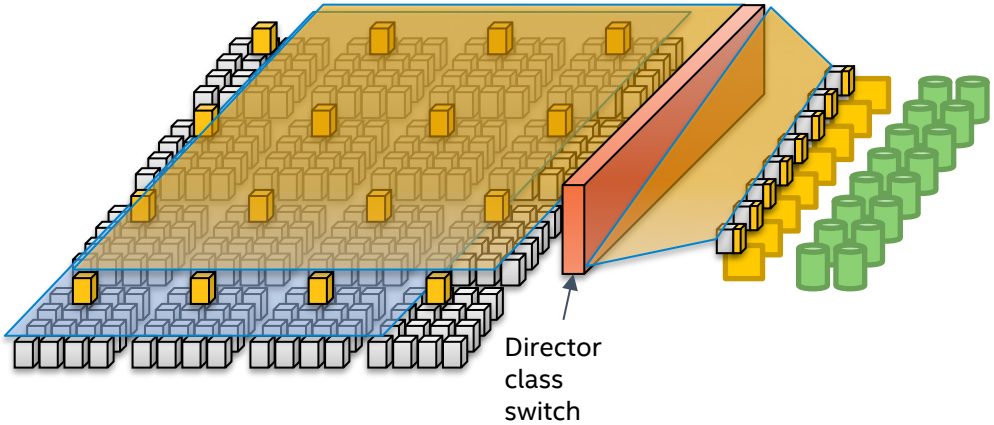
Lustre* evolving in response to:

- 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
- Raising level of abstraction



Legal Information

THANK-YOU
LUSTRE.INTEL.COM


intel®