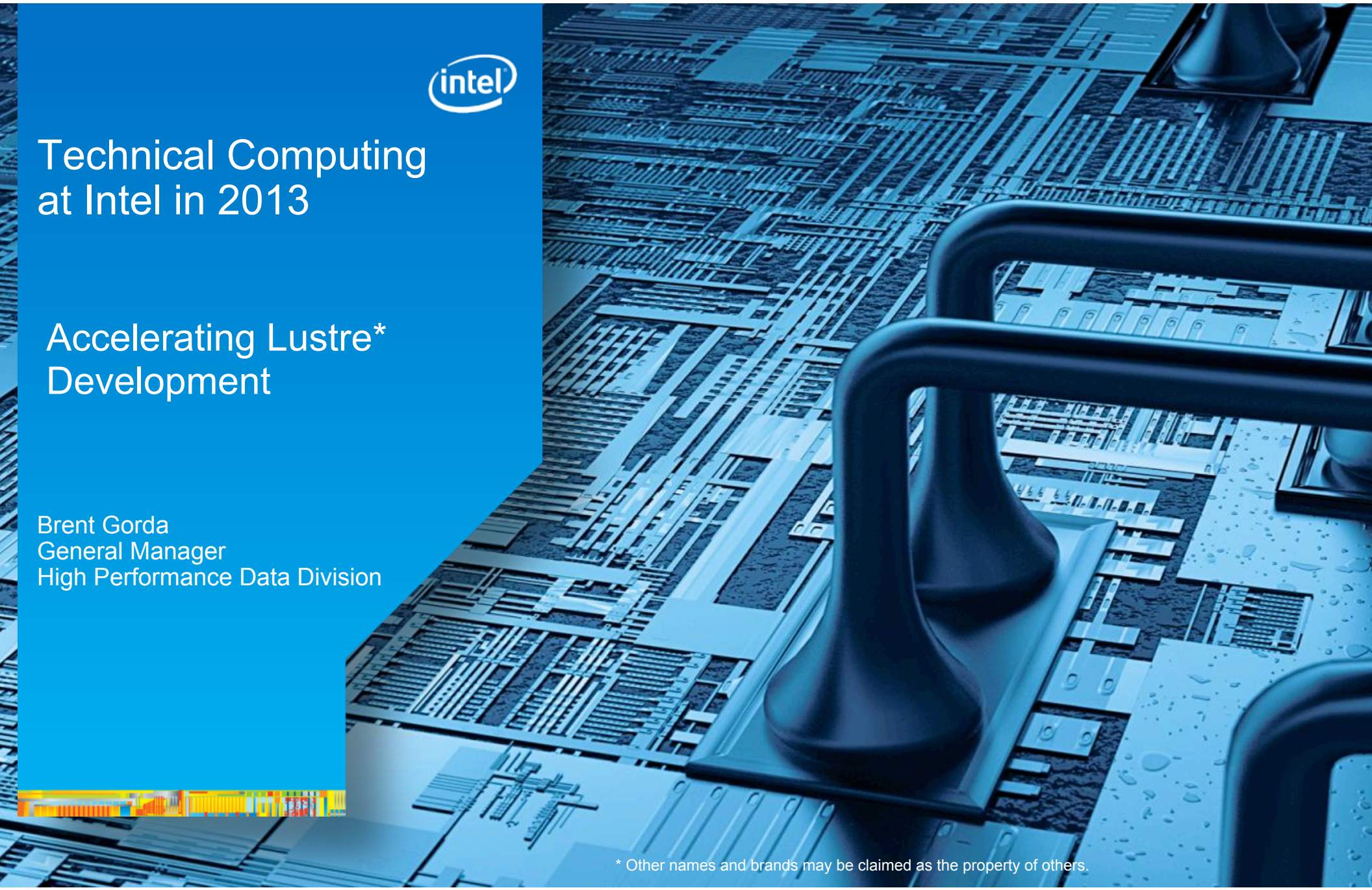


Technical Computing at Intel in 2013

Accelerating Lustre* Development

Brent Gorda
General Manager
High Performance Data Division

From Whamcloud to Intel

Started July 16, 2010

- Brent Gorda – CEO
- Eric Barton – CTO

Founded Whamcloud to keep Lustre* in play and vendor-neutral for HPC

- Recognized by OpenSFS and EOFS as the maintainer of open source repositories

Acquired by Intel in July 2012

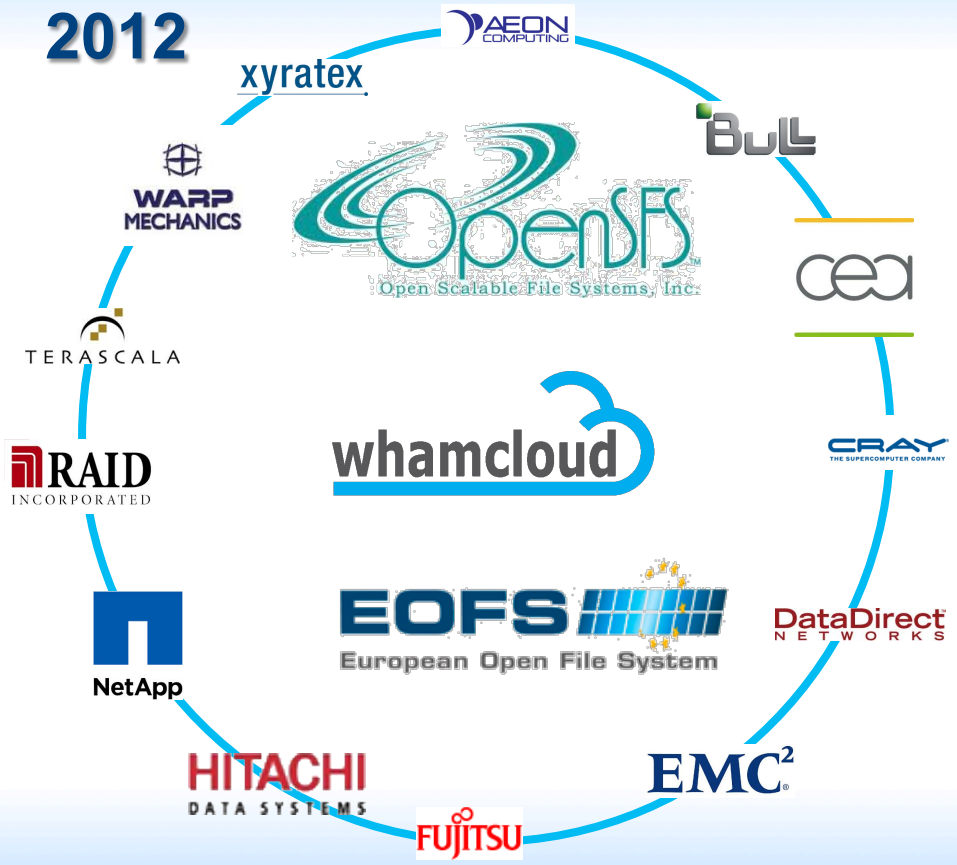
- Becomes the High Performance Data Division
- Same team, same mission, more resources

Development of a Vibrant Ecosystem

2010

ORACLE

2012



Moving Lustre Forward

- Continue to focus on traditional HPC requirements…
  - Drive open and collaborative development
  - Mange the open source tree on behalf of the community
  - Rigorously tested to ensure high quality
  - Member of EOFS and OpenSFS Board of Directors
- Work to penetrate the enterprise and Big Data' markets
- Provide support services to both Lustre communities
  - Open source releases on a regular cadence
  - Worldwide, multi-vendor support for any user

"Intel bought Whamcloud to be Whamcloud…"

- Boyd Davis, VP and General Manager, Datacenter Software Division

Lustre* Partner and Solution Ecosystem

Advocacy

- EOFS European Open File System
- OpenSFS
- intel

Storage

- DataDirect Networks
- NetApp
- HITACHI Data Systems
- DELL
- EMC²
- FUJITSU
- RAID Incorporated
- TERASCALA
- xyratex

Compute

- CRAY The Supercomputer Company
- AEON Computing
- BULL
- sgi
- Penguin Computing
- APPRO Supercomputer Solutions
- Advanced HPC

Integration

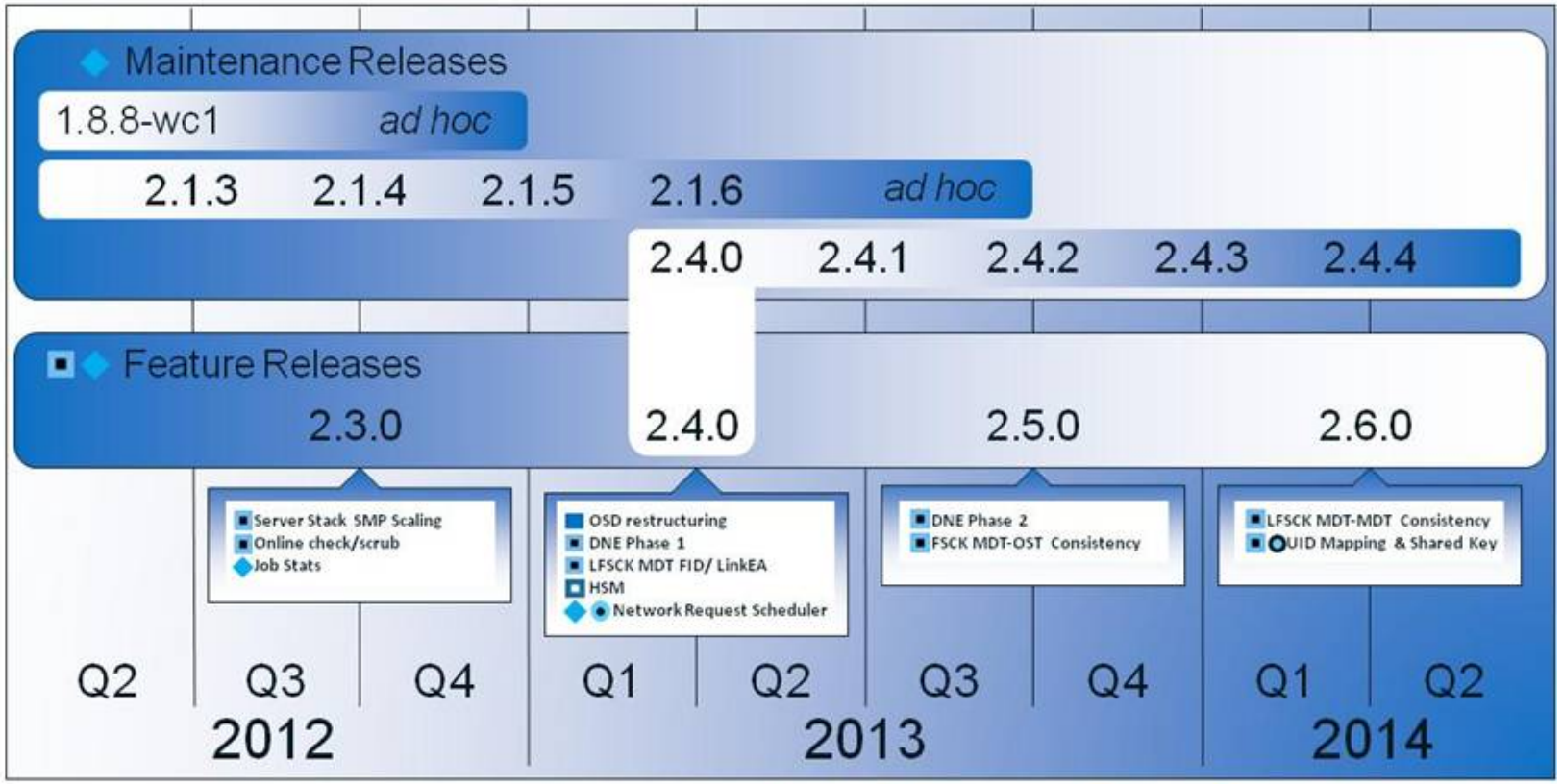
- ParTec Cluster Competence Center
- iNET 2000
- ALCES Software
- Red Hat
- VPAC Infinite Innovation
- Eucalyptus Systems

Development

- Indiana University
- Oak Ridge National Laboratory
- Lawrence Livermore National Laboratory
- NASA
- TACC
- cea
- SDSC San Diego Supercomputer Center
- ASC

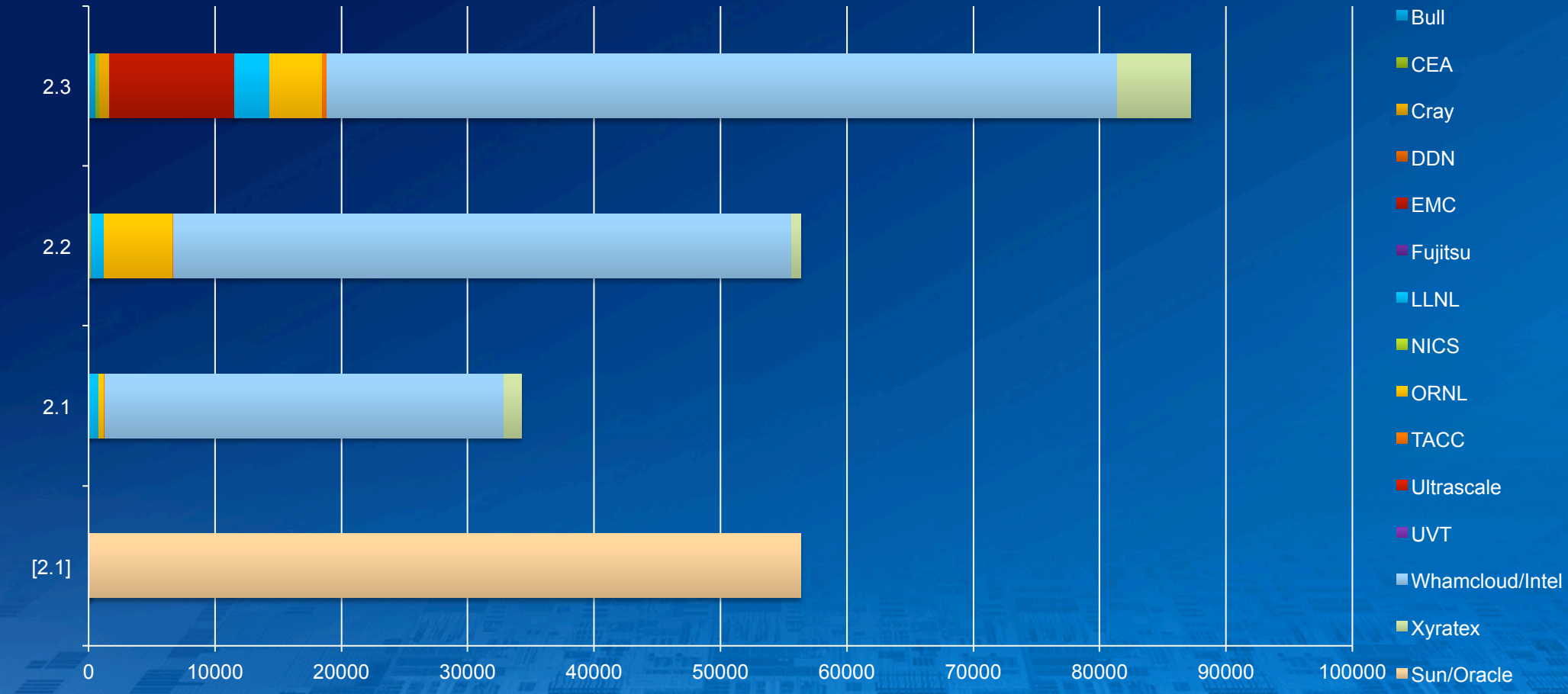
COMMUNITY LUSTRE*

Community Lustre* Roadmap



Sponsor for Intel Development and Releases: ORNL, OpenSFS, LLNL, Intel
Third Party Development: CEA, Xyratex, Indiana University

Increasing Community Participation



el internal statistics related to the lines of approved code per contributor per release.

Notable New Features and Enhancements

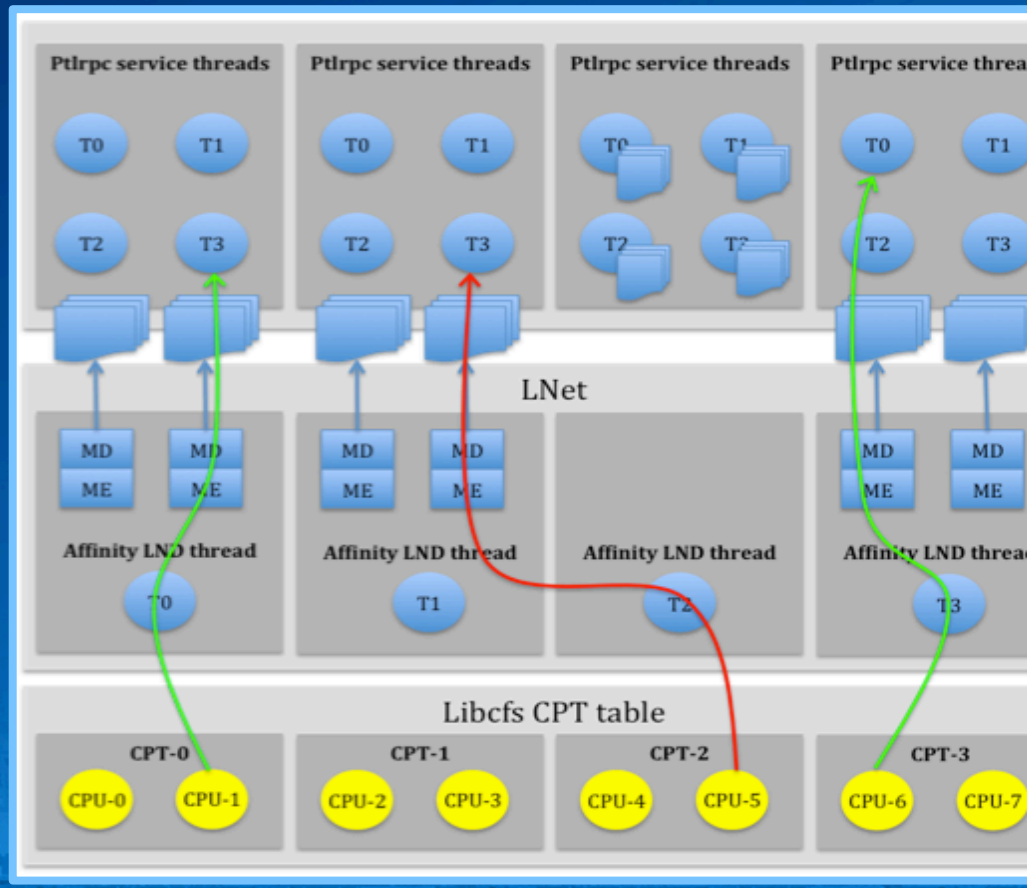
- Multiple performance enhancements
  - SMP Scaling
  - Multiple MDS support via Distributed Namespace (DNE)
- Object Storage Device API
- Hierarchical Storage Management API
- Distributed, automated test infrastructure

  maloo.whamcloud.com
- URL to Community Development details:

  wiki.whamcloud.com/display/PUB/Lustre+Community+Development+in+Progress

Metadata Server Performance

- CPU Partition (CPT)
  - Similar to cpuset in Linux
  - Easily used by kernel thread
- Partitioned LNET (LND)
  - LND thread-pool for each CPT
  - Core LNet has partition data
- Partitioned ptlrpc service
  - Ptlrpc service thread-pool for each CPT
  - Request-queue & wait-queue for each CPT



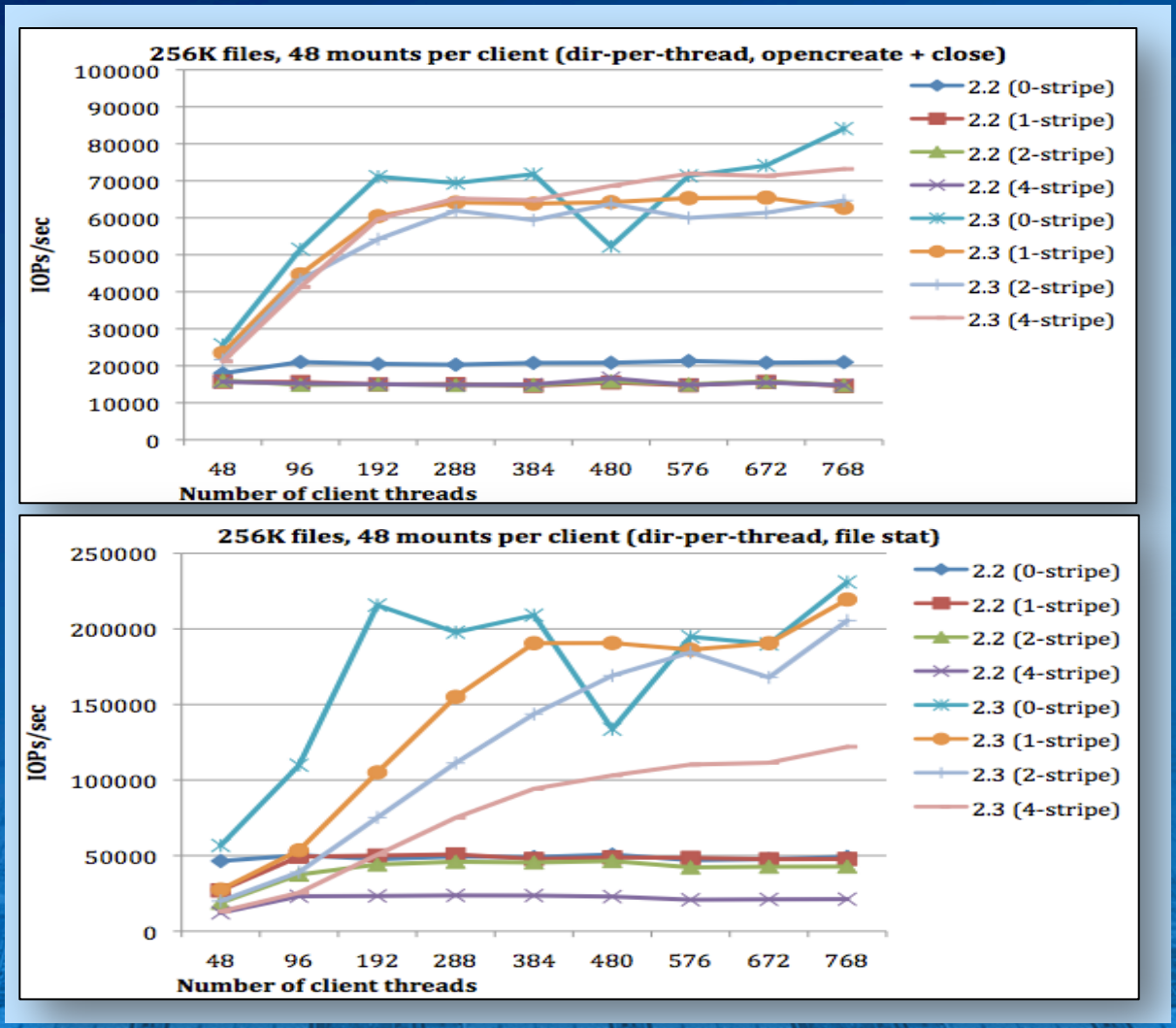
Improved Metadata Server Performance

4X improvement in file creation performance

- 15k-20k → 60k-80k

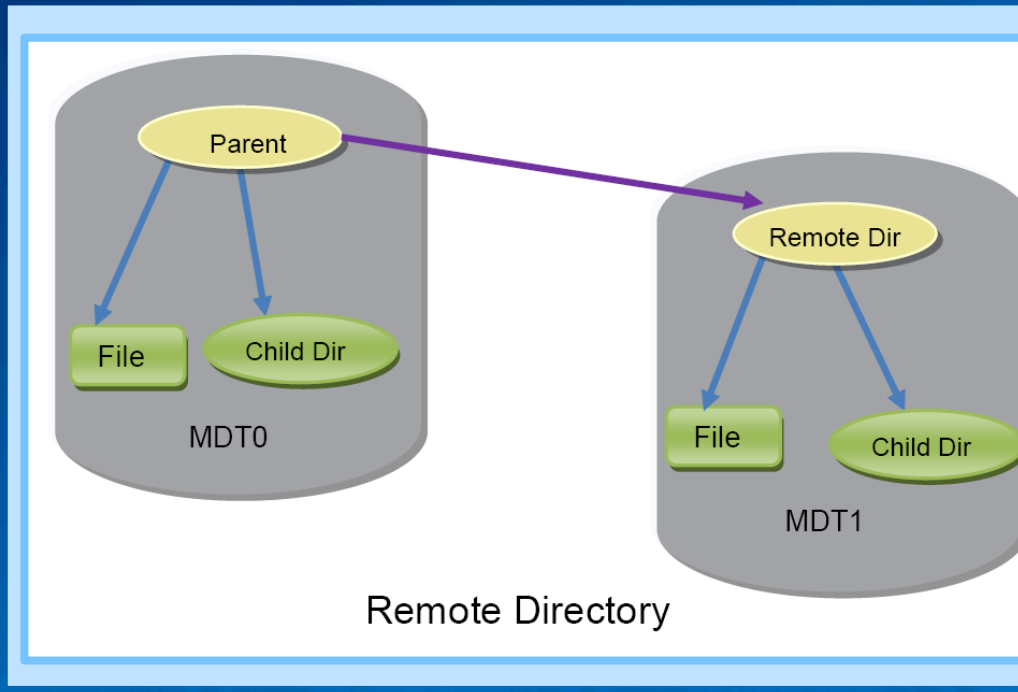
4X improvement in file stat performance

- 50K → 200K

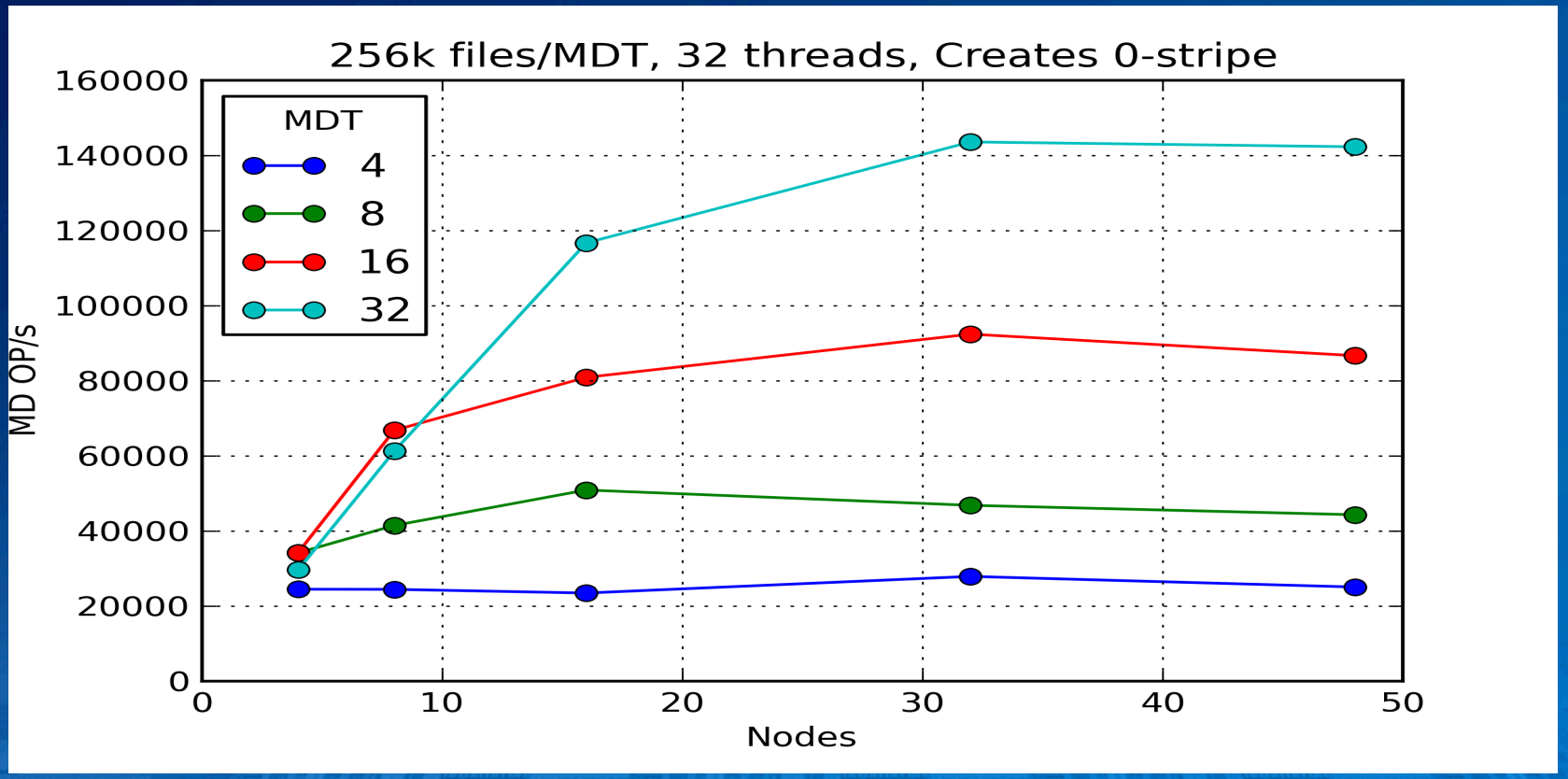


Distributed Namespace

- Distributes namespace by remote directory
- Supports active/active failover use
  - Allows multiple MDT to be exported from one MDS
  - Supports active/active failover for metadata and data
- Linear performance improvements seen
- Root command to mkdir on secondary MDS
- Additional features targeted for upcoming releases



Scalable Metadata Performance (DNE)



Object Storage Device API

An abstraction layer between the storage ‘file system’ and Lustre usage

Allows Lustre to support other file systems types as backing store

- Start with ZFS integration
- Potential future use with btrfs

Lustre 2.4 can leverage many ZFS features

- Mature and robust file system
- Scales beyond current file system limits
  - Object count and size
  - File system size
- Easier management of many disks, commodity JBODs without RAID hardware
- Integrated with flash storage cache (L2ARC read cache)

Hierarchical Storage Management

- Important feature for traditional and commercial HPC
  - Move data between tiered storage to meet performance, capacity and availability
  - Classes of storage can include SSD, disk and tape
- Uses *Robin Hood* policy engine developed by CEA
  - Leverages ChangeLog for minimal impact on performance
- Client-side implemented in Lustre 2.4
  - Layout lock, copytools API, RPC protocol
- Server changes currently under development, targeted for 2.5
- Infrastructure for Intel proposed data migration and replication features

Project Maloo – Distributed test infrastructure

- Foundation for consistent, repeatable testing
- Improved quality assurance
- Easily review coverage
- Supports upgrade/downgrade and interoperability testing
- Key to ensuring stability and predictable releases
- Go to maloo.whamcloud.com for more details

Project Maloo Dashboard

s | results | reports & stats | queues | user admin | **Chris Gearing** [settings | logou

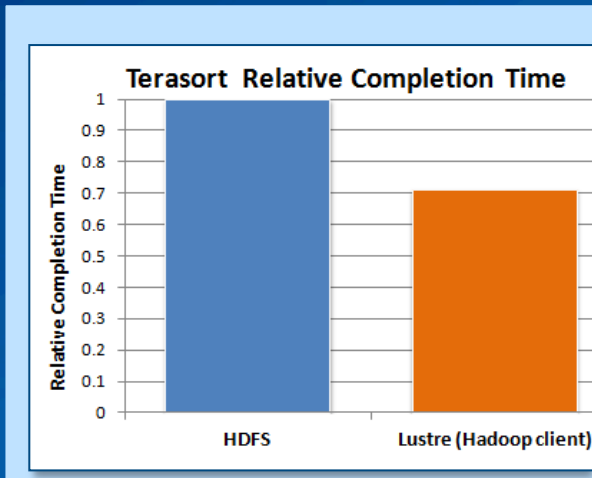
rate report for lustre-release - master

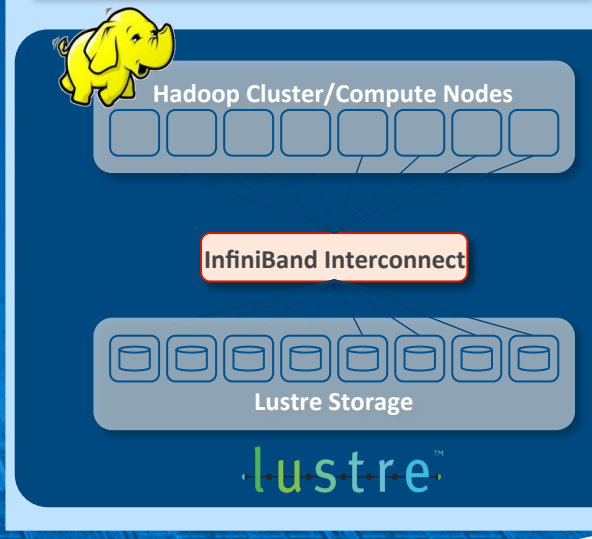
on-lustre test sets

	2.3.62 87ee788 2013-03-06	2.3.61 2c6702b 2013-02-10	2.3.59 7677269 2013-01-19	2.3.58 1f77320 2012-12-31	2.3.56 e72ffc3 2012-11-19	2.3.54 241615b 2012-10-29	2.3.53 5f9e428 2012-10-08	2.2.93 861105f 2012-08-16	2.2.92 fee5548 2012-07-30	2.2.91 cae478c 2012-07-19	2.2.90 1934a98 2012-07-10	2.2.59 84a414b 2012-07-02	2.2.57 b3b8bc5 2012-06-19	2.2.56 68eb992 2012-06-18	2.2.55 4ae3e06 2012-06-14	2.2. 240 201
t_upgrade										⊕ 1/1						
_upgrade										⊕ 1/1						
y	⊕ 2/4	⊕ 1/1	⊕ 2/5	⊕ 3/8	⊕ 1/3	⊕ 1/1		⊕ 6/6	⊕ 9/10	⊕ 4/6	⊕ 10/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 4/4	⊕ 5
	⊕ 4/4	⊕ 1/1	⊕ 3/5	⊕ 7/7	⊕ 3/3	⊕ 1/1		⊕ 6/6	⊕ 10/10	⊕ 6/6	⊕ 10/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 0/4	⊕ 7
e	⊕ 4/4	⊕ 1/1	⊕ 3/5	⊕ 7/7	⊕ 3/3	⊕ 1/1		⊕ 6/6	⊕ 6/9	⊕ 4/6	⊕ 10/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 3/4	⊕ 6
	⊕ 3/4	⊕ 0/1	⊕ 5/5	⊕ 7/8	⊕ 2/3	⊕ 0/1		⊕ 2/6	⊕ 2/10	⊕ 2/6	⊕ 3/10	⊕ 4/6	⊕ 1/6	⊕ 1/3	⊕ 0/4	⊕ 3
st	⊕ 3/4	⊕ 1/1	⊕ 5/5	⊕ 7/7	⊕ 3/3	⊕ 1/1		⊕ 6/6	⊕ 6/9	⊕ 5/6	⊕ 10/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 3/4	⊕ 6
c-test	⊕ 3/4	⊕ 1/1	⊕ 3/5	⊕ 6/7	⊕ 2/3	⊕ 1/1		⊕ 5/5	⊕ 6/9	⊕ 4/5	⊕ 7/8	⊕ 5/5	⊕ 5/5	⊕ 2/2	⊕ 3/4	⊕ 5
y	⊕ 3/3	⊕ 1/1	⊕ 5/5	⊕ 7/7	⊕ 3/3	⊕ 1/1		⊕ 5/5	⊕ 3/7	⊕ 3/4	⊕ 7/7	⊕ 4/4	⊕ 4/4	⊕ 2/2	⊕ 2/3	⊕ 2
updates	⊕ 4/4	⊕ 1/1	⊕ 3/5	⊕ 7/7	⊕ 2/3	⊕ 1/1		⊕ 6/6	⊕ 10/10	⊕ 5/6	⊕ 10/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 3/4	⊕ 7
	⊕ 4/4	⊕ 1/1	⊕ 3/6	⊕ 7/9	⊕ 3/3	⊕ 1/1		⊕ 6/6	⊕ 7/10	⊕ 5/8	⊕ 10/14	⊕ 6/8	⊕ 6/7	⊕ 3/3	⊕ 4/5	⊕ 7
urvey	⊕ 4/4	⊕ 1/1	⊕ 3/5	⊕ 7/7	⊕ 2/3	⊕ 1/1		⊕ 1/6	⊕ 6/9	⊕ 5/6	⊕ 10/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 3/4	⊕ 6
	⊕ 4/4	⊕ 0/1	⊕ 3/5	⊕ 3/7	⊕ 1/3	⊕ 1/1		⊕ 0/6	⊕ 9/10	⊕ 5/6	⊕ 10/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 3/4	⊕ 7
ale	⊕ 4/4	⊕ 1/1	⊕ 2/5	⊕ 4/7	⊕ 2/3	⊕ 1/1		⊕ 5/6	⊕ 6/10	⊕ 4/6	⊕ 7/10	⊕ 6/6	⊕ 5/6	⊕ 3/3	⊕ 3/4	⊕ 4
ale-nfsv3	⊕ 4/4	⊕ 1/1	⊕ 1/5	⊕ 7/7	⊕ 1/3	⊕ 0/1		⊕ 4/5	⊕ 0/8	⊕ 2/5	⊕ 6/9	⊕ 4/5	⊕ 4/5	⊕ 2/3	⊕ 2/3	⊕ 5
ale-nfsv4	⊕ 1/4	⊕ 0/1	⊕ 2/5	⊕ 7/7	⊕ 2/2	⊕ 1/1		⊕ 5/5	⊕ 0/7	⊕ 1/4	⊕ 7/9	⊕ 0/5	⊕ 1/5	⊕ 1/3	⊕ 0/3	⊕ 2
ce-sanity	⊕ 4/4	⊕ 1/1	⊕ 2/5	⊕ 6/7	⊕ 3/3	⊕ 1/1		⊕ 6/6	⊕ 7/10	⊕ 4/6	⊕ 9/10	⊕ 6/6	⊕ 6/6	⊕ 3/3	⊕ 4/4	⊕ 6
	⊕ 1/4	⊕ 0/1	⊕ 3/5	⊕ 5/7	⊕ 0/2	⊕ 1/1		⊕ 3/4	⊕ 2/6	⊕ 0/3	⊕ 0/7	⊕ 0/4				

ccelerating Hadoop Workloads

- Bringing Hadoop analytics to HPC
- Initial work demonstrates the advantaged of shared storage
  - Exploit the superior performance, scalability and management simplicity of shared storage
  - Scale storage and compute nodes separately
- On track to deliver the combined benefits of Lustre* with Intel® Distribution of Apache Hadoop*





Intel® Manager for Lustre* Software

Experience the benefits of Lustre* powered workloads faster and easier

Simplifies installation, configuration, monitoring and management

Extensible and easy to integrate using plug-in architecture and REST interface

Management Console

- Provides graphical and command-line interfaces for file system management
- Central repository of file system details and statistics

Storage Servers

- Intelligent management software layered over object server nodes

Exascale File System

Integrated I/O Stack

- Epoch transaction model
- Non-blocking scalable object I/O

HDF5/other schema

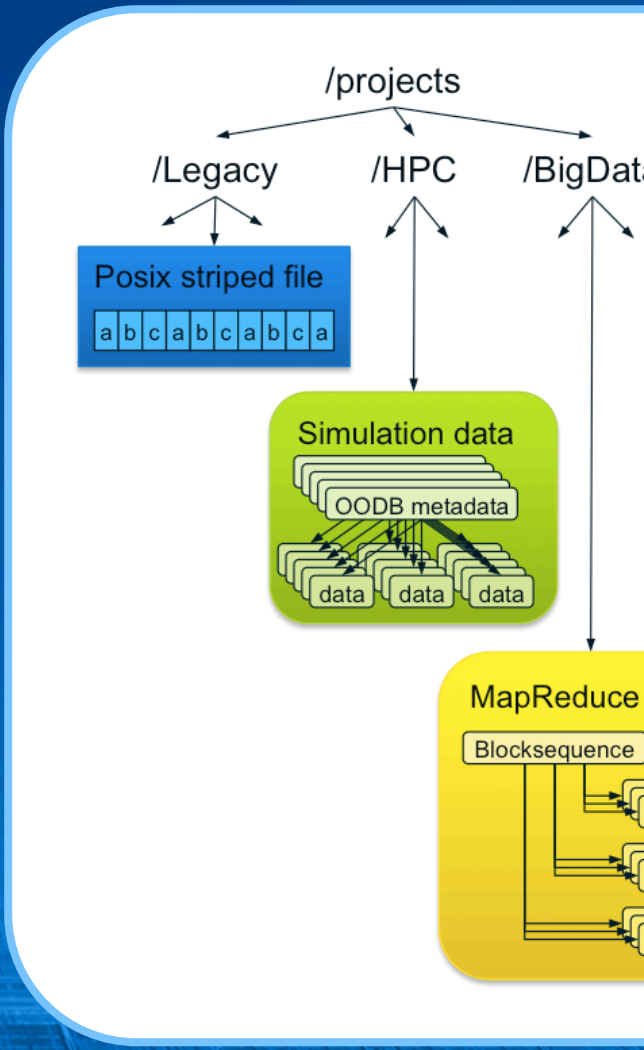
- High level application object I/O model
- I/O forwarding

I/O Dispatcher

- Burst Buffer management
- Impedance match application I/O performance to storage system capabilities

DAOS

- Conventional namespace
- DAOS container files for transactional, scalable, object I/O




Thank You.