SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2014

LustreFS and its ongoing Evolution for High Performance Computing and Data Analysis Solutions

Roger Goff

Senior Product Manager

DataDirect Networks, Inc.

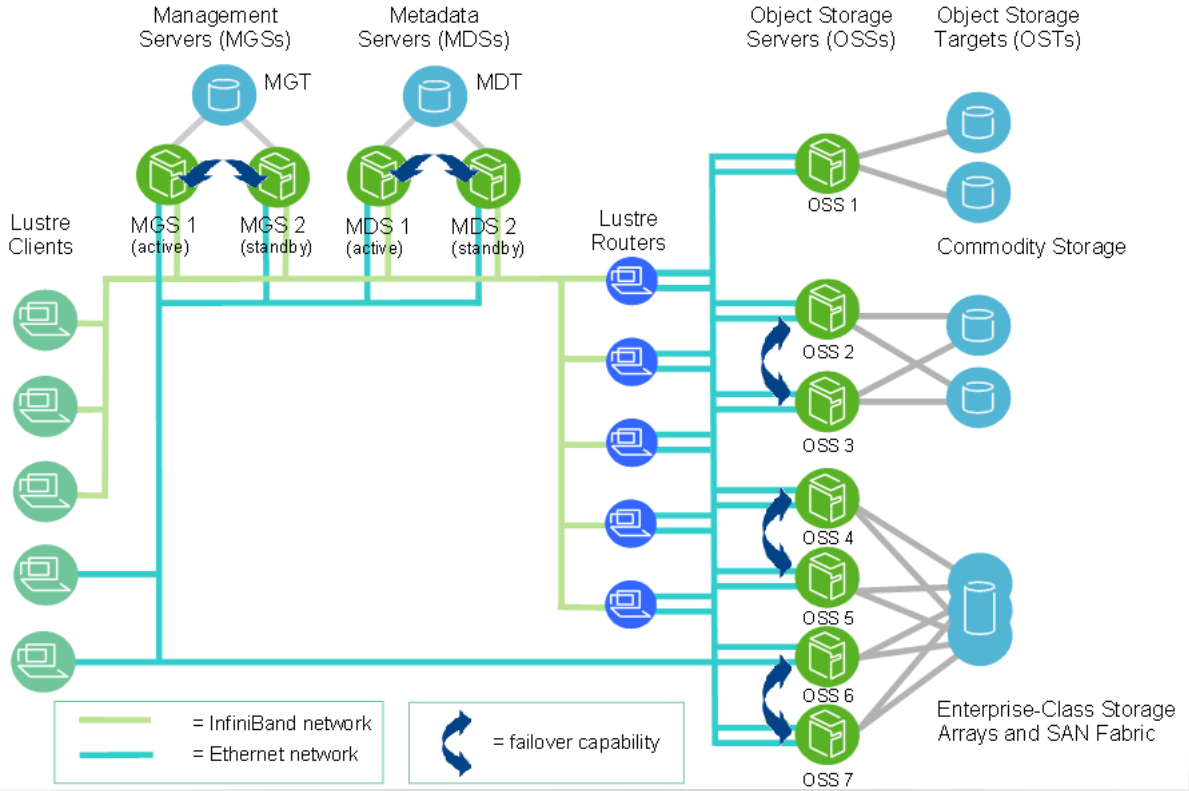
What is Lustre?

- Parallel/shared file system for clusters
  - Aggregates many systems into one file system
  - Hugely scalable - 100s of GB/s throughput, 10s of thousands of native clients Scales I/O throughput and capacity
- Widely adopted in HPC community
- Open source, community driven development
- Supports RDMA over high speed low latency interconnects and Ethernet
- Runs on commodity storage and purpose built HPC storage platforms
  - Abstracts hardware view of storage from software
- Highly available

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre Node Types

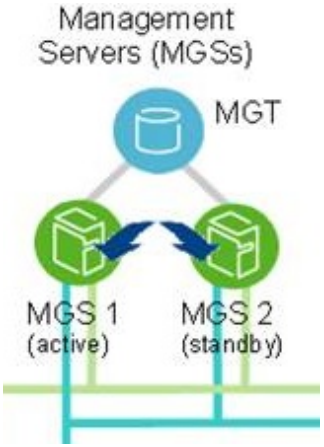
- Management Servers (MGS)
- Meta-Data Servers (MDS)
- Object Storage Servers (OSS)
- Clients
- LNET Routers



SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre MGS

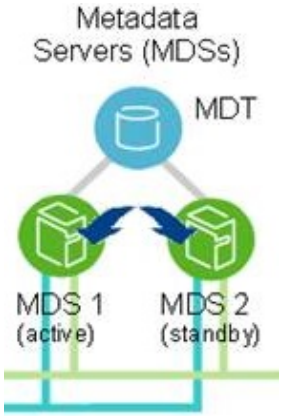


- ManaGement Server
- Server node which manages cluster configuration database
- All clients, OSS and MDS need to know how to contact the MGS
- MGS provides all other components with information about the Lustre file system(s) it manages
- MGS can be separated from or co-located with MDS

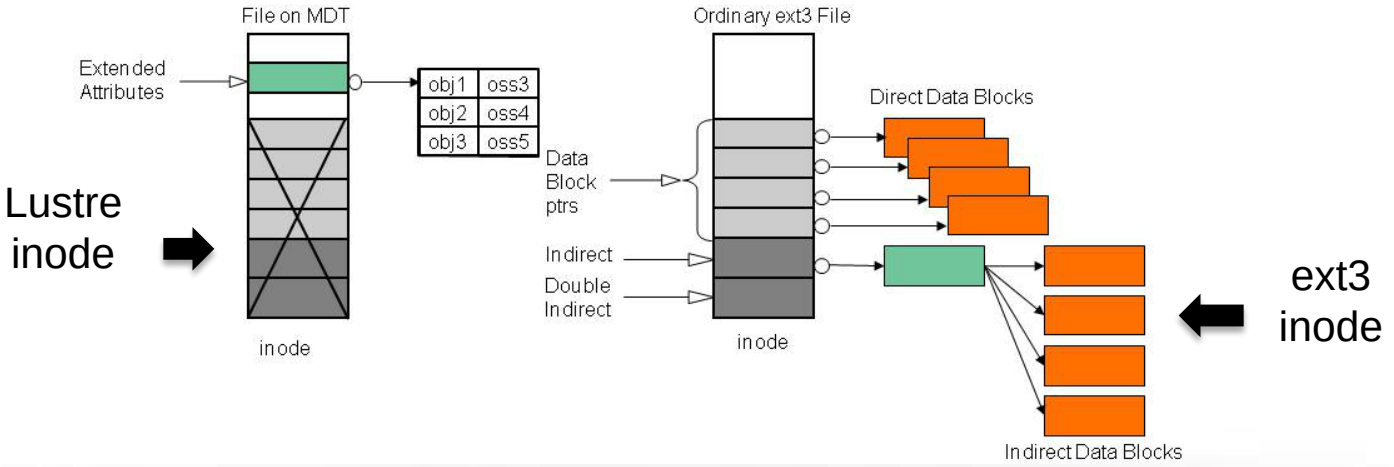
SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre MDS



- Meta Data Server
- Manages the names and directories in the Lustre file system
- Stores metadata (such as filenames, directories, permissions and file layout) on a MetaData Target (MDT)
- Usually MDSs are configured in an active/passive failover pair

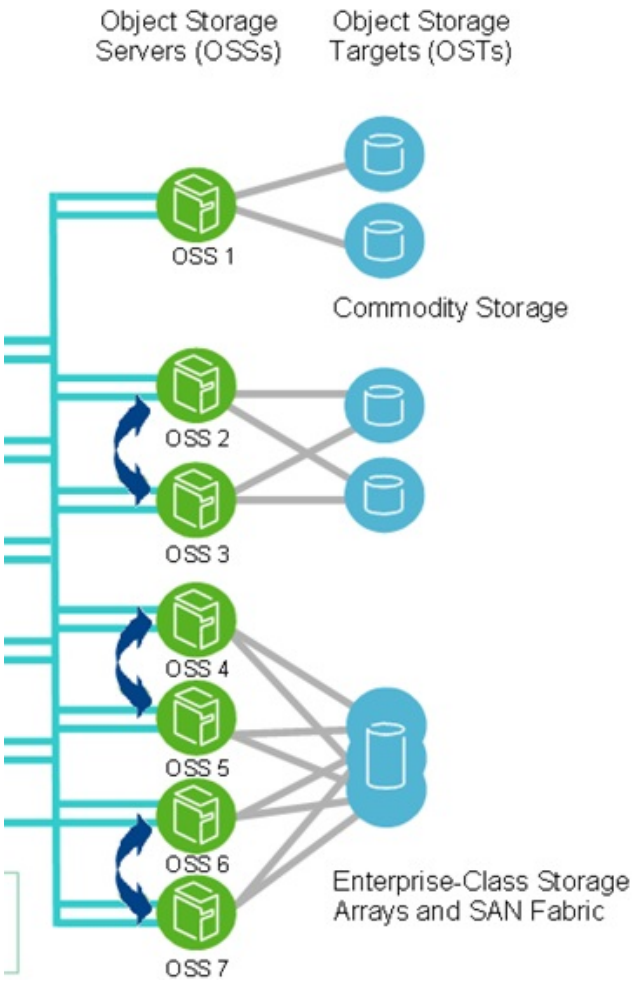


SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre OSS

- Object Storage Server
- Provides file I/O service, and network request handling for one or more Object Storage Target (OST)
- Can be configured for high availability
- Storage agnostic

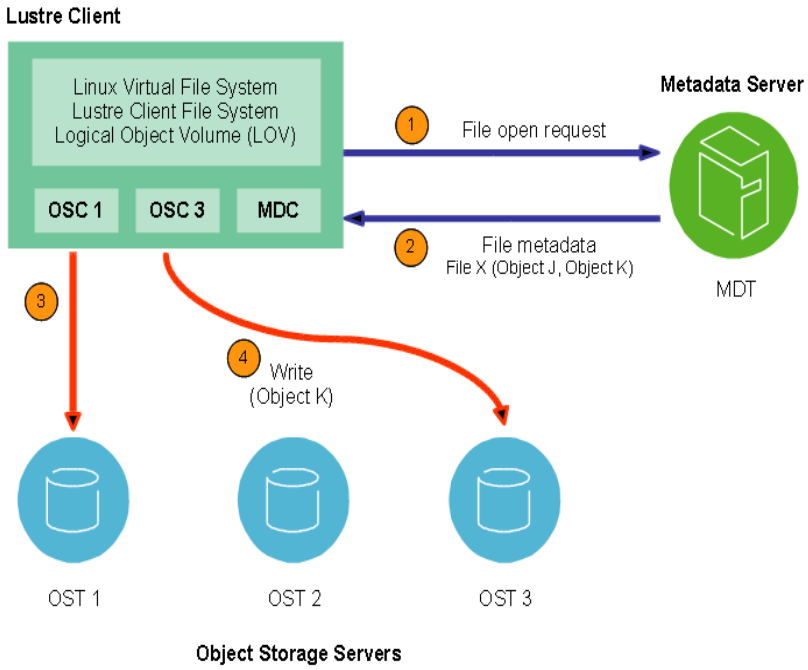


SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre Clients

- POSIX compliant file system access layer
- Has Logical Object Volume (LOV) layer that manages file striping across multiple OSTs.
- Has client layers for each of the servers e.g. Object Storage Client (OSC) to interact with OSS and Metadata Client (MDC) to interact with MDS

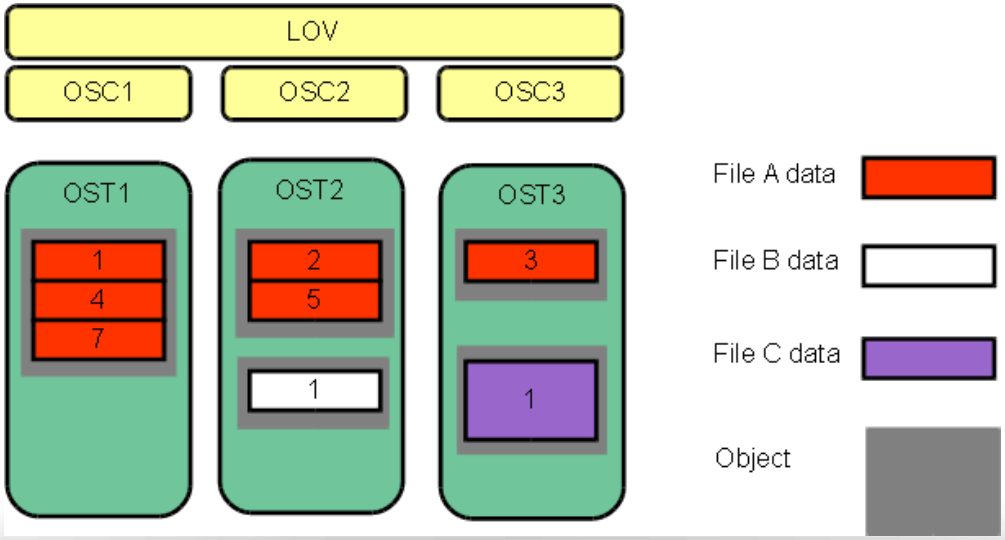


SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre File Striping

- The ability to stripe data across multiple OSTs leads to high performance
- Stripes used to improve performance when the desired aggregate bandwidth to a single file exceeds the bandwidth of a single OST
- stripe_count and stripe_size can be set per file



SDC 14

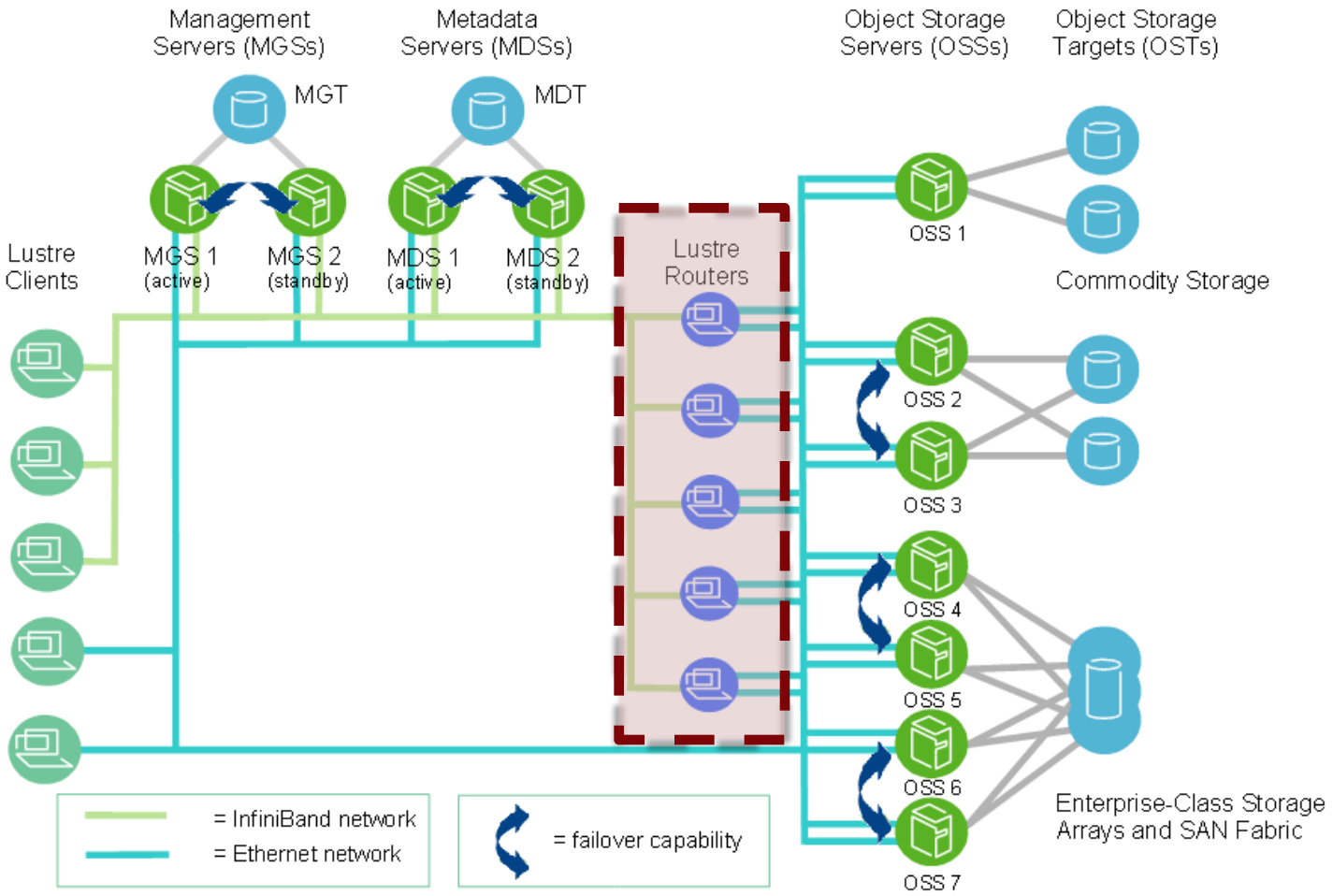
2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre Networking (LNET)

- LNET Routers provide simultaneous availability of multiple network types with routing between them.
- Support for many commonly-used network types such as InfiniBand and TCP/IP
- RDMA when supported by underlying networks
- High availability and recovery features enabling transparent recovery in conjunction with failover servers
- LNET can use bonded networks
- Supported Networks:
  - InfiniBand, TCP, Cray: Seastar, Myrinet: MX, RapidArray, Quadrics: Elan

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

LNET Routers in Lustre Solutions



SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre File Systems

- Public Sector Lustre Systems:
  - Lawrence Livermore National Lab – 64PB, 1.5TB/s, 96,000 clients
  - Texas Advanced Computing Center – 20PB, 250+GB/s
  - Oakridge “Spider” – 10PB, 240GB/s, 26,000 clients
  - Environmental Molecular Science Lab @ PNNL – 75GB/s
- Commercial Lustre Systems:
  - Energy Company (unnamed) – 10.5PB, 200GB/s
  - Energy Company (unnamed) – 270GB/s
  - Wellcome Trust Sanger Institute (life sciences) – 22.5PB, 20GB/s
  - ANU – (Bureau of Meteorology + climate science labs) 6PB, 152GB/s

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre Small File Performance Improvement

- Array Cache and Flash Cache Acceleration
- Lustre Distributed Namespace (DNE)
- Lustre Data on Metadata

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Cache Optimization

DDN ReACT Intelligent Cache Management

- Caches unaligned IO
- Allows full stripe writes pass directly to disk
- Faster, safer data

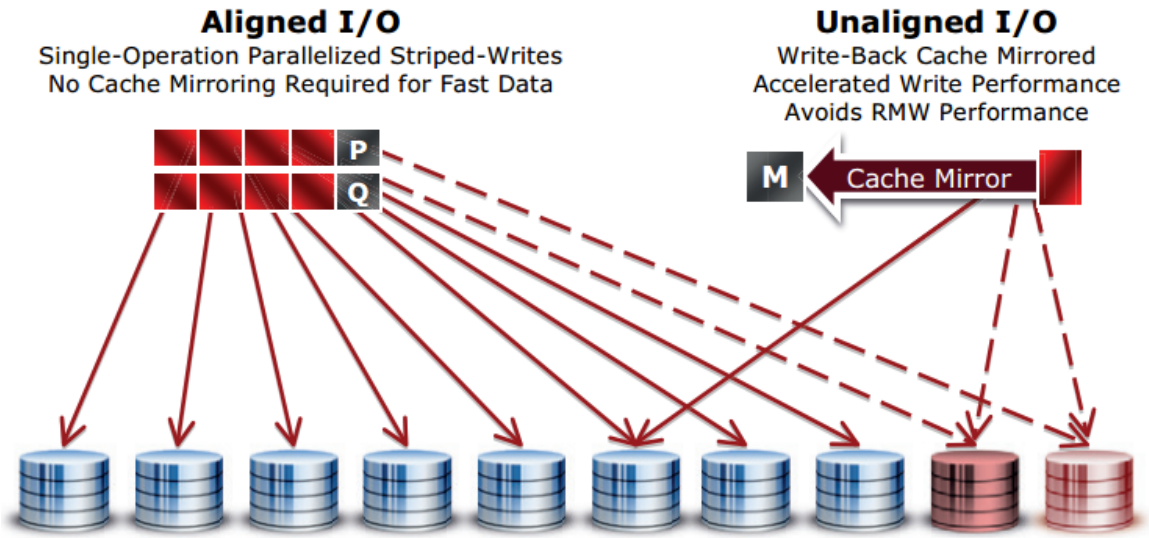
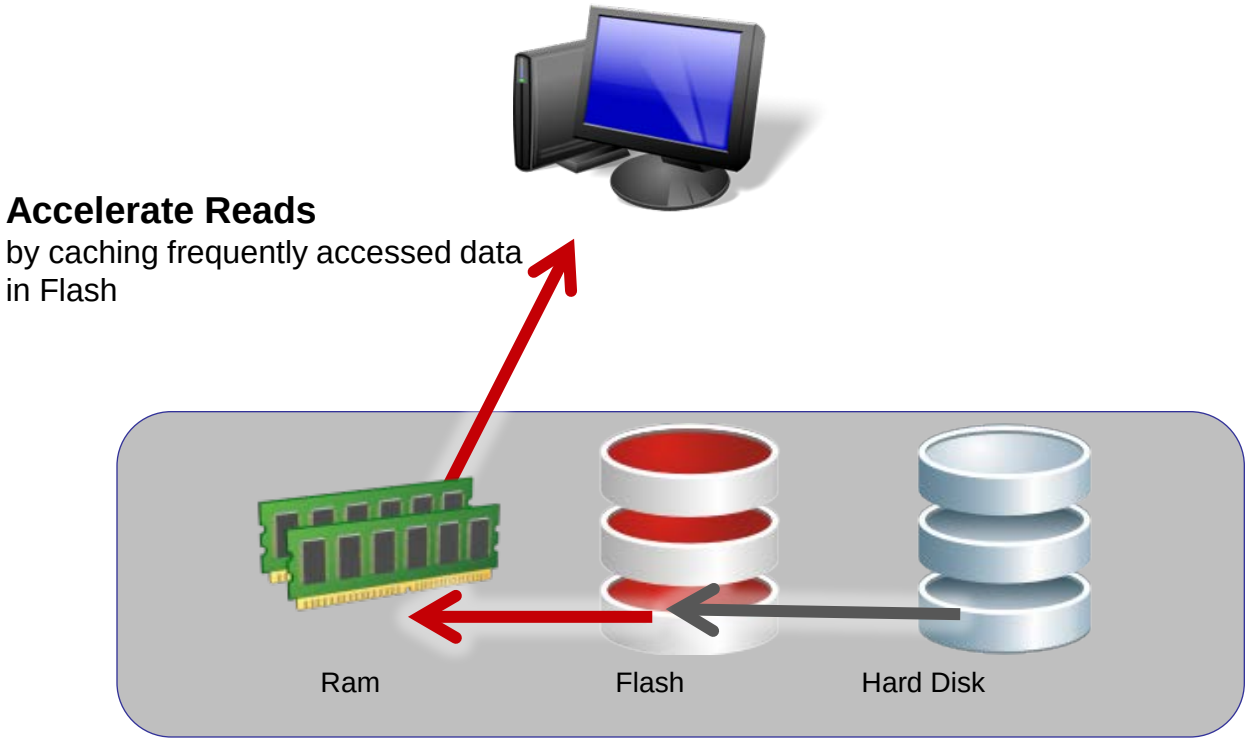


Figure 5 – Optimizing Cache Utilization with ReACT

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Flash Cache Acceleration

Re-read

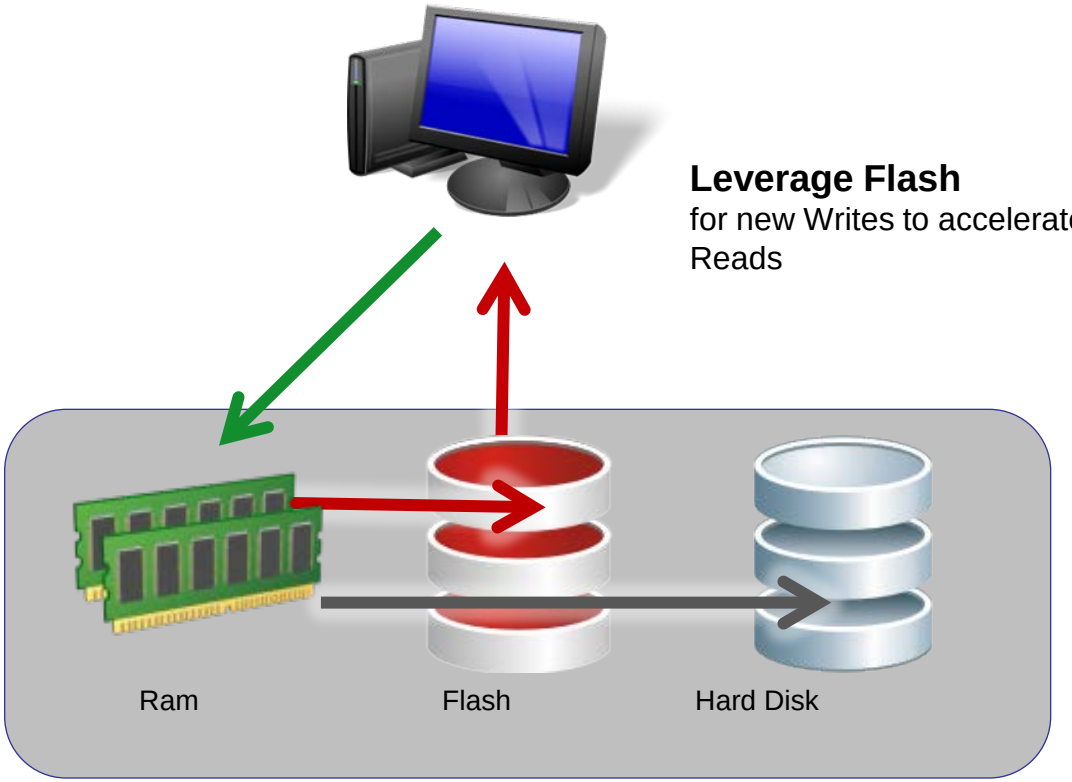


SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Flash Cache Acceleration

DDN Storage Fusion Xcelerator™ (SFX) - Instant Commit



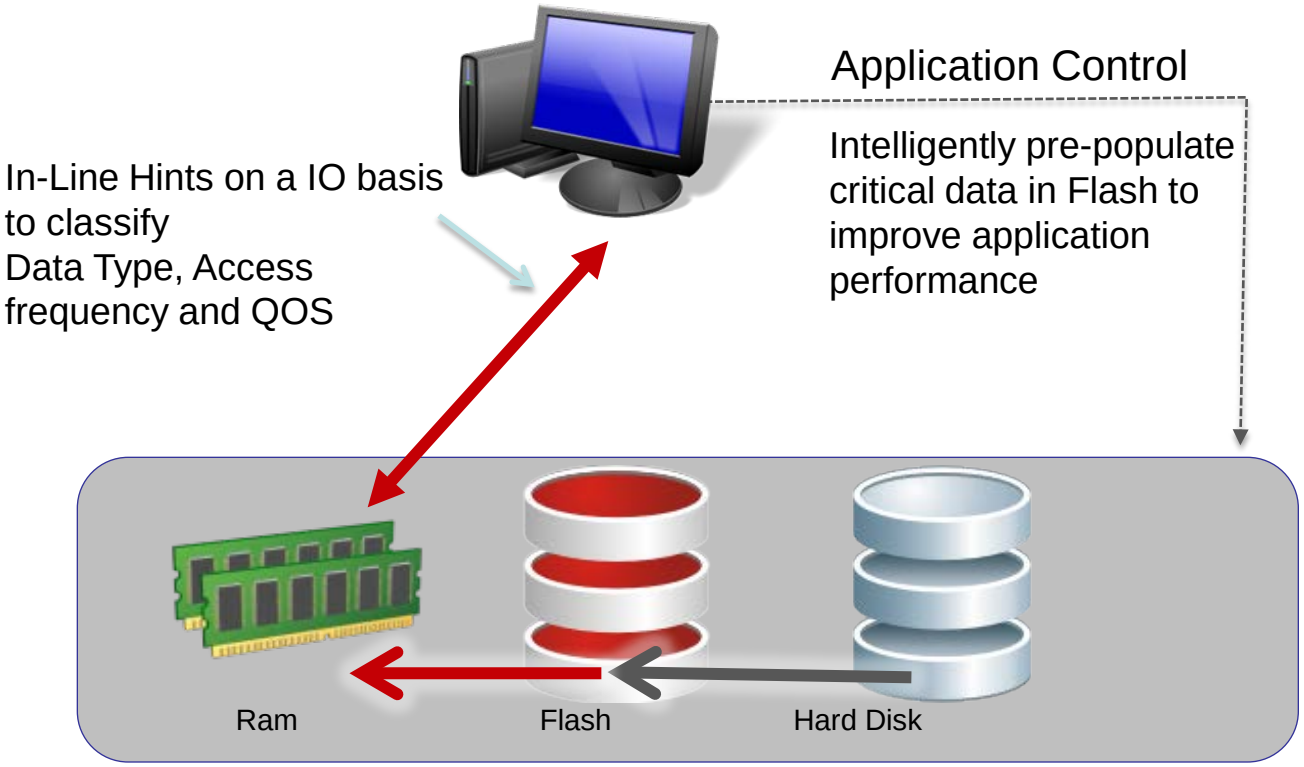
Leverage Flash
for new Writes to accelerate subsequent Reads

SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Flash Cache Acceleration

DDN Storage Fusion Xcelerator™ (SFX) - Context Commit



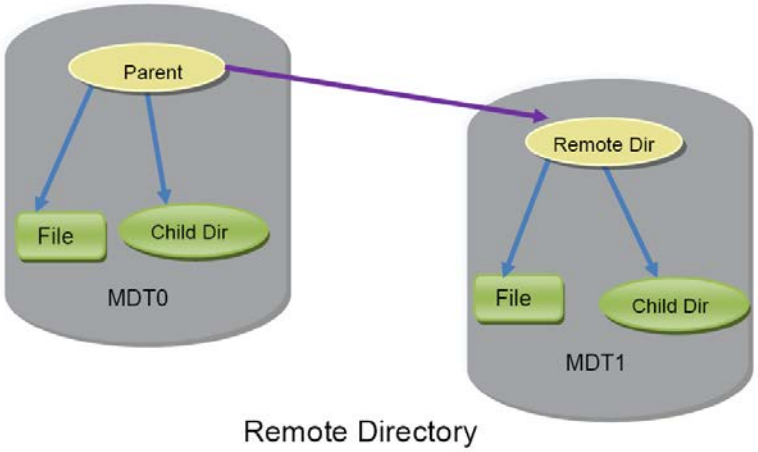
SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre Distributed NamespacE (DNE)

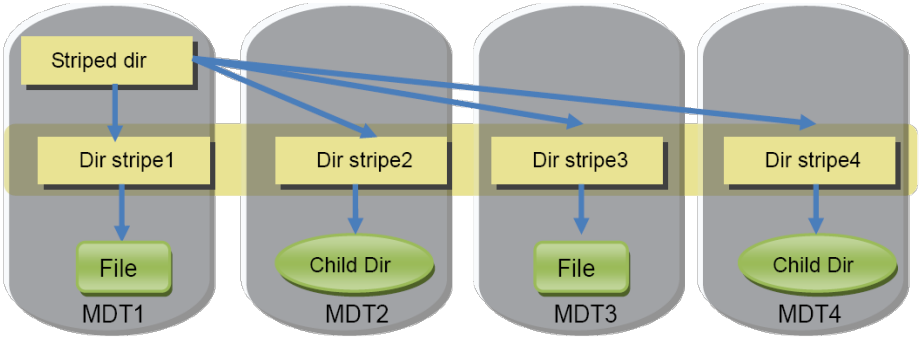
- DNE allows the Lustre namespace to be divided across multiple metadata servers
- Enables the size of the namespace and metadata throughput to be scaled with the number of servers
- An administrator can allocate specific metadata resources for different sub-trees within the namespace

Phase 1 - 1x MDT per directory (available now)



Remote Directory

Phase 2 - Multiple MDTs per directory (available in Lustre 2.7)



Striped Directory

SDC 14

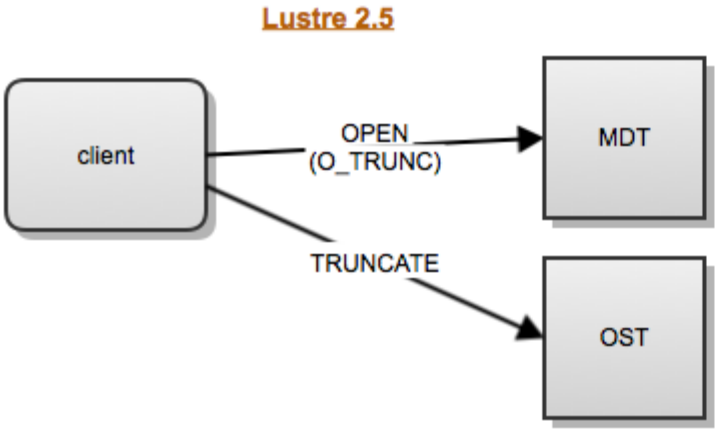
2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

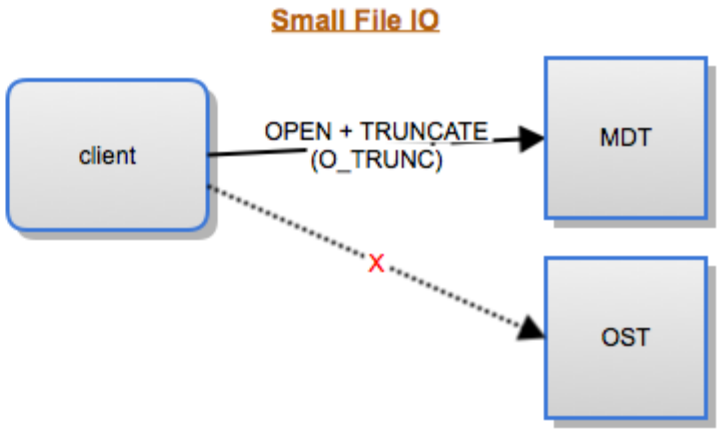
Lustre Data on Metadata

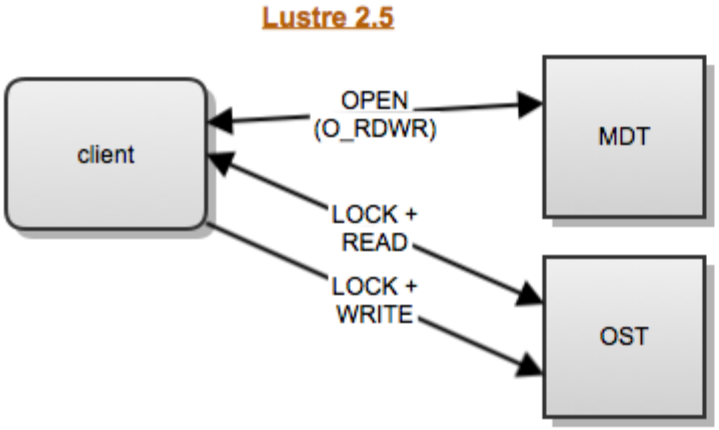
- Aims to improve small file performance by allowing the data for small files to be placed only on the MDT
- 3 Phase Implementation
  1. Basic DOM mechanism
  2. Auto migration when file becomes too large
  3. Performance optimizations
     - readahead for readdir, stat
     - First write detection so files are only created on MDT when appropriate
     - Reserve OST objects in advance even for small files

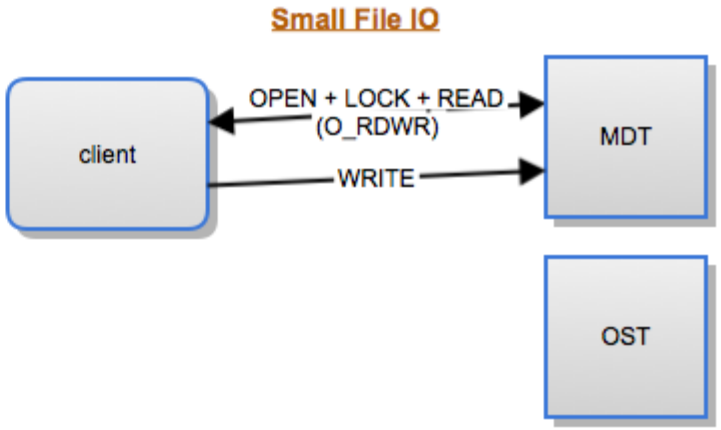
2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Reduced RPCs with Data on Metadata









SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Estimated Performance Improvement with Data on Metadata

part of operation	server	time for operation, usec	Data-on-MDT goal
Write at the end of file			
open + lock	MDT	1296 + N	1296 + N
glimpse	OST	392 + N	0
IO lock	OST	1737 + N	0
IO	OST	1760 + N	1760 + N
		5185 + 4N	3056 + 2N
Read small file			
open + lock	MDT	844 + N	844 + IO read + N
glimpse	OST	333 + N	0
IO lock	OST	726 + N	0
IO read	OST	1392 + N	0
		3295 + 4N	~ 1200 + N
Stat of existent file with data			
getattr + lock	MDT	954 + N	954 + N
glimpse	OST	654 + N	0
		1608 + 2N	954 + N

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

High Performance Data Analysis (HPDA)

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

High Performance Data Analysis (HPDA)

HPC workloads create and keep LOTS of data…

Hadoop uses local, direct-attached storage

But, HPC nodes are diskless

- PAIN POINT→ storage efficiency and management complexity

Recent IDC research uncovered:

- ~67% of HPC sites are running Hadoop workloads on their HPC systems
- Hadoop workloads consume about 30% of their computing cycles

IDC

SPECIAL STUDY

IDC HPC End-User Special Study of High-Performance Data Analysis (HPDA): Where Big Data Meets HPC

Steve Conway Chirag Dekate, Ph.D.
Earl C. Joseph, Ph.D.

IDC OPINION

This study is part of the third edition of IDC's end-user special study of the worldwide high-performance computing (HPC) market. IDC coined the term high-performance data analysis (HPDA) to refer to data-intensive ("Big Data") workloads that require or benefit greatly from HPC resources, even though not all HPDA beneficiaries consider themselves HPC users. These workloads include established data-intensive simulations and newer advanced analytics problems. The common denominator for HPDA problems is a degree of algorithmic complexity that is atypical for operational business workloads. Findings include:

- Especially during the rapid rise of clusters since 2002, HPC vendors have aggressively advanced the processor peak performance of their systems while paying less attention to data-intensive applications, HPC storage, and I/O capabilities. The areal densities of magnetic disks have increased dramatically, but improvements to disk I/O performance and access density have greatly lagged behind advances in disk capacity and processor speeds. And as HPC users have deployed ever-larger parallel servers, this fundamental imbalance — the gap between the server and storage sides of HPC — has grown worse. The road maps of HPC cluster vendors indicate that in relation to the continuing data explosion, tomorrow's HPC systems will be even more unbalanced (Top sided)

+18% CAGR for HPDA storage, twice HPC storage growth

* Some names and brands may be claimed as the property of others.

Slide courtesy of Intel, LUG2014 "Moving Lustre Forward" presentation

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre* + Hadoop: Open Platform for High Performance Data Analytics

Value Prop: Features, Functions, and Benefits

Performance

- Bring compute to the data: Run MapReduce* on Lustre without code changes
- Run MapReduce faster: Avoid the intermediate file shuffle with shared storage

Efficiency

- Avoid Hadoop* islands in the sea of HPC systems
- Run MapReduce jobs alongside HPC workloads with full access to the cluster resources

Manageability

- Use the seamless integration to manage one common platform for Hadoop and HPC
- Develop with multiple programming models and deploy on shared storage

* Some names and brands may be claimed as the property of others

Slide courtesy of Intel, LUG2014 “Moving Lustre Forward” presentation

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Extending Lustre beyond the data center

- HSM & policy engines
- Automated cloud tiering
- Research collaborations
- Data processing as a service
- Global business data sharing

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

HSM and Policy Managers

- Lustre 2.5 introduced HSM capabilities
  - An interface for data movement within Lustre namespace
  - Does not include policy engines
- Policy Engines
  - iRODS – Integrated Rule Oriented Data System; http://irods.org/
    - Open source data management software
    - Functions independently of storage resources and abstracts data control away from storage devices and device location
  - Robinhood
    - Open source policy engine to monitor and schedule actions on file systems
    - Developed at CEA; http://www-hpc.cea.fr/index-en.htm
  - Versity Storage Manager
    - An enterprise-class storage virtualization and archiving system that runs
    - Based on Open-source SAM-QFS

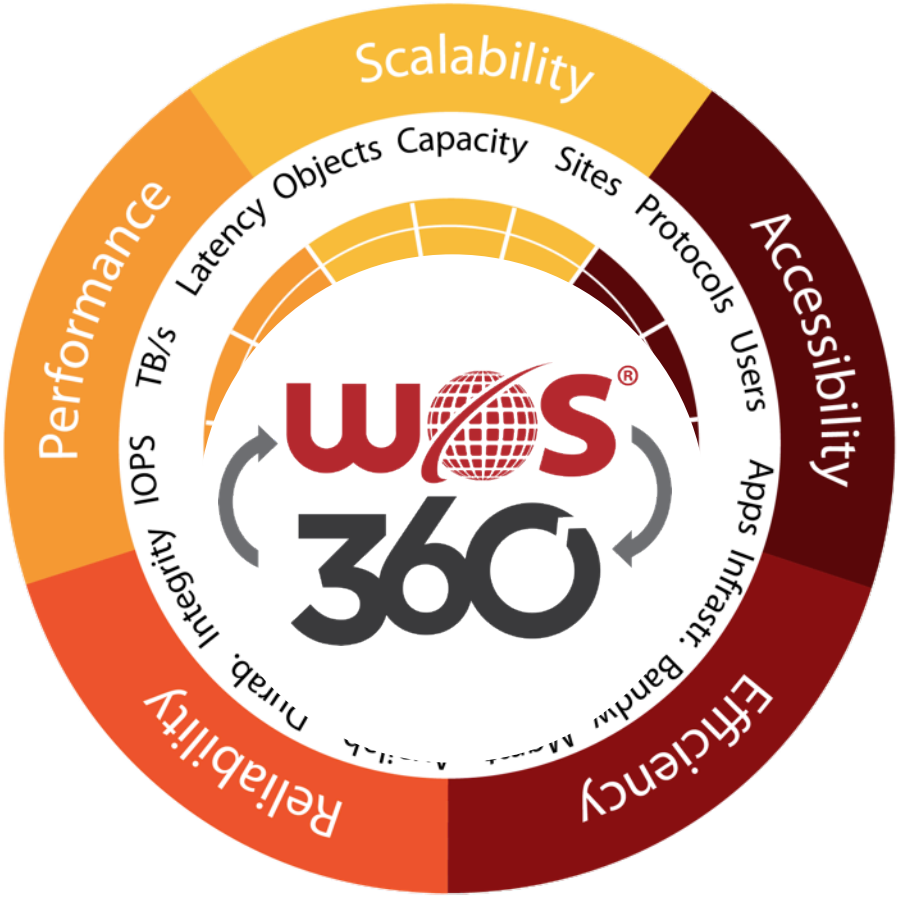
SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Bridging Lustre to a Global Object Store

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

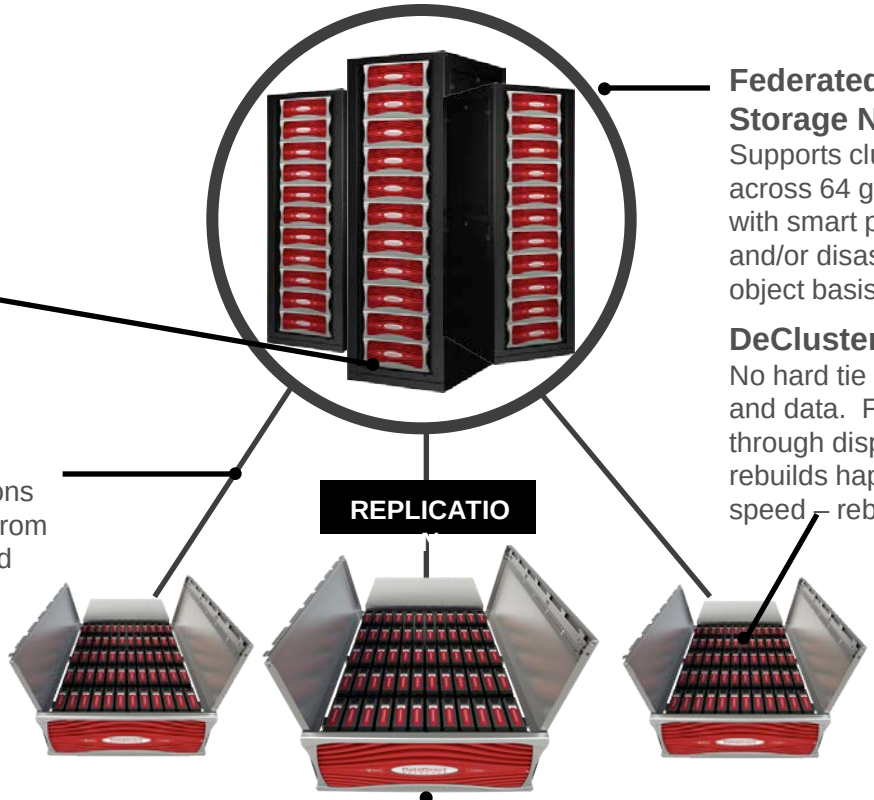
DDN Web Object Scaler
The Ultimate in Flexibility & Scale



- Architecture: choose any number of sites
- Platform: choose integrated DDN Hardware, or software-only
- Data Protection: choose from a wide range of protection schemes
- Integration: choose natively integrated applications, direct integration through native object API's or file gateways

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

WOS Feature Baseline



Federated, Global Object Storage Namespace
Supports clustering of up to 256 nodes across 64 geographies, replicate data with smart policies for performance and/or disaster recovery on a per-object basis

Flexible Data Protection
Chose between replication, erasure coding or both

DeClustered Data Protection
No hard tie between physical disks and data. Failed drives are recovered through dispersed data placement – rebuilds happen at read, not write, speed – rebuilding only data.

Latency-Aware Access Manager
WOS intelligently makes decisions on the best geographies to get from based upon location access load and latency

Fully-Integrated Object Storage Appliance
60 Drives in 4U WOS7000

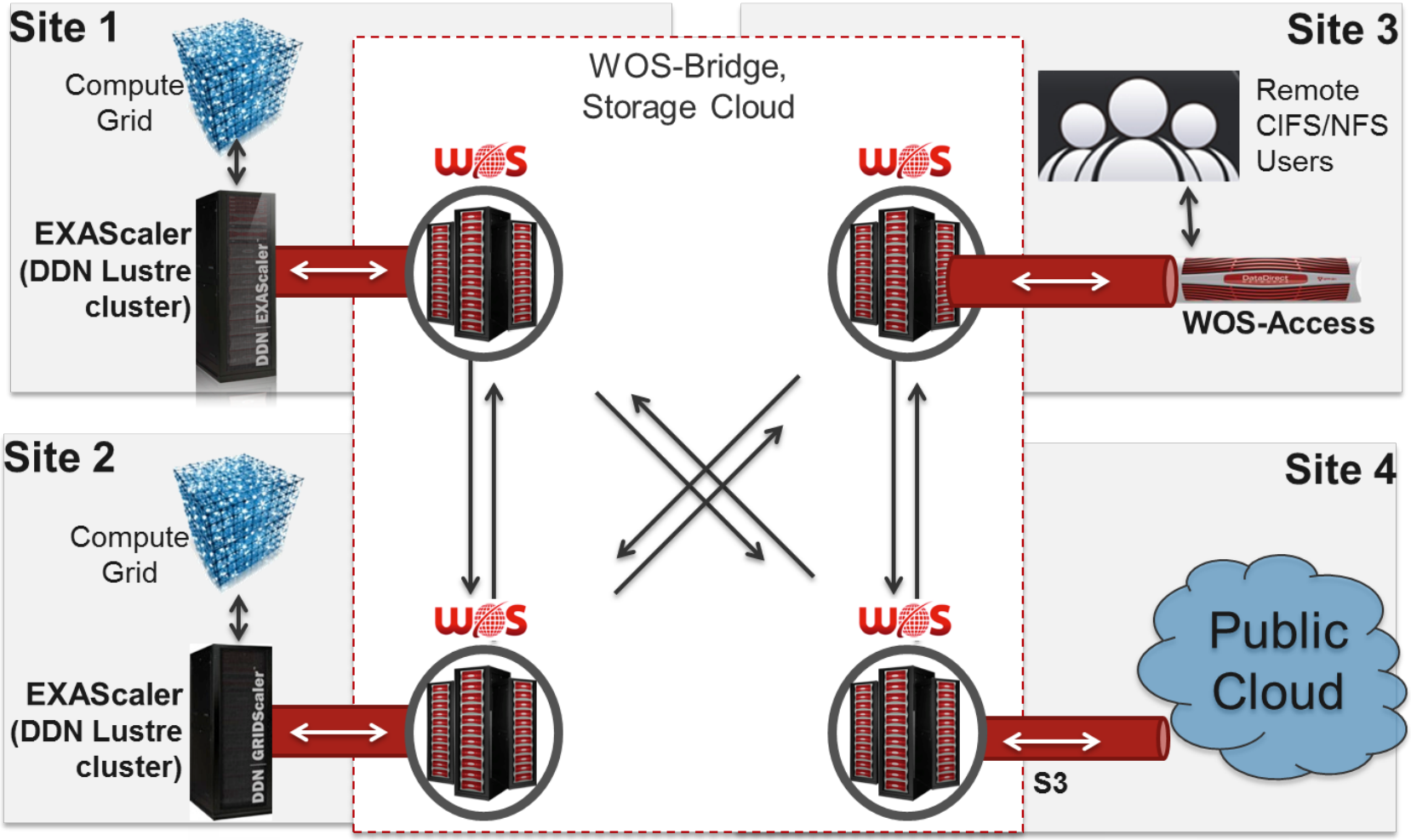
Object Disk Architecture
Formats drives with custom WOS disk file system, no Linux file I/Os, no fragmentation, fully contiguous object read and write operations for maximum disk efficiency

User Defined Metadata and Metadata Search
Applications can assign their own metadata via object storage API, WOS now also supports simple search of WOS user metadata

SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

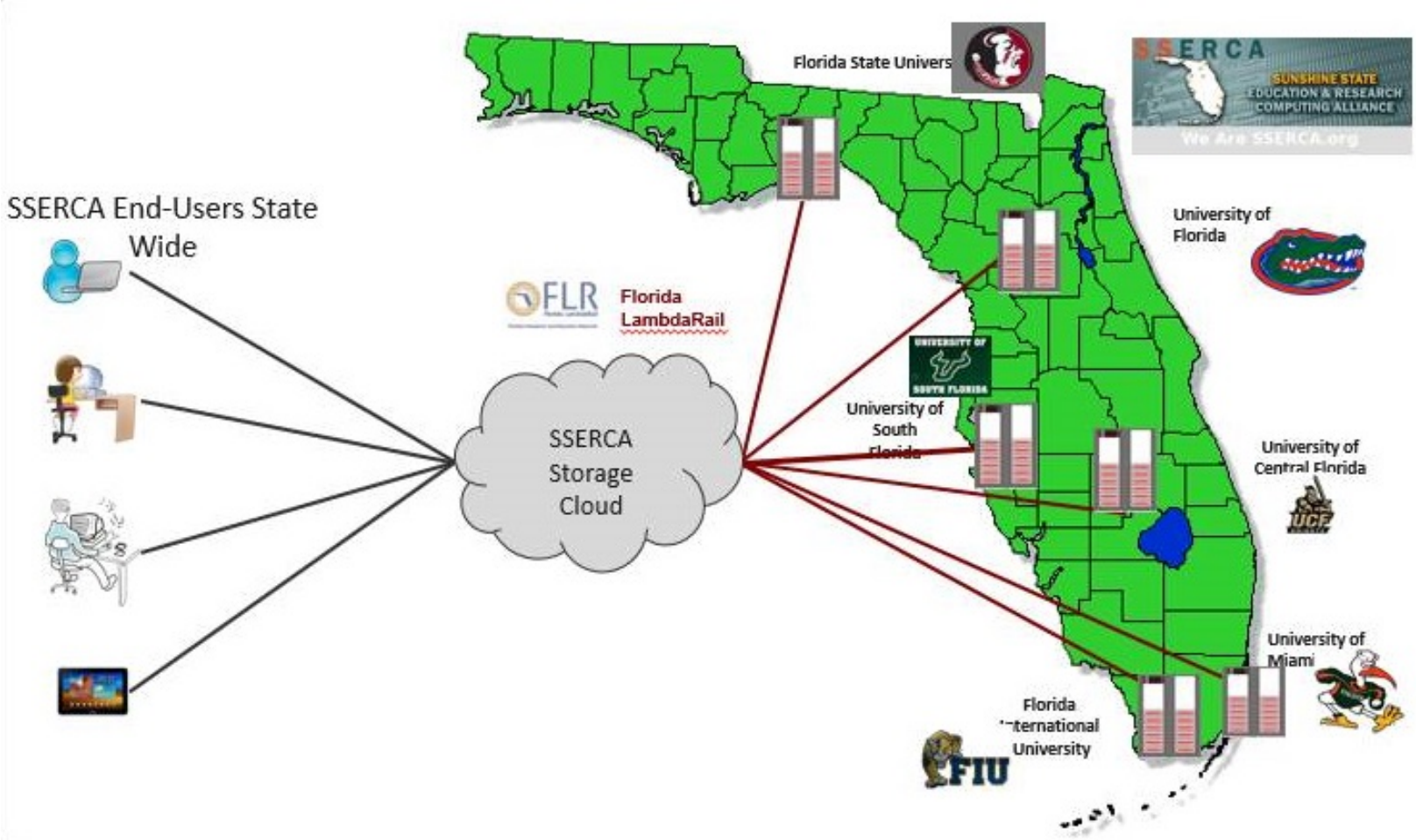
Collaborate, Distribute, Federate



SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

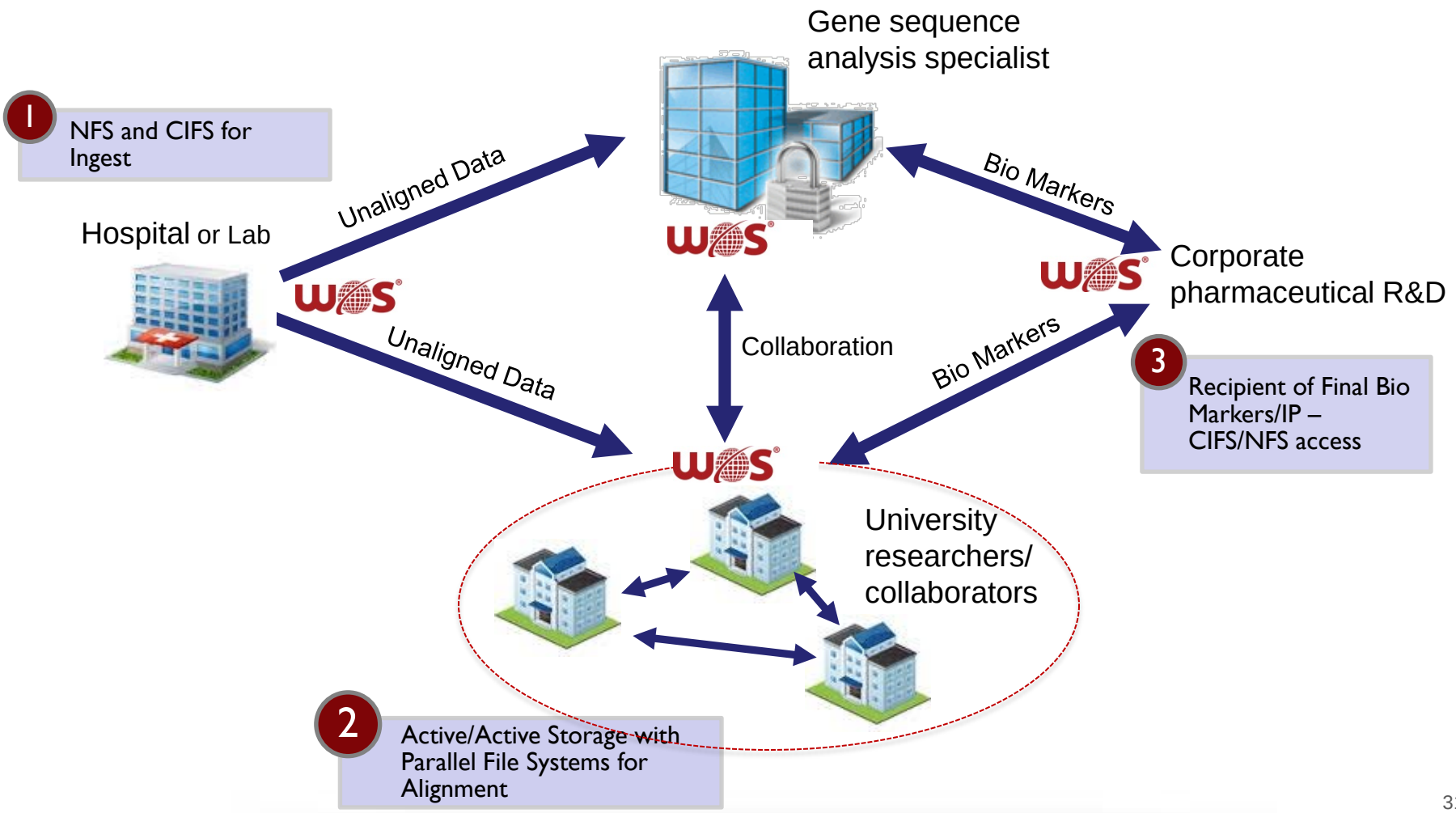
SSERCA: Cross-university Data Sharing



SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

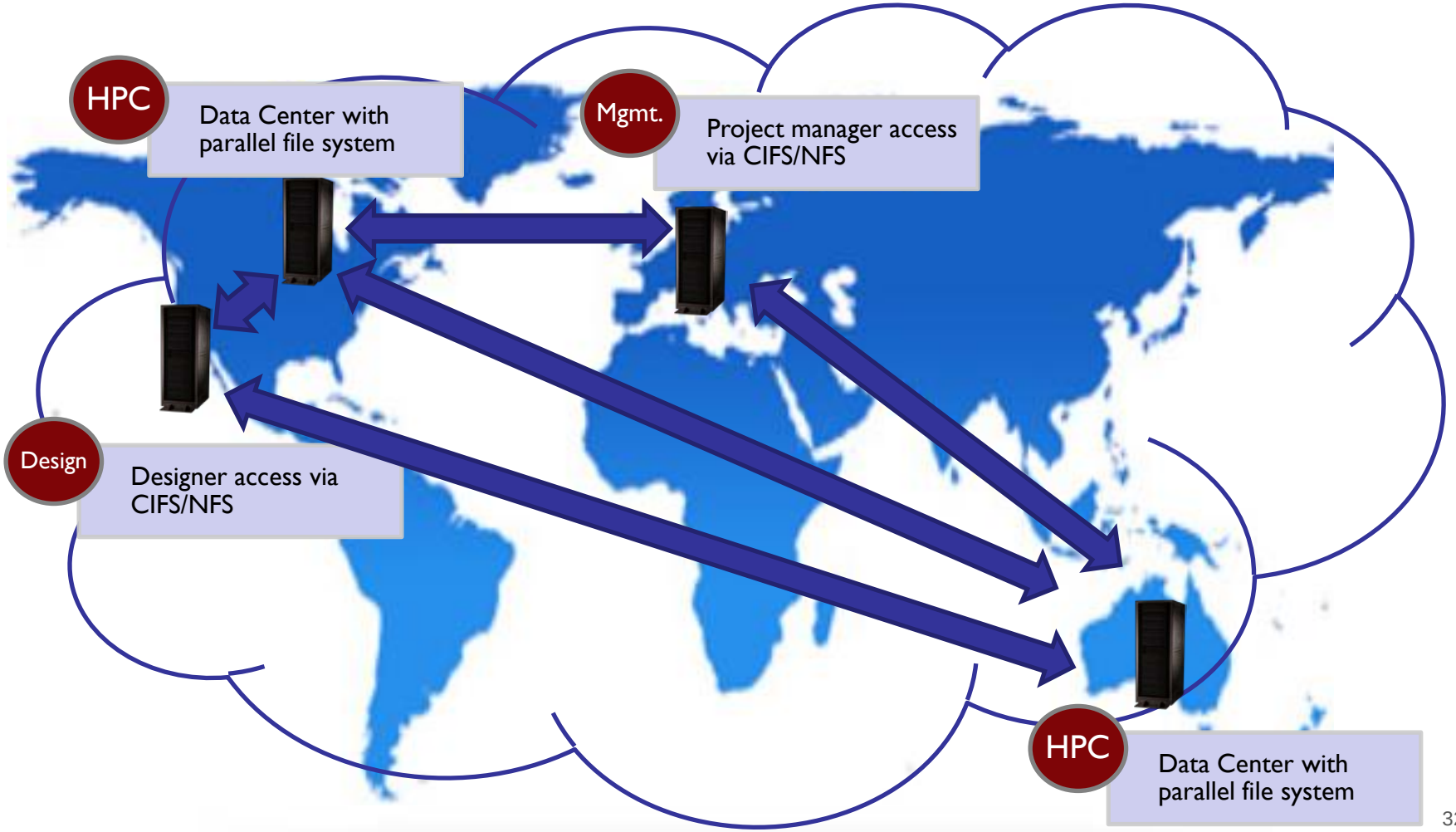
Gene Sequencing as a Service



SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Global Business Data Sharing



SDC 14

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Summary

- Lustre is open source
- Lustre has a proven ability to scale
- Lustre is incorporating new features to better support small file IO
- Lustre's reach extends from the data center to the globe

Could Lustre be the foundation for your software-defined-storage future?

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Thank You!

Keep in touch with us

sales@ddn.com

2929 Patrick Henry Drive
Santa Clara, CA 95054

@ddn_limitless

1.800.837.2298
1.818.700.4000

company/datadirect-networks

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Lustre Scalability and Performance

Feature	Current Practical Range	Tested in Production
Client Scalability	100-100000	50000+ clients, many in the 10000 to 20000 range
Client Performance	*Single client:* I/O 90% of network bandwidth *Aggregate:* 2.5 TB/sec I/O	*Single client:* 2 GB/sec I/O, 1000 metadata ops/sec *Aggregate:* 240 GB/sec I/O
OSS Scalability	*Single OSS:* 1-32 OSTs per OSS, 128TB per OST *OSS count:* 500 OSSs, with up to 4000 OSTs	*Single OSS:* 8 OSTs per OSS, 16TB per OST *OSS count:* 450 OSSs with 1000 4TB OSTs 192 OSSs with 1344 8TB OSTs
OSS Performance	*Single OSS:* 5 GB/sec *Aggregate:* 2.5 TB/sec	*Single OSS:* 2.0+ GB/sec *Aggregate:* 240 GB/sec
MDS Scalability	*Single MDT:* 4 billion files (ldiskfs), 256 trillion files (ZFS) *MDS count:* 1 primary + 1 backup Introduced in Lustre 2.4Up to 4096 MDTs and up to 4096 MDSs	*Single MDT:* 1 billion files *MDS count:* 1 primary + 1 backup
MDS Performance	35000/s create operations, 100000/s metadata stat operations	15000/s create operations, 35000/s metadata stat operations
File system Scalability	*Single File:* 2.5 PB max file size *Aggregate:* 512 PB space, 4 billion files	*Single File:* multi-TB max file size *Aggregate:* 55 PB space, 1 billion files

2014 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.