1-2013

DataDirect™ NETWORKS
INFORMATION IN MOTION™

Exploring Efficiencies in Data Reduction, Analysis, and Distribution in the Exascale Era

CERN
31 January 2013

tbeckers@ddn.com
dfellinger@ddn.com

Dave Fellinger
Chief Scientist
Office of Strategy and Technology

DDN | We Accelerate Information Insight

DataDirect NETWORKS Information in Motion

DDN provides a competitive advantage by maximizing your datacenter investment while mitigating growth challenges over your discovery process.

- **Established:** 1998
- **Revenue: >**$300M – Profitable, Fast Growth
- **Main Office:** Sunnyvale, California, USA
- **Employees:** 600+ Worldwide
- **Worldwide Presence:** 16 Countries
- **Installed Base:** 1,000+ End Customers; 50+ Countries
- **Go To Market:** Global Partners, Resellers, Direct

World-Renowned & Award-Winning

DDN TOP 500 Presence

70%	7	Of the	Top10	
65%	13	Of the	Top20	
64%	32	Of the	Top50	
61%	61	Of the	Top100	
29%	143	Of the	Top500	
over	50%	Of the	Top500	Bandwidth
over	75%	Of the	Top500	Luster Sites
over	60%	Of the	Top500	GPFS Sites

Sample HPC Partners & Customers

CRAY
IBM
sgi
Bull
DELL
hp invent

Lawrence Livermore National Laboratory
NOAA
Sandia National Laboratories
National Geospatial-Intelligence Agency
NASA
KIT Karlsruhe Institute of Technology
cea
DKRZ
CINECA
AWE
ZIH Zentrum für Informationsdienste und Hochleistungsrechnen
PGS
WesternGeco
TOTAL
ارامكو السعودية Saudi Aramco
NERSC
Berkeley Lab
ARL
NCSA
Oak Ridge National Laboratory

Accelerating Accelerators

DDN is the leading provider of affordable, high-availability storage for the next generation of particle physics research.

DDN Supplied over 30PB of Storage to the LHC Community in the last 3 years

KIT Karlsruhe Institute of Technology · CERN · Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas · SFU Simon Fraser University Thinking of the World · TRIUMF · BiG Grid the dutch e-science grid · NIKHEF · DESY · SciNet · INFN

LHC Customer Base

- Tier 0
  - CERN-LHCb (1*S2A9900➔ SFA10K, 100TB)
- Tier 1
  - SARA/NIKHEF (13*S2A9900, 6 PB)
  - KIT-2009 (10*S2A9900, 16PB)
  - KIT (SFA10K, 0.6PB)
  - IN2P3 (7*DCS9550, 1.5PB)
  - PIC (2*S2A9900, 2.4PB)
  - INFN-CNAF (5*S2A9900, 1*SFA10K, 10PB)
  - TRIUMF (2*DCS9900, 0.6PB)
- Tier 2
  - DESY (2*S2A9900, 1.2PB)
  - DESY (2*SFA10K, 1.8PB)
  - NBI (1*S2A6620, 60*2TB)
  - INFN-PISA (2*S2A9900, 1*SFA12K, 1PB)
  - INFN-PADOVA (1*S2A9900, 240TB)
  - IFCA (1*S2A9900, 1.2PB)
  - SFU (1*S2A9900, 1PB)
  - UNIV. ALBERTA (1*S2A9550, 100TB)
  - UNIV. VICTORIA (1*S2A9900, 500TB)
  - SCINET (2*S2A9900, 1PB)
  - McGill UNIV. (2*SFA10K, 1PB)

DDN I 15 Years in HPC
Investment In Scale & Innovation

DataDirect Networks

DDN Incorporated

DataDirect Networks — Information in Motion

First HPC Customer

NASA

U.S. Department of Energy

SFA Project Inception

WOS Project Inception

Argonne National Laboratory

Largest private storage co. (IDC)

CERN

500+ EMPLOYEES

1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012

Technology

S2A6000

S2A3000

S2A8000

S2A9550

S2A9900

SFA Storage Fusion Architecture

6620

10K

12K

WOS

AWARDS

NAB SHOW

HPCwire

ISC

Frost & Sullivan

Deloitte. Technology Fast 500

Excellence in Cluster Technology

STORAGE

Inc.

The 100 Fastest Growing Private Companies

Challenges of Extracting Knowledge From Data

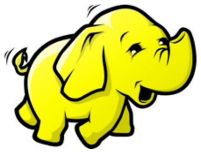
DataDirect NETWORKS Information in Motion

- Data reduction and distribution is a critical process in the feedback loop of iterative science.
  - Huge amounts of data must be captured and stored
  - Processes used to execute data conversions or reductions
  - Reduced data must be distributed and will be analyzed locally by globally distributed researchers
  - Collaborations are generally established to visualize the results
  - The entire process can then feedback required changes to the process
- Every process has inherent latencies that must be reduced or eliminated if possible.

DDN HPC Portfolio

DataDirect NETWORKS
INFORMATION IN MOTION

Apache Hadoop

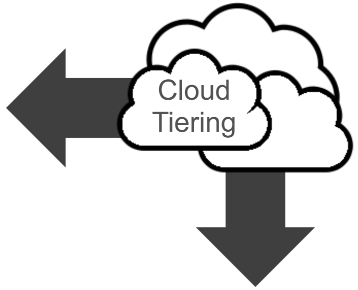


Hscaler

Parallel File Storage

EXAScaler™
10Ks of Clients
1TB/s+, HSM
NFS, CIFS

GRIDScaler™
1Ks of Clients
1TB/s+, HSM
NFS, CIFS



Cloud Storage

WOS® 2.5
256 Billion Objects
GeoReplicated
Cloud Foundation
Mobile Cloud Access

DirectMon
Enterprise Platform Management

Storage Fusion Architecture Storage Appliances

7700

10GB/s, 600K IOPS
60 Drives in 4U; 396 Drives in 20U
Embedded Computing (tba)

12K

40GB/s/1.7M IOPS
1,680 Drives: 2 Racks
Embedded Computing

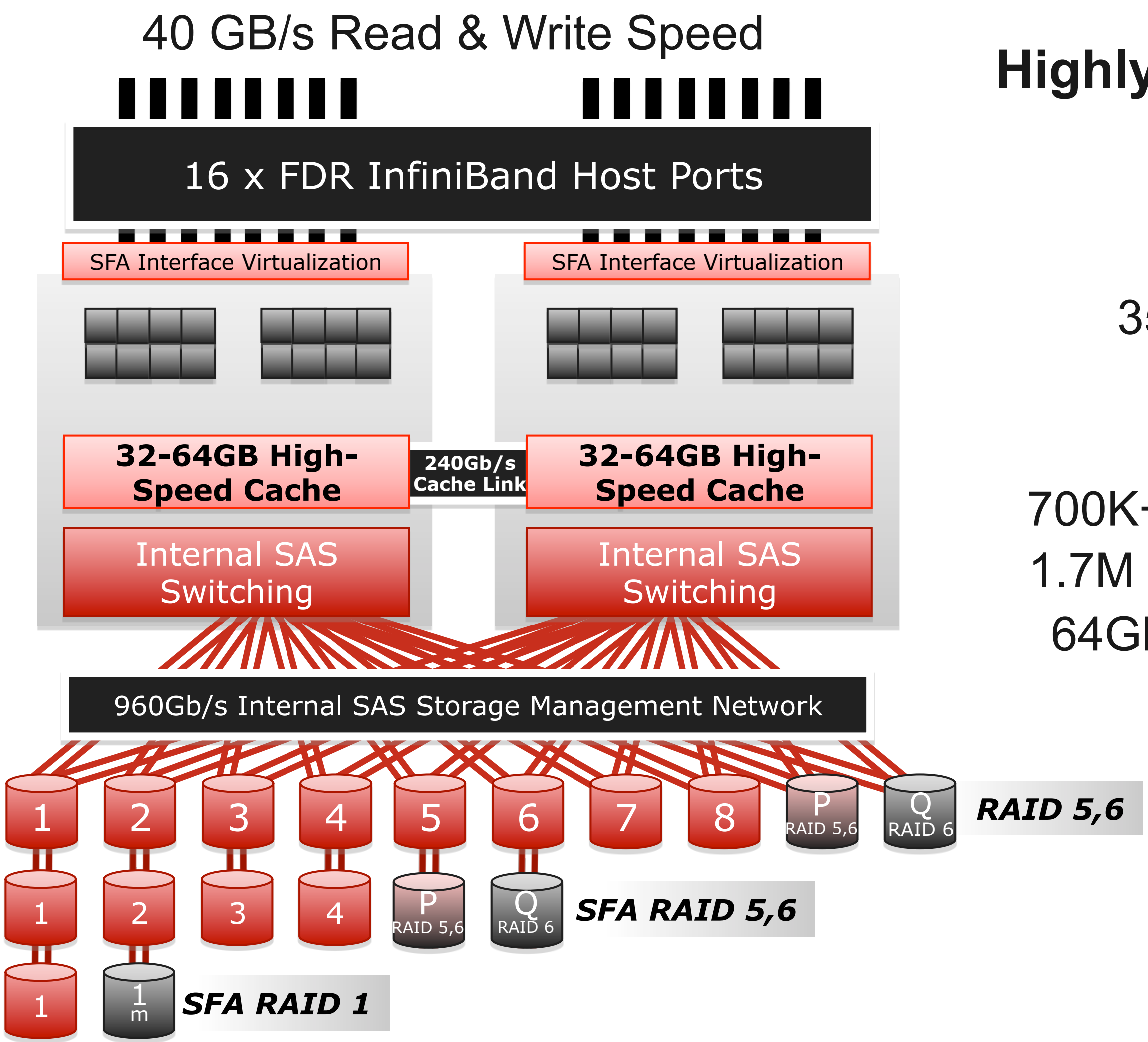
Storage Fusion Xcelerator (SFX) Flash Acceleration

- SFX Read
- SFX Write
- SFX Context Commit
- SFX Instant Commit

Flexible Media Configuration

SAS SATA SSD

SFA12K-40 (Block Appliance)



Highly Parallelized SFA Storage Processing Engine

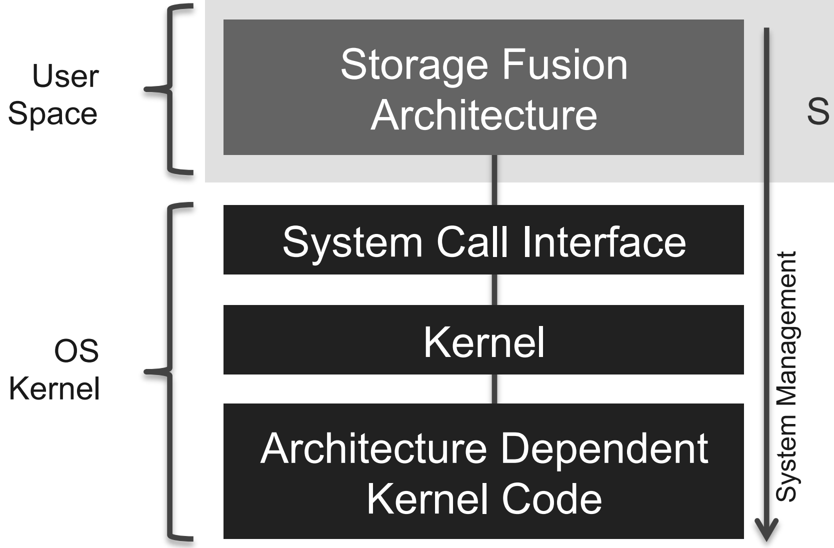
Active/Active Storage Design
35-40GB/s Read & Write Speed
Up to 6.7PB of Disk
2.4+ Million Burst IOPS
700K+ Random Spinning Disk IOPS
1.7M Sustained Random SSD IOPS
64GB+ Mirrored Cache (Protected)
RAID 1/5/6
Intelligent Block Striping
DirectProtect™
GUI, SNMP, CLI, API
16 x FDR IB Host-Ports
8RU Height

SFA User Space Storage OS

A Scalable Storage OS Implemented In User Space

SFA Implementation

Tight Integration W/ Virtualization is Ideal for Converged Storage



SFA User Space Storage Operating System

SFA Boots off Linux Kernel & Takes Control of System

SFA Device Drivers & Scheduler Handle All I/O in User Space

SFA State Machine is Parallel and Portable in Design

Benefits:

- System is highly efficiency due to elimination of context switching and high-speed routing infrastructure.
- User space implementation is easier to develop than kernel-level.
- DDN back/front-end drivers streamline IO
- Easy to port SFA OS to other platforms

Barriers to Entry:

- Full storage OS is 1M+ lines of tightly integrated big data optimized code
- Implementation is very sophisticated – requiring integrated drivers and I/O layers

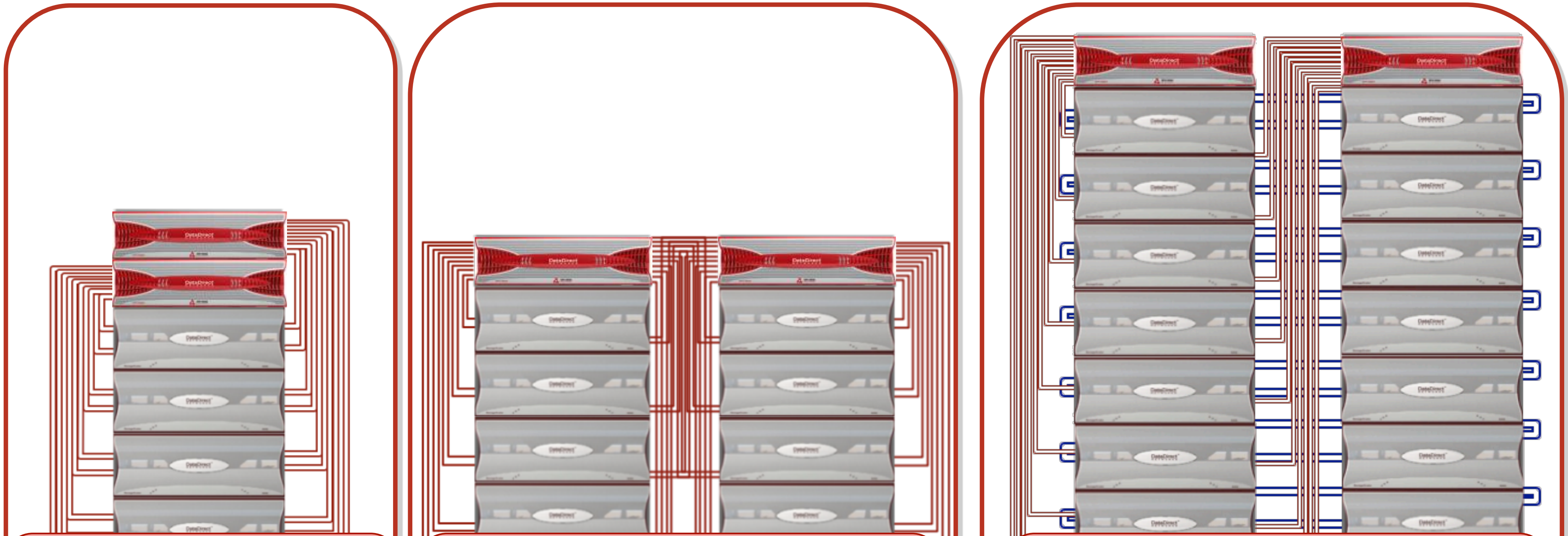
All of the benefits of an in-kernel implementation, none of the limitations of kernel/HW dependency.

SS8460 – Highest Density Enclosure

84 Drives – SSD, SAS, SATA - in 4 rack units

Up to 336 TB

Variable System Size



5 Enclosure System
Up to 420 Drives
2 BBUs, 28U

10 Enclosure System
Up to 840 Drives
2 BBUs, 48U

20 Enclosure System
Up to 1,680 Drives
2 BBUs, 88U

High Availability Drive Channel & Enclosure RAIDing

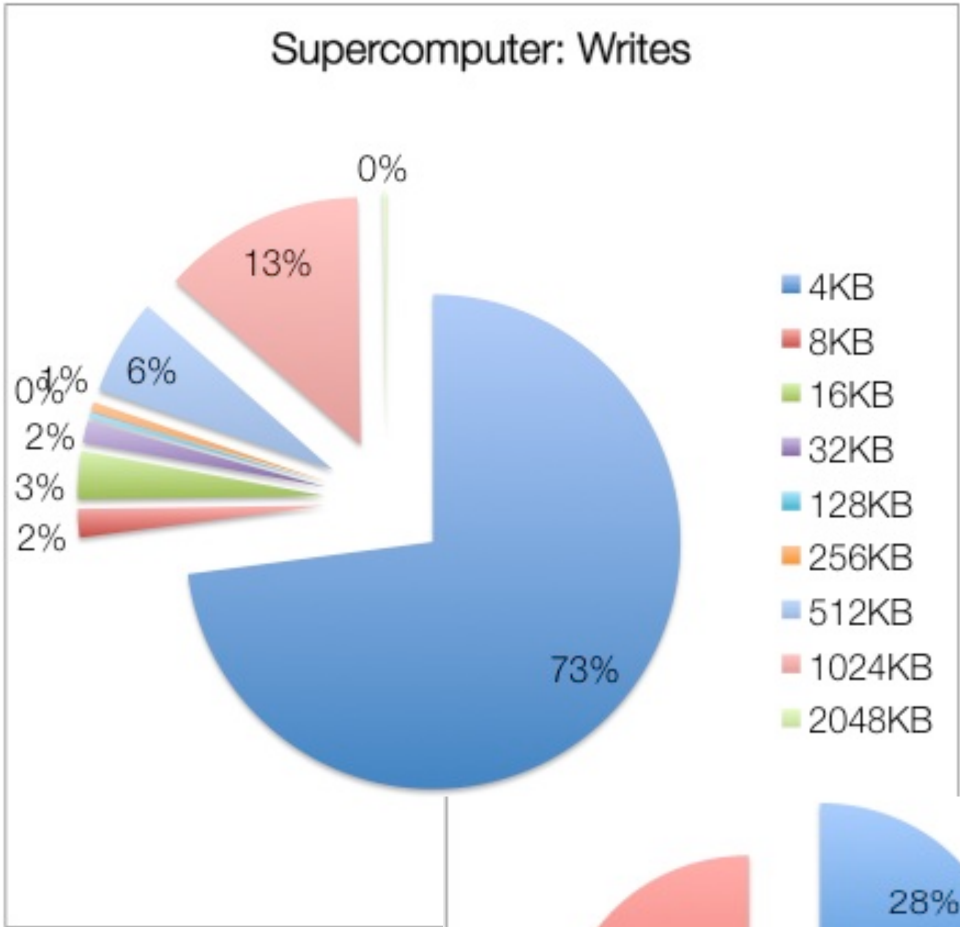
SKA 12Ke Embedded Processing

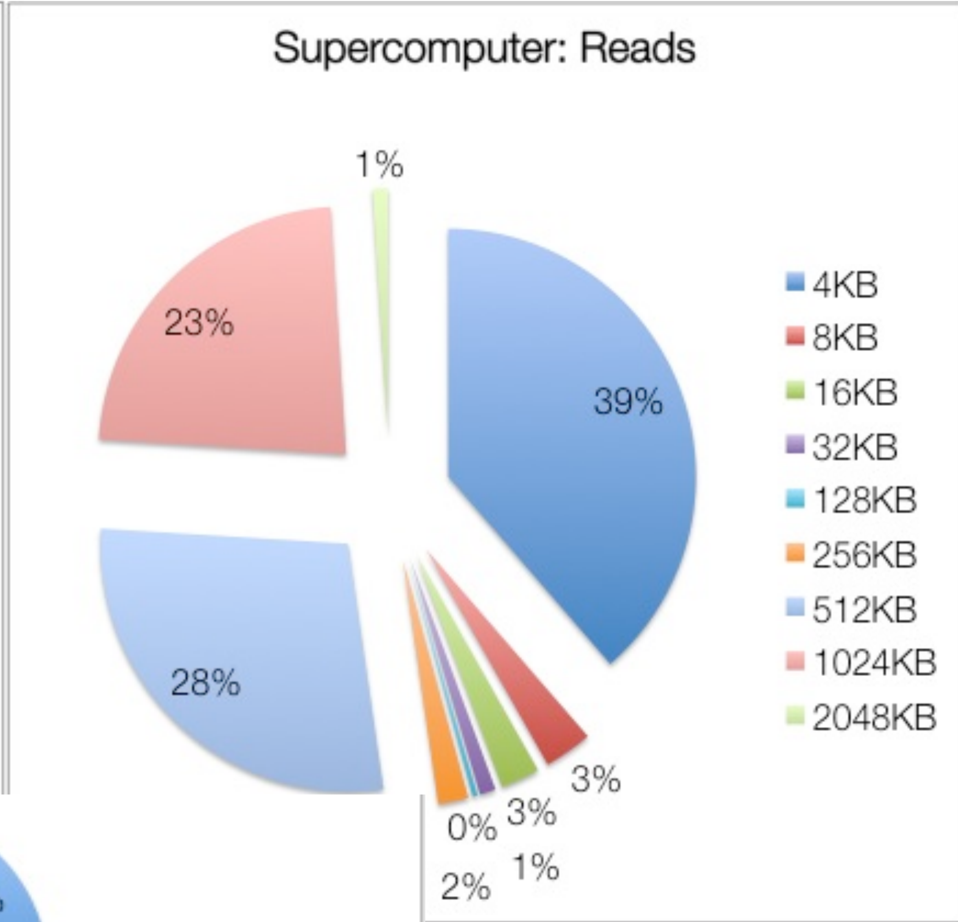
System Level Sources of Latency

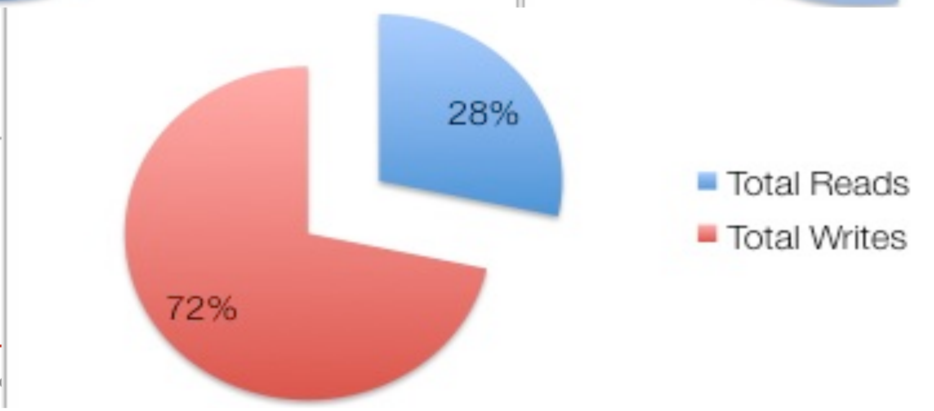
- Hardware Chain
  - Disk drive servo operation
  - Multiple SCSI layers
  - Multiple bus transitions
  - Memory bandwidth limitations
  - Network service latencies
- Software Chain
  - Memory copies
  - Kernel operations
  - Layers of consecutive operations including the service of V-nodes, I-nodes and FAT
  - Serial data transport processes

HPC Data Patterns

DataDirect NETWORKS
Information in Motion

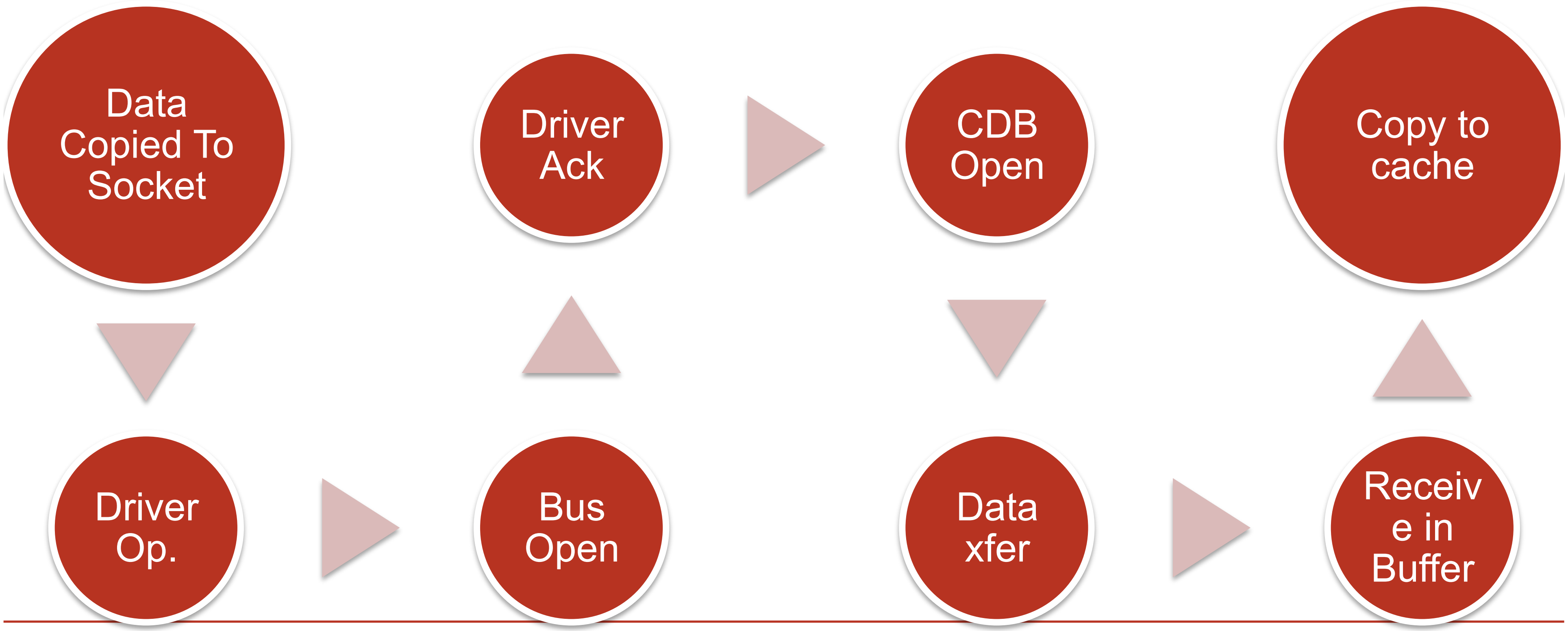






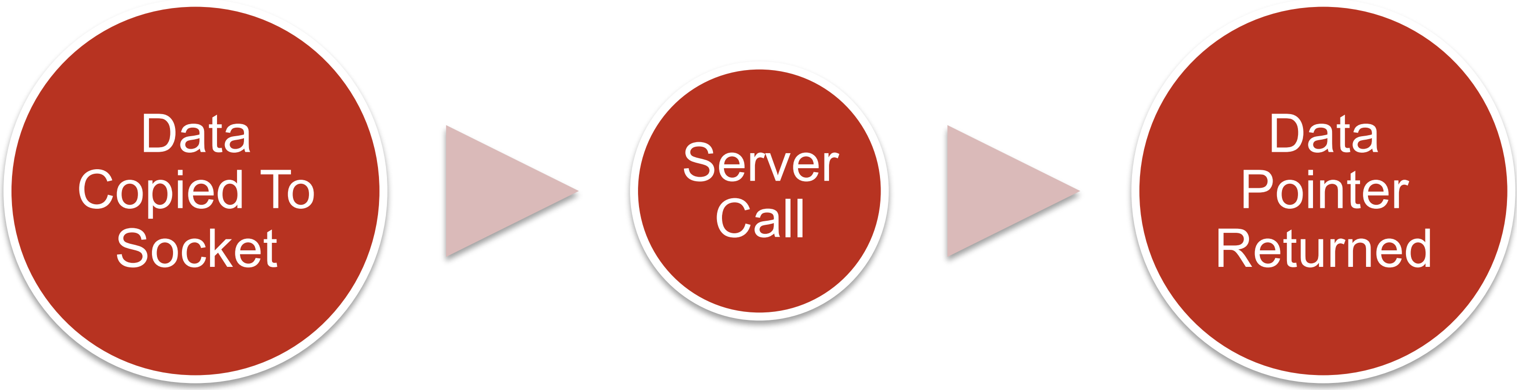
"Traditional" File Access

- File servers are connected to storage devices by serializing devices such as HBAs or HCAs
- Multiple steps executed to move data to/ from a server;



Efficient Alternative File Access

- File servers are run as virtual machines within the storage system in a shared memory environment with the storage cache
- The steps to move data to or from the storage;



“Traditional” Data Reduction

- Data reduction, manipulation, filtering, resolution shifting, etc. is done in an external network connected processor.
- The steps required to execute a process must first include moving data to the processor;
    1. The file server builds a front end socket
    2. A routine is called that appends headers and footers to data packets so that a TCP transport layer can move the data from the server to the processor
    3. The processor receives the frames from a switch, strips the headers and footers, reorders the packets if required, and places the data in user space for manipulation

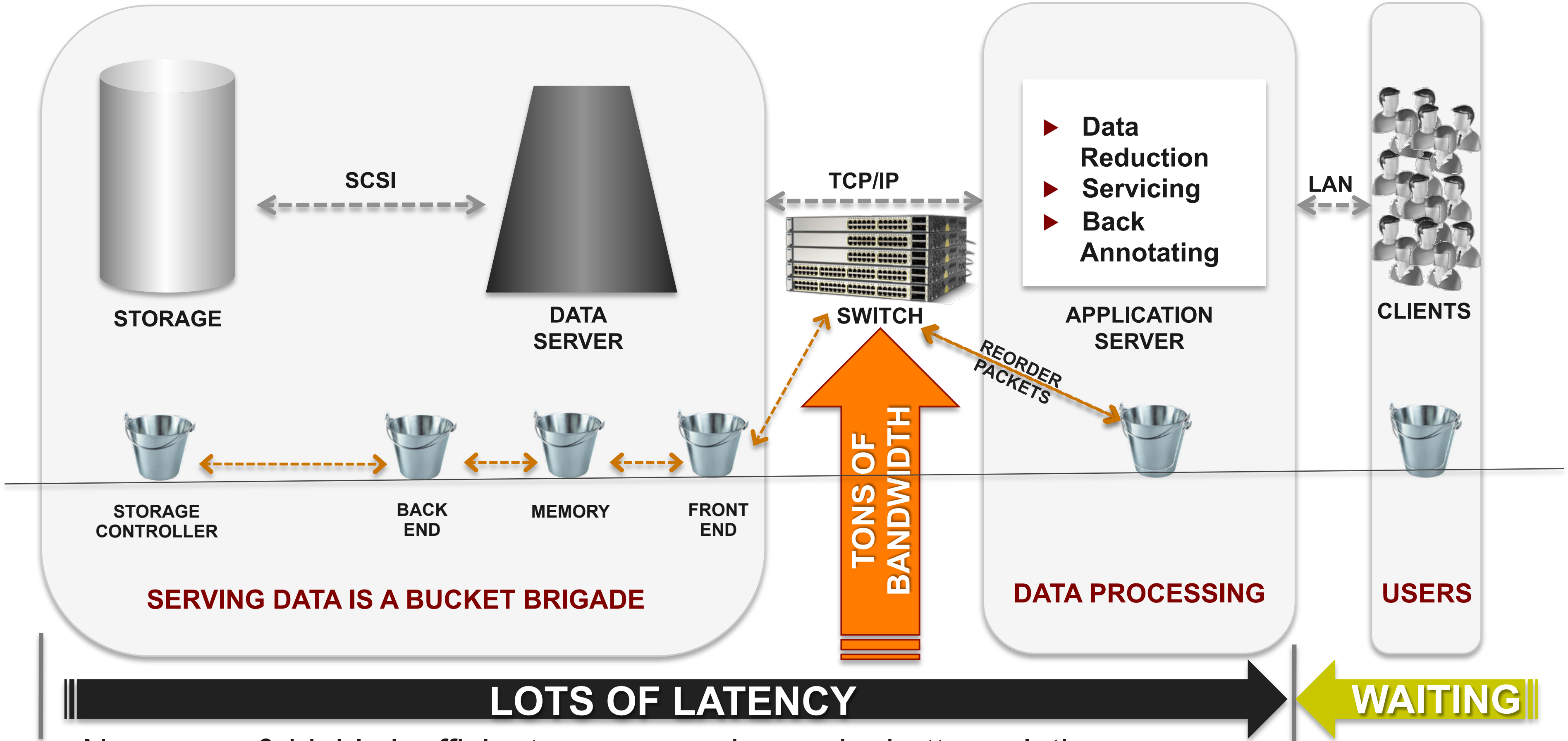
Efficient Alternative Data Reduction

- An image including the file system and the data reduction process is run as a virtual machine on the storage system in multiple cores with dedicated cache.
- Steps to obtaining reduced data are executed removing both the SCSI bus transaction and the TCP/IP layer;
  - The user requests data through the process
  - Reduced data is transported through the TCP layer
- An alternative is to start a scheduled process on the storage that processes data from a "raw" data LUN to a LUN containing processed data.

i·R·O·D·S
Integrated Rule-Oriented Data System

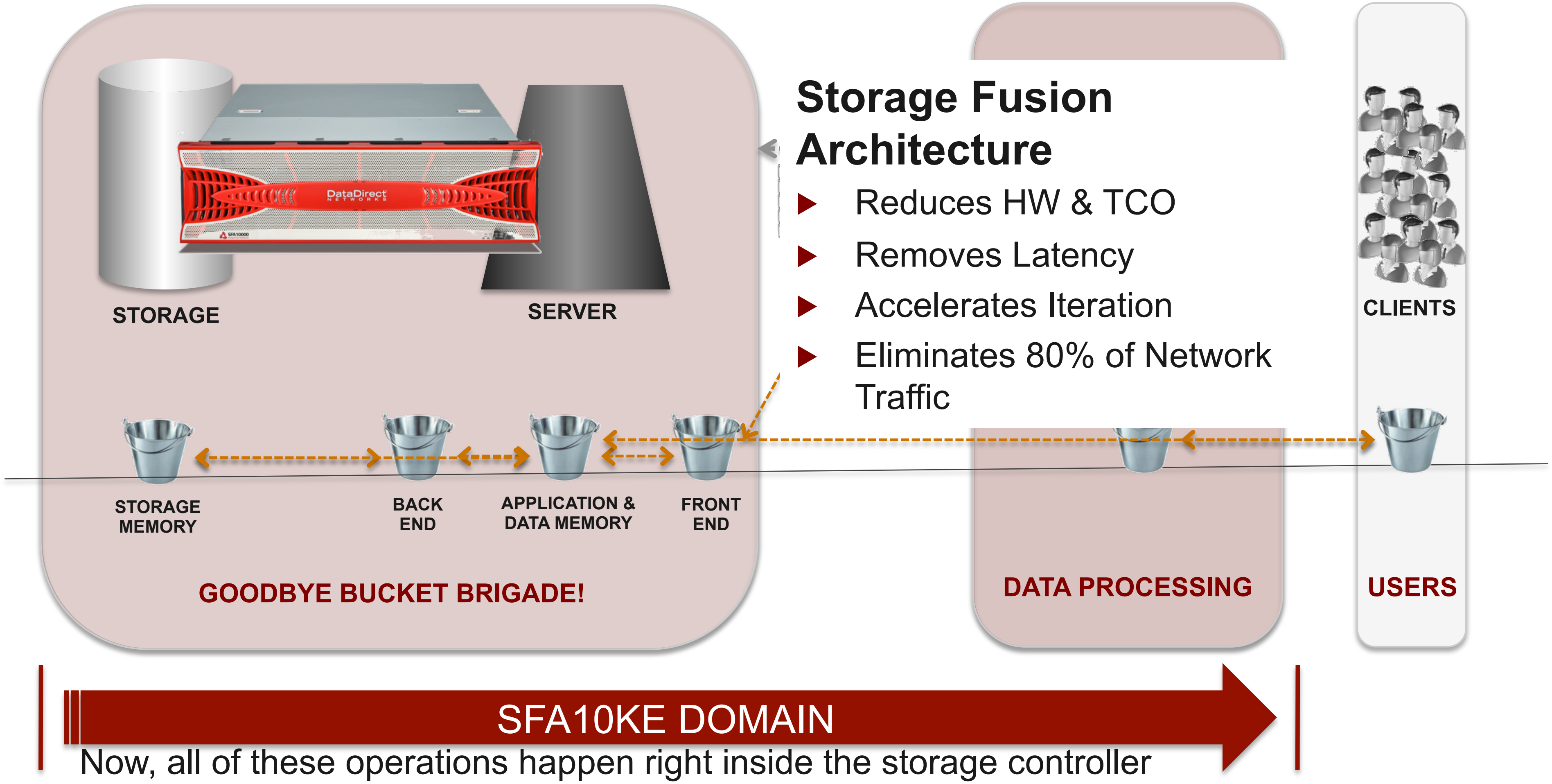
DDN has partnered with RENCI & Dice to simplify & accelerate data reduction with DDN in-storage processing with iRODS.

Inefficient Bucket Brigade of Protocols



Numerous & highly inefficient processes demand a better solution

Reduction of Latency and Network Traffic



Storage Fusion Architecture

- Reduces HW & TCO
- Removes Latency
- Accelerates Iteration
- Eliminates 80% of Network Traffic

SFA12K™ | Models

DataDirect NETWORKS
INFORMATION IN MOTION

	SFA12K-20	SFA12K-20E	SFA12K-40
Maximum Drives	1,680[1]	1,680[1]	1,680[1]
System Interface	FDR IB 16Gb FC[2]	FDR IB 10/40GbE	FDR IB 16Gb FC[2]
Drive Types	3.5" & 2.5" SSD, SAS & SATA (inter-mixable)		
System Capacity	6.72PB (w/ 4TB HDDs)[1]		
Bandwidth	20GB/s (raw I/O)	20GB/s (file I/O)	40GB/s (raw I/O)
Cache IOPS	850K	850K	1.7M
Flash IOPS	700K	700K	1.4M
In-Storage Processing™	N/A	Yes. ExaScaler, GridScaler Customer Provided	N/A

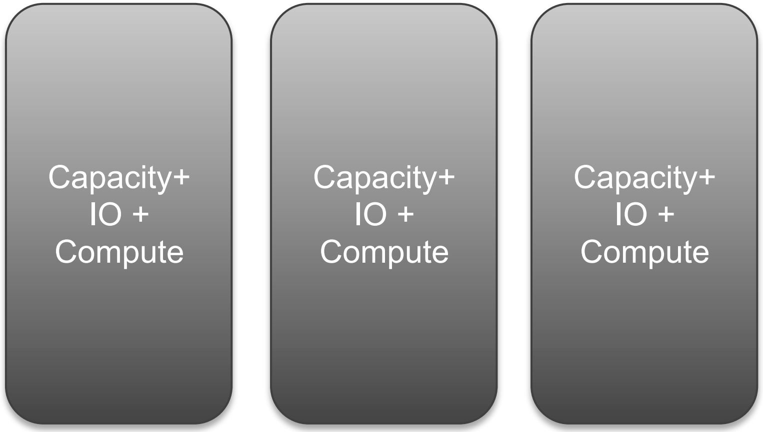
[1] 840 Drives Until Dec12
[2] 16Gb FC available Q1'13

hScaler

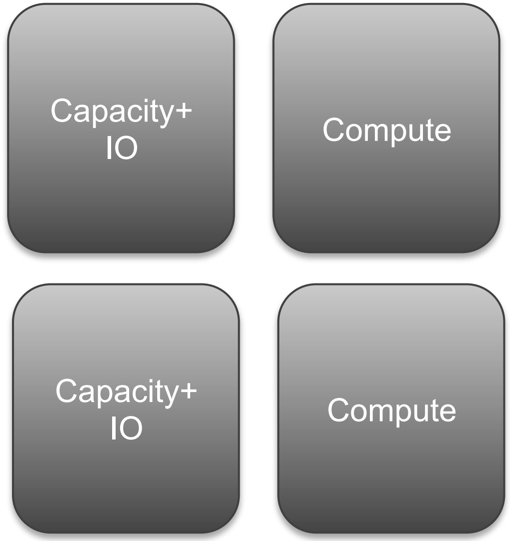
Common Hadoop Architecture

DataDirect NETWORKS Information in Motion

Hadoop Model



DDN Model



Pros

- Low entry cost

Cons

- High operations overhead
- No optimized performance, broad range

Pros

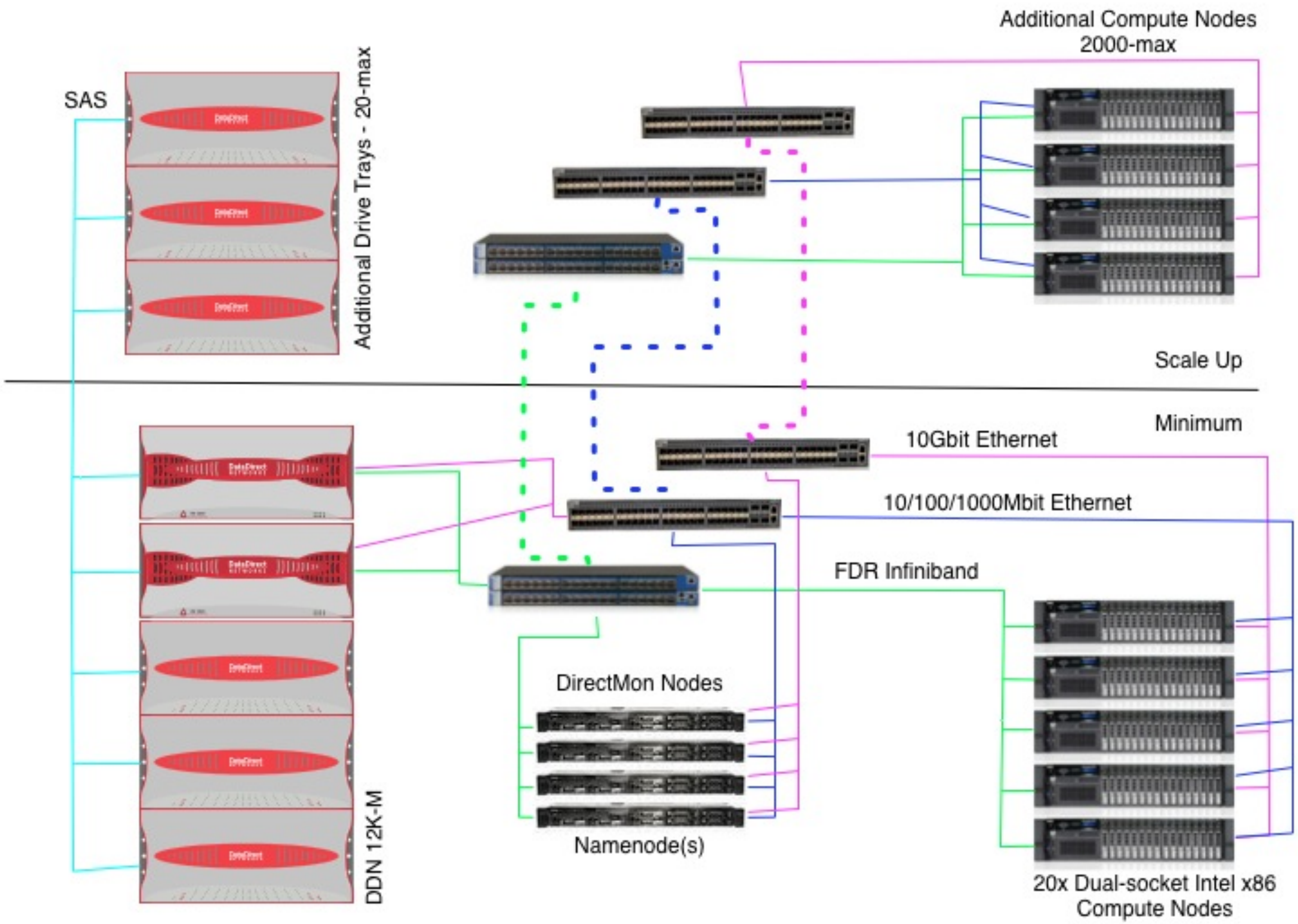
- Independent Scaling
- Lower TCO
- Software Support Model
- Optimized Performance

Cons

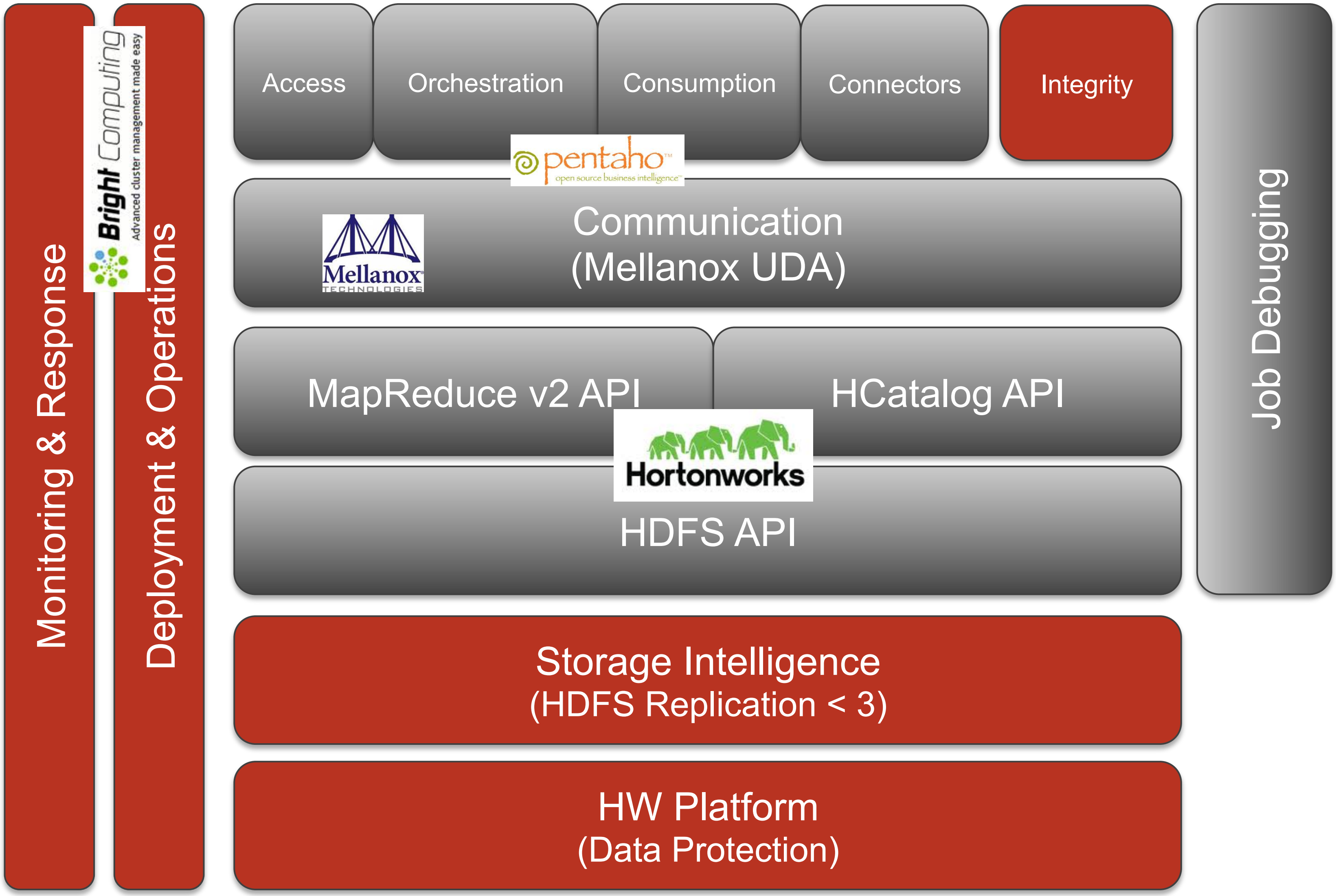
- High Entry Cost

hScaler Architecture

DataDirect NETWORKS
INFORMATION IN MOTION



Technology Ecosystem & DDN IP hScaler Product Delivery



Throughput Performance

GOAL = 40% performance gain   Comparing bare metal to hScaler

TestDFSIO is a Distributed i/o benchmark tool which write and read data from HDFS. If is the preferred tool which will validate our hScaler performance value.

Minimum performance increase was observed on small files:

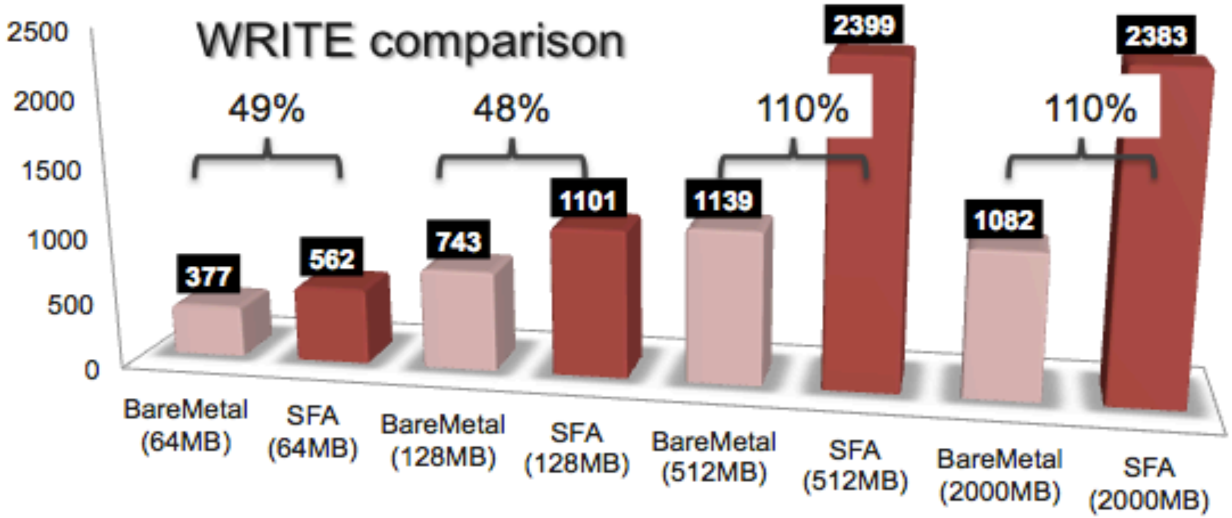
- **WRITE = +48%**
- **READ = +54%**

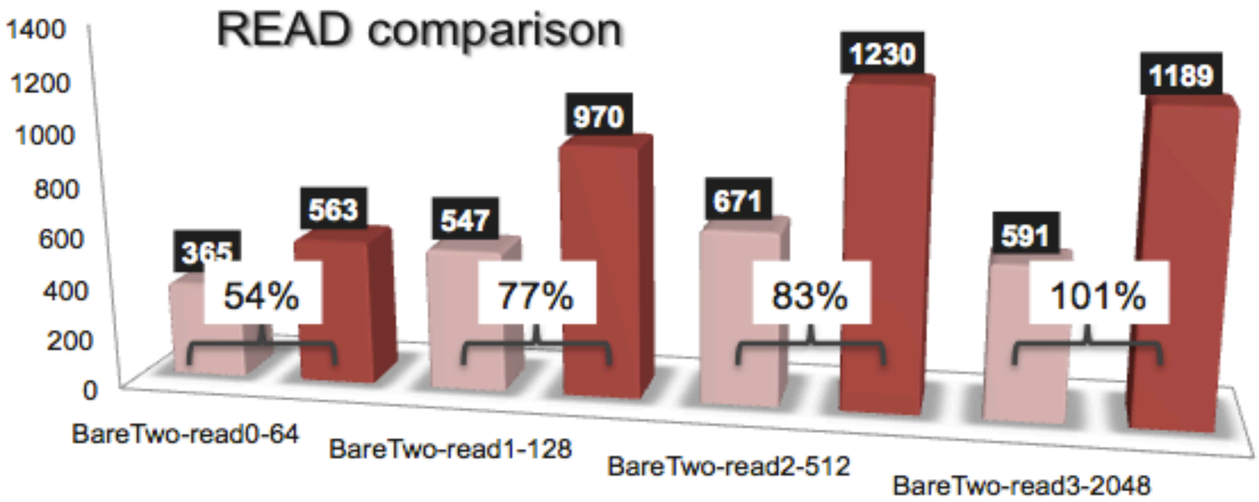
1.5X Increase

Maximum performance increase is noticeable using large files

- **WRITE = +120%**
- **READ = +101%**

2.0X Increase





DataDirect™
NETWORKS
INFORMATION IN MOTION™

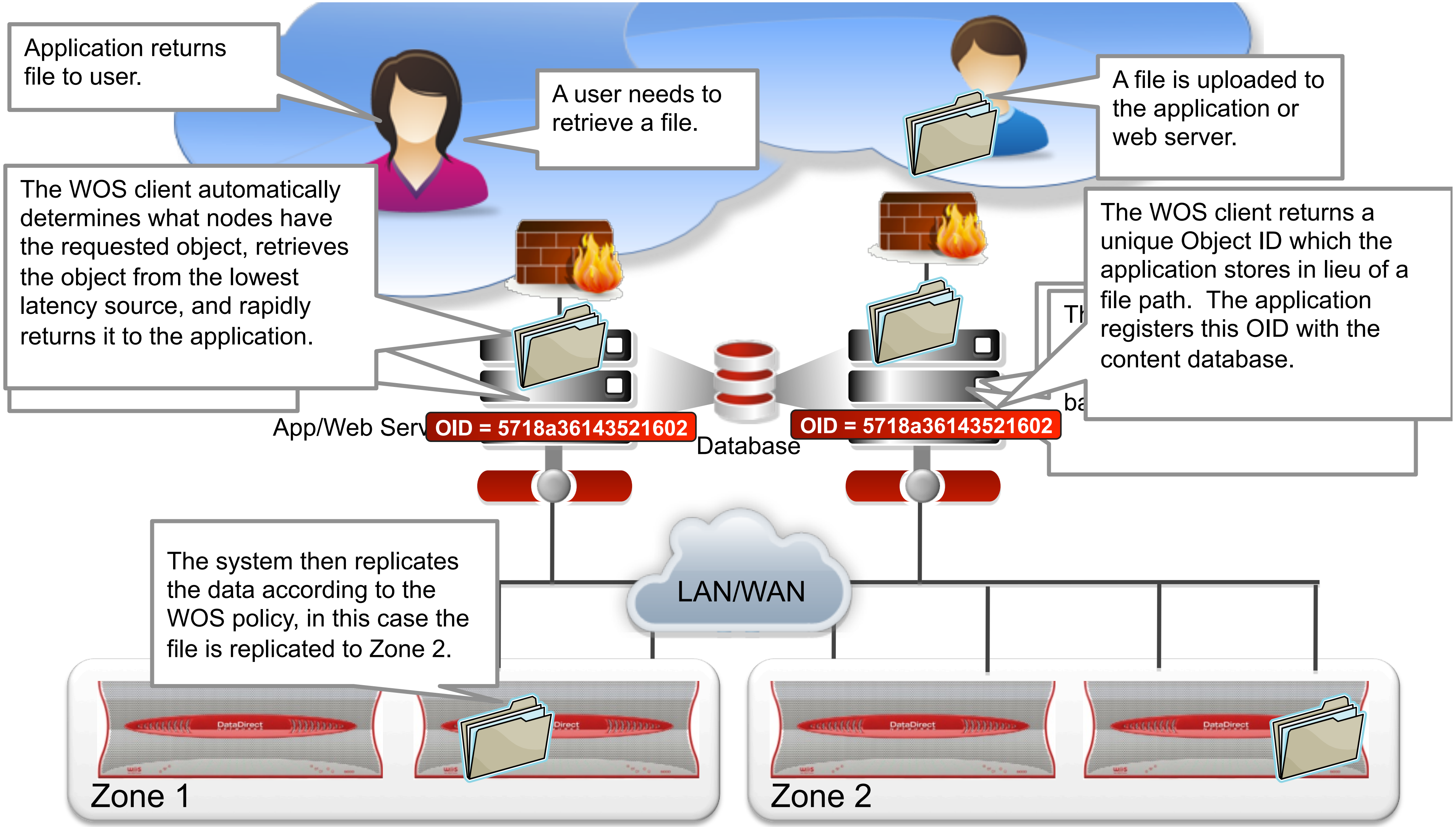
WOS

DDN | Hyperscale Initiative

DataDirect NETWORKS
Information in Motion

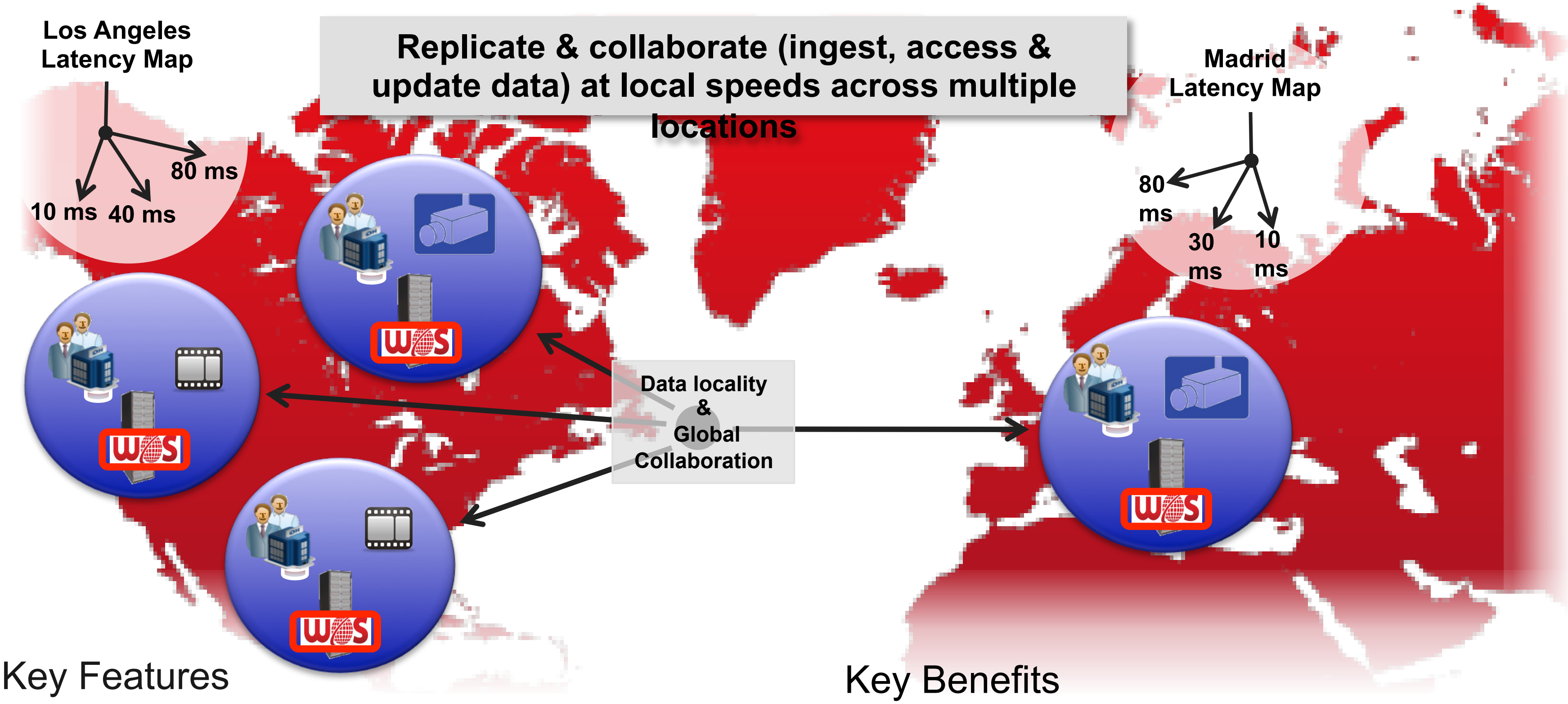
- Understand the data usage model in a collaborative environment where data is shared and studied
- A simplified data access system
- Eliminates the concept of FAT, extent lists to maximize efficiency
- Reduce the instruction set to only PUT, GET, & DELETE
- Add the concept of locality based on latency to data and load balance
- Abandons storage convention entirely

WOS Puts & Gets



Distributed Hyperscale Collaborative Storage

DataDirect NETWORKS
INFORMATION IN MOTION



Key Features

- ▶Asynchronous or Synchronous Replication across up to 4 sites
- ▶Geographic, location, & latency intelligence
- ▶NAS data access @ LAN speeds
- ▶Data and DR protected

Key Benefits

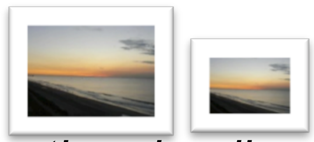
- Users can access and update data simultaneously across multiple sites
- Increases performance & optimizes access latency
- No risk of data loss

Intelligent WOS Objects

DataDirect NETWORKS
INFORMATION IN MOTION

Sample Object ID (OID): `ACuoBKmWW3Uw1W2TmVYthA`

WOS Signature	A random 64-bit key to prevent unauthorized access to WOS objects
WOS Policy	`Eg. Replicate Twice; Zone 1 & 3`
WOS Checksum	Robust 64 bit checksum to verify data integrity during every read.
User Metadata Key Value or Binary	`Object = Photo` `Tag = Beach`
Full File or Sub-Object	



thumbnails

Why is WOS so *Fast*?

DataDirect NETWORKS
INFORMATION IN MOTION

Traditional storage does a lot of work to write/read data

- Expends excess disk operations
  - 5-12 Disk Operations per File Read
- Multiple levels of translation and communication
  - Metadata lookups and directory travelling
  - Extent list fetches
  - RAID & block operations

WOS delivers performance through simplicity

- None of the constructs of traditional systems
- Single-Disk-Operation Reads, Dual-Operation Writes
- Reduced latency from SATA Disks since seeks are minimized
- Millions of file/ops per second with ¼ of the disks

High Resiliency

WOS provides highest data availability

- Self healing - WOS automatically corrects disk failures & data corruption problems
- Data Locality - If the closest instance of an object isn't available, WOS will automatically & transparently return the next closest instance

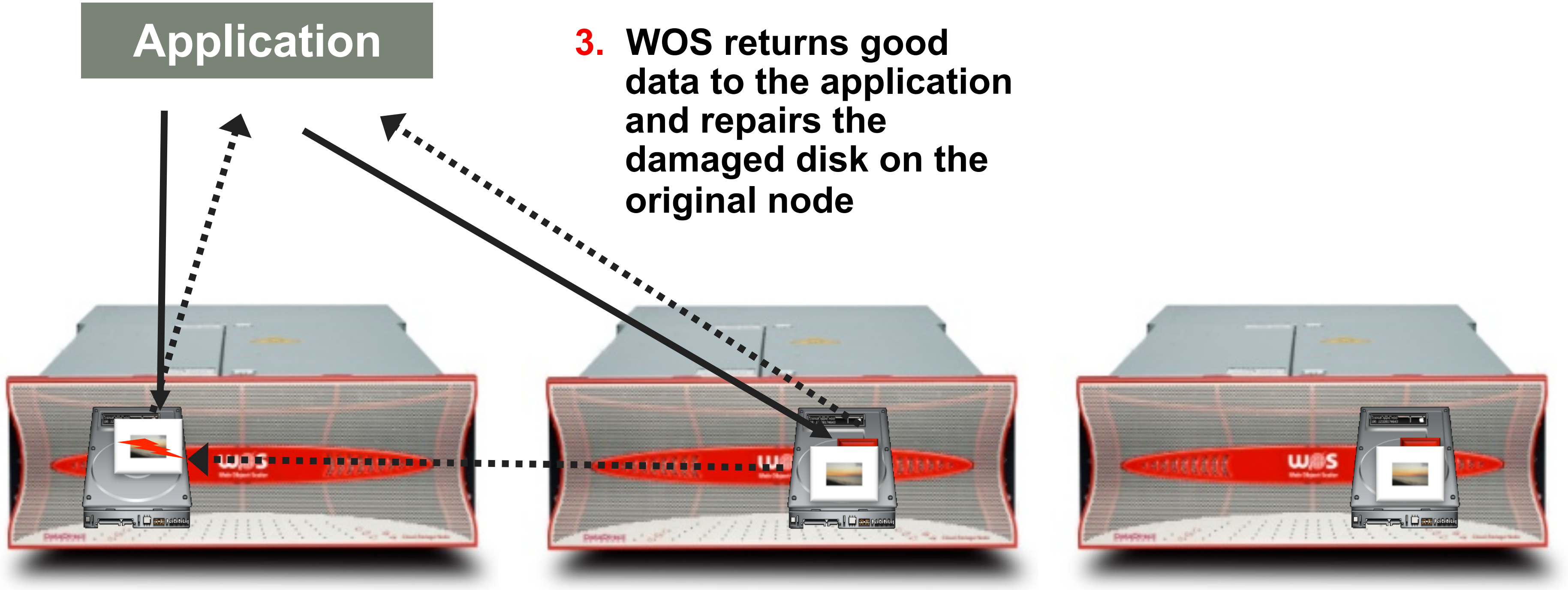
Instantaneous recovery from disk failure, not days

Built in data integrity, no silent data corruption

Replication built in, not added on

WOS Self Healing Example

1. **WOS reads an Object. Checksum shows that data is corrupt**
2. **WOS tries another node. Checksum indicates that the Object is good**
3. **WOS returns good data to the application and repairs the damaged disk on the original node**



Intelligent Fail in Place Architecture

Policy: Object replicated 2 times (3 total)

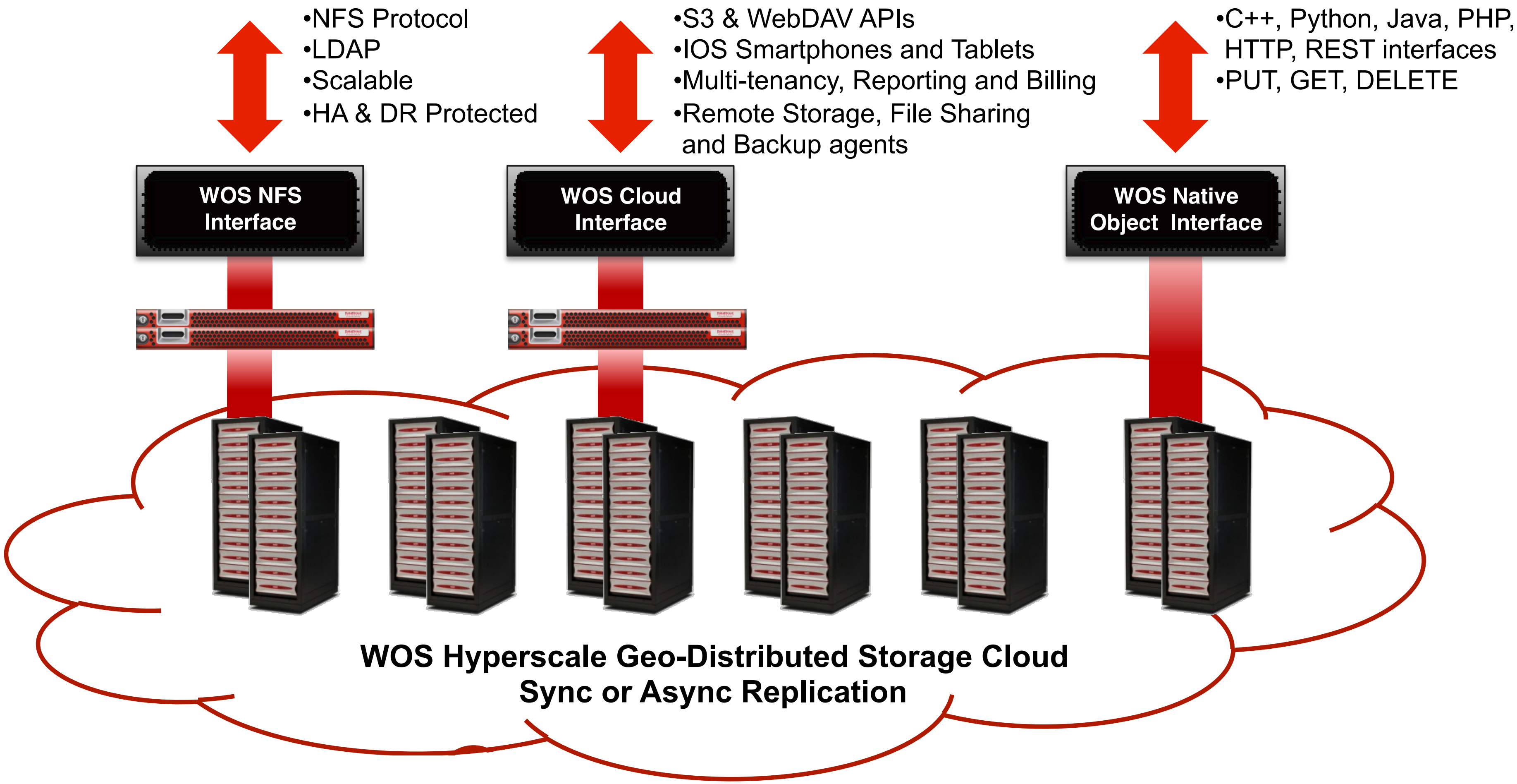
Disk Failure

Another Failure

WOS automatically recognizes and rectifies this state and brings the system back into policy compliance by replicating the object again

Node 1 Node 2 Node 3

WOS 2.5 Provides Enhanced Access Through Standard Interfaces



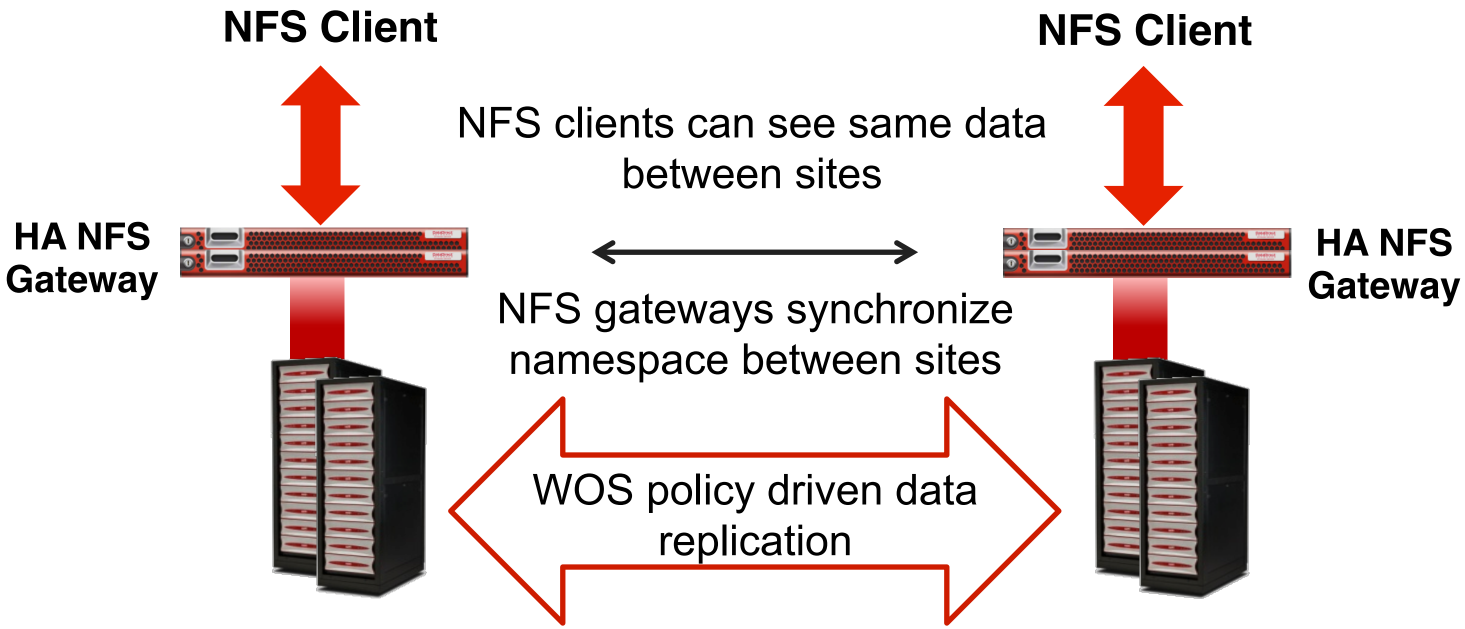
WOS Access NFS

- WOS Access NFS is a bundled software/server solution that supports NFS V3 and V4
- NFS Access Control Lists (ACLs), group/user levels, and identity authorization
  - Authorization is per file/directory
- Multiple namespaces and mount points
- Synchronizes NFS Gateways across multiple sites
  - Single Federated NFS Namespace to 23PB
- Local read/write cache
- HA protected
  - Active/passive failover
- NFS Database DR backup to WOS
- Intuitive Graphical User Interface (GUI)
- Available as software only
- Has been installed and running at a custome

WOS Access NFS GUI

WOS Access NFS

Geographically Distributed Single NFS Name Space



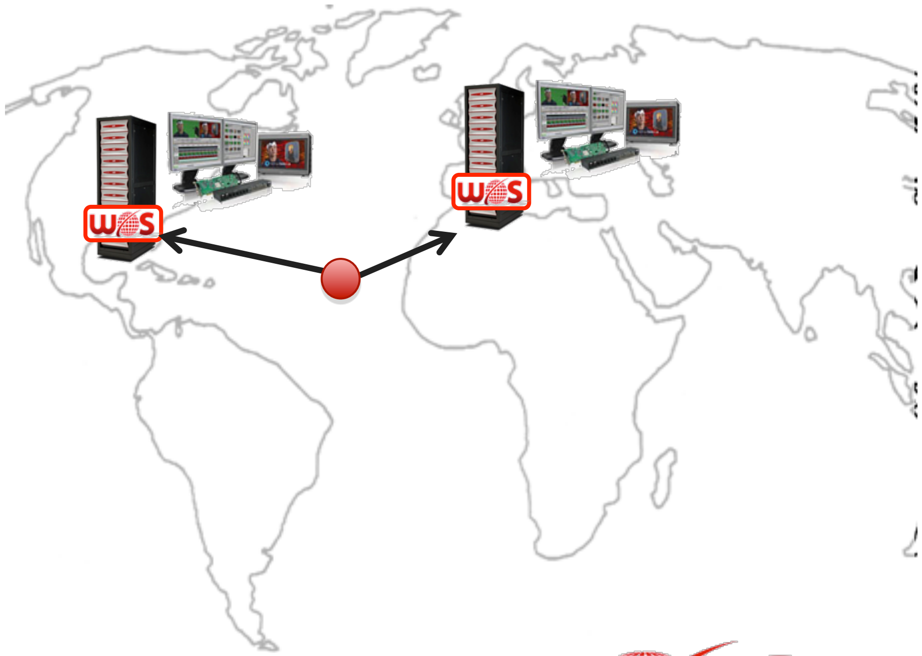
- WOS Access NFS will have significant manageability, robustness and TCO advantages v. other solutions when there is
  - Immutable, unstructured data at scales > 2PB
  - Multisite access and/or disaster recovery requirements

Globally Distributed Organizations Can Collaborate using WOS

DataDirect NETWORKS
Information in Motion

Access and Update Data Simultaneously Across Multiple Sites

- Multiple copies of data can be replicated globally for both disaster recovery and low latency access
- Everyone, at every location, has immediate access to the latest versions of the project
- Enabling globally distributed users to collaborate as part of a powerful workflow
- Speeding discovery and time to market
- Managed as a single entity, lowering IT costs

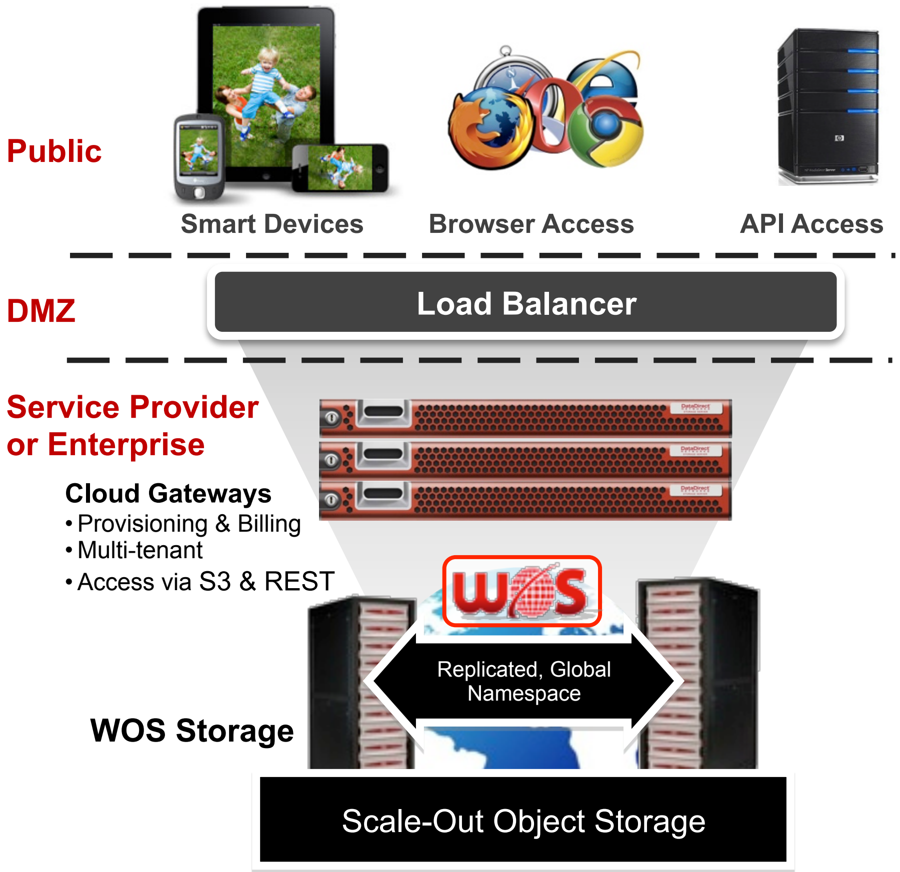


WOS Cloud

Efficiently build hyperscale storage for Public and Private Clouds

- **Offer industry leading, differentiated service**
  - Better service delivery, support, performance, cost, robustness, and SLAs
  - Flexible, pay as you grow scale
  - Remote manageability with no physical access

- **Cloud Platform Software**
  - Multi-tenancy support
  - S3 compatible & WebDAV APIs
  - Full CDMI Compliance
  - Integrates w/ existing provisioning & billing systems
  - Geographic location controls
  - Native smart client access



ObjectAssure Single Copy Data Protection

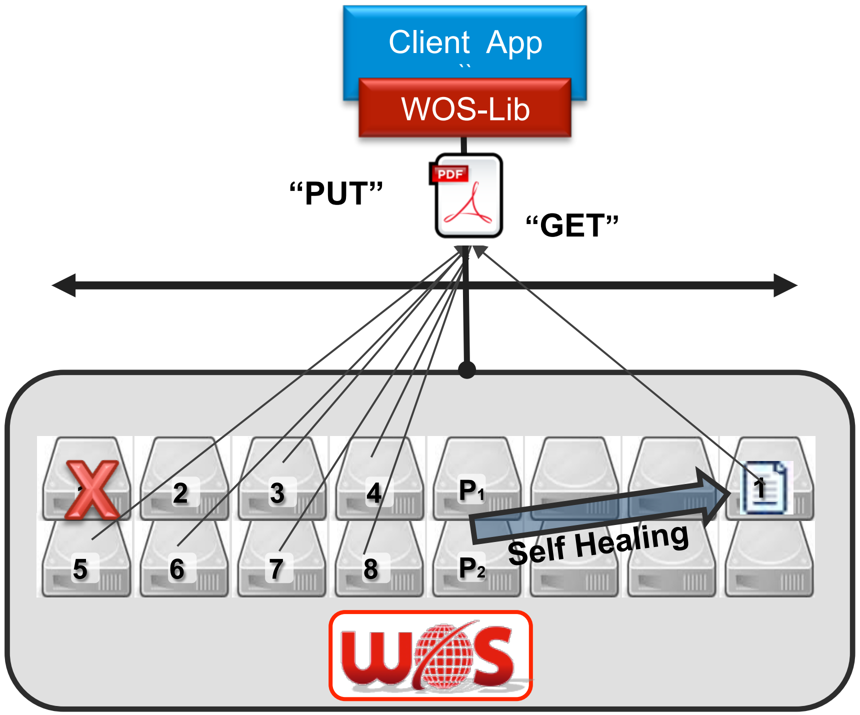
- ObjectAssure erasure-code based declustered data protection
  - An erasure code provides redundancy by breaking objects into smaller fragments and storing the fragments across different disks
  - Data can be recovered from any smaller combination of fragments
- ObjectAssure is the first erasure code protection mechanism for hyper-scale, high-performance cloud storage
- With ObjectAssure, each WOS node can wit… to two concurrent drive failures without loss data availability
  - Data protection without the cost of replication

ObjectAssure Single Copy Data Protection

- Erasure coding vs.RAID Benefits
  - Only rebuild data, not whole disks
  - Rehydrate to all available resources, not rebuild to a specific drive
- ObjectAssure vs. Dispersal Methods
  - All ObjectAssure data is locally available, speeding access
  - Dispersed data has to come across WAN, significantly slowing access
  - WOS is up to 25 times faster than dispersal-based storage platforms

- WOS supports mixed mode (replication or (policies
  - Replicate data that requires fastest access
  - Erasure code data that isn’t as frequently used

ObjectAssure Single Copy Data Protection



Best viewed in Slide Show mode

Operational Specifics

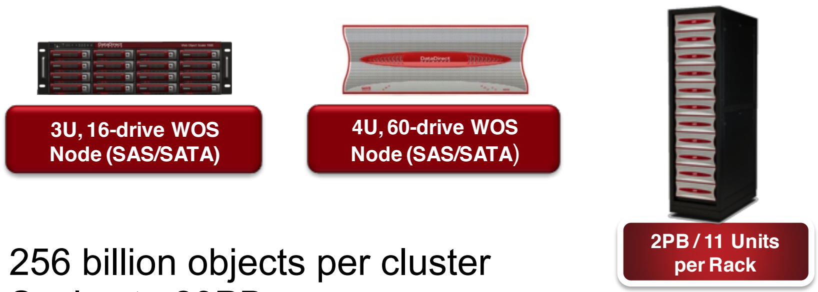
- Object Assure works with both WOS-Lib & REST API's
- Operates within a single WOS node
- Enabled by specifying a single (1) replica in a WOS storage policy
- OA & replica storage methods can be mixed inside a WOS cluster
- Detects concurrent multi-disk errors & corrects for 2 separate concurrent disk errors per-WOS node

OA Process Flow

- During PUT operations, OA splits objects across 8 drives & generates 2 parity drives (8+2)
- If a drive fails or object is corrupted, WOS uses the parity drives to rebuild corrupted data on other drives(self healing)
- WOS OA corrects data in-flight during reads (GETs) as needed

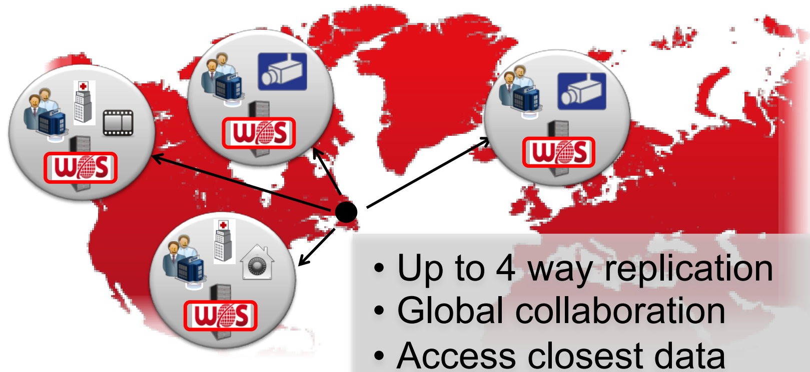
WOS – Architected for Big Data

Hyper-Scale



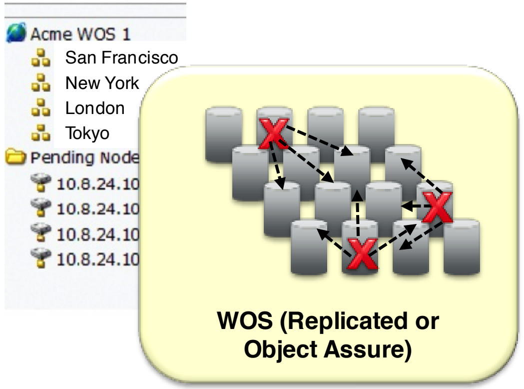
- 256 billion objects per cluster
- Scales to 23PB
- Start small, grow to tens of Petabytes
- Network & storage efficient

Global Reach & Data Locality



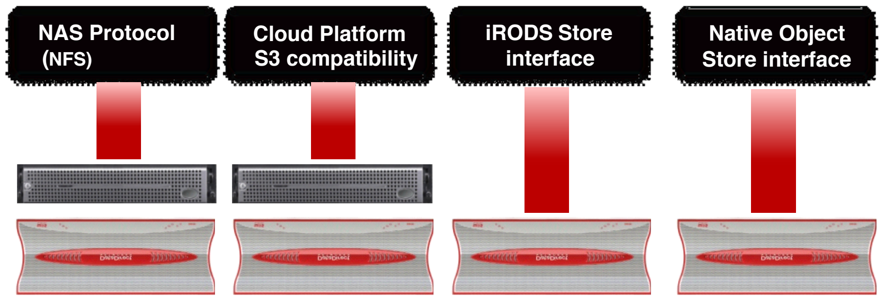
- Up to 4 way replication
- Global collaboration
- Access closest data
- No risk of data loss

Resiliency with Near Zero Administration



- Self healing
- All drives fully utilized
- 50% faster recovery than traditional RAID
- Reduce or eliminate service calls

Universal Access



WOS NAS Access
- CIFS/NFS protocols
- Scalable
- HA & DR Protected
- Federation across existing NAS

Cloud Store Platform
- S3-Compatible & WebDAV APIs
- Multi-tenancy
- Reporting & Billing
- Personal storage

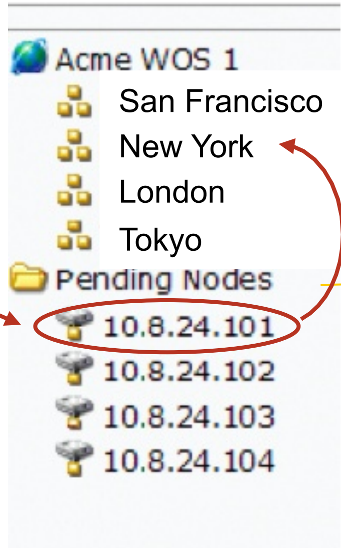
IRODS
- Custom rules – a rule for every need
- 28 access methods

Native Object Store
- C++, Python, Java, PHP, HTTP REST interfaces
- Simple interface

DDN | WOS™
Deployment & Provisioning

DDN | WOS building blocks are easy to deploy & provision – in 10 minutes or less

- Provide power & network for the WOS Node
- Assign IP address to WOS Node & specify cluster name ("Acme WOS 1")
- Go to WOS Admin UI. WOS Node appears in "Pending Nodes" List for that cluster
- Drag & Drop the node into the desired zone
- Assign replication policy (if needed)



Simply drag new nodes to any zone to extend storage

Policies

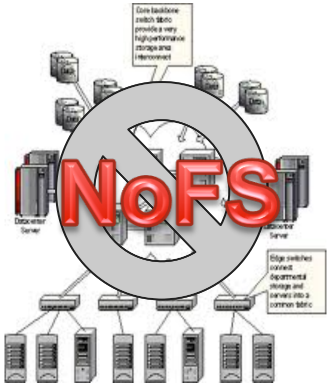
Policy Name (ID) **Zone Replication**

Policy Name: UnitedStates

Create Policy

Cancel

Zone	Replica Count
San Francisco	1
New York	1
London	0
Tokyo	0



It's that simple to add 180TB to your WOS cluster!

GridScaler – WOS Integration

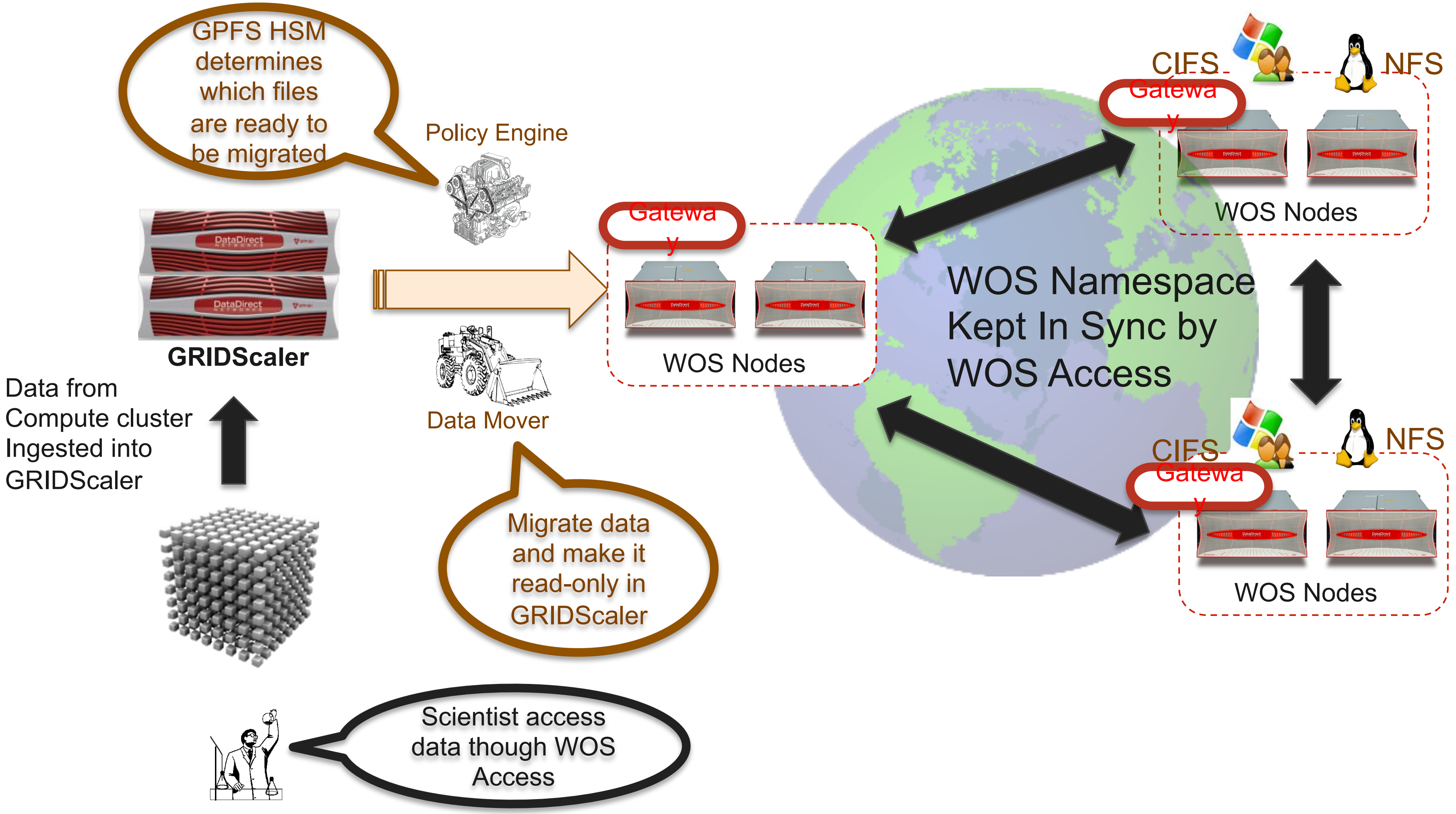
Objective

- Enable WOS to be used as back-end storage for GPFS/Gridscaler systems, either to offload data from GPFS filesystems or to enable collaboration between GPFS and distributed WOS Access applications /users

Use Cases

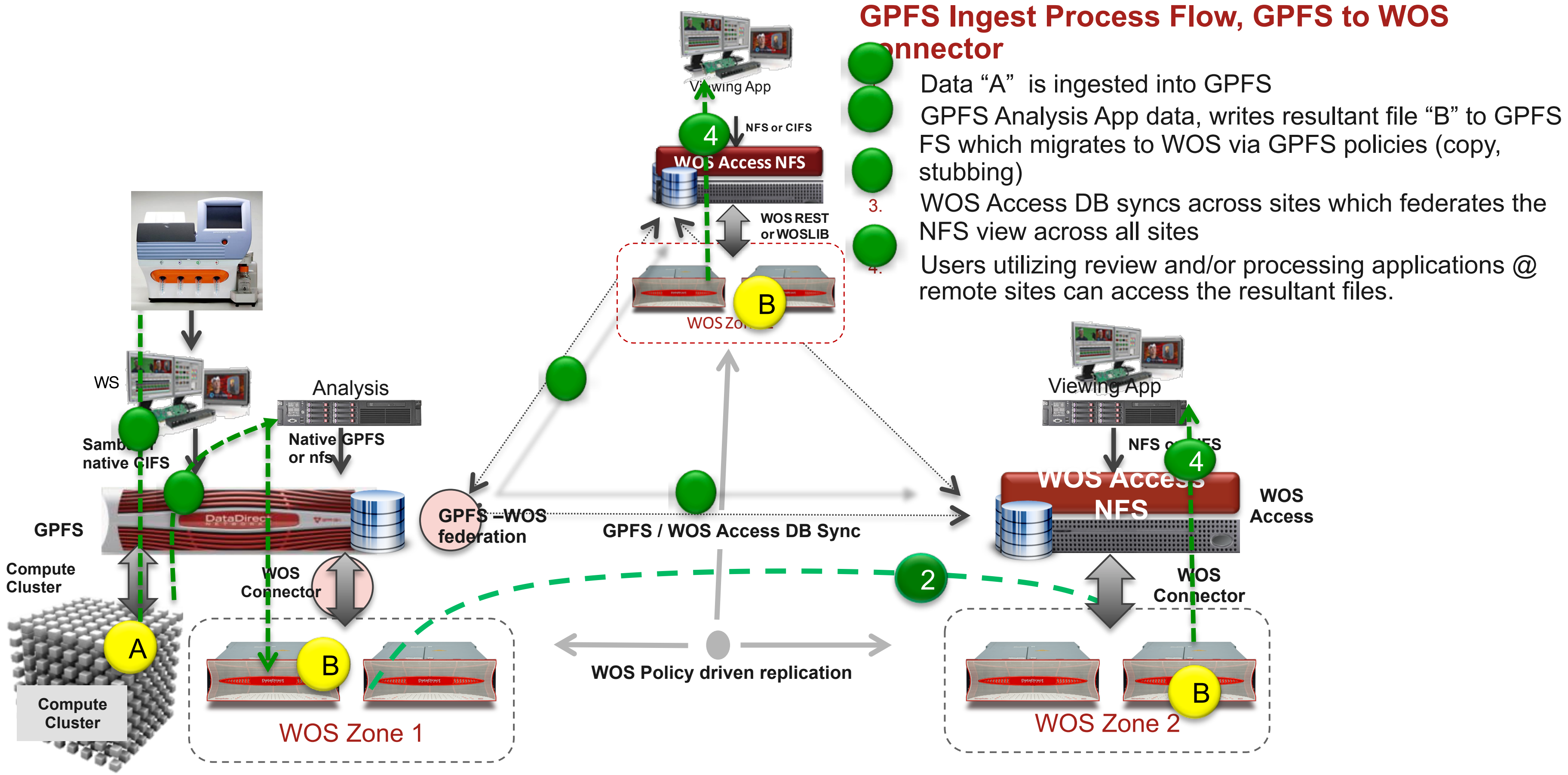
1. Archive -GPFS files are archived to WOS based on GPFS migration policy (copy/ stub) to free up disk space on GPFS. Archived files are viewable / accessible from GPFS as a WOS (NFS) mount point and can either be accessed directly from WOS or pre-staged into GPFS storage
2. Collaboration- Data is ingested / processed in GPFS, and resultant file is written directly to WOS (exposed as nfs mount point) & federated across WOS Access sites (available via NFS or CIFS clients)
3. Collaboration - Data is ingested / processed in GPFS, resultant file is written directly to GPFS & also copied to WOS (exposed as nfs mount point & federated across WOS Access sites) based on GPFS Policy
4. Collaboration -Data is ingested into WOS at any WOS Access location (via NFS or CIFS), federated across other WOS Access and/or GPFS locations, and accessible to GPFS storage as an NFS mount point

Overview of System – Phase I



GPFS – WOS Integration
Archive & Local Ingest Use Cases

Use Case 1: Archive GPFS data into WOS
Use Case 2: Ingest to GPFS for analysis, GPFS w/connector distributes to WOS for viewing/processing, GPFS to WOS DB Sync federates GPFS & WOS



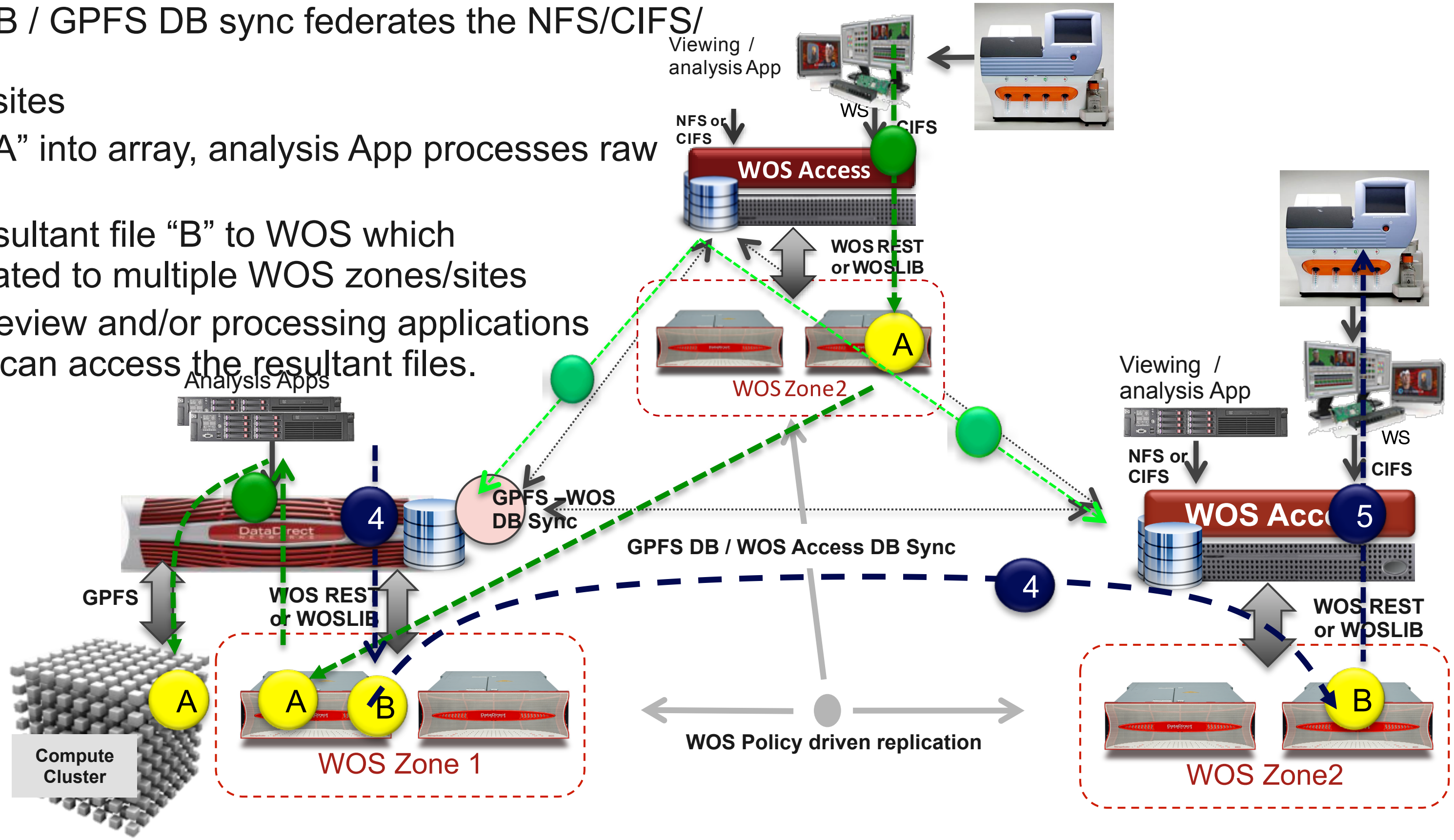
GPFS – WOS Integration

Phase 2:Remote Ingest Use Case

Use Case 3: Ingest to WOS via WOS Access, distribute to GPFS for analysis, migrate back to WOS for distribution, viewing/processing

Process Flow

- Ingest raw data "A" into WOS via WOS Access
- WOS Access DB / GPFS DB sync federates the NFS/CIFS/ GPFS view across all sites
3. GPFS ingests "A" into array, analysis App processes raw data
4. GPFS writes resultant file "B" to WOS which then gets replicated to multiple WOS zones/sites
5. Users utilizing review and/or processing applications @ remote sites can access the resultant files.



Conclusion

- As data sets and transaction densities grow, data systems must become more efficient at every level.
- Latencies are expensive in multiple dimensions often requiring additional hardware as a “work around” to enable usable performance.
- The fundamental concept of “just enough” is not a luxury but rather a necessity for data capture, reduction, and distribution.
- Processing must be adjacent to the data.
- Service times to end users should become a portion of the overall figure of merit for any research system.

Thank You