

Object storage in Cloud Computing and Embedded Processing

Jan Jitze Krol

Systems Engineer

DDN is a Leader in Massively Scalable Platforms and Solutions for Big Data and Cloud Applications

- ▶ **Established:** 1998
- ▶ **Revenue:** \$226M (2011) – 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

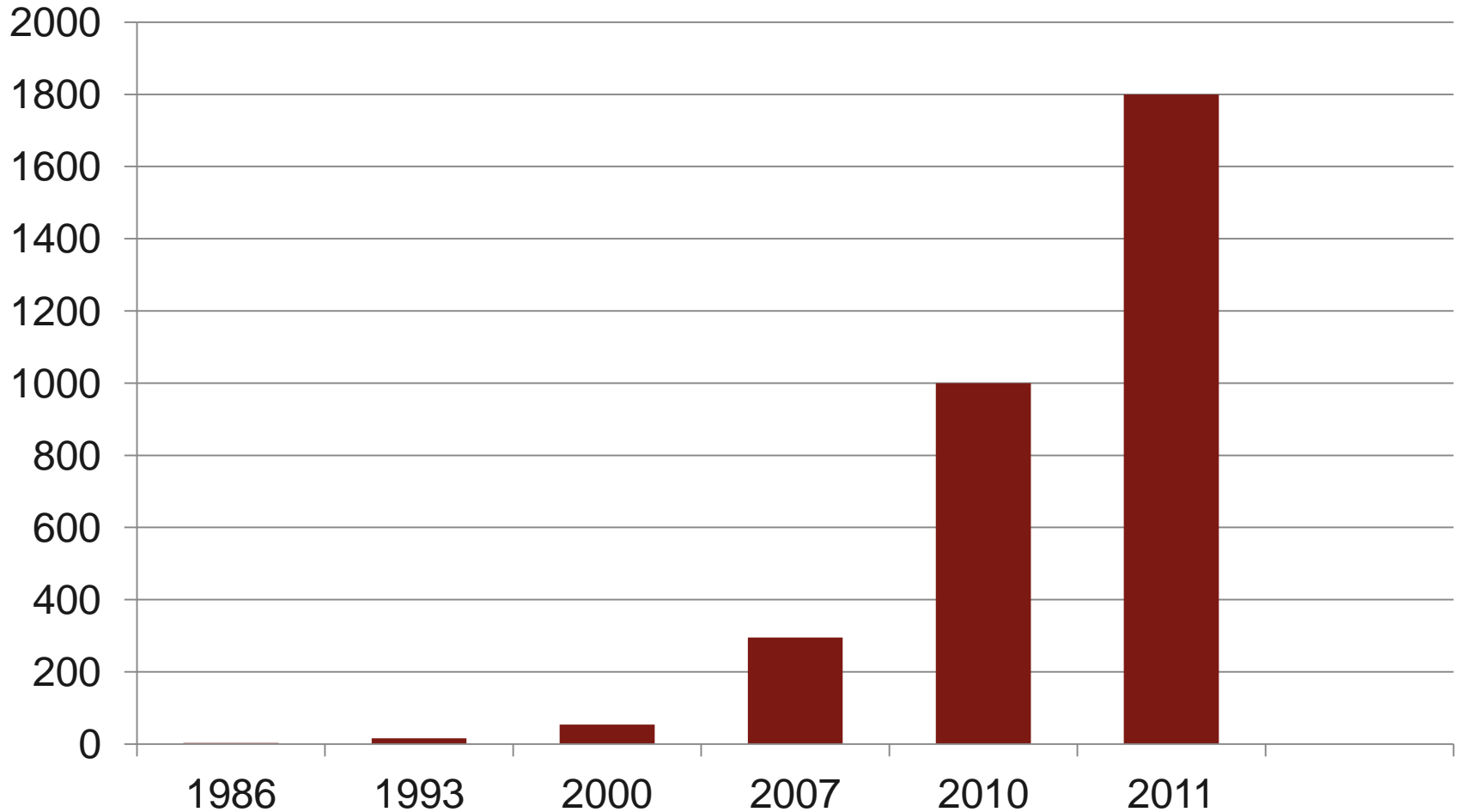


World-Renowned & Award-Winning



Fact: amount of data is growing, fast

data stored world wide in Exa Bytes



Disk drives grow bigger, not faster or better

- ▶ Disk drives haven't changed that much over the last decade
- ▶ They just store more, ~40GB in 2002, 4000 GB in 2012
- ▶ Access times are about the same, ~ 5 – 10 ms
- ▶ Write/read speeds are about the same ~ 50 -100MB/s
- ▶ Read error rate is about the same 1 error per 10^{14} bits read, or one guaranteed read error per 12 TB read.

- ▶ **Highlight two of DDN's initiatives to deal with large repositories of data:**
- ▶ The Web Object Scaler, WOS
- ▶ Embedded data processing, aka in-storage processing

Challenges of ExaByte scale storage

- ▶ Exponential growth of data
- ▶ The expectation that all data will be available everywhere on the planet.
- ▶ Management of this tidal wave of data becomes increasingly difficult with regular NAS :
 - ▶ Introduced in the 90's (when ExaByte was a tape drive vendor)
 - » With 16TB file system sizes that many still have
 - ▶ Management intensive
 - » LUNs, Volumes, Aggregates,...
 - » Heroically management intensive at scale
 - ▶ Antiquated resiliency techniques that don't scale
 - » RAID (disk is a unit in RAID, whereas drive vendors consider a sector a unit)
 - » Cluster failover, "standby" replication, backup
 - » File Allocation Tables, Extent Lists
 - ▶ Focused on structured data transactions (IOPS)
 - » File locking overhead adds cost and complexity

Hyperscale Storage | Web Object Scaler

DataDirect[™]
NETWORKS
INFORMATION IN MOTION[™]



WOS

DataDirect[™]
NETWORKS

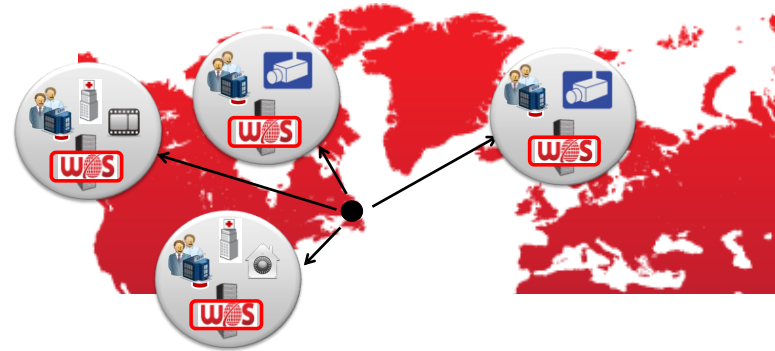
NoFS

**Hyperscale
Distributed
Collaboration**

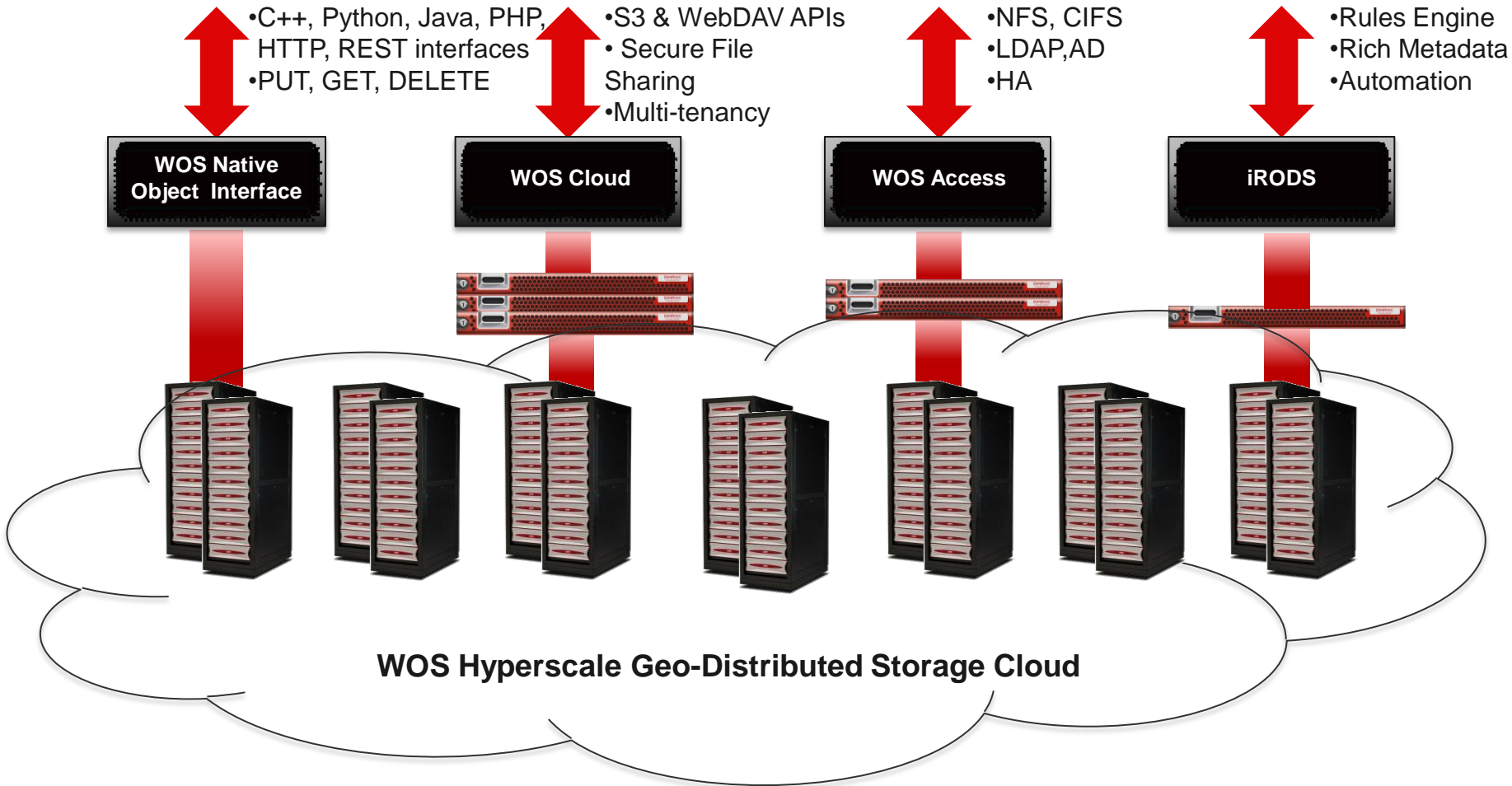


What is WOS

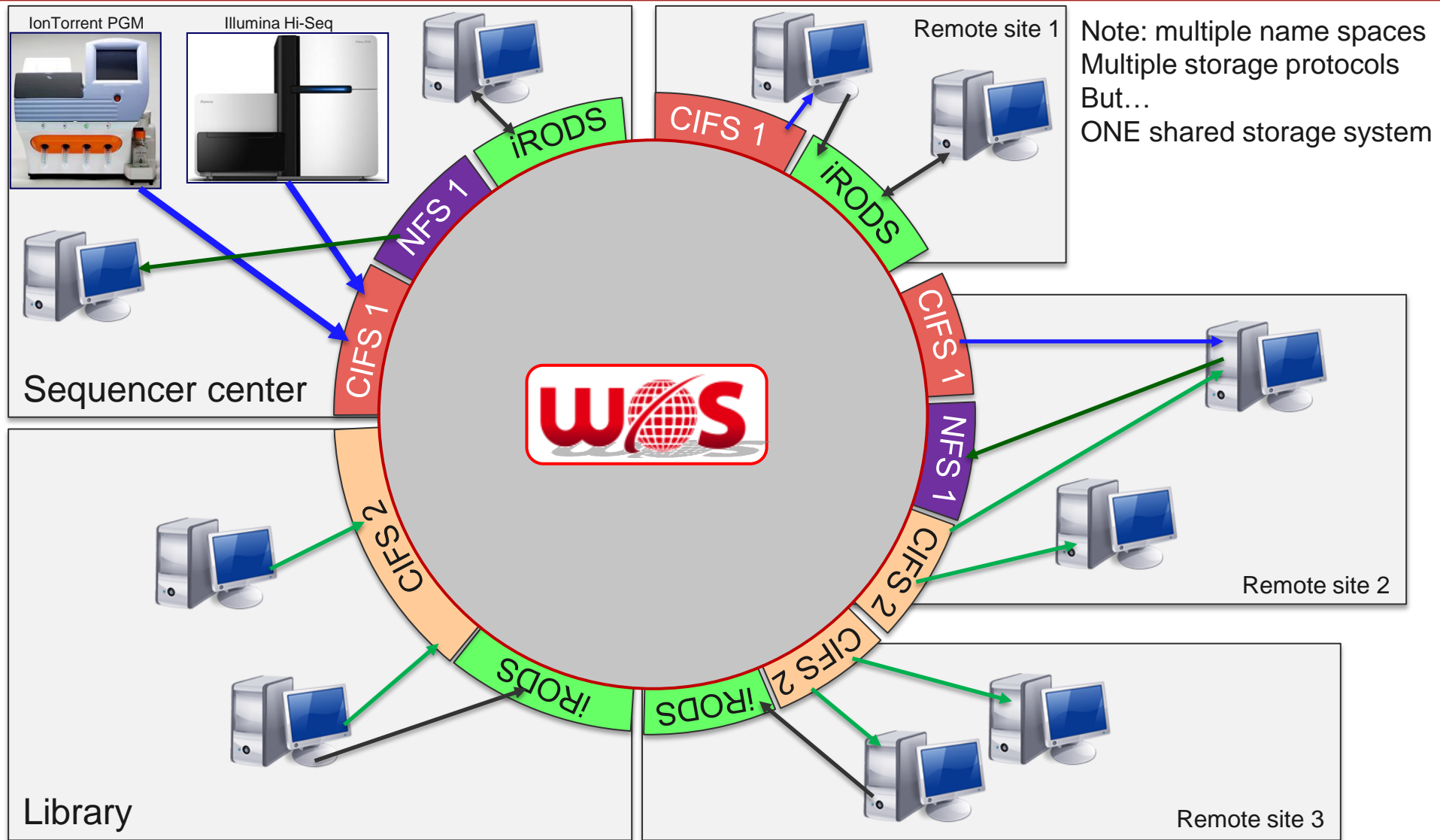
- ▶ A data store for immutable data
 - That means we don't need to worry about locking
 - No two systems will write to the same object
- ▶ Data is stored in objects
 - Written with policies
 - Policies drive replication
- ▶ Objects live in 'zones'
- ▶ Data protection is achieved by replication or erasure coding
 - Replicate within a zone or between zones
 - Data is available from every WOS node in the cluster
- ▶ Only three operations possible, PUT, GET and DELETE



Universal Access to Support a Variety of Applications



What can you build with this?





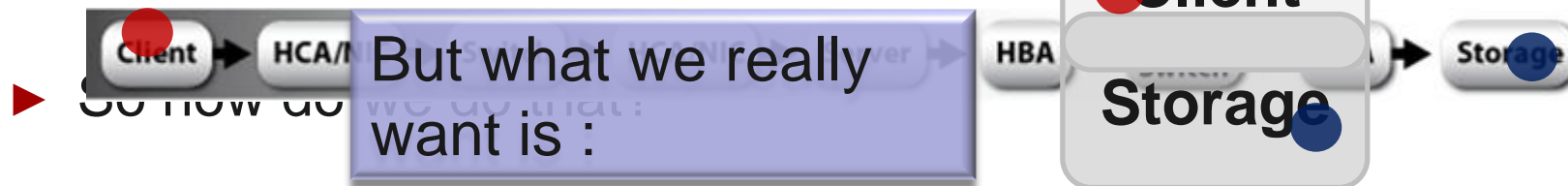
Storing ExaBytes, ZettaBytes or JottaBytes of data is only part of the story. The data needs to be processed too, which means as fast as possible access.

What is 'Embedded Processing'?

And why ?

- ▶ Do data intensive processing as 'close' to the storage as possible.
 - Bring computing to the data instead of bring data to computing
- ▶ HADOOP is an example of this approach.
- ▶ Why Embedded Processing?
- ▶ Moving data is a lot of work
- ▶ A lot of infrastructure needed

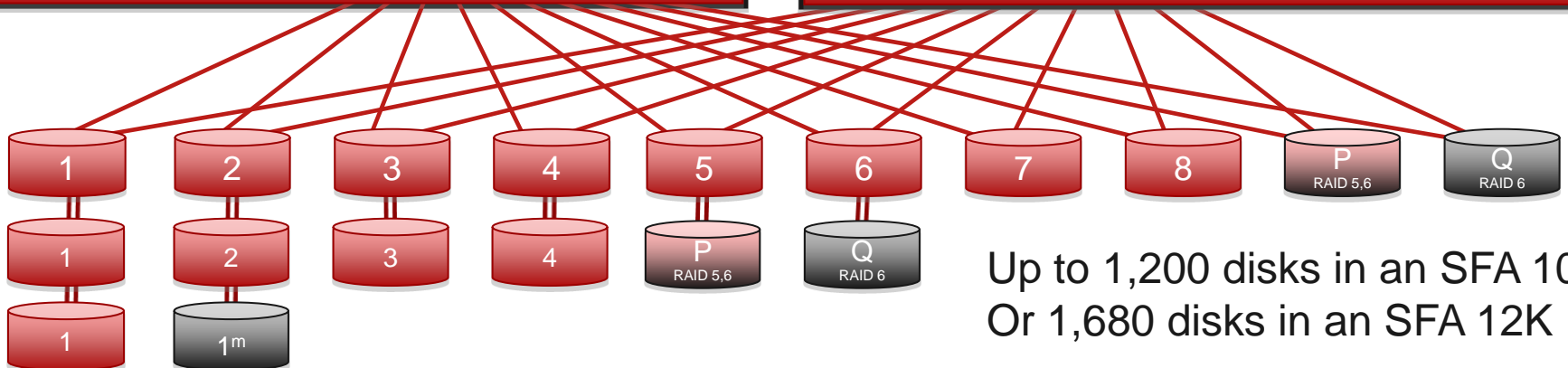
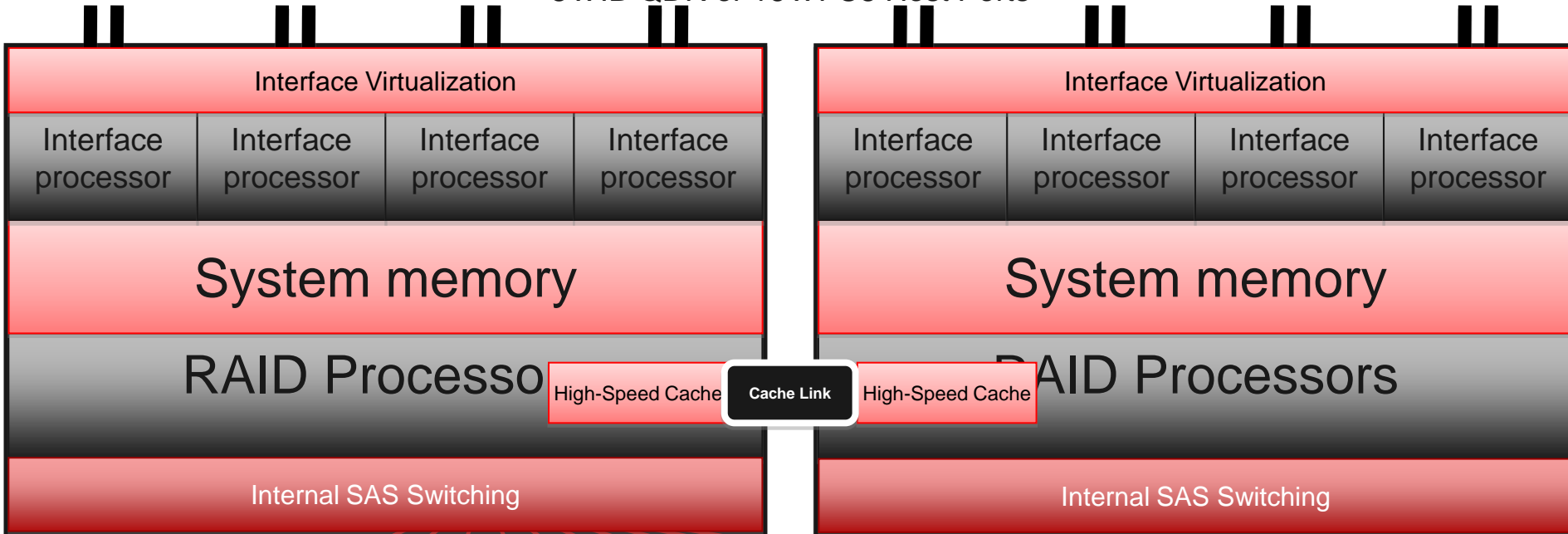
Client sends a request to storage (red ball)



Storage responds with data (blue ball)

Storage Fusion Architecture (SFA)

8 x IB QDR or 16 x FC8 Host Ports



Up to 1,200 disks in an SFA 10K
Or 1,680 disks in an SFA 12K

Repurposing Interface Processors

- ▶ In the block based SFA10K platform, the IF processors are responsible for mapping Virtual Disks to LUNs on FC or IB
- ▶ In the SFA10KE platform the IF processors are running Virtual Machines
- ▶ RAID processors place data (or use data) directly in (or from) the VM's memory
- ▶ One hop from disk to VM's memory
- ▶ Now the storage is no longer a block device
- ▶ **It is a storage appliance with processing capabilities**

One SFA-10KE controller

8 x IB QDR/10GbE Host Ports (No Fibre Channel)

Interface Virtualization

Virtual
Machine

Virtual
Machine

Virtual
Machine

Virtual
Machine

System memory

RAID Processors

High-Speed Cache

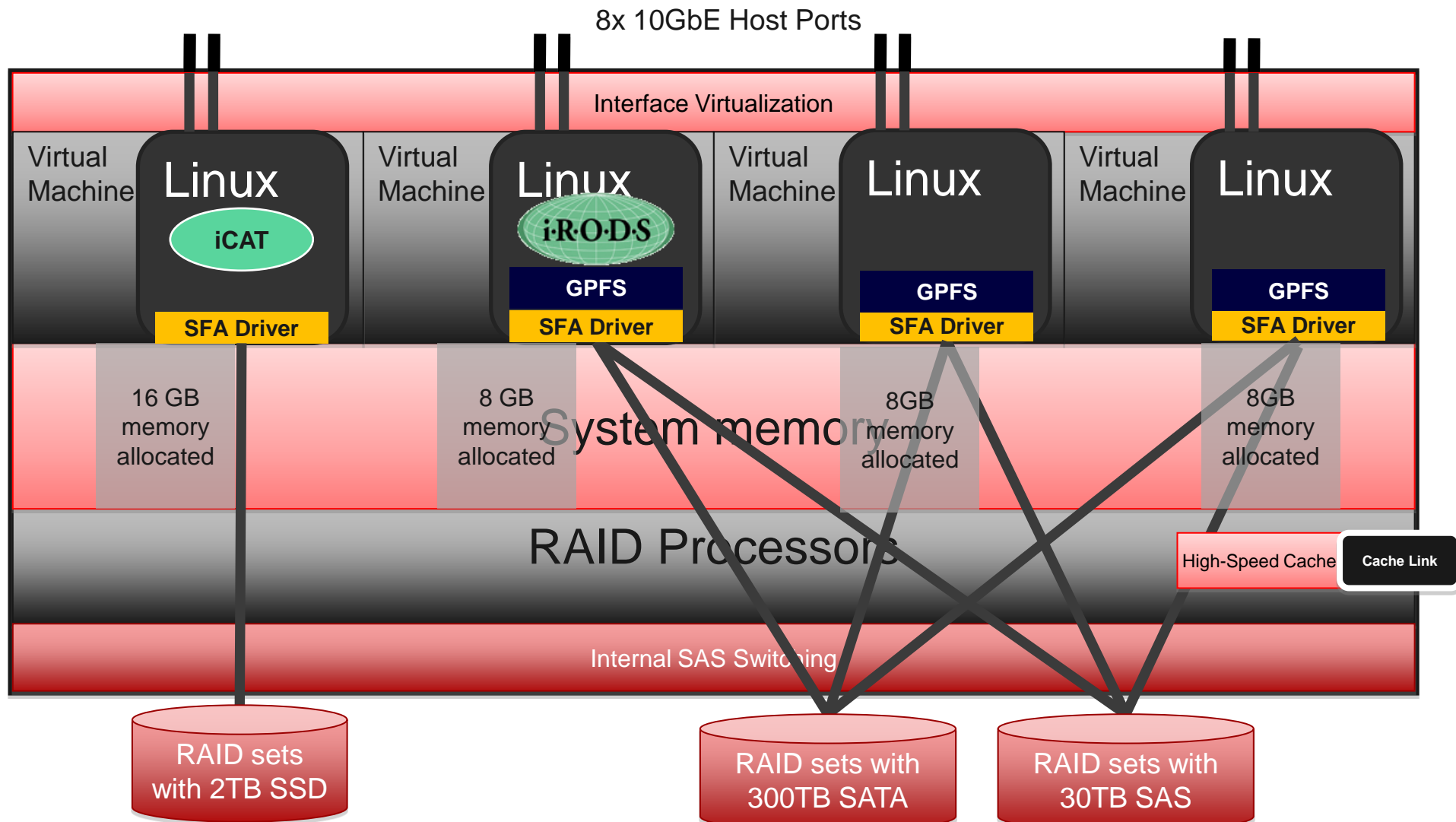
Cache Link

Internal SAS Switching

Example configuration

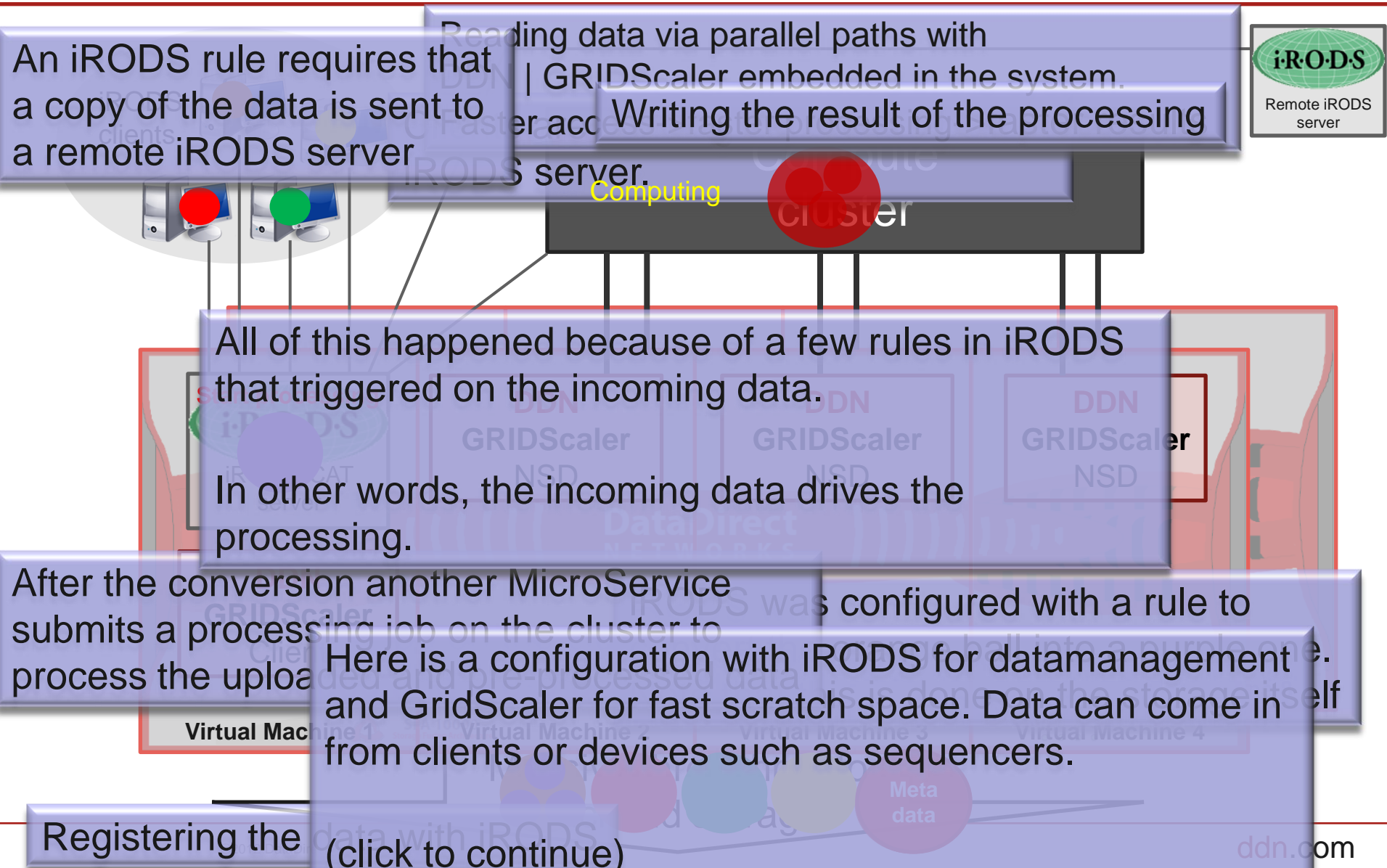
- ▶ Now we can put iRODS inside the RAID controllers
 - This give iRODS the fastest access to the storage because it doesn't have to go onto the network to access a fileserver. It lives **inside** the fileserver.
- ▶ We can put the iRODS catalogue, iCAT, on a separate VM with lots of memory and SSDs for DataBase storage
- ▶ The following example is a mix of iRODS with GPFS
 - The same filesystem is also visible from an external compute cluster via GPFS running on the remaining VMs
- ▶ This is only one controller, there are 4 more VMs in the other controller need some work too
 - They see the same storage and can access it at the same speed.
- ▶ On the SFA-12K we will have 16 VM's available running on Intel Sandy Bridge processors. (available Q3 2012)

Example configuration



- ▶ Since iRODS runs inside the controller we now can run iRODS MicroServices right on top of the storage.
- ▶ The storage has become an iRODS appliance ‘speaking’ iRODS natively.
- ▶ We could create ‘hot’ directories that kick off processing depending on the type of incoming data.

With iRODS and GridScaler parallel filesystem





Thank You

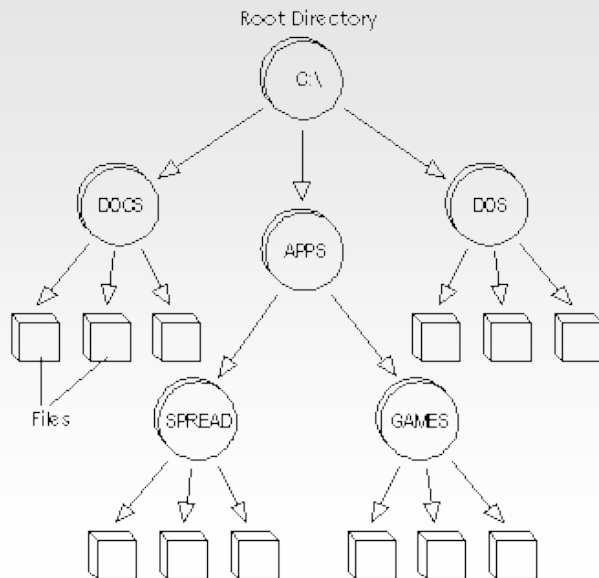
Backup slides



Object Storage Explained

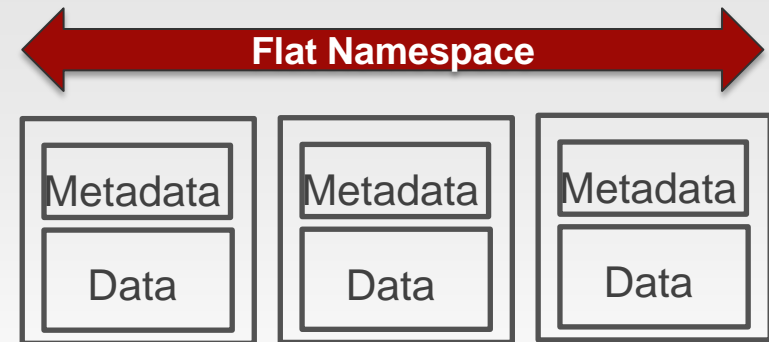
- ▶ Object storage stores data into containers, called objects
- ▶ Each object has both data and user defined and system defined metadata (a set of attributes describing the object)

File Systems



File Systems were designed to run individual computers, then limited shared concurrent access, not to store billions of files globally

Objects



Objects are stored in an infinitely large flat address space that can contain billions of files without file system complexity

Intelligent WOS Objects

Sample Object ID (OID): ACuoBKmWW3Uw1W2TmVYthA

WOS Signature

A random 64-bit key to prevent unauthorized access to WOS objects

WOS Policy

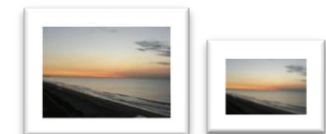
Eg. Replicate Twice; Zone 1 & 3

WOS Checksum

Robust 64 bit checksum to verify data integrity during every read

User Metadata
Key Value or Binary

Object = Photo
Tag = Beach



thumbnails

Full File or
Sub-Object



WOS Performance Metrics

Network	Performance Per Node				Maximum System Performance (256 Nodes)	
	Large Objects		Small Objects (~20KB)		Small Objects (~20KB)	
	Write MB/s	Read MB/s	Object Writes/s	Object Reads/s	Object Writes/Day	Object Reads/Day
1 GbE	200	300	1200	2400	25,214,976,000	50,429,952,000
10 GbE	250	500	1200	2400	25,214,976,000	50,429,952,000

- Benefits
 - Greater data ingest capabilities
 - Faster application response
 - Fewer nodes to obtain equivalent performance

WOS – Architected for Big Data

Hyper-Scale



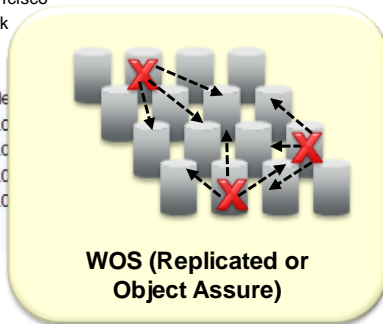
- 256 billion objects per cluster
- 5 TB max object size
- Scales to 23PB
- Network & storage efficient
- 64 zones and 64 policies

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

