Storage IO Challenges

Ghislain de Jacquelot
Technical Director HPC EMEA

DataDirect™
N E T W O R K S
I N F O R M A T I O N   I N   M O T I O N™

Overview of the SFA10000
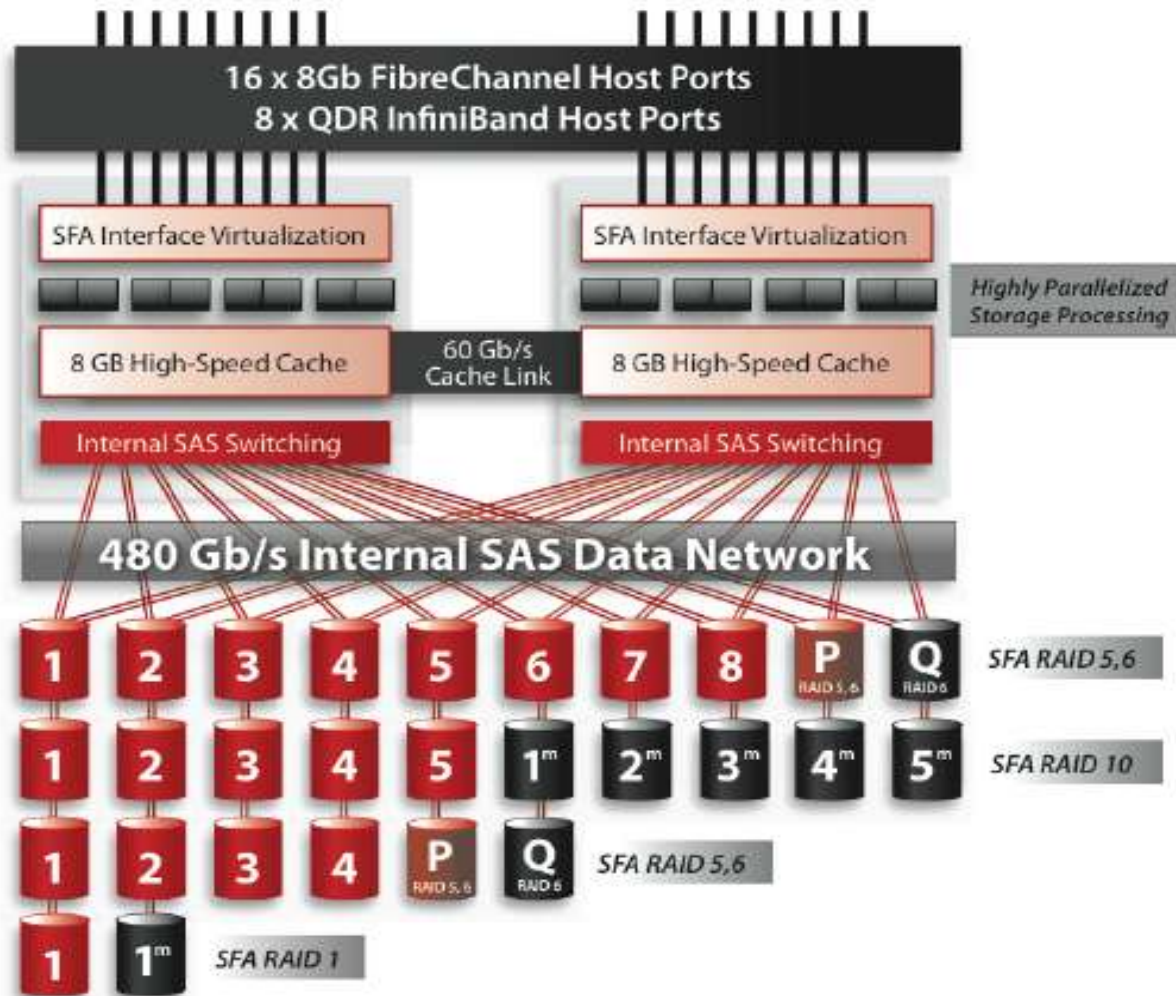
Hardware IO challenges

Provide simultaneously
- High-Availability
- Reliability of data
- High Bandwidth
- High IO rates
- Undegraded performance
- Mixing cheap components and fast devices

- **Read error rates**

- **Rebuild performance**

- **Mix of small & large IO**

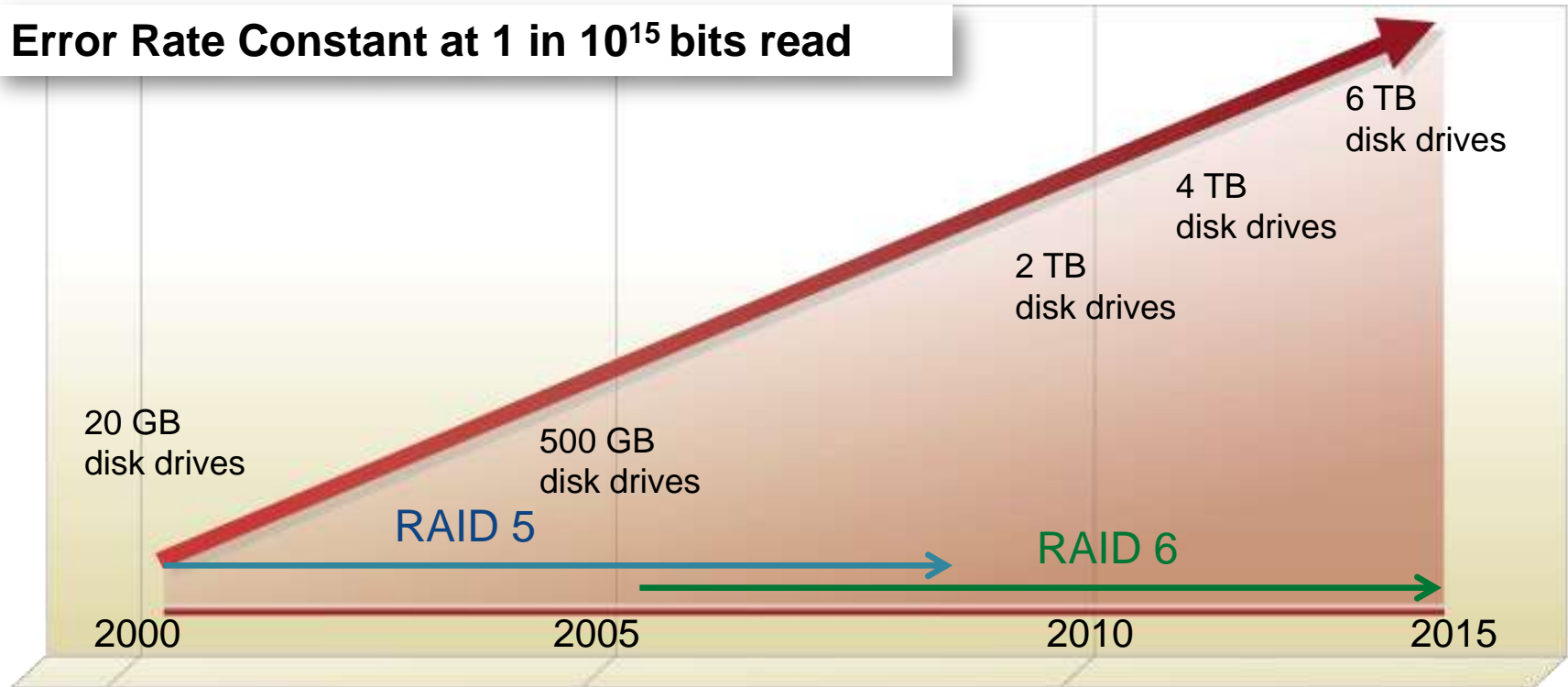- **In-Store compute**

DataDirect
N E T W O R K S
INFORMATION IN MOTION

Dealing with read error rate

Disks: binary devices ?

- **Either new or full**

- **Working or dead ?**

- **Stores 0s and 1s ?**

DataDirect
NETWORKS
INFORMATION IN MOTION

Reliability & Performance Challenges

Error Rate Constant at 1 in 10^{15} bits read

6 TB
disk drives

4 TB
disk drives

2 TB
disk drives

20 GB
disk drives

500 GB
disk drives

RAID 5

RAID 6

2000   2005   2010   2015

- Disks are >100 times denser than a decade ago, but media error rate is constant

- *More Media + More Media Errors per Drive*

DataDirect
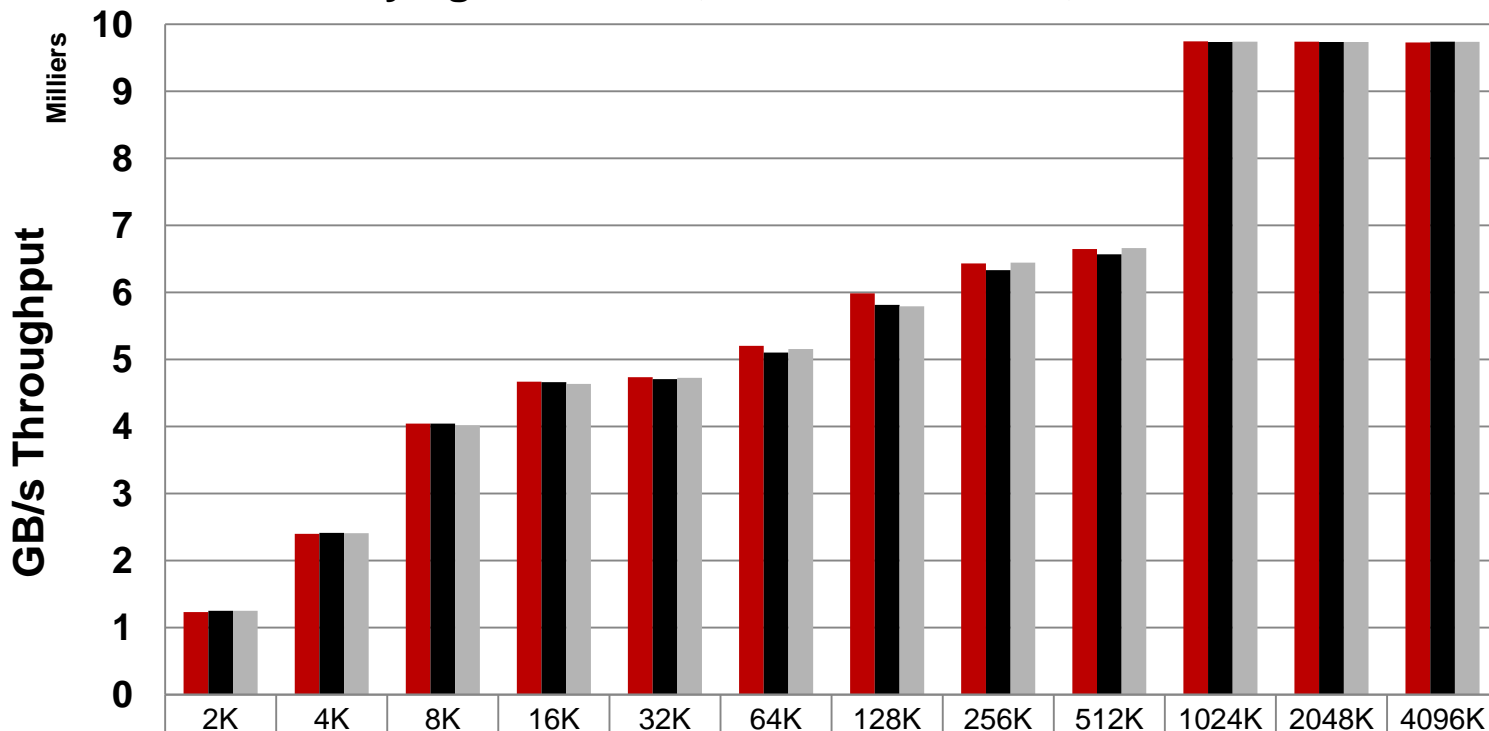N E T W O R K S
INFORMATION IN MOTION

Defense-In-Depth Solutions

- **DDN's SFA Protects, Detects, Corrects & Enables Performance in Multiple Ways**
  - Bit Error Rates Get Eliminated Through Parity
    - Checksums & Parity Calculation on Read Options
    - Foundation for T10-Dif (like) Protection Without Perf Loss
  - Background Parity Scrubbing (cycle data at rest)
  - Robust cache protection eliminates corruption on write

Much of this technology also serves as basis for future real-time capabilities.

DataDirect
NETWORKS
INFORMATION IN MOTION

Defense In Depth: SATAssure

300 x 1TB SATA Drives, 30 x 8+2, Write Throughput, Varying Block Size, 128K Chunk Size, 16 x FC-8

	2K	4K	8K	16K	32K	64K	128K	256K	512K	1024K	2048K	4096K
■ SATAssure Parity	1230	2398	4044	4667	4735	5203	5985	6432	6647	9745	9739	9729
■ SATAssure DIF	1251	2412	4043	4662	4706	5102	5815	6332	6568	9736	9736	9741
■ SATAssure off	1251	2409	4015	4634	4725	5154	5794	6443	6662	9739	9736	9738

6Gb SS7000 Enclosure Increases Performance 10%

DataDirect
N E T W O R K S
INFORMATION IN MOTION

Dealing with rebuilds

DataDirect™
N E T W O R K S
INFORMATION IN MOTION™

10PFlop Deployment Scenario

- **50 PB Useable Storage System**
  - 15,000 4TB Drives = 1,500 LUNs
  - 1TB/s Storage Cluster
  - Drive AFR = 3%
    - AFR is observed data from all DDN installation base
  - 1.2 drive failures per day
    - Without fast rebuilds – file system is always rebuilding
- **Without Performance Protection…**
  - A striped I/O will only run as fast as the slowest LUN which will be the LUN being rebuilt
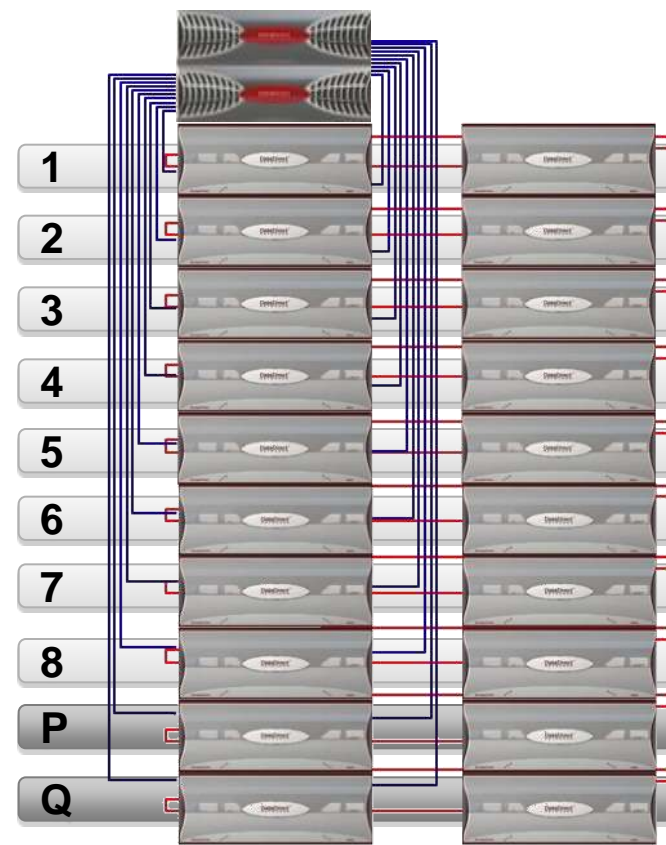
DataDirect
NETWORKS
INFORMATION IN MOTION

Reconstruction Comparison

	SFA Software	Software RAID
Capacity Utilized	60%	60%
Rebuild Speed	40MB/s	25MB/s
MTTR	28 hours (all blocks)	26 hours (data only)
Performance Impact	10%	50% When Optimized (Currently 63%)
Striped IO Perf In Reconstruction	820GB/s	455GB/s (best case) 233GB/s (unoptimized)

1.2 drive failures per day with rebuild taking a full day:
The file system is always slow
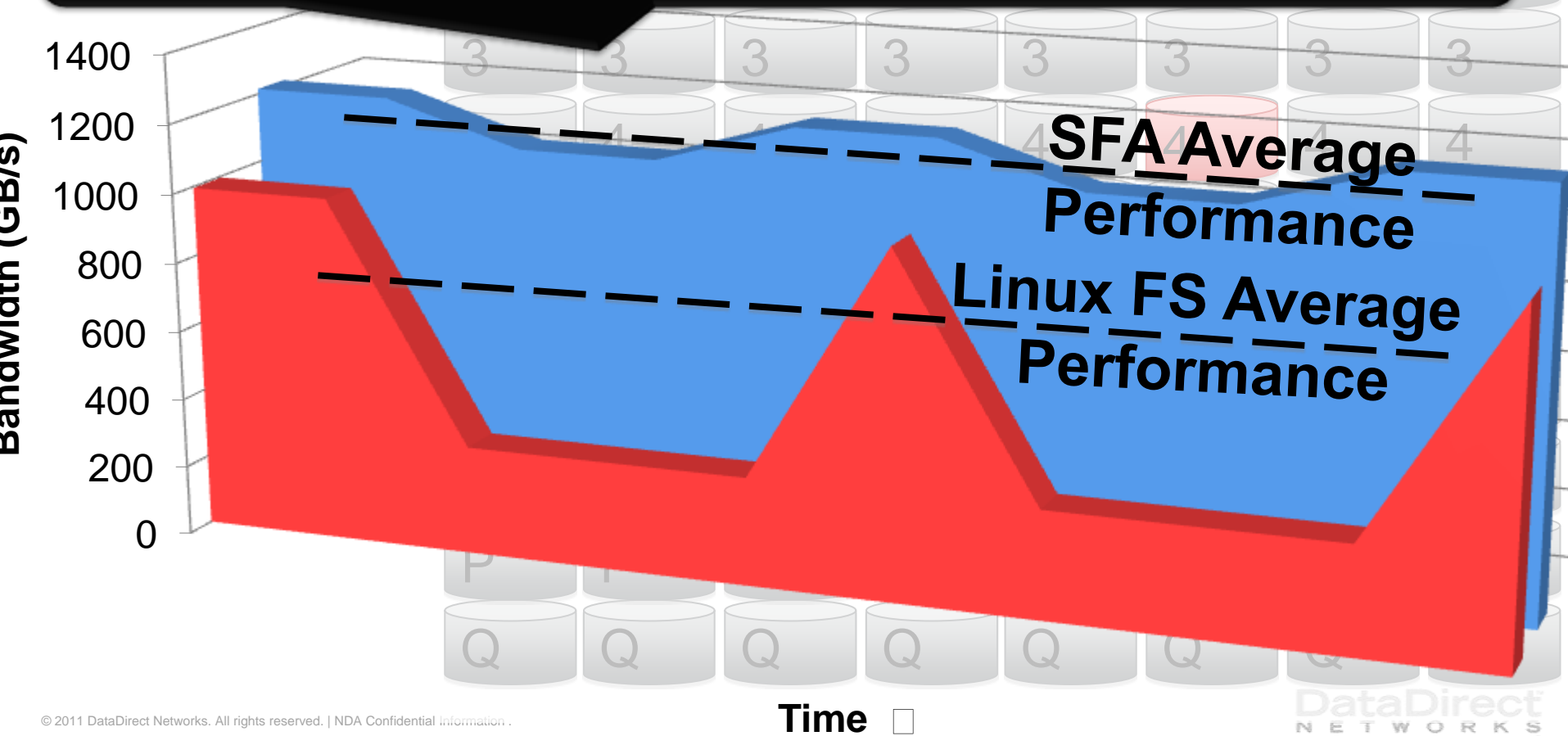
DataDirect
NETWORKS
INFORMATION IN MOTION

SFA10000

- **Maximum Data Availability & Performance**
  - Direct cabling avoids daisy chaining & latency
  - Data is striped across channels/enclosures for maximum bandwidth to/from HDDs
- **Designed To Withstand 20% Loss without Service Impact**
  - Drive Channels are RAIDed 8+2
  - Drive Enclosures are RAIDed 8+2
- **Oversized 60GB/s Backend**
  - Preserves performance even during rebuilds
  - Supports data protection and management IO

DataDirect
N E T W O R K S
INFORMATION IN MOTION

Ahmdal's Law & Striped File I/O

No More Than A 10% Performance Drop For Rebuilds, Compared To As Much As 66% With Other Systems

Bandwidth (GB/s)

1400
1200
1000
800
600
400
200
0

SFA Average Performance

Linux FS Average Performance

Time

Mixing large and small IO

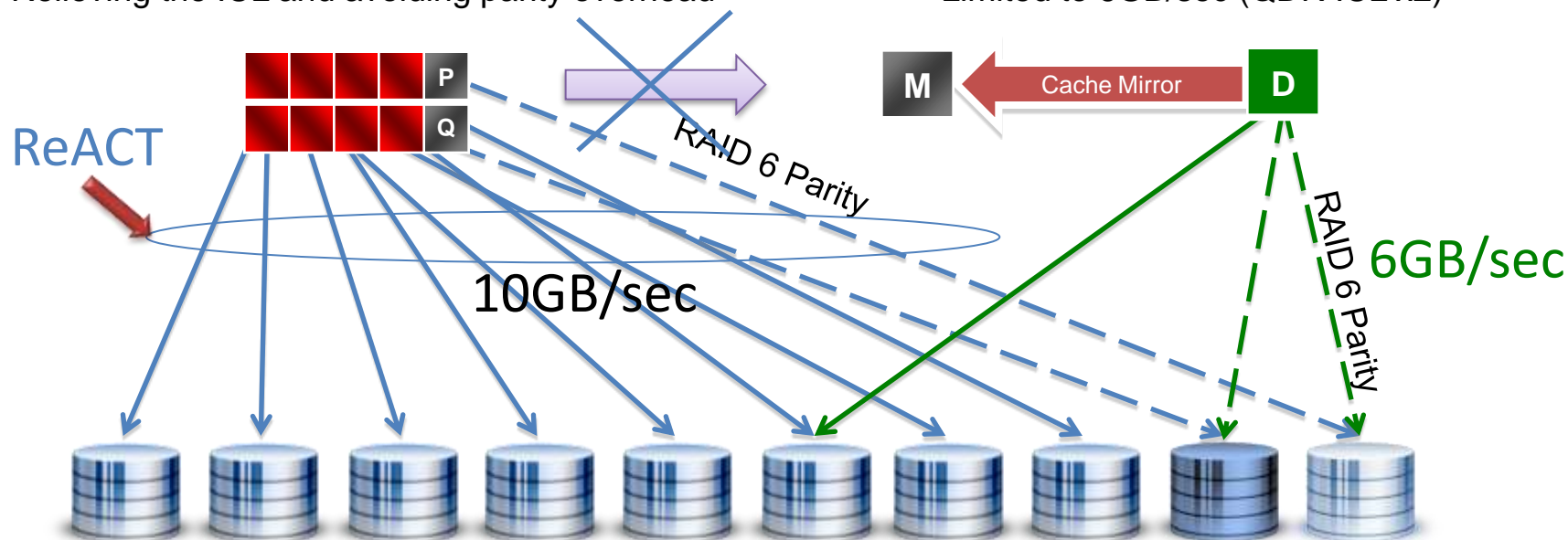# ReACT technology
Bringing out the performance of SFA10K

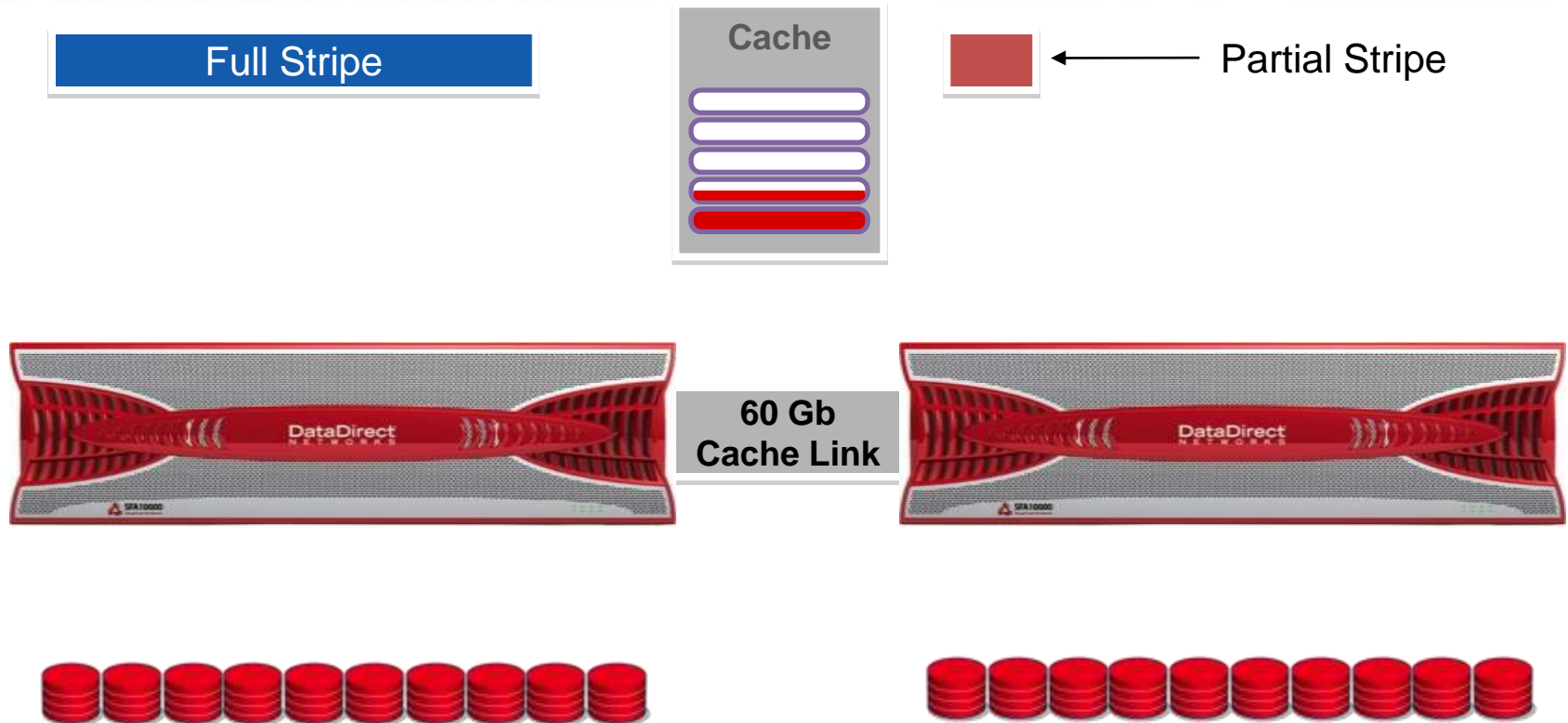## Write-Back Cache ≒ Bottleneck in Mirror & Interconnect

Aligned IO

- Written through cache and striped directly to disk
- Relieving the ICL and avoiding parity overhead

Non-Aligned IO

- Mirrored Write-Back Cache
- Limited to 6GB/sec (QDR ICL x2)

P

Q

ReACT

RAID 6 Parity

10GB/sec

M ← Cache Mirror — D

RAID 6 Parity

6GB/sec

Performance Efficiency

Full Stripe

Cache
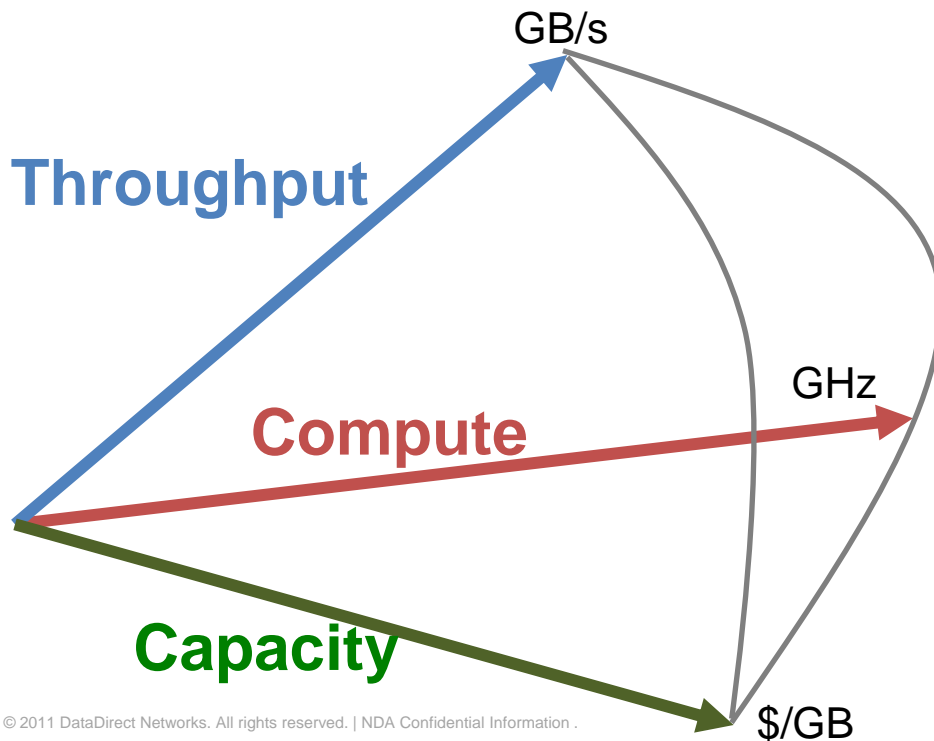
Partial Stripe

60 Gb Cache Link

With ReACT, partially striped I/O is given more cache headroom to achieve full stripe & optimize applications. SFA Algorithms ensure max. performance per array, per cache GB.

DataDirect
N E T W O R K S
INFORMATION IN MOTION

In-Store compute

DataDirect™
N E T W O R K S
I N F O R M A T I O N   I N   M O T I O N™

In-Store Compute

- **Announced In 2009**
- **Optimized For Parallel File Systems**
- **Supports File I/O, Pre and Post Processing**

GB/s

Throughput

Compute

GHz

Capacity

$/GB

DataDirect
NETWORKS
INFORMATION IN MOTION

IO Path Acceleration
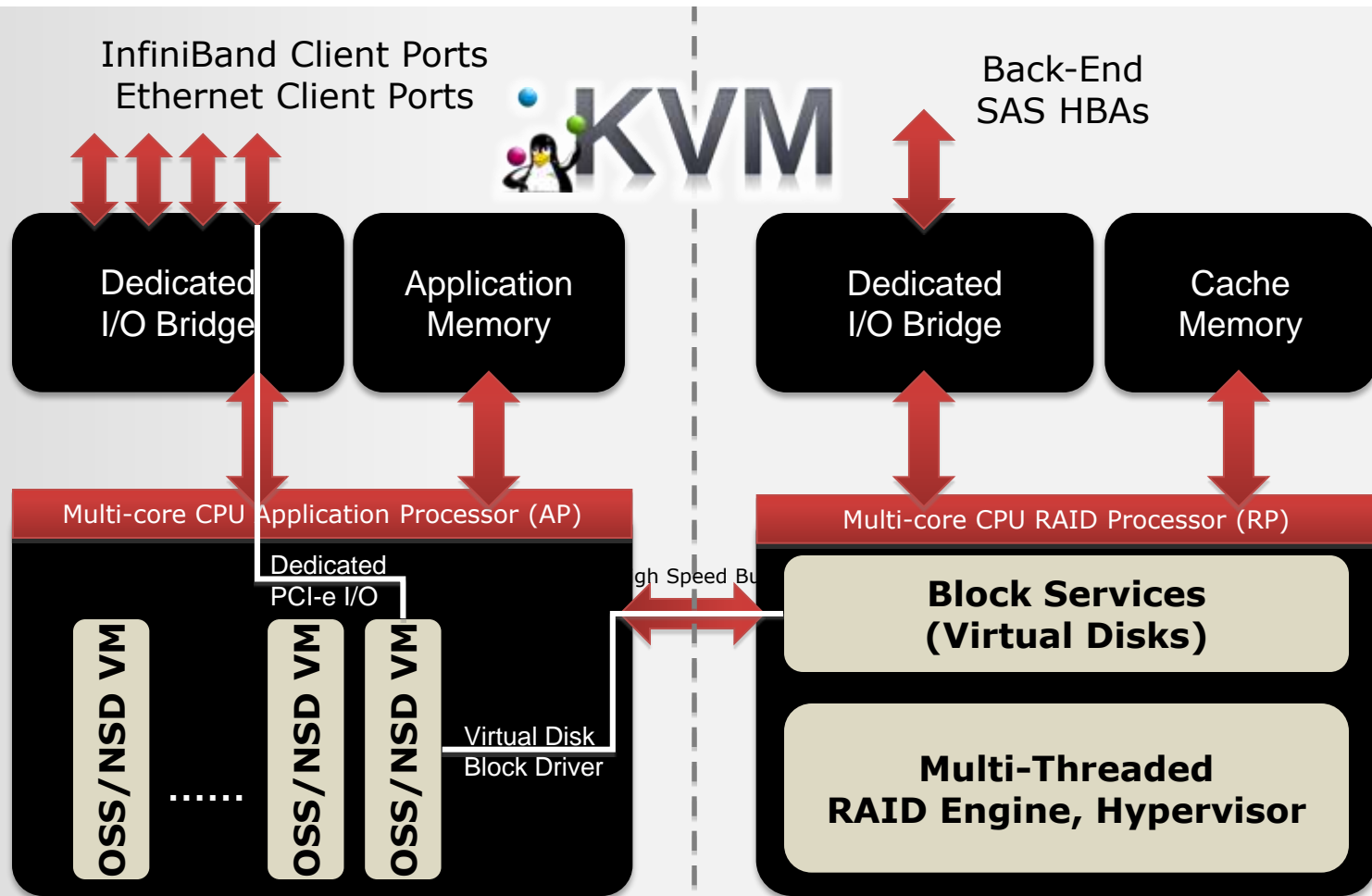

Storage Fusion Architecture shortens the IO path from the application to storage, reducing latency and increasing IOPS performance.

DataDirect NETWORKS
INFORMATION IN MOTION

Appliances, We've Got 'Em

- **Embed Apps Within the Storage Array**
  - Reduces complexity, infrastructure and administration
  - Reduces cost as well as lower operational cost
  - Increases performance – access to disk is direct

- **Appliance-ize File Systems**
  - Eliminate external networking and filer heads
  - Reduce file system licenses
    - Virtualization Enables Multi-Port Storage Servers
  - Increases performance, particularly for small I/O

DataDirect
N E T W O R K S
INFORMATION IN MOTION

In-Store Compute Virtualization

InfiniBand Client Ports
Ethernet Client Ports

KVM

Back-End
SAS HBAs

Dedicated
I/O Bridge

Application
Memory

Dedicated
I/O Bridge

Cache
Memory

Multi-core CPU Application Processor (AP)

Multi-core CPU RAID Processor (RP)

Dedicated
PCI-e I/O

gh Speed Bu

Block Services
(Virtual Disks)

OSS/NSD VM

......

OSS/NSD VM

OSS/NSD VM

Virtual Disk
Block Driver

Multi-Threaded
RAID Engine, Hypervisor

DataDirect
N E T W O R K S
INFORMATION IN MOTION

GPFS Embedded on SFA

FAST
Traditional block storage, direct attached to Parallel FS Servers

FASTER, SIMPLER
Servers embedded in 10K, improving performance and reducing network and servers

Parallel File System Clients

Parallel File System Clients

IB or 10Gig-E

IB or 10 Gig-E

File System
• Networking
• Server Infrastructure
• NSD Licenses

IB

IB or 10Gig-E

DataDirect
N E T W O R K S
INFORMATION IN MOTION

Thank You

Questions?

DataDirect™
N E T W O R K S
INFORMATION IN MOTION™