



Embedding dCache in a Scalable Storage System



The 10th International dCache Workshop
Barcelona, Spain

April 11, 2016

Dave Fellingner

Chief Scientist, DDN

dfellinger@ddn.com

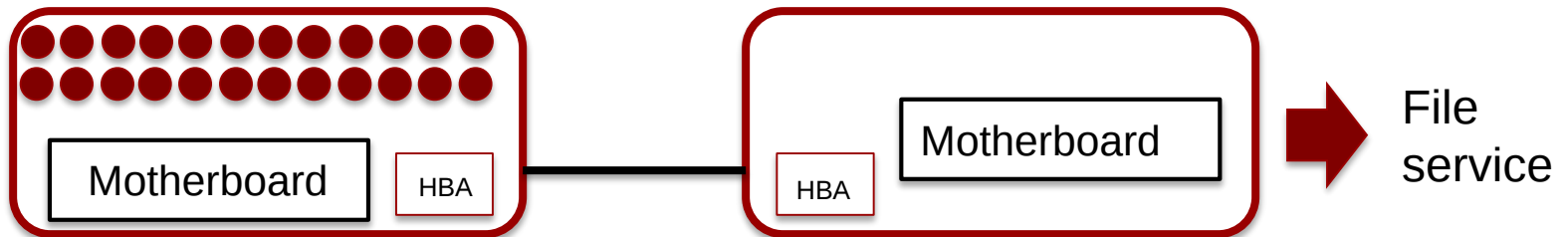
Traditional Data Service Model

► Server

- Dual Socket
- 600W Motherboard
- HBA or HCA

► Storage System

- Dual Socket
- 600W Motherboard
- HBA or HCA

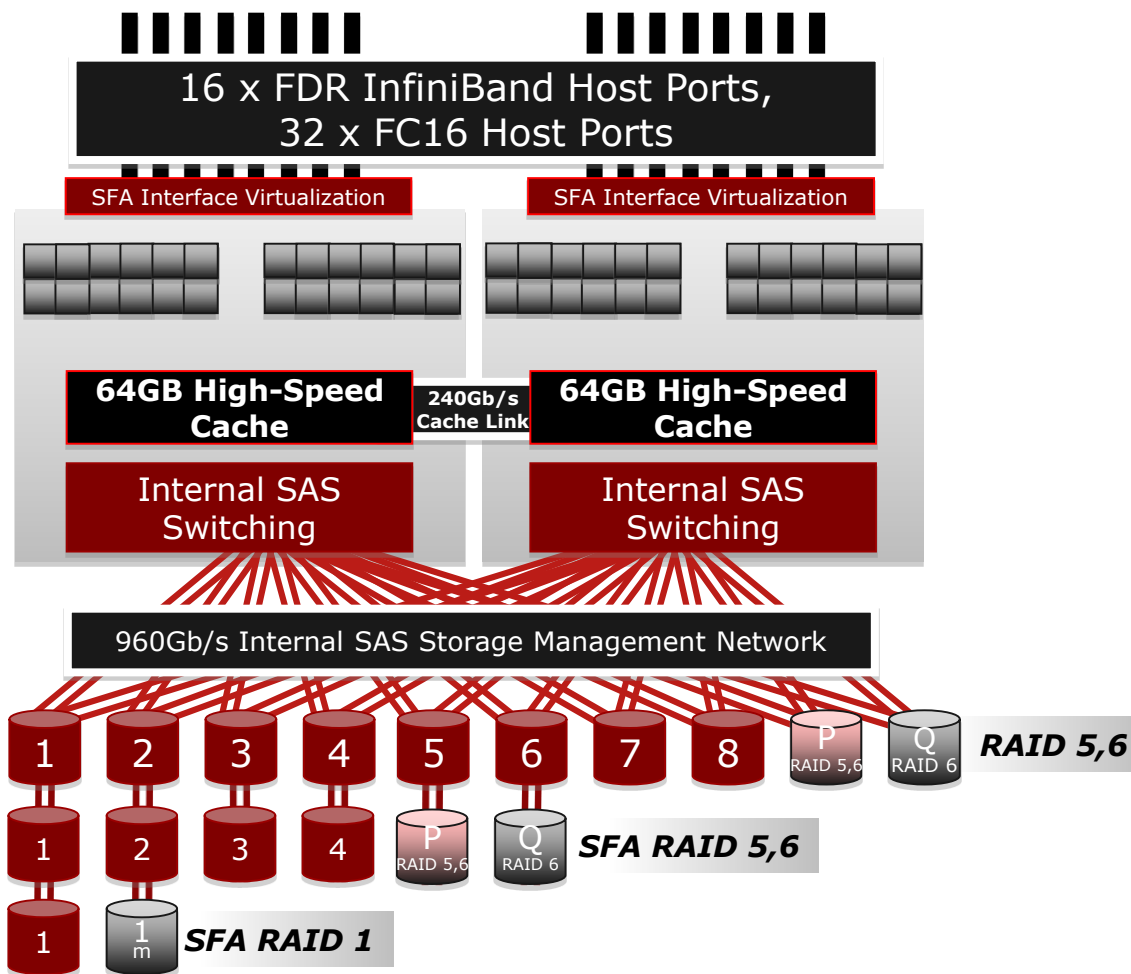


dCache12K Platform

Built Atop SFA12KX-E

Highly Parallelized SFA Storage Processing Engine

Active/Active Storage Design
 40-44GB/s Read & Write Speed
 Up to 6.7PB of Disk
 2.4+ Million Burst IOPS
 700K+ Random Spinning Disk IOPS
 1.4M Sustained Random SSD IOPS
 64GB Mirrored Cache (Protected)
 Up to 12.8 TB Flash Cache using SFX
 RAID 1/5/6
 Intelligent Block Striping
 DirectProtect™
 GUI, SNMP, CLI, API
 16 x FDR IB Host-Ports
 32 x FC16 Host-Ports
 8RU Height

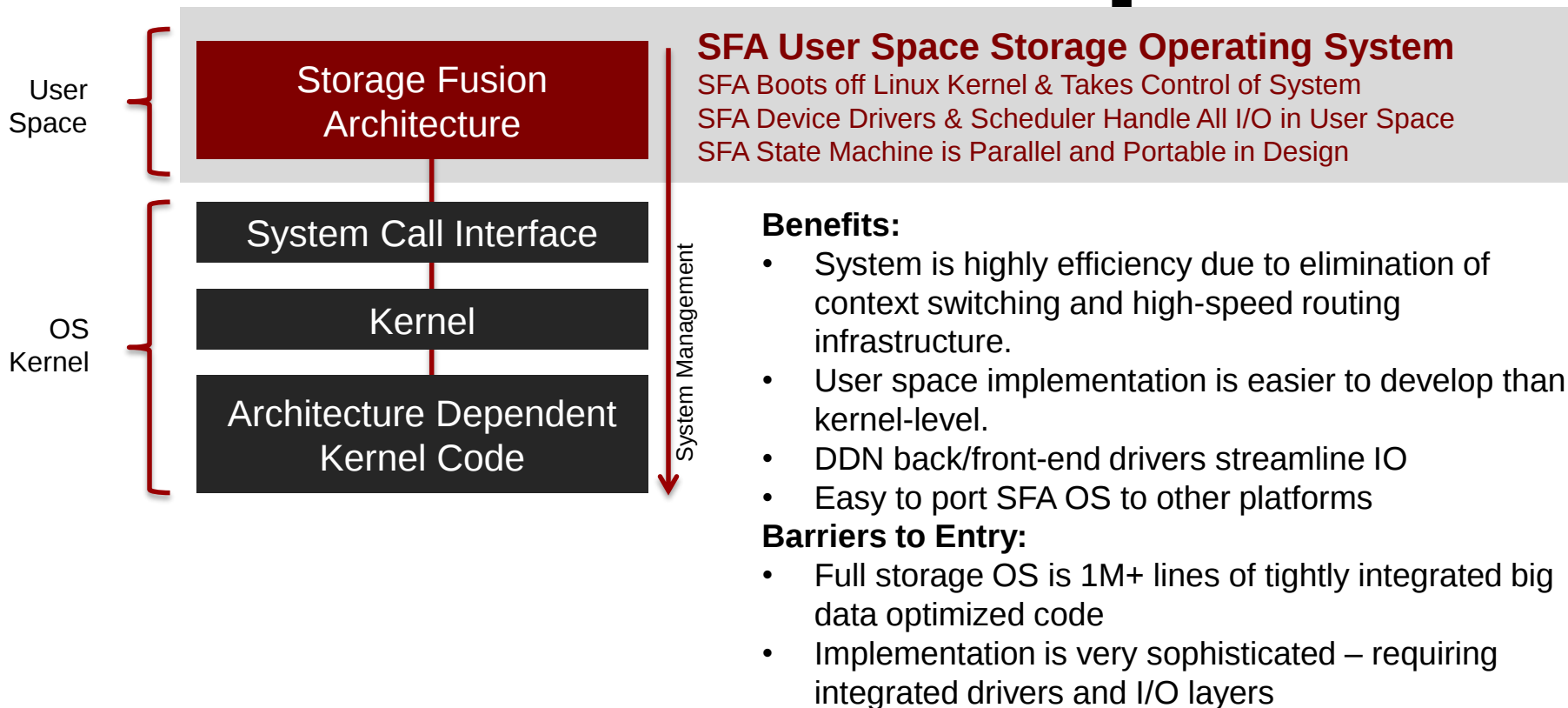


SFA User Space Storage OS

A Scalable Storage OS Implemented In User Space

SFA Implementation

Tight Integration W/ Virtualization is Ideal for Converged Storage



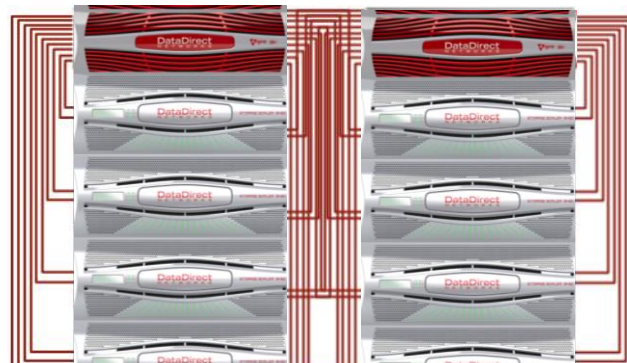
ALL of the benefits of an in-kernel implementation • NONE of the limitations of kernel/HW dependency.

dCache 12K Platform

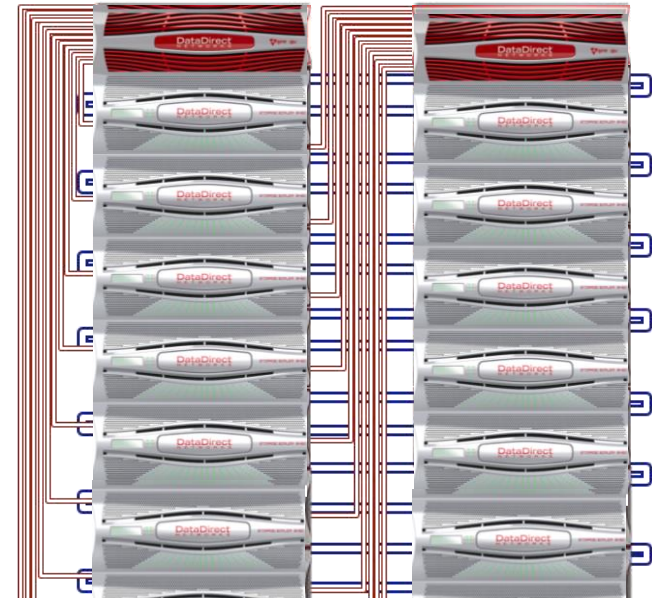
Variable System Sizing



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



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

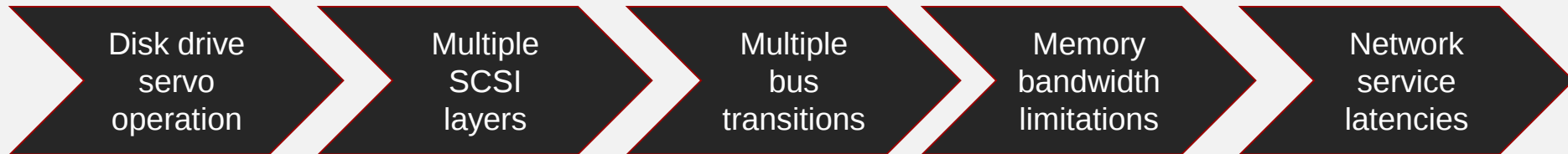


20 Enclosure System
Up to 1,680 Drives
2 BBUs, 88U

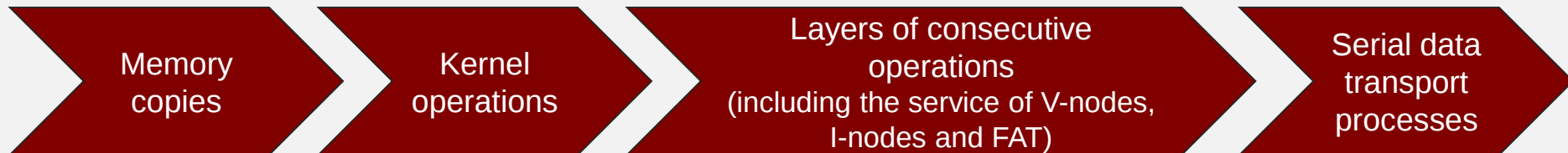
High Availability Drive Channel & Enclosure RAIDing

System Level Sources of Latency

HARDWARE CHAIN

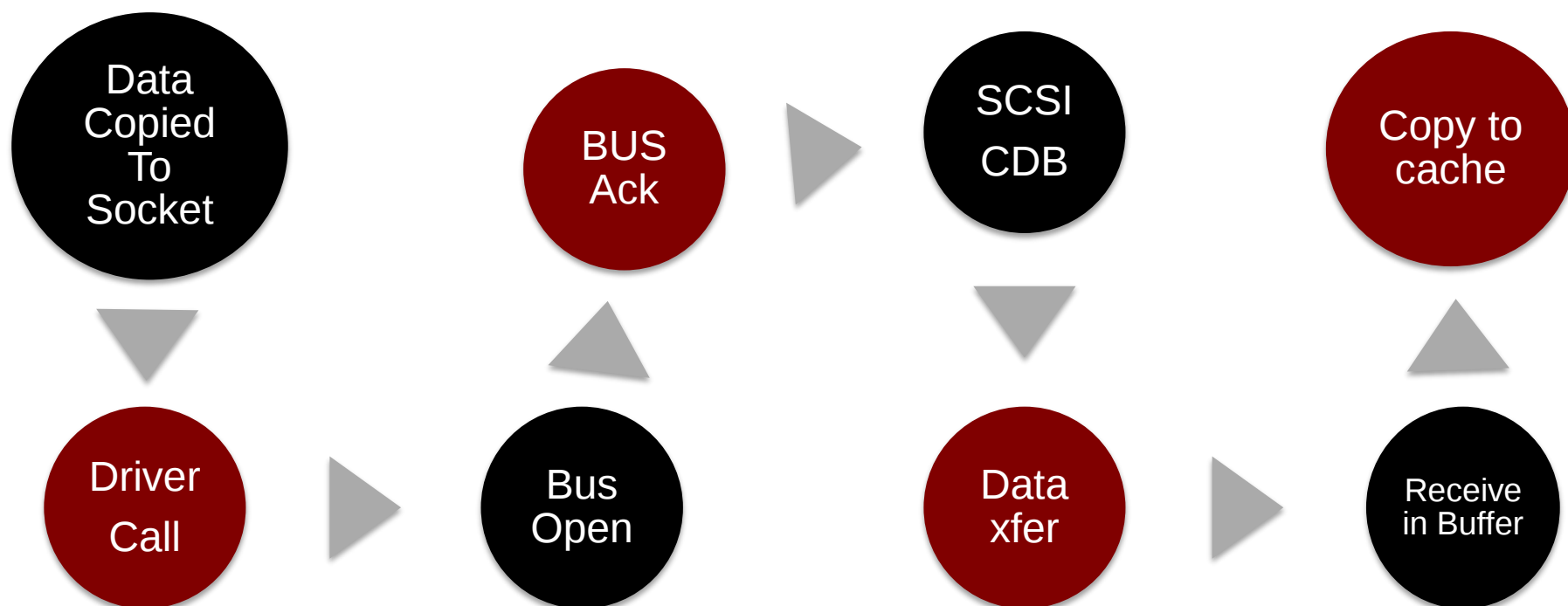


SOFTWARE CHAIN



“Traditional” File Access

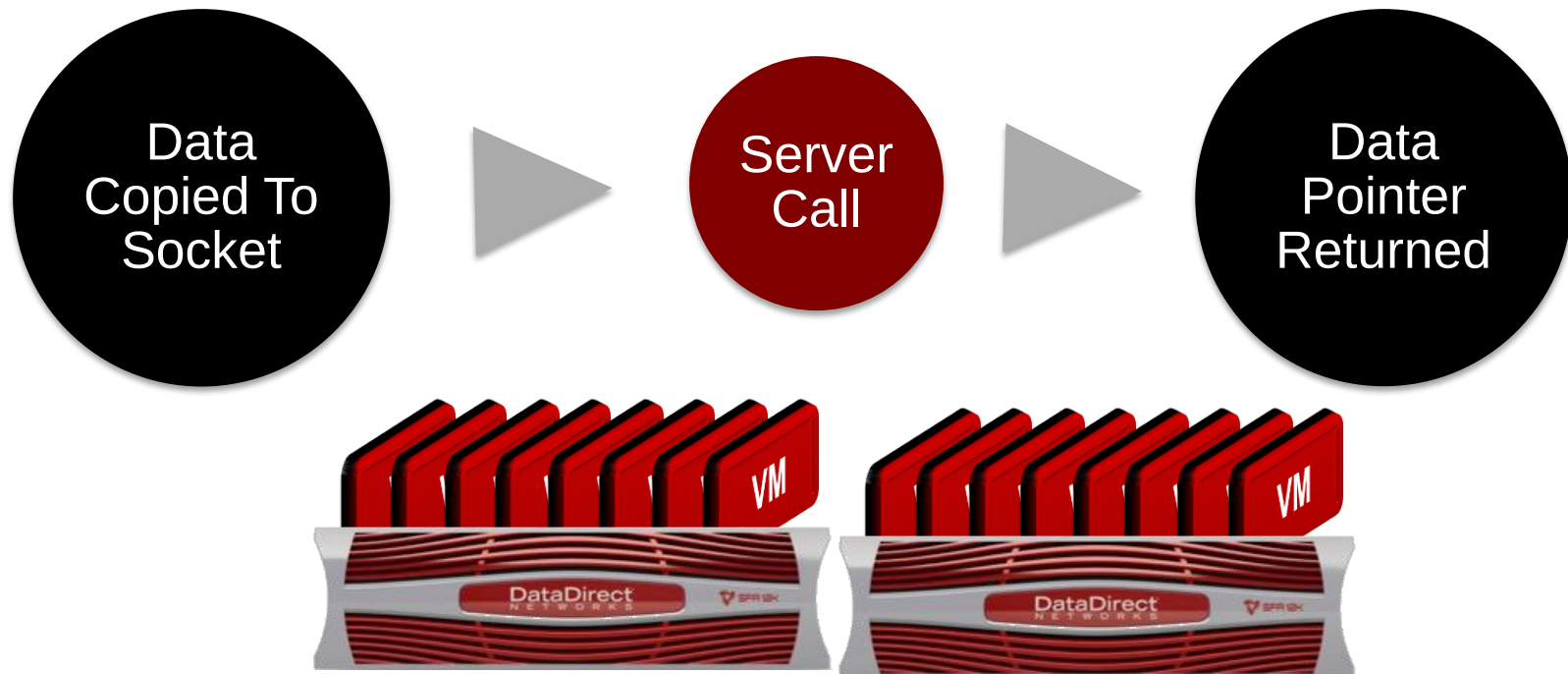
- ▶ File servers are connected to storage devices by serializing devices such as HBAs or HCAs
- ▶ Multiple steps executed to move data to/from a server



dCache 12K Platform

More 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



Case Study: TRIUMF ATLAS

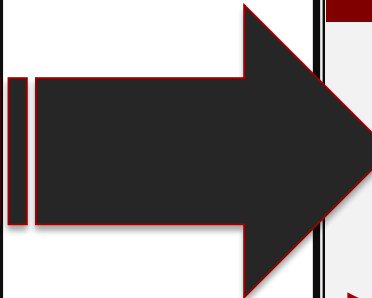
Tier-1 Data Center



BEFORE



- ▶ Qty 21 Sunfire® Disk Servers (in 4 half racks)
- ▶ Qty 1,008 HDDs (both 1TB & 2TB)
- ▶ Total 1.87PB (Raw)
- ▶ Connections: Qty 21 4GigE



AFTER

**28
RACK
UNITS!**



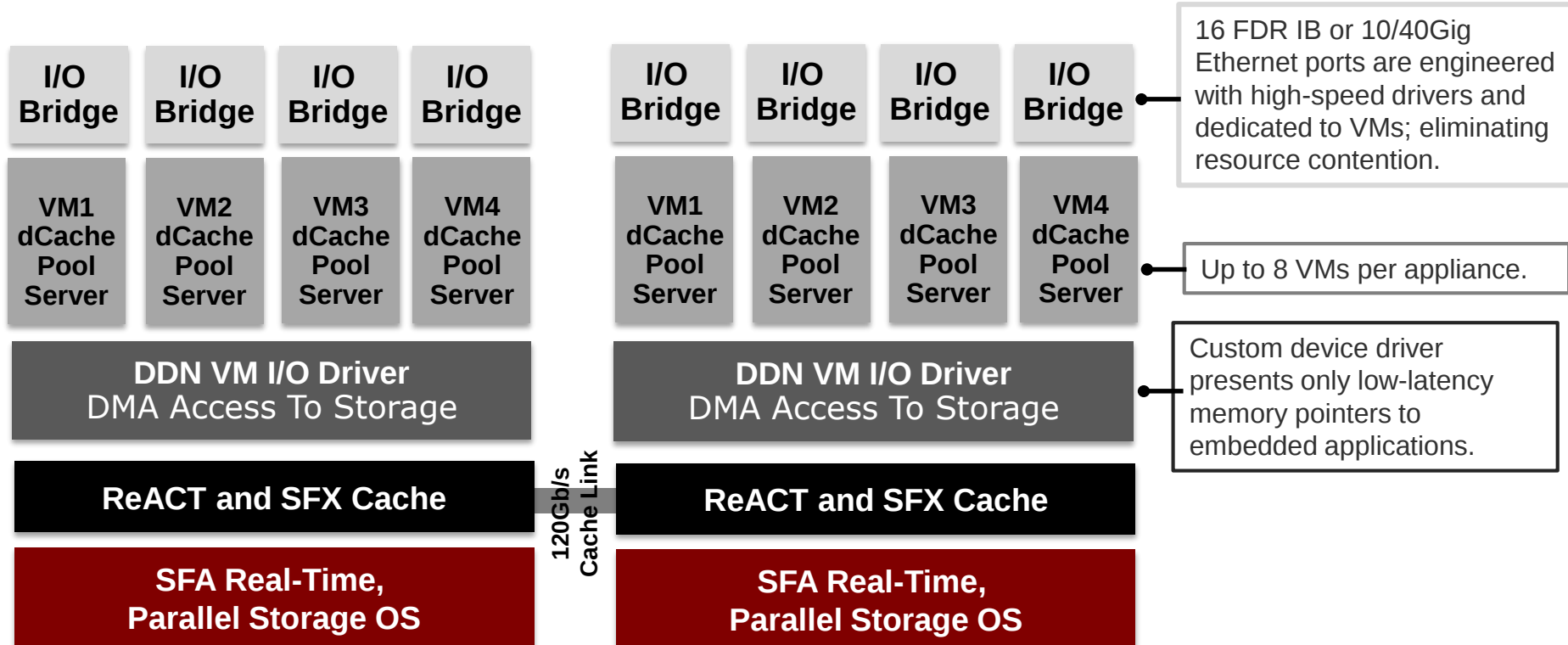
- ▶ Qty 1 SFA with 5 84 slot HDD Enclosures
- ▶ Total 1.6PB (Raw)
 - ▶ Qty 400 4TB SATA
 - ▶ Qty 4 400GB SSDs
- ▶ Connections: Qty 8 10GigE

dCache 12K Platform

Server-less Scale-Out Storage

1st Open Array Architecture to Collocate Compute with Storage

Eliminates file system gateways • Reduces license expenditures • Simplifies management



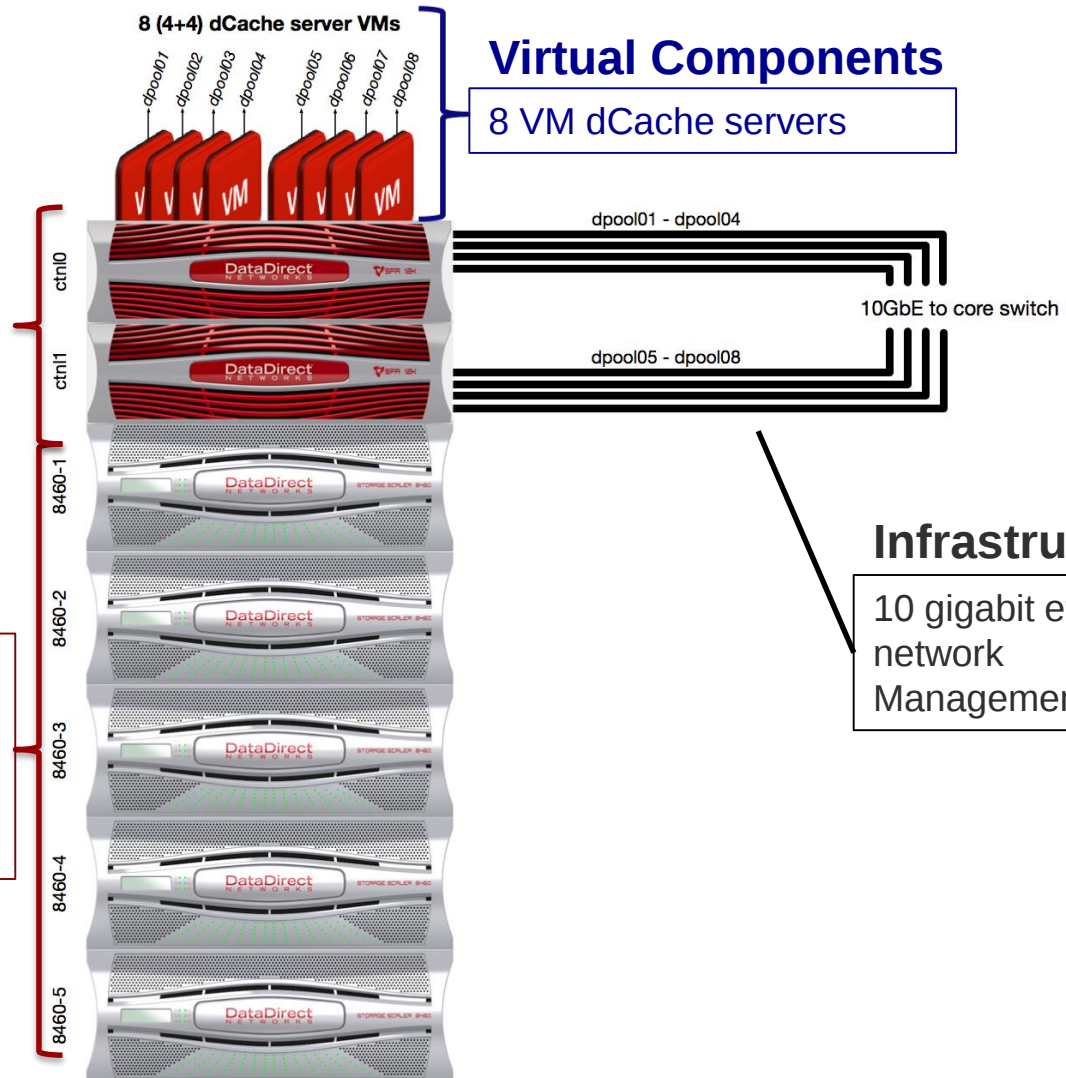
dCache 12K Embedded – Hardware

Physical Components

SFA12K20E Active-Active
Controller Pair

QTY 5 84 Slot Disk
Enclosures

QTY 400 NL SAS 3TB HDDs
QTY 4 400GB SSDs



dCache Embedded – LUN Allocation



dCache Embedded

Stack specifics / Tuning

dCache Pool Server VM O/S and stack component versions

- Scientific Linux 6.4
- Kernel 2.6.32-431.29.2.el6.x86_64
- Mellanox mlx_en driver 1.5.10, MTU 9000
- SFA block driver initially 2.0.0.3-17400 – recently updated to 2.3.0.1-2555
- xfsprogs 3.1.1-10

dCache specifics

- dCache version 2.6.34
- dcache.java.memory.heap=2048m
- Java version 1.7.9-17
- 2 movers for gridftp, 3 for dccp
- 1 VM = 1 dCache pool domain, 5 pools/VM
- Checksum type – ADLER32
- Some background impact from production client nodes during testing during lcg-cp testing

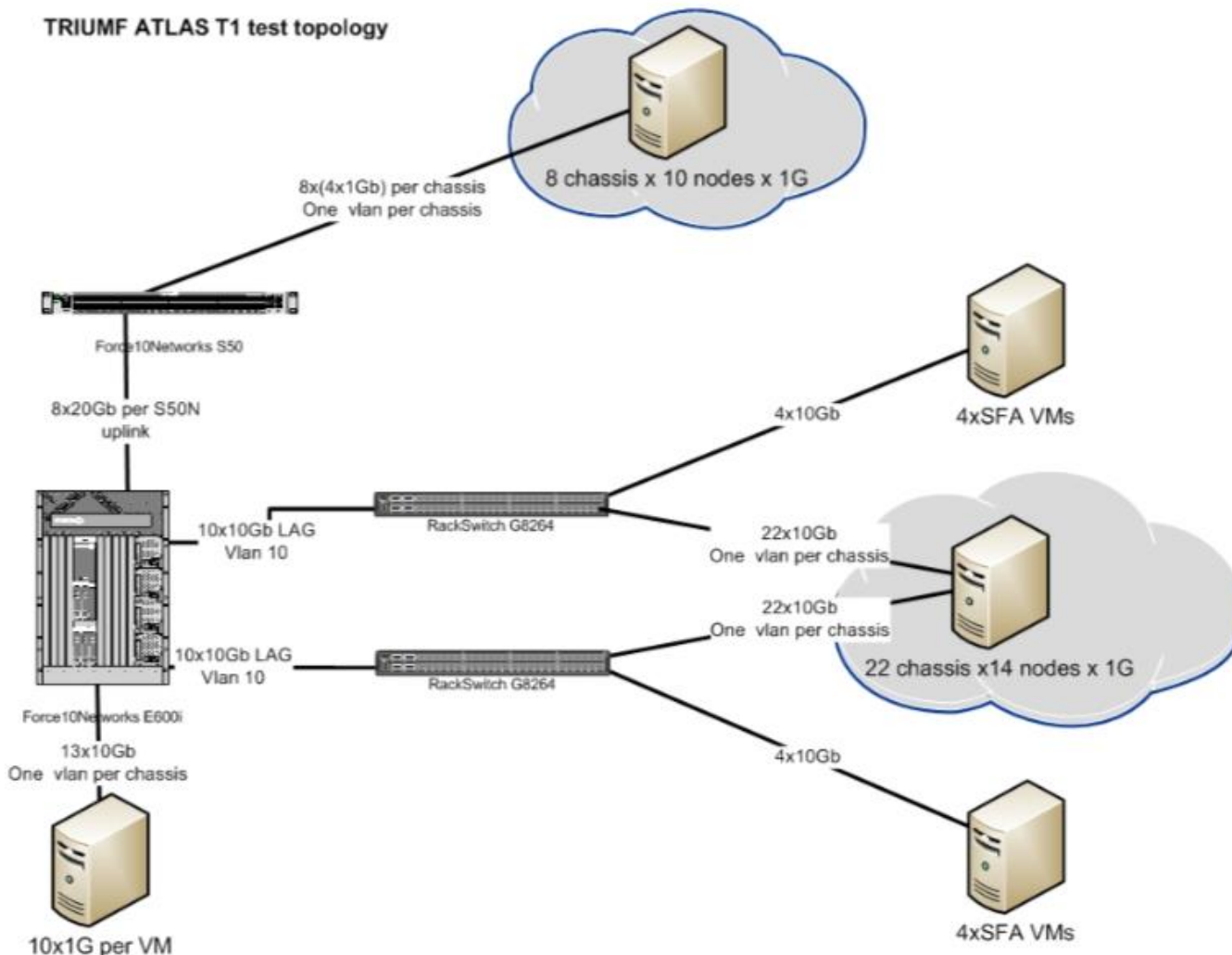
sysctl.conf tuning

```
net.ipv4.tcp_timestamps=0 net.ipv4.tcp_sack=1
net.ipv4.tcp_window_scaling=1
net.ipv4.tcp_low_latency=1
net.core.netdev_max_backlog = 250000
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.ipv4.tcp_rmem = 4096 87380 524288
net.ipv4.tcp_wmem = 4096 65536 524288
vm.swappiness=60
vm.dirty_expire_centisecs=1000
vm.dirty_writeback_centisecs=500
vm.dirty_background_ratio=5
vm.dirty_ratio=80
vm.min_free_kbytes = 262144
```

SFAOS version and settings

- SFAOS version 2.3.0.1-2555
- 8460 f/w – 0121-145
- DiF – DirectProtect
- ReACT – TRUE
- Read_ahead – TRUE
- Write Back Cache - TRUE
- Cache Mirroring – TRUE
- Pool verify priority – 10%
- 1 hot spare pool

dCache Embedded Topology at TRIUMF

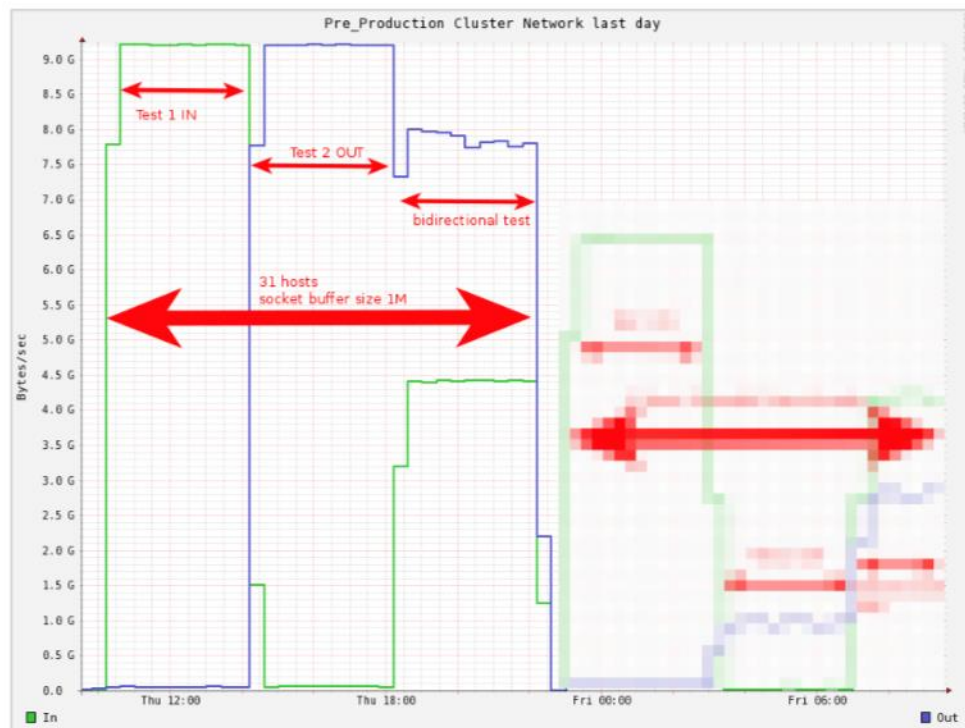


courtesy of TRIUMF
Simon Liu

dCache Embedded Network Performance Testing



Client	Server	Bandwidth	Duration	Threads	Direction	Tuning	
31xWN	8xDDN VM	8 GB/s	8 hour	1	IN	Window sz 1M	Test 1
8xDDN VM	31xWN	9 GB/s	8 hour	1	OUT	Window sz 1M	Test 2
31xWN	8xDDN VM	7.5 GB/s	8 hour	1	IN	Window sz 1M	Bi-dir
8xDDN VM	31xWN	4.9 GB/s	8 hour	1	OUT	Window sz 1M	Bi-dir



Typical aggregate network iperf using:

- QTY 8 DDN VM server hosts
- 31 blade client hosts

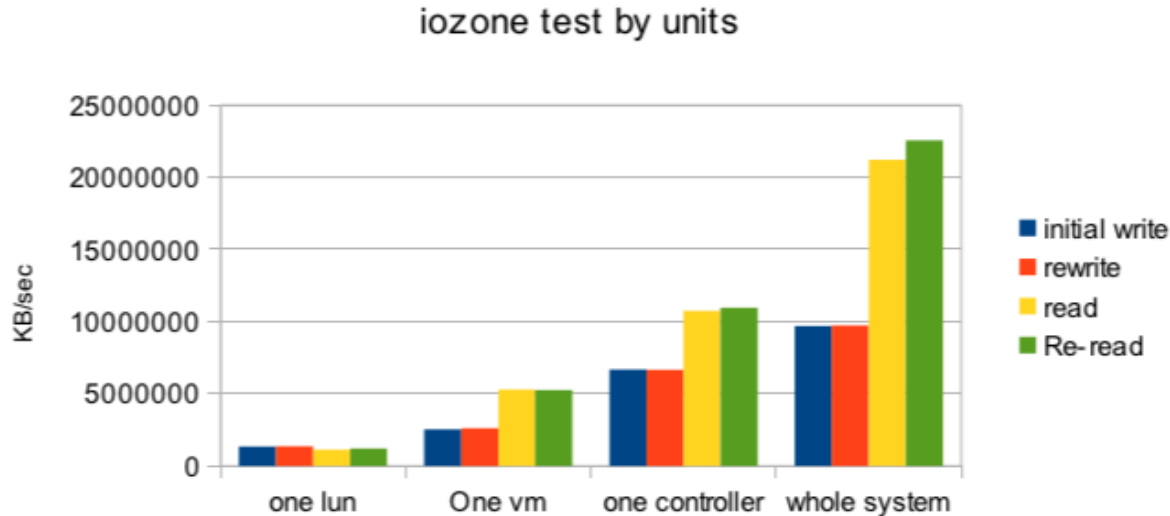
Ganglia plot shows GB/s in 8 hour segments

- From clients to servers
- Servers to clients
- Bi-directionally

NOTE: During subsequent storage tests we decided to move the VM 10 Gbit connections to the core router.

courtesy of TRIUMF
Simon Liu

dCache Embedded Storage Performance Testing (iozone)



One VM has 5 luns mounted, one controller has 4 VMs, whole system has 8 VMs with 39 luns mounted

"From a dCache performance perspective, we tuned the VM server configuration such that it matches the performance we get from physical server implementations in our datacentre. We are happy with the performance and reliability of this system configuration thus far. It is being integrated into our production system as we decommission the end-of-life Thor cluster."

TRIUMF

courtesy of TRIUMF
Simon Liu

dCache Embedded storage performance testing (lcg-cp)



courtesy of TRIUMF

Writes: 8VMs, 80 *10 writing threads
5.3-5.6 GB/s sustained

Reads: 8VMs, 120 dcp processes
7-8 GB/s sustained

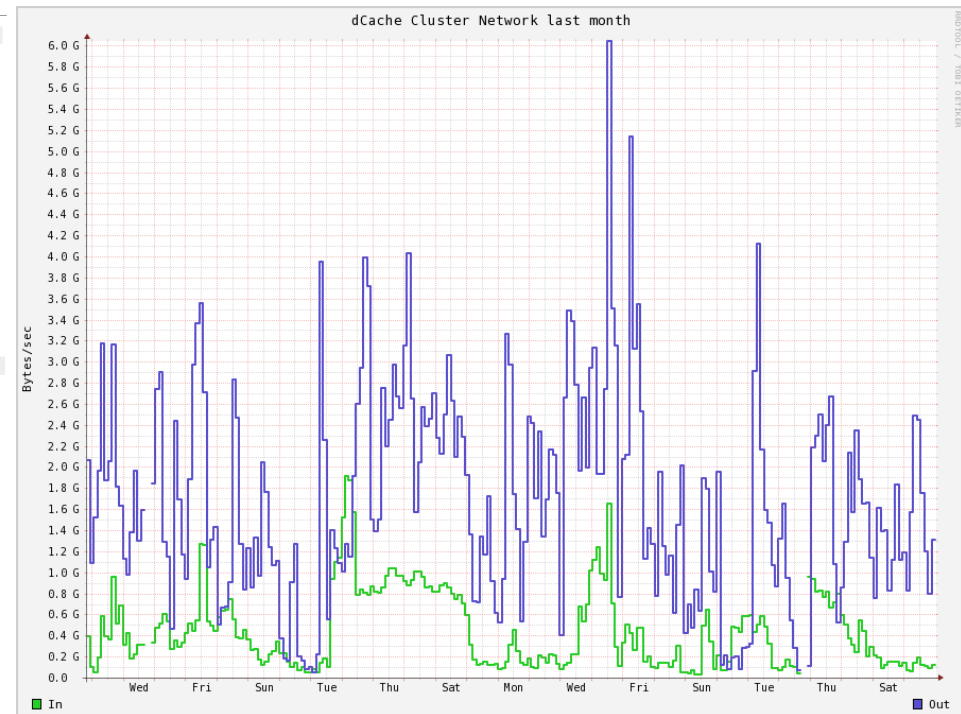
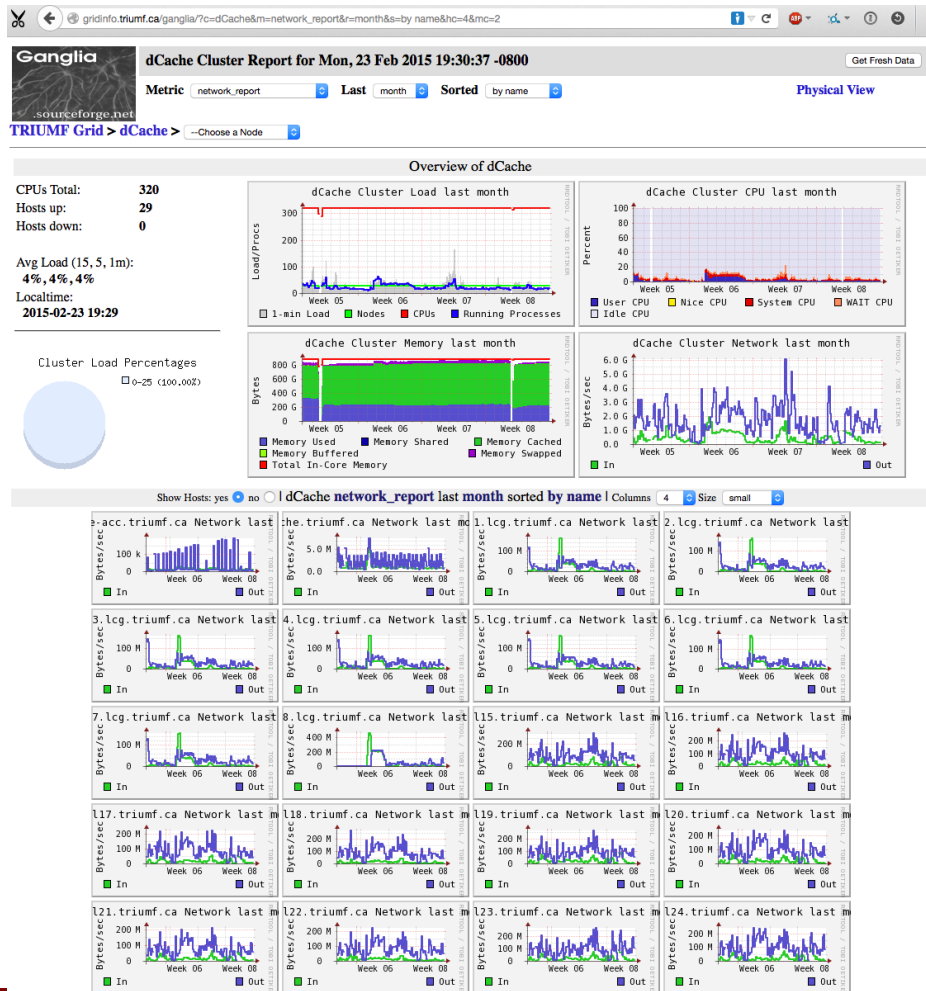
Read/Write: 4-4.5 + 2.8-3.4 GB/s sustained

18

dCache Embedded Monitoring



<http://gridinfo.triumf.ca/ganglia/?r=month&s=by%2520name&c=>



dCache 7K Platform

Looking for a Smaller Scalable Building Block?

2 x Active Active Storage Appliances Embedded

- Dual Socket 12 core Ivy Bridge @2.8GHz
- 128GB memory
- Mellonox ConnectX-3 cards

Up to 60 Slots in Base Unit

- SAS, SATA, SSD



Tested with 8 clients
140 NLSAS drives

14 Pools RAID6 8+2
128KB chunk size
DIF enabled
WB cache enabled
Mirror Cache enabled
Read Ahead Enabled
ReAct Enabled

2VMs 14 disk pools 6 cores/VM
Curl test results
2GB/sec/VM read and write

dCache 14K

Looking for a Larger Scalable Building Block?

2 Compute Controllers

- SFAOS 3.0
- 4x 14-core Haswell
- 256GB DDR4 memory
- Redundant Power

Connectivity

- 8 x IB EDR/FDR ports
- 10/40/100 GbE
- DDN Omni-Connect

Hyperconverged

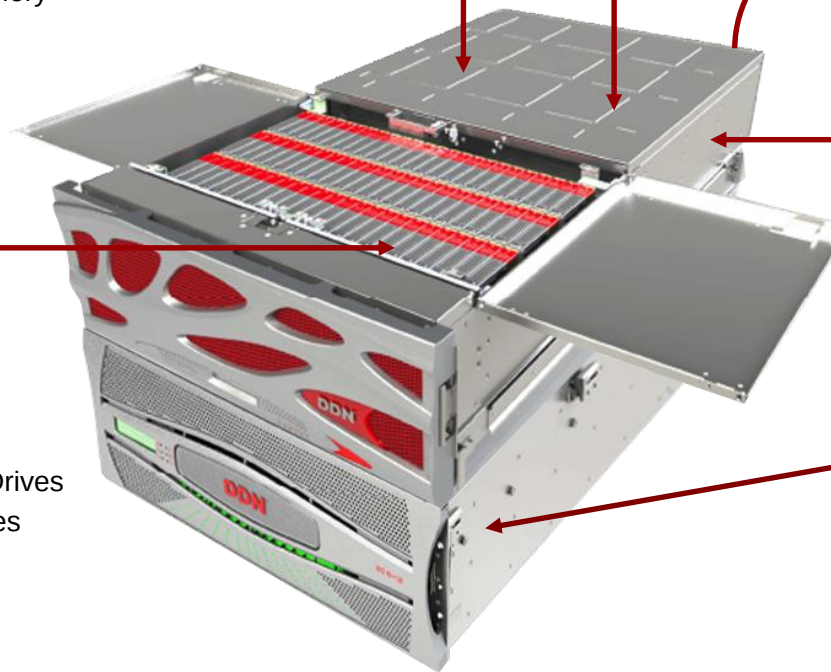
- Embedded PFS
- Applications

Drives

- 72 x 12Gb/s SAS 2.5" slots (48 x PCIe NVMe)
- 4 x 480GB SATA SSD System Drives
- 4 x 480GB SATA SSD App Drives

84 Slot Expansion Enclosure

- Add 1, 2, 4, 5, 10 or 20
- 84 Drives Each



Scale Flexibly & Efficiently

Three Right-sized, At-scale Building Blocks

	dCache 7700	dCache 12K	dCache 14K
Throughput For dCache	4GB/s	8GB/s	20GB/s
Host Ports	4X 56Gbit/s FDR <u>or</u> 4X 40GbitE	4X 56Gbit/s FDR <u>or</u> 16X 10/40GbitE	8X EDR/FDR <u>or</u> 8X DDN OmniConnect
Drive Types	SAS/SATA SSDs Performance SAS HDD Capacity SAS, SATA HDDs	SAS/SATA SSDs Performance SAS HDD Capacity SAS, SATA HDDs	NVMe, SAS SSDs Performance SAS HDD Capacity SAS HDD
Drive Size	Base: 3.5" Enclosure: 3.5"	Enclosure: 3.5"	Base: 2.5" Enclosure: 3.5"
Drives in Base Unit	60	N/A	72
Total Drives (w/ expansion enclosures)	396	1,680	1,752

Conclusion

**The utilization of
dCache in a virtual
storage system
environment**

**Is a very low latency
implementation of
software defined
storage**

Eliminating external servers enables . . .

**1.
Ease
of
system
management**

**2.
Reduction of
external bus and
switch
architecture**

**3.
Consistent performance
for all transfer sizes
through the reduction of
bus protocol**

Thank You!

Keep in touch with us

Questions?

dfellinger@ddn.com



sales@ddn.com



2929 Patrick Henry Drive
Santa Clara, CA 95054



[@ddn_limitless](https://twitter.com/ddn_limitless)



1.800.837.2298
1.818.700.4000



[company/datadirect-networks](https://www.linkedin.com/company/datadirect-networks)