LUG2011

Virtualization of Lustre for QA and Benchmarking

DataDirect Networks Japan, Inc.

Senior Solutions Architect

Shuichi Ihara

DataDirect™ NETWORKS

Infrastructure Challenges

- Massive number of dedicated servers required for testing.
- Challenges with setting up Networking for all the servers.
- Proliferation of hardware and operating system configurations to be tested.
- Infrastructure requires massive amounts of Rack Space and Power.
- CPU cores are increasing – Single threaded applications need to take advantage of it.

Virtualization technology can help us here

KVM (Kernel-based Virtual Machine)

- Strong candidates for Lustre QA infrastructure
- The code is already merged in the Linux mainstream
- Supported by many Linux distributions (SLES11, RHEL5.x, 6, Ubuntu)
- Supports full virtualization
  - No special patched kernel needed on VMs
  - Ability to create an quasi-real server environment
  - Supports CPU affinity and PCI passthrough
  - SR-IOV (future)
- CLI is available and it's scriptable!
  - Easy automation and administration

Benefits

•Less need for physical work in the lab ☺

•Fast implementation of Lustre QA infrastructure

✓For Lustre sanity testing on many types of H/W configuration.

✓For function testing (HA, LNET routing, etc..).

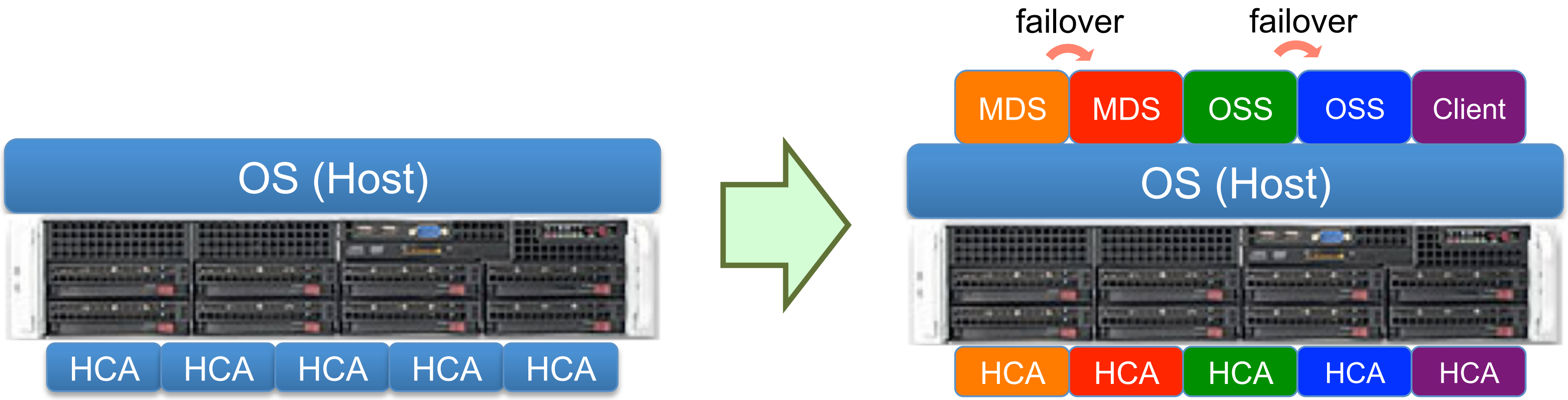
✓For benchmark use.

• To build Lustre RPMs

✓Build systems for different Linux distributions

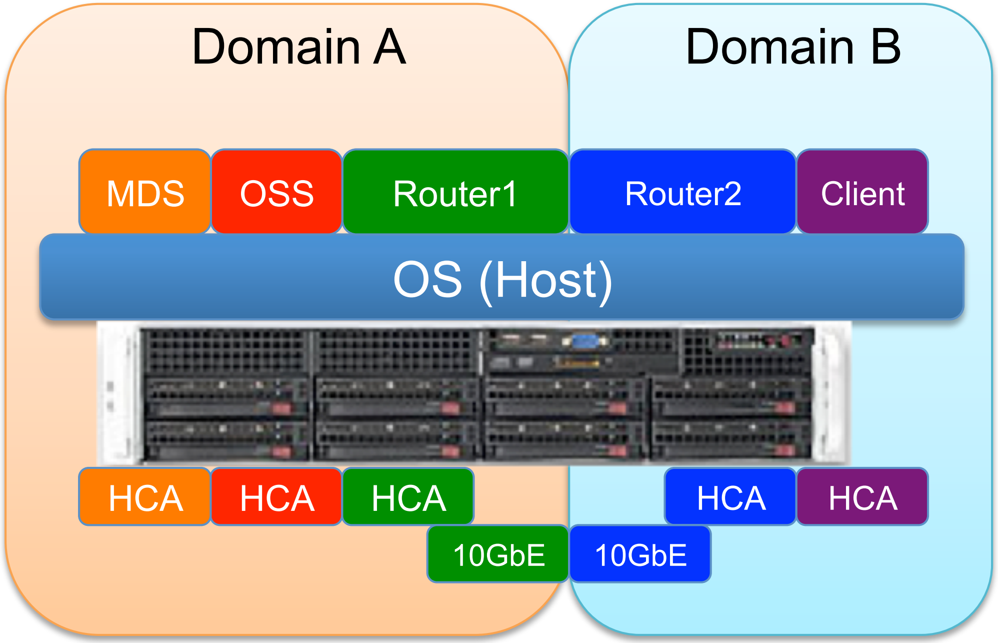
Lustre on VM example(1)
- HA testing -

- **Lustre HA testing**
  - VMs : 2 x MDS, 2 x OSS and 1 x Client
  - Attach QDR Infiniband HCA to each VM



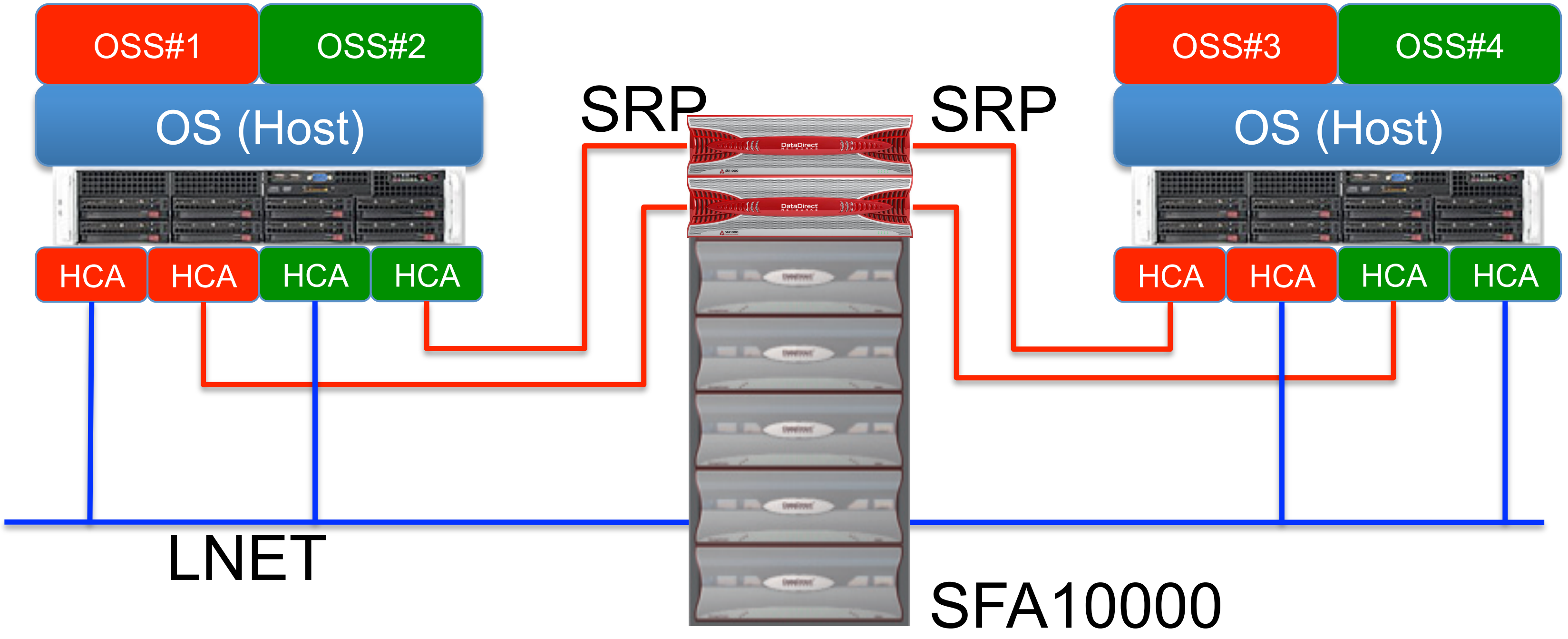
Lustre on VM example(2) - N-hop Routing -

- **Testing for 2-hop routing (IB <-> 10GbE <-> 10GbE <-> IB)**
  - 5 VMs : 1 x MDS, 1 x OSS, 2 x Router and 1 x Client
  - Attach a QDR Infiniband HCA to each VM; Router VM also have 10GbE connections



Lustre on VM example(3)
- Benchmark use -

- **For Lustre Benchmark**
  - 2 VMs : 2 x OSS per physical server
  - Attach two QDR Infiniband HCAs to each VM. (One for connecting to Storage, another one for LNET)



Lustre performance on KVM

DataDirect™
NETWORKS

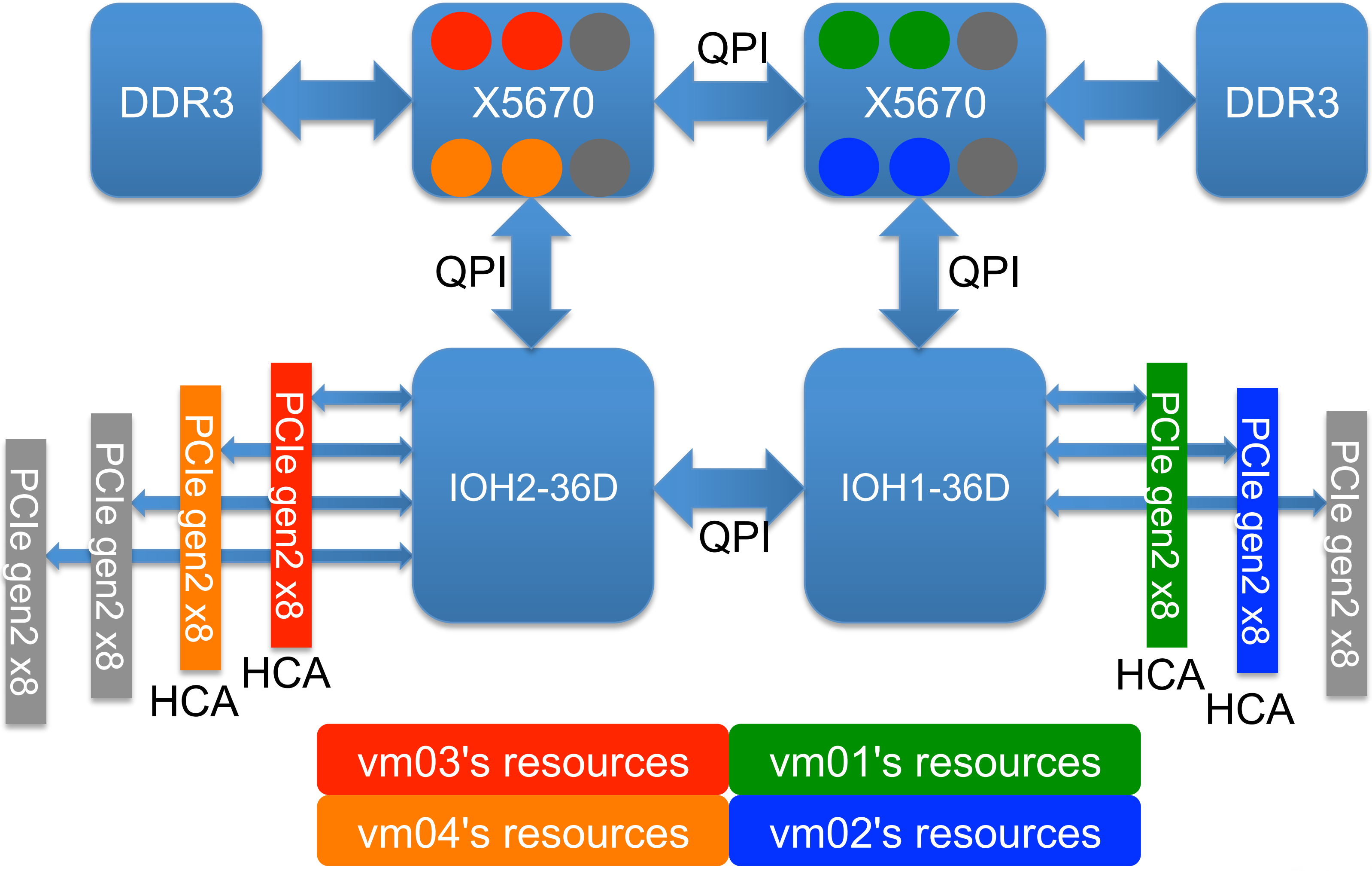
Benchmark configuration

- **SuperMicro's SuperServer**
  - 2 x Intel Xeon (X5670, 2.93GHz), 48GB Memory
  - 2 x Tylersburg (IOH-36D), 6 x PCIe gen2 slots
- **PCI devices**
  - 4 x Mellanox QDR HCA
  - Dual ports and works as 10GbE/QDR hybrid network card
- **Software**
  - RHEL6 (Host), CentOS5.5(VM)
  - Lustre-1.8.4

Type of benchmark

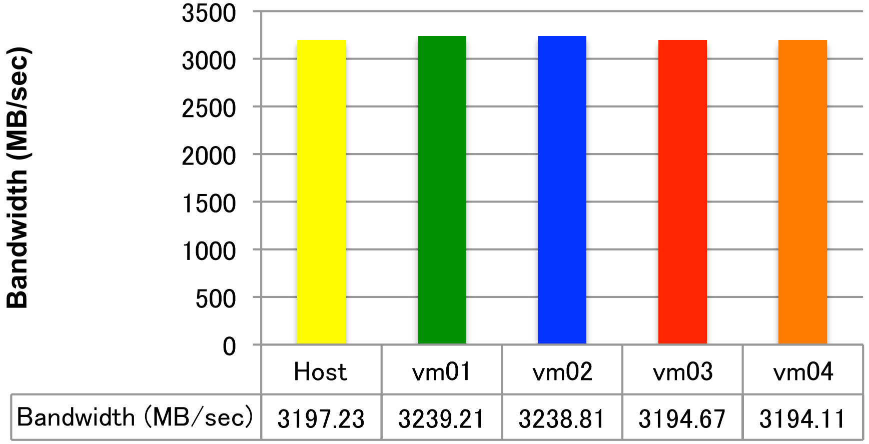
- **Network performance**
  - RDMA (bandwidth, latency)
  - LNET Selftest
- **Lustre backend performance**
  - obdfilter-survey
- **Lustre performance from the clients**
  - IOR

Network performance
- Physical resource assignment for VMs -

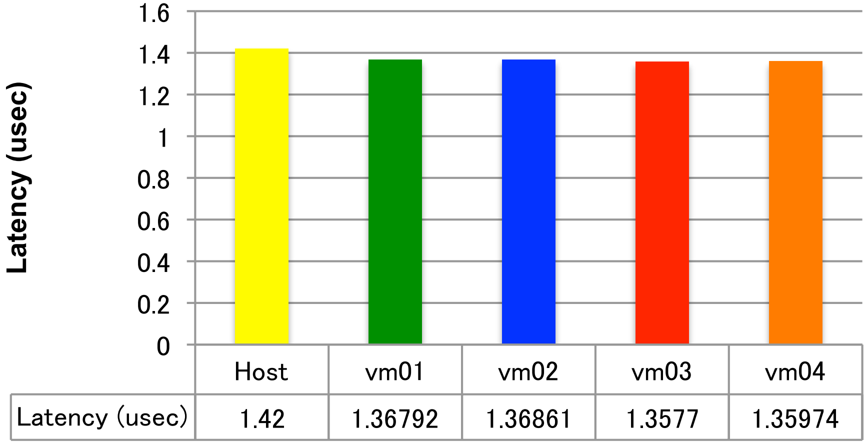


Network performance
- RDMA Benchmark results -

- Tested with rdma_bw, rdma_lat
- No performance differences when compared to non-VM (Host).

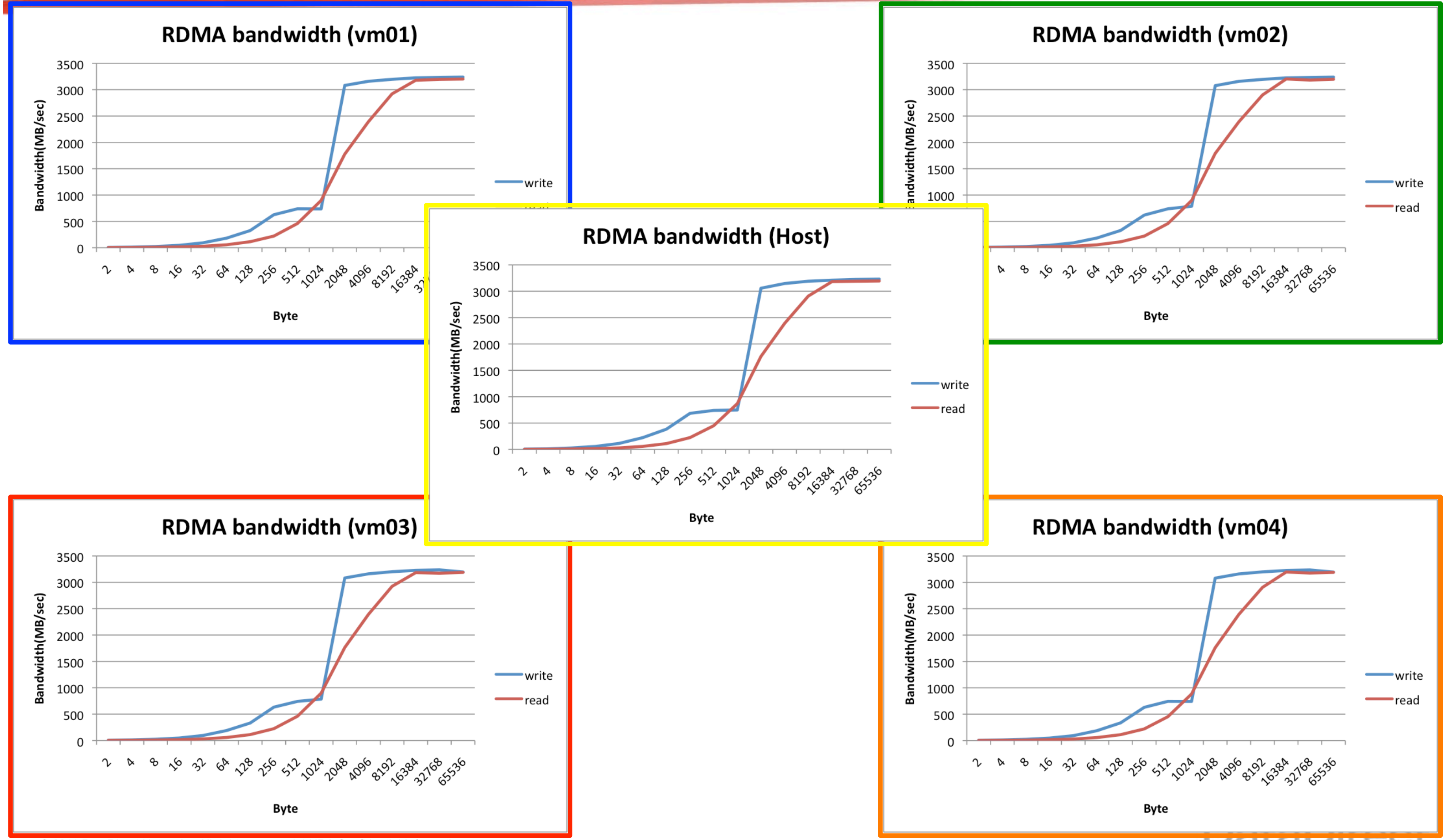

RDMA Bandwidth

	Host	vm01	vm02	vm03	vm04
Bandwidth (MB/sec)	3197.23	3239.21	3238.81	3194.67	3194.11

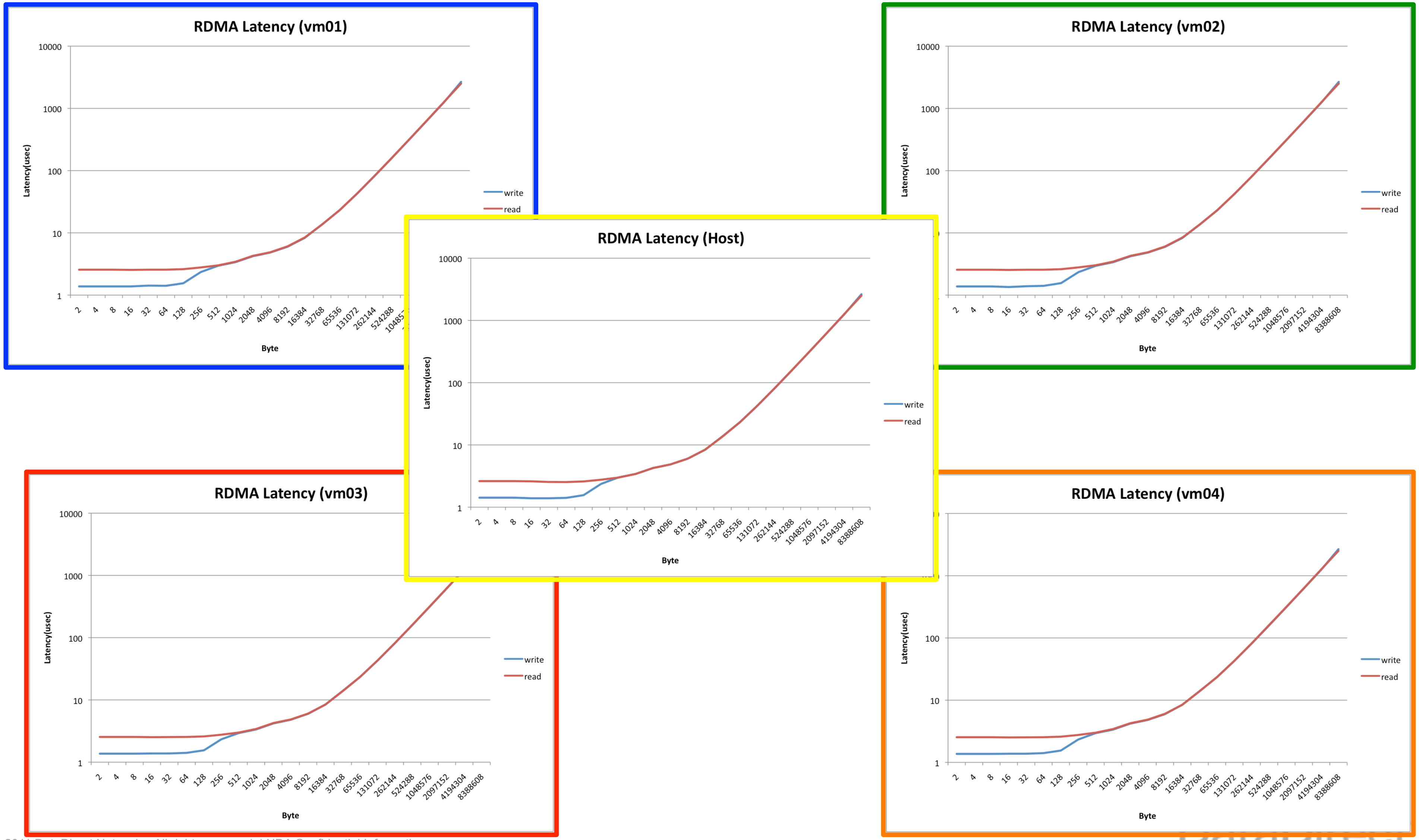

RDMA Latency

	Host	vm01	vm02	vm03	vm04
Latency (usec)	1.42	1.36792	1.36861	1.3577	1.35974

Network performance
- RDMA Benchmark (Bandwidth) in detail -



Network performance
- RDMA Benchmark (Latency) in detail -



RDMA Latency (vm01)

RDMA Latency (vm02)

RDMA Latency (Host)

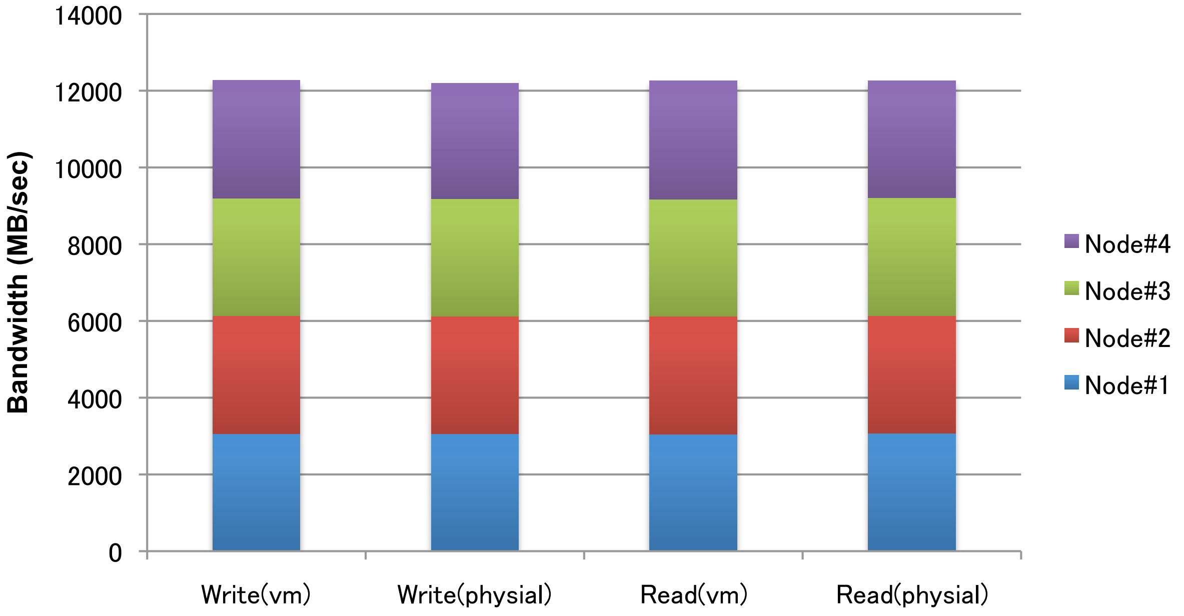
RDMA Latency (vm03)

RDMA Latency (vm04)

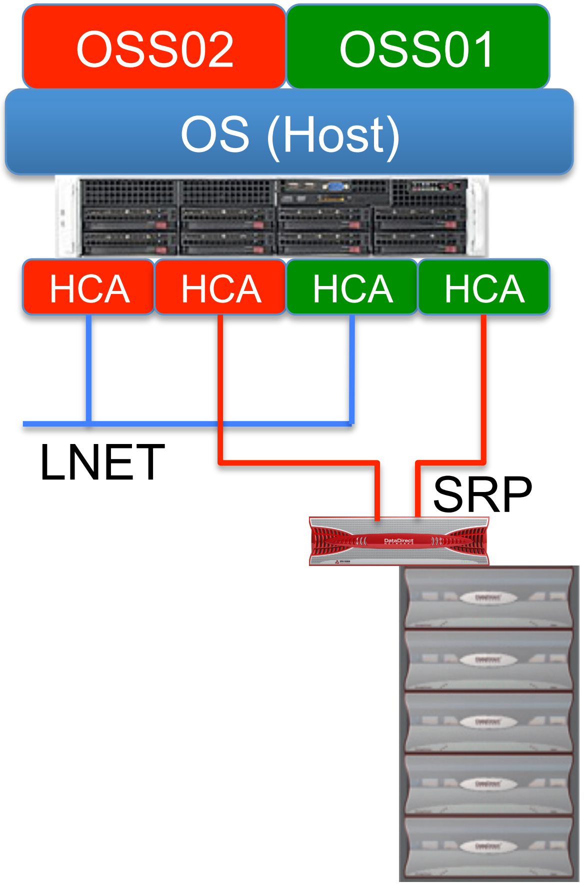
Network performance
- LNET selftest results -

- Tested on 4 servers and 4 clients
- Compared the Lustre clients on VM and non-VM

LNET selft (comparing VM and nonVM)

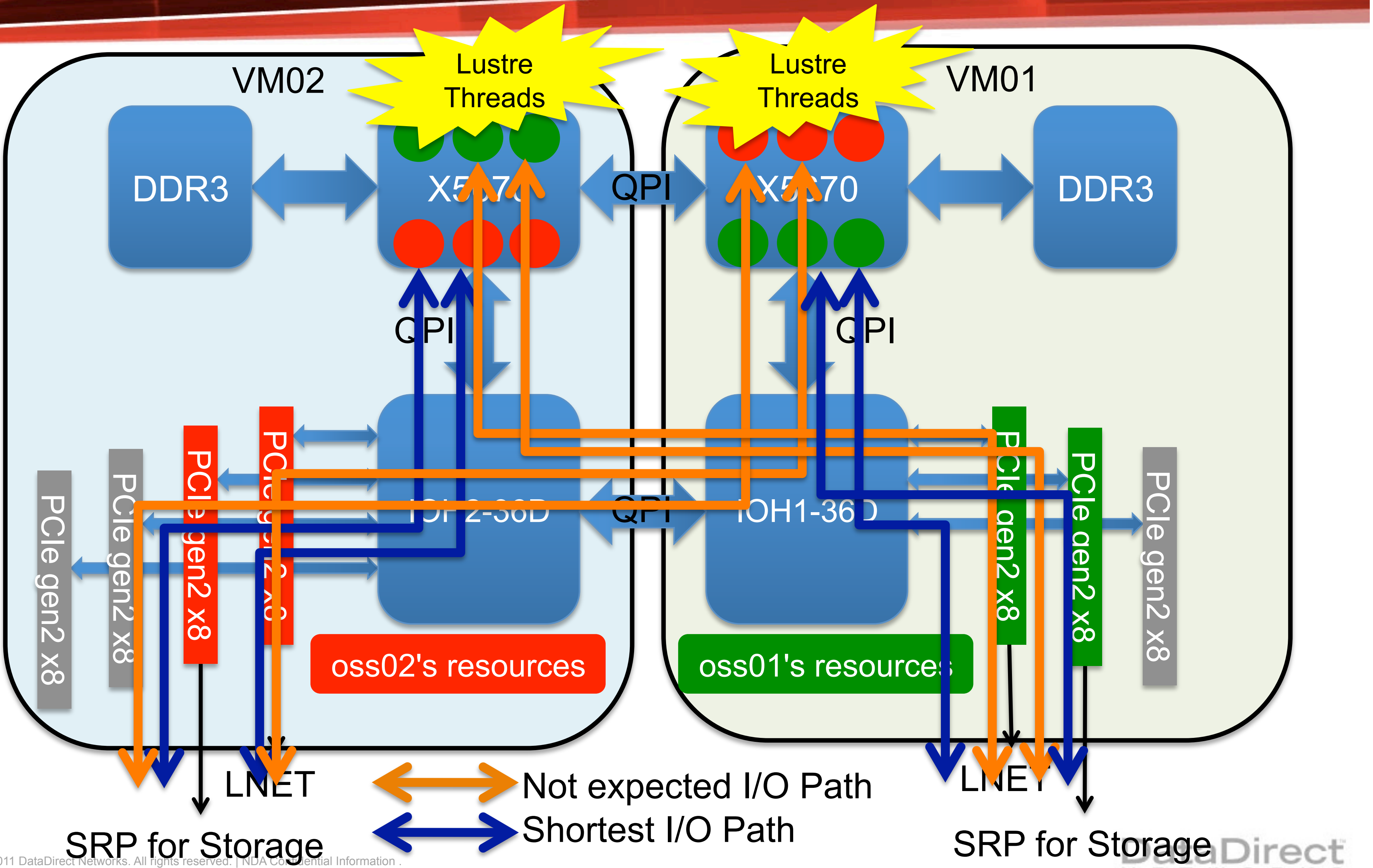


Lustre backend performance

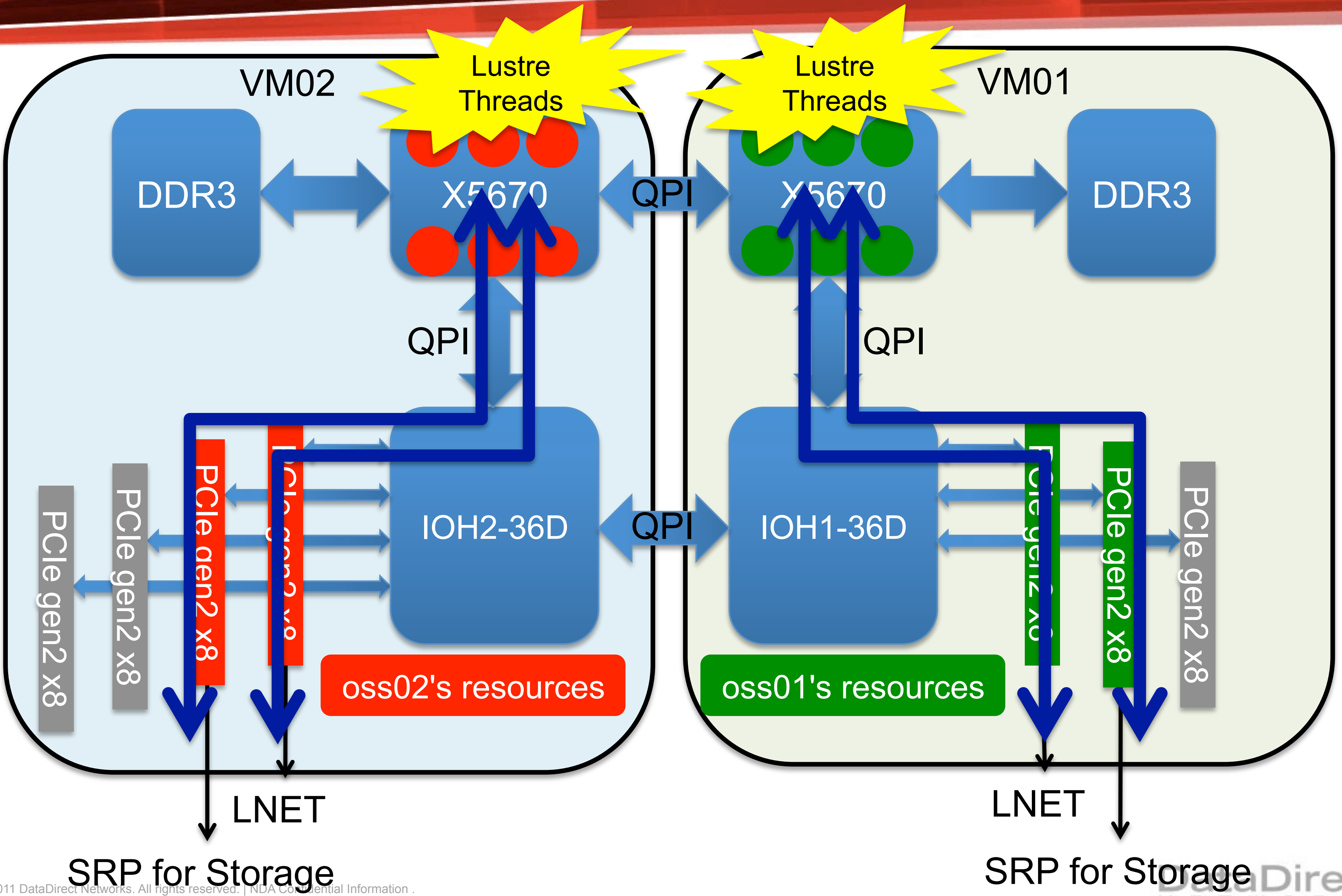


- **OSS (two VMs)**
  - 6 cores per VM
  - 12GB memory per VM
  - NUMA & NUMIOA aware
  - Two HCA (for SRP and LNET) are assigned with PCI pass-through
- **Storage**
  - SFA10000 (Single Controller)
  - 140 x SATA disk
  - 2 x QDR connections

Lustre backend performance
- I/O path without CPU affinity -

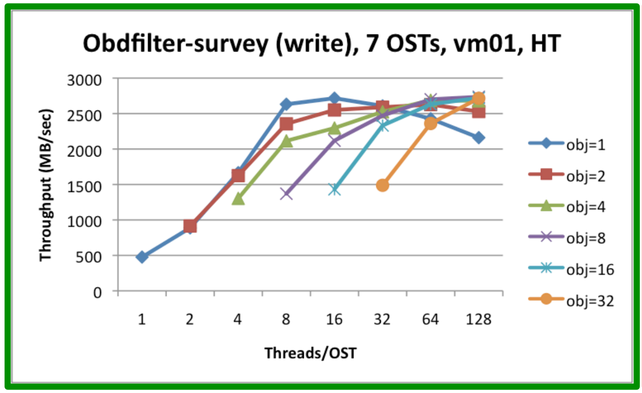


Lustre backend performance
- NUMA & NUMIOA aware with VMs -

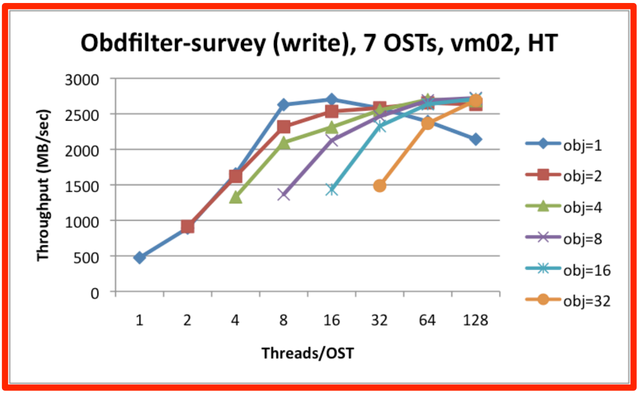


Lustre backend performance - Obdfilter-survey results -

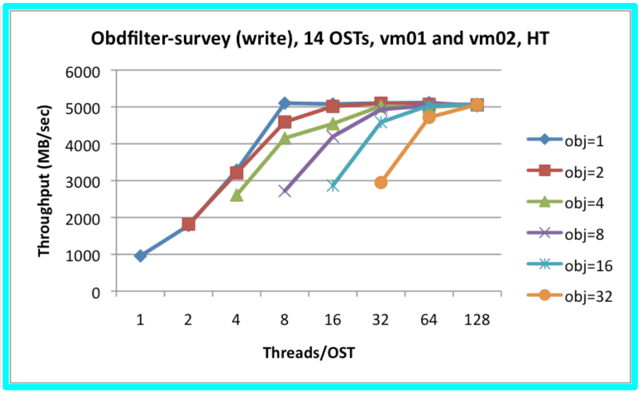
Write


Obdfilter-survey (write), 7 OSTs, vm01, HT

vm01: 2.6GB/sec

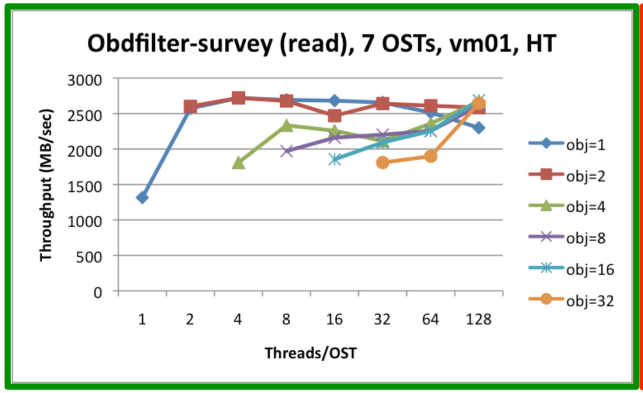

Obdfilter-survey (write), 7 OSTs, vm02, HT

vm02: 2.6GB/sec

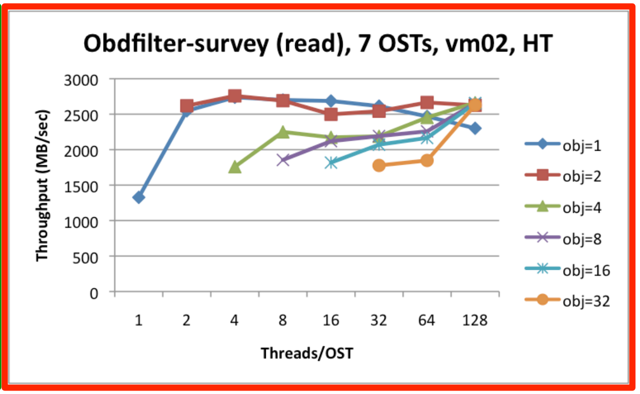

Obdfilter-survey (write), 14 OSTs, vm01 and vm02, HT

vm01+vm02: 5GB/sec

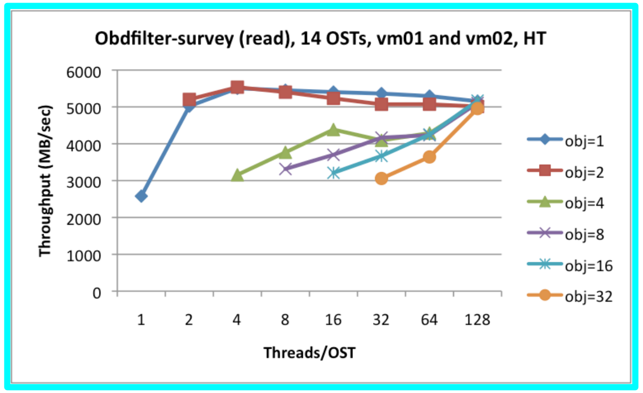
Read


Obdfilter-survey (read), 7 OSTs, vm01, HT

vm01: 2.7GB/sec


Obdfilter-survey (read), 7 OSTs, vm02, HT

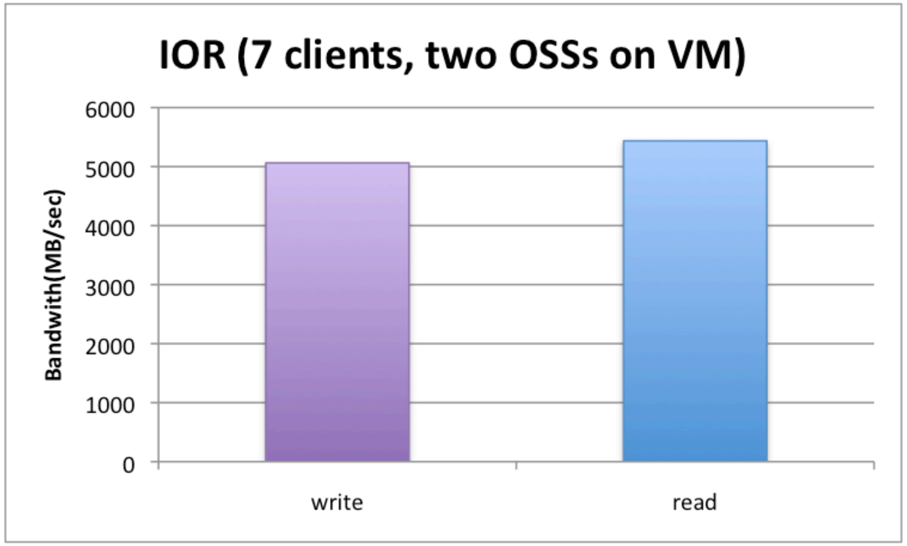
vm02: 2.7GB/sec


Obdfilter-survey (read), 14 OSTs, vm01 and vm02, HT

vm01+vm02: 5.4GB/sec

Lustre performance from the clients
- IOR -

- **Run IOR from 7 x Lustre clients**



Write: 5GB/sec Read: 5.4GB/sec

- This is almost the same results which we are seeing on the physical Lustre servers with SFA10000 (single controller).
- Could see double performance by using 4 VMs and dual SFA10000 controllers.

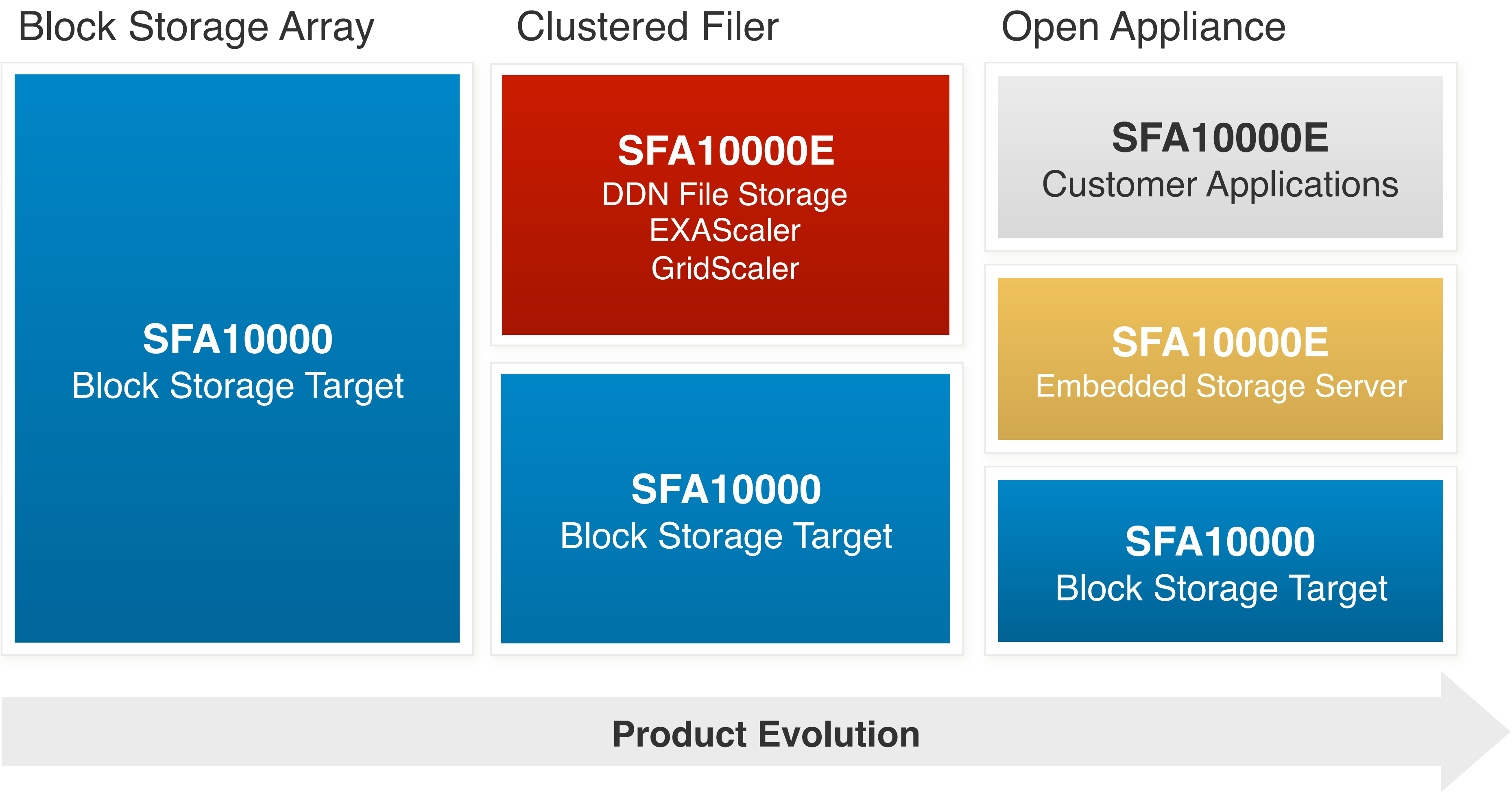
Summary

- **A Virtualized Infrastructure based on KVM works well**
  - Only a couple of minutes needed for all server setup!
  - Various types of Lustre testing are possible.
  - Achieves almost equal performance numbers when compared to physical servers without VMs.
  - SR-IOV will provide a basis for much more flexible configurations!
  - Will investigate SR-IOV, FC and testing on more servers in future!
  - Will continue to invest Lustre on KVM

Introducing the DDN SFA10000E

DataDirect™
NETWORKS

Multi-Platform Architecture



Block Storage Array

SFA10000
Block Storage Target

Clustered Filer

SFA10000E
DDN File Storage
EXAScaler
GridScaler

SFA10000
Block Storage Target

Open Appliance

SFA10000E
Customer Applications

SFA10000E
Embedded Storage Server

SFA10000
Block Storage Target

Product Evolution

Flexible Deployment Options: 3 System Modalities

SFA10000E Appliance

- **SFA10000E initially available with DataDirect Networks' s parallel clustered file system solutions**
- **Integrate multiple appliances to scale to over 200GB/s and 10' s of Petabytes**

ExaScaler
SFA10000E

Up to 6GB/s
Up To 900TB

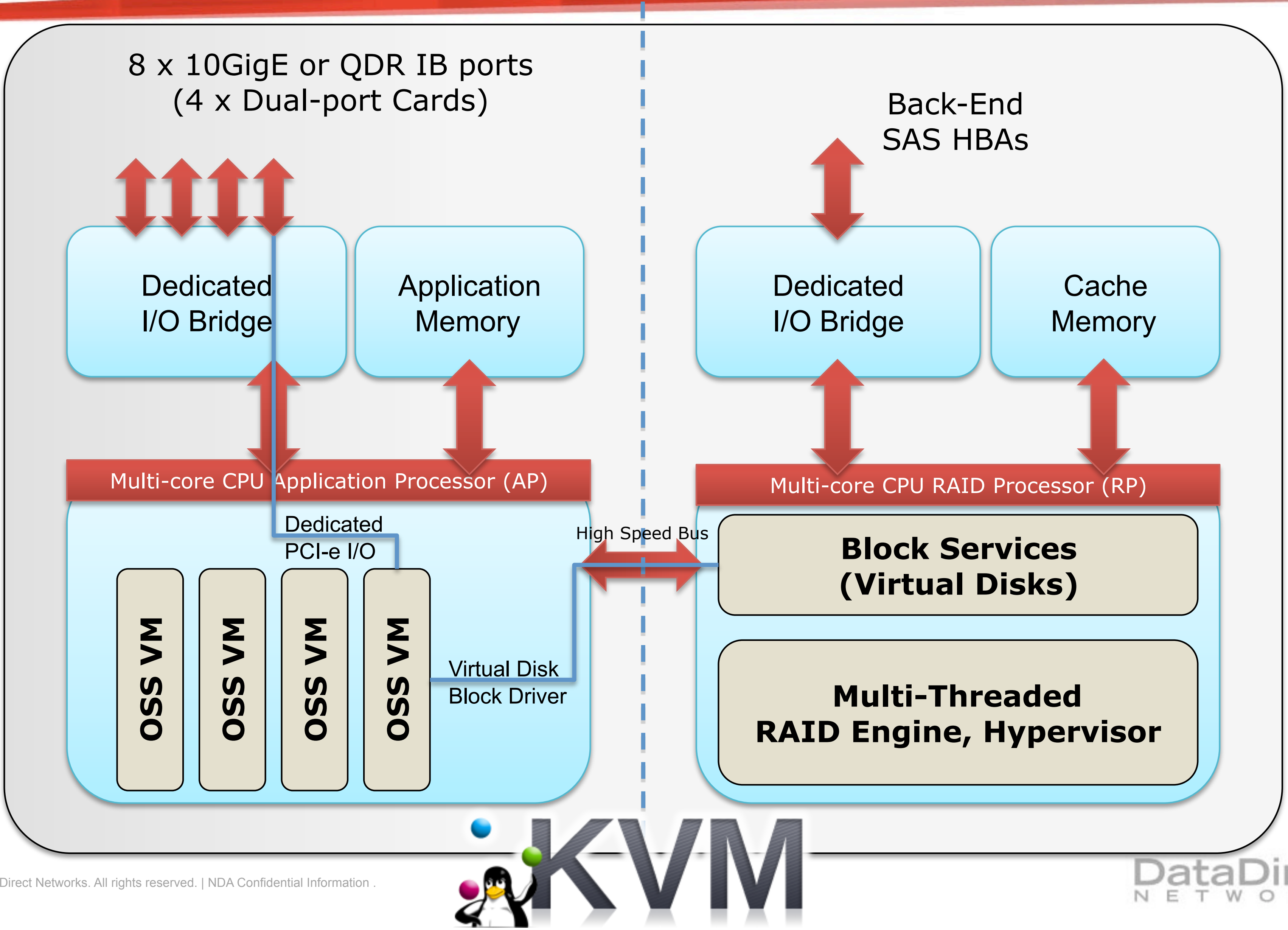
✓ Reduce complexity, infrastructure and administration

✓ Reduce cost as well as lower operational cost

✓ Increase performance for latency sensitive applications

  - ✓Shared Memory
  - ✓Eliminate SCSI Overhead

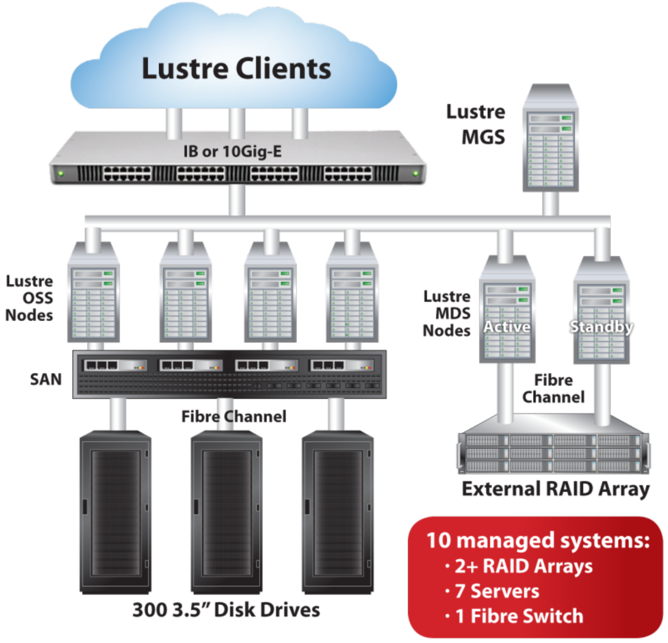
SFA10000 Embedded ExaScaler



HPC Storage on the SFA10000E Appliance

Storage Fusion Architecture

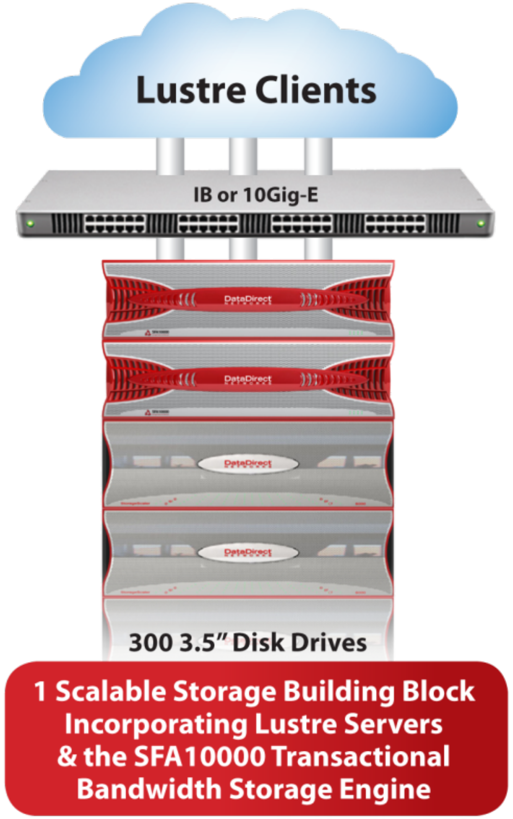
Traditional Lustre Deployment to Achieve 5GB/s



10 managed systems:

- 2+ RAID Arrays
- 7 Servers
- 1 Fibre Switch

SFA10000E, Embedded EXAScaler, 5GB/s



1 Scalable Storage Building Block Incorporating Lustre Servers & the SFA10000 Transactional Bandwidth Storage Engine

Storage Fusion Architecture
not only reduces complexity, it streamlines IO
by reducing latency and protocol conversions

Sample Customers

lustre user group !

Efficient Storage Lustre Users Worldwide – Nearing our 10-Yr Lustre Deployment Milestone

TB/s of Lustre Performance Powering HPC Worldwide

Thanks To DDN Customers For Your Partnership in HPC!

Thank You

DataDirect™
NETWORKS

Backup Slides

DataDirect™
NETWORKS

How to setup Lustre on KVM with IB
- Quick example -

It's basic KVM setup, no any special operations!

1. Create first VM and Install Operating system (and Lustre)

```
# virt-install –name=oss01 –vcpus=6 --ram=8192 --os-type=linux –hvm --connect=qemu:///system --network
bridge:br0 --location /var/www/html/os_images/centos5.5 --file /vmimage/oss01.img -s 10 –accelerate –nographics --
mac=52:54:00:aa:aa:00 --extra-args='console=tty0 console=ttyS0,115200n8 ks=http://192.168.122.1/centos5.ks'
```

2. Create clone VM image for Second VM

```
# virt-clone –original oss01 –name oss02 --mac=52:54:00:aa:aa:02 --file /vmimage/oss02.img
.....
```

3. Find PCI device ID of HCA

```
# lspci | grep InfiniBand
02:00.0 InfiniBand: Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR / 10GigE](rev b0)
...
# virsh nodedev-list | grep pci_0000_02
pci_0000_02_00_0
....
```

4. Detach HCAs from Host

```
# virsh nodedev-dettach pci_0000_02_00_0
```

How to setup Lustre on KVM with IB - Quick example - (cont'd..)

5. Create an definition file of HCA

```
# cat mellanox_hca_bus02_00_0.xml
<hostdev mode='subsystem' type='pci'>
<source><address bus='0x02' slot='0x00' function='0x00'/></source>
</hostdev>
```

6. Attach HCA to VM

```
# virsh attach-device vm01 mellanox_hca_bus02.xml
```

7. CPU assailment and affinity setting

```
# virsh vcpupin vm01 0 0
# virsh vcpupin vm01 1 1
...
```

8. Now, VMs with Infiniband is ready. Move forward formatting the Lustre.

This all procedures are scriptable and don't many typing ☺