April 15th - 19th, 2013

DataDirect™ NETWORKS
INFORMATION IN MOTION™

LUG13

Active-Active LNET Bonding Using Multiple LNETs and Infiniband partitions

Shuichi Ihara

DataDirect Networks, Japan

Today's H/W Trends for Lustre

DataDirect NETWORKS
Information in Motion

- Powerful server platforms emerging
  - 16 CPU cores and growing, high memory bandwidth, PCI gen3, etc.
  - The number of OSS is important in order to obtain high throughput performance, but power and management cost are also critical.
- Fast disks and next generation devices are coming
  - New 2.5 inch form-factor with 12 Gbps mixed SAS/PCI connector.
  - Prices for high speed devices are decreasing rapidly.
  - Infiniband is providing high bandwidth to storage.
- High bandwidth network is available
  - Infiniband is now the most common interconnect Lustre Networking
  - Efficient bit encoding rates

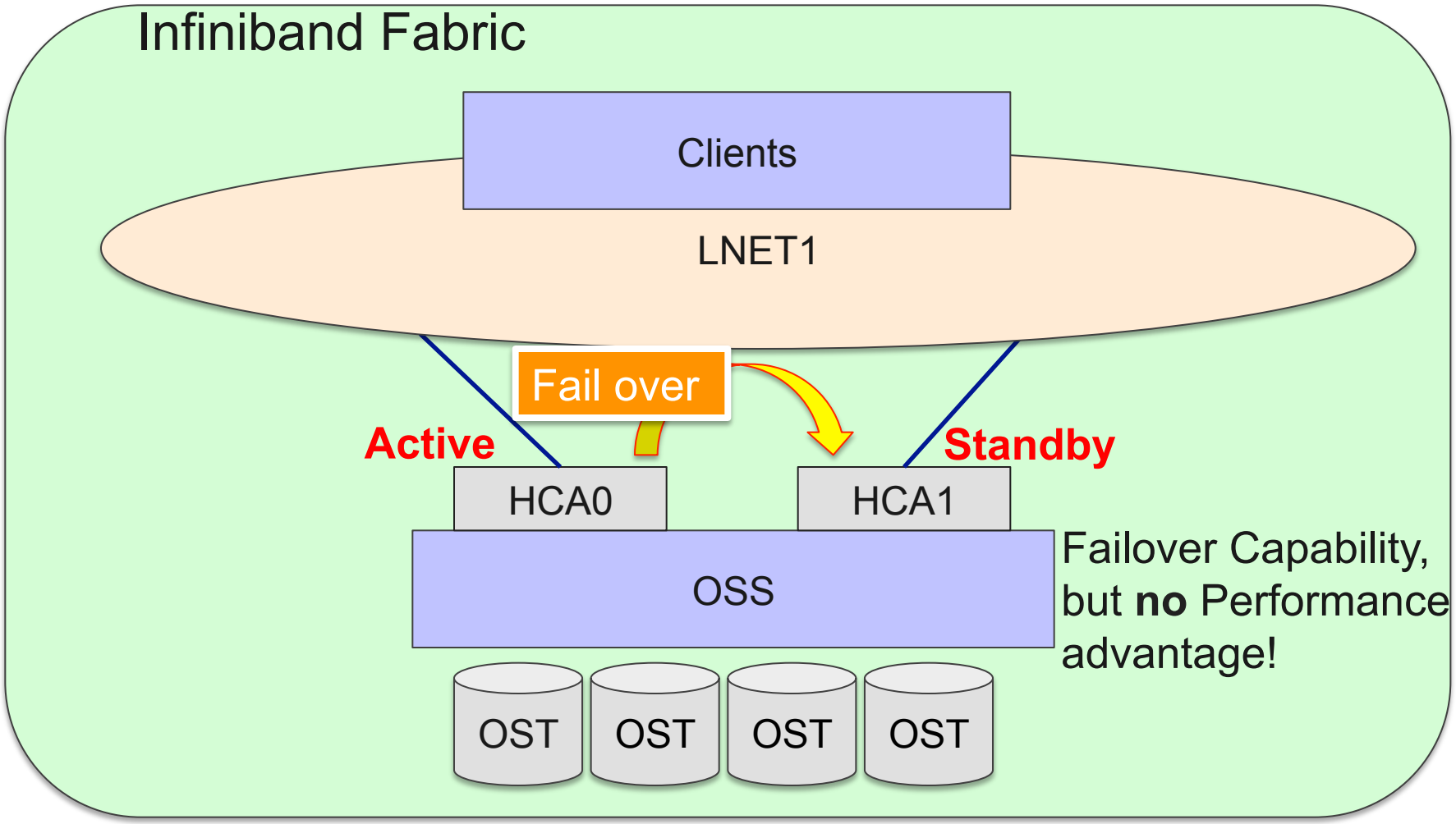
Conclusion: The performance of all components is increasing drastically. To achieve optimal H/W and S/W performance is challenging!

LND/LNET Bonding

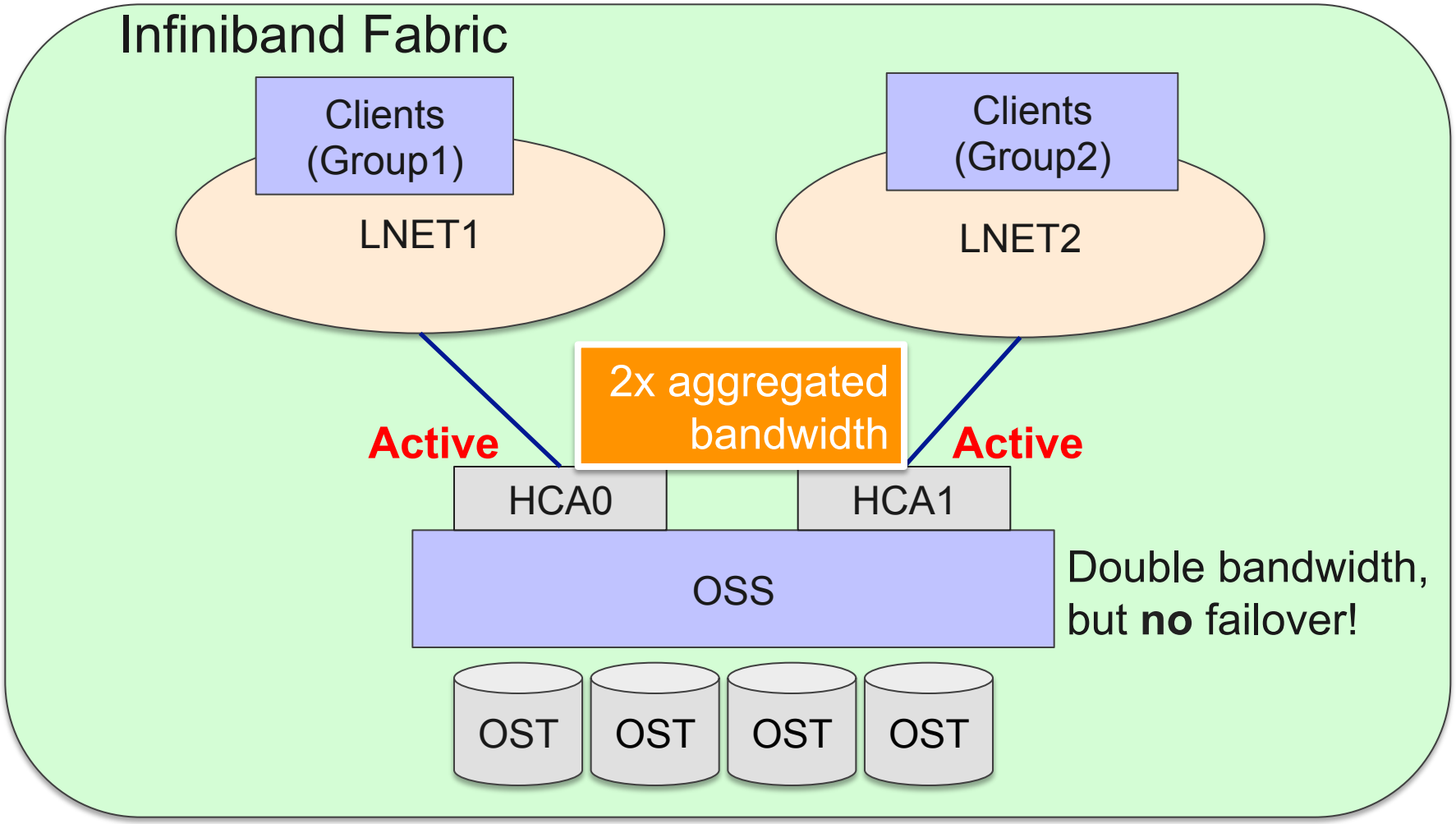
DataDirect NETWORKS
Information in Motion

- Lustre LNET performance
  - 6GB/sec on Single Infiniband FDR link
  - More bandwidth would help if the storage system is very powerful
  - Configurations with *less* Lustre servers become possible.
- Channel bonding
  - LND active/active channel bonding is not supported in mainstream Lustre today.
  - Infiniband multi-rail configuration is supported.
  - Lustre supports Active/Standby bonding with Infiniband.

Active/Standby IB LND configuration



Infiniband Multi-Rail Configuration



Novel Approach: Active/Active LND/LNET Configuration

DataDirect NETWORKS
Information in Motion

- 2 x Active/Standby = **Active/Active**
  - IB partitions create virtual (child) Interface on a HCA
  - Multiple LNETs with o2iblnd are created on an IB fabric
  - o2iblnd LND layer provides Lustre failover capability for Infiniband
- What's advantages?
  - No Lustre modification necessary—simply enabling IB partitions on SM (Subnet Manager), bonding, and LNET configuration.
  - Basically, no additional hardware on clients and server (More hardware increases performance!)
  - NUMA aware optimized OST access.
  - Auto failback, manual active network link control is possible.

Infiniband Partition

DataDirect NETWORKS
Information in Motion

- Partitioning enforces isolation among systems sharing an Infiniband fabric
  - The concept is similar to VLAN (802.1Q)
  - Enforced on Host and Switch
- Partitions are represented by P-key
  - Subnet Manager creates P-KEY tables for HCAs and switches in the network
  - Two membership configuration are available:
    - Full access
    - Limited access.
  - IPoIB uses P-keys for creating “child” interfaces associated with the P-key

```
/etc/opensm/partitions.conf
Default=0x7fff,ipoib      :ALL=full;
LNET0=0x8001,ipoib   :ALL=full;
LNET1=0x8002,ipoib   :ALL=full;
```

Active/Active LNET Configuration

DataDirect NETWORKS
Information in Motion

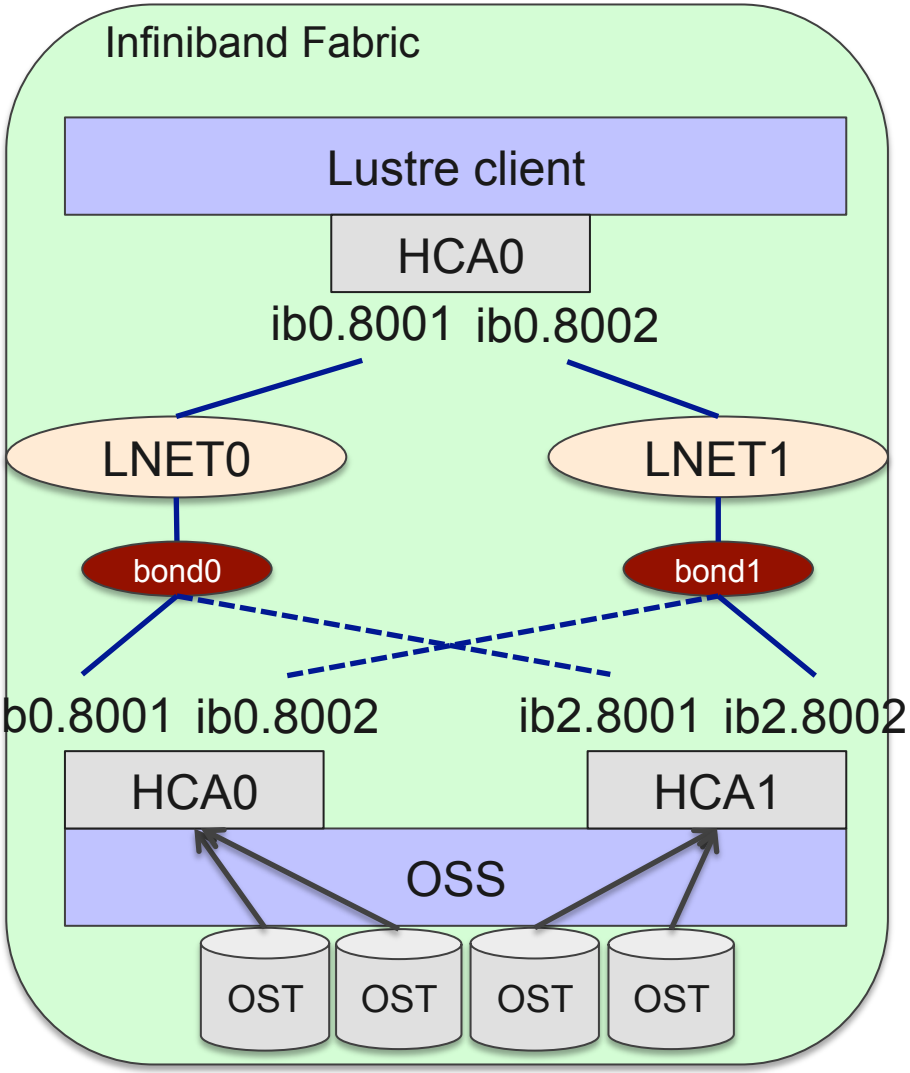
- Same Physical H/W Configuration
- Two P_KEY are created for IPoIB child interface on OSSs and clients.
- Two bond interfaces are enabled with IPoIB child interfaces.

  e.g) bond0 is active on HCA0
  bond1 is active on HCA1
- Two LNETs with o2iblnd are created using bond interfaces

  oss: options lnet networks=o2ib0(bond0), \
  o2ib1(bond1)

  client: options lnet networks=o2ib0(ib0.8001) \
  o2ib1(ib0.8002)
- Restricted OST access by LNET

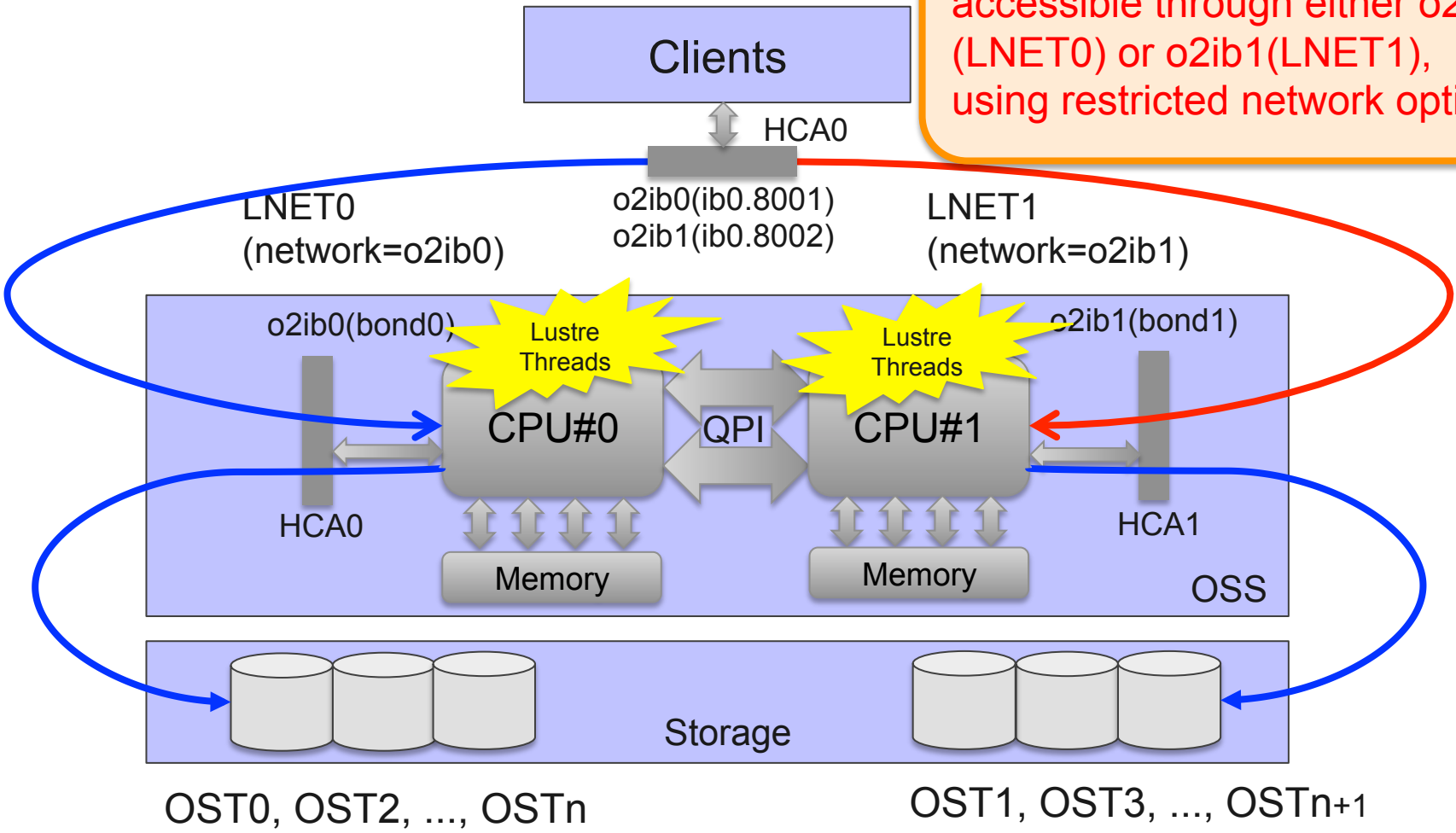
  mkfs.lustre –ost .. --network=o2ib



LNET Round-robin for OST access

DataDirect NETWORKS
Information in Motion

Even/odd OSTs are only accessible through either o2ib0 (LNET0) or o2ib1(LNET1), using restricted network option.



Benchmark and Failover Testing

DataDirect NETWORKS
INFORMATION IN MOTION

Test Configuration

Storage

- 1 x SFA12K-40
- 160 x 15Krpms SAS disk

Server

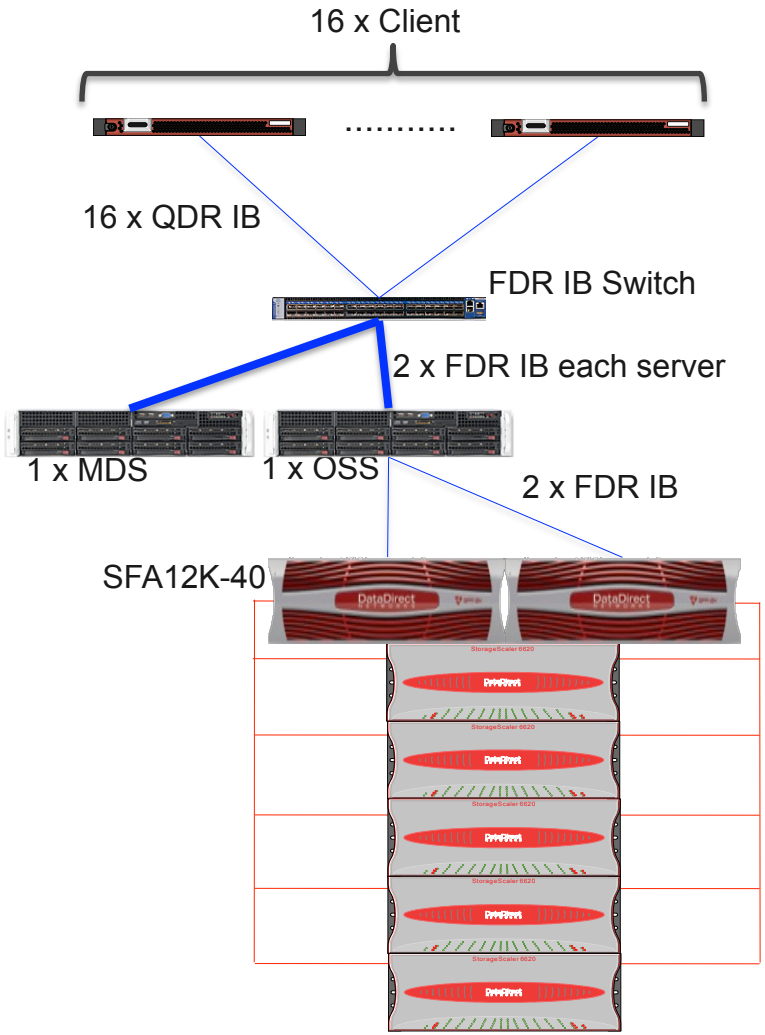
- 1 x MDS, 1 x OSS
- 2 x 2.6GHz E5-2670, 64GB Memory
- 2 x FDR IB Dual port HCA
- (2 ports for LNET and 2 ports for Storage)

Client

- 16 x Client
- 1 x 2.0GHz, E5-2650, 16GB Memory
- 1 x QDR IB Single port HCA

Software

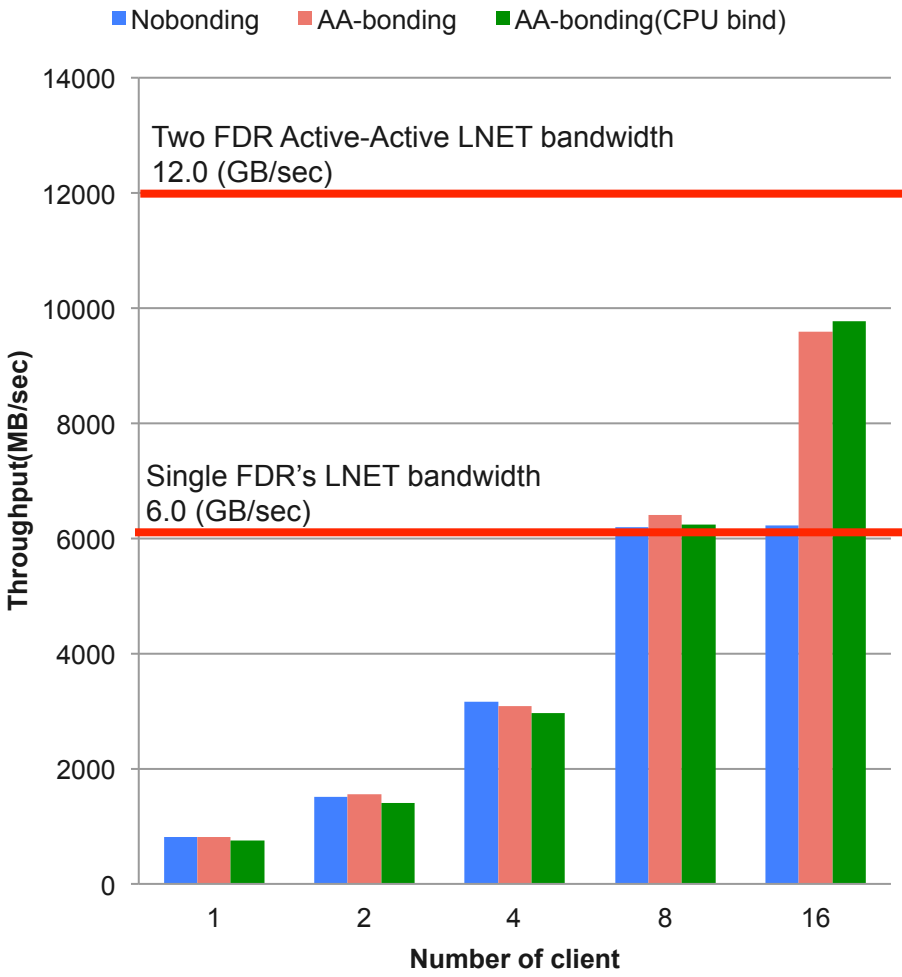
- CentOS6.3
- Lustre-2.3.63
- Mellanox OFED-1.5.3



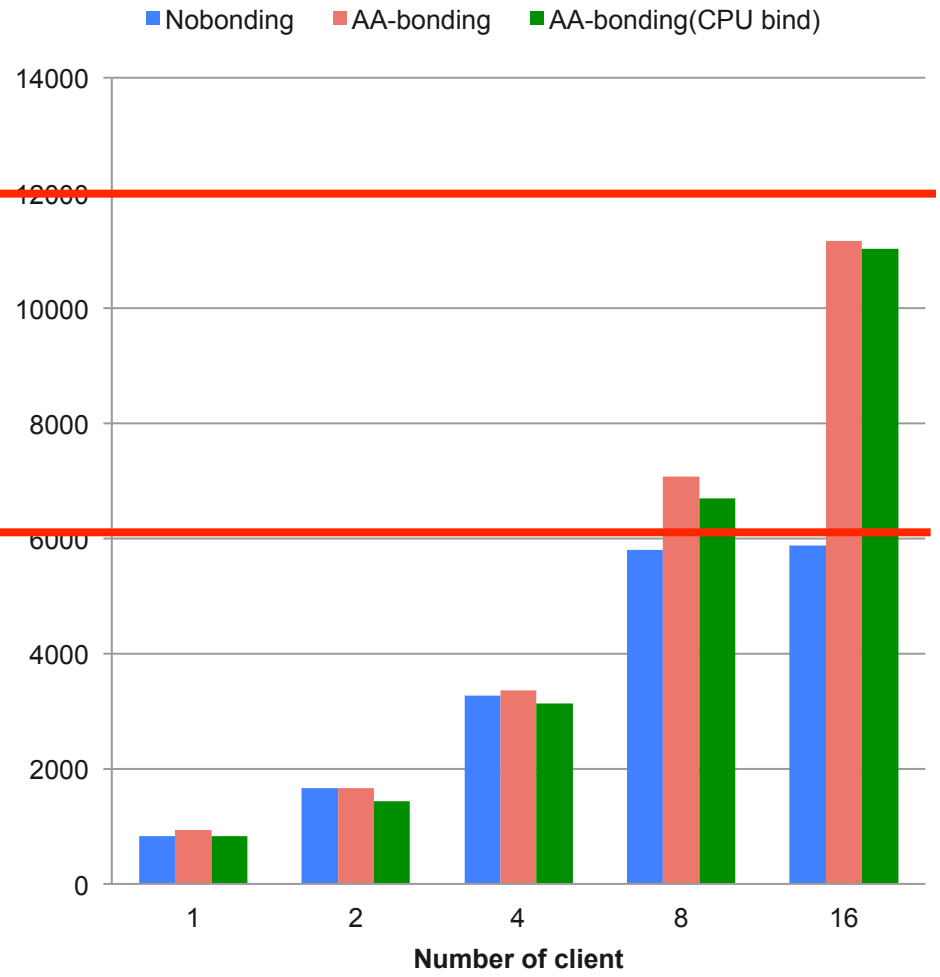
Active/Active LNET Bonding Performance

DataDirect NETWORKS
INFORMATION IN MOTION

Single OSS's Throughput (Write)

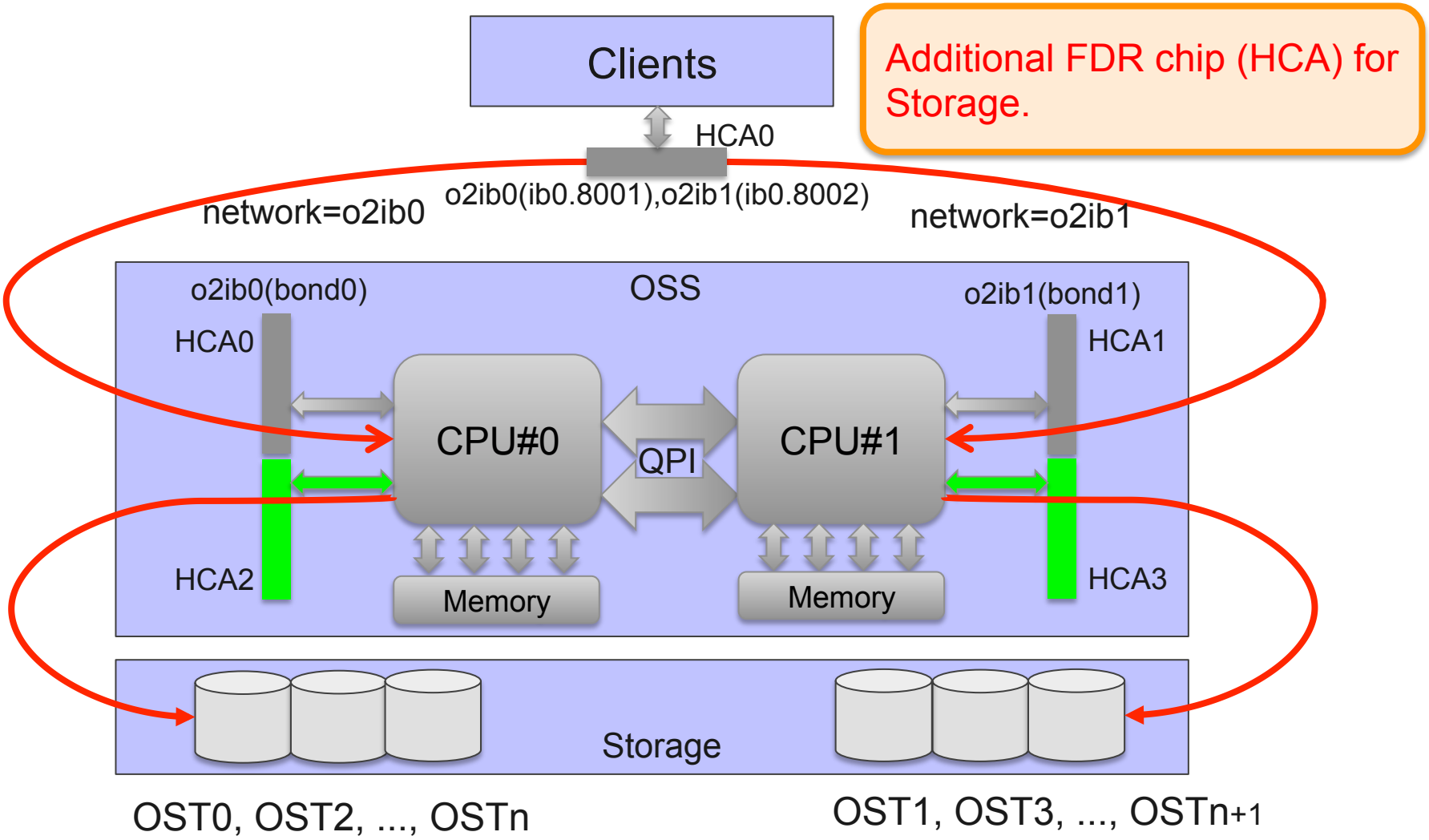


Single OSS's Throughput (Read)



Performance Evaluation on Enhanced Hardware Configuration

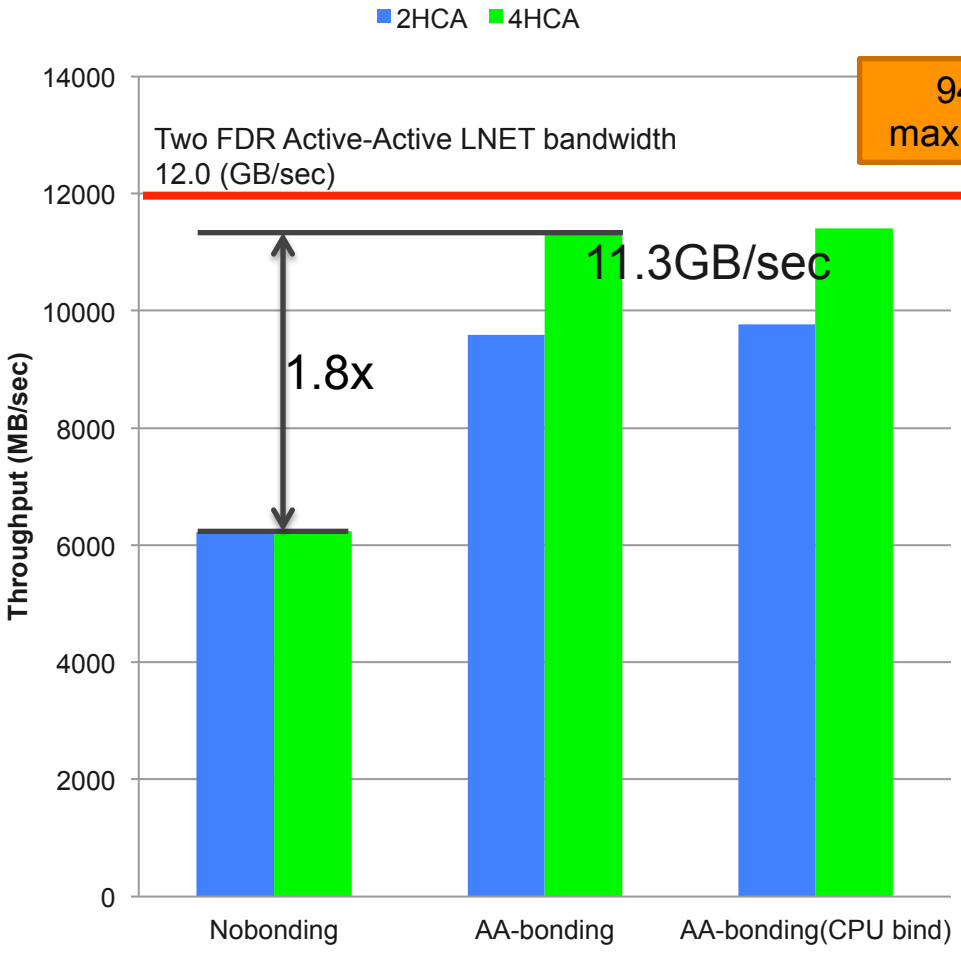
DataDirect NETWORKS
Information in Motion



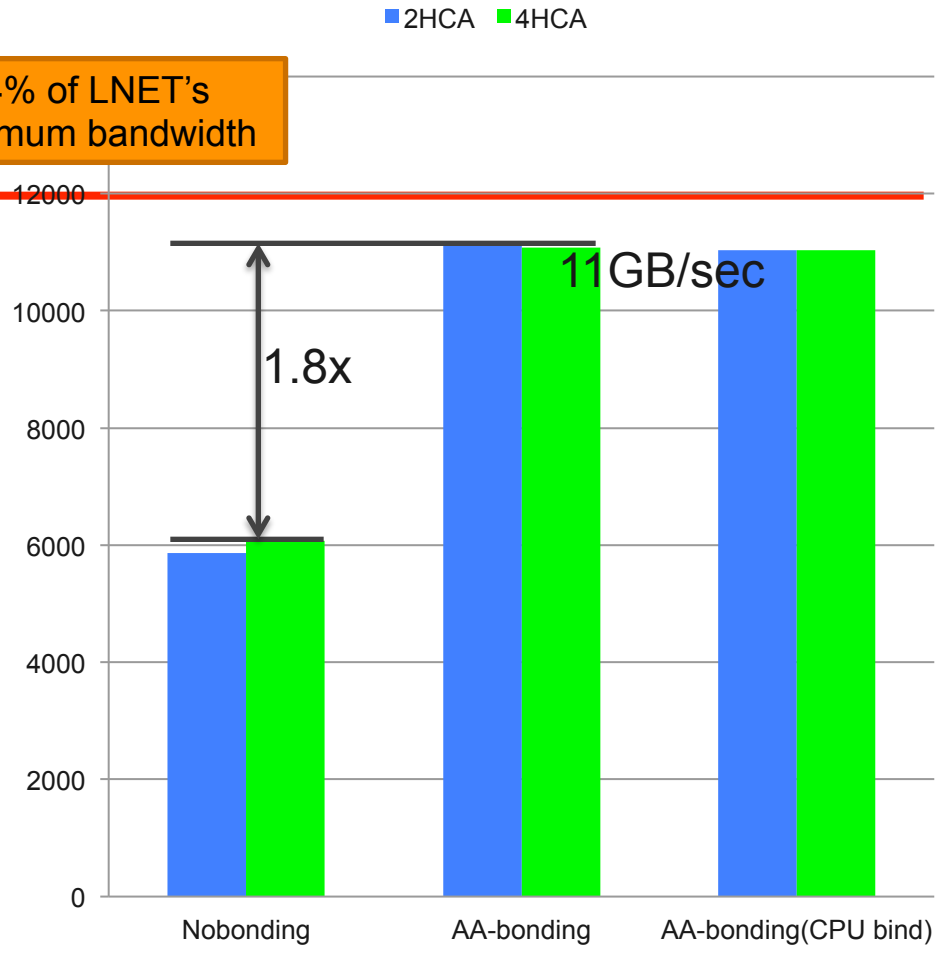
Active/Active LNET Bonding Performance (4 x HCA Configuration)

DataDirect NETWORKS
Information in Motion

Single OSS's Throughput (Write)

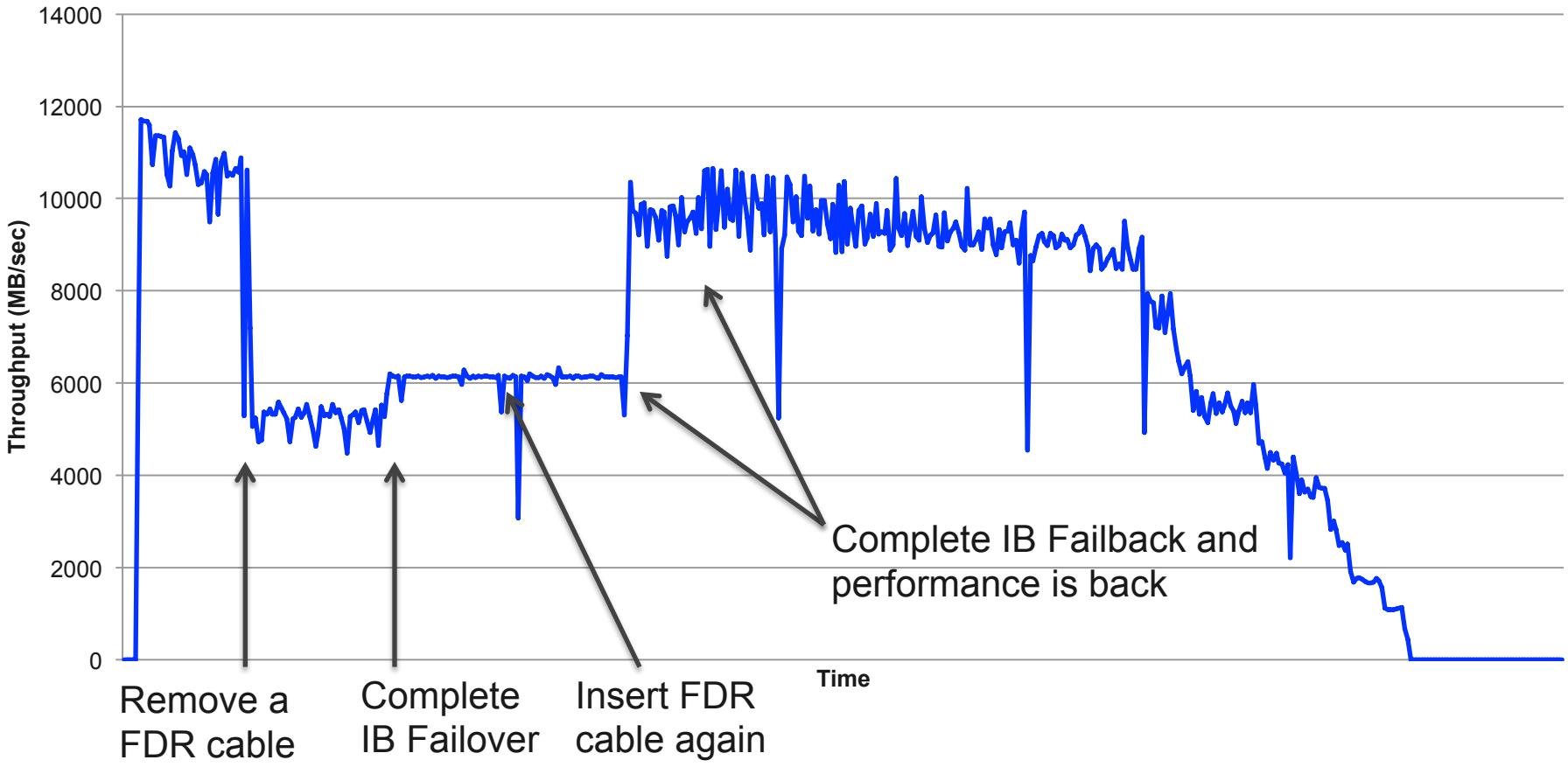


Single OSS's Throughput (Read)



Failover testing and behavior(1)

- Failover Test during IOR from 16 clients to OSS -



Failover Testing and Behavior(2)

DataDirect NETWORKS
INFORMATION IN MOTION

- Administrator controlled active IB links
  - AA LNET bonding can be configured on clients as well as server
  - "ifenslave" command helps to switch active Interface

e.g) # ifenslave bond0 -c ib2.8001

cat /proc/net/bonding/bond0

```
Ethernet Channel Bonding Driver: v3.6.0 (September 26, 2009)

Bonding Mode: fault-tolerance (active-backup) (fail_over_mac active)
Primary Slave: ib0.8001 (primary_reselect always)
Currently Active Slave: ib2.8001
MII Status: up
MII Polling Interval (ms): 50
Up Delay (ms): 5000
Down Delay (ms): 0
.....
```

Lustre OSTs try to reconnect Once active slave Interface changed.

Conclusion

DataDirect NETWORKS
Information in Motion

- Demonstrated LNET active/active configuration using IB partitions and it performed well.
  - 2 x Active/Standby bonding configuration works for LNET.
  - Achieved more than 94% of 2 x FDR LNET bandwidth from a single OSS.
  - Max Performance: **11.3GB/sec (WRITE), 11.0GB/sec (READ)**
  - Failover and failback works well after IB and bonding driver detect link failure/up status.
  - User controlled failover and client side Active/Active configurations are possible. Application job “aware” network control might be possible using this approach.

DataDirect™
NETWORKS
INFORMATION IN MOTION™

LUG13

