Active-Active LNET Bonding Using Multiple LNETs and Infiniband partitions

Shuichi Ihara
DataDirect Networks, Japan
Today’s H/W Trends for Lustre

► 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

▶ 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 Fabric

Clients

LNET1

Fail over

Active

HCA0

Failover Capability, but **no** Performance advantage!

Standby

HCA1

OSS

OST

OST

OST

OST
Infiniband Multi-Rail Configuration

Infiniband Fabric

Clients (Group 1)
LNET1

Clients (Group 2)
LNET2

2x aggregated bandwidth

Active

HCA0  HCA1

OST  OST  OST  OST

Double bandwidth, but **no** failover!
Novel Approach: Active/Active LND/LNET Configuration

▶ 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

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

- 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

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

OST0, OST2, ..., OSTn

OST1, OST3, ..., OSTn+1
Benchmark and Failover Testing

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

**Single OSS's Throughput (Write)**

- **Nobonding**
- **AA-bonding**
- **AA-bonding(CPU bind)**

**Single OSS's Throughput (Read)**

- **Nobonding**
- **AA-bonding**
- **AA-bonding(CPU bind)**

- **Two FDR Active-Active LNET bandwidth**
  - 12.0 (GB/sec)

- **Single FDR's LNET bandwidth**
  - 6.0 (GB/sec)
Performance Evaluation on Enhanced Hardware Configuration

Clients

Additional FDR chip (HCA) for Storage.

CPU#0

HCA0

network=o2ib0

o2ib0(ib0.8001),o2ib1(ib0.8002)

HCA1

network=o2ib1

OSS

CPU#1

QPI

HCA2

network=o2ib0

OSS

HCA3

CPU#0

CPU#1

Memory

OST0, OST2, ..., OSTn

Memory

OST1, OST3, ..., OSTn+1

Storage

OST0, OST2, ..., OSTn

OST1, OST3, ..., OSTn+1

Clients

Additional FDR chip (HCA) for Storage.
Active/Active LNET Bonding Performance (4 x HCA Configuration)

Single OSS's Throughput (Write)

- 2HCA
- 4HCA

94% of LNET's maximum bandwidth

Two FDR Active-Active LNET bandwidth
12.0 (GB/sec)

Throughput (MB/sec)

- Nobonding
- AA-bonding
- AA-bonding (CPU bind)

1.8x

11.3 GB/sec

Single OSS's Throughput (Read)

- 2HCA
- 4HCA

Throughput (MB/sec)

- Nobonding
- AA-bonding
- AA-bonding (CPU bind)

1.8x

11 GB/sec
Failover testing and behavior(1)

- Failover Test during IOR from 16 clients to OSS -

Throughput (MB/sec) vs. Time

- Remove a FDR cable
- Complete IB Failover
- Insert FDR cable again

Complete IB Failback and performance is back
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

<table>
<thead>
<tr>
<th>Ethernet Channel Bonding Driver: v3.6.0 (September 26, 2009)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonding Mode: fault-tolerance (active-backup) (fail_over_mac active)</td>
</tr>
<tr>
<td>Primary Slave: ib0.8001 (primary_reselect always)</td>
</tr>
<tr>
<td>Currently Active Slave: ib2.8001</td>
</tr>
<tr>
<td>MII Status: up</td>
</tr>
<tr>
<td>MII Polling Interval (ms): 50</td>
</tr>
<tr>
<td>Up Delay (ms): 5000</td>
</tr>
<tr>
<td>Down Delay (ms): 0</td>
</tr>
<tr>
<td>......</td>
</tr>
</tbody>
</table>

Lustre OSTs try to reconnect Once active slave Interface changed.
Conclusion

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.