Lustre*/Lnet Usage Pattern

Doug Oucharek, Intel Corp.

#OFAUserGroup

* Some names and brands may be claimed as the property of others.
Overview

• Look at how Lustre uses the Fabric and OFED
• LNet Routing
• LNet Configuration Verification
• Channel Bonding
• Monitoring Connections
Lustre Overview

Compute Nodes (Clients)

10k - 100k nodes

Backend Storage

Lustre Servers

Fabric

Up to 500 OSSs

Up to 4000 OSTs
## By the Numbers

<table>
<thead>
<tr>
<th></th>
<th>2012</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>10-100K</td>
<td>100K-1M</td>
</tr>
<tr>
<td>Threads/node</td>
<td>~10</td>
<td>~1000</td>
</tr>
<tr>
<td>Total concurrency</td>
<td>100K-1M</td>
<td>100M-1B</td>
</tr>
<tr>
<td><strong>Object create</strong></td>
<td><strong>100K/s</strong></td>
<td><strong>100M/s</strong></td>
</tr>
<tr>
<td>Memory</td>
<td>1-4PB</td>
<td>30-60PB</td>
</tr>
<tr>
<td>FS Size</td>
<td>10-100PB</td>
<td>600-3000PB</td>
</tr>
<tr>
<td>MTTI</td>
<td>1-5 Days</td>
<td>6 Hours</td>
</tr>
<tr>
<td>Memory Dump</td>
<td>&lt; 2000s</td>
<td>&lt; 300s</td>
</tr>
<tr>
<td><strong>Peak I/O BW</strong></td>
<td><strong>1-2TB/s</strong></td>
<td><strong>100-200TB/s</strong></td>
</tr>
<tr>
<td>Sustained I/O BW</td>
<td>10-200GB/s</td>
<td>20TB/s</td>
</tr>
</tbody>
</table>
Markets

• HPC
  – Well established

• Enterprise
  – Large Scale Data Analytics (i.e. Oil & Gas)
  – Big Data -> Hadoop Adapter

• Cloud
  – As a series of VMs
More Details

Management Servers (MGSs)

Metadata Servers (MDSs)

Object Storage Servers (OSSs)

Object Storage Targets (OSTs)

Lustre Clients

MGS 1 (active)

MGS 2 (standby)

MDS 1 (active)

MDS 2 (standby)

Object Storage Arrays and SAN Fabric

Commodity Storage

Lustre Routers

OSS 1

OSS 2

OSS 3

OSS 4

OSS 5

OSS 6

OSS 7

InfiniBand network

Ethernet network

= failover capability
Usage Patterns

• **Striping**
  – RAID0 Behaviour

• **Peer Connections**
  – Created on demand: RC QPs
  – Multiple Server Messages in Parallel
  – Monitor Connections via LNet Ping

• **DNE : Multiple Meta-Data Servers**

• **Self Imposed Flow Control**
  – Credits (for network) and Peer Credits (for each connection)
  – Latency is OK, message loss is not
LNet Routing

Compute Node (Client)  LNet Router  Lustre Servers
Why Route?

• From ORNL paper: Network Contention and Congestion Control: Lustre Fine-Grained Routing – Matt Ezell
  – Control bandwidth by varying the number of routers
  – Control I/O paths by selecting which routers to use
  – Control route computation by partitioning fabrics
  – If you are using multiple network types, you have to route traffic
Configuration

• Server Discovery
  – Use IPv4 (do not use IPoIB)
  – Can not support IPv6 any time soon (see 2012 LUG presentation: LNET Support for IPv6 is Long Overdue – Isaac Huang)

• Currently done via module parameters

• However…
Dynamic LNet Config

- Adding/Deleting networks
- Adding/Deleting routes
- Configuring router buffer pools
- Enabling/Disabling routing.
- Showing routing information
- Importing/exporting configuration in YAML format
- See Lustre 2.7 Manual for details
Testing Configuration

• Start with IB ping
• Then use LNet Ping ("Ist ping <nid>")
• Finally, use LNet Selftest
• See last year’s presentation for example
Monitoring Connections

• Two Types of Pings
  – OST Pings
  – Router Pings

• OST Pings not good for scalable fault detection
  – See 2012 LUG paper: Lustre Ping Evictor Scaling in LNET Fine Grained Routing Configurations - Cray
  – Found a 4%-11% reduction in throughput
  – Example: 25,000 clients, 360 OSSs, 4 OSTs per OSS:
  – 36M pings every 75s
  – Suppress via ptrlrpc module parameter: suppress_pings
Channel Bonding

• Coming later this year
• Provides both:
  – Bandwidth
  – Redundancy
• Can Mix Fabrics supported by OFED
Network Stack

- ptlrpc
- LNet
- gnilnd
- Gemini
- sockInd
- Ethernet
- o2ibInd
- OFED
- Provider 1
- Provider 2
OFED Usage

• Only PUTs Supported in o2iblnd
  – Therefore: only RDMA writes
  – Done to avoid limitations of RDMA reads

• Use Reliable Connections

• Choices:
  – <4k = Use Immediate
  – >4k, <1M = Use RDMA write
Example: Sending a File

LNet  o2ibInd  OFED  OFED  o2ibInd  LNet

LNet PUT

PUT Hdr  PUT Hdr  PUT Hdr

PUT ACK  PUT ACK  PUT ACK

Register Buffer  Register Buffer

Mem ID  Mem ID

RDMA Write  RDMA

MD

Register Buffer

Mem ID

RDMA Write
Thank You

OpenFabrics Software
User Group Workshop

#OFSUserGroup