AGENDA

- Overview of Multi-Rail Features
- Base Multi-Rail
- Multi-Rail Health and Resiliency
- Multi-Rail Routing
- Multi-Rail Network Selection Policies
- Summary
MULTI-RAIL FEATURE SET OVERVIEW
● Lustre has its own networking abstraction layer, LNet
● Different types of LNet networks are configured to encapsulate traffic
● Each network type has its own driver, LND
  – IB/RoCE/OPA (verbs) Traffic, o2ibInd - o2ibX
  – Ethernet traffic, sockInd - tcpX
Traditionally, LNet allowed only one network interface per LNet network.

If a node had multiple interfaces, multiple LNet Networks need to be configured.
WHAT IS MULTI-RAIL

- **LNet Level Multi-Rail Solution**

- **Multi-Rail allows nodes to communicate across multiple interfaces:**
  - Using Multiple interfaces connected to one network
  - Using multiple interfaces connected to different networks
  - These interfaces are used simultaneously
MULT-RAIL GOALS

- Goals
  - Allow multiple interfaces to be configured in the same LNet Network
  - Allow automatic interface discovery of peers
  - Allow all interfaces in the same network to be used in Active/Active mode
  - Allow heterogeneous networks (ex: tcp3 and o2ib1) to be used simultaneously
  - Monitor Interface/network health and use the healthiest interface
  - Apply selection criteria on which interface to use
  - Apply user specified network selection policies
MULTI-RAIL LNET
LNET LEVEL MULTI-RAIL

- LNet Level Multi-Rail Solution

- Advantages
  - Simpler configuration and automatic discovery
  - Support different HW (MLX, ETH, OPA, ...)
  - Aggregate throughput of interfaces
  - Resiliency against network failure
  - Control over interface selection based on internal consideration such as NUMA configuration, health, credits
  - Control over network pathway selection based on physical network characteristics
BASE MULTI-RAIL
Why LNet Multi-Rail not lower level interface bonding?
- Finer grained control over network interface selection
- Finer grained control over network path selection
- Finer grained control over network interface health monitoring
- LNet level message control
- Other Interface types can be configured and leverage the Multi-Rail capabilities

The Multi-Rail solution was implemented in multiple phases
Phase one and two was to get LNet using multiple interfaces and simplify configuration
What criteria should be used to select an interface?

Keep RDMA performance in mind

Criteria
- NUMA closeness
- Credits available per interface
- Round Robin

Flexibility:
- Algorithm needs to be flexible to allow other criteria.
  - Health
  - Buffer source restrictions beside NUMA
When configuring an LNet network specify the interfaces on this network

- options lnet network="o2ib(ib0,ib1),tcp(eth0,eth1)"
  - or
  - Inetctl net add --net o2ib --if ib0,ib1
  - Inetctl net add --net tcp --if eth0,eth1

First interface configured on the node becomes its Primary NID
- The Primary NID becomes the unique identifier of the node

Nodes can automatically discover the list of interfaces of other peers. No extra configuration required

Considerations:
- Group interfaces on the same subnet in the same LNet
- Group homogeneous interfaces in the same LNet.
- **Network interface performance is aggregated**
  - EX: 2x EDR IB interfaces with 12.5GB/s performance --> ~23 GB/s LNet level Performance (almost line rate)
  - 1MB block size RDMA write
LUSTRE PERFORMANCE

- Lustre File system doesn't approach line rate but performance is still improved
  - 32 socket of Xeon Processors
  - 16 TB of memory
  - 8 Omni-Path network interfaces
  - 8 C2112-GP2-EX Object Storage Systems (OSS)
  - 4 P3700 NVMe devices LDISKFS Object Storage Target (OST) per OSS

- Theoretical maximum performance of the system:
  - P3700 Sequential Write: 34560 MB/s
  - Sequential Read: 86400 MB/s

- Multi-Rail performance:
  - Sequential Write: 31990.18 MB/s
  - Sequential Read: 68593.35 M
MULTI-RAIL HEALTH & RESILIENCY
Need to monitor health in order to use the healthiest interface

Assign a maximum health value to each interface

Whenever failure occurs on the interface decrement the health value

When selecting an interface prefer the healthiest interface
  - Add this as a criteria to the interface selection algorithm

Handle protocol layer events, such as:
  - IB_EVENT_DEVICE_FATAL
  - IB_EVENT_PORT_ERR

The above two IB events lead to the interface going out of service until the corresponding up events are sent.
LNET LEVEL RETRIES

- Lustre Level RPCs are composed of one or more LNet messages.
- LNet message send failures can be handled at the LNet level before passing the failure up to Lustre for handling.

There are restrictions on failure handling:
- Local send failures are handled. IE: messages didn’t make it to the wire.
- Remote messages are not received. IE: remote didn’t process the message.
- Retry only if multiple interfaces are available.

- In this case an LNet message can be retried on a different interface.
- Maximum number of retries is configurable.
- Ensure retries do not over flow Lustre timeouts in order not to introduce further delays.
MULTI-RAIL ROUTING
What are LNet Routers?
- They route LNet messages to across different types of networks: tcp, o2ib

What are they used for?
- There are cases where two clusters separated by great geographical distance need to be connected
- Each cluster can use IB but messages traversing the clusters go over ethernet
- Routers are used to route IB LNet traffic over ethernet from one cluster to another

What is an MR Router?
- An MR node acting as a router with multiple interfaces
- Always referenced by its Primary NID
LNET ROUTERS

Clients Location 1 (tcp)

Gateway (tcp<>o2ib)

Clients Location 2 (o2ib)

Lustre servers Servers (o2ib)
A route is usually configured as follows:

- `lnetctl route add --net <remote net> --gateway <gateway NID> [--hop <number of hops --priority <prio>]`

- The remote net is a network we are not directly connected to which we want to reach.
- The gateway NID is the NID to send messages destined to the remote NID to.
- hop is the number of hops to the final destination.
- priority is the priority of that route.
- Multiple routes can be configured to the same remote network over different gateways.
- LNet will select the route with the highest priority or least number of hops.
- If all is the same, it’ll round robin.
Multi-Rail Routing Goals

- Deal with gateway as Multi-Rail nodes in order to leverage MR advantages, higher throughput, performance
- Can reduce the number of gateways if we just need to increase the throughput
- Use existing health mechanism to monitor the health of the gateway instead of having a separate mechanism
- Simplify routing configuration
  - No need to configure multiple routes which go to different interfaces of the same gateway
  - Use only the Primary NID of the gateway node
  - LNet will use all the gateway’s interfaces
LNET ROUTERS

Clients Location 1 (tcp)

Multiple Ethernet interfaces

Gateway (tcp<>o2ib)

Multiple IB interfaces

Clients Location 2 (o2ib)

Lustre servers (o2ib)
NETWORK SELECTION POLICIES
What are network selection policies

- Policies designed to allow the administrator to fine grained control traffic
- They govern the selection of:
  - Networks
  - Interfaces
  - Pairs of Networks or interfaces
  - Gateway interfaces

Why do we need it?

- There are some scenarios where the cluster administrators might want to configure two networks but keep one of them in standby
  - EX: o2ib network should be used for all traffic, unless it's not available then use tcp
- There could be physical network limitation which create a specific bottle neck which we try to avoid
This is a single fabric with a bottleneck.

Client1: 10.10.10.2@o2ib
Client2: 10.10.10.3@o2ib
MGS-1: 10.10.10.4@o2ib
MGS-2: 10.10.10.5@o2ib
MDS-1: 10.10.10.6@o2ib
MDS-2: 10.10.10.7@o2ib
OSS1-1: 10.10.10.8@o2ib
OSS1-2: 10.10.10.9@o2ib
OSS2-1: 10.10.10.10@o2ib
OSS2-2: 10.10.10.11@o2ib
This rule makes *Client1* avoid the red path:

- type: peer
- local: 10.10.10.2@o2ib
- remote: 10.10.10.[4-10/2]@o2ib
- priority: 0 # highest priority

*Client1* will only use the red path if there is no other option.
● Configuration is done from user space tool: lnctl
  - Add/Delete/Show policies

● Policies are created in user space, serialized and passed to LNet kernel module

● Policies are stored and applied on existing LNet constructs
  - This is done in order not traverse the policy tree on the fast path

● When new constructs are added, like Networks or Peers, the stored policies are automatically applied to them.
SUMMARY
Multi-Rail feature set was designed for the following main purposes
- Increase throughput
- Increase resiliency
- Simplify Configuration
- Fine control over traffic

Multi-Rail allows for intelligent selection of interfaces to maximize performance
- NUMA awareness is one example
- But if other RDMA sources introduce other criteria, they can be integrated into the selection algorithm

Multi-Rail was designed in LNet to allow for using heterogeneous networks
Other Network Interface types can be added later and benefit from the Multi-Rail feature without having to implement their own.
QUESTIONS
THANK YOU

John Smith, President and CEO

COMPANY XYZ