Whamcloud

LNet Update

Amir Shehata
Agenda

► Overview of Multi-Rail features
► Configuration Example
► Feature Updates
Multi-Rail Impact

- The Multi-Rail feature impacted the LNet code significantly
- Prior to MR each NID was considered a separate Peer
- MR adds the concept of Peers/Nets and Peer NIs/Local NIs
  - Peers and Nets become a collection of Remote and Local interfaces respectively
  - This gives the ability for LNet to utilize multiple interfaces for the same peer
Mult-Rail Overview
Impacts Enumerated

► Increased performance (Multi-Rail/Dynamic Discovery features)
  - LU-7734/LU-9480

► Improved resiliency (LNet Health)
  - LU-9120

► Traffic Control (User Defined/Network Selection Policies)
  - LU-9121

► Multi-Rail Routing (MR- Routing)
  - LU-11297

► Better Statistics (Sysfs)
  - LU-9667
Drawbacks

- Lustre still has a few areas where it deals with NIDs directly and makes certain assumptions, e.g:
  - Retrieving MDT nids from client log
- Goal is to eventually localize all peer lookup in LNet or under LNet APIs
Integrated Features

► The set of features listed are intended to be configured together to yield the best network performance/reliability.

► The next few slides will show an example configuration to illustrate how to configure these features.
Example Setup

- All nodes have 2x IB and 2x 100 GE interfaces
- Routers have 4x IB interfaces & 4x 100 GE interfaces
- Router Set A is more optimal for Client Set A and C
- Router Set B is more optimal for Client Set B
- Router Set B is more optimal to the MDSs
- Router Set A is more optimal to the OSS/OSTs
Configuration Requirements

- Maximize Performance
- Ensure reliability
- Ensure traffic goes over the optimal path
- Use IB as primary network, only use 100 GE if IB is not healthy
Client Configuration

modprobe lnet

lnetctl lnet configure

# configure networks
lnetctl net add --net o2ib --if ib0, ib1
lnetctl net add --net tcp --if eth0, eth1

# configure o2ib to be preferred
lnetctl policy add --src o2ib --priority 0

# configure router preference
lnetctl policy add \\n    --src 10.30.20.[150-154,160-164]@o2ib

    --rte 10.30.20.[2-8]@o2ib
Client Configuration

```
lnetctl policy add \n  --src 10.40.20.[150-154,155-159,160-164]@tcp
  --rte 10.40.20.[2-8]@tcp
lnetctl policy add \n  --src 10.30.20.[155-159]@o2ib
  --rte 10.30.20.[7-13]@o2ib
```
Client Configuration

# Route Configuration
lnetctl route add --net tcp1 --gateway 10.40.20.[2-8]@tcp
lnetctl route add --net o2ib1 --gateway 10.30.20.[2-8]@o2ib

# Health Configuration
lnetctl set retry_count 3
lnetctl set transaction_timeout 10
lnetctl set health_sensitivity 100
lnetctl set recovery_interval 1
Router Configuration

modprobe lnet
lnetctl lnet configure
# configure networks
lnetctl net add --net o2ib --if ib0, ib1
lnetctl net add --net o2ib1 --if ib2, ib3
lnetctl net add --net tcp --if eth0,eth1
lnetctl net add --net tcp1 --if eth2,eth3

# Health Configuration
lnetctl set retry_count 3
lnetctl set transaction_timeout 10
lnetctl set health_sensitivity 100
lnetctl set recovery_interval 1
Server Configuration

modprobe lnet
lnetctl lnet configure
# configure networks
lnetctl net add --net o2ib1 --if ib0, ib1
lnetctl net add --net tcp1 --if eth0, eth1
# configure o2ib to be preferred
lnetctl policy add --src o2ib1 --priority 0
# configure router preference
lnetctl policy add "
    --src 10.10.20.[100-105]@o2ib1
    --rte 10.10.20.[7-13]@o2ib1
Server Configuration

```bash
lnetctl policy add \
   --src 10.10.20.[106-110]@o2ib1
   --rte 10.10.20.[2-8]@o2ib1
lnetctl policy add \
   --src 10.20.20.[106-110]@tcp1
   --rte 10.20.20.[2-8]@tcp1
lnetctl policy add \
   --src 10.10.20.[100-105]@o2ib1
   --rte 10.10.20.[2-8]@o2ib1
lnetctl policy add \
   --src 10.20.20.[100-105]@tcp1
   --rte 10.20.20.[7-13]@tcp1
```
Server Configuration

# Route Configuration
lnetctl route add --net tcp --gateway 10.20.20.[7-13]@tcp1
lnetctl route add --net o2ib --gateway 10.10.20.[7-13]@o2ib1
lnetctl route add --net tcp --gateway 10.20.20.[2-8]@tcp1
lnetctl route add --net o2ib --gateway 10.10.20.[2-8]@o2ib1

# Health Configuration
lnetctl set retry_count 3
lnetctl set transaction_timeout 10
lnetctl set health_sensitivity 100
lnetctl set recovery_interval 1
Progress and Updates

- Multi-Rail/Dynamic Discovery/LNet Health have all landed
- We’re aiming to get UDSP and MR Routing in 2.13
UDSP

► Requirements:
  - https://wiki.whamcloud.com/display/LNet/Multi-Rail+User+Defined+Policies

► HLD:

► Code:
  - https://review.whamcloud.com/#/c/34580
UDSP Overview

Progress:

- Requirements: Complete
- Design: Complete
- Implementation: Complete
- Testing: 65% Complete
MR Routing

► Patches on gerrit on the Multi-Rail branch
  - https://review.whamcloud.com/#/c/34772
► Will be merged to master with the UDSP feature
► Refer to each router with its primary NID
► Multiple interfaces can exist on a single gateway
► No need to define a separate route for each NID on the gateway
► Select best gateway NI for message sending
MR Routing

- UDSP can be used to assign priority for individual gateway NIs
- LNet Health is used to maintain gateway health
- Discovery is used to maintain gateway aliveness
  - Discovery protocol uses ping. It’s backwards compatible
- Much of the code is simplified by reusing existing mechanisms
- Patch Description: [https://wiki.whamcloud.com/display/LNet/Patch+Description](https://wiki.whamcloud.com/display/LNet/Patch+Description)
MR Routing

► Progress:
  - Requirements: Complete
  - Design: Complete
  - Implementation: Complete
  - Testing: Complete
LNet Sysfs

► Progress:

- Requirements: Complete
- Design: Complete
- Implementation: Complete
- Review: Complete
- Testing: Complete
LNet Unit Test Framework

► Intent is to thoroughly test LNet functionality
► C/Python Hybrid
► Scripts written in Python
► Exercises LNet through the InetConfig API (same API used by Inetctl)
► Currently the code for the LUTF is on the Multi-Rail branch
  • [https://review.whamcloud.com/#/c/33181](https://review.whamcloud.com/#/c/33181)
LNet Unit Test Framework

Progress:
- Requirements: Complete
- Design: Complete
- Implementation: 65% (worked on in the background)
- Review: In Progress
- Testing: 30%
Roadmap

► IPv6 Support
► LUTF
► 4K message performance optimization
► LNet/Lustre Top (performance measurements)
► o2ibInd verbs update (Integrate new APIs added)
► Load control (QoS) depending on network/NID
► self-test enhancement
  – better statistics, different traffic flow
Summary

► Multi-Rail Routing is planned for 2.13
► UDSP is planned for 2.13
► LNet Sysfs statistics is planed for 2.13
► Next high-priority items:
  • Initial investigation of IPv6 implementation
  • Complete the LUTF test suite
Thank You

Questions?