Lustre® networking (LNET)

Isaac Huang 2008-03-12
LNET Overview

- Support for many commonly-used network types such as InfiniBand and TCP/IP
- RDMA, when supported by underlying networks such as Elan, Myrinet (MX), and InfiniBand
- Routing between multiple networks
- Almost full raw bandwidth with low CPU utilization (except on TCP), even over WAN
LNET features

MDS disk storage containing Metadata Targets (MDT)

Pool of clustered Metadata Servers (MDS) 1-100

Object Storage Servers (OSS) 1-1000’s

OSS storage with Object Storage Targets (OST)

Commodity Storage

Shared storage enables failover OSS

Enterprise-Class Storage Arrays and SAN Fabric

Lustre Clients 1 - 100,000

Elan Myrinet InfiniBand

Router

GigE

Simultaneous support of multiple network types

= failover
LNET Architecture

LNET users (PTLRPC, selftest)

Generic LNET

LND

LND

LND

LNET API

LND API

Network – Specific APIs
Lustre Network Drivers (LNDs)

- **Kernel**
  - socklnd – TCP/IP sockets
  - \{cib,open\}iblnd – Topspin IB
  - iiblnd – Silverstorm IB
  - viblnd – Voltaire IB
  - o2iblnd – OFA IB
  - ptllnd – Cray Portals
  - ralnd – Cray RapidArray
  - qswlnd – Quadrics Elan
  - gmlnd – Myricom GM (no RDMA)
  - mxlnd – Myricom MX

- **Userspace**
  - socklnd – TCP/IP sockets
  - o2iblnd – OFA IB
  - ptllnd – Cray Portals
Routing Overview

- Mostly static network topology
  - Route table built at startup from LNET config
  - Routers can be enabled and disabled
  - Utility can add and remove routers (currently used very rarely)

- Store-and-forward
  - LNET message is forwarded after it has been received completely

- Forwarding Buffer Credits
  - Back-pressure on buffer contention

- Resilience
  - Avoid dead routers
  - Re-use newly available routers
  - Router Checker

- Load Balance
  - Routed sends may be re-ordered in the network
Sample Routed Networks

- Clients
- Cluster 1
  - Switch
  - Elan4
- Routers
- Server Farm
  - OSS
  - MDS
- Cluster 2
  - Switch
  - IP Network
- Switch
- Storage Network
- Storage Island
LNET in Userspace

- Run LNET in userspace – liblnet
  - Linux & Solaris
  - Not all LNDs are available
  - Uo2iblnd available soon
  - No router in userspace
- Access kernel LNET via libula
LNET Selftest

• Overview
  • Easy to setup and use (everything done at test console)
  • Ping, bulk read/write (with data integrity checks)
• Smoke test
• Performance measurement
• Selftest in userspace
Anticipated Features

• IPv6 support

• Multiple Interfaces
  > Aggregation and failover
  > Currently rely on underlying network bonding, e.g. Ethernet and Elan4, or static aggregation with no failover by creating multiple networks
    – OFED ib-bonding is a misnomer – it only bonds IPoIB interfaces

• iWARP