Lustre Over BXI Update

Grégoire Pichon
Atos HPC software R&D
09/2021

Atos

Content overview

01. LNet and Portals4 LND overview

02. Portals4 LND enhancements

03. LND integration and packaging

04. Lustre over BXI clusters

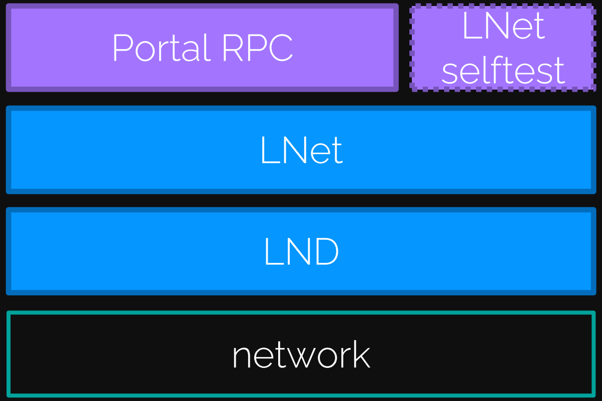
01. LNet and Portals4 LND overview

Lustre Network & Lustre Network Driver

Overview

- LNet
  - communication infrastructure between Lustre clients and servers
  - allows routing between networks through Lustre routers
  - key features
    - RDMA transfers
    - routers high availability and recovery
    - interfaces high availability and aggregation on multi-rail nodes

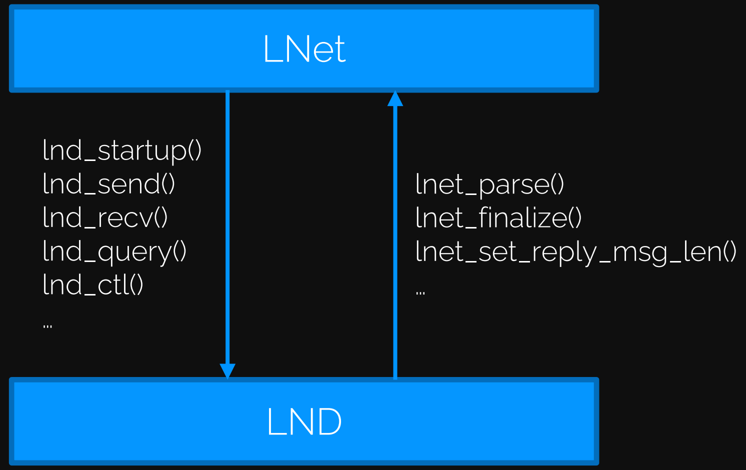
- LND
  - allows support for specific network hardware
  - supports many commonly-used network types: Infiniband, Ethernet
  - transports LNet requests and responses



LNet – LND interface

What are the requirements of a LND ?

- register to LNet
  - lnet_register_lnd()
  - lnet_unregister_lnd()
- provide a lnet_lnd structure
  - lnd type (SOCKLND, O2IBLND, LOLND, GNILND, PTL4LND)
  - startup/shutdown network communication on the interface
  - send/receive LNet messages
  - notify /query on peer health / aliveness
  - handle control commands
- use LNet callbacks
  - parse received message for LNet matching
  - finalize message transmission for LNet event generation



Portals 4 LND

Overview

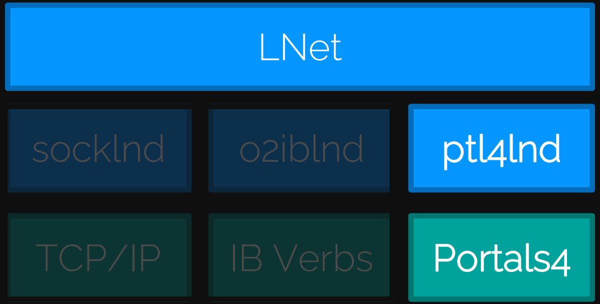
- Bull eXascale Interconnect (BXI)
  - 100Gb/s NIC, BXI V1 (2018), BXI V2 (2021)
  - hardware implementation of Portals4

- ptl4lnd
  - rely on Portals4 network API
  - network name: ptlf - adapter identified by its device number

    networks = ptlf2(0),ptlf6(1)
  - LNet address built from BXI network id

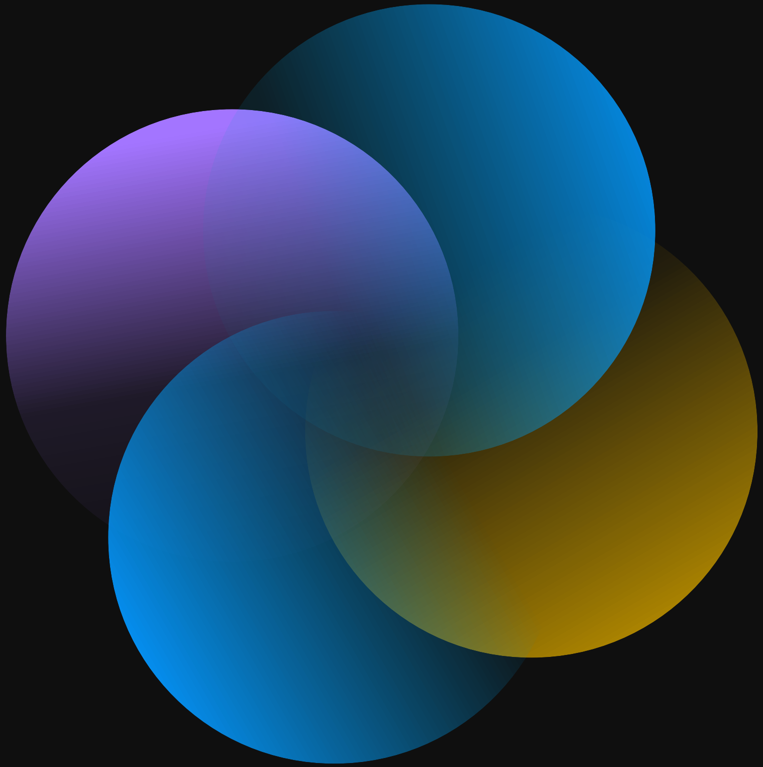
    42@ptlf2, 43@ptlf6

- LND key features
  - immediate and rendez-vous (RDMA) transmissions
  - peer status management
  - flow control and resource management



See LAD'17 presentation "Overview of the new Portals4 LND"

02. Portals4 LND enhancements



Portals4 LND improvements

What has been improved since initial version ?

Performance

Separation of immediate and rendez-vous traffic

- use distinct network channels (PTE) for bulk-io data transmissions
- reduce list matching overhead

Performance

Parallelization and NUMA Binding

- independent worker threads for LND internal processing
- CPT aware LND worker threads

Robustness

Robustness

- improve reliability of LND to unexpected/malformed messages and unexpected Portals events

Platform

ARM support

- handle 64K pages
- impacts on memory allocations, iovec segments limit, …
- BXI V2 hardware required

Portals4 LND improvements

Which “recent” LNet features have required changes to the LND ?

- LNet Multi-rail and Health status
  - need to report interface up/down through lnet_ni_t->ni_fatal_error_on
  - need to report transmission status through lnet_msg_t->msg_health_status

msg_health_status	Case
LNET_MSG_STATUS_OK	Transmission succeed
LNET_MSG_STATUS_LOCAL_TIMEOUT	Transmission is stuck in LND Send-Queue for too long (missing tx credits or hello handshake hung)
LNET_MSG_STATUS_LOCAL_ERROR	LND resources are exhausted (missing tx descriptors, peer table full, memory allocation failed)
LNET_MSG_STATUS_REMOTE_DROPPED	Response from remote LND indicates that LNet did not find a matching ME
LNET_MSG_STATUS_REMOTE_ERROR	Remote LND resources are exhausted, or peer is unreachable
LNET_MSG_STATUS_NETWORK_TIMEOUT	Transmission is stuck in LND Active-Queue for too long

Portals4 LND Multi-rail

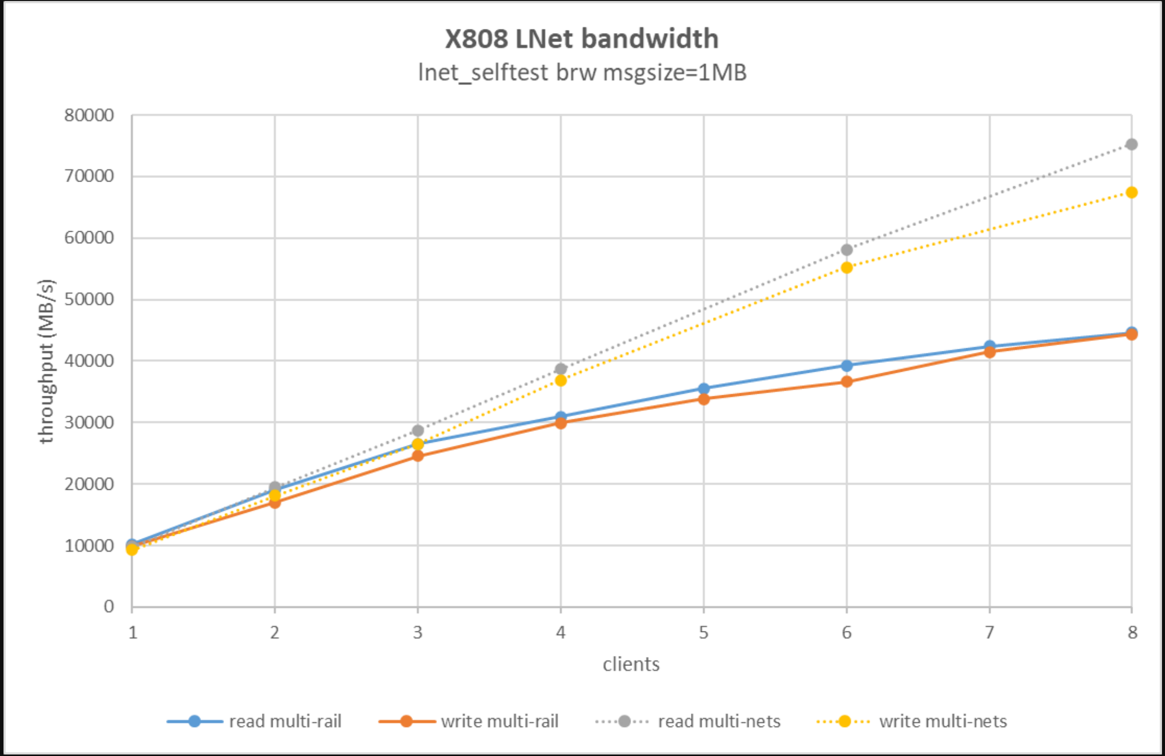
Multi-rail Performance

Configuration

- 1 X808 8-sockets server, with 8 BXI V2 adapters
- 8 2-sockets servers, with 1 BXI V2 adapter each
- take care of multi-rail interface binding (LU-14875)
- Redhat 8, Lustre 2.12.6

Tests

- Lnet selftest : X808 in group1, 1-8 clients in group2
- multi-rail: 1 ptlf lnet network
- multi-nets: 8 ptlf lnet networks



With multi-rail, bandwidth scales up to 45 GB/s … but should be able to reach 70-80 GB/s

03. LND integration and packaging

LND with numeric address

Integration issue

- Correctly handled by lnet kernel module
  - interface name is parsed by the LND itself
  - nid processing managed by libcfs routines declared in libcfs_netstrfns table
    either libcfs_num_xxx() or libcfs_ip_xxx()

- Issues with Dynamic LNet configuration
  - when handling a numeric interface name
    lnetctl net add --net ptlf --if 0
    lnetctl import <file>" with "interfaces: 0: 0
  - when handling a numeric address nid
    lnetctl route add --net o2ib0 --gateway [42-43]@ptlf0

- Issues reported or to be reported to the community
  - LU-11860, patch "lnet: support config of LNDs with numeric intf name" integrated in Lustre 2.13

Building LNDs in separate packages

Packaging issue

- Example
  - Lustre packages configured with o2ib built against Mellanox OFED
  - target cluster contains some nodes with IB adapters and some nodes with Ethernet or BXI adapters
  - Lustre installation will require Mellanox OFED packages
  - Why should we have to install Mellanox OFED on nodes that have no IB hardware ?

- Optionally package LNDs in their own RPM (LU-11824, Sébastien Piechurski)
  - limit dependency on third-party network packages to the LND package
  - administrators can select Lustre & LND packages that need to be installed on each node
    - configure option: --with-separate_lnds=o2ib
    - separate RPM: kmod-lustre-lnd-o2ib

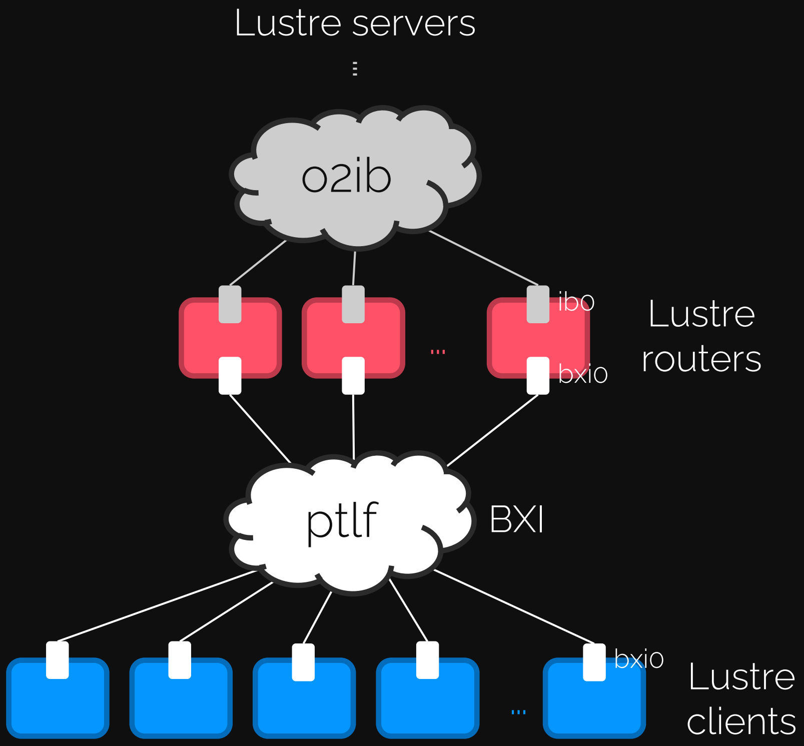
04. Lustre over BXI clusters

T1K at CEA (2018)

Cluster description

- Lustre clients
  - BullSequana X1110 Intel Xeon Phi KNL 68-cores
  - 1 BXI V1 adapter
- Lustre routers
  - BullSequana R423-E4 2-sockets Intel Xeon Broadwell 14-cores
  - 1 BXI V1 adapter, 1 IB EDR adapter
- Redhat 7.9, Lustre 2.12.6
- 30 groups of 276 clients + 5 routers
  - routers attached to L1 switches with 10m to 25m cables

Similar installation for Joliot-Curie cluster at TGCC (2018)



T1K at CEA

Tuning and Performance

- Lustre clients tuning

```
lnet networks=ptlf(0) routes='o2ib [1,2,3,4,5]@ptlf'
        lnet_peer_discovery_disabled=1 lnet_health_sensitivity=0
        check_router_before_use=1 live_router_check_interval=107 dead_router_check_interval=50
kptl4lnd peer_credits=32
```

- Lustre routers tuning

```
lnet networks=ptlf(0)[0],o2ib(ib0)[1] forwarding=enabled
        lnet_peer_discovery_disabled=1 lnet_health_sensitivity=0
kptl4lnd peer_credits=32 ntx=8192
```

Performance (BXI V1, FEC activated)	Read Bandwidth	Write Bandwidth
LNet 1 client – 1 router (10m cable)	8,0 GB/s	7,8 GB/s
LNet 1 client – 1 router (25m cable)	6,0 GB/s	6,0 GB/s
IOR 272 clients – 5 routers (4ppn, FPP, directIO)	39,5 GB/s	37,5 GB/s
IOR 128 clients – 5 routers (8ppn, SSF, MPIIO)	32,8 GB/s	32,5 GB/s

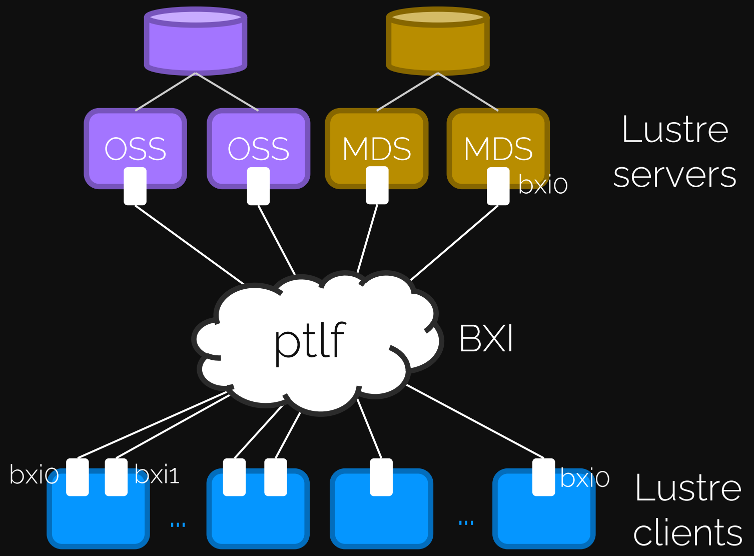
FEC: Forward Error Correction, FPP: File Per Process, SSF: Single Shared File

Romeo at Université de Reims Champagne-Ardenne (2018)

Cluster description

- Lustre clients
  - 70 GPU compute nodes
    BullSequana X1125 - Intel Xeon Skylake 2-sockets, 14-cores, 2 BXI V1 adapters, 4 Nvidia Tesla P100 GPUs
  - 45 compute nodes
    BullSequana X1120 - Intel Xeon Skylake 2-sockets, 14-cores, 1 BXI V1 adapter
  - 2 login nodes
    BullSequana X410-E5 – Intel Xeon Skylake 2-sockets, 6-cores, 1 BXI V1 adapter
- Lustre servers
  - 4 service nodes
    BullSequana X430-E5 – Intel Xeon Skylake 1-socket, 12-cores, 1 BXI V1 adapter
  - 1 filesystem with 12 OSTs, 1 MDT, 1 MGT

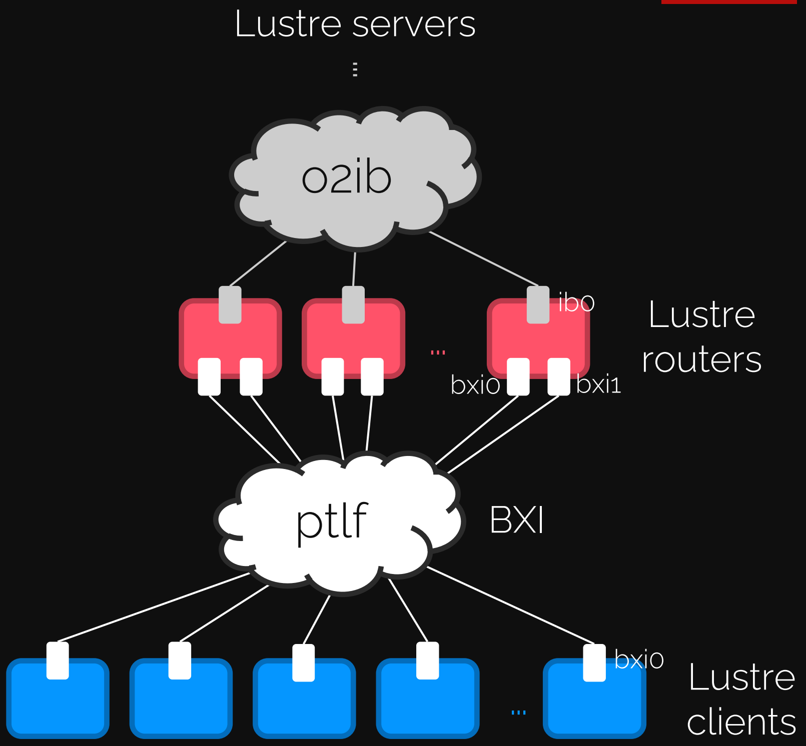
- Redhat 7.9
- Lustre 2.12.6



Exa1-HFi at CEA (2021)

Cluster description

- Lustre clients
  - BullSequana X2410 AMD EPYC 7763 2-sockets 64-cores
  - 1 BXI V2 adapter
- Lustre routers
  - BullSequana X431 AMD EPYC 7452 1-socket 32-cores
  - 2 BXI V2 adapters, 1 IB HDR adapter
- Redhat 8.3, Lustre 2.12.6
- clients/routers ratio = 576/4
  - routers attached to dedicated L1 switches



Exa1-HFi at CEA

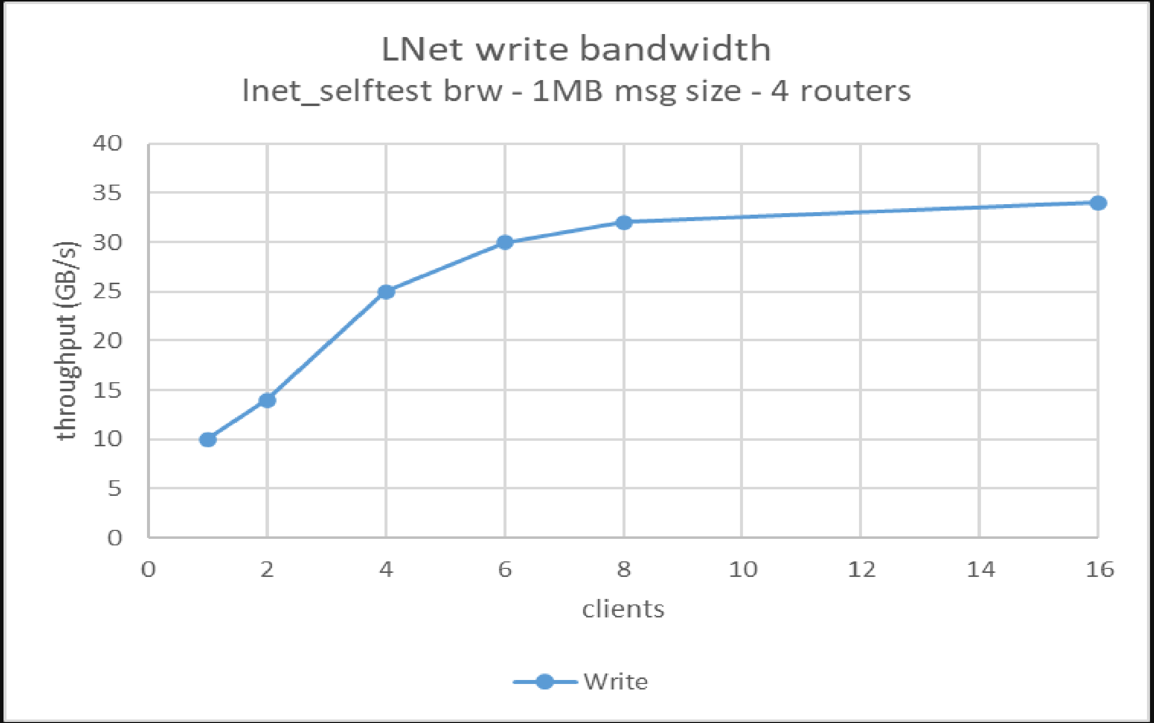
Tuning and Performance

- Lustre clients tuning

```
lnet networks=ptlf(0)[0]
        routes='o2ib [1,2,3,4,5]@ptlf'
        lnet_peer_discovery_disabled=1
        lnet_health_sensitivity=0
        check_router_before_use=1
        live_router_check_interval=107
        dead_router_check_interval=50
kptl4lnd peer_credits=32
```

- Lustre routers tuning

```
lnet networks=ptlf(0)[0],o2ib(ib0)[1]
        forwarding=enabled
        lnet_peer_discovery_disabled=1
        lnet_health_sensitivity=0
kptl4lnd peer_credits=32 ntx=8192
```



- Plan to enhance the setup of Lustre routers by using the 2nd bxi interface with multi-rail

Wrap-Up

Lustre over BXI is running and performing well on HPC production clusters

Integration effort of Portals4 LND within Lustre sources needs to be carried on

Portals4 LND will continue to be updated and tested with new LNet features of latest Lustre versions

Thank you!

For more information please contact:
gregoire.pichon@atos.net