



Remote Lustre Access

Thomas Roth 2. September 2024

- Lustre at GSI
- Remote access?
- Lustre mounting: LNET, module config, routing
- LNET routing between sites
- Node mapping
- User mapping

Lustre at GSI

- since 2008
- currently Lustre 2.12.5
- Main storage for GSI compute cluster
- WLCG (Alice) with 6 PiB

Bulk data

- Net capacity: 57 PiB
- $\sim 38\%$ utilized

Metadata

- Capacity: $11.55 * 10^9$ inodes
- 800 M inodes ($\sim 7\%$) utilized

What if ...

- ... you need to access your data from *somewhere else*?
- ... and you have a high bandwidth connection to GSI?
- \Rightarrow Long range Lustre mount

Mounting Lustre on client

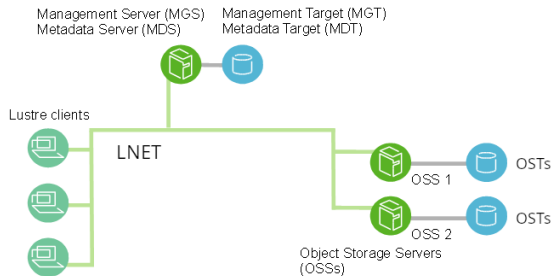
- means loading several kernel modules
 - thereby starting many kernel threads
- ⇒ *root access* on that client
- 👑 *root access* to Lustre ???

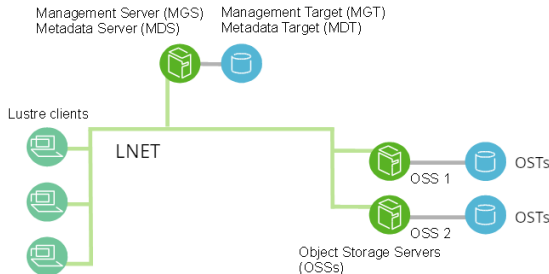
Mounting Lustre on client

- means loading several kernel modules
 - thereby starting many kernel threads
- ⇒ *root access* on that client
- 👤 *root access* to Lustre ???

... high bandwidth connection

- across different network types?





LNET

- LNET = virtual networking layer between Lustre nodes
- Lustre Network Identifier: *IP-Address @ HW-driver*, e.g. 10.10.24.313@tcp0 or 10.5.20.400@o2ib3
- *tcp* = used for Ethernet
- *o2ib* = used for Infiniband (Nvidia)
- ... other network technologies

LNET module options

- LNET and interfaces are specified as *modprobe* - options
 /etc/modprobe.d/lnet.conf:
 options lnet networks=o2ib7(ib0)
 - *ib0* = name of the local IB interface

LNET module options

- LNET and interfaces are specified as *modprobe* - options
 `/etc/modprobe.d/lnet.conf:`
 `options lnet networks=o2ib7(ib0)`
 - *ib0* = name of the local IB interface

LNET NIDs

- `10.6.23.11@o2ib3`: *IP-Address @ HW-driver Number*
- Several LNETs can co-exist in the same IB fabric: *o2ib1*, *o2ib5*
- \Rightarrow several distinct Lustre FS mounted on the same node

LNET module options

- LNET and interfaces are specified as *modprobe* - options
`/etc/modprobe.d/lnet.conf:`
`options lnet networks=o2ib7(ib0)`
 - *ib0* = name of the local IB interface

LNET NIDs

- `10.6.23.11@o2ib3`: *IP-Address @ HW-driver Number*
- Several LNETs can co-exist in the same IB fabric: *o2ib1*, *o2ib5*
- \Rightarrow several distinct Lustre FS mounted on the same node

LNET module options

```
/etc/modprobe.d/lnet.conf:  
options lnet networks="o2ib7(ib0),o2ib8(ib0)"
```

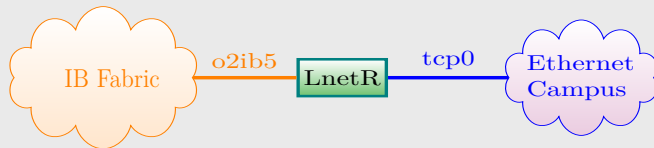
LNET Routing

- LNET accross different network technologies requires LNET routing
- Example GSI: Lustre FS in Infiniband fabric, batch farm in the same fabric: LNET config simple
- GSI campus is on Ethernet:
- Task: mount Lustre on Ethernet client!

LNET Routing

- LNET accross different network technologies requires LNET routing
- Example GSI: Lustre FS in Infiniband fabric, batch farm in the same fabric: LNET config simple
- GSI campus is on Ethernet:
- Task: mount Lustre on Ethernet client!

LNET client



LNET Routing

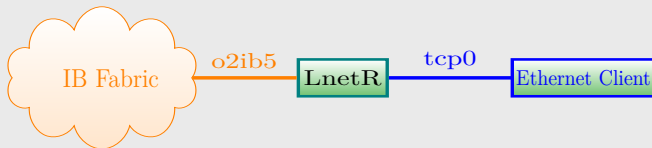
- LNET accross different network technologies requires LNET routing
- Example GSI: Lustre FS in Infiniband fabric, batch farm in the same fabric: LNET config simple
- GSI campus is on Ethernet:
- Task: mount Lustre on Ethernet client!

LNET Router

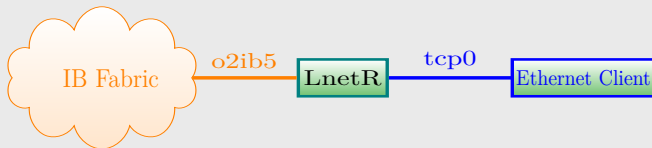
- Set up dual-homed Linux box: Infiniband (*ib0*) + Ethernet (*eth0*)

```
options lnet networks="o2ib5(ib0),tcp0(eth0)"  
options lnet forwarding="enabled"  
options lnet accept=all
```
- First line specifies two LNETs
- Other lines switch on routing

LNET client



LNET client



- LNET config on Ethernet client
`/etc/modprobe.d/lnet.conf:`
`options lnet networks="tcp0(eno1)"`
`options lnet routes="o2ib5 10.10.24.206@tcp0"`
- First line: client lives on *tcp0*
- Second line: client can reach *o2ib5* by going through 10.10.24.206@*tcp0*
- Lustre mount command identical to IB clients: *mount -t lustre 10.20.34.1@o2ib5:/lustre /mountpoint*


Remote LNET client

- Location of Ethernet client unimportant:
 - Across campus
 - 100 km away at other institute

Remote LNET client

- Location of Ethernet client unimportant:
 - Across campus
 - 100 km away at other institute

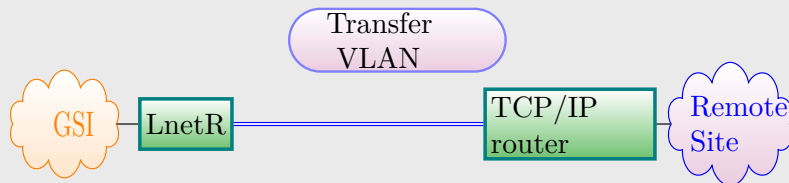
1 Remote network structure

-  Far away institute might not want to expose their clients directly to outside
 - Need for routers at remote site
 - Use a Transfer VLAN between the two sites
 - Remote router has to translate network packages from the transfer net to the designated local net (at the remote site)







Remote LNET client

- Location of Ethernet client unimportant:
 - Across campus
 - 100 km away at other institute


LNET client



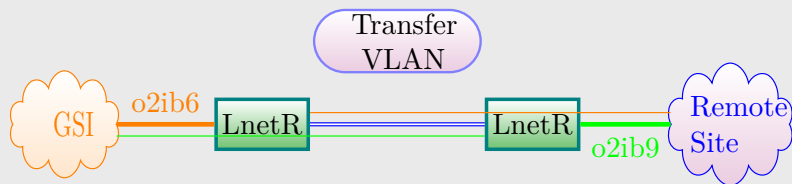
2 Remote network structure

-  Far away institute does not want to route their internal network to the connecting box.
-  But: There is another Lustre at the remote institute!
 -  The desired connection is between *GSI* and the *LNET clients* at the *remote site*
 -  Set up the remote box as another LNET router
 -  Remote LNET router now transports remote LNET to the connecting point
 -  Local LNET routes gets another *route*

```

/etc/modprobe.d/lnet.conf:
options lnet networks="o2ib6(ib0),tcp0(enp23s0d1)"
>> options lnet routes="o2ib9 192.168.5.216@tcp0" <<
options lnet forwarding="enabled"
options lnet accept=all
                    
```
 -  where *192.168.5.216@tcp0* = *NID of the remote LNET router*

👤 Remote LNET router



Example from Lustre Manual

- Some University runs a Lustre FS, e.g. in the Physics department
- Group of Biologists doing bird research hear about it and want to store data
- But bird researchers don't want Physicist accounts and ssh-access to Physics computers etc.
- Bird researchers want that Lustre on their own work station!

Problems

- 1 Bird researcher workstations on different network
- 2 Bird researcher workstations administrated by someone = unknown!
- 3 Bird researcher UIDs not known at Physics site

Problems

- ❶ Bird researcher workstations on different network
- ❷ Bird researcher workstations administrated by someone = unknown!
- ❸ Bird researcher UIDs not known at Physics site
 - Problem 1 ✓ (LNET routing)
 - What about alien users and alien administrators?

Lustre nodemap: Policy engine for NIDs and for UIDs

- When an operation is made from a NID, Lustre decides if that NID is part of a nodemap.
- If no policy group exists for that NID, access is squashed to user nobody by default.
- Properties of policy groups (e.g. *trusted*, *admin*) manage access.
 - Each policy group also has *identity maps*.
 - *idmaps* determine how UIDs and GIDs on the client are translated into the canonical user space on Lustre.

Number of commands

- Switch feature on: `mgs# lctl nodemap_activate 1`
- Add a nodemap: `mgs# lctl nodemap_add Birds`
- Add an IP (range) to this map:
 - `mgs# lctl nodemap_add_range -name Birds -range 192.168.10.6@o2ib1`

Number of commands

- Switch feature on: `mgs# lctl nodemap_activate 1`
- Add a nodemap: `mgs# lctl nodemap_add Birds`
- Add an IP (range) to this map:
 - `mgs# lctl nodemap_add_range -name Birds -range 192.168.10.6@o2ib1`

Properties

- Two important properties of a nodemap: *trusted* and *admin*
- *trusted* hosts see UIDs and GIDs without any translation
- *admin* hosts are not root-squashed
- Important step: define a *Admin* nodemap that contains **all** Lustre servers and has both *trusted* = 1 and *admin* = 1 - otherwise, one will eventually disable inter-server communication
- Also include some administrative clients in this nodemap, there to be root on the mounted filesystem.

Default nodemap

- Nodes that are not mentioned - neither in *Birds* nor in *Admin* - can still mount Lustre
- By default, they are not *trusted* nor *admin*, so *root* becomes *nobody* and users cannot see UIDs/GIDs of the system
- Seems to be a good idea to set *trusted=1* on the Default nodemap: new (batch) clients can mount and users on these new clients can see Lustre, but root is squashed to nobody

This addresses problem no 2 = control which clients connect and who has administrative permissions.



Of course, it becomes necessary to obtain the visible NIDs of the potential remote clients, to add them to the respective nodemap.

Identity mapping

- Identity maps or idmaps are kept for each nodemap identity mapping policy group
- These idmaps determine how UIDs and GIDs on the client are translated into the canonical user space of the local Lustre file system.

Identity mapping

- Identity maps or idmaps are kept for each nodemap identity mapping policy group
- These idmaps determine how UIDs and GIDs on the client are translated into the canonical user space of the local Lustre file system.

Bird researcher mapping

- Suppose among the bird researcher there is the user *eagle* = 1030.425
- If *eagle* writes files to Lustre, only he or members of his group should be able to access them.
- And the bird researchers should not see the uids/gids or have access to the files of the regular users.

Bird researcher mapping

 An *idmap* is added:

- `mgs# lctl nodemap_add_idmap -name Birds -idtype uid -idmap 1030:30120`
- The *uid* 1030 from the bird watchers user space is mapped to the *uid* 30120
- (A similar command maps remote and local GIDs.)

⇒ The local *uid* 30120 is a kind of dummy account.

- Reserve a *uid* range in your local user data base for this purpose
- E.g. *uids* 30120 – 30300 = *lnmbird001* – *lnmbird181*



Needs its own database for bookkeeping, and on the MGS it's all manual commands.

- On a GSI Lustre *Admin* client:

```
root@lxbk0469:~# ls -la /lustre/
total 20
drwxr-xr-x  6 root          root      4096 Oct 19 18:30 .
drwxr-xr-x 32 root          root      4096 Oct  9 09:48 ..
drwxr-xr-x  3 lnmbird001 lnmbird  4096 Oct 19 18:53 birdsnest
drwxr-xr-x  4 troth        hpc       4096 Oct  1 11:04 hpc
```

- On the Bird researchers workstation:

```
root@BigNest:~# ls -la /lustre/
total 20
drwxr-xr-x  6      99          99 4096 Oct 19 18:30 .
drwxr-xr-x 32   root   root      4096 Oct  9 10:14 ..
drwxr-xr-x  3 eagle   birds     4096 Oct 19 18:53 birdsnest
drwxr-xr-x  4      99          99 4096 Oct  1 11:04 hpc
```


- GSI *default* client, *trusted=0*: everybody is squashed

```
root@lxbk0472:~# ls -la /lustre/
total 20
drwxr-xr-x  6   99   99 4096 Oct 19 18:30 .
drwxr-xr-x 32 root root 4096 Oct  9 09:56 ..
drwxr-xr-x  3   99   99 4096 Oct 19 18:53 birdsnest
drwxr-xr-x  4   99   99 4096 Oct  1 11:04 hpc
```

- GSI *default* client, *trusted = 1*: only *root* is squashed

```
root@lxbk0472:~# ls -la /lustre/
total 20
drwxr-xr-x  6   99          99    4096 Oct 19 18:30 .
drwxr-xr-x 32 root          root    4096 Oct  9 09:48 ..
drwxr-xr-x  3 lnmbird001 lnmbird 4096 Oct 19 18:53 birdsnest
drwxr-xr-x  4 troth        hpc     4096 Oct  1 11:04 hpc
```