Lustre Networking at Cray

Chris Horn
hornc@cray.com

Agenda

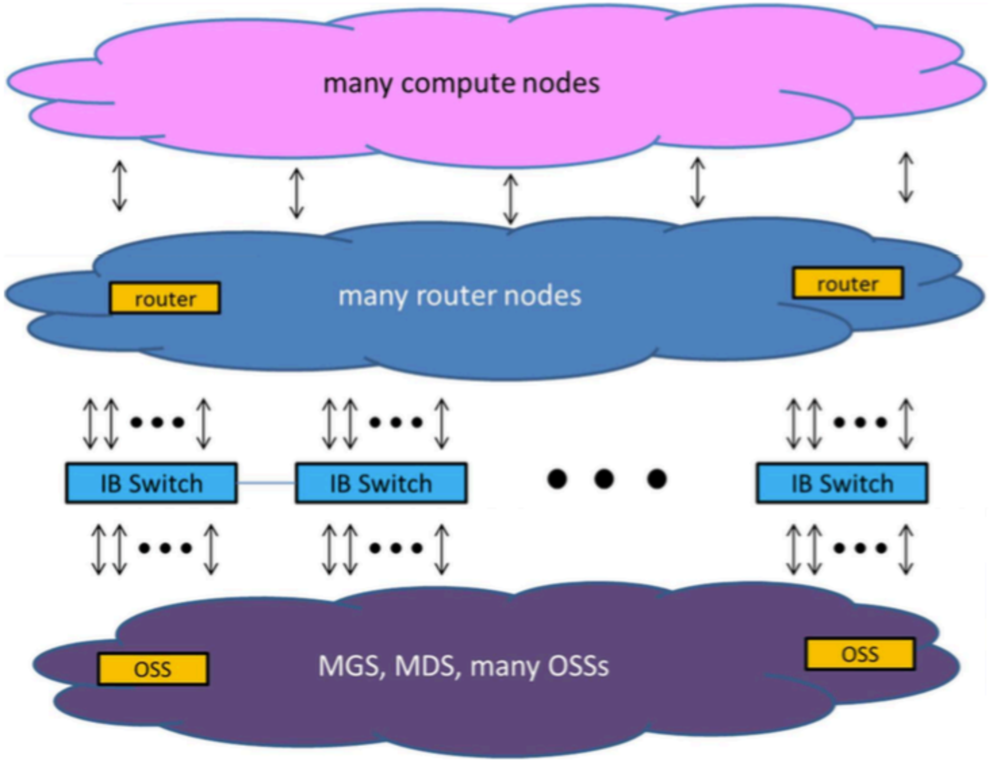
- **Lustre Networking at Cray**
  - LNet Basics
  - Flat vs. Fine-Grained Routing
  - Cost Effectiveness - Bandwidth Matching
  - Connection Reliability – Dealing with ARP Flux
  - Serviceability – Generating and Emplacing Configuration
- **Recent LNet Work in the Lustre Community**
  - Support for new Mellanox Hardware
  - Multiple Fabric Support
- **Summary**
- **Q&A**

LNet Basics

- **LNet is Lustre Networking layer**
- **Network type agnostic**
  - Lustre Network Drivers (LNDs) provide interface to specific network drivers
    - gnilnd (Aries/Gemini)
    - o2iblnd (InfiniBand/OPA)
    - socklnd (Ethernet)
- **LNet routers bridge clients on Cray's high speed network with external Lustre servers**
  - Gemini/Aries ←→ InfiniBand
  - Two types of routing: Flat and Fine-Grained

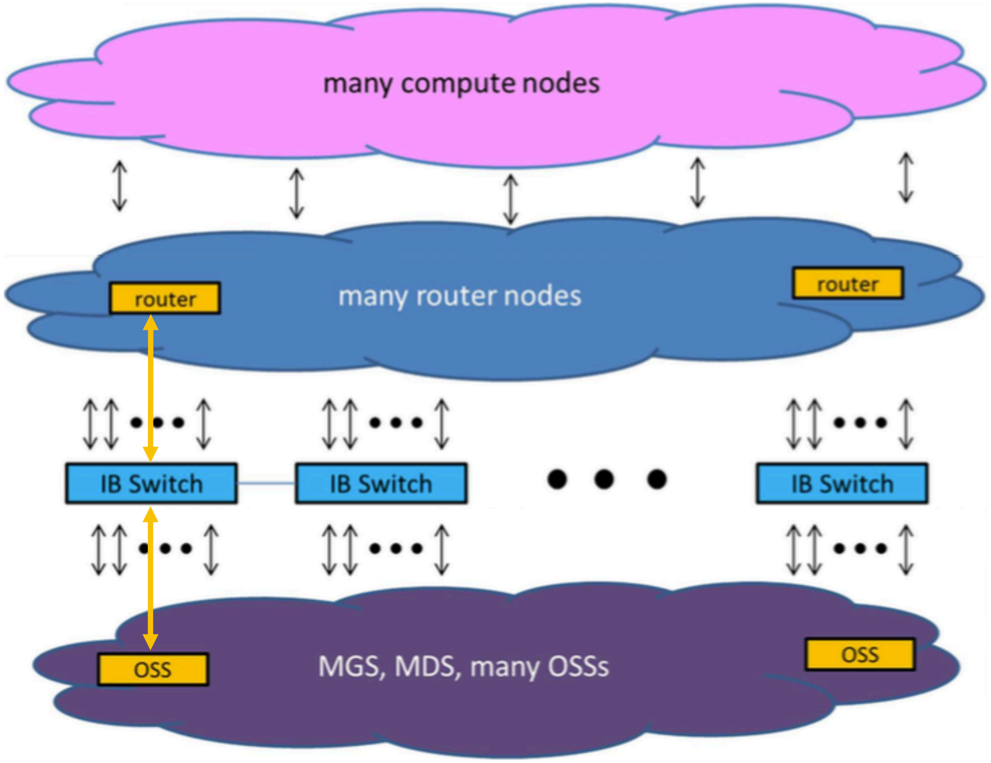
Flat LNet

- Simple configuration
- Any router can talk to any other peer



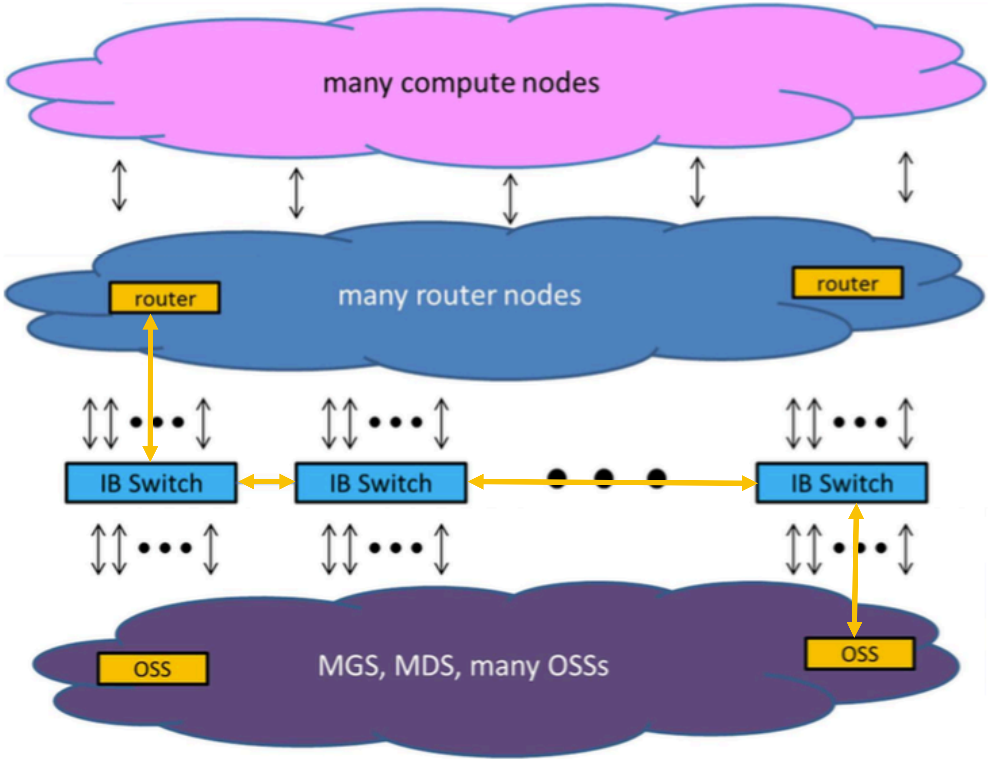
Flat LNet

- Performance can be optimal at small scale



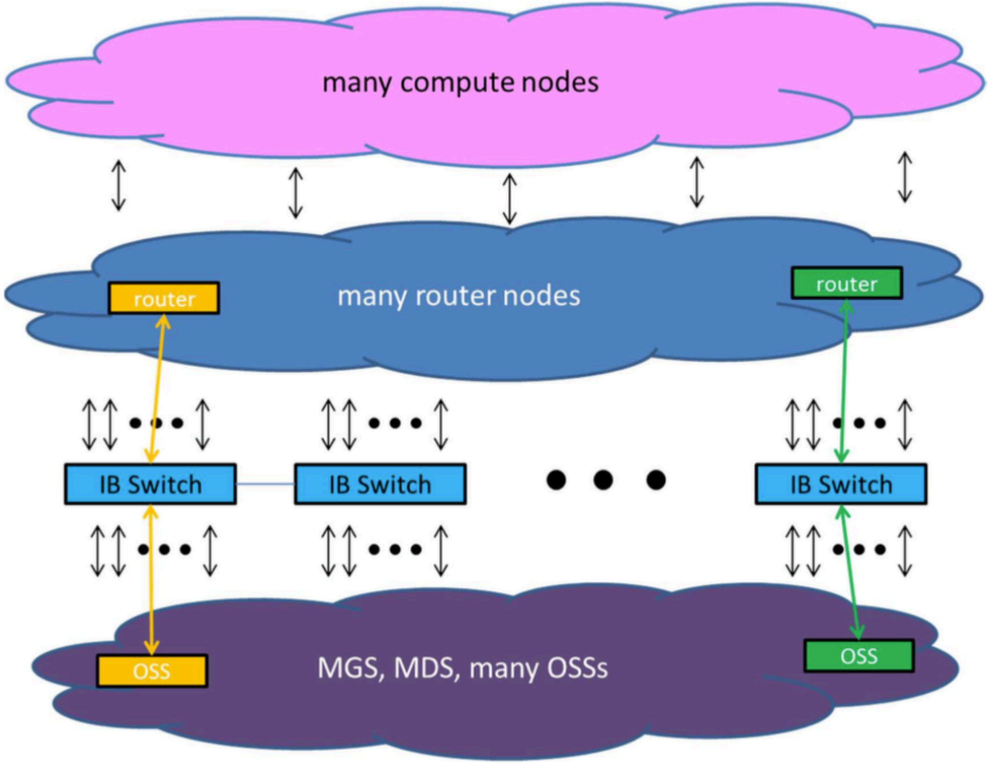
Flat LNet

- **Performance suffers at large scale from need to traverse inter-switch links**



Fine-Grained Routing

- Define groups of peers
- Best performance at scale by avoiding ISLs
- Complex configuration
  - # Groups is total # of servers divided by # servers in each group



Cost Effectiveness and Bandwidth Matching

	1*	2	3	4	5	6
Sonexion-1600	3.00	6.00	9.00	12.00	15.00	18.00
Sonexion-2000	3.75	7.50	11.25	15.00	18.75	22.50
Single HCA	5.50	11.00	16.50	22.00	27.50	33.00
Dual HCA	4.20	8.40	12.60	16.80	21.00	25.20

- **Need to provide sufficient IB bandwidth in cost-effective manner**
  - No network bottlenecks
  - Minimize excess bandwidth

*** Average throughput of 1 Server or IB link; 2 Servers or IB links; etc.**

Bandwidth Matching

	1*	2	3	4	5	6
Sonexion-1600	3.00	6.00	9.00	12.00	15.00	18.00
Sonexion-2000	3.75	7.50	11.25	15.00	18.75	22.50
Single HCA	5.50	11.00	16.50	22.00	27.50	33.00
Dual HCA	4.20	8.40	12.60	16.80	21.00	25.20

- **Single HCA == Bandwidth of one IB port on XC40 LNet router node with one IB HCA**
- **Dual HCA == Bandwidth of one IB port on XC40 LNet router node with two IB HCAs**

*** Average throughput of 1 Server or IB link; 2 Servers or IB links; etc.**

Bandwidth Matching

	1*	2	3	4	5	6
Sonexion-1600	3.00	6.00	9.00	12.00	15.00	18.00
Sonexion-2000	3.75	7.50	11.25	15.00	18.75	22.50
Single HCA	5.50	11.00	16.50	22.00	27.50	33.00
Dual HCA	4.20	8.40	12.60	16.80	21.00	25.20

- **6 Sonexion 2000 OSSes (3 SSUs) ~ 22.5 GB/s**
- **5 IB Links (from single HCA routers) ~ 27.50**
- **Servers using ~ 82% of available network bandwidth**

Bandwidth Matching

	1*	2	3	4	5	6
Sonexion-1600	3.00	6.00	9.00	12.00	15.00	18.00
Sonexion-2000	3.75	7.50	11.25	15.00	18.75	22.50
Single HCA	5.50	11.00	16.50	22.00	27.50	33.00
Dual HCA	4.20	8.40	12.60	16.80	21.00	25.20

- **6 Sonexion 2000 OSSes (3 SSUs) ~ 22.5 GB/s**
- **6 IB Links (from dual HCA routers) ~ 25.2 GB/s**
- **Servers using ~ 90% of available network bandwidth**
- **Ideal ratio *n:n***

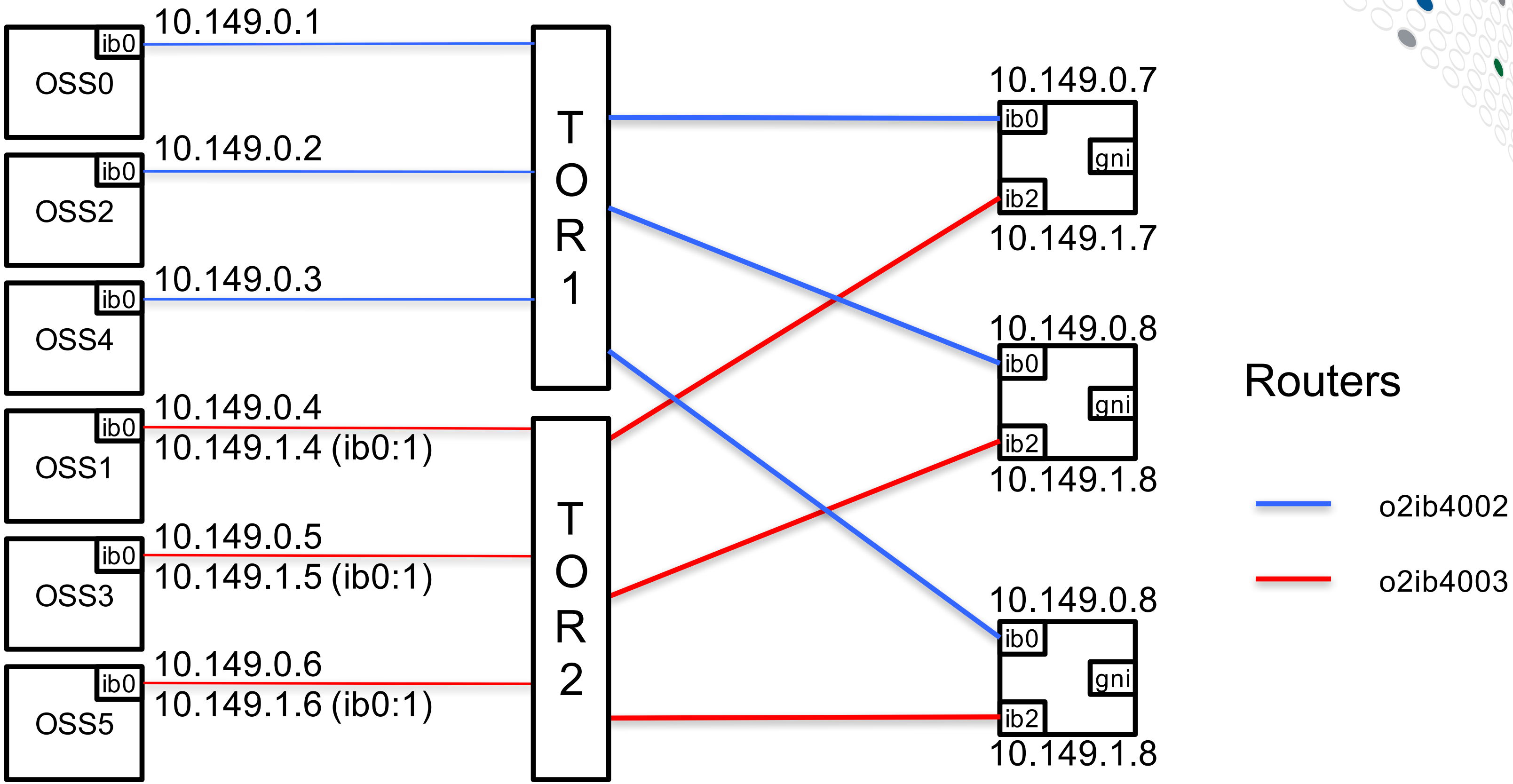
Connection Reliability – Dealing with ARP Flux

- **Address Resolution Protocol (ARP)**
  - Maps Network layer address (e.g. IPv4) to link layer address (e.g. MAC address)
  - Broadcasts ARP "who-has" request to all peers, "Who has IP w.x.y.z?"
  - Peer who-has IP w.x.y.z responds with its MAC address
- **"Flux" occurs when multiple interfaces are on a single host**
  - Both interfaces may respond to ARP request
    - non-deterministic population of the ARP cache (a.k.a. neighbor table)
  - Breaks IPoIB ☹

ARP Flux cont.

- **Can workaround by issuing “lctl ping” from routers to servers**
  - Routers populate server’s ARP cache
- **Investigated using kernel IP tunables but found it insufficient**
  - net.ipv4.conf.all.arp_ignore = 1
  - net.ipv4.conf.all.arp_announce = 2
- **Currently recommend placing interfaces on separate subnets**
  - More complexity

LNet Configuration



Serviceability - Dealing with Complexity

- Cray LNet Configuration and Validation Tool
- Simple and descriptive input file
- Knowledge of Cray Sonexion IB switch configuration
- Generates "ip2nets" and "routes" LNet module parameters
  - Typically stored in files: "ip2nets.dat" and "routes.dat"
- Validates configuration
  - Validate IB connectivity
  - Validate LNet group membership
  - Validate LNet destinations

Add/Remove IP alias to ib0 on module load

```
/sbin/ip -o -4 a show ib0 | \
/usr/bin/awk '/inet/{s=$4;
    sub("10\\.149\\.0\\.","10.149.1.",s);
    print "/sbin/ip address add dev ib0 label ib0:1", s}' | \
/bin/sh
/sbin/modprobe --ignore-install lnet
```

```
/sbin/modprobe -r --ignore-remove lnet &&
/sbin/ip -o -4 a show label ib0:1 | \
awk '{print "/sbin/ip address del dev ib0 label ib0:1", $4}' | \
/bin/sh
```

Hat tip to Dave McMillen

LNet Design/Config Overview

- Use bandwidth matching to get router:server ratio
- Determine IP addressing scheme
- Use *clcvt* to generate ip2nets and routes configuration
- Configure interfaces
- Plug in cables
- Emplace LNet configuration
  - ip2nets, routes, other module parameters

Configuration Emplacement

- **Sharedroot in CLE < 6.0**
  - Access sharedroot from bootnode: xtopview -c lnet
  - Edit modprobe.conf.local:
    - options lnet ip2nets = “/path/to/ip2nets.dat”
    - options lnet routes = “/path/to/routes.dat”
- **Config sets in CLE >= 6.0**
  - Run *cfgset* command on smw:
    - `cfgset update --service cray_lnet --mode interactive CONFIGSET`
    - See slides at end of deck for example
  - Advanced users can manipulate worksheets

Recent LNet Work in the Lustre Community

Memory Registration in o2iblnd

- **Historically supported PMR and FMR APIs**
  - Physical Memory Region (PMR) dropped
  - Fast Memory Region (FMR) deprecated
- **"Fast Registration API" is the new (Linux 2.6.27) hotness**
- **Mellanox hardware utilizing mlx5 drivers do not support FMR**
- **LU-5783: Adds support for Fast Registration API**
  - Fallback for FMR
  - Landed for upcoming Lustre 2.9 release

Mixed Fabric Concerns

- **How to optimize ko2iblnd in presence of multiple HCAs?**
  - OPA ←→ EDR; EDR ←→ FDR; Aries ←→ FDR(ib0) & EDR(ib2)
- **LU-7101: per NI map_on_demand values**
  - FMR enhances performance of OPA
  - FMR enabled by setting: 0 < map_on_demand <= 256
  - MLX5 does not support FMR, so needs map_on_demand = 0
  - Works in conjunction with LU-3322 to allow optimal settings
  - Landed for upcoming Lustre 2.9 release
- **LU-3322: Allow different peer_credits and map_on_demand values**
  - Available in just released Lustre 2.8

Summary

- **Covered some LNet basics:**
  - Flat vs. Fine Grained Routing
- **Cost/Reliability/Serviceability:**
  - Bandwidth Matching
  - ARP Flux
  - Cray LNet Configuration and Validation Tool - clcvt
- **New configuration emplacement**
  - Bye Bye Sharedroot! Hello config sets!
- **Recent changes in Lustre for new IB technology**
  - LU-5783, LU-3322, others
- **Mixed fabric**
  - Dealing with different HCAs that use ko2iblnd

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, REVEAL, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.

Q&A

Chris Horn
hornc@cray.com

What is Multi-Rail

- Use multiple independent networks, or “rails”, to overcome bandwidth limitations or increase fault tolerance
- Allow communication between two hosts across multiple interfaces
  - One or more networks
  - Interfaces used concurrently
- Cray utilizes multiple interfaces in non-multi-rail configuration

Multi-Rail LNet

- **Basic capability**
  - Multiplex across interfaces, as opposed to striping
  - Need multiple streams to see any benefit
- **Extend peer discover to simplify configuration**
  - Discover a peer's interfaces and multi-rail capability
- **Enable run-time configuration changes**
  - add/remove interfaces, etc., via lnetctl
- **Compatibility with non-multi-rail nodes**
- **Increase resiliency by using alternate paths**
- **Targeted for Lustre 2.10**
- **http://wiki.lustre.org/Multi-Rail_LNet**

```
smw:~ # cfgset update --service cray_lnet --mode interactive hornc-p2
<snip>
Service Configuration Menu (Config Set: hornc-p2, type: cle)

  cray_lnet        [ status: enabled ]  [ validation: valid ]

--------------------------------------------------------------------------------------------------
  Selected     #      Settings               Value/Status (level=basic)
--------------------------------------------------------------------------------------------------
                      ko2iblnd
               1)       peer_credits         63
               2)       concurrent_sends     63

                      local_lnet
               3)       lnet_name            gni4
               4)       ip_wildcard          10.129.*.*

               5)     flat_routes            [ 6 sub-settings unconfigured, select
                                             and enter C to add entries ]

               6)     fgr_routes             [ 5 sub-settings unconfigured, select
                                             and enter C to add entries ]

--------------------------------------------------------------------------------------------------
<snip>
```

```
 Selected     #      Settings               Value/Status (level=basic)
---------------------------------------------------------------------------------------------------
<snip>

               5)     flat_routes             [ 6 sub-settings unconfigured, select
                                              and enter C to add entries ]

         *     6)     fgr_routes              [ 5 sub-settings unconfigured, select
                                              and enter C to add entries ]

---------------------------------------------------------------------------------------------------
**** Select Options ****
  a: all                          n: none                        c: configured
  u: unconfigured                 #: toggle #

**** Actions on Selected (1 settings) ****
  C: configure                    @: show guidance

**** Other Actions ****
  ?: help                         l: switch level                E: toggle enable
  I: toggle inherit               ^^: go to service list         r: refresh
  $: view changelog               Q: save & exit                 x: exit without save


Cray Lustre Networking (LNet) Menu [default: configure - C] $ C
```

- Enter “6”
- Enter “C”

```
**********************************  cray_lnet.settings.fgr_routes  ***********************************

      fgr_routes
        Enter all external LNets which will be reached via Fine-Grained Routing
        (FGR). The information entered for each of these flat LNets will be
        used to set up ip2nets on the routers and routes to reach the external
        LNets through the routers on the clients.

    Configured Values:

        (none)


    Inputs: menu commands (? for help)


|--- Information
|  *    Multiple 'fgr_routes' entries can be added using this menu
|---

cray_lnet.settings.fgr_routes
[<cr>=set 0 entries, +=add an entry, ?=help, @=less] $ +
```

- Enter "+"

```
*************************** cray_lnet.settings.fgr_routes.data.dest_name ***************************

    fgr_routes
        Enter all external LNets which will be reached via Fine-Grained
        Routing (FGR). The information entered for each of these flat LNets
        will be used to set up ip2nets on the routers and routes to reach the
        external LNets through the routers on the clients.

      dest_name -- Destination name
         Enter the name of the destination. This is not functionally
         important. A good convention would be to use the name of the
         destination. For example, if the destination is the husk2 external
         file system, enter 'husk2'.

       Default:          Current:
          (none)             not configured yet

    Value: string, blank values not allowed
           level=basic, state=unset

    Inputs: <string>  -- OR --  menu commands (? for help)

cray_lnet.settings.fgr_routes.data.dest_name
[<cr>=set '', <new value>, ?=help, @=less] $ snx8675309
```

- Enter “snx8675309”

```
********************** cray_lnet.settings.fgr_routes.data.snx8675309.routers ***********************

    fgr_routes (current key: snx8675309)
         Enter all external LNets which will be reached via Fine-Grained
         Routing (FGR). The information entered for each of these flat LNets
         will be used to set up ip2nets on the routers and routes to reach the
         external LNets through the routers on the clients.

      routers -- LNet router nodes
          Enter a list of router cnames which will be used to route from the
          source LNet to the destination LNet. If the router nodes are managed
          externally (e.g. you are currently configuring LNet on servers) this
          can be left empty.

        Default:           Current:
            (none)         (none)

    Value: list, blank values allowed, regex=^c(\d+)-(\d+)c([0-2])s(\d[0-5]?)n([0-3])$|^(\d{1,3})(\.\d{1,3}){3}$
           level=basic, state=unset

    Inputs: menu commands (? for help)


cray_lnet.settings.fgr_routes.data.snx8675309.routers
[<cr>=set 0 entries, +=add an entry, ?=help, @=less] $ +
```

- Enter “+”

```
Add routers (Ctrl-d to exit) $ c0-0c0s2n1
Add routers (Ctrl-d to exit) $ c0-0c0s2n2
Add routers (Ctrl-d to exit) $ c0-0c0s3n1
Add routers (Ctrl-d to exit) $ c0-0c0s3n2
Add routers (Ctrl-d to exit) $ c0-0c1s2n1
Add routers (Ctrl-d to exit) $ c0-0c1s2n2
Add routers (Ctrl-d to exit) $
```

```
********************  cray_lnet.settings.fgr_routes.data.snx8675309.ip2nets_file  ********************

    fgr_routes (current key: snx8675309)
         Enter all external LNets which will be reached via Fine-Grained
         Routing (FGR). The information entered for each of these flat LNets
         will be used to set up ip2nets on the routers and routes to reach the
         external LNets through the routers on the clients.

      ip2nets_file -- FGR ip2nets file
          Enter the name of the ip2nets file for this FGR config.  The file
          must be placed in the config_set at
          smw:/var/opt/cray/imps/config/sets/<config_set>/files/roles/lnet/.
          This file must be generated using an external tool, such as clcvt.

        Default:                Current:
            (none)                  not configured yet

    Value: string, blank values not allowed, regex=^[!-.0-~]+$
           level=basic, state=unset

    Inputs: <string>  -- OR --  menu commands (? for help)

cray_lnet.settings.fgr_routes.data.snx8675309.ip2nets_file
[<cr>=set '', <new value>, ?=help, @=less] $ ip2nets.dat
```

```
******************** cray_lnet.settings.fgr_routes.data.snx8675309.routes_file *********************

    fgr_routes (current key: snx8675309)
         Enter all external LNets which will be reached via Fine-Grained
         Routing (FGR). The information entered for each of these flat LNets
         will be used to set up ip2nets on the routers and routes to reach the
         external LNets through the routers on the clients.

      routes_file -- FGR routes file
          Enter the name of the routes file for this FGR config.  The file must
          be placed in the config_set at
          smw:/var/opt/cray/imps/config/sets/<config_set>/files/roles/lnet/.
          This file must be generated using an external tool, such as clcvt.

        Default:            Current:
            (none)              not configured yet

    Value: string, blank values not allowed, regex=^[!-.0-~]+$
           level=basic, state=unset

    Inputs: <string>  -- OR --  menu commands (? for help)

cray_lnet.settings.fgr_routes.data.snx8675309.routes_file
[<cr>=set '', <new value>, ?=help, @=less] $ routes.dat
```

```
*************** cray_lnet.settings.fgr_routes.data.snx8675309.ko2iblnd_peer_credits  ****************

   fgr_routes (current key: snx8675309)
       Enter all external LNets which will be reached via Fine-Grained
       Routing (FGR). The information entered for each of these flat LNets
       will be used to set up ip2nets on the routers and routes to reach the
       external LNets through the routers on the clients.

     ko2iblnd_peer_credits -- ko2iblnd peer_credits
        The number of concurrent sends allowed to a single peer. Cray
        recommends setting this to 126. peer_credits must be consistent
        across all peers on the IB network. This means it must be the same on
        the routers and the Lustre servers. If there is a mismatch, the file
        system will be unmountable. This value is specific to the routers
        specified in this FGR config, and it will override the general
        ko2iblnd peer_credits setting specified earlier.

      Default:             Current:
          126                  not configured yet

   Value: integer, blank values allowed, regex=^[1-9]\d*$
         level=basic, state=unset

   Inputs: <integer>  -- OR --  menu commands (? for help)

cray_lnet.settings.fgr_routes.data.snx8675309.ko2iblnd_peer_credits
[<cr>=set '126', <new value>, ?=help, @=less] $ 63
```

```
************* cray_lnet.settings.fgr_routes.data.snx8675309.ko2iblnd_concurrent_sends  *************

  fgr_routes (current key: snx8675309)
     Enter all external LNets which will be reached via Fine-Grained
     Routing (FGR). The information entered for each of these flat LNets
     will be used to set up ip2nets on the routers and routes to reach the
     external LNets through the routers on the clients.

   ko2iblnd_concurrent_sends -- ko2iblnd concurrent_sends
      Determines send work-queue sizing. If this option is omitted, the
      default is calculated based on peer_credits and map_on_demand. Cray
      recommends setting this to 63. concurrent_sends must be consistent
      across all peers on the IB network. This means it must be the same on
      the routers and the Lustre servers. If there is a mismatch, the file
      system will be unmountable. This value is specific to the routers
      specified in this FGR config, and it will override the general
      ko2iblnd concurrent_sends setting specified earlier.

    Default:         Current:
      63                 not configured yet

  Value: integer, blank values allowed, regex=^[1-9]\d*$
         level=basic, state=unset

  Inputs: <integer>  -- OR --  menu commands (? for help)

cray_lnet.settings.fgr_routes.data.snx8675309.ko2iblnd_concurrent_sends
[<cr>=set '63', <new value>, ?=help, @=less] $
```

```
********************************** cray_lnet.settings.fgr_routes ***********************************

      fgr_routes
        Enter all external LNets which will be reached via Fine-Grained Routing
        (FGR). The information entered for each of these flat LNets will be
        used to set up ip2nets on the routers and routes to reach the external
        LNets through the routers on the clients.

    Configured Values:

        1) 'snx8675309'
            a) routers:
                   c0-0c0s2n1
                   c0-0c0s2n2
                   c0-0c0s3n1
                   c0-0c0s3n2
                   c0-0c1s2n1
                   c0-0c1s2n2
            b) ip2nets_file: ip2nets.dat
            c) routes_file: routes.dat
            d) ko2iblnd_peer_credits: 63
            e) ko2iblnd_concurrent_sends: 63



    Inputs: menu commands (? for help)


|--- Information
|  *    Multiple 'fgr_routes' entries can be added using this menu
|---

cray_lnet.settings.fgr_routes
[<cr>=set 1 entries, +=add an entry, ?=help, @=less] $
```