Hewlett Packard
Enterprise

KFILND & UDSP

Chris Horn, Lustre Software Engineer

October, 2023

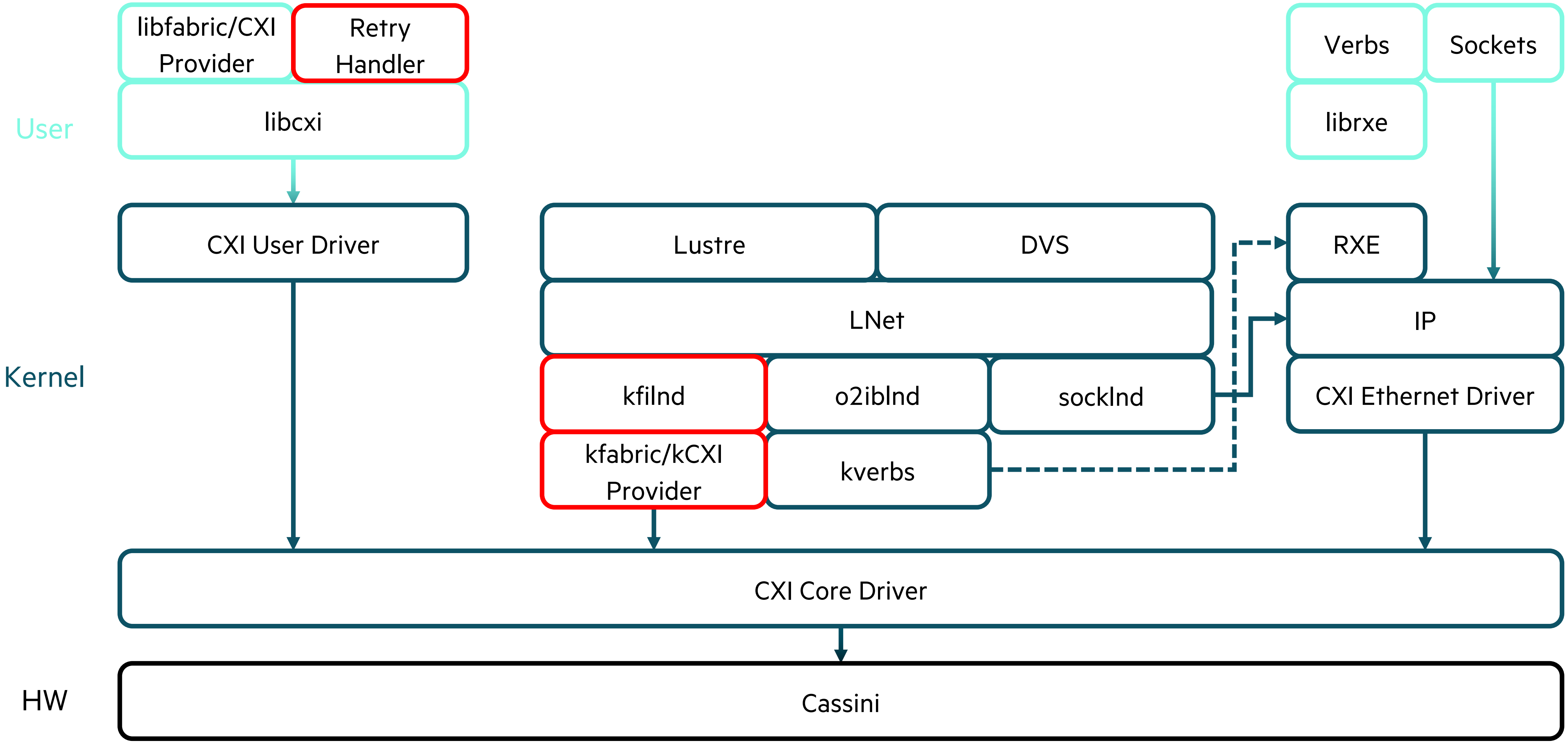
OUTLINE

- kfilnd overview
- UDSP
  - Overview
  - Local Net Selection Example
  - UDSP + LNet health
  - Other rule types
  - YAML config
- socklnd + kfilnd + UDSP

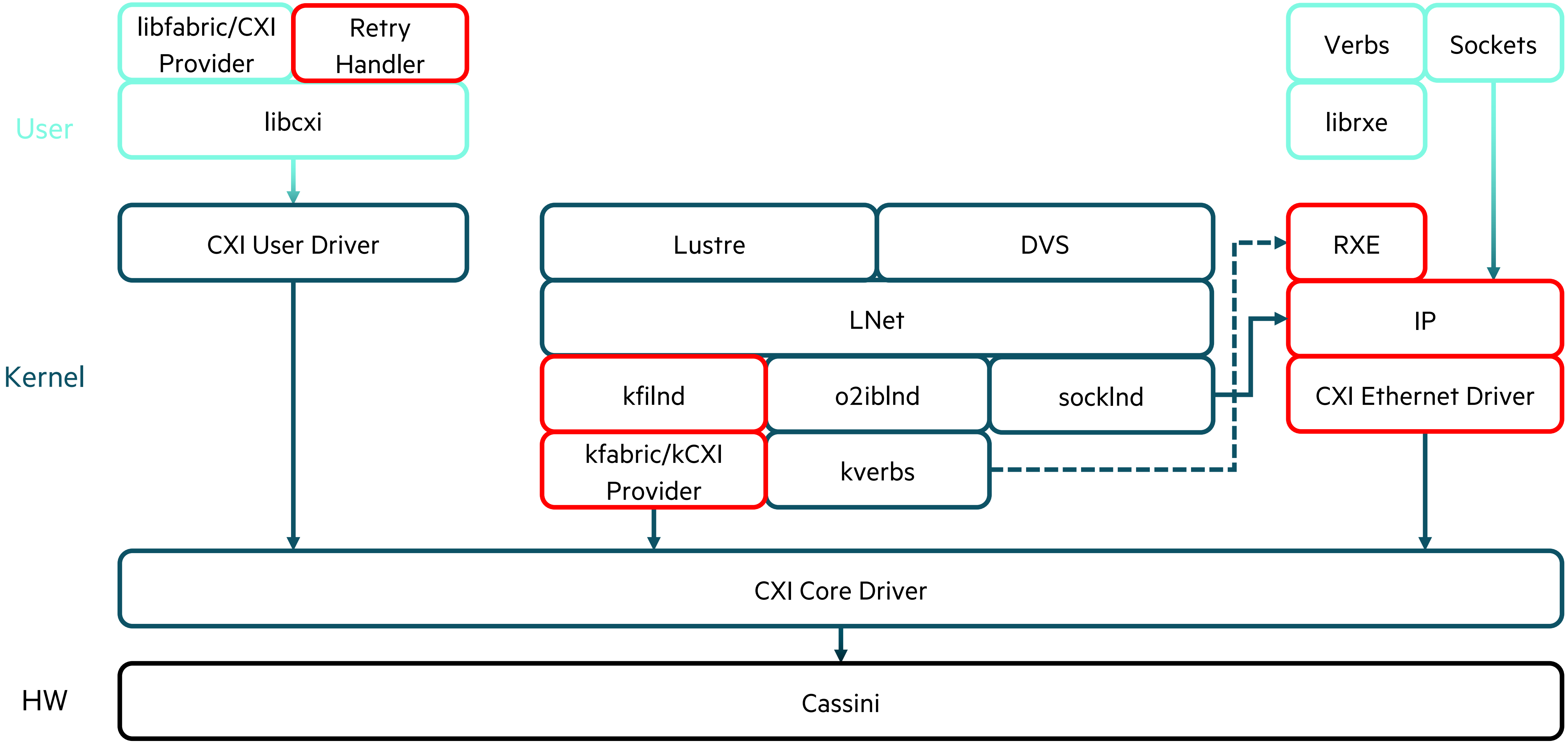
KFILND OVERVIEW

- kfilnd == kfabric Lustre network driver
  - Uses numeric LNet NIDs: 1@kfi, 2@kfi, …
  - NID number == Destination Fabric Address (DFA)
    - Reflects group, switch and port numbers
  - Implements LND api (lnd_send, lnd_recv, etc.) using kfabric
  - Lustre 2.16
- kfabric
  - Network-agnostic API used for RDMA in the kernel
  - Envisioned as common mid layer for multiple ULPs
  - Providers map kfabric API to lower-level network software/hardware
    - kfi_cxi is the only provider
- Cassini
  - Ethernet L1/L2
  - Portals 4 RDMA based extensions
  - PCIe Gen. 4, 200 Gbps, Virtualization with SR-IOV
  - CXI – Cray eXascale Interface

SOFTWARE STACK



SOFTWARE STACK



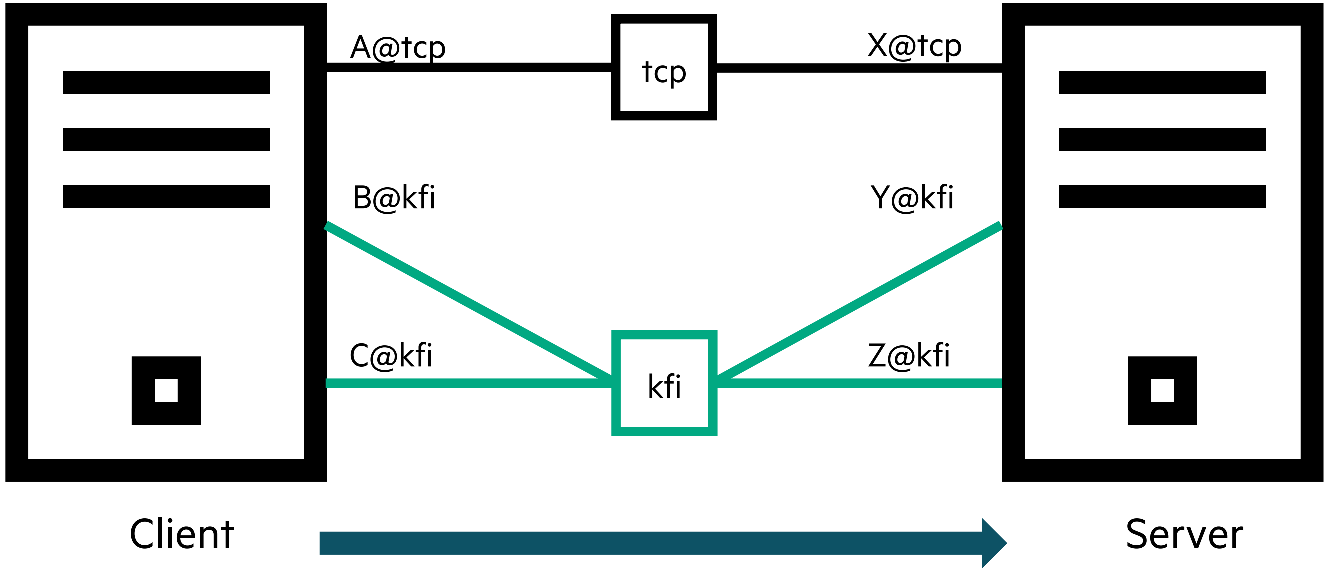
LNET USER DEFINED SELECTION POLICY (UDSP)

- New LNet feature in Lustre 2.15
  - Shout out Amir Shehata, Sonia Sharma and Serguei Smirnov
- Motivation:
  - Multi-Rail peers may have multiple paths
  - Some paths may be better than others
- lnetctl CLI
  - lnetctl udsp add
  - lnetctl udsp del
  - lnetctl udsp show
  - YAML config
- Rule types:
  - Local net/NID selection priority
  - Peer NID selection priority
  - NID-Pair selection
  - Peer-Router selection

LNET USER DEFINED SELECTION POLICY (UDSP)

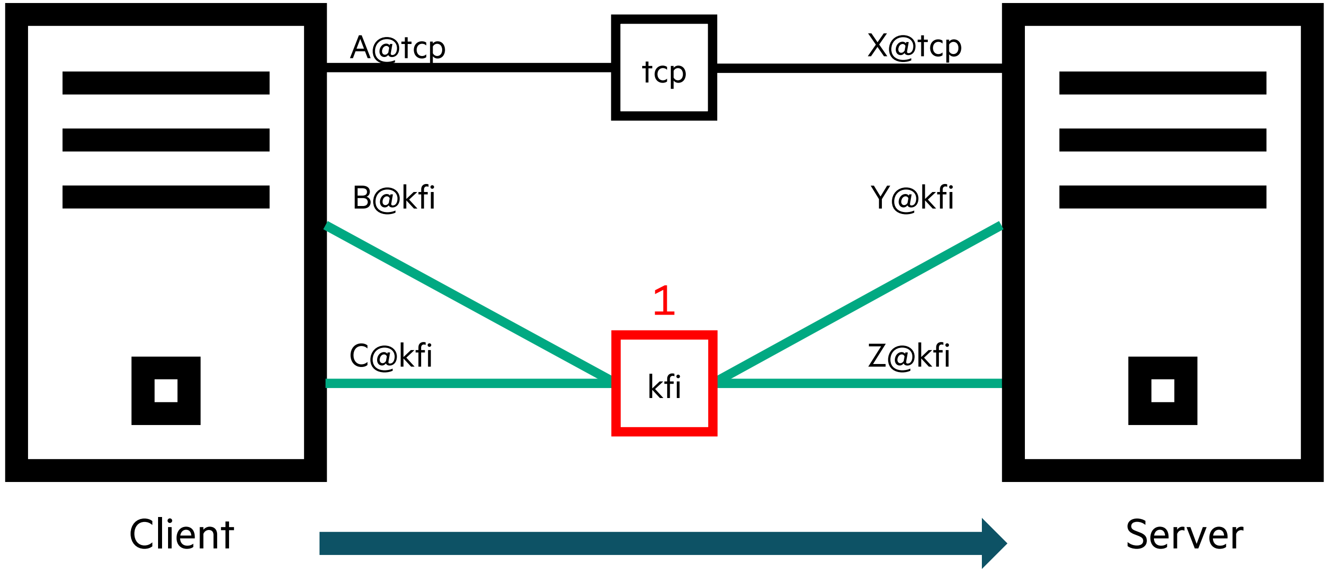
- New in Lustre 2.15
- Motivation:
  - Multi-Rail peers may have multiple paths
  - Some paths are better than others
- lnetctl CLI
  - lnetctl udsp add
  - lnetctl udsp del
  - lnetctl udsp show
  - YAML config
- Rule types:
  - Local net/NID selection priority
  - Peer NID selection priority
  - NID-Pair selection
  - Peer-Router selection

LNET PATH SELECTION



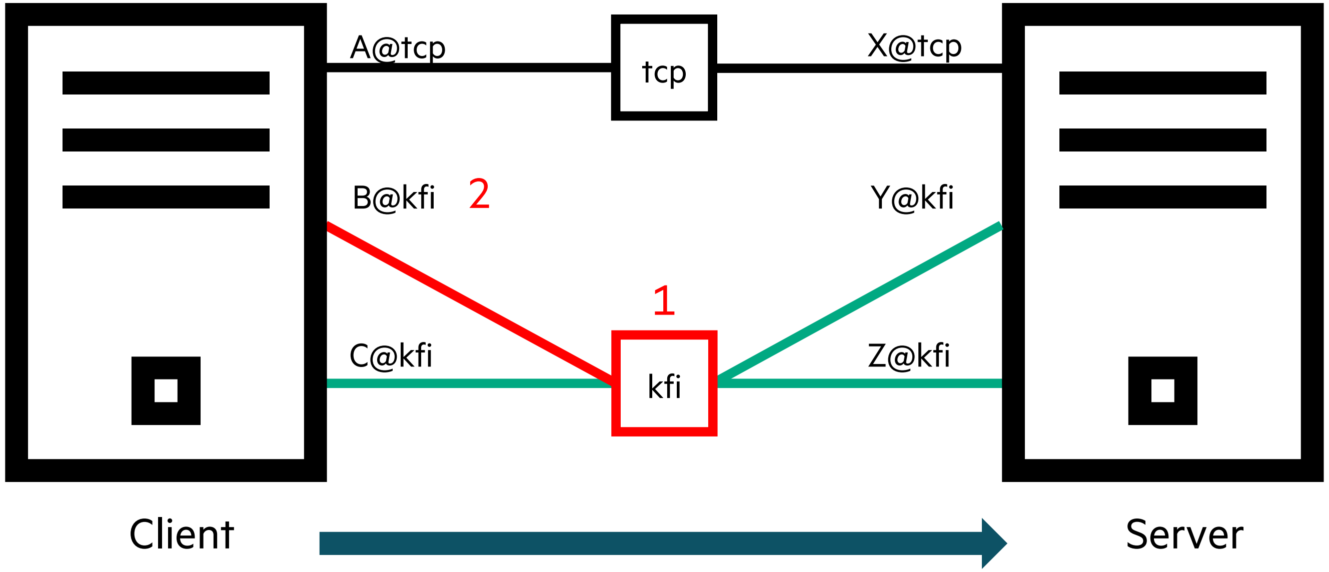
- Local LNet Path Selection (PUT or GET)

LNET PATH SELECTION



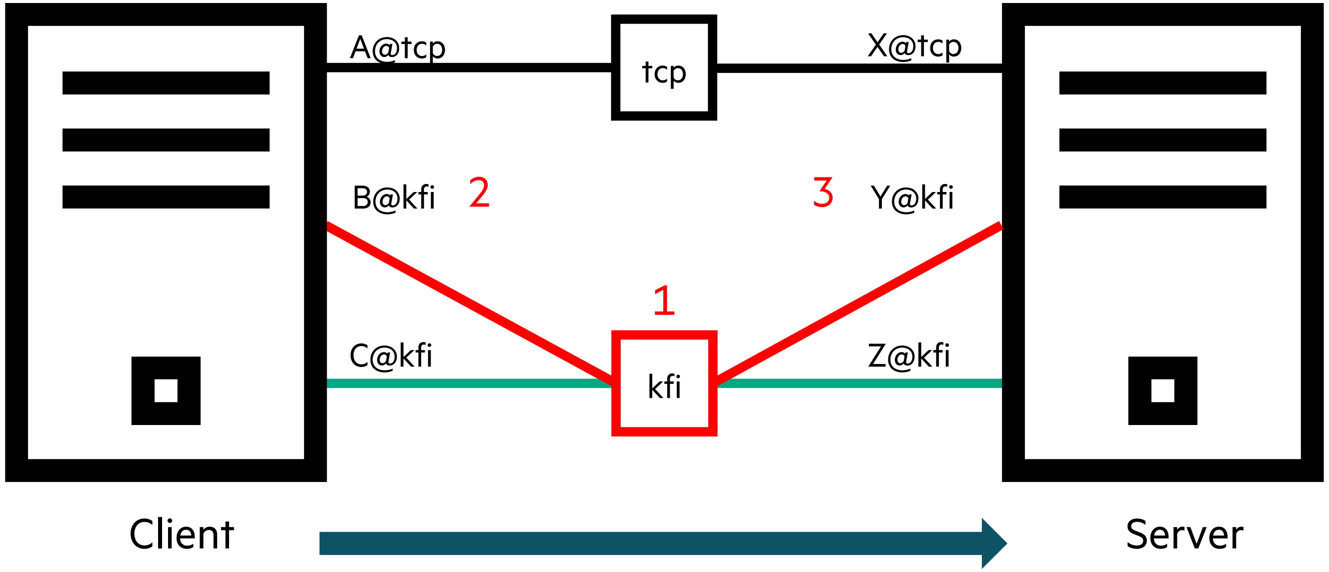
- Local LNet Path Selection (PUT or GET)
  1. Select local network

LNET PATH SELECTION



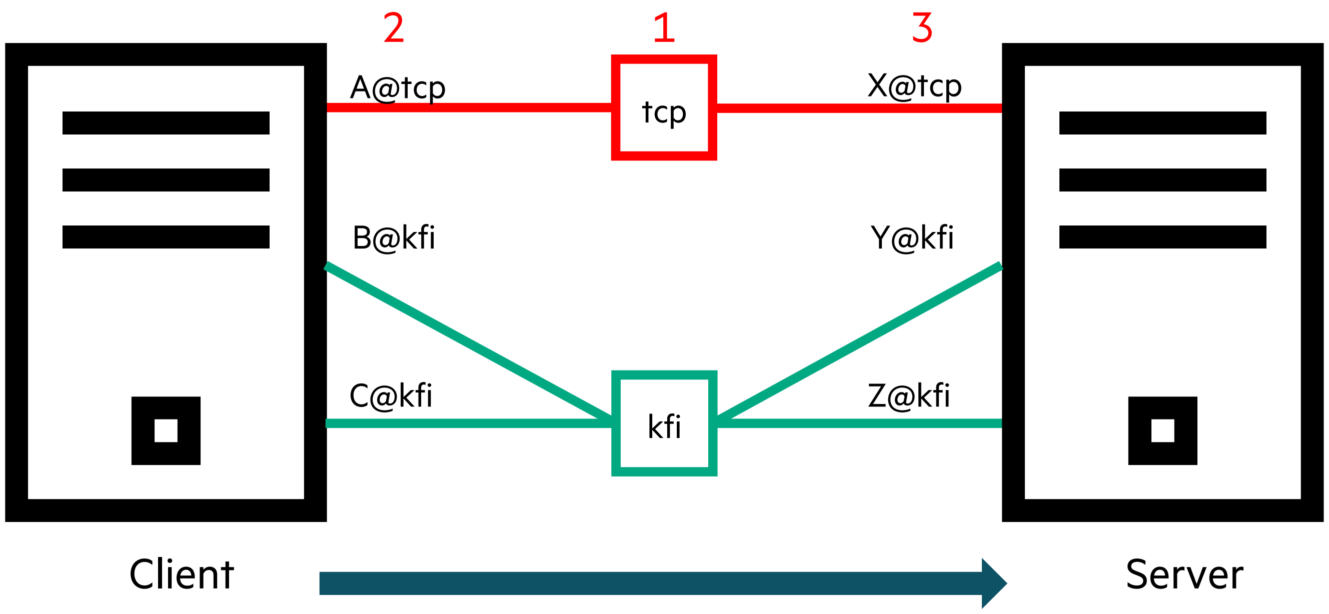
- Local LNet Path Selection (PUT or GET)
  1. Select local network
  2. Select source NID

LNET PATH SELECTION



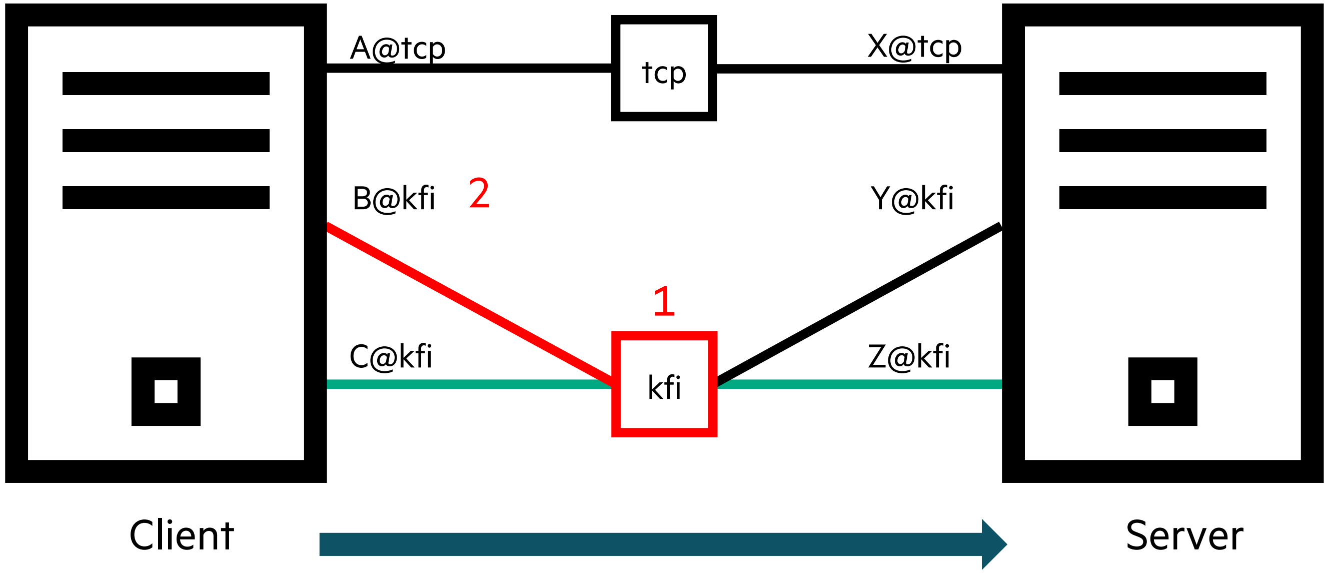
- Local LNet Path Selection (PUT or GET)
  1. Select local network
  2. Select source NID
  3. Select destination NID
     - On same network as above
  - At each step consider:
    - Health
    - Priority
    - Credits
    - etc.
    - Round robin when all else equal

LNET PATH SELECTION



- Local LNet Path Selection (PUT or GET)
  1. Select local network
  2. Select source NID
  3. Select destination NID
     - On same network as above
  - Round robin when all else equal
  - tcp slow relative to kfi

LOCAL NET/NID SELECTION PRIORITY



- Local LNet Path Selection (PUT or GET)
  1. Select local network
  2. Select source NID
  3. Select destination NID
     - On same network as above
  - Round robin when all else equal
- Local Net selection rules affect (1)
  - Prefer kfi over tcp
  - lnetctl udsp add --src kfi --priority 0
- Local NID selection rules affect (2)
  - Prefer B@kfi over C@kfi
  - lnetctl udsp add --src B@kfi --priority 0

EXAMPLE

```
n00 $ cat ~/setup.sh
#!/bin/bash

# <Load Modules>

pdsh -w n0[0-1] lnetctl lnet configure
pdsh -w n0[0-1] lnetctl net add --net tcp --if eth0
pdsh -w n0[0-1] lnetctl net add --net kfi --if cxi0
pdsh -w n0[0-1] lnetctl net add --net kfi --if cxi1
pdsh -w n0[0-1] insmod /home/hornc/lustre-wc-rel/lnet/selftest/lnet_selftest.ko

lnetctl net show -v 4 | grep -P 'nid|priority’
n00 $
```

EXAMPLE - DEFAULT BEHAVIOR

```
n00 $ bash setup.sh
        - nid: 10.214.131.25@tcp
              net priority: -1
              nid priority: -1
        - nid: 17@kfi
              net priority: -1
              nid priority: -1
        - nid: 5@kfi
              net priority: -1
              nid priority: -1
n00 $ lst.sh -t 10.214.131.25@tcp -f 10.214.129.92@tcp -m read -g servers
...
[LNet Bandwidth of servers]
[R] Avg: 0.04     MB/s  Min: 0.04     MB/s  Max: 0.04     MB/s
[W] Avg: 235.48   MB/s  Min: 235.48   MB/s  Max: 235.48   MB/s
...
n00 $ lnetctl net show -v | grep -P 'nid|send_count|recv_count'
...
        - nid: 10.214.131.25@tcp
              send_count: 3430
              recv_count: 1719
        - nid: 17@kfi
              send_count: 1715
              recv_count: 859
        - nid: 5@kfi
              send_count: 1711
              recv_count: 857
```

- Default priorities
- Awful performance due to slow tcp
- Traffic split ~ evenly
  - 5149 on tcp
  - 5142 on kfi

EXAMPLE - LOCAL NET PRIORITY

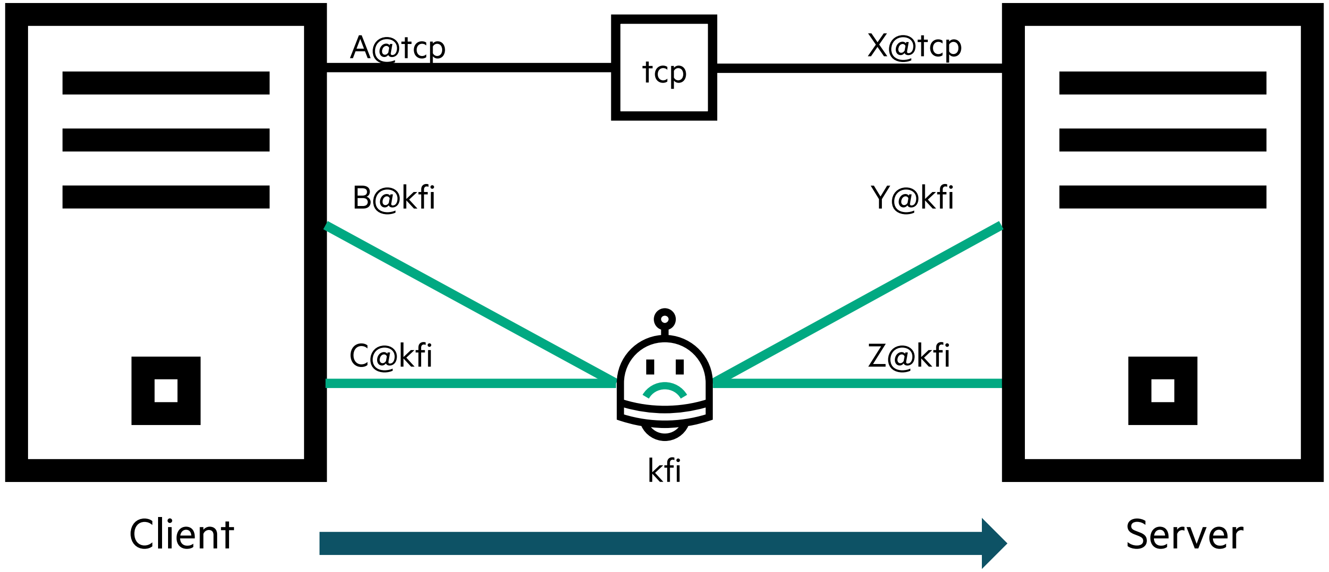
```
n00 $ cat ~/setup.sh
#!/bin/bash

# <Load Modules>

pdsh -w n0[0-1] lnetctl lnet configure
pdsh -w n0[0-1] lnetctl net add --net tcp --if eth0
pdsh -w n0[0-1] lnetctl net add --net kfi --if cxi0
pdsh -w n0[0-1] lnetctl net add --net kfi --if cxi1
pdsh -w n0[0-1] insmod /home/hornc/lustre-wc-rel/lnet/selftest/lnet_selftest.ko

pdsh -w n0[0-1] lnetctl udsp add --src kfi --priority 0
lnetctl net show -v 4 | grep -P 'nid|priority’
n00 $
```

EXAMPLE - LOCAL NET PRIORITY

```
n00 $ bash setup.sh
       - nid: 10.214.131.25@tcp
            net priority: -1
            nid priority: -1
       - nid: 17@kfi
            net priority: 0
            nid priority: -1
       - nid: 5@kfi
            net priority: 0
            nid priority: -1
n00 $ lst.sh -t 10.214.131.25@tcp -f 10.214.129.92@tcp -m read -g servers
...
[LNet Bandwidth of servers]
[R] Avg: 7.23     MB/s  Min: 7.23     MB/s  Max: 7.23     MB/s
[W] Avg: 47413.76 MB/s  Min: 47413.76 MB/s  Max: 47413.76 MB/s
...
n00 $ lnetctl net show -v | grep -P 'nid|send_count|recv_count’
...
       - nid: 10.214.131.25@tcp
            send_count: 2
            recv_count: 2
       - nid: 17@kfi
            send_count: 679592
            recv_count: 339798
       - nid: 5@kfi
            send_count: 679737
            recv_count: 339870
```

- Priority assignments
- Performance greatly improved
- Traffic traverses fast HSN links

UDSP AND HEALTH



- What happens when our preferred network fails?
- LNet always considers every path

EXAMPLE - UDSP AND HEALTH

```
n00 $ cat ~/setup.sh
#!/bin/bash

# <Load Modules>

pdsh -w n0[0-1] lnetctl lnet configure
pdsh -w n0[0-1] lnetctl net add --net tcp --if eth0
pdsh -w n0[0-1] lnetctl net add --net kfi --if cxi0
pdsh -w n0[0-1] lnetctl net add --net kfi --if cxi1
pdsh -w n0[0-1] insmod /home/hornc/lustre-wc-rel/lnet/selftest/lnet_selftest.ko

pdsh -w n0[0-1] lnetctl udsp add --src kfi --priority 0

# Simulate failure of kfi network
lnetctl set health_sensitivity 0
lnetctl net set --health 0 --nid 17@kfi
lnetctl net set --health 0 --nid 5@kfi
lnetctl net show -v 4 | grep -P 'nid|priority|health value’
n00 $
```

EXAMPLE - UDSP AND HEALTH

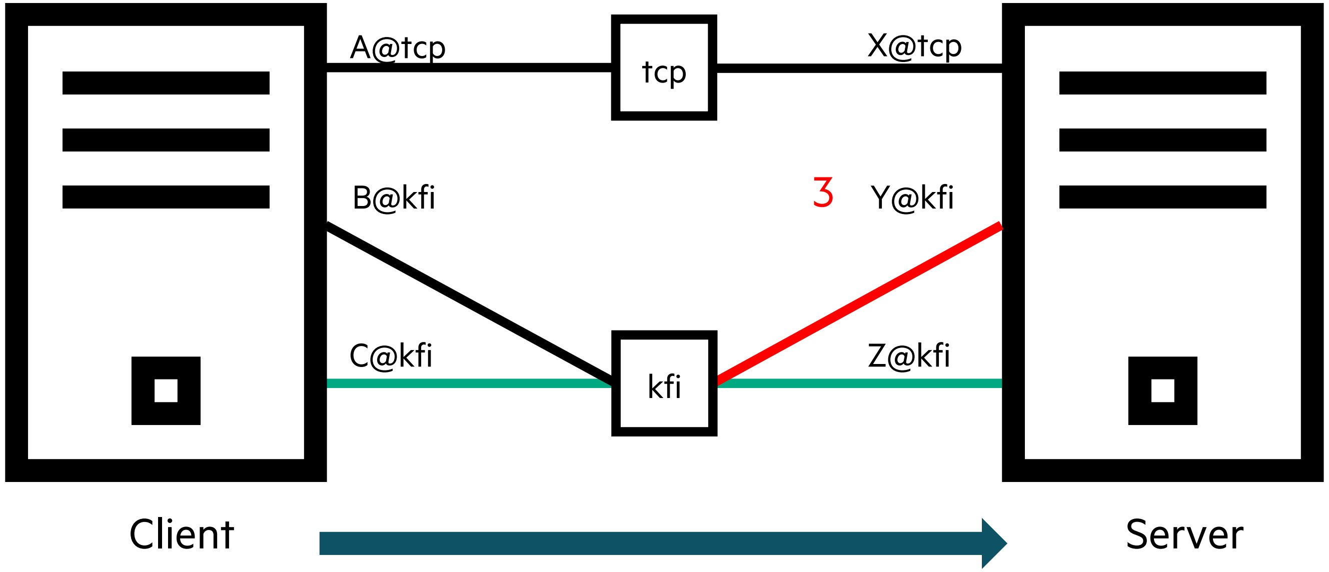
```
n00 $ bash setup.sh
        - nid: 10.214.131.25@tcp
              net priority: -1
              nid priority: -1
              health value: 1000
        - nid: 17@kfi
              net priority: 0
              nid priority: -1
              health value: 0
        - nid: 5@kfi
              net priority: 0
              nid priority: -1
              health value: 0
n00 $ lst.sh -t 10.214.131.25@tcp -f 10.214.129.92@tcp -m read -g servers
...
[LNet Bandwidth of servers]
[R] Avg: 0.02     MB/s  Min: 0.02     MB/s  Max: 0.02     MB/s
[W] Avg: 117.70   MB/s  Min: 117.70   MB/s  Max: 117.70   MB/s
...
n00 $ lnetctl net show -v | grep -P 'nid|send_count|recv_count’
        - nid: 10.214.131.25@tcp
              send_count: 5058
              recv_count: 6747
        - nid: 17@kfi
              send_count: 0
              recv_count: 0
        - nid: 5@kfi
              send_count: 0
              recv_count: 0
```

- LNet always selects the healthiest networks and interfaces

LNET USER DEFINED SELECTION POLICY (UDSP)

- New in Lustre 2.15
- Motivation:
  - Multi-Rail peers may have multiple paths
  - Some paths are better than others
- lnetctl CLI
  - lnetctl udsp add
  - lnetctl udsp del
  - lnetctl udsp show
  - YAML config
- Rule types:
  - Local net/NID selection priority
  - Peer NID selection priority
  - NID-Pair selection
  - Peer-Router selection

PEER NID SELECTION PRIORITY

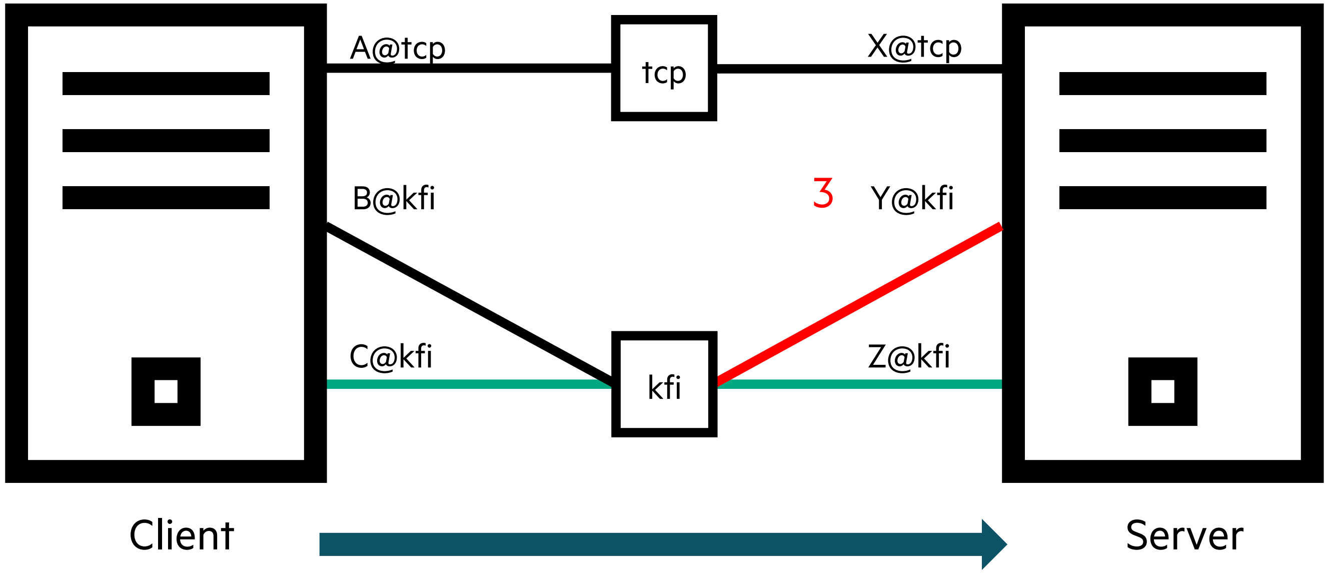


- Local LNet Path Selection (PUT or GET)
  1. Select local network
  2. Select source NID
  3. Select destination NID
     - On same network as above
  - Round robin when all else equal
- Peer NID selection rules affect (3)
  - Prefer Y@kfi over Z@kfi
  - lnetctl udsp add --dst Y@kfi --priority 0

LNET USER DEFINED SELECTION POLICY (UDSP)

- New in Lustre 2.15
- Motivation:
  - Multi-Rail peers may have multiple paths
  - Some paths are better than others
- lnetctl CLI
  - lnetctl udsp add
  - lnetctl udsp del
  - lnetctl udsp show
  - YAML config
- Rule types:
  - Local net/NID selection priority
  - Peer NID selection priority
  - NID-Pair selection
  - Peer-Router selection

NID-PAIR SELECTION

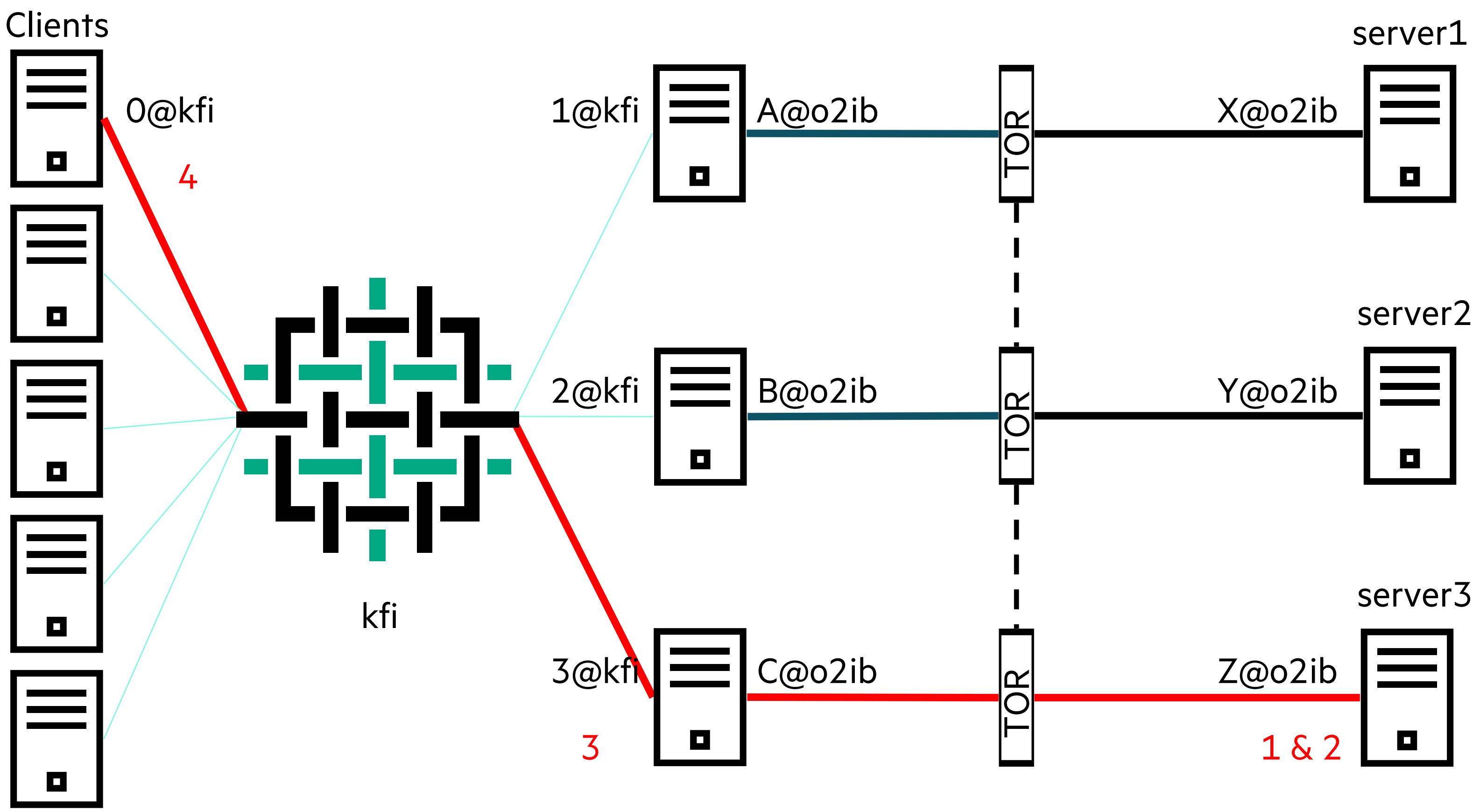


- Local LNet Path Selection (PUT or GET)
  1. Select local network
  2. Select source NID
  3. Select destination NID
     - On same network as above
  - Round robin when all else equal
- NID-Pair selection rules affect (3)
  - Prefer Y@kfi when using B@kfi
  - lnetctl udsp add --src B@kfi --dst Y@kfi
  - Prefer Z@kfi when using C@kfi
  - lnetctl udsp add --src C@kfi --dst Z@kfi

LNET USER DEFINED SELECTION POLICY (UDSP)

- New in Lustre 2.15
- Motivation:
  - Multi-Rail peers may have multiple paths
  - Some paths are better than others
- lnetctl CLI
  - lnetctl udsp add
  - lnetctl udsp del
  - lnetctl udsp show
  - YAML config
- Rule types:
  - Local net/NID selection priority
  - Peer NID selection priority
  - NID-Pair priority
  - Peer-Router priority

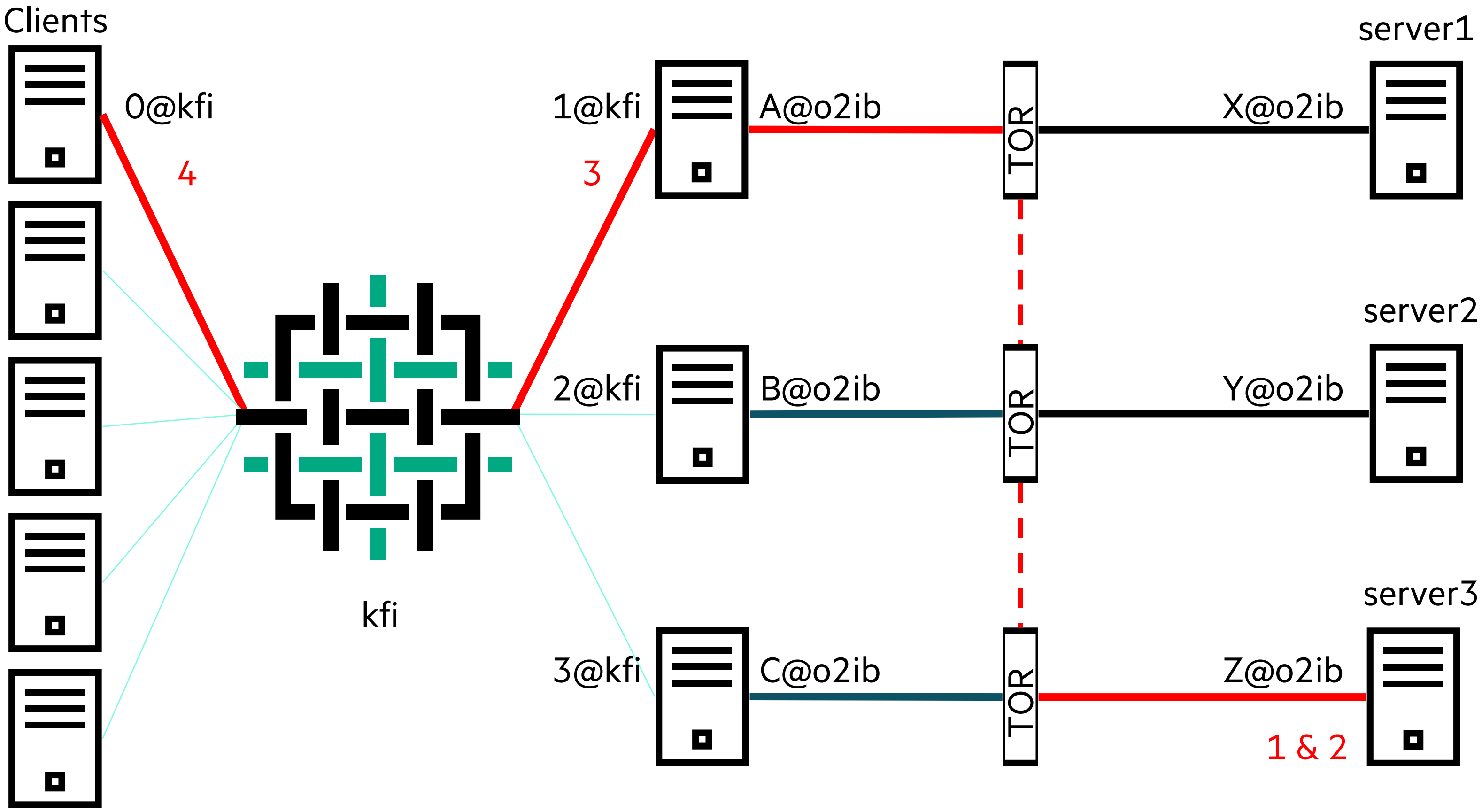
OPTIMAL PATH



Remote Net	Gateway
o2ib	[1-3]@kfi

- Routed LNet Path Selection
  1. Select destination network (o2ib)
  2. Select destination NID (Z@o2ib)
  3. Select router NID (3@kfi)
  4. Select local NID (0@kfi)
  - Round robin when all else equal

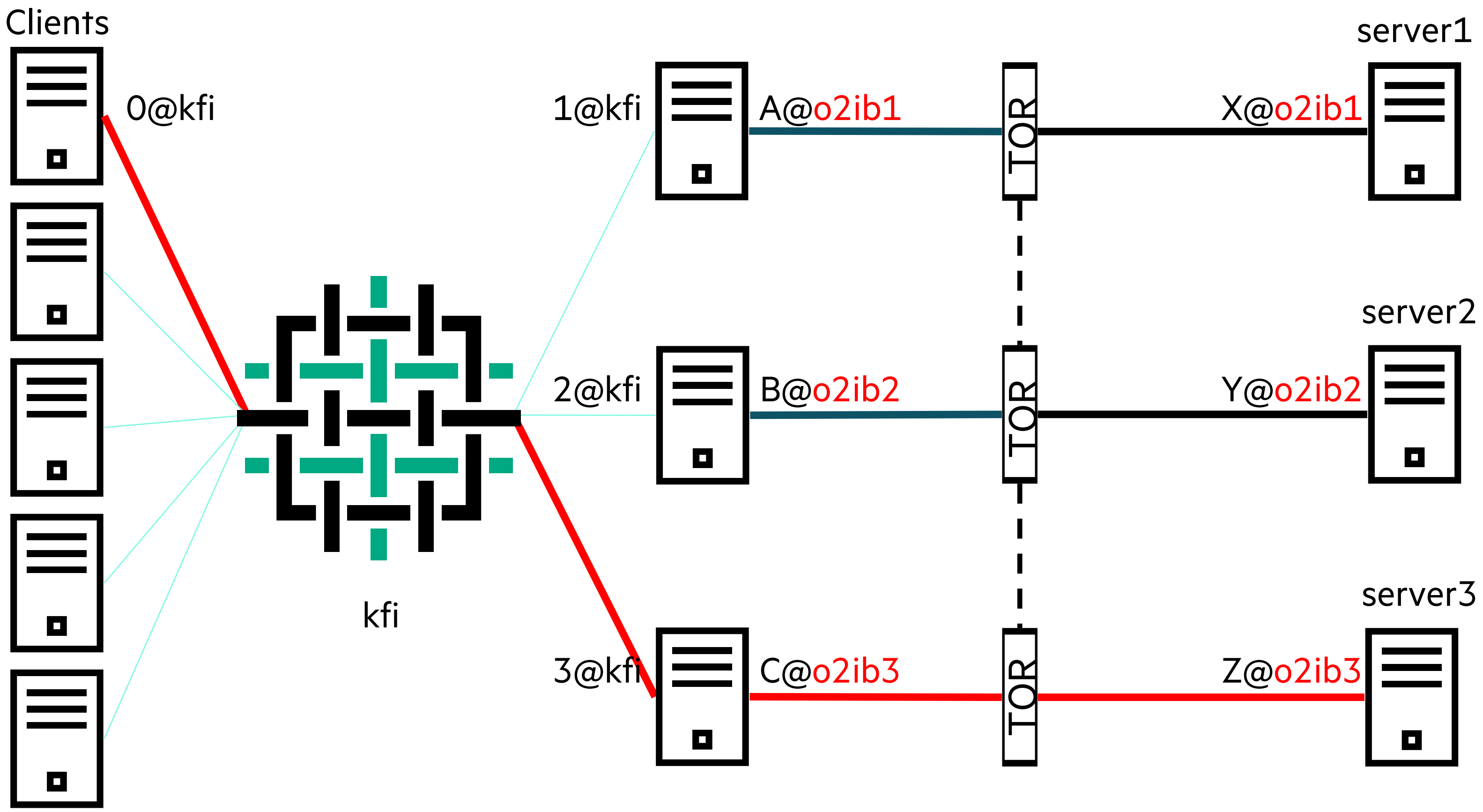
WORST PATH



Remote Net	Gateway
o2ib	[1-3]@kfi

- Routed LNet Path Selection
  1. Select destination network (o2ib)
  2. Select destination NID (Z@o2ib)
  3. Select router NID (1@kfi)
  4. Select local NID (0@kfi)
  - Round robin when all else equal

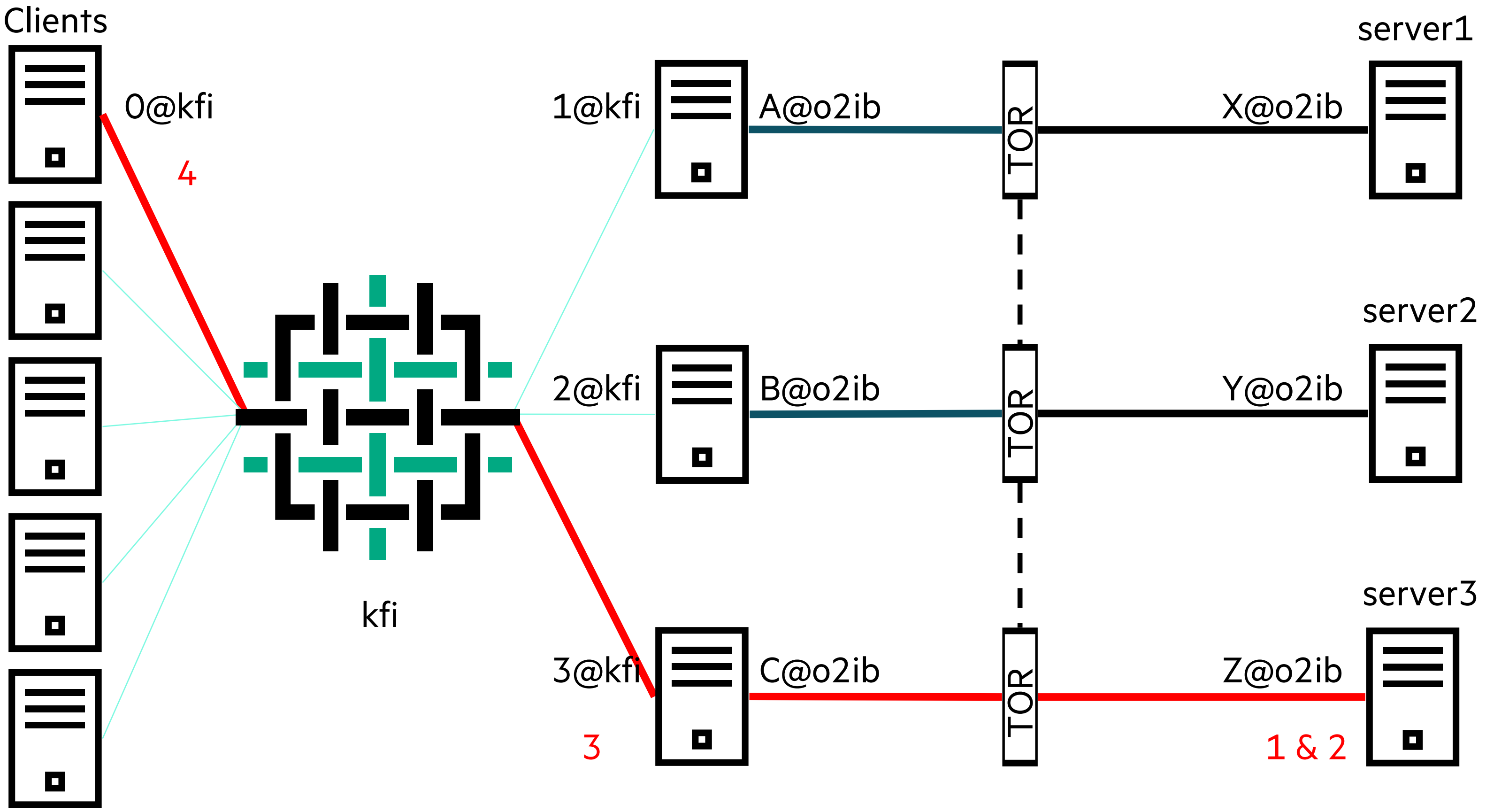
FINE GRAINED ROUTING (FGR)



Remote Net	Gateway
o2ib1	1@kfi
o2ib2	2@kfi
o2ib3	3@kfi

- Routed LNet Path Selection
  1. Select destination network
  2. Select destination NID
  3. Select router NID
  4. Select local NID
- FGR defines optimal path via route table
- Reduces total number of available paths

UDSP CAN DEFINE OPTIMAL PATHS



Remote Net	Gateway
o2ib	[1-3]@kfi

```
lnetctl udsp add --dst X@o2ib --rte 1@kfi
lnetctl udsp add --dst Y@o2ib --rte 2@kfi
lnetctl udsp add --dst Z@o2ib --rte 3@kfi
```

- Routed LNet Path Selection
  1. Select destination network
  2. Select destination NID
  3. Select router NID
  4. Select local NID
- Peer Router rules define optimal paths by influencing (3)
  - Preferred routers added to list on peer NI
- Other routers can be used as failback

UDSP YAML CONFIG

- Issue 1:
  - lnetctl export --backup output cannot be used for import
  - Workaround - Manually remove “NA” lines

```
$ cat /etc/lnet.conf.good
udsp:
    - idx: 0
      src: kfi
      action:
          priority: 0
```

```
# lnetctl udsp add --src kfi --priority 0
# lnetctl export --backup
udsp:
    - idx: 0
      src: kfi
      dst: NA <<<<< “NA” is not understood by import
      rte: NA
      action:
          priority: 0
#
```

UDSP YAML CONFIG

- Issue 2:
  - Different rule types cannot be combined in obvious way
  - Workaround - Separate every rule with “udsp:”

```
$ cat /etc/lnet.conf.good
udsp:
    - idx: 0
      src: kfi
      action:
          priority: 0
udsp:
    - idx: 1
      dst: 867@kfi
      action:
          priority: 0
```

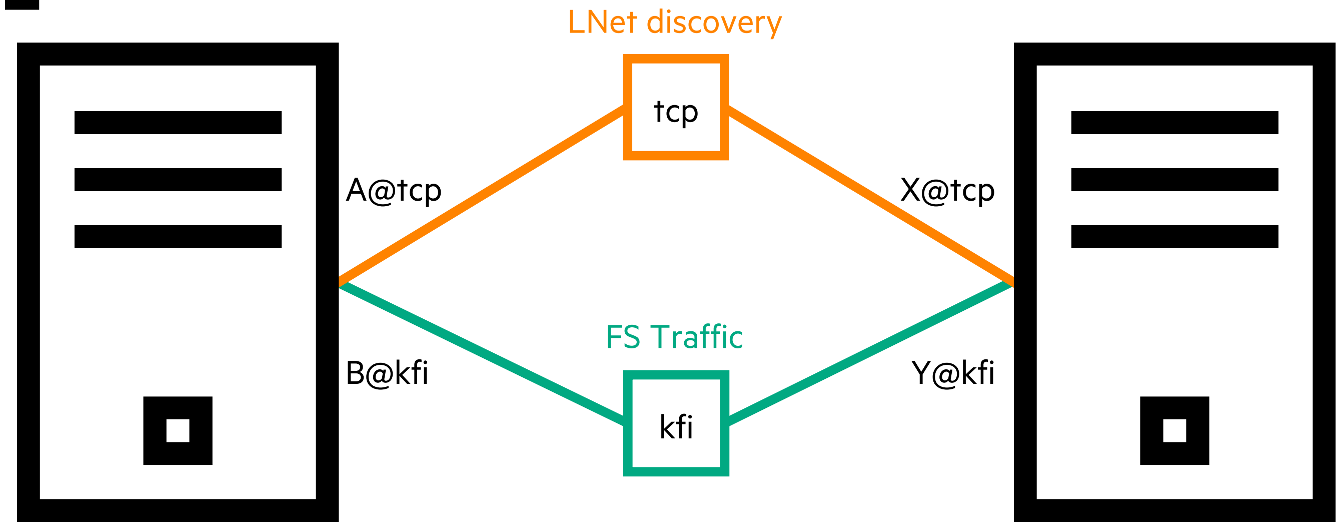
```
$ cat /etc/lnet.conf.bad
udsp:
    - idx: 0
      src: kfi
      action:
          priority: 0
    - idx: 1
      dst: 867@kfi
      action:
          priority: 0
$ lnetctl import /etc/lnet.conf.bad
$ lnetctl udsp show
udsp:
    - idx: 0
      src: kfi
      dst: 897@kfi  <<<< Malformed rule
      rte: NA
      action:
          priority: 0
    - idx: 1
      src: NA
      dst: 897@kfi
      rte: NA
      action:
          priority: 0
```

KFILND ADMINISTRATIVE CHALLENGE

- kfilnd NID number == Destination Fabric Address (DFA)
- DFAs change with re-cabling (including cable swap)
  - Can’t swap while LNet is running
- New DFA == new LNet NID
  - On a Lustre server, new NIDs require writeconf (or lctl replace_nids)
  - MGS NID changes -> All clients must update /etc/fstab
  - Router gets a new NID it invalidates routing table on other peers

SOCKLND + KFILND + MULTI-RAIL

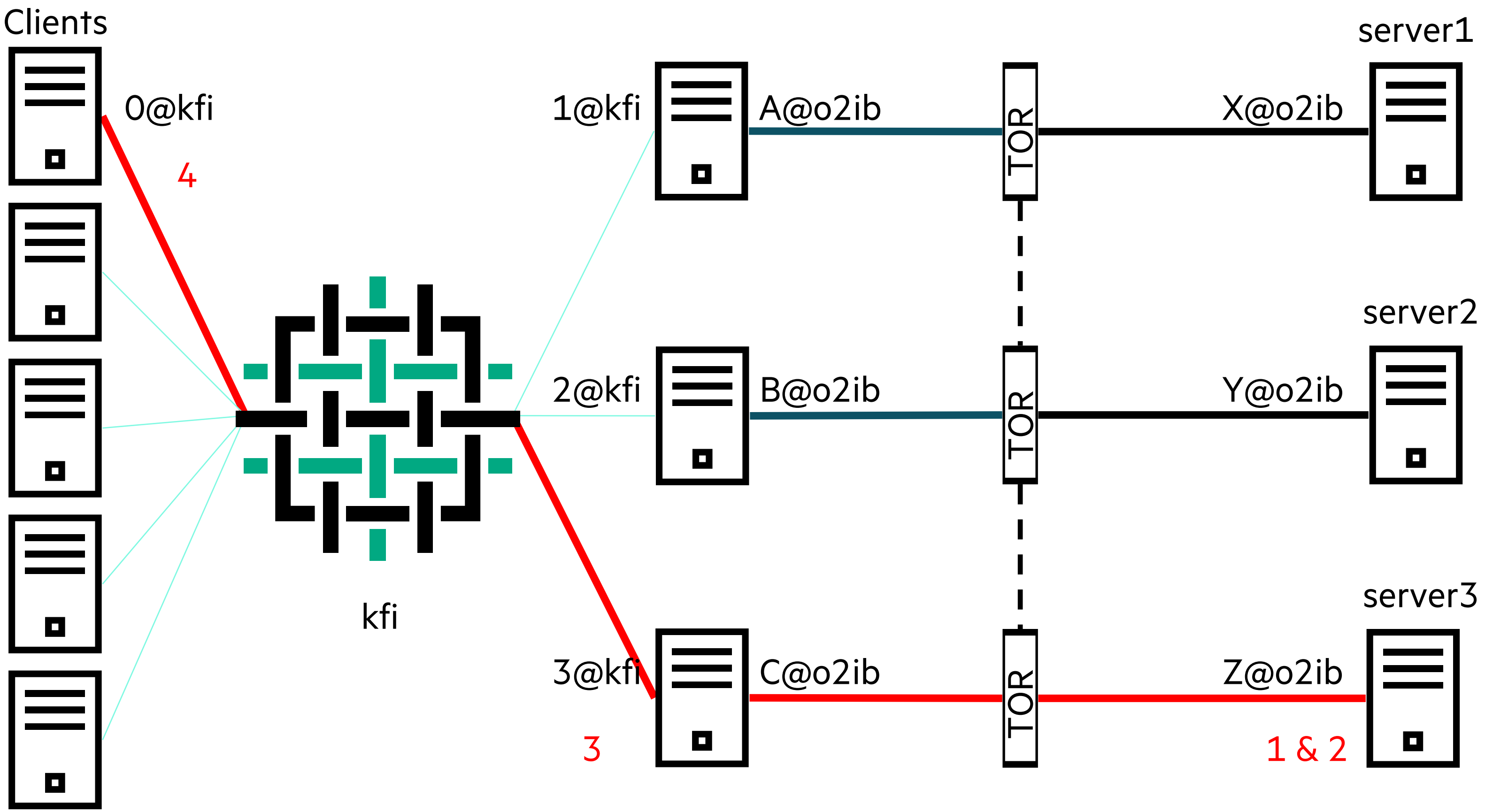
- Multi-LND Configuration:
  - Format filesystem using only tcp NIDs
  - Define routes using only tcp NIDs
  - Client fstab only reference tcp NIDs
  - No DFAs in config log
  - No DFAs in /etc/fstab
  - No DFAs in route configuration
- LNet Multi-Rail magic:
  - LNet peer discovery traffic over tcp
  - Discovery finds the kfi NIDs
  - UDSP prioritizes future traffic on kfi
- Serviceability (tcp/ip) + Performance (kfi)



```
net:
    - net type: tcp
      local NI(s):
        - interfaces:
              0: cxi0
    - net type: kfi
      local NI(s):
        - interfaces:
              0: cxi0
udsp:
    - idx: 0
      src: kfi
      action:
          priority: 0
```

TICKETS

- lst.sh & lst-survey
  - LNet Selftest wrapper
  - LU-16217
- Peer net selection priority
  - Code exists but is LBUGgy
  - LU-15944
  - LU-16573
    - lnetctl udsp add --dst o2ib --priority 0
- Small memory leak
  - LU-16575
- YAML Issues
  - LU-16572



- LNet Path Selection
  1. Select destination network
  2. Select destination NID
  3. Select router NID
  4. Select local NID
- Peer Net selection rules affect (1)

THANK YOU

Chris Horn
chris.horn@hpe.com