Advancing Digital Storage Innovation

LNET Routing Enhancements and Extracting Maximum Performance

Isaac Huang, Xyratex
David Dillow, ORNL

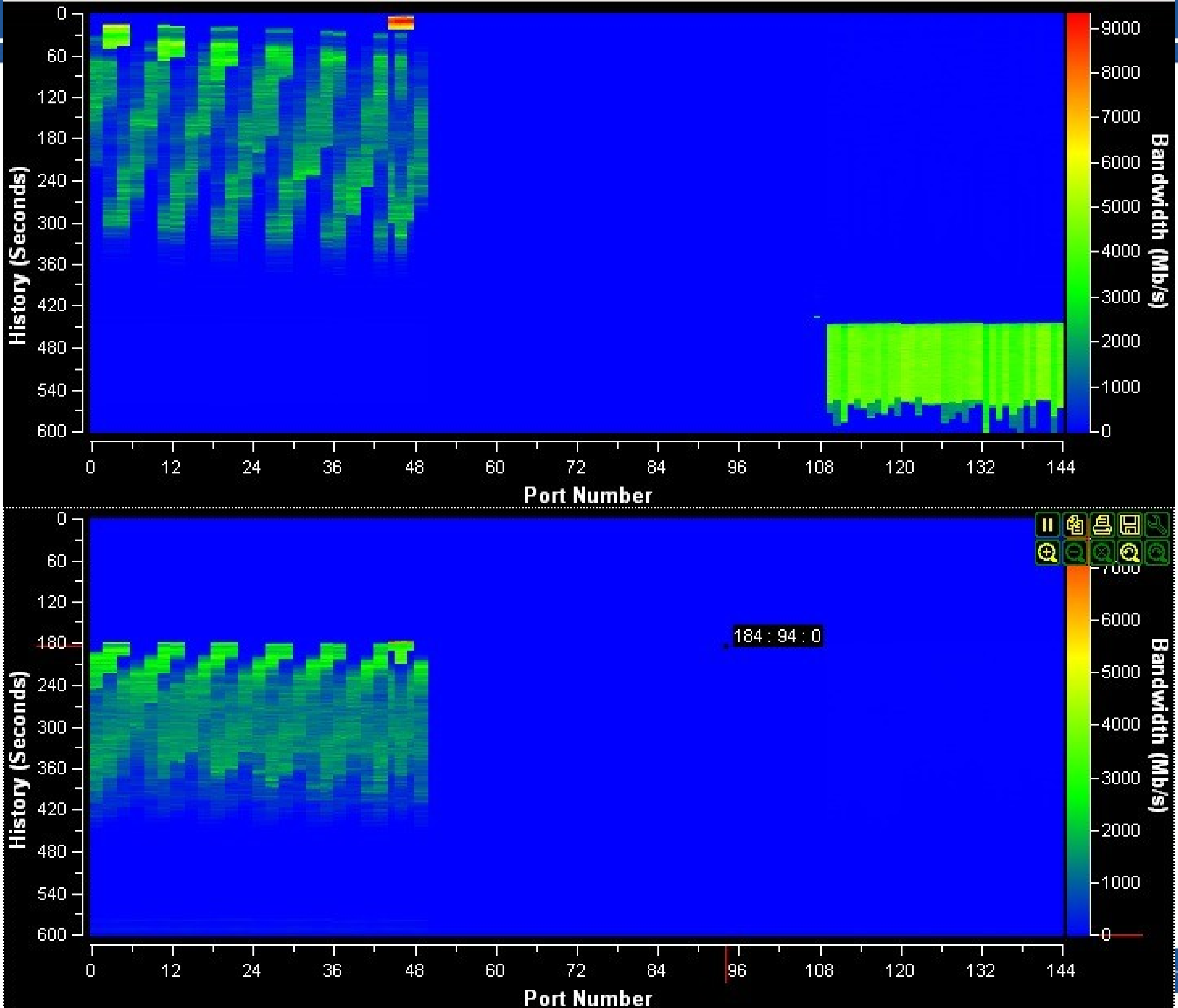
Presentation Overview

- LNet routing enhancements, more throughput and better reliability:
  - Better load balancing among routers: Router shuffler
  - Avoid router failures smartly: Asymmetric router pinger
  - Exploit locality in network: Fine grain routing
- Using multiple LNets to extract maximum performance

Router shuffler

- Clients/servers initialized with a same set of routers in a same order
  - We thought the order would be randomized over time
  - Clients are more synchronized than we believed
- Now set up with a same set of routers but randomly ordered

Router shuffler: how the problem was found

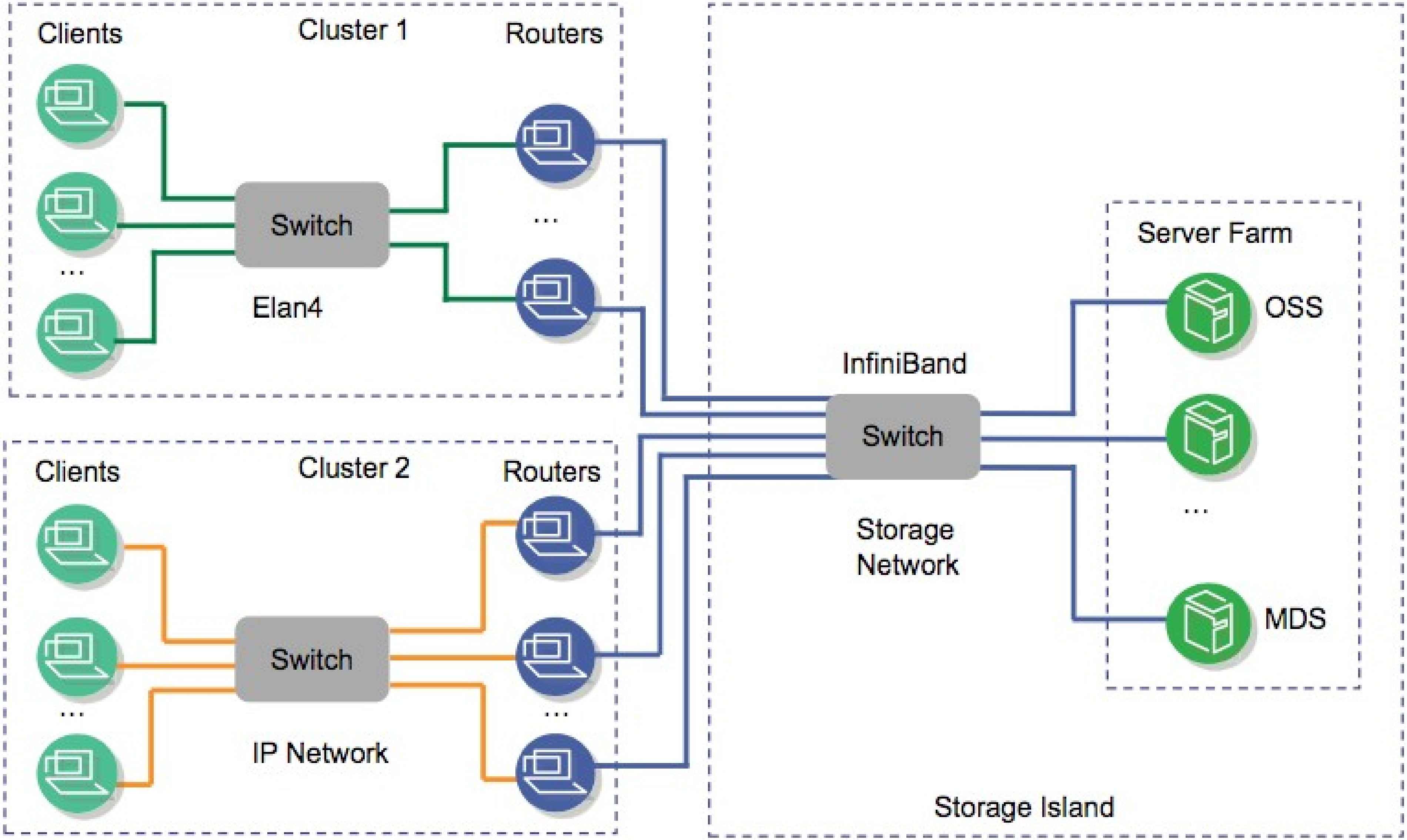


Asymmetric router pinger

- LNet pings routers and avoids bad ones
  - But can't handle interface failures on the other side
  - *options lnet live_router_check_interval=60 \ dead_router_check_interval=60 router_ping_timeout=60*
- Asymmetric pinger solves the problem by routers returning interface status in ping replies
  - Advanced feature: a router with a bad interface is avoided only if the interface is needed to forward a message
  - *options lnet avoid_asym_router_failure=1*

Asymmetric router pinger

Sample routed networks



Weighted routing

- Routers to a same network were treated equally:

```
options lnet routes="tcp0 132.6.1.[1-8]@o2ib0; \
                     tcp1 132.6.2.[1-8]@o2ib0; \
                     o2ib0 192.168.0.[1-8]@tcp0; \
                     o2ib0 10.1.1.[100-109]@tcp1"
```

- With weighted routing, they are divided into subsets of different weights (priorities):

```
options lnet routes="tcp0 1 132.6.1.[1-4]@o2ib0; \
                     tcp0 2 132.6.1.[5-8]@o2ib0; \
                     tcp1 1 132.6.2.[1-4]@o2ib0; \
                     tcp1 2 132.6.2.[5-8]@o2ib0; \
                     o2ib0 1 192.168.0.[1-4]@tcp0; \
                     o2ib0 2 192.168.0.[5-8]@tcp0; \
                     o2ib0 1 10.1.1.[100-104]@tcp1 # preferred routers; \
                     o2ib0 2 10.1.1.[105-109]@tcp1 # backup routers to o2ib0"
```

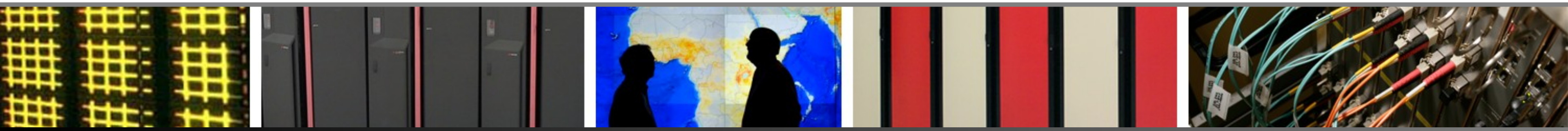
Extracting Maximum Performance (FGR in a nutshell)

20 Years of Excellence in Computational Science

OLCF

Oak Ridge Leadership Computing Facility

1992–2012



What is “Fine-Grained Routing?”

- Using LNets to control network flows
  - No routers required!

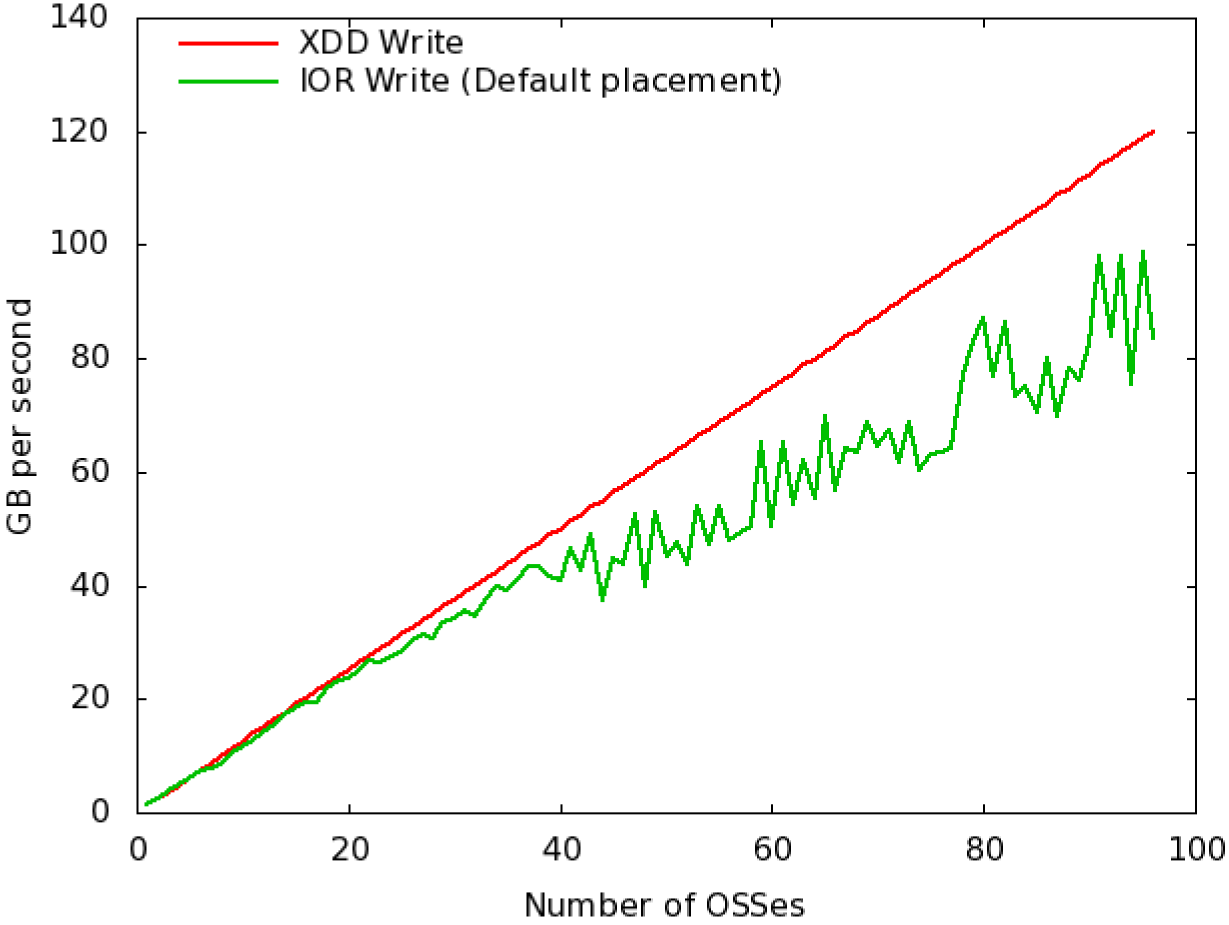
- More information:
  - IPCCC 2011: Enhancing I/O throughput via efficient routing and placement for large-scale parallel file systems

    http://doi.ieeecomputersociety.org/10.1109/PCCC.2011.6108062
  - CUG 2011: I/O Congestion Avoidance via Routing and Object Placement

    http://info.ornl.gov/sites/publications/files/Pub30140.pdf

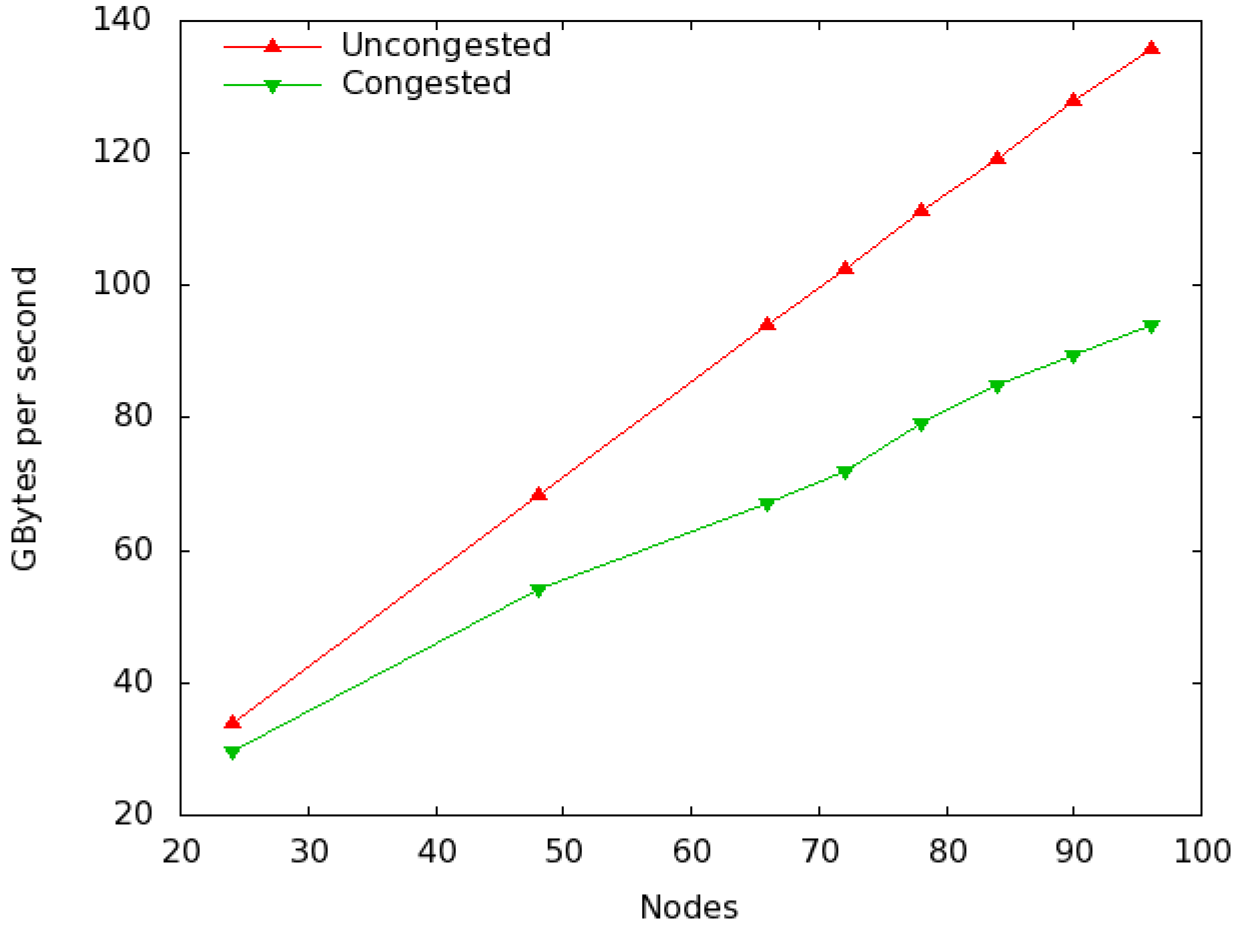
Why do we need to control network flows?

Torus congestion



Why do we need to control network flows?

InfiniBand congestion



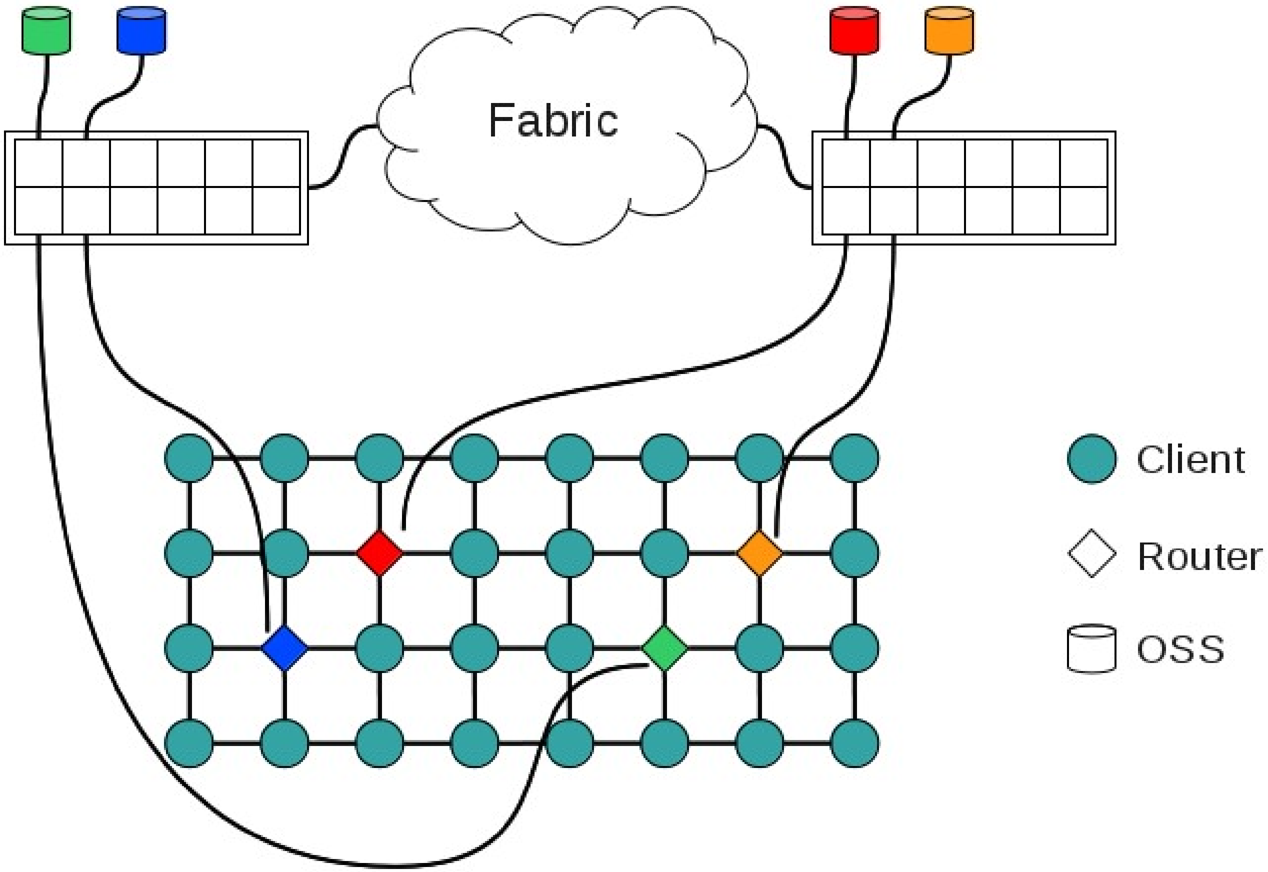
How does LNet contribute?

- Basic LNet routing gives worst of both worlds!
- LNet routing is dispersive
  - Subsequent messages to a remote LNet round-robin through the available routers
  - Impossible to predict path of traffic
  - Impossible for applications/libraries to optimize

Our approach: Use multiple LNets

- Allows us to steer traffic
- One LNet per IB leaf switch
  - 32 groups of 6 routers/6 OSSes
  - Nearest-neighbor
  - Round-robin allocation
- One LNet per OSS
  - 192 o2ib networks
  - “Projected” routing

“Projected” routing



OSS Configuration

```
options lnet ip2nets="o2ib(ib2)     10.36.227.*;\
                      o2ib201(ib2)  10.36.227.*;"
options lnet routes="ptl1  1  10.36.229.1@o2ib201;\
                     ptl1  11 10.36.229.97@o2ib201;\
                     ptl1  21 10.36.229.7@o2ib201;\
                     ptl0  1  10.36.223.[1-48]@o2ib;\
                     o2ib3 1  10.36.222.[81-85]@o2ib;"
```

```
options lnet ip2nets="o2ib(ib2)     10.36.227.*;\
                      o2ib201(ib2)  10.36.227.*;"
options lnet routes="ptl1  1  10.36.229.97@o2ib202;\
                     ptl1  11 10.36.229.1@o2ib202;\
                     ptl1  21 10.36.229.103@o2ib202;\
                     ptl0  1  10.36.223.[1-48]@o2ib;\
                     o2ib3 1  10.36.222.[81-85]@o2ib;"
```

Router Configuration (o2ib201)

- Router 10.36.229.1 (nid5716)

  options lnet networks="ptl1,o2ib,o2ib201,o2ib202,o2ib208"

- Router 10.36.229.97 (nid5719)
  - options lnet networks="ptl1,o2ib,o2ib201,o2ib202,o2ib206"
- Router 10.36.229.7 (nid5924)
  - options lnet networks="ptl1,o2ib,o2ib201,o2ib206,o2ib208"

Client Configuration (config file)

```
# LNET, Primary, Secondary, Tertiary routers
o2ib201 5716 5719 5924
o2ib202 5719 5716 5927
o2ib203 1623 1620 1831
o2ib204 1620 1623 1828
o2ib205 1540 1543 1560
o2ib206 5927 5924 5719
o2ib207 1543 1540 1563
o2ib208 5924 5927 5716
o2ib209 1828 1831 1620
o2ib210 1831 1828 1623
```

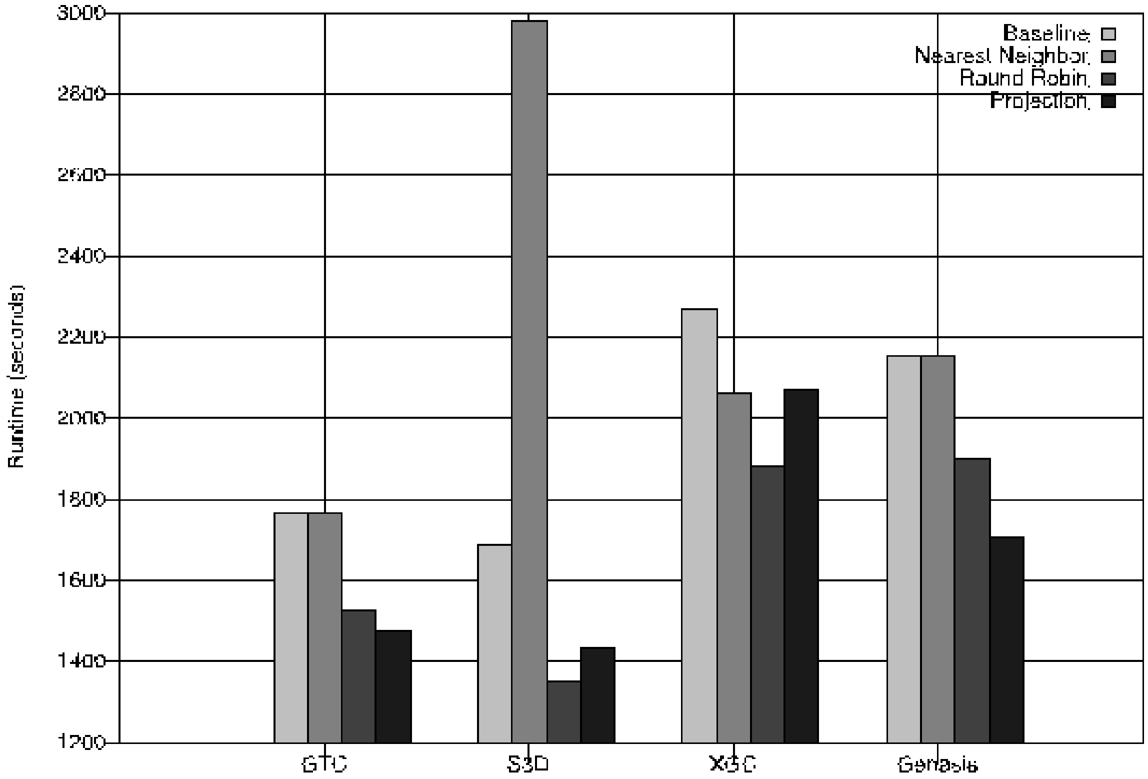
Client Configuration (scripting)

```
cat $MAP | while read lnet rtrs; do
    set -- $rtrs
    weight=1
    while [[ $1 ]]; do
        /sbin/lctl --net ${lnet} add_route ${1}@ptl1 $weight
        weight=$((weight + 10))
        shift
    done
done
```

Effective commands for o2ib201:

```
/sbin/lctl –net o2ib201 add_route 5716@ptl1 1
/sbin/lctl –net o2ib201 add_route 5719@ptl1 11
/sbin/lctl –net o2ib201 add_route 5924@ptl1 21
```

Application Results (briefly)



Questions?

20 Years of Excellence in Computational Science

OLCF

Oak Ridge Leadership Computing Facility

1992–2012

xyratex

Advancing Digital Storage Innovation