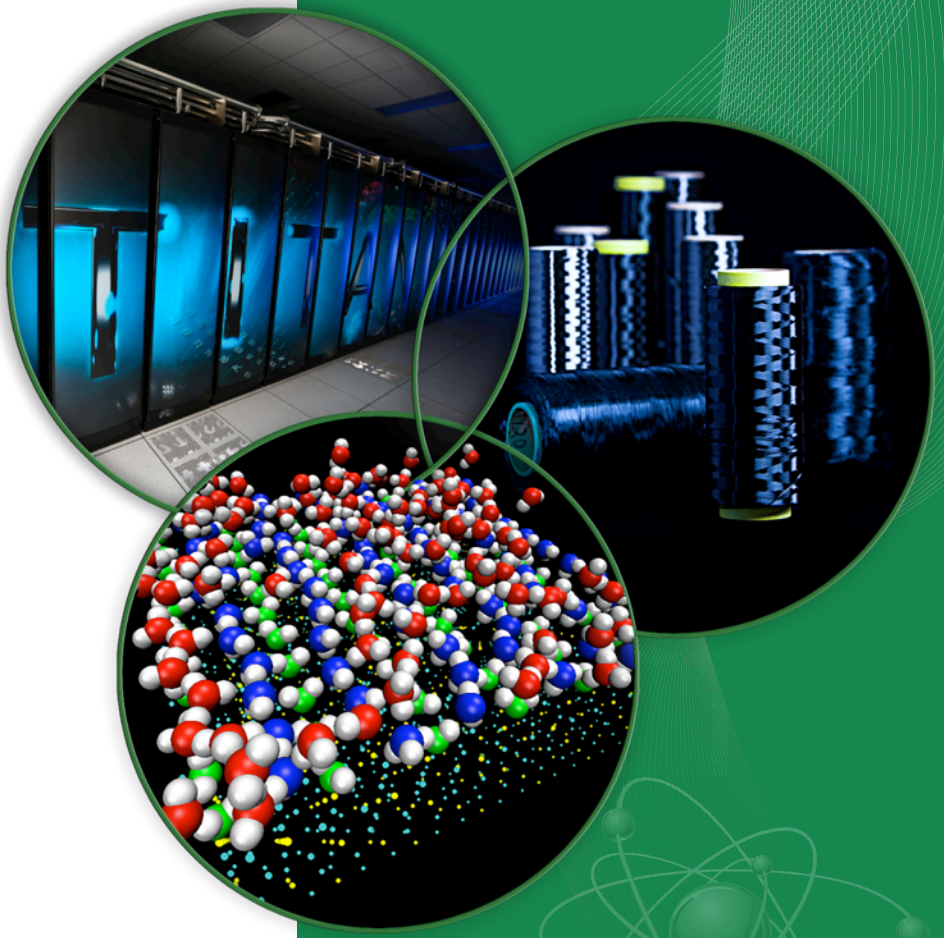
Network Contention and Congestion Control: Lustre Fine-Grained Routing



Matt Ezell
HPC Systems Administrator

International Workshop on The Lustre Ecosystem: Challenges and Opportunities
Annapolis, Maryland *March 3-4, 2015*

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

Congestion

Performance

Fat Tree Network

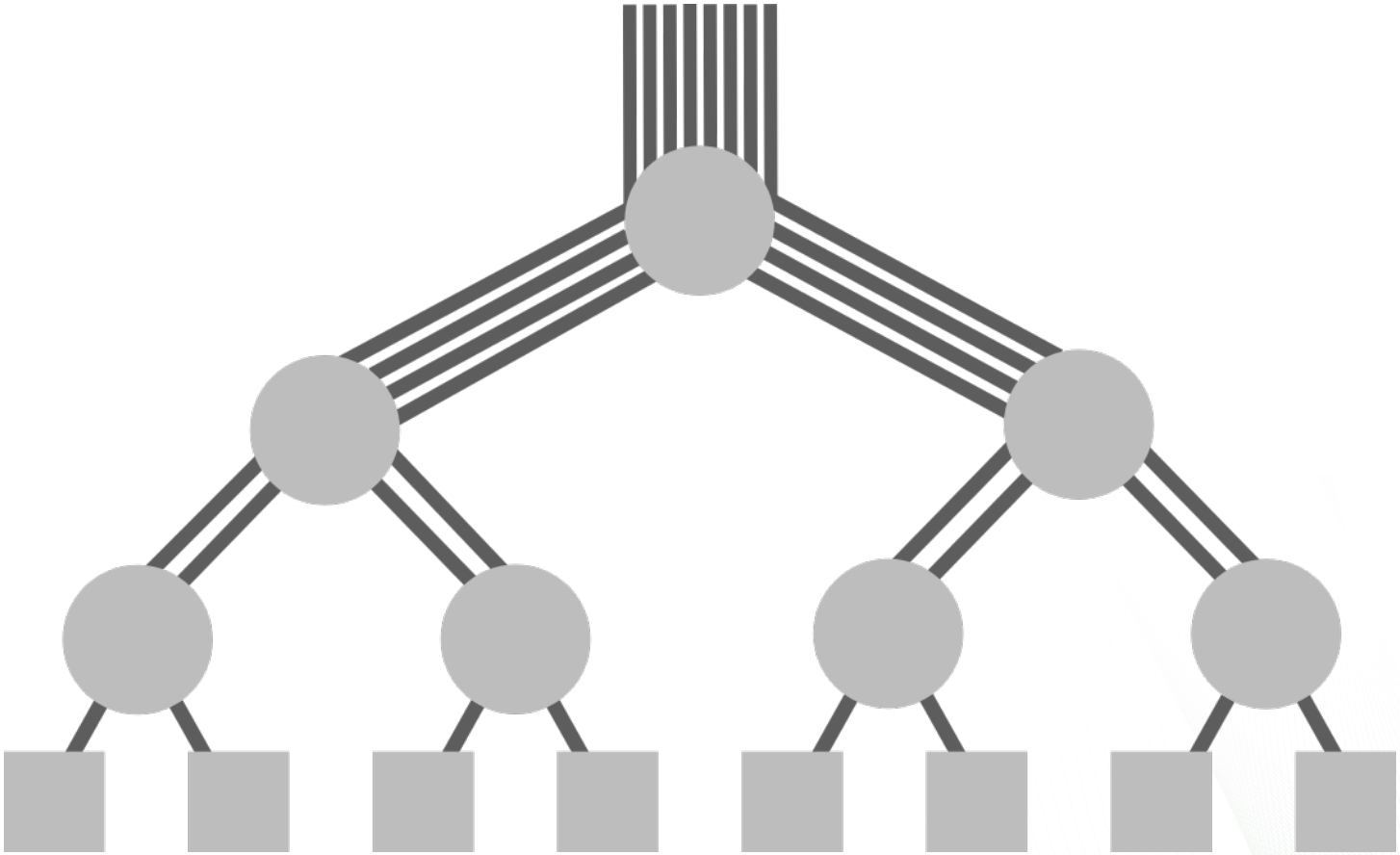


Image from http://clusterdesign.org/fat-trees/fat_tree_varying_ports/

Fat Tree Network (Folded Clos Network)

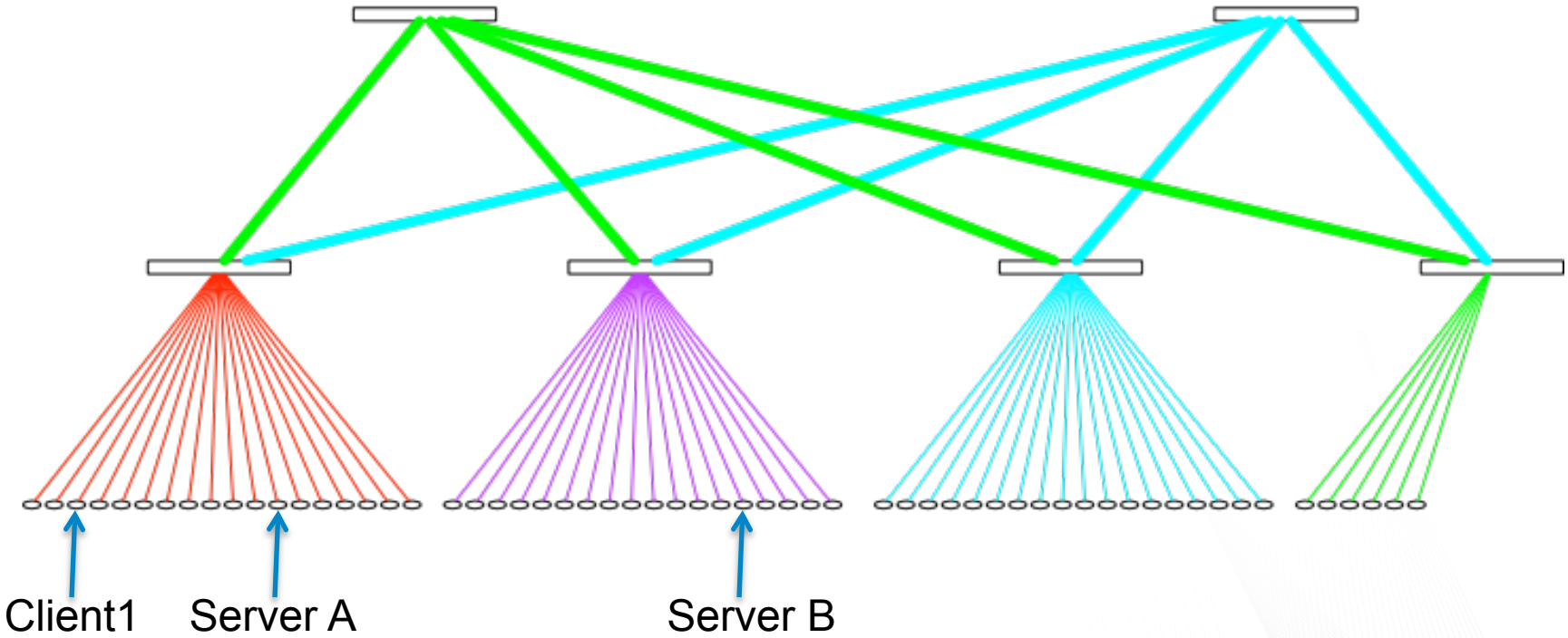


Image from http://clusterdesign.org/fat-trees/60_nodes_labeled/

Fat Tree Network (Folded Clos Network)

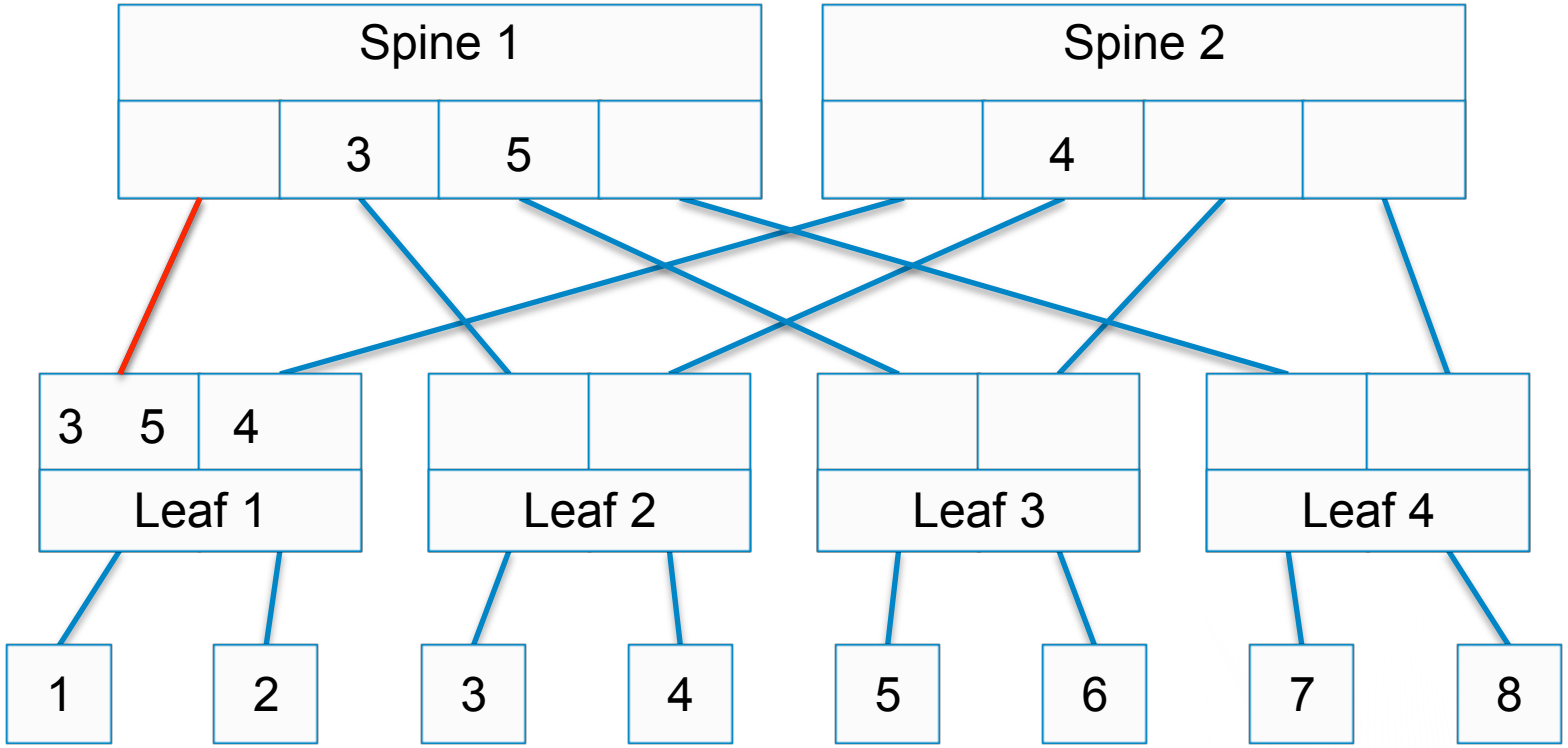
- Network can be
  - Non-blocking (1:1)
  - Blocking (x:1)
- Congestion and hot spots can (significantly) reduce performance
  - Even in an expensive full-bisection-bandwidth setup
  - Worse in under-provisioned or irregular networks

InfiniBand Routing

- InfiniBand uses a Linear Forwarding Table
  - A given destination always uses the same output port

Destination LID	Output Port
7	2
59	3
66	3
101	5
121	11
200	11
232	20
299	27
341	32
351	35

Hot Spot Example – 4 Port Switches



What happens when
node 1 => node 3
node 2 => node 5

Gemini 3D Torus

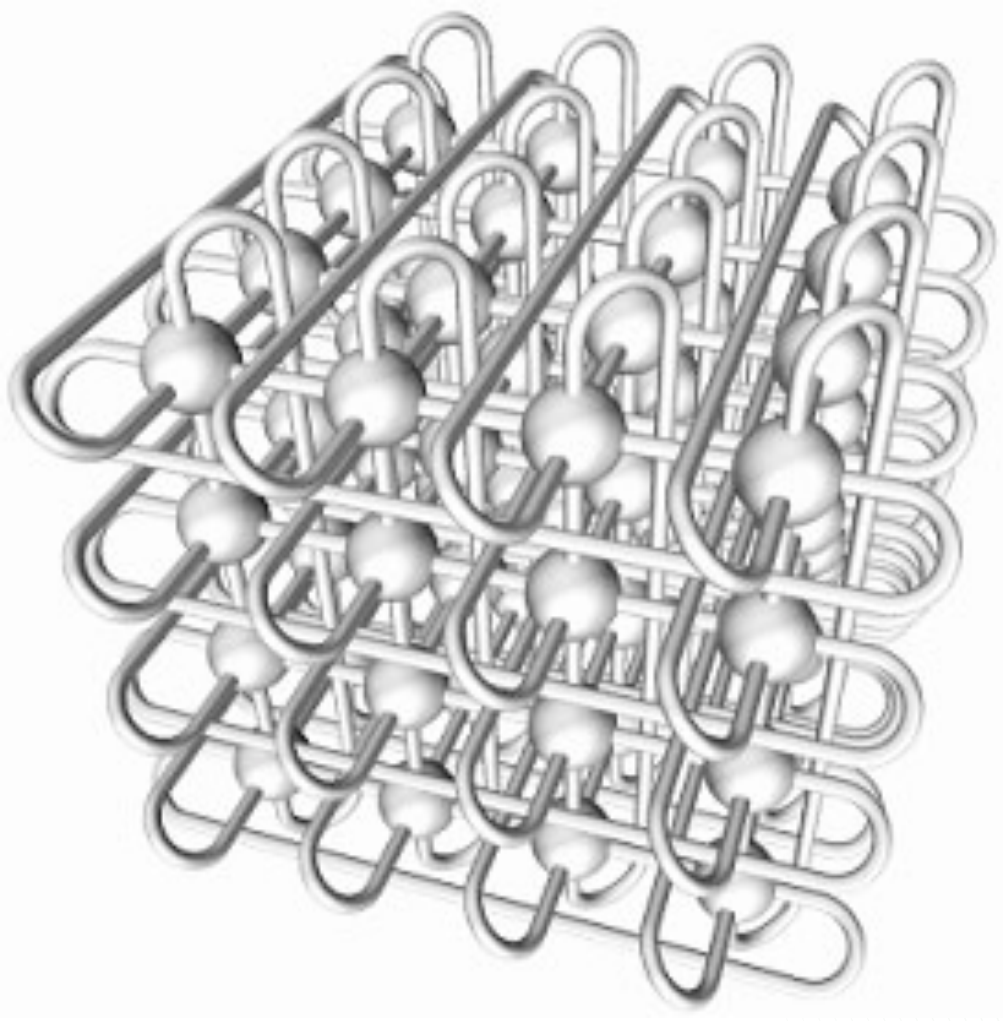


Image from http://clusterdesign.org/torus/

Gemini Dimension Ordered Routing

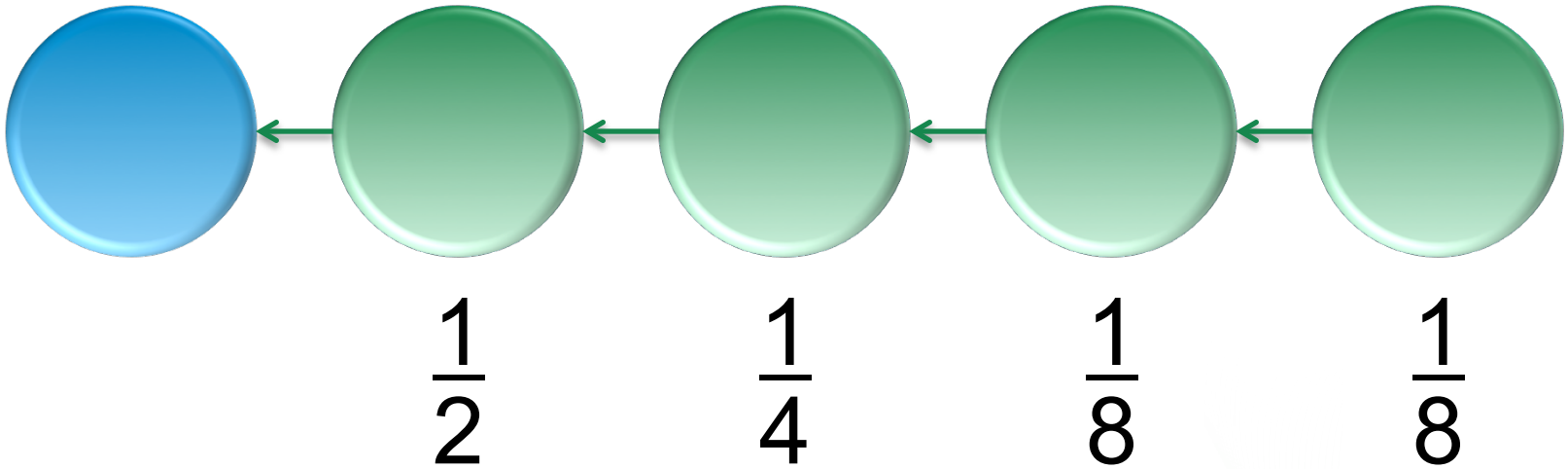
X → Y → Z

How fast are Gemini Links?

- It depends!
- Link speed depends on link type
- Protocol overhead is around 35% for large messages

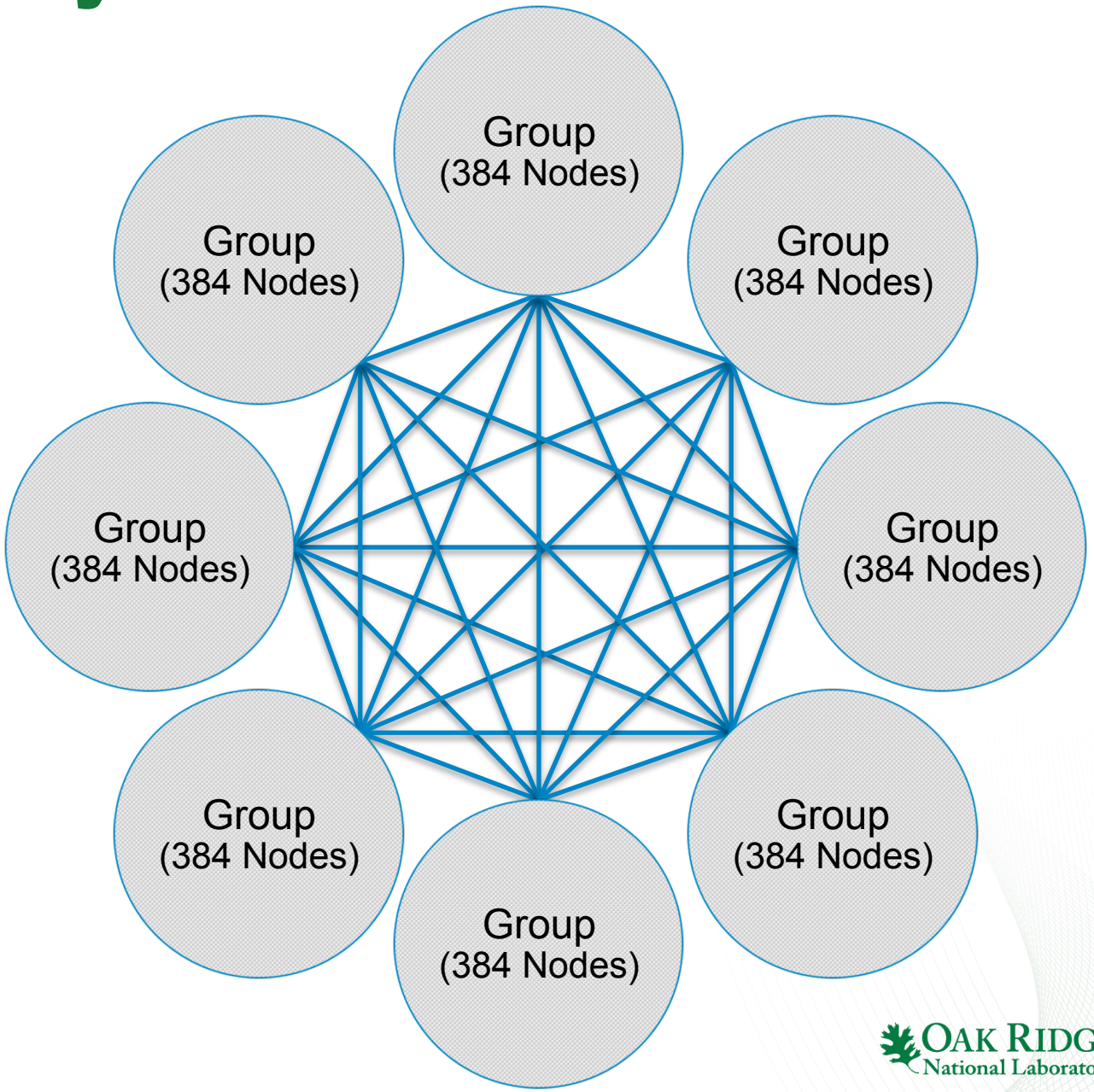
Link Type	Link Data Rate	Number Links	Raw Bitrate	Data Rate
Y-Mezzanine	6.25 gbps	12	9.375 GB/s	~ 6 GB/s
Z-Backplane	5.0 gbps	24	15 GB/s	~ 9.75 GB/s
X, Z Cable	3.125 gbps	24	9.375 GB/s	~ 6 GB/s
Y Cable	3.125 gbps	12	4.6875 GB/s	~ 3 GB/s

Geometric Bandwidth Reductions



Titan is a 25 x 16 x 24 torus

Dragonfly Network



Dragonfly Network

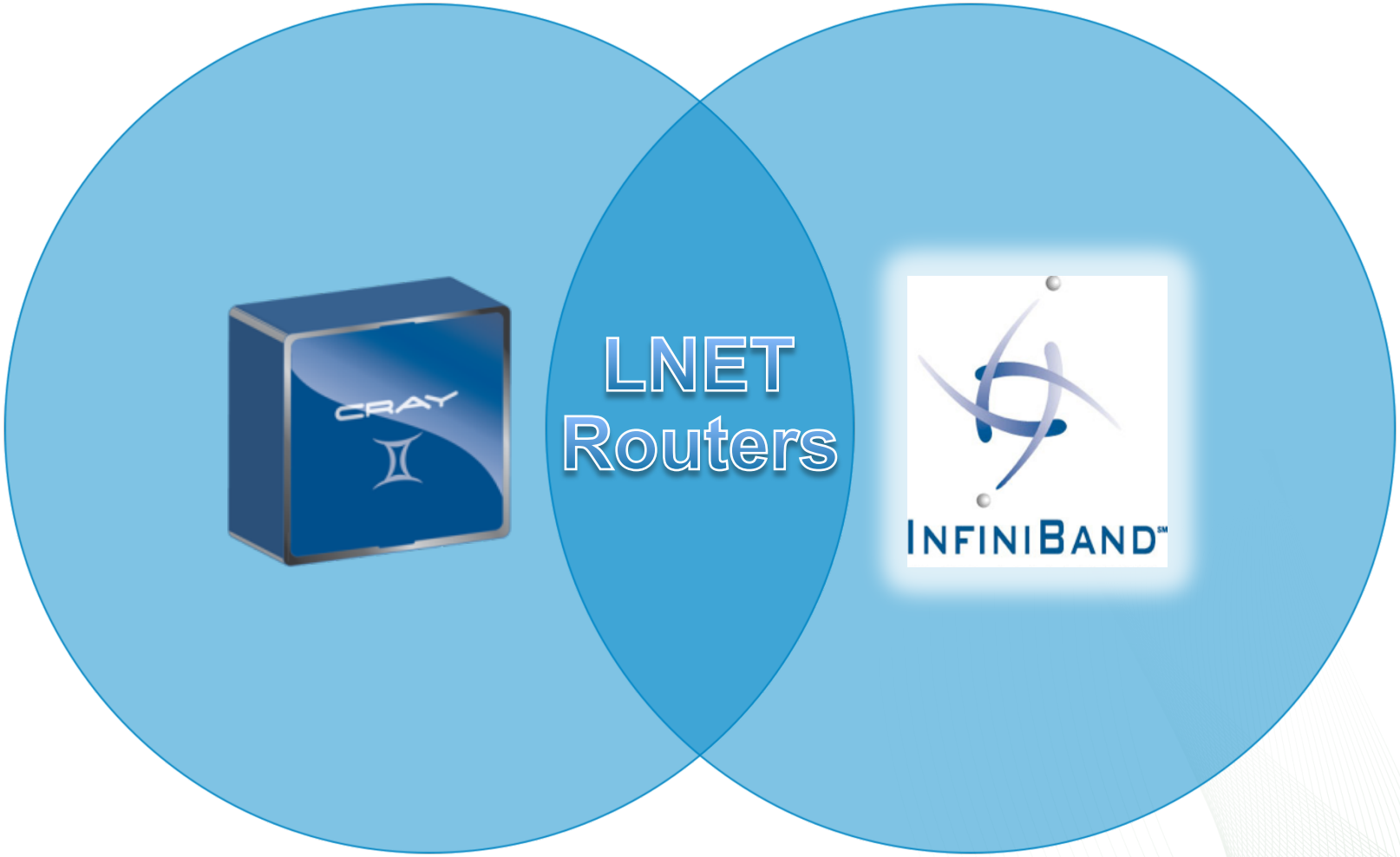
- In an Aries network, global bandwidth can be increased by purchasing more links
- Adaptive routing can ease congestion
  - But it makes it difficult to predict the path that a packet will take

The Million Dollar Question

?

lustre™

LNET Routing



LNET Routing

identifier@network

10.10.10.101@o2ib0

10.10.10.101@o2ib201

9409@gni0

9409@gni101

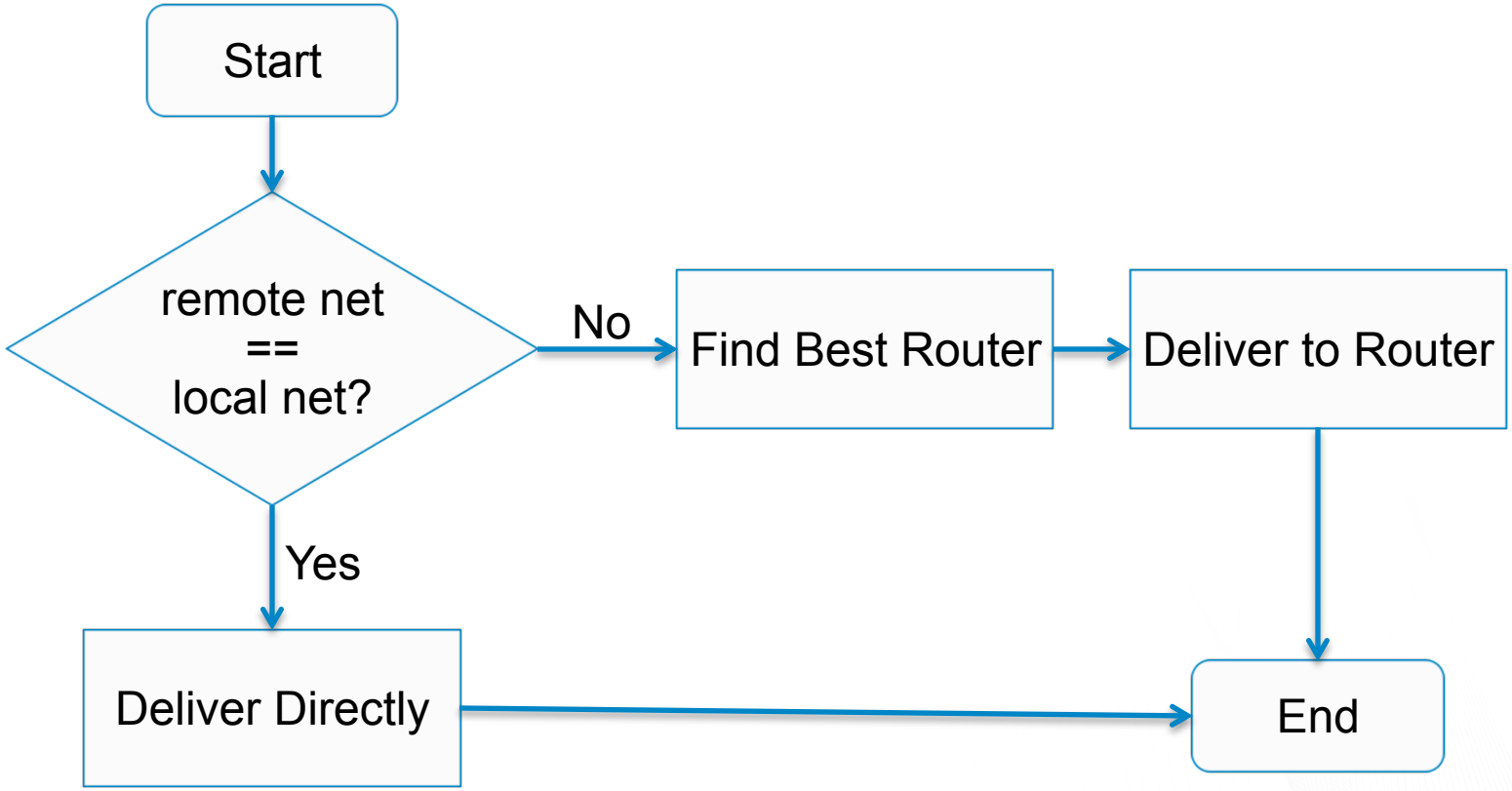
LNET Routing

options lnet routes="remoteNet hops id@localNet"

lctl --net remoteNet add_route id@localNet hops

```
net              o2ib225 hops 10 gw                18329@gni102 up
net              o2ib225 hops 1 gw                 10871@gni102 up
net              o2ib234 hops 10 gw                 5991@gni102 up
net              o2ib234 hops 10 gw                18247@gni102 up
net              o2ib234 hops 1 gw                 10921@gni102 up
net              o2ib208 hops 10 gw                15218@gni102 up
net              o2ib208 hops 10 gw                 2946@gni102 up
net              o2ib208 hops 1 gw                 7788@gni102 up
net              o2ib217 hops 10 gw                15212@gni102 up
net              o2ib217 hops 10 gw                 2908@gni102 up
```

LNET Routing



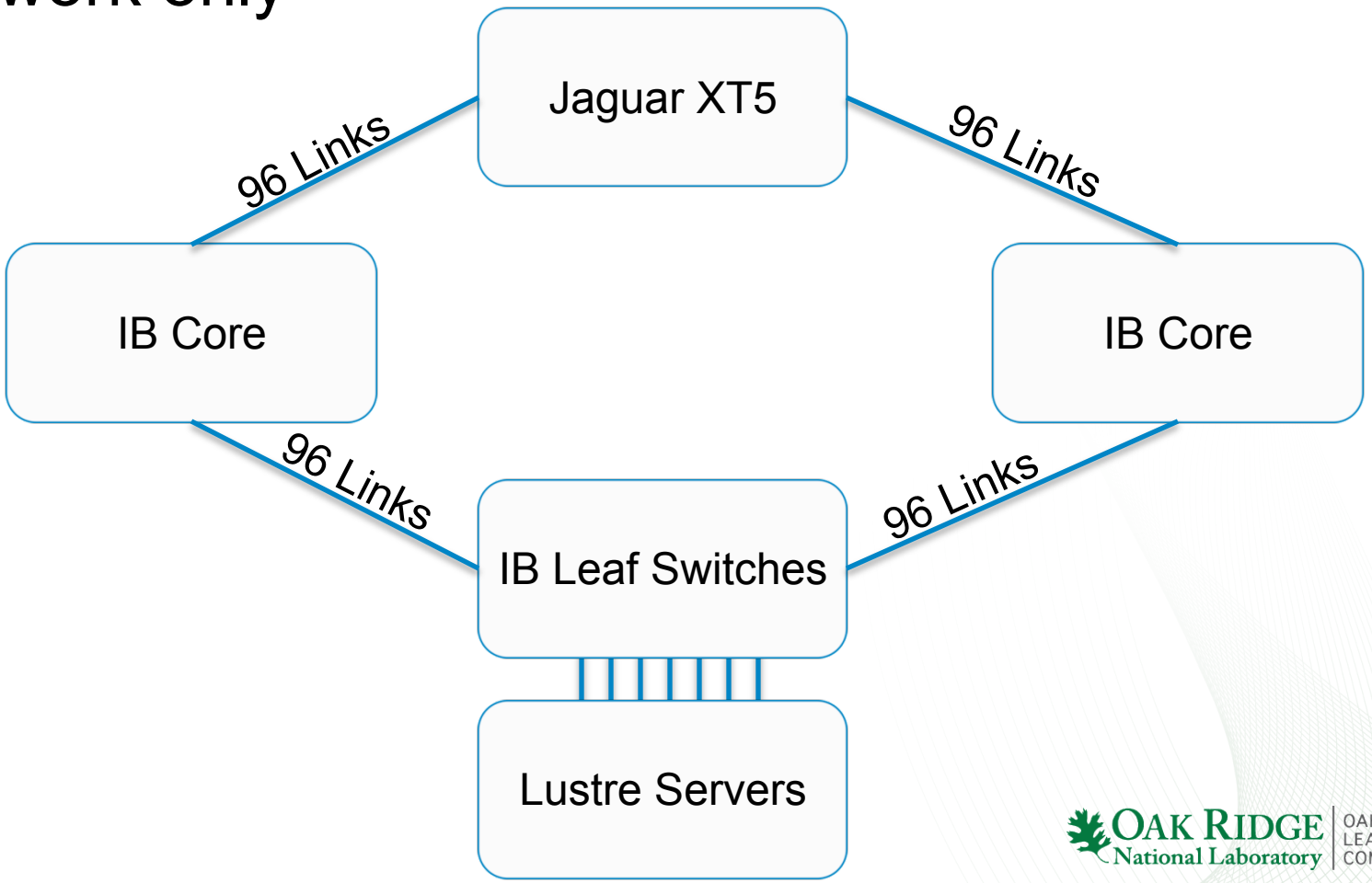
Why Route?

- Control bandwidth by varying the number of routers
- Control I/O paths by selecting which routers to use
- Control route computation by partitioning fabrics

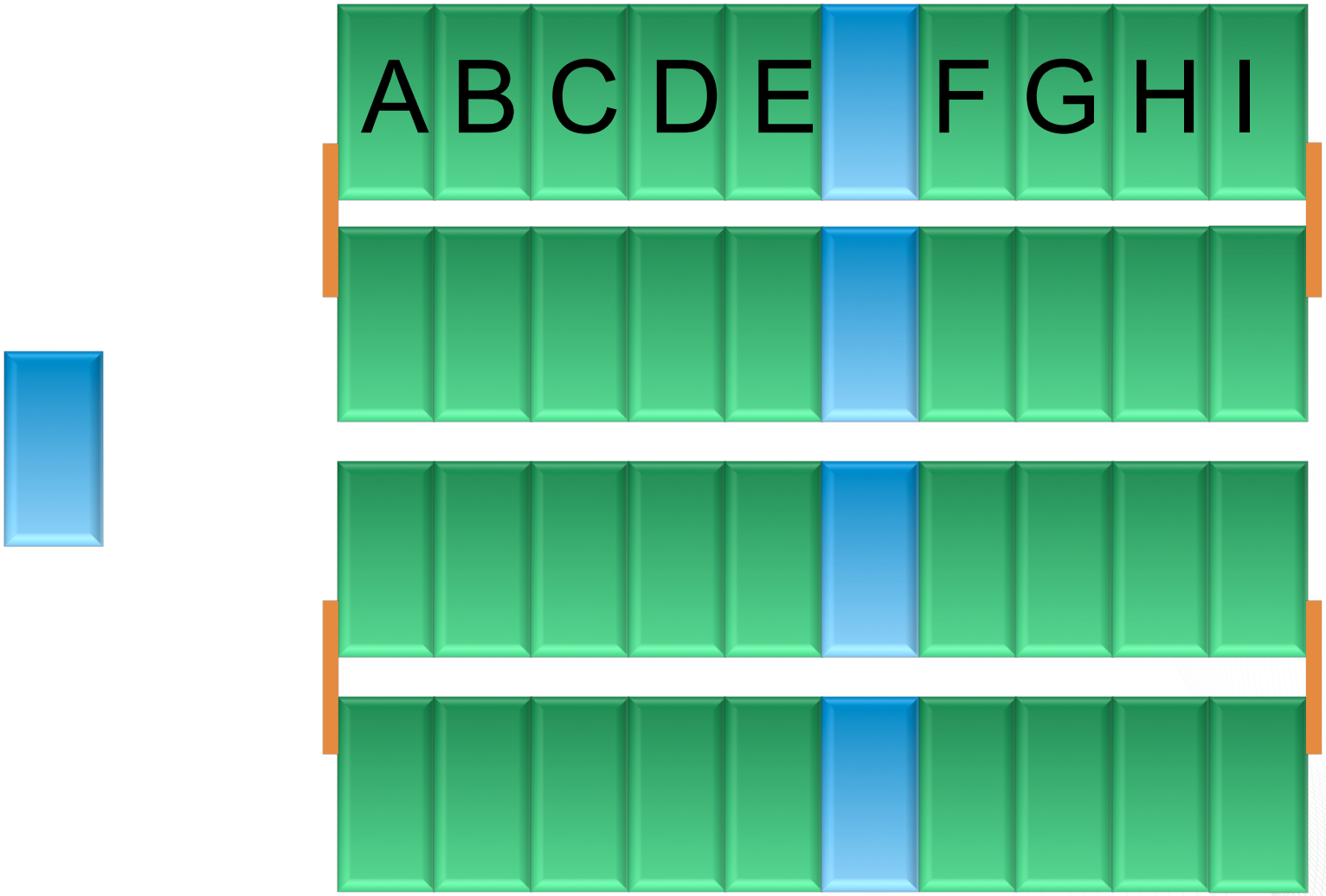
… and if you are using multiple network types,
you have to route traffic

Initial Implementation for Jaguar

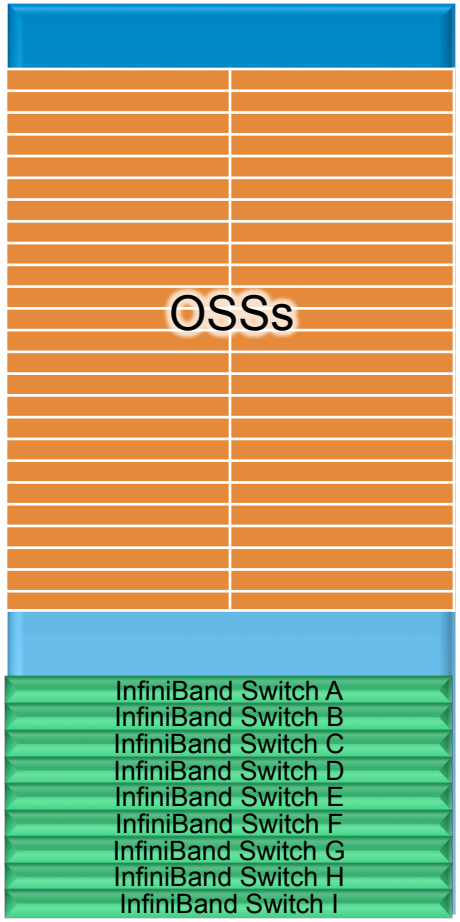
- Focused on relieving contention within the IB network only



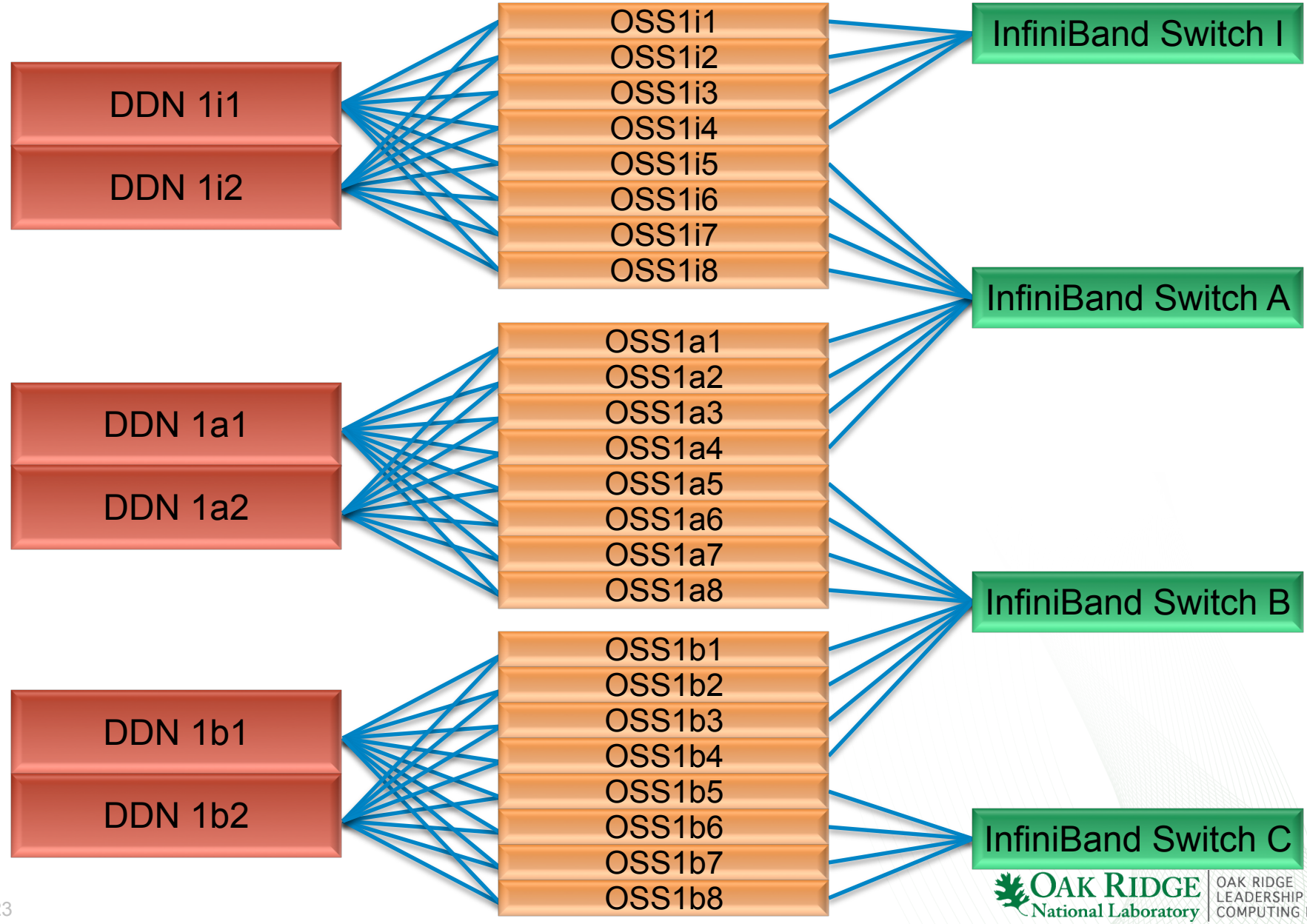
Atlas Layout



Atlas Layout



OSS Connections



How many routers?

Performance Goal: > 1TB/s

÷

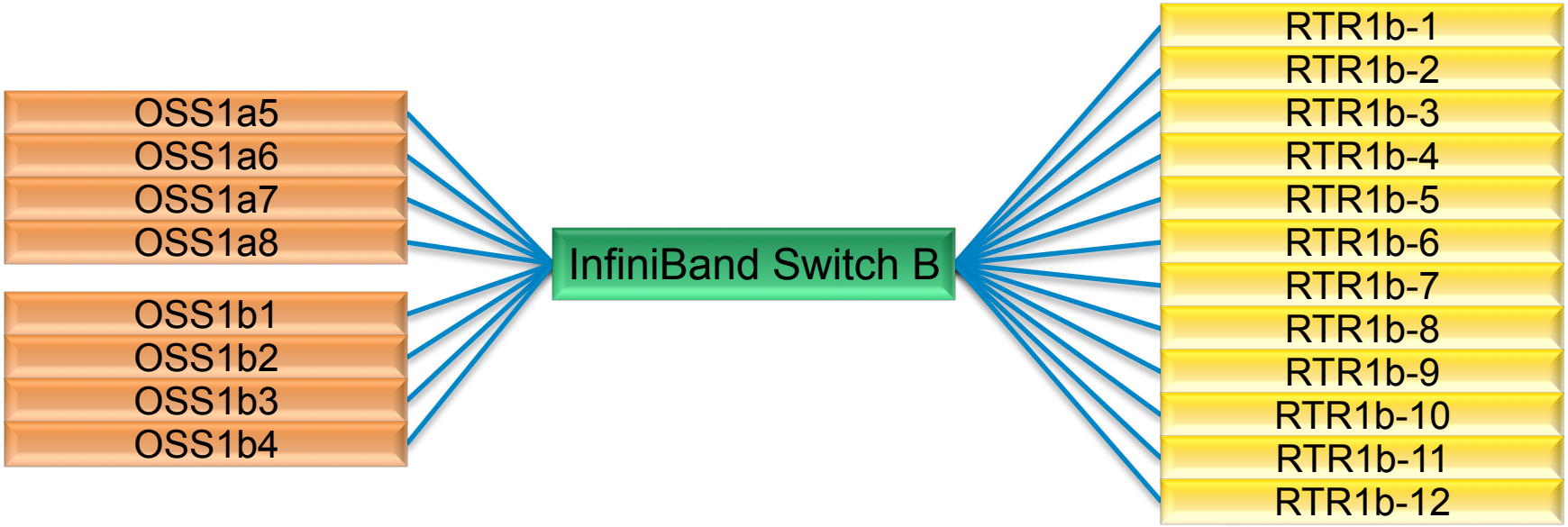
Router Performance: 2.6GB/s

=

Required Router Count: > 385

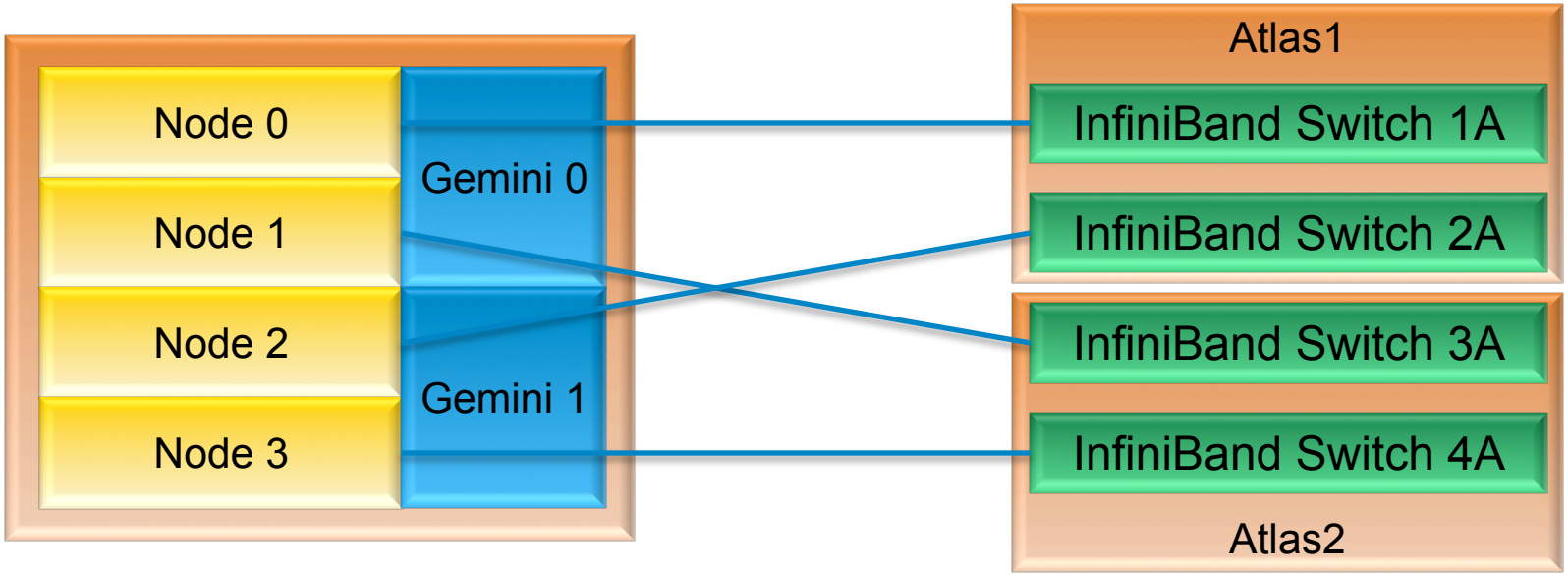
Deployed Router Count: > 432

FGR Group

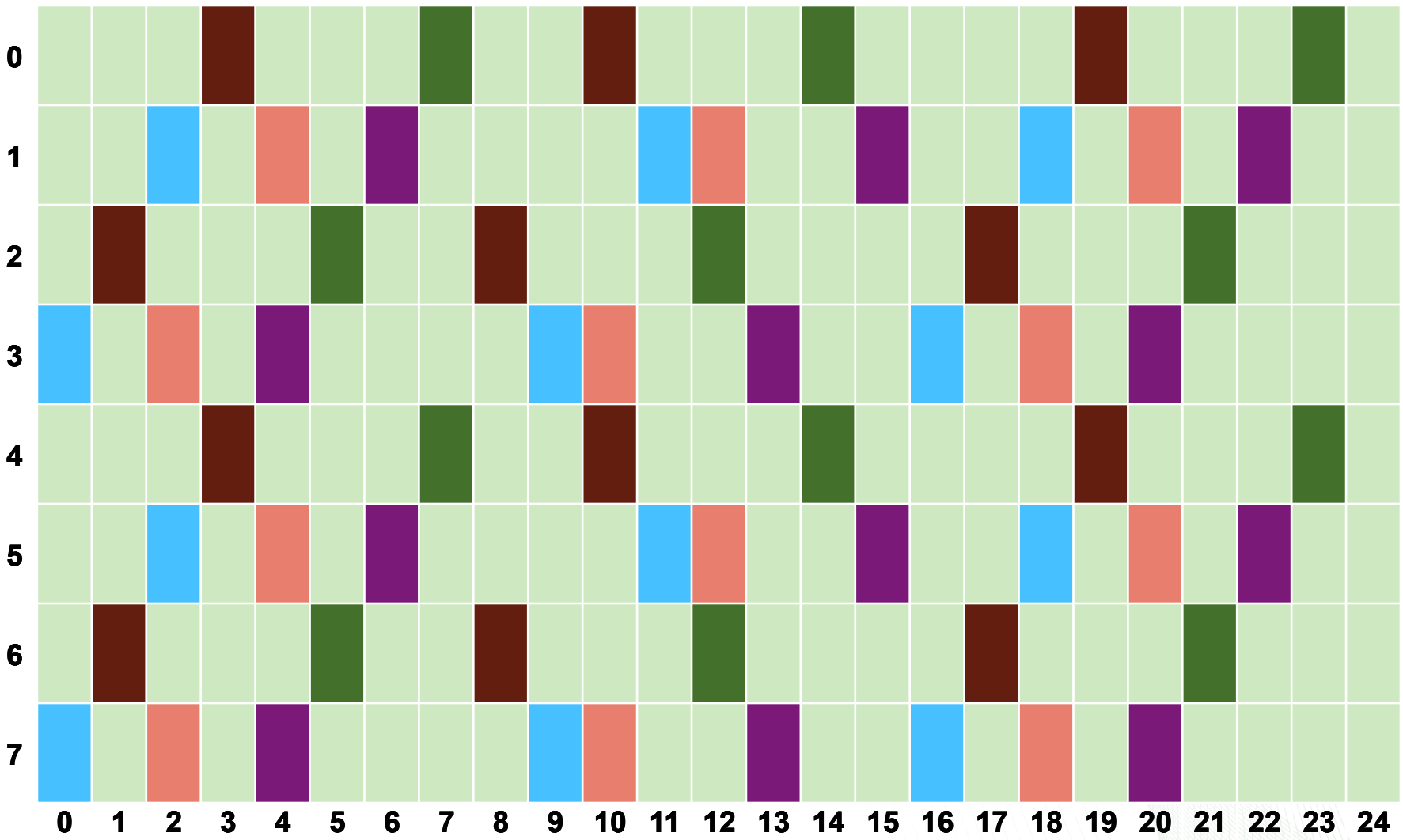


8:12 Ratio

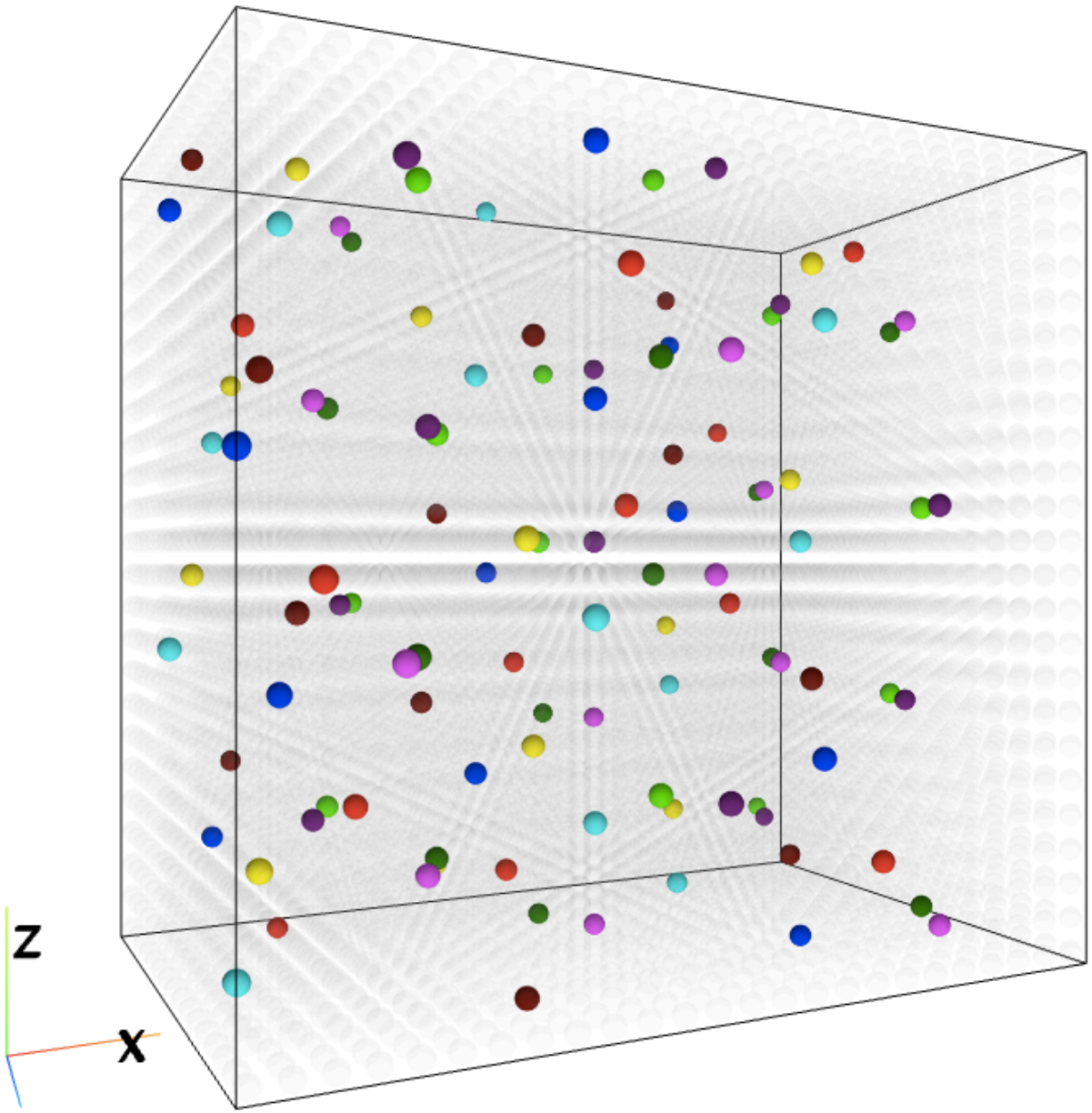
Router Module



Router Layout



Router Placement

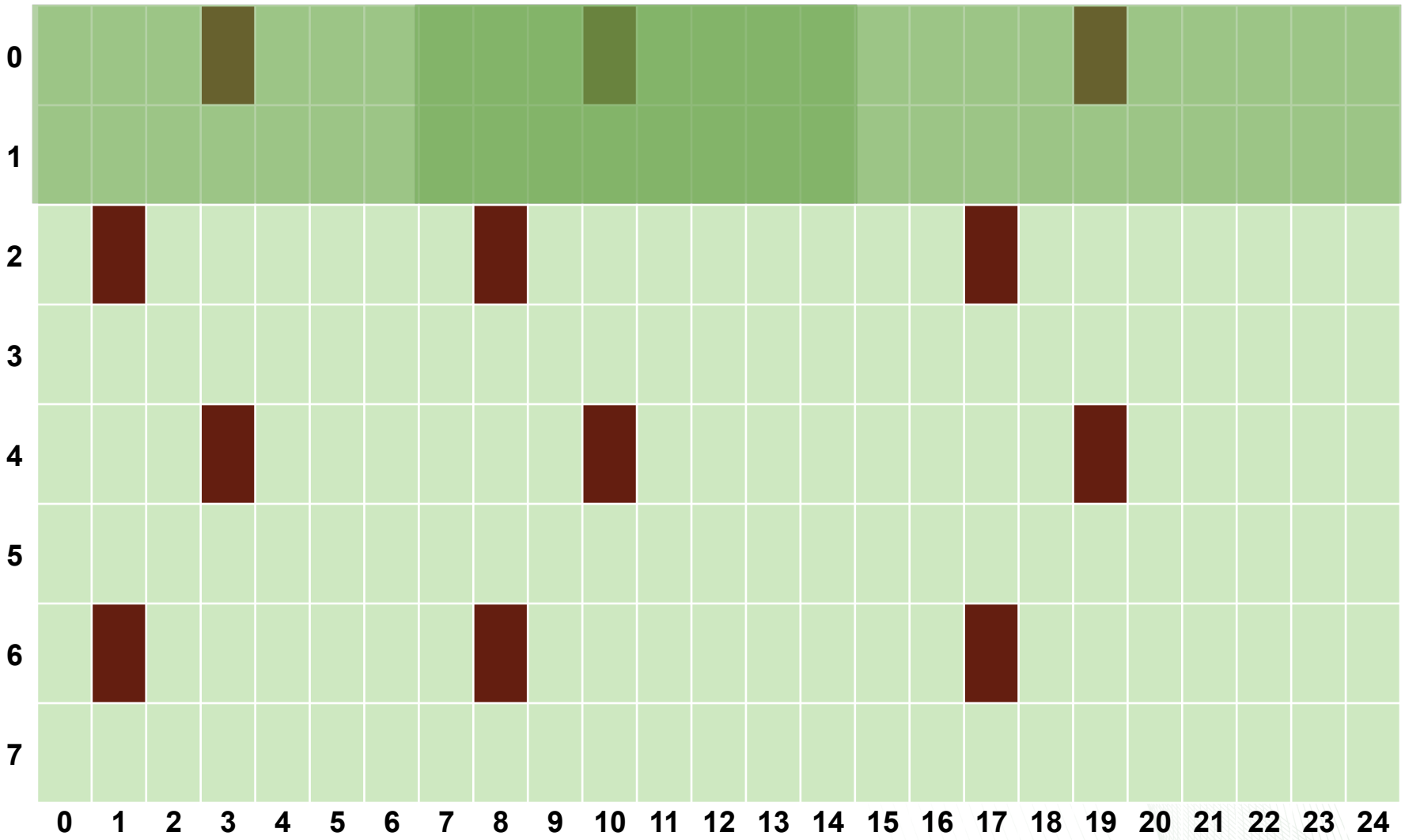


Fine-Grained Routing Configuration

36

12

Routing Group



FGR On an XC30

- Our XC30 is only two groups
  - We have not measured significant contention in the Aries network
- 9 routers with 4 interfaces each, to provide connectivity to all 36 TOR switches
- Routers provide higher percentage of peak bandwidth, compared to our XK7

FGR on an IB Cluster

- Compute nodes and routers connect to to a 648-port core switch
- 36 routers connect directly to the 36 TOR switches

FGR Best Practices

- Understand your topology
- Understand your application I/O workload
- Create large FGR groups
- Create FGR groups in both directions
- Verify all settings and physical connections

Questions?

ezellma@ornl.gov