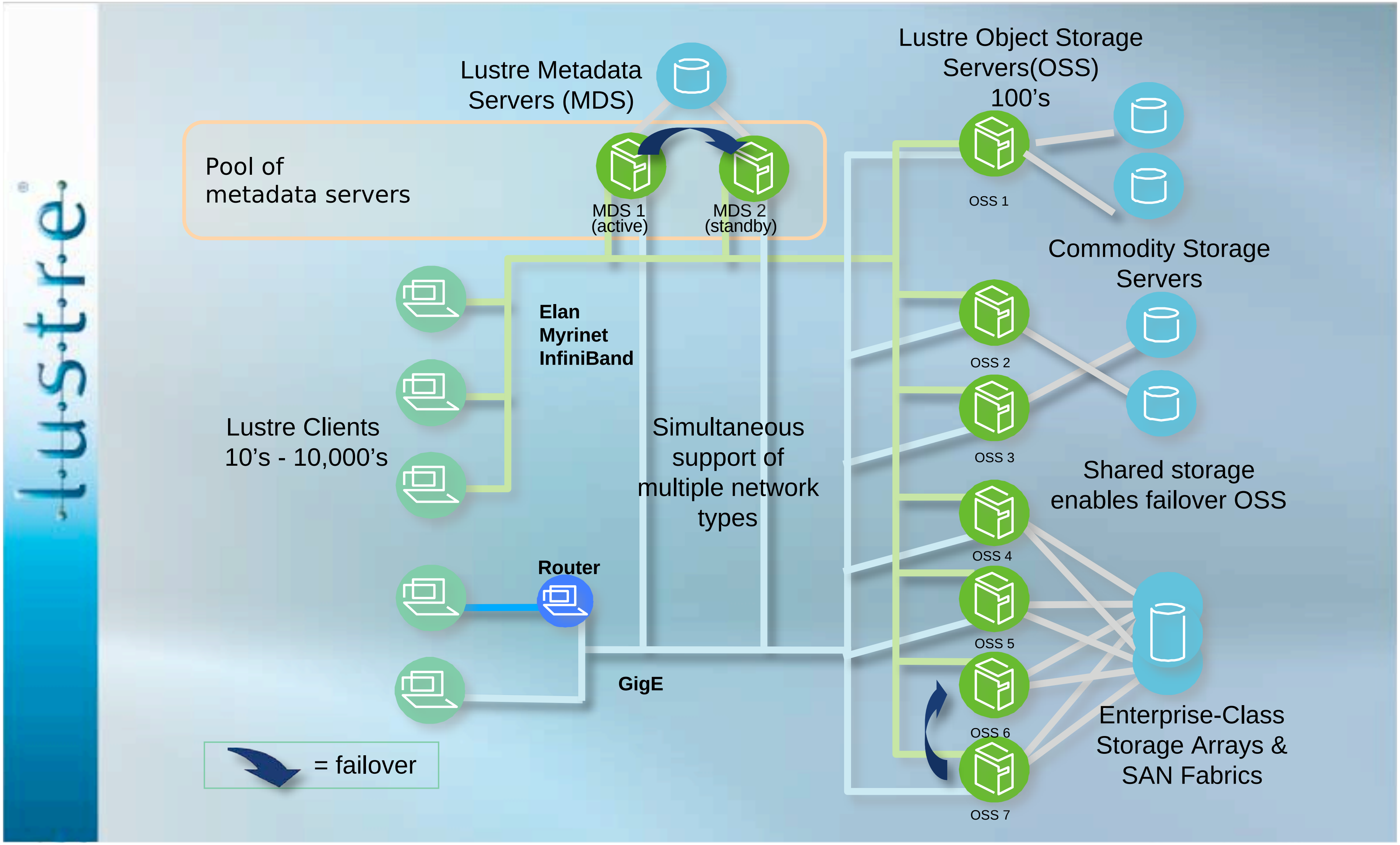
Lustre Networking - an overview
LUG 2007

Cluster
File
Systems, Inc™

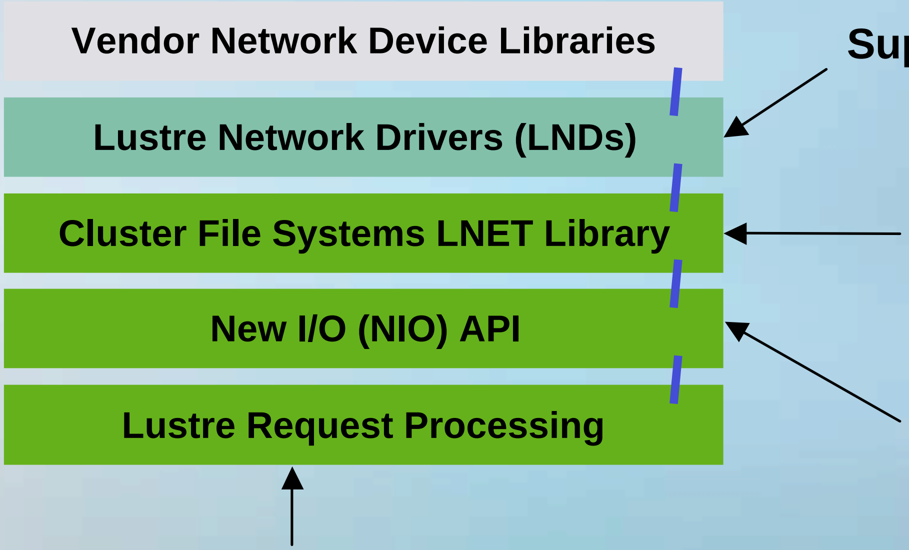
Lustre Deployment Overview


lustre
Lustre Metadata Servers (MDS)
Pool of metadata servers
MDS 1 (active)
MDS 2 (standby)
Lustre Object Storage Servers(OSS) 100's
OSS 1
Commodity Storage Servers
OSS 2
OSS 3
Elan
Myrinet
InfiniBand
Lustre Clients 10's - 10,000's
Simultaneous support of multiple network types
Shared storage enables failover OSS
OSS 4
Router
OSS 5
GigE
OSS 6
Enterprise-Class Storage Arrays & SAN Fabrics
OSS 7
= failover

Network features

- Scalability - network 10,000’s nodes
- Support for multiple networks
  - TCP
  - IB - many flavors
  - Elan3,4
  - Myricom GM, MX
  - Cray Seastar & RA
- Routing nodes between networks

Modular Network Implementation



Vendor Network Device Libraries

Lustre Network Drivers (LNDs)

Cluster File Systems LNET Library

New I/O (NIO) API

Lustre Request Processing

Support for multiple network types
Including routing API

Similar to Sandia Portals with
some new and different features

Move small and large buffers
Use RDMA
Generate events

Zero-copy marshalling libraries
Service framework and request dispatch
Connection and address naming
Generic recovery infrastructure

Key:

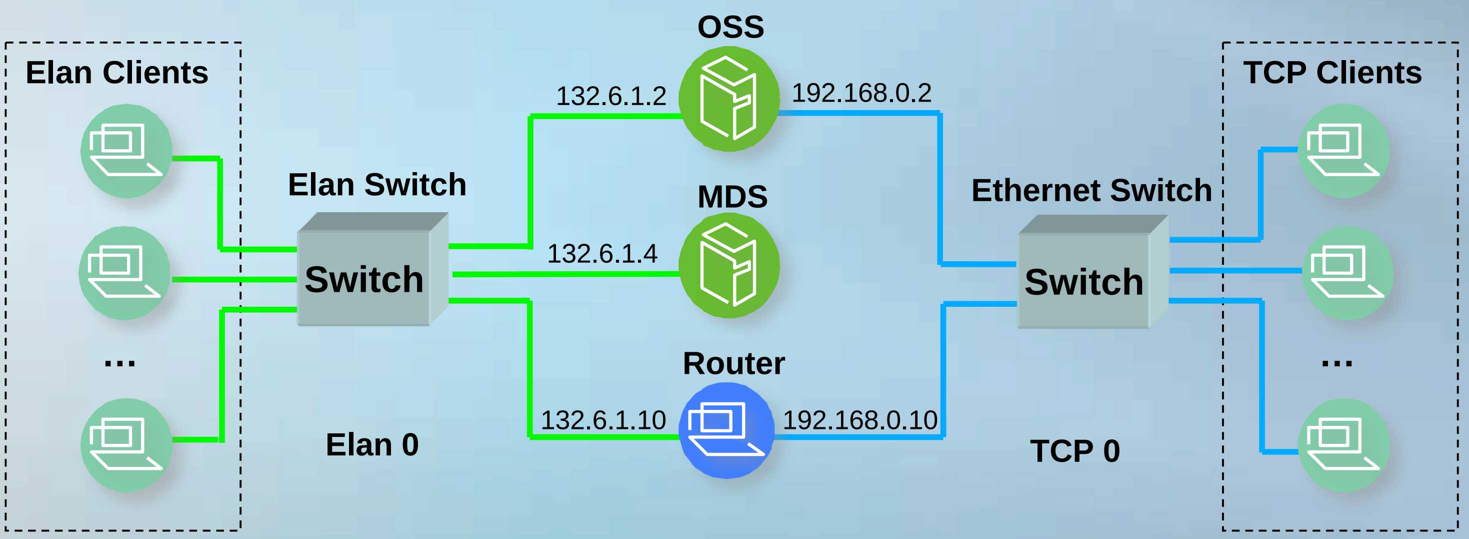
/ Protocol

🟩 Portable Lustre component

🟩 Not portable

□ Not supplied by CFS

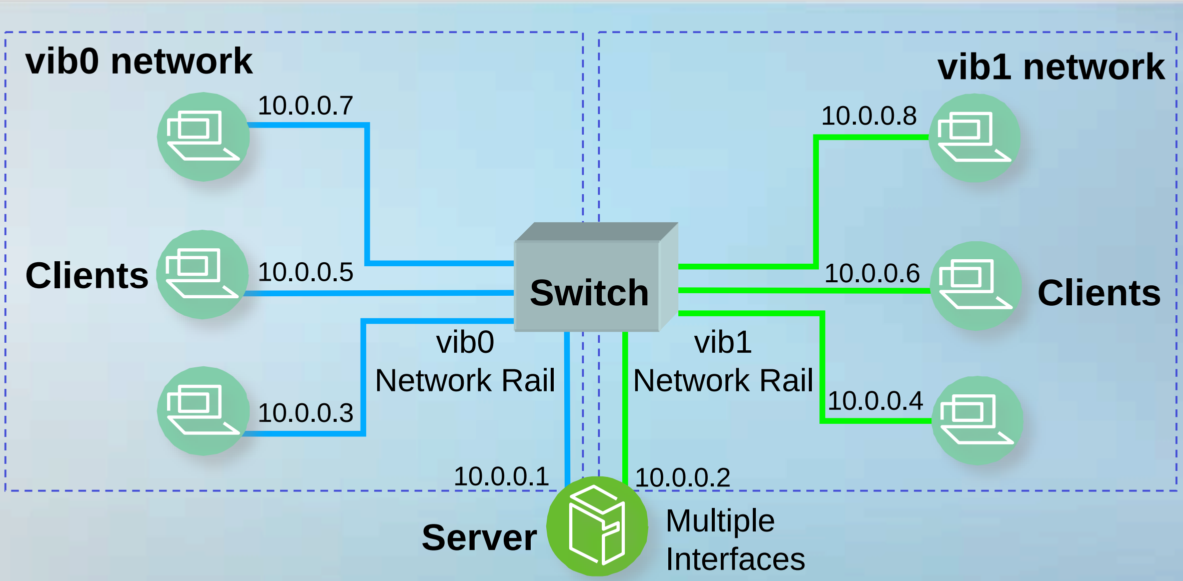
Routing - an example



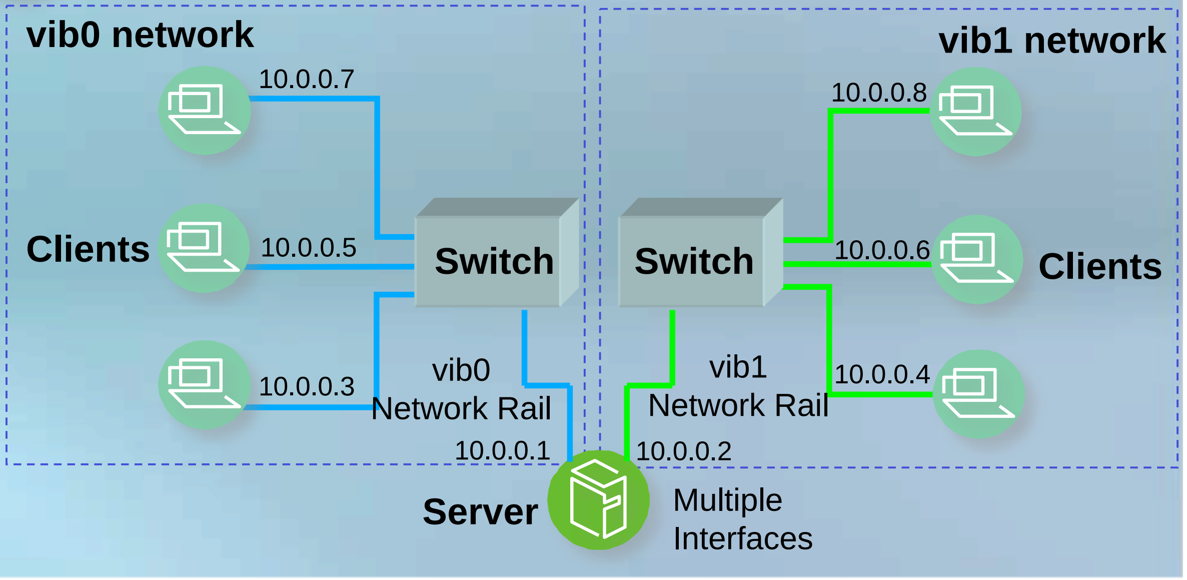
Configuration:

```
options lnet 'ip2nets="tcp0 192.168.0.*; elan0 132.6.1.*"'
   'routes="tcp0 [2,10]@elan0; elan0 192.168.0.[2,10]@tcp0"’
```

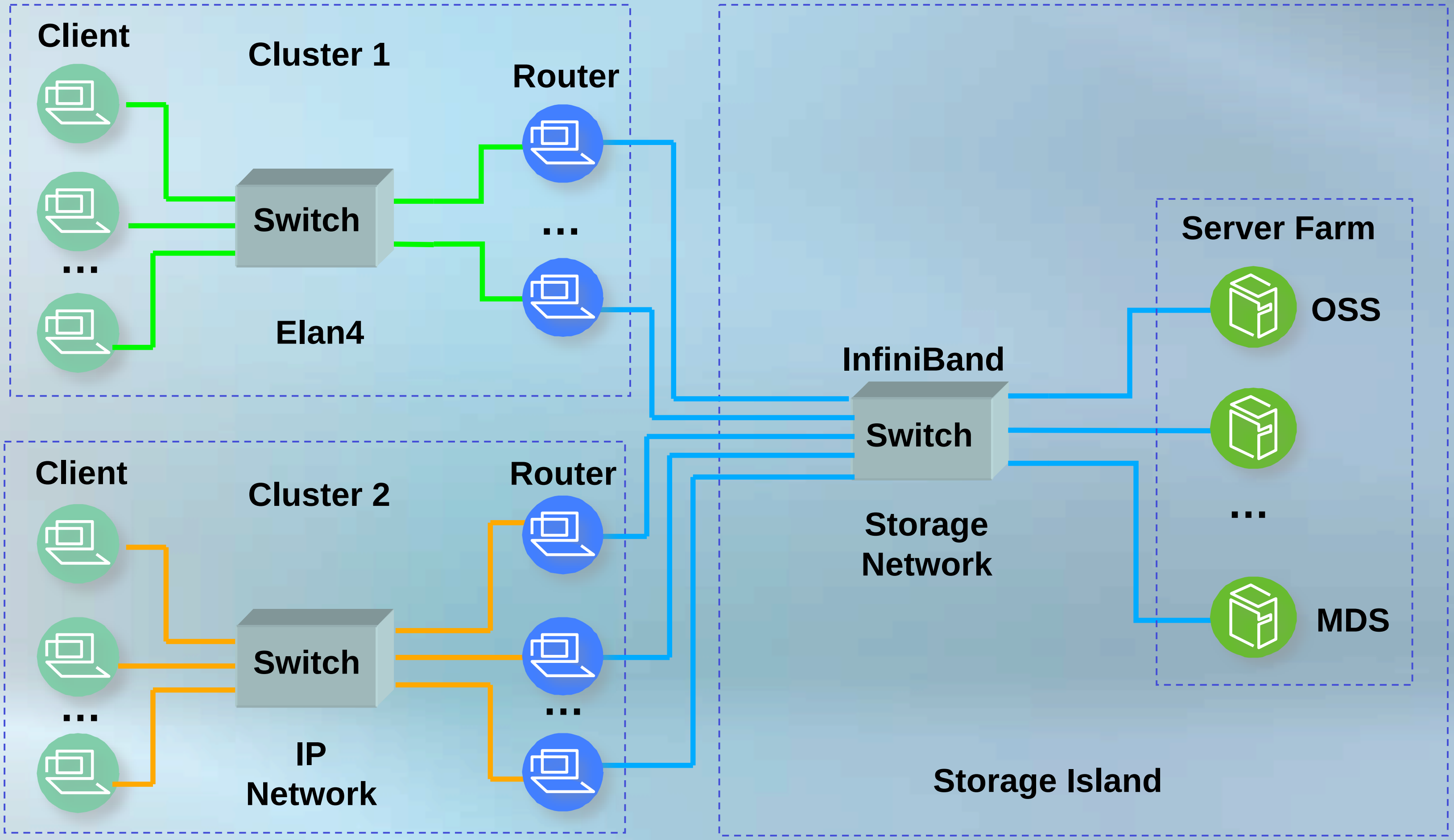
Multiple interfaces and LNET



Support through:
- multiple Lustre networks
- on one or two physical networks
- requires clients to load balance



Site wide file systems



Router features

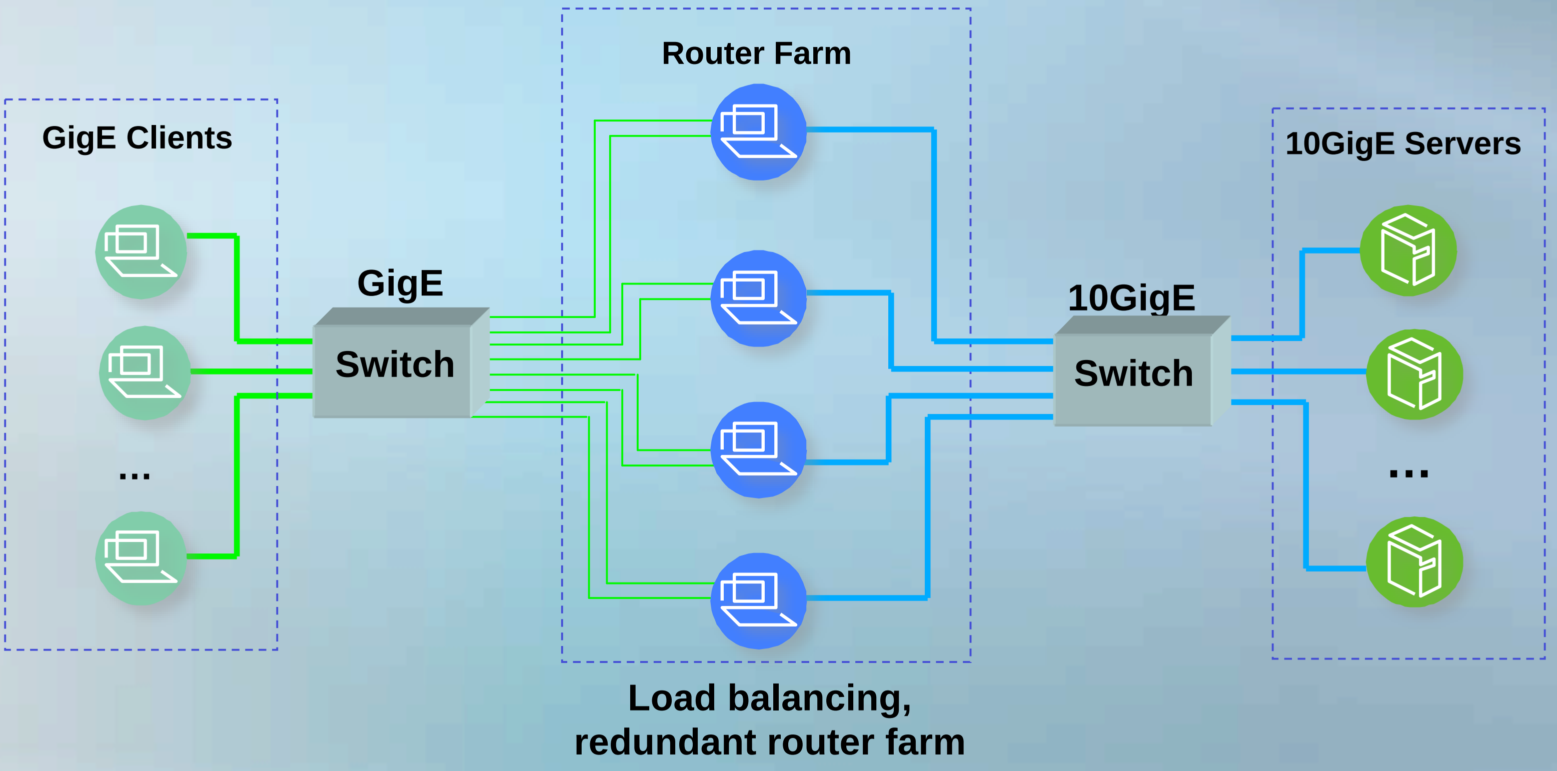
- Redundant routers
- Sophisticated buffer level load balancing
- Failed routers are avoided
- Failed routers are pinged & recoverable

- Future router features may be:
  - Control plane - adjust policies
  - Dynamic router addition
  - All of these require LNET access to the management node

Router uses

- Site wide file systems
  - Build a storage island
  - Add routers with multiple interface types
  - Connect clusters with different interconnects
- Accessing fast servers
  - Build a small fast server network
  - Attach routers for large slower client network
  - Utilize server bandwidth without switches

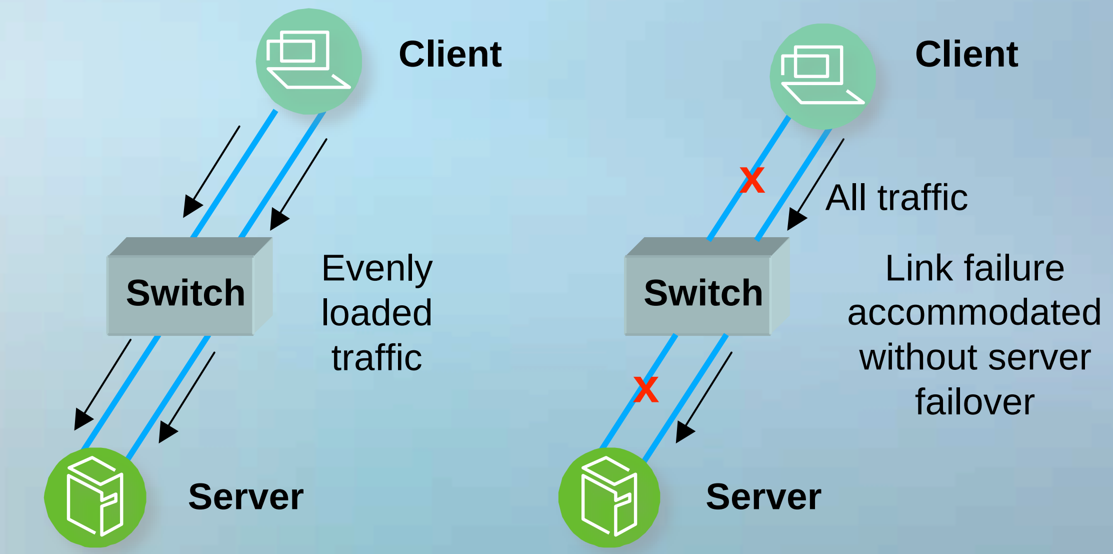
Routers act as 10GigE - 1Gige switch



Multiple interface handling

- Future work
- Desirable features
  - Link level load balancing
  - Link failover
  - N-to-K link handling

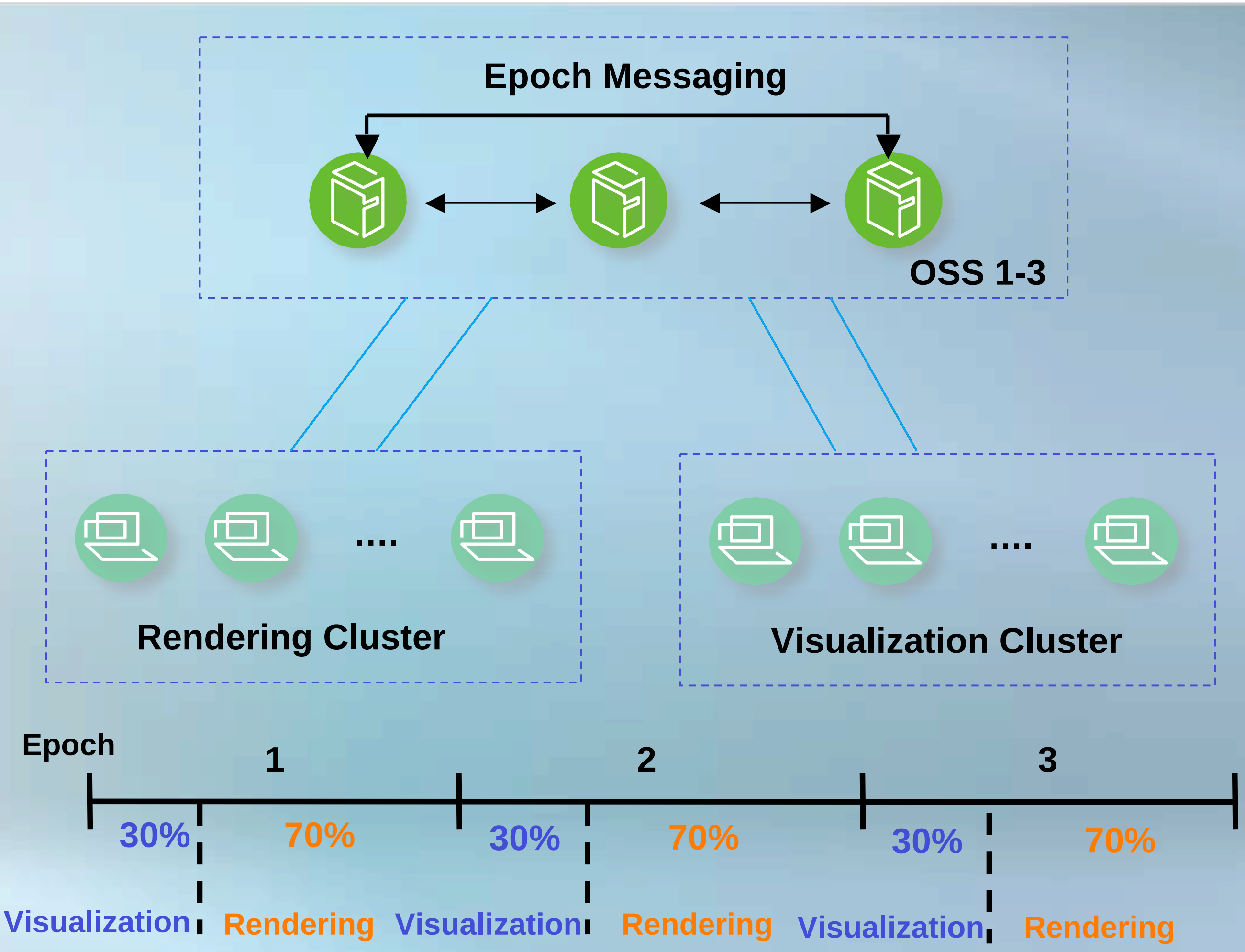
Multiple interface features



Server level load balancing

- Multiple clusters will compete for storage
- Coordinate access to the servers
  - Policy
- Lustre’s task is:
  - Not to decide policy
  - Enable policies

Server level Load balancing



Epoch Messaging

OSS 1-3

Rendering Cluster

....

Visualization Cluster

....

Epoch

1

2

3

30% Visualization

70% Rendering

30% Visualization

70% Rendering

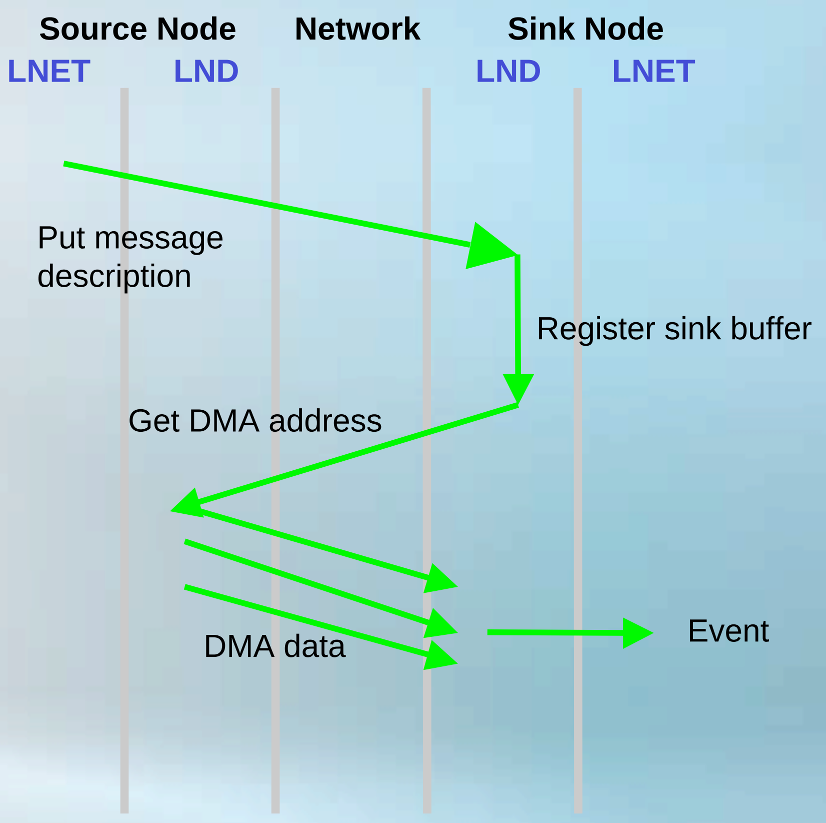
30% Visualization

70% Rendering

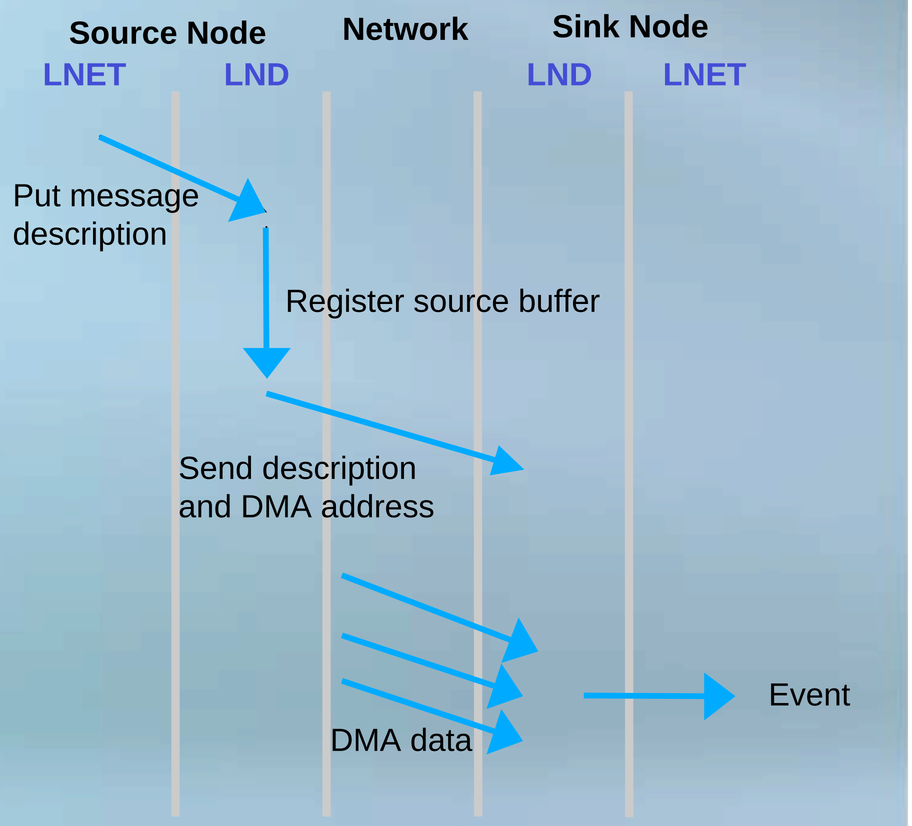
LRS policy allocates 30% of each epoch time slice to visualization and 70% to rendering.

Interrupt free asynchronous IO

- If clients register RDMA buffers before IO
  - Data can be sent or drained while they compute
  - No interrupts on the clients
- Requires LNET changes



Sending message with DMA handshake



Sending message without DMA handshake

Thank you