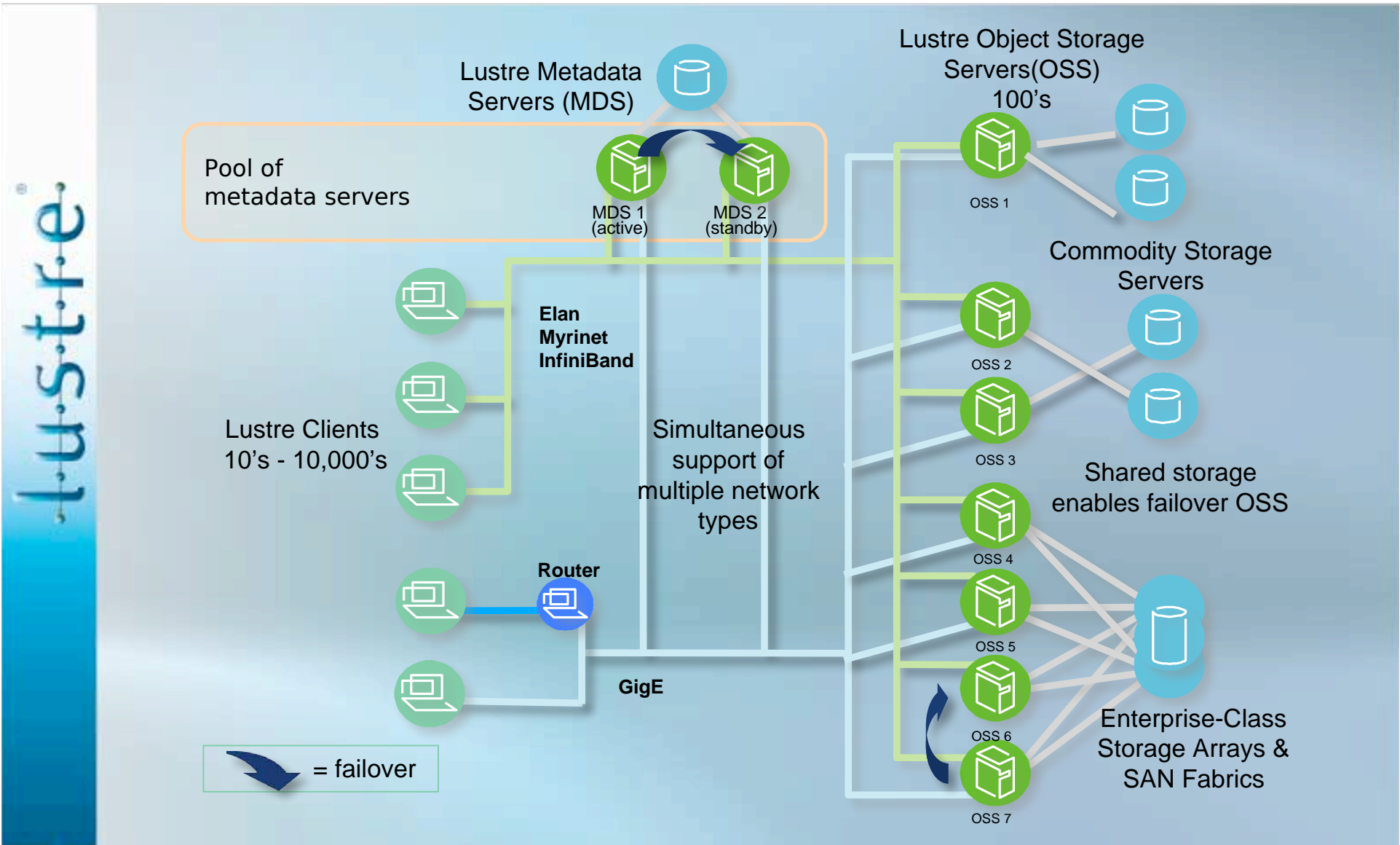




Lustre Networking - an overview

LUG 2007

Lustre Deployment Overview



lustre

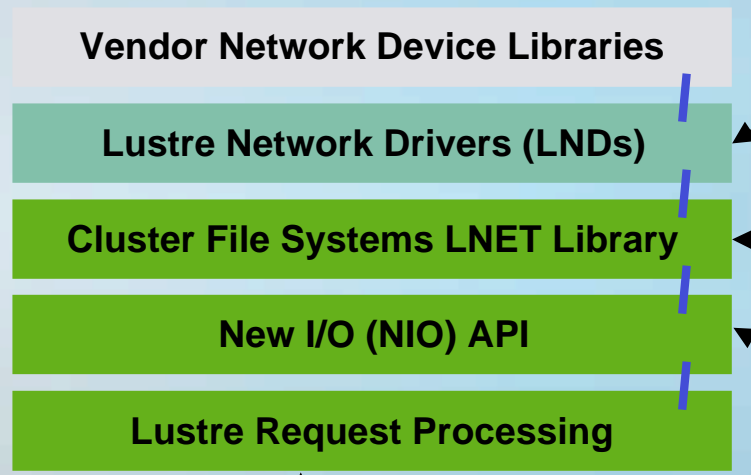
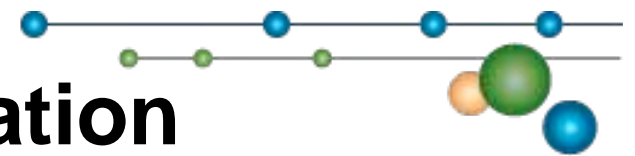


Network features

lustre

- 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



Support for multiple network types
Including routing API

Similar to Sandia Portals with
some new and different features

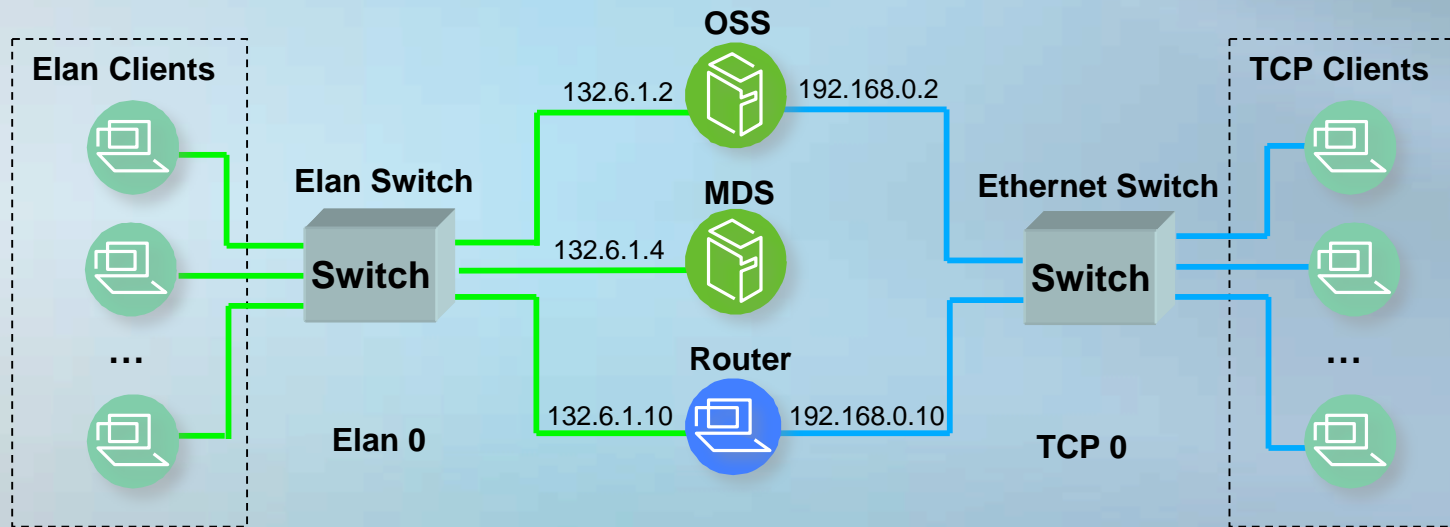
Move small and large buffers
Use RDMA
Generate events

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



lustre

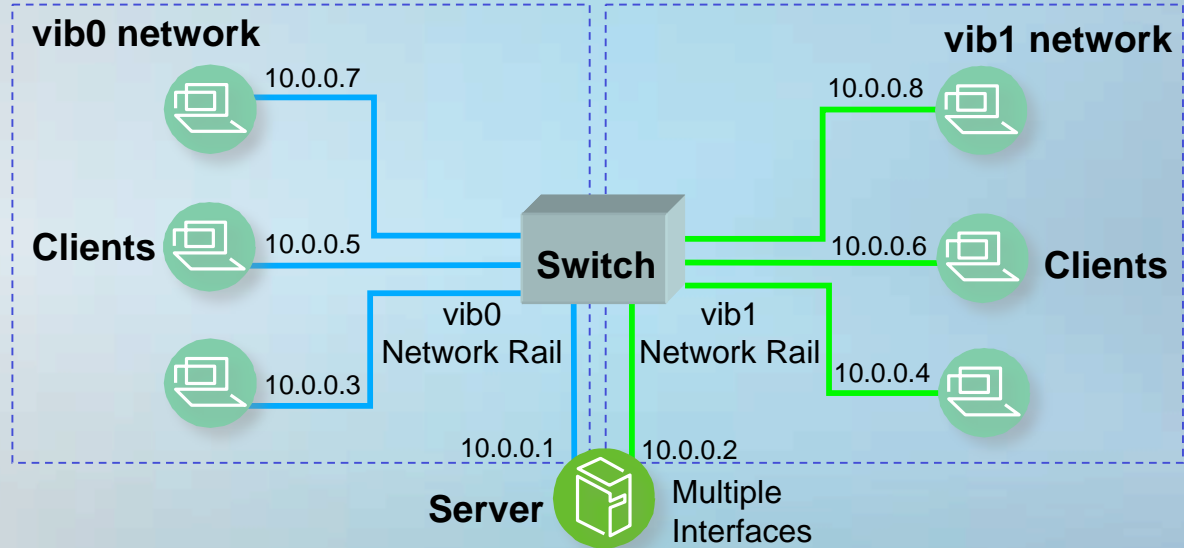
Configuration:

```
options Inet '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

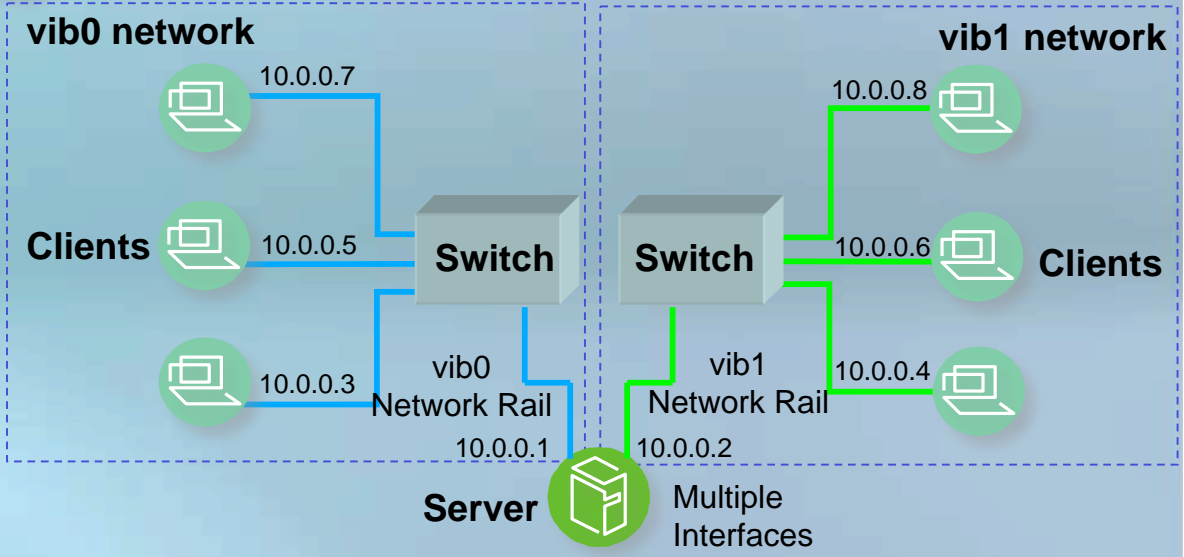


Lustre

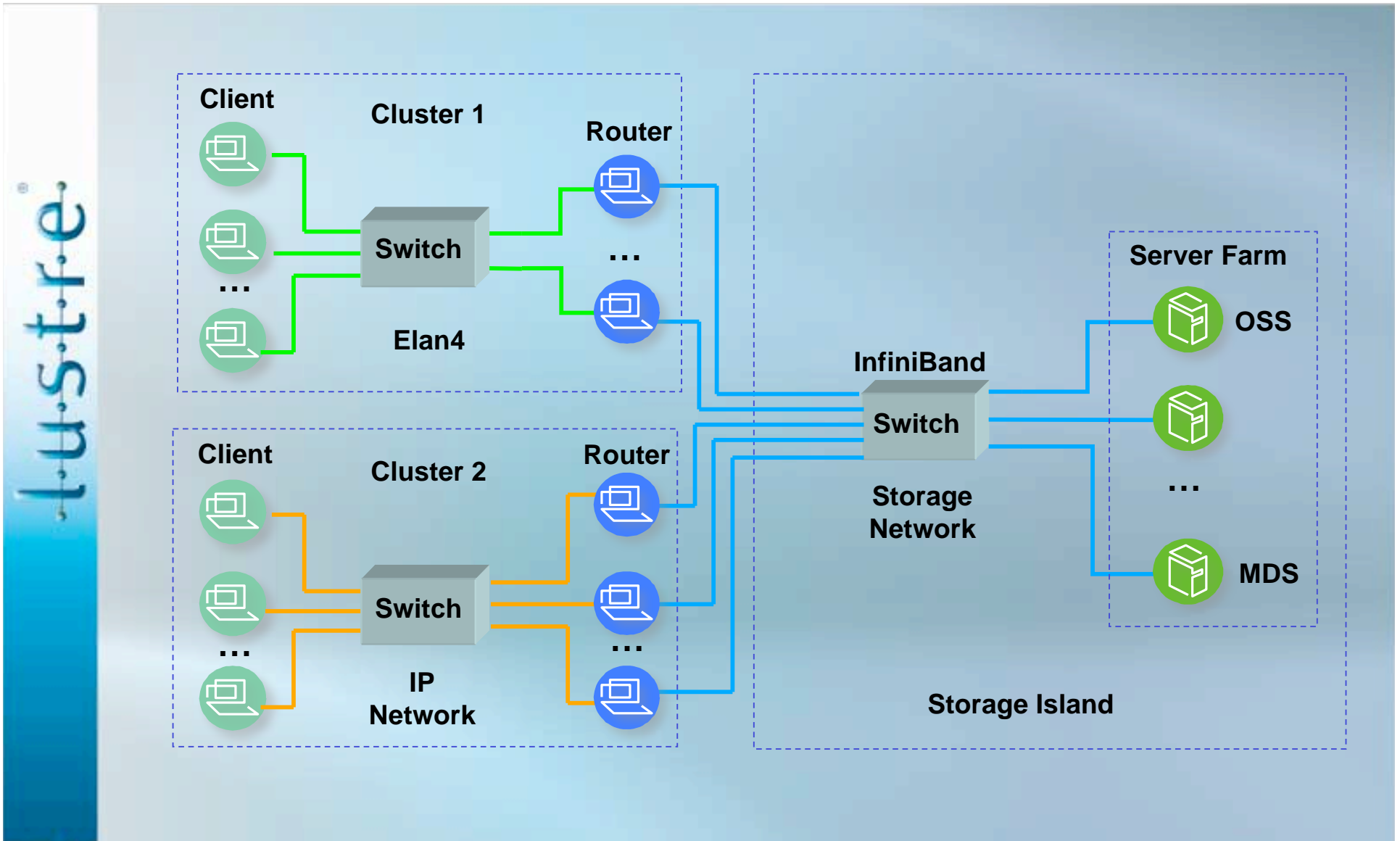


Support through:

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



Site wide file systems



luster



Router features

lustre

- 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

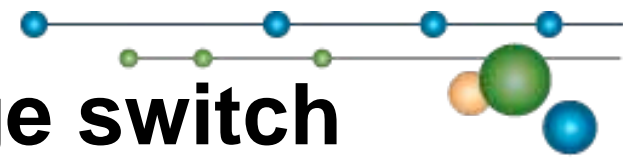


Router uses

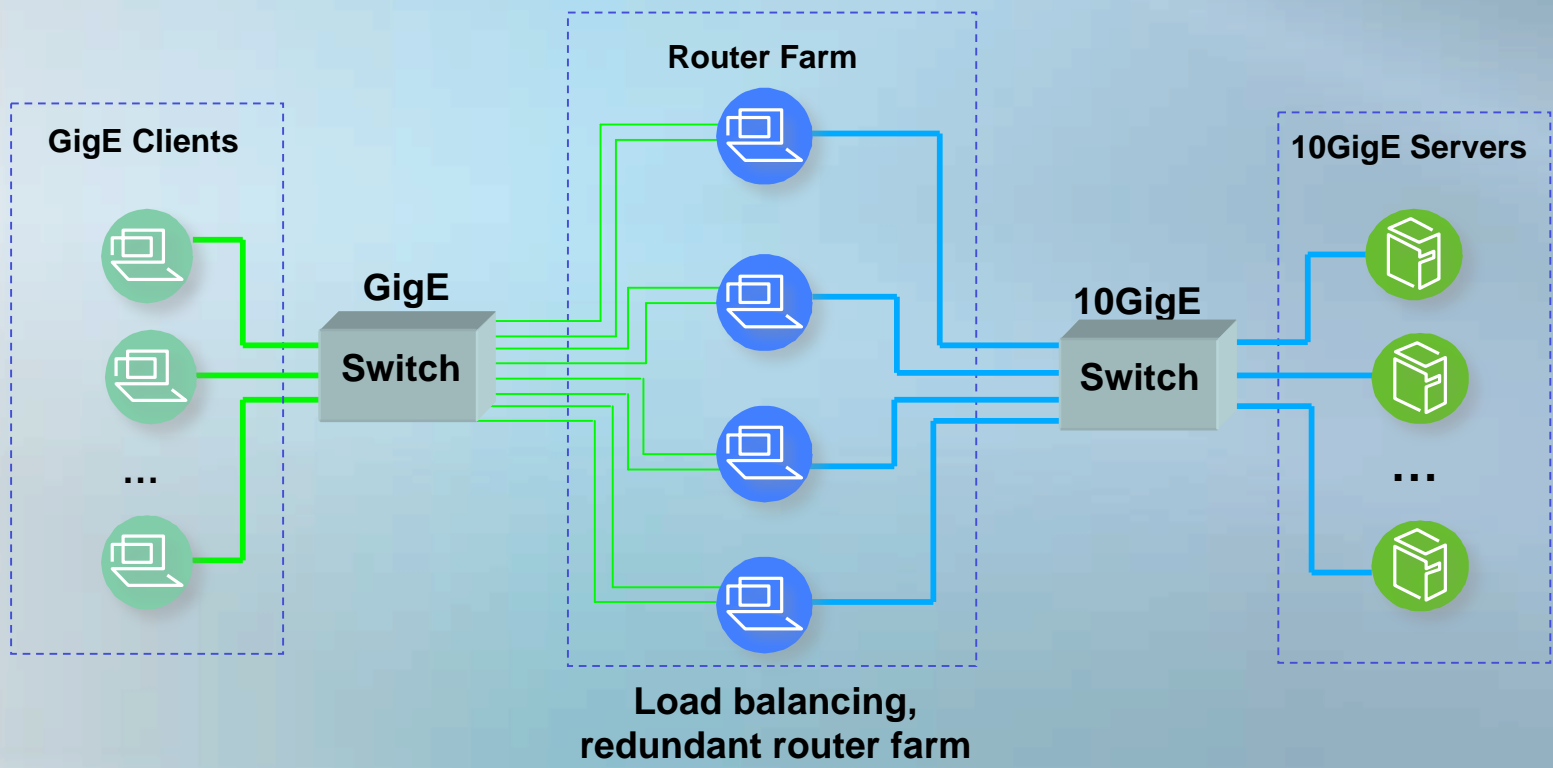
lustre

- 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



lustre



Multiple interface handling



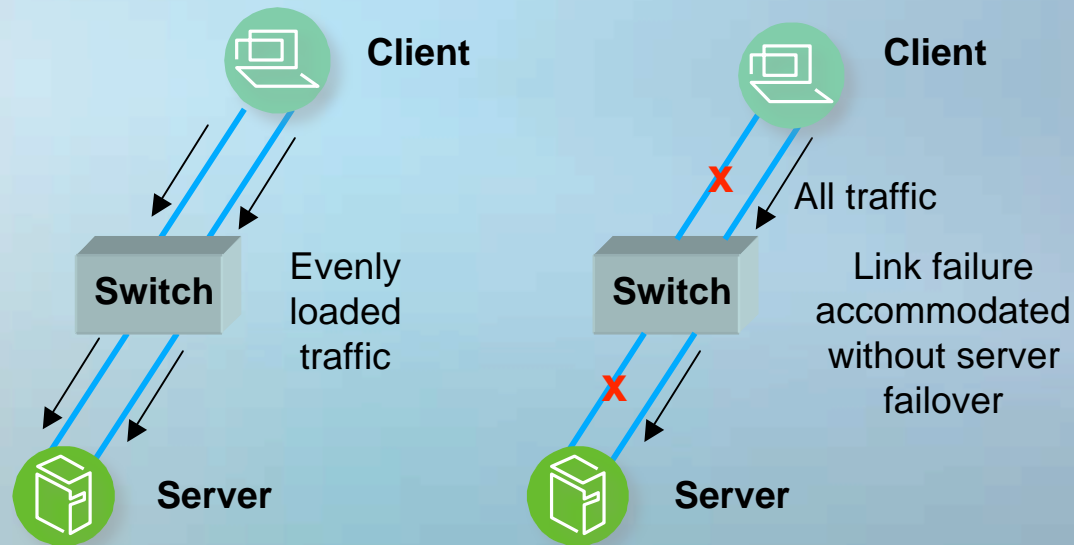
- Future work
- Desirable features
 - Link level load balancing
 - Link failover
 - N-to-K link handling

lustre®

Multiple interface features



lustre



Server level load balancing



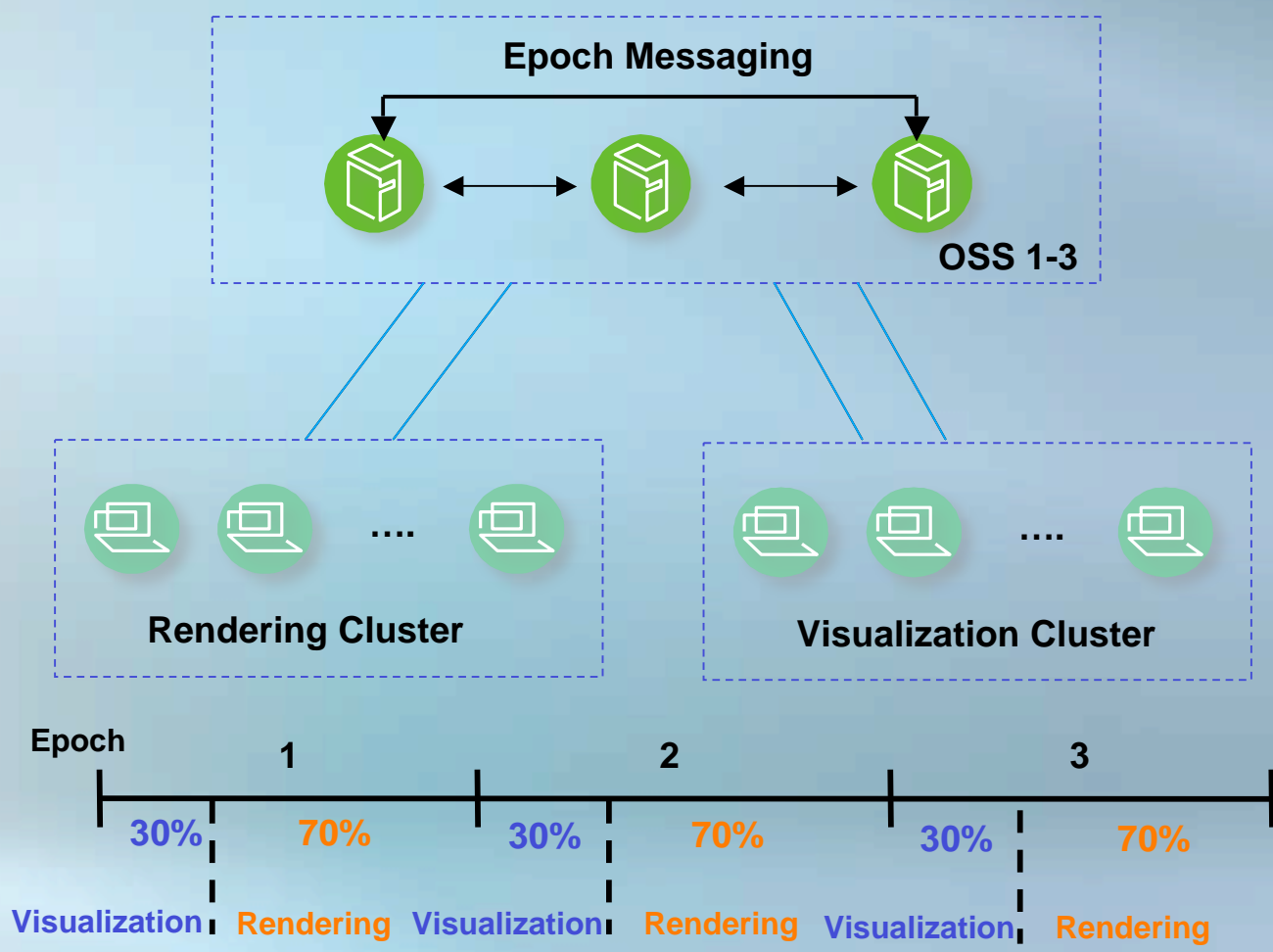
lustre®

- 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



lustre



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

Interrupt free asynchronous IO

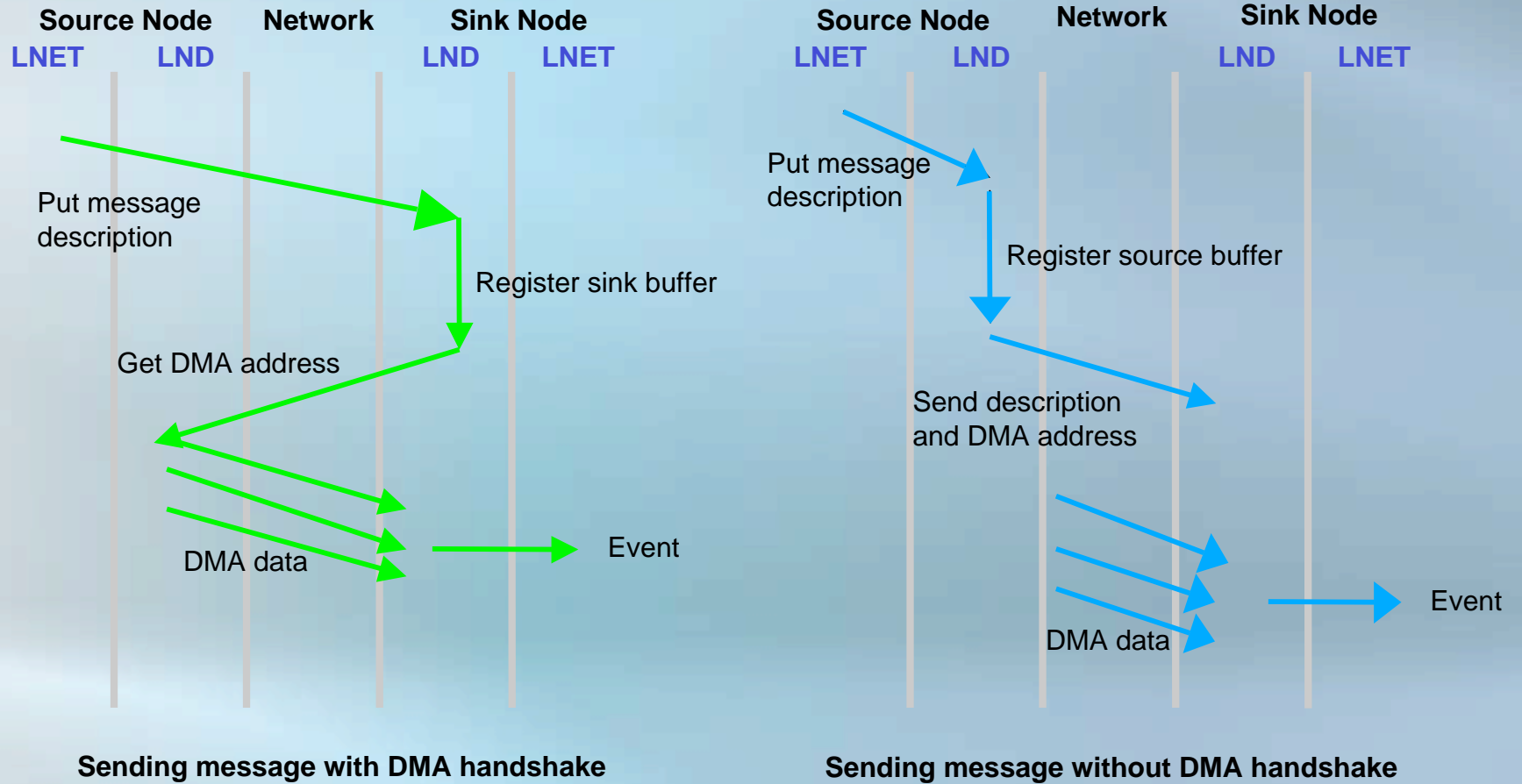


- If clients register RDMA buffers before IO
 - Data can be sent or drained while they compute
 - No interrupts on the clients
- Requires LNET changes

lustre®



lustre



Thank you

