OPENFABRICS ALLIANCE

Lustre*/Lnet Usage Pattern

OpenFabrics Software User Group Workshop

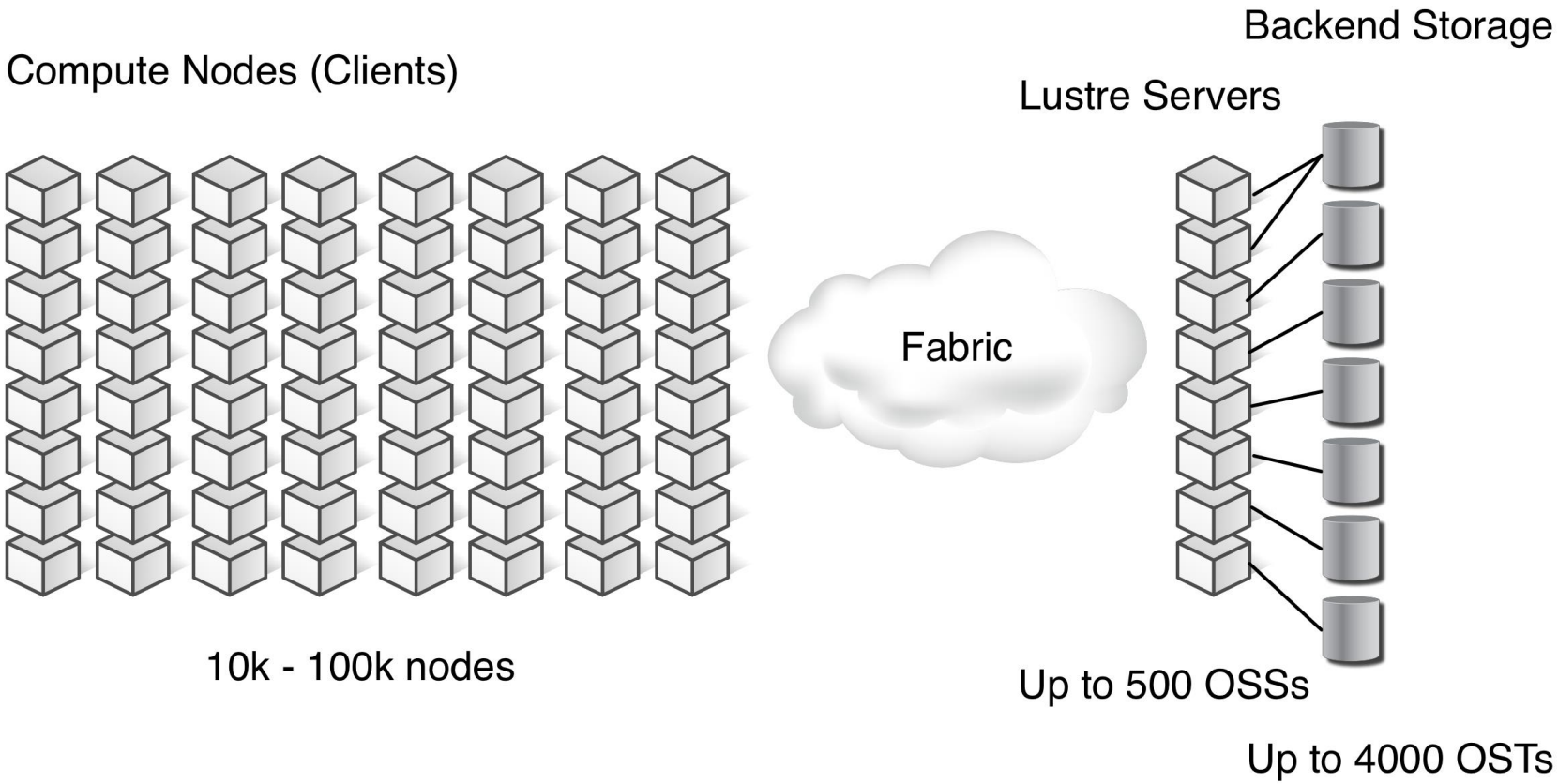
Doug Oucharek, Intel Corp.

#OFAUserGroup

Overview

- Look at how Lustre uses the Fabric and OFED
- LNet Routing
- LNet Configuration Verification
- Channel Bonding
- Monitoring Connections

Lustre Overview



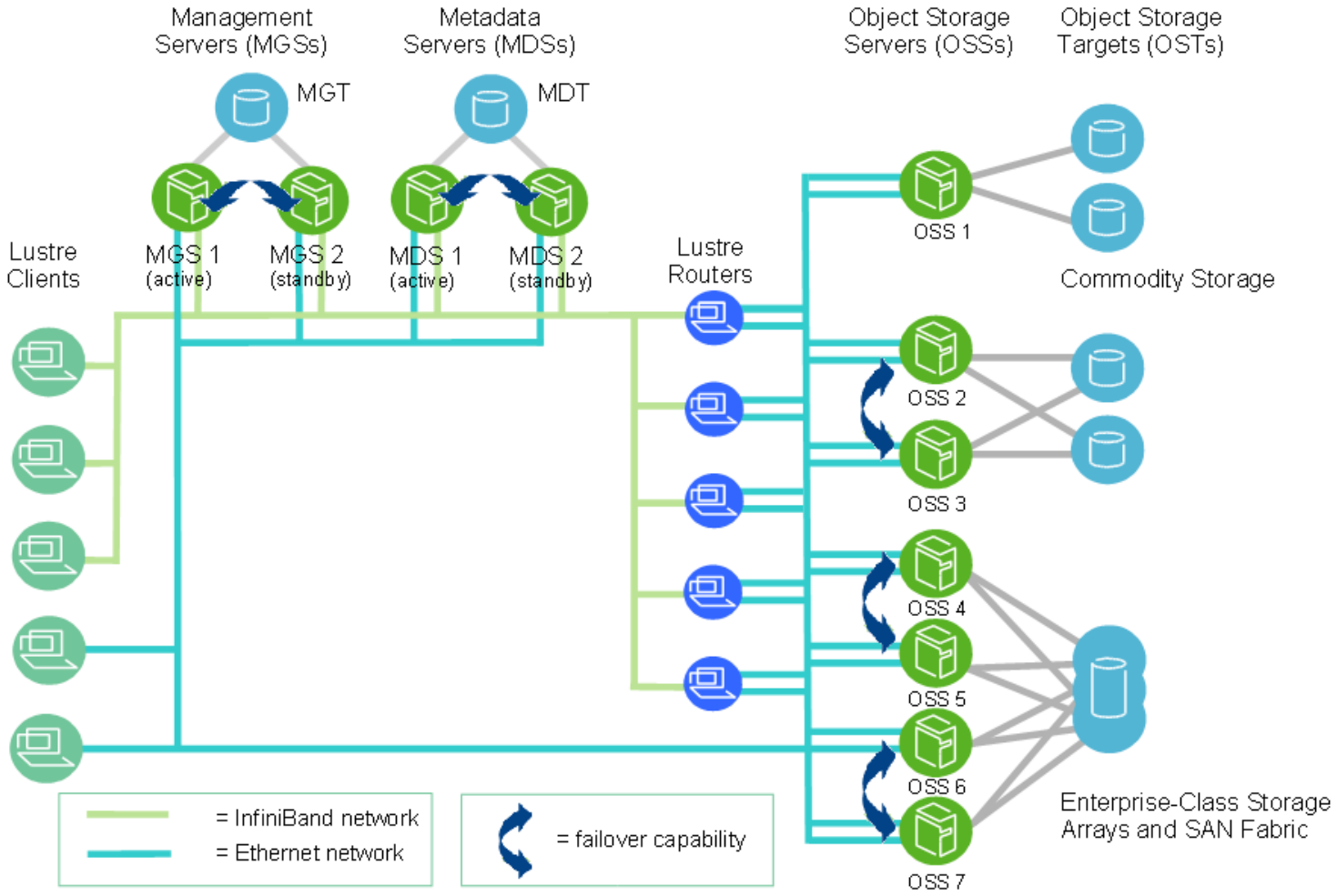
By the Numbers

	2012	2020
Nodes	10-100K	100K-1M
Threads/node	~10	~1000
Total concurrency	100K-1M	100M-1B
Object create	**100K/s**	**100M/s**
Memory	1-4PB	30-60PB
FS Size	10-100PB	600-3000PB
MTTI	1-5 Days	6 Hours
Memory Dump	< 2000s	< 300s
Peak I/O BW	**1-2TB/s**	**100-200TB/s**
Sustained I/O BW	**10-200GB/s**	**20TB/s**

Markets

- HPC
  - Well established
- Enterprise
  - Large Scale Data Analytics (i.e. Oil & Gas)
  - Big Data -> Hadoop Adapter
- Cloud
  - As a series of VMs

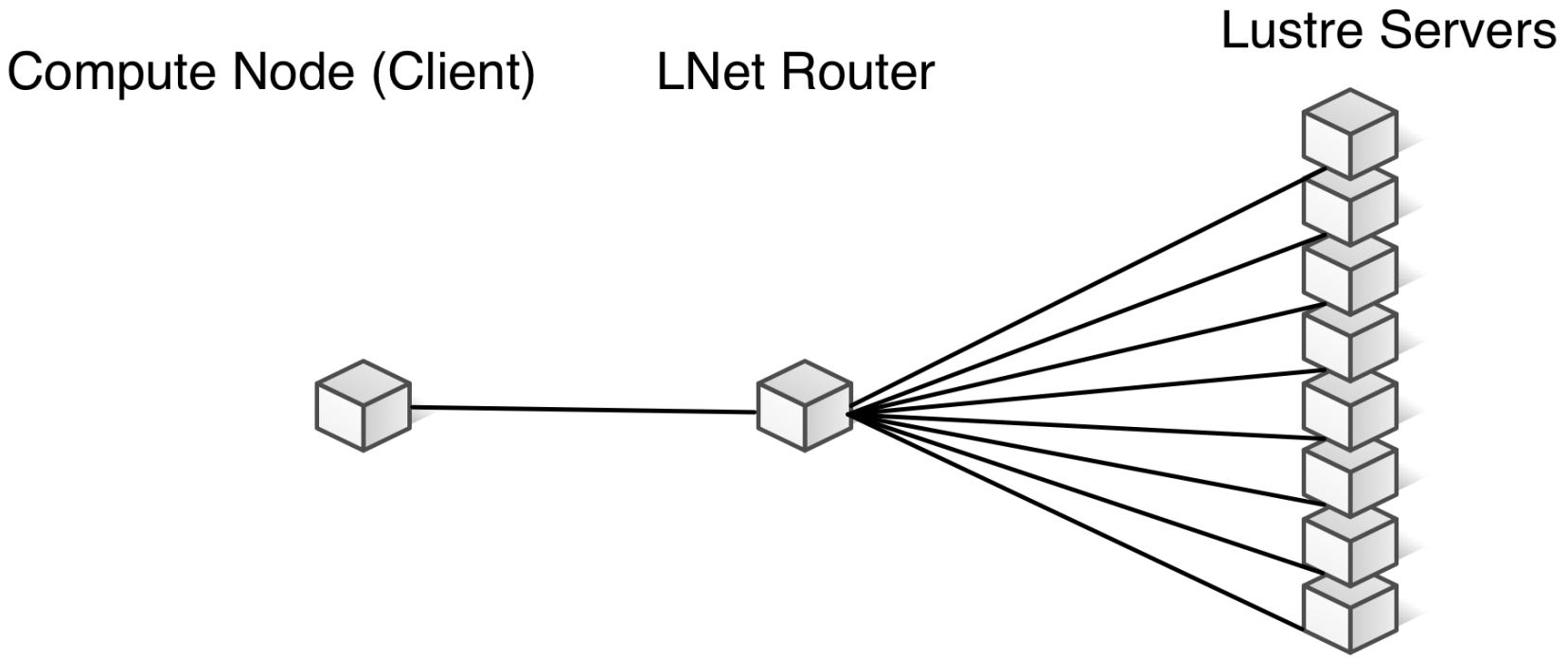
More Details



Usage Patterns

- Striping
  - RAID0 Behaviour
- Peer Connections
  - Created on demand: RC QPs
  - Multiple Server Messages in Parallel
  - Monitor Connections via LNet Ping
- DNE : Multiple Meta-Data Servers
- Self Imposed Flow Control
  - Credits (for network) and Peer Credits (for each connection)
  - Latency is OK, message loss is not

LNet Routing



Why Route?

- From ORNL paper: Network Contention and Congestion Control:Lustre Fine- Grained Routing – Matt Ezell
  - Control **bandwidth** by varying the number of routers
  - Control **I/O paths** by selecting which routers to use
  - Control **route computation** by partitioning fabrics
  - If you are using multiple network types, you have to route traffic

Configuration

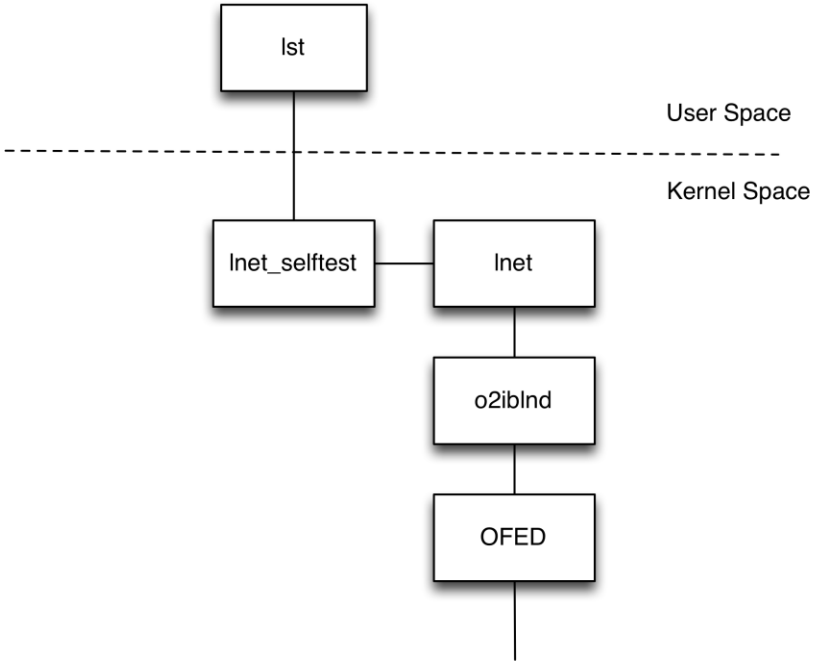
- Server Discovery
  - Use IPv4 (do not use IPoIB)
  - Can not support IPv6 any time soon (see 2012 LUG presentation: LNET Support for IPv6 is Long Overdue – Isaac Huang)
- Currently done via module parameters
- However…

Dynamic LNet Config

- Adding/Deleting networks
- Adding/Deleting routes
- Configuring router buffer pools
- Enabling/Disabling routing.
- Showing routing information
- Importing/exporting configuration in YAML format
- See Lustre 2.7 Manual for details

Testing Configuration

- Start with IB ping
- Then use LNet Ping ("lst ping <nid>")
- Finally, use LNet Selftest
- See last year's presentation for example

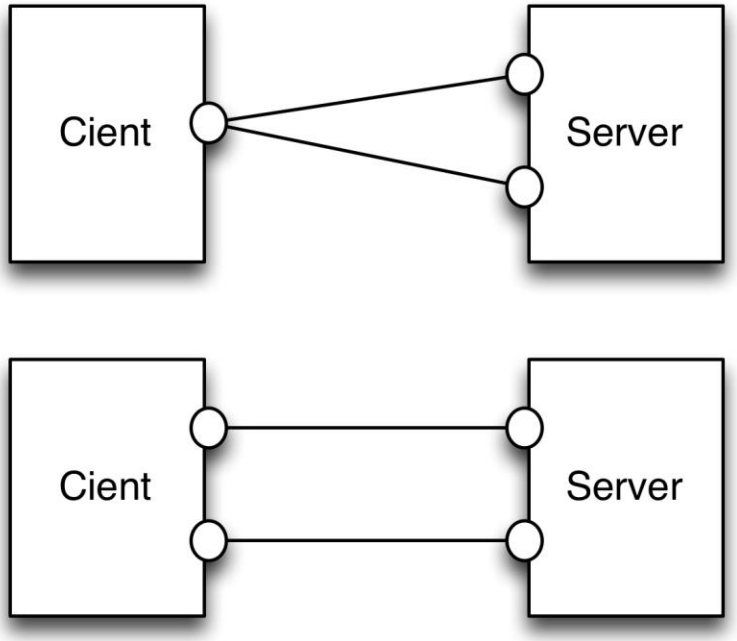


Monitoring Connections

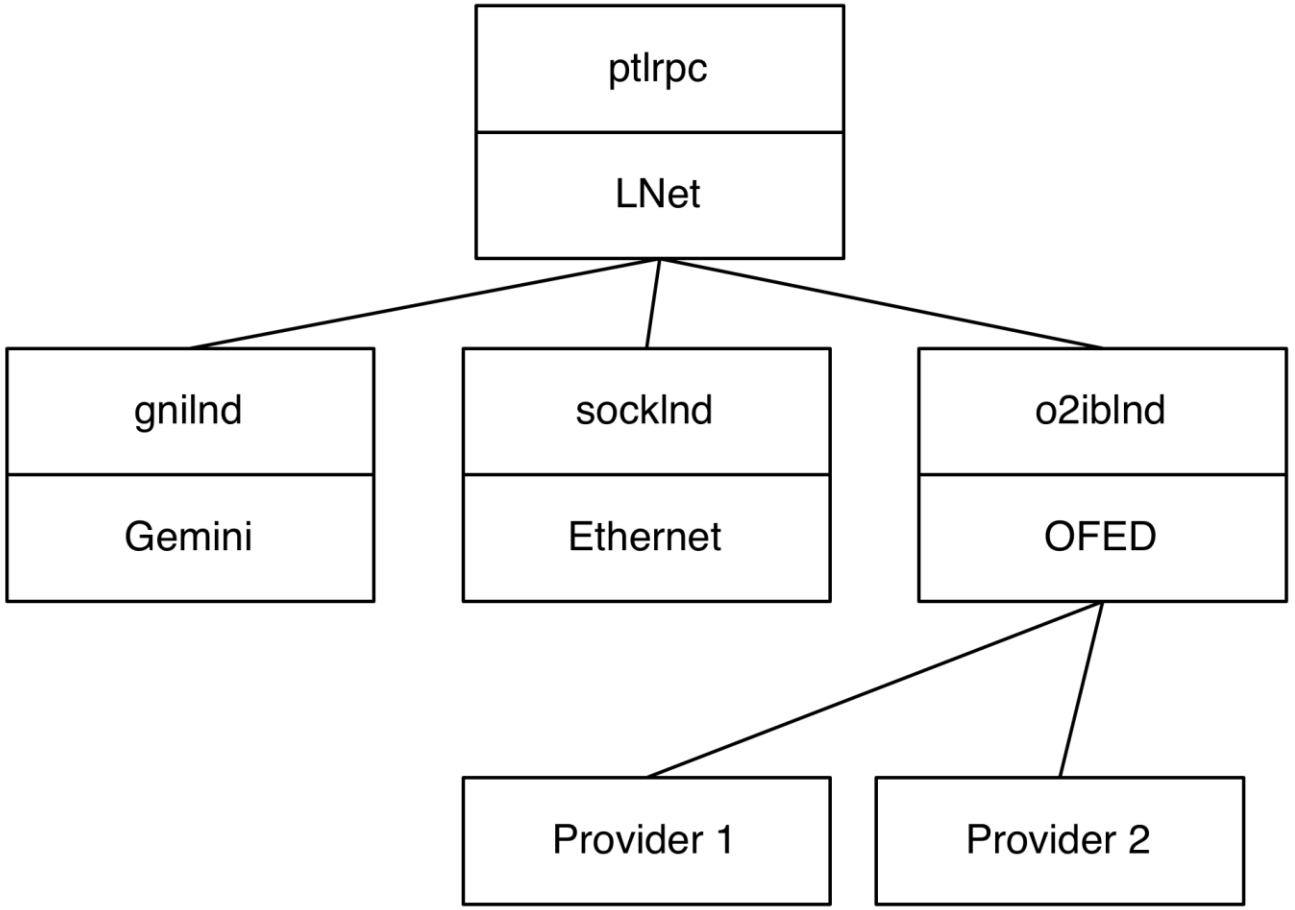
- Two Types of Pings
  - OST Pings
  - Router Pings
- OST Pings not good for scalable fault detection
  - See 2012 LUG paper: Lustre Ping Evictor Scaling in LNET Fine Grained Routing Configurations - Cray
  - Found a 4% - 11% reduction in throughput
  - Example: 25,000 clients, 360 OSSs, 4 OSTs per OSS:
  - 36M pings every 75s
  - Suppress via ptlrpc module parameter: suppress_pings

Channel Bonding

- Coming later this year
- Provides both:
  - Bandwidth
  - Redundancy
- Can Mix Fabrics supported by OFED



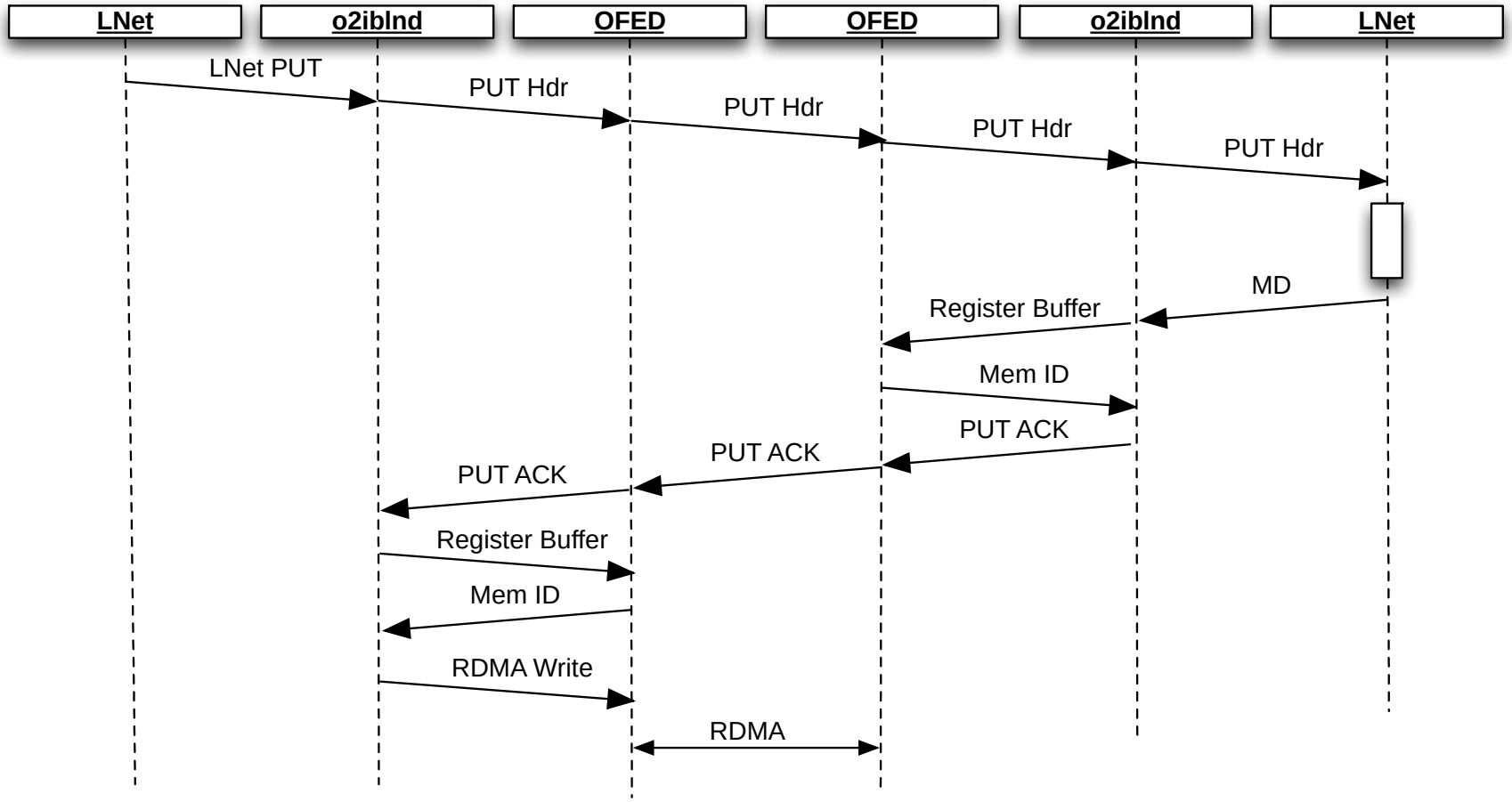
Network Stack



OFED Usage

- Only PUTs Supported in o2iblnd
  - Therefore: only RDMA writes
  - Done to avoid limitations of RDMA reads
- Use Reliable Connections
- Choices:
  - <4k = Use Immediate
  - >4k, <1M = Use RDMA write

Example: Sending a File



Thank You

OPENFABRICS ALLIANCE

OpenFabrics Software
User Group Workshop