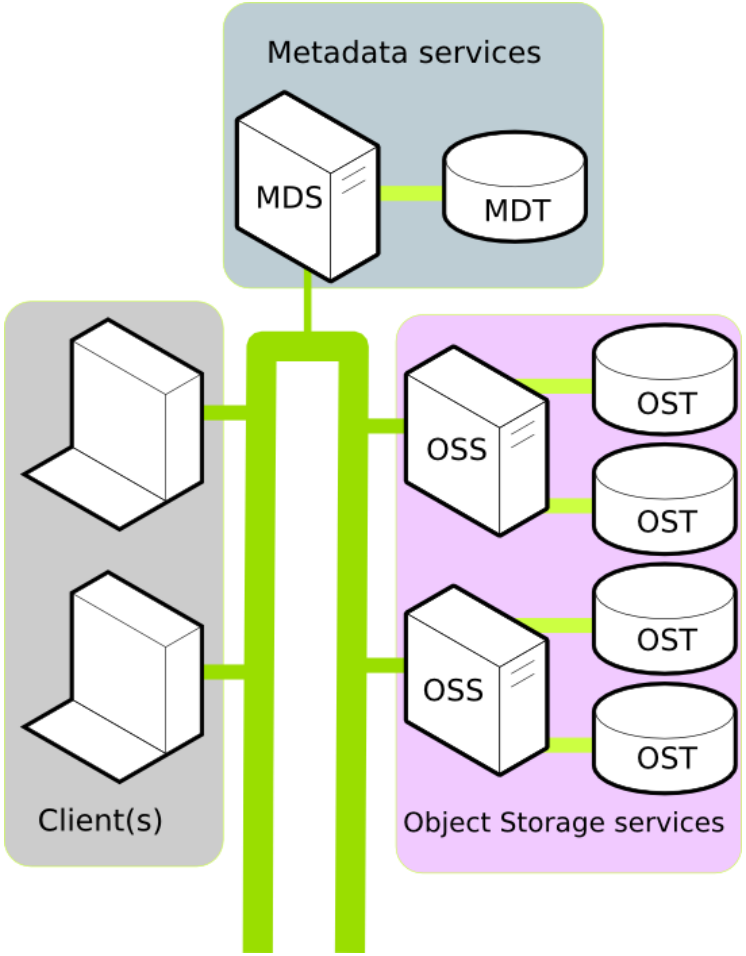
OPENFABRICS ALLIANCE

Lustre* Filesystem

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP

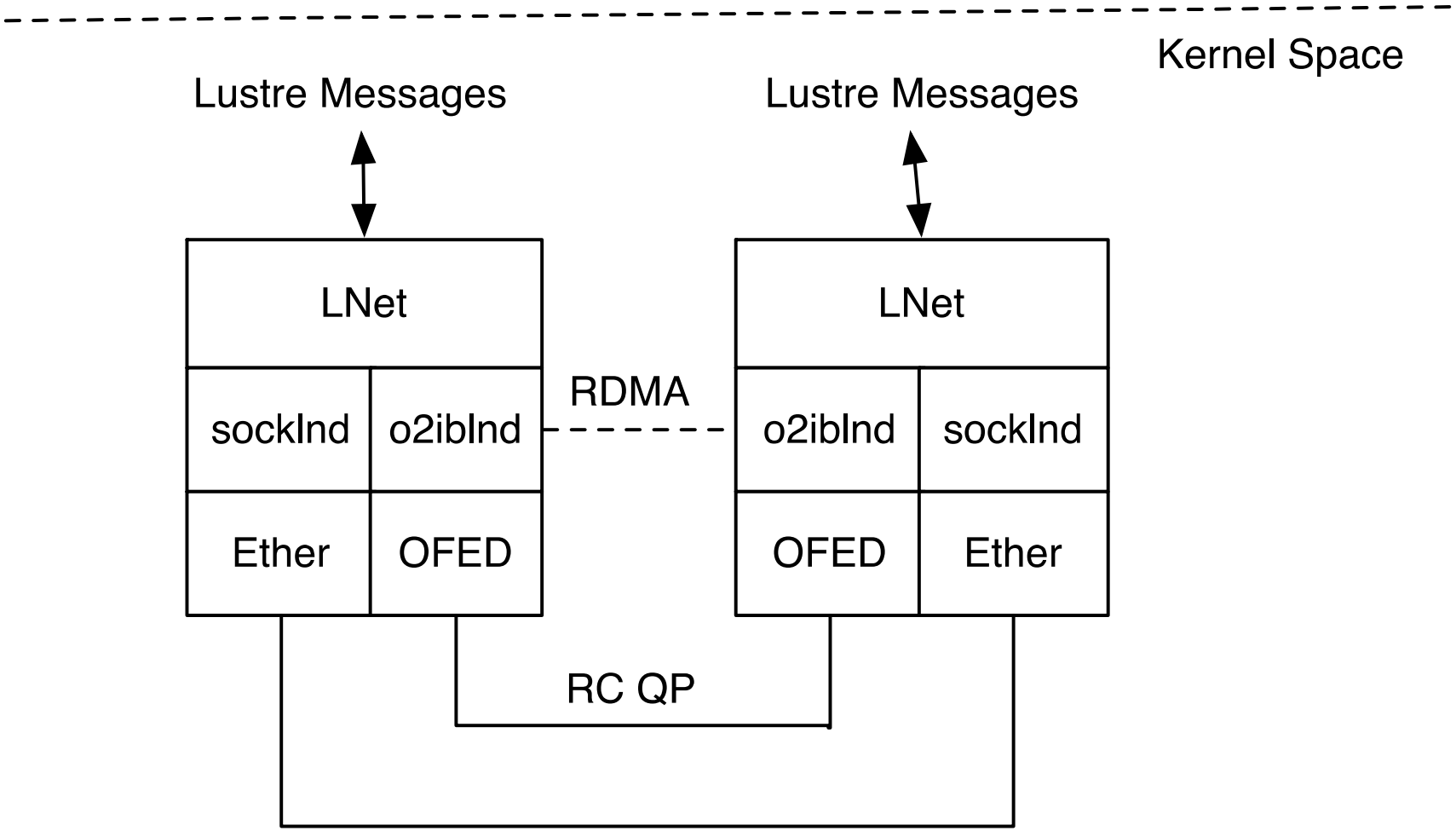
Doug Oucharek
Intel® High Performance Data Division

Lustre and LNet

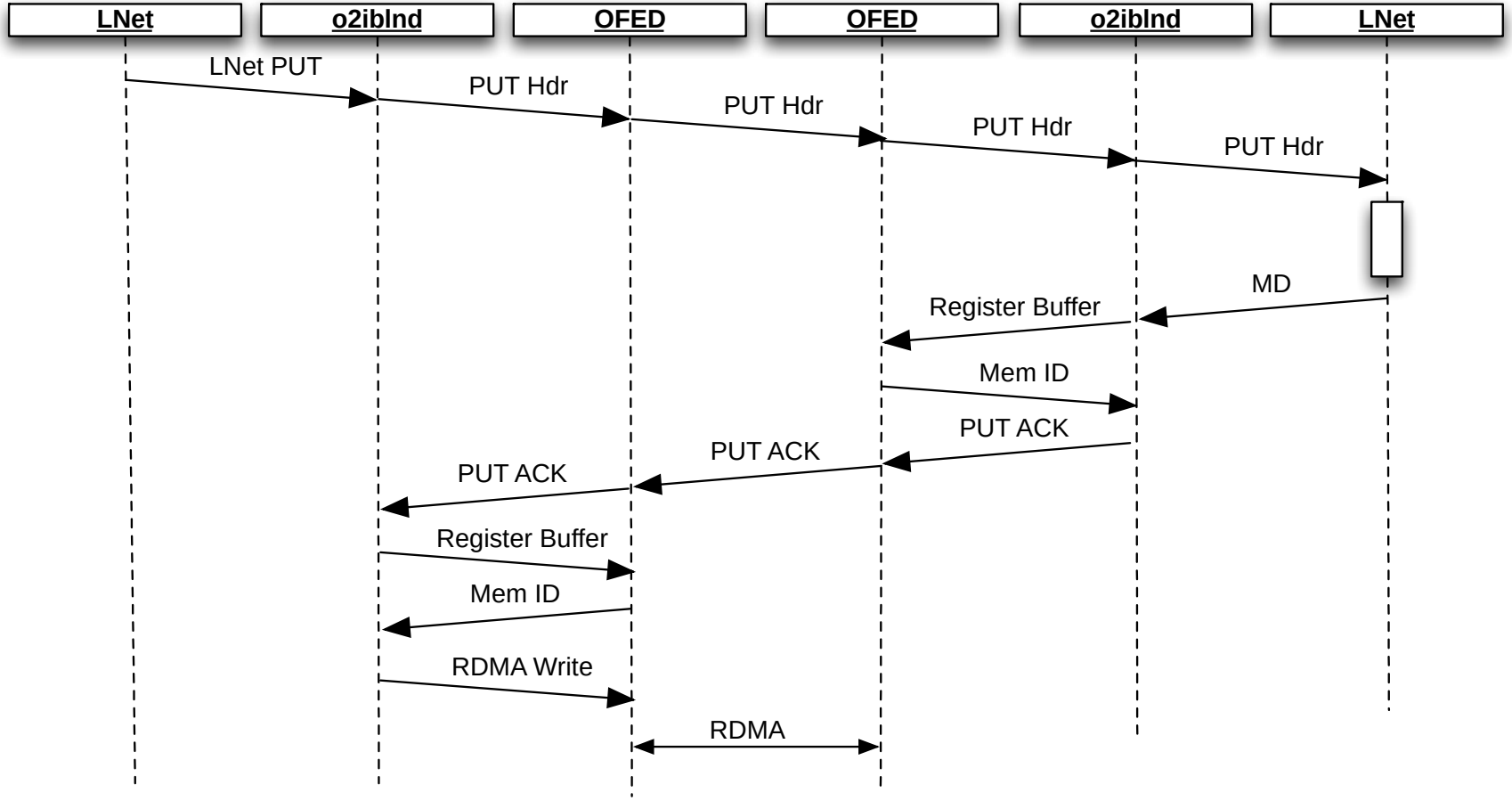


Node-specific Lustre code
ptlrpc
LNet
LND (o2iblnd)
OFED

LNet and OFED

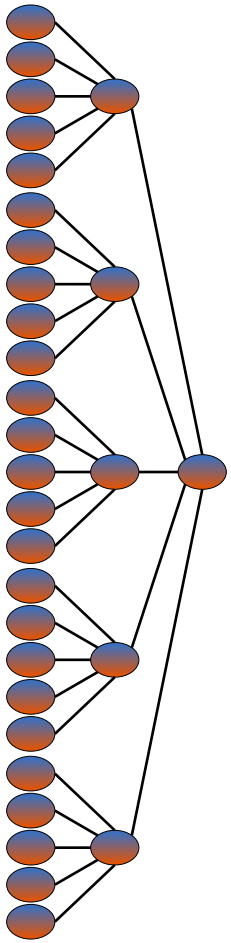


Example: Sending File Data



Health Network: Now and Future

- Today: Peer to Peer pings (O(n))
- Future: Gossip Protocol
- Fault tolerant O(log n) global state distribution latency



What LNet Needs Going Forward

- Reliable Datagram
  - No order preservation required
  - Small max size (e.g. 1k)
  - Receivers are guaranteed eager
  - Senders will adhere to injections limits
  - Network only drops datagrams on transmission failure (dead peer, broken link), not on congestion
- RMA
  - PUT & GET
  - Passive side buffers guaranteed pre-registered (network may discard if no match)
  - 1k <= size <= sensible max (e.g. 1MB)
- High Priority Messages for Health Network
  - Vastly reduced injection rate
  - These messages come before anything else in all devices (i.e. switches)

Thank You

OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP