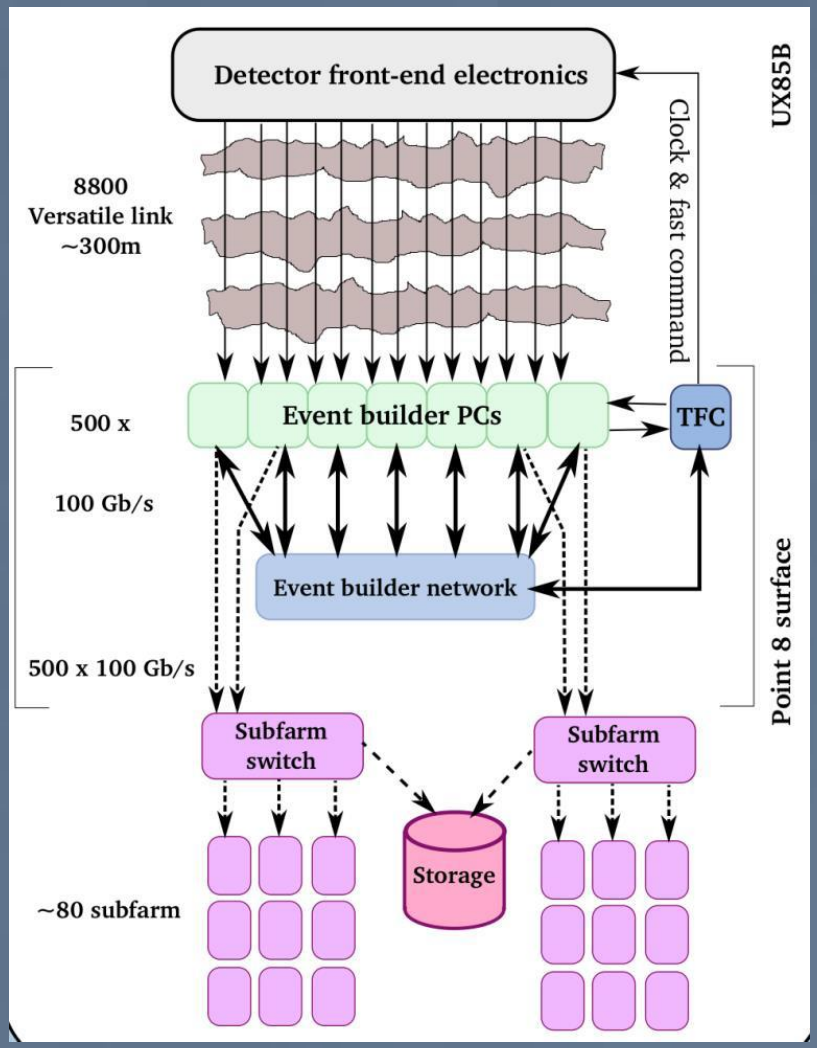
RDMA optimizations on top of 100 Gbps Ethernet for the upgraded data acquisition system of LHCb

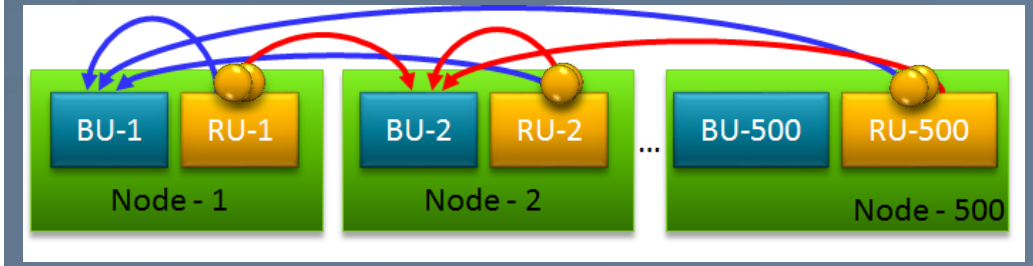
Balazs Voneki
CERN/EP/LHCb Online group

TIPP2017, Beijing
23.05.2017

LHCb Upgrade



- To improve detectors and electronics such that the experiment can run at higher instantaneous luminosity
- Increase the event rate from 1 MHz to 40 MHz
- Selection in software
- Key challenges:
  - Relatively large chunks
  - Everything goes through the network
- RU=Readout Unit
  BU=Builder Unit



Network technologies

Possible 100 Gbps solutions:

- Intel® Omni-Path
- EDR InfiniBand
- 100 Gbps Ethernet

Possible 200 Gbps solution:

- HDR InfiniBand

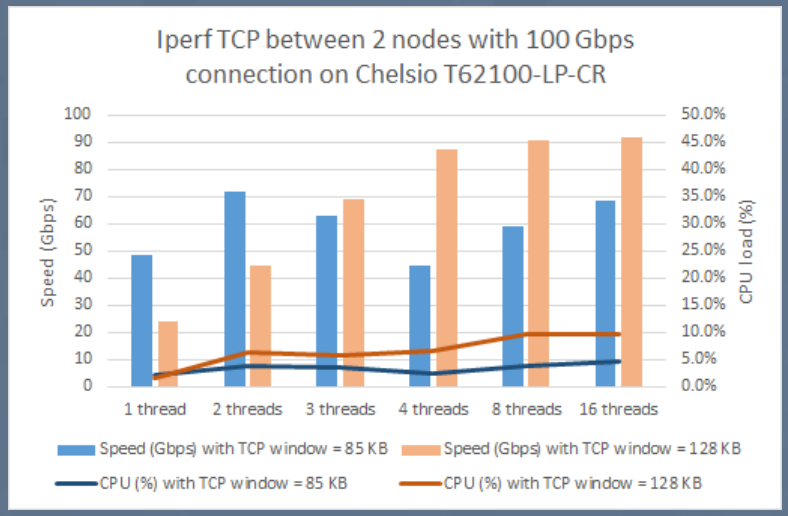
Arguments for Ethernet:

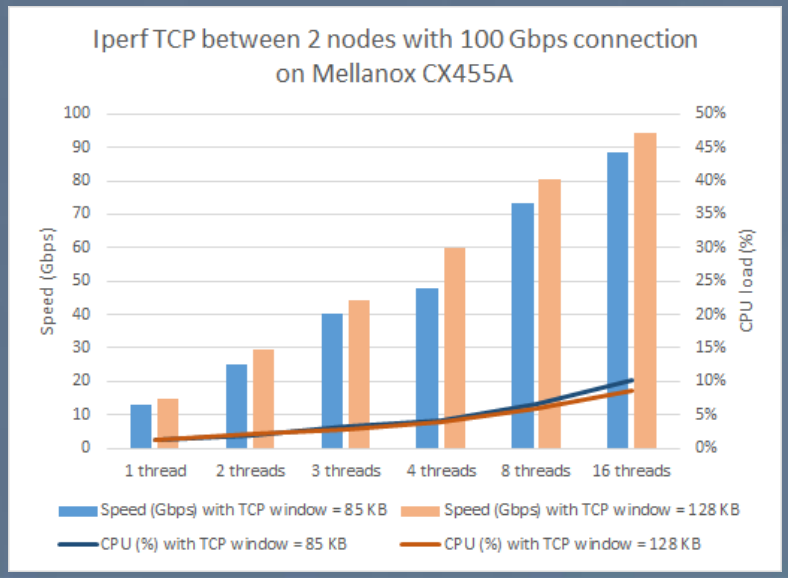
- Widely used
- Old, mature, well-tried
- OPA and IB are single vendor technologies, Ethernet is multi vendor

- Ethernet is challenging at these speeds

Iperf result charts

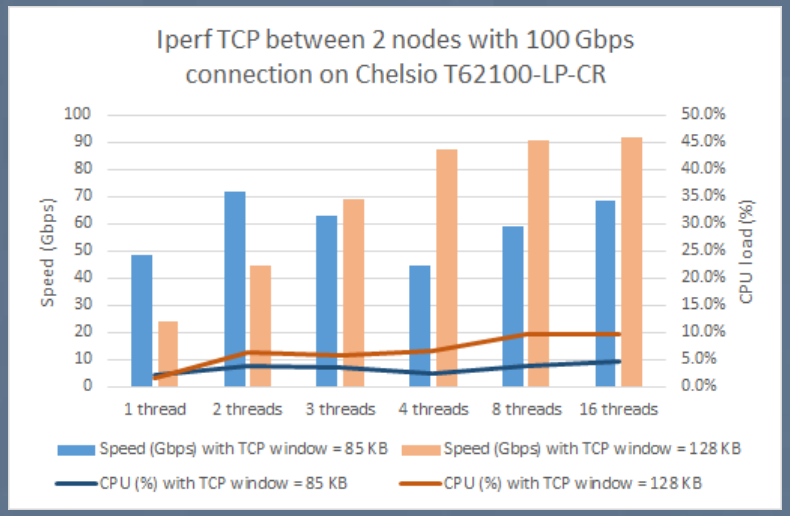
TCP

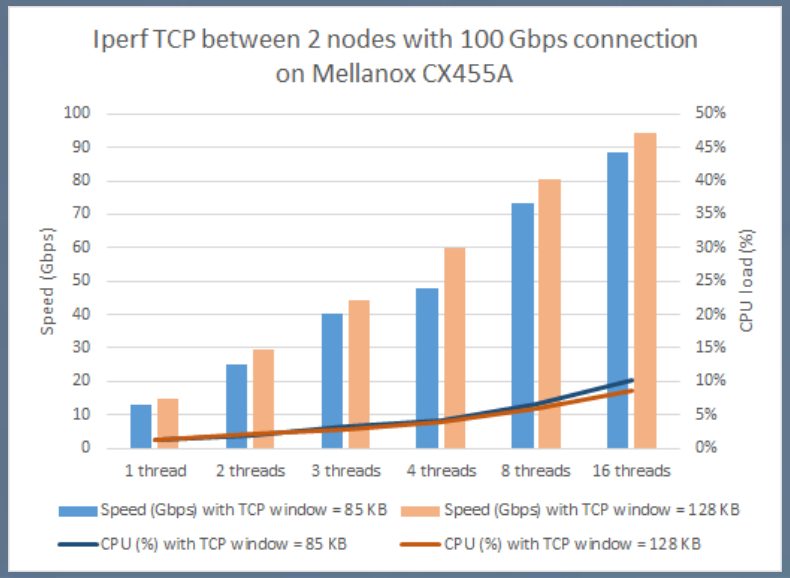




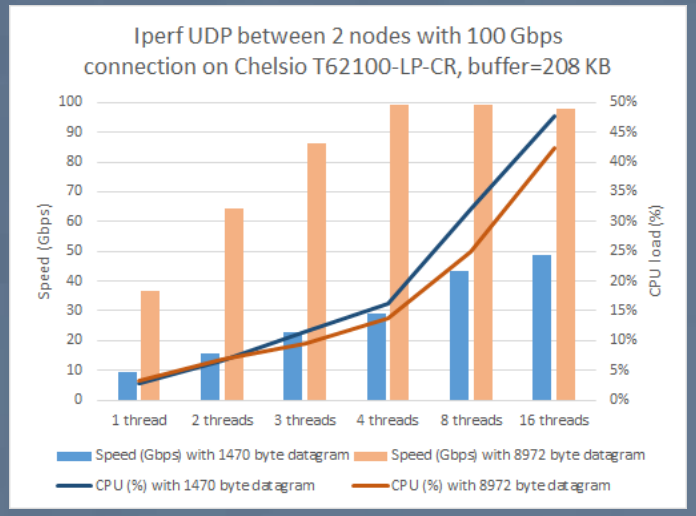
Iperf result charts

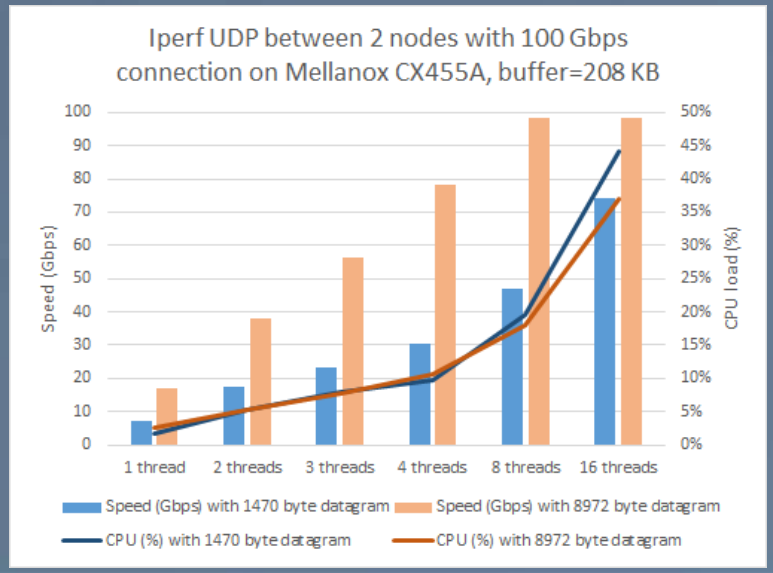
TCP





UDP

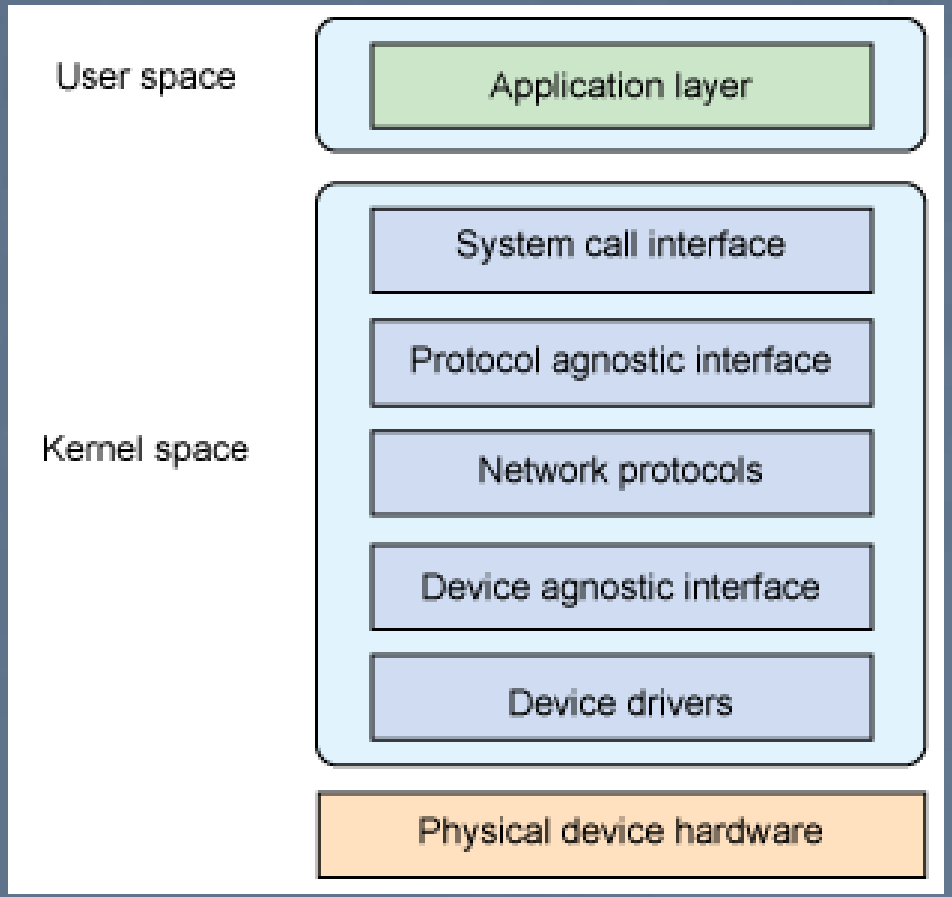




High CPU load

Major difference between vendors

Linux network stack



- *Source: http://www.linuxscrew.com/2007/08/13/linux-networking-stack-understanding/*

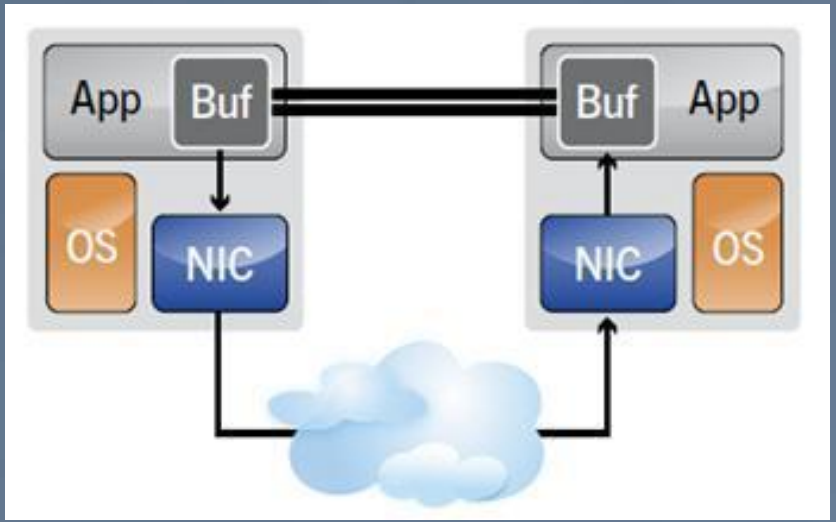
What is RDMA?

- Remote Direct Memory Access
- DMA from the memory of one node into the memory of another node without involving either one's operating system
- Performed by the network adapter itself, no work needs to be done by the CPUs, caches or context switches

Benefits:

- High throughput
- Low latency

These are especially interesting for High Performance Computing!



Source: https://zcopy.wordpress.com/2010/10/08/quick-concepts-part-1-%E2%80%93-introduction-to-rdma/

RDMA Technologies

Available solutions:

- RoCE (RDMA over Converged Ethernet)
- iWARP (internet Wide Area RDMA Protocol)

RoCE needs custom settings on the switch (priority queues to guarantee lossless L2 delivery). iWARP does not need that, only the NICs has to support it.

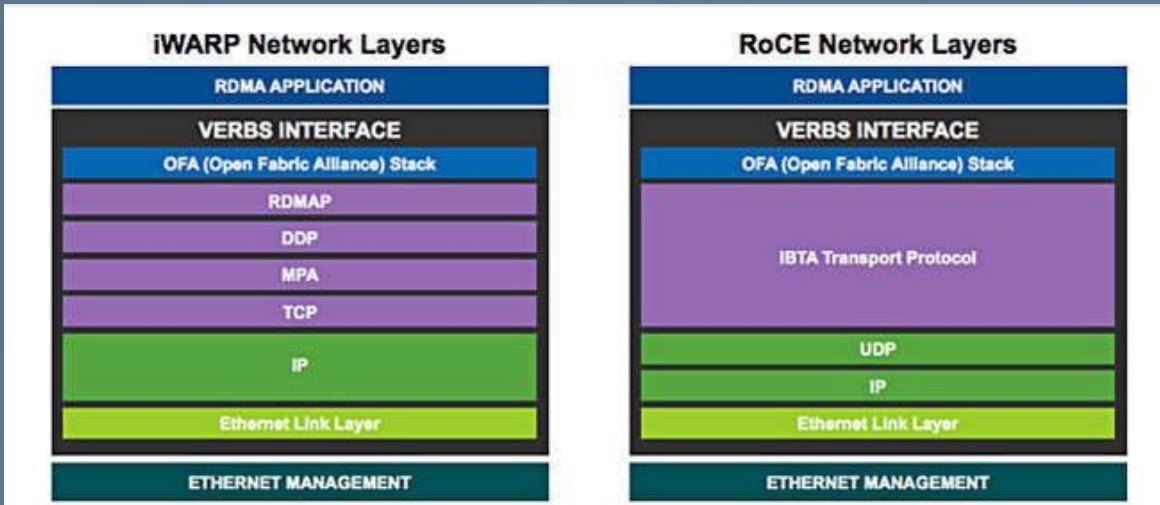


Figure 1. *iWARP's complex network layers vs. RoCE's simpler model*

Source: https://www.theregister.co.uk/2016/04/11/nvme_fabric_speed_messaging/

Test made using:

- Chelsio T62100-LP-CR NIC
- Mellanox CX455A ConnectX-4 NIC
- Mellanox SN2700 100G Ethernet switch

Testbed details

iWARP testbench elements:

4 nodes of:

- Dell PowerEdge C6220
- 2x Intel® Xeon® CPU E5-2670 at 2.60GHz (8 cores, 16 threads)
- 32 GB DDR3 memory at 1333 MHz
- Chelsio 100G T62100-LP-CR NIC

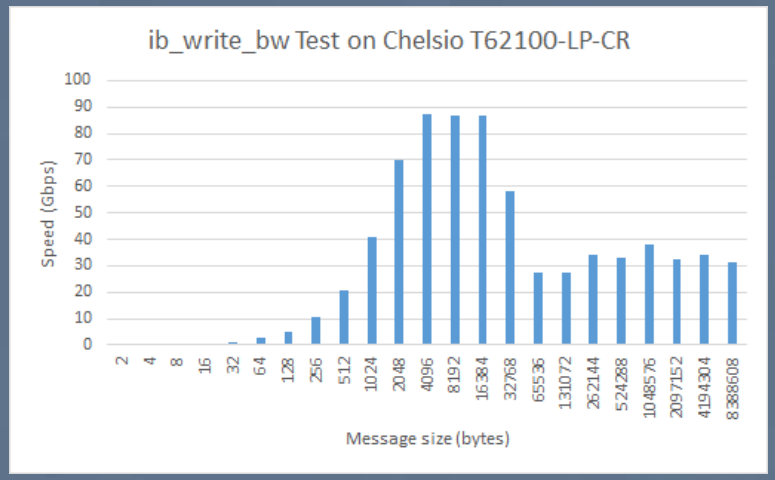
RoCE testbench elements:

4 nodes of:

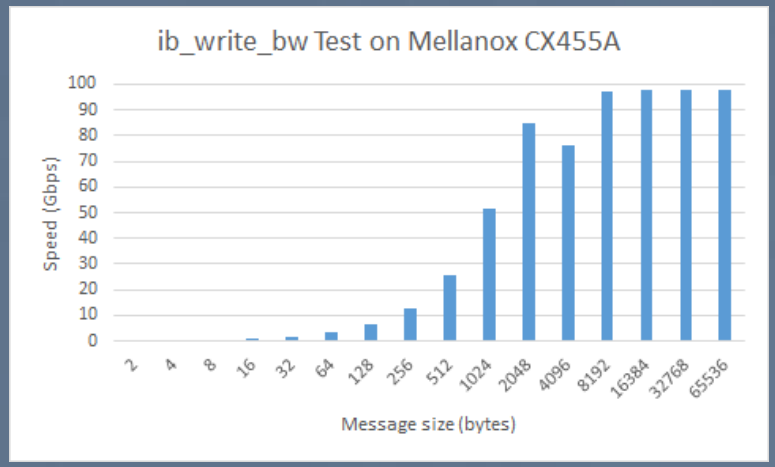
- Intel® S2600KP
- 2x Intel® Xeon® CPU E5-2650 v3 at 2.30GHz (10 cores, 20 threads)
- 64 GB DDR4 memory at 2134 MHz
- Mellanox CX455A ConnectX-4 100G NIC

Result charts

iWARP

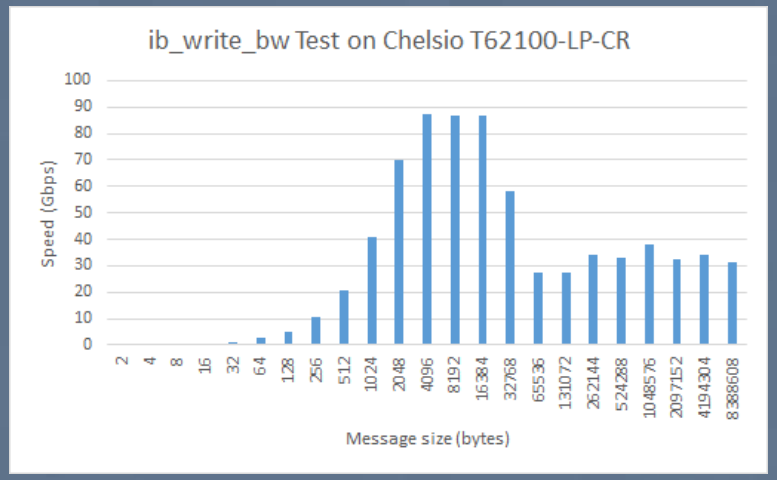


RoCE

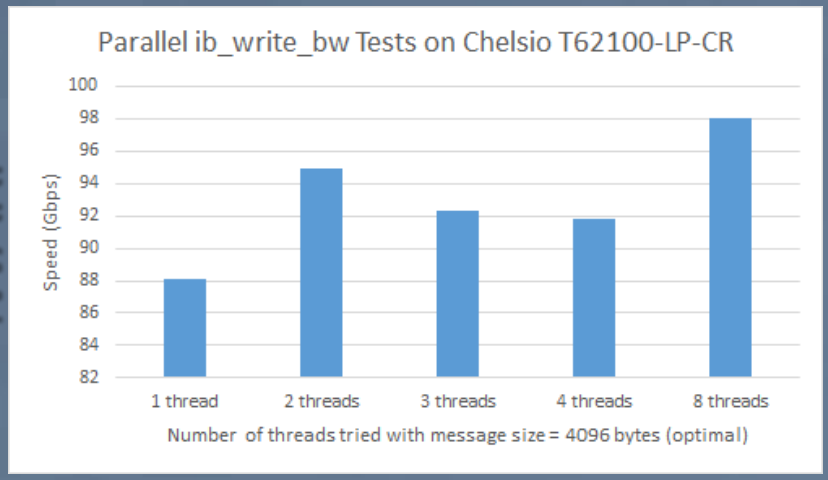


Result charts

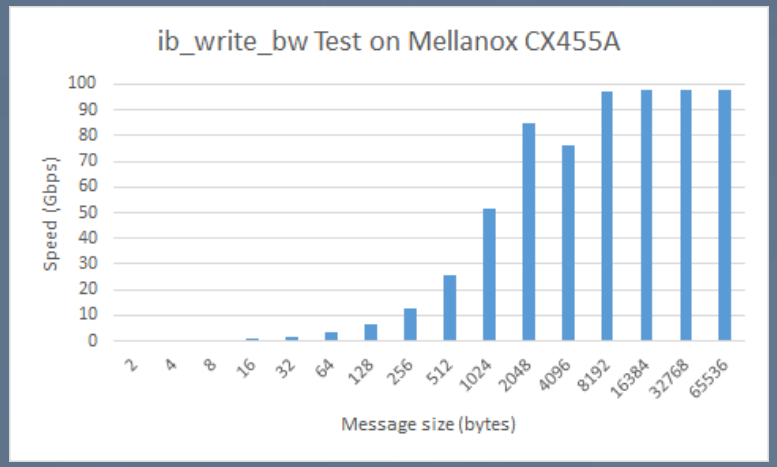
iWARP



iWARP

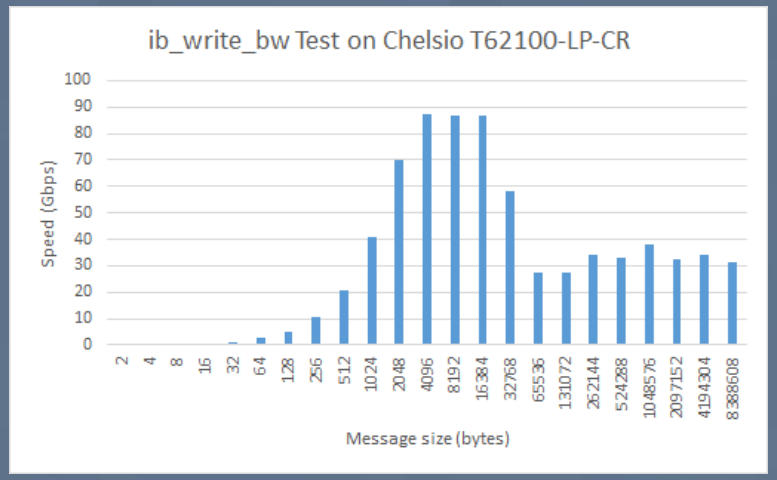


RoCE

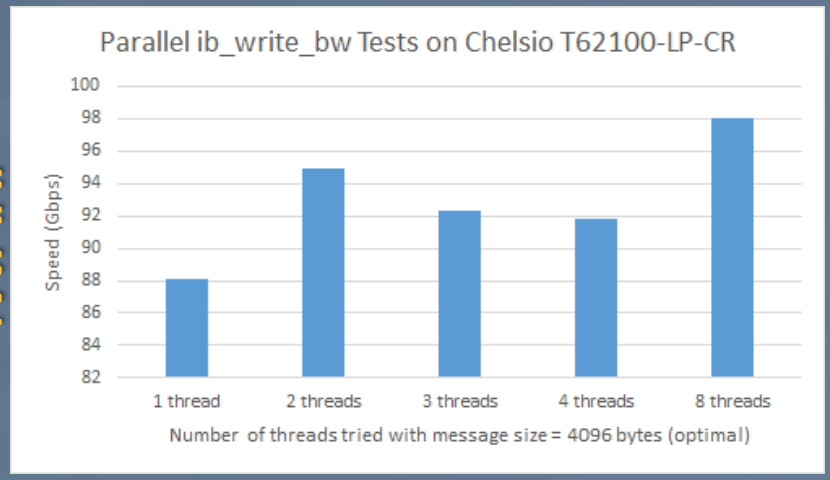


Result charts

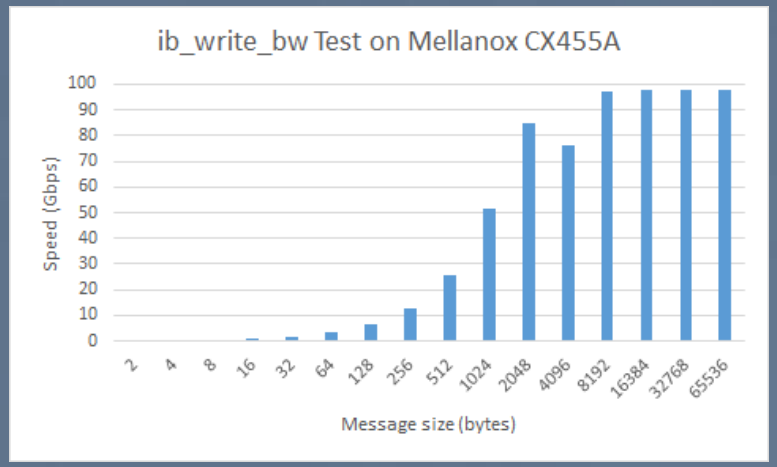
iWARP



iWARP



RoCE

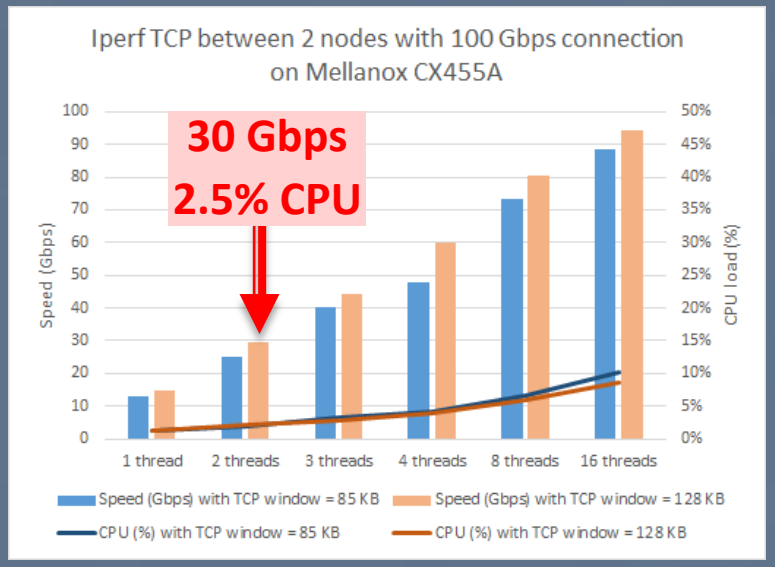


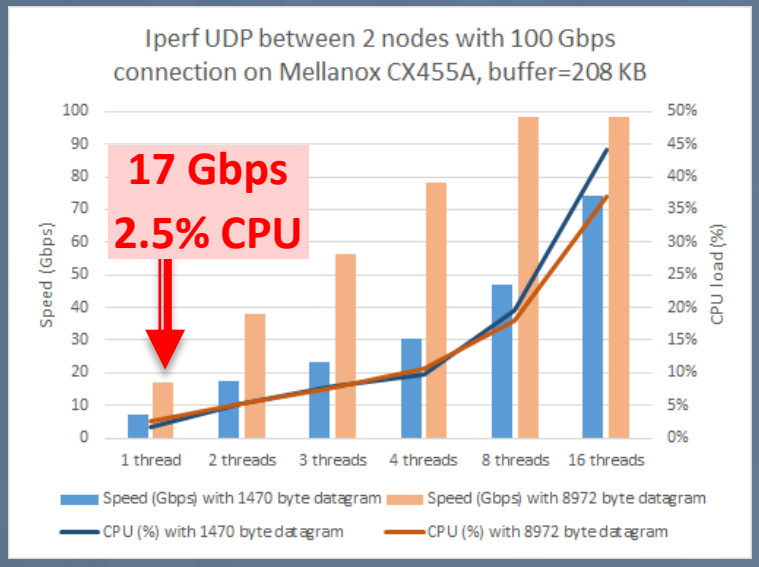
By 1 single thread

2.5% CPU for all message sizes

Result charts

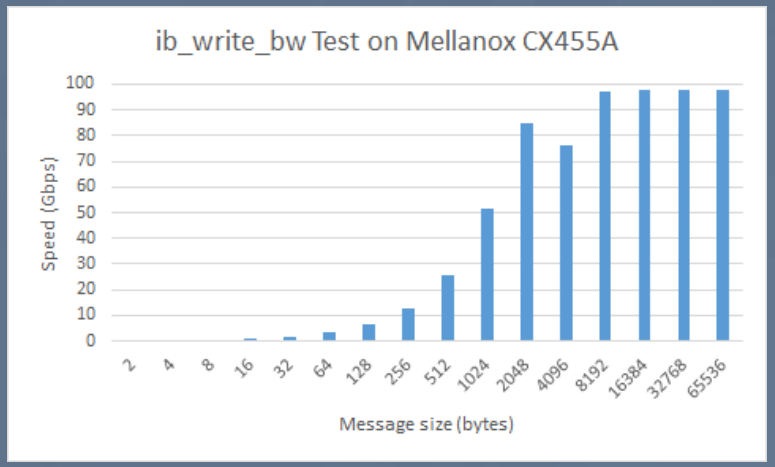
TCP





UDP

RoCE



2.5% CPU for all message sizes

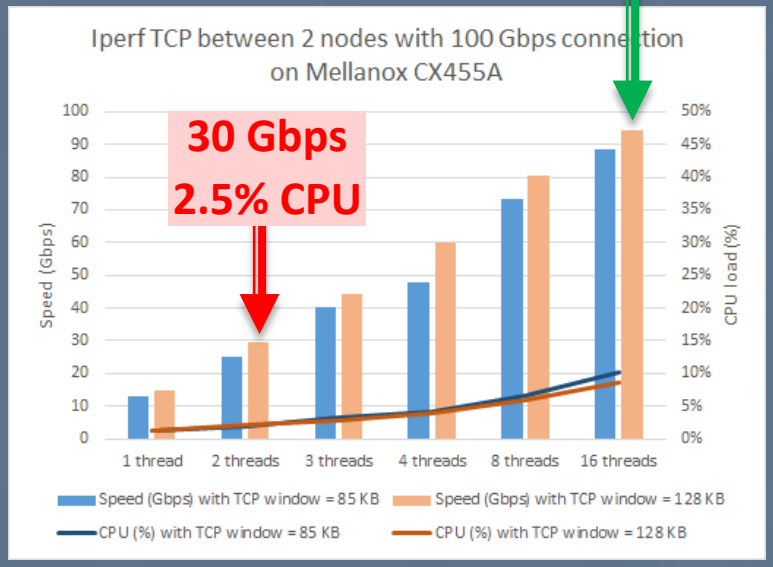
By 1 single thread

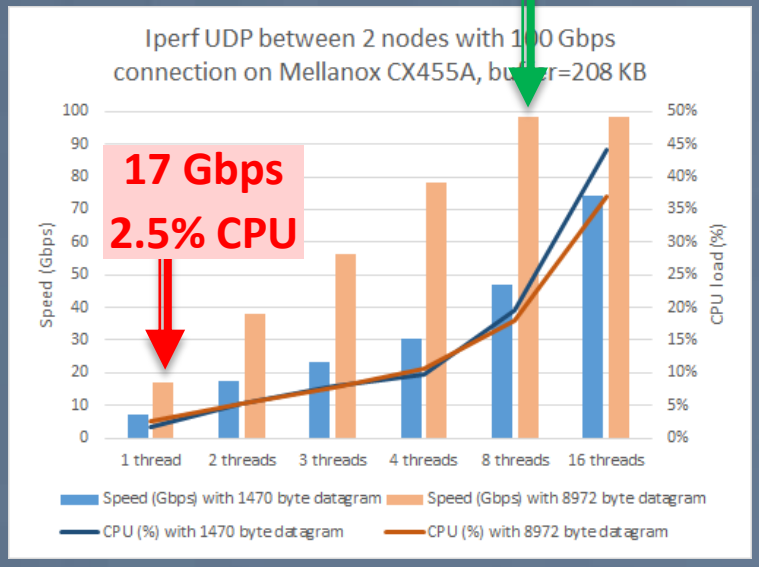
Result charts

94.4 Gbps
8.6% CPU

98 Gbps
18% CPU

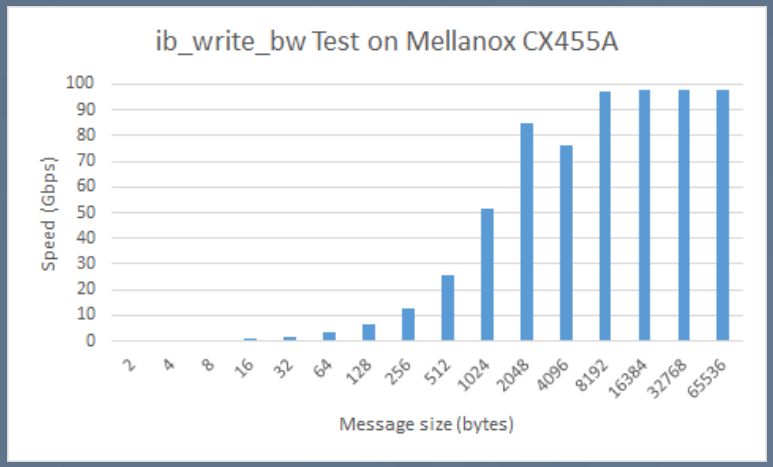
TCP





UDP

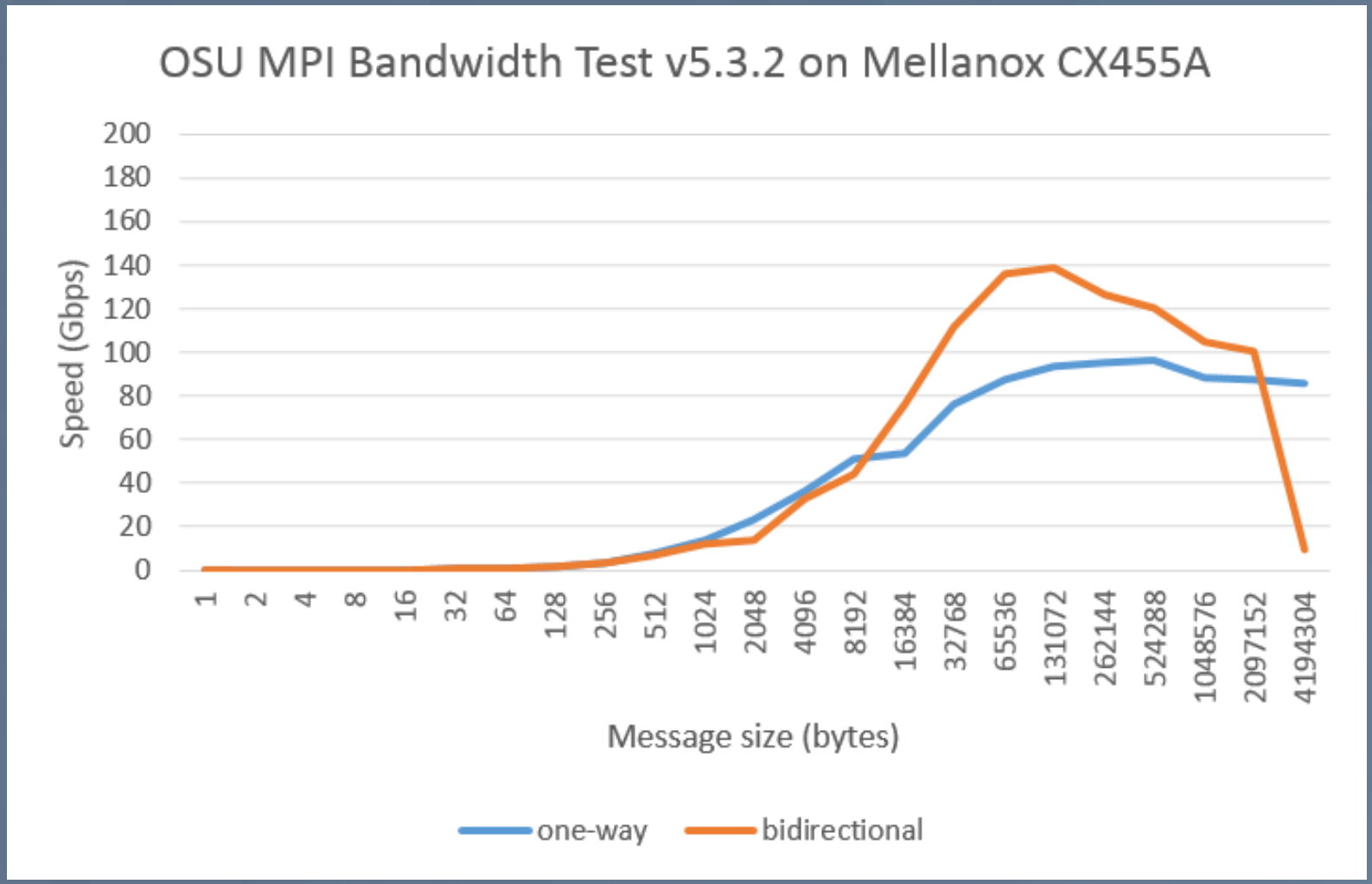
RoCE



2.5% CPU for all message sizes

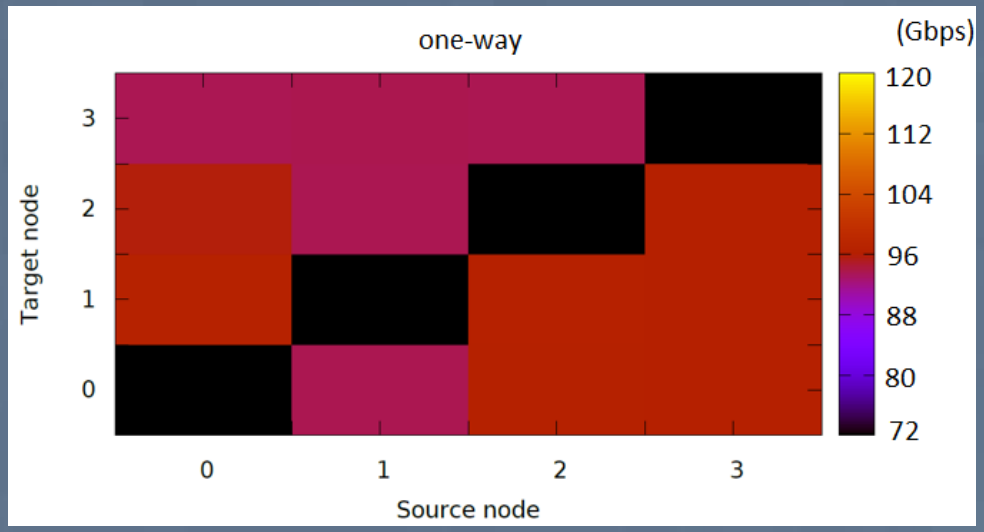
By 1 single thread

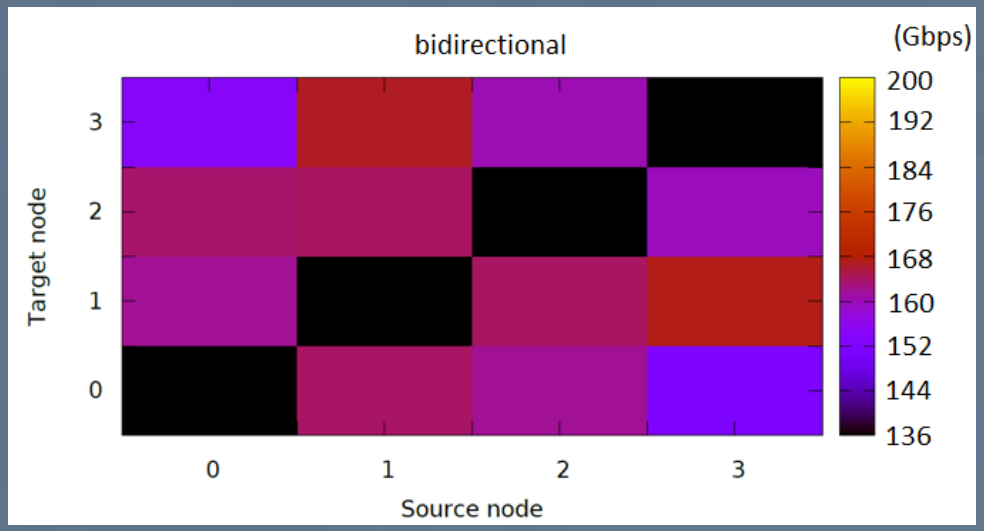
MPI result charts with RoCE



RoCE

Result charts 2 with RoCE





Purpose of heat maps:

- To check stability

Summary

- Promising results
- Pure TCP/IP is inefficient
- Zero-copy approach is needed
- Need to understand why the bidirectional heat map is not homogeneous

Thank you!