

Non-threaded and Threaded Approaches to MultiRail Communication with uDAPL

Jie Cai, Alistair P. Rendell, Peter E. Strazdins

School of Computer Science
The Australian National University

NPC 2009 @ Gold Coast, Australia
21/09/2009

Outline

- Problems and motivations
- Background
 - User Direct Access Programming Library (uDAPL)
 - InfiniBand (IB) MultiRail Configurations
- Bandwidth Micro-Benchmarks
 - Non-threaded and Threaded
- Performance Results
- Conclusion

Problem and Motivation

- Network bandwidth is a bottleneck for most applications using message passing or inter-nodes communication middleware on different networks.
 - Ethernet
 - High performance interconnects (e.g. InfiniBand)
- A well-known solution to this problem is to utilize “multirail networks”.
 - A multirail network is a network containing multiple parallel physical connections or “rails”.

Problem and Motivation (Cont.)

- However, current solutions are not generic or portable, either Sockets API or IB Verbs are targeted, such as:
 - MVAPICH2, Open MPI
- The first question investigated and answered in this presentation:
 - Can significant bandwidth improvement be achieved through a portable and platform independent communication library compared to single rail?
 - uDAPL is chosen as a candidate.

Problem and Motivation (Cont.)

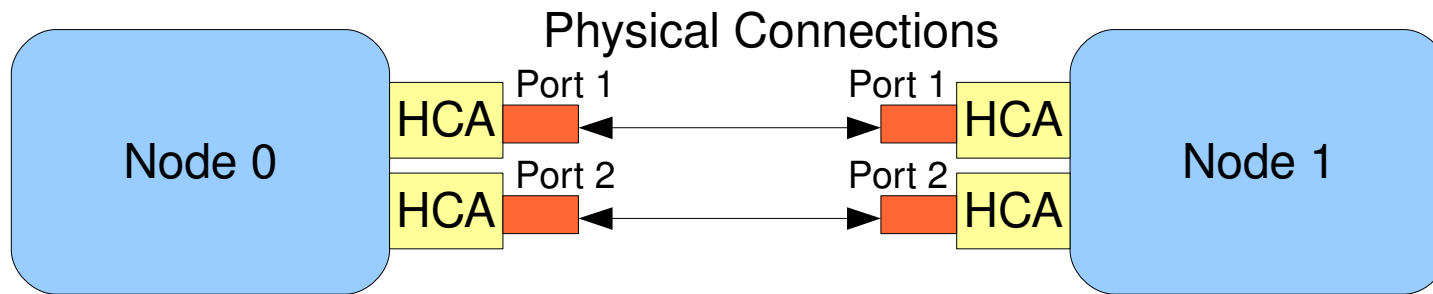
- Besides the above issue, a modern cluster system consists of:
 - Multi-ports network interface cards
 - Multicore workstations
- The second question investigated and answered:
 - What is the best way to pursue the bandwidth improvement on different multirail configured clusters?
 - Non-threaded and threaded approaches are investigated on InfiniBand (IB) multirail network.

uDAPL

- uDAPL is a portable and platform independent communication library that provides Remote Direct Memory Access (RDMA) and send/recv operations.
- Some well-known message passing and communication middlewares have attempted to take the portability advantage from uDAPL.
 - Intel MPI
 - Intel Cluster OpenMP
 - Open MPI
 - MVAPICH2
 - HP MPI

IB MultiRail Configurations

Multi-HCA:



Network Interface Card of IB is known as Host Channel Adapter (HCA).
Configuration (a)

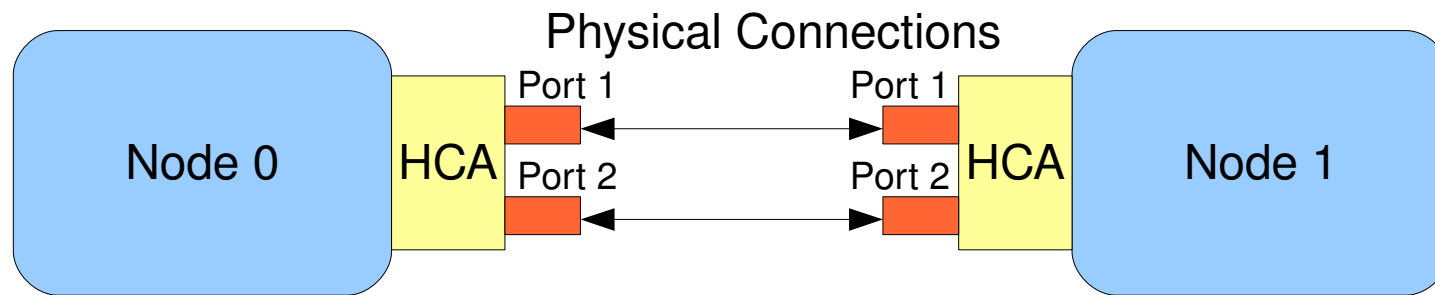
Bandwidth of configuration (a) can be calculated
by:

$$B_a = N_r \min(B_{IB}, B_{HCA}, B_{host})$$

N_r is number of rails, B_{IB} is IB speed (SDR/DDR/QDR),
 B_{HCA} is HCA PCI speed, B_{host} is host PCI speed

IB MultiRail Configurations (Cont.)

Multi-port Single-HCA:



Network Interface Card of IB is known as Host Channel Adapter (HCA).
Configuration (b)

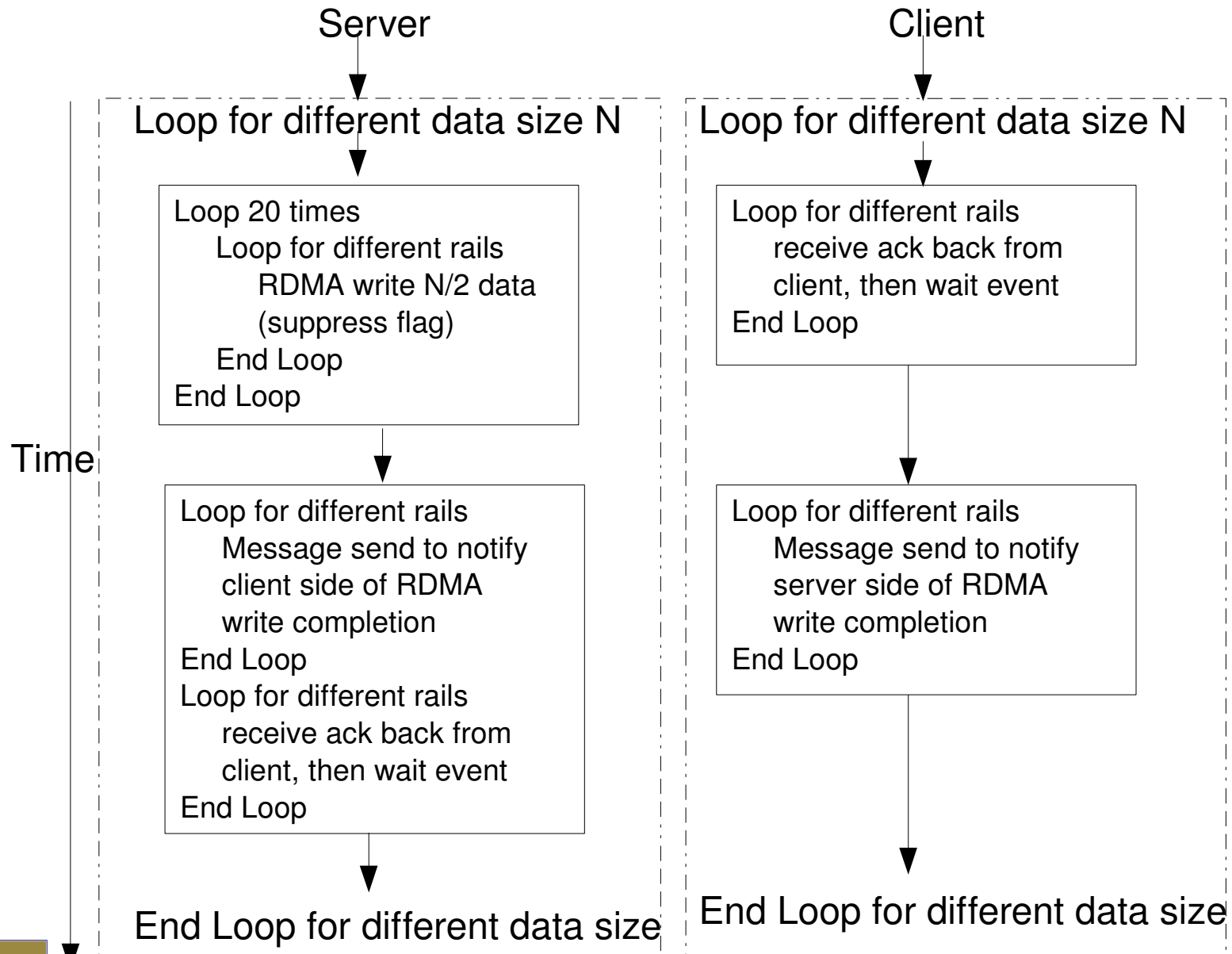
Bandwidth of configuration (a) can be calculated

by:

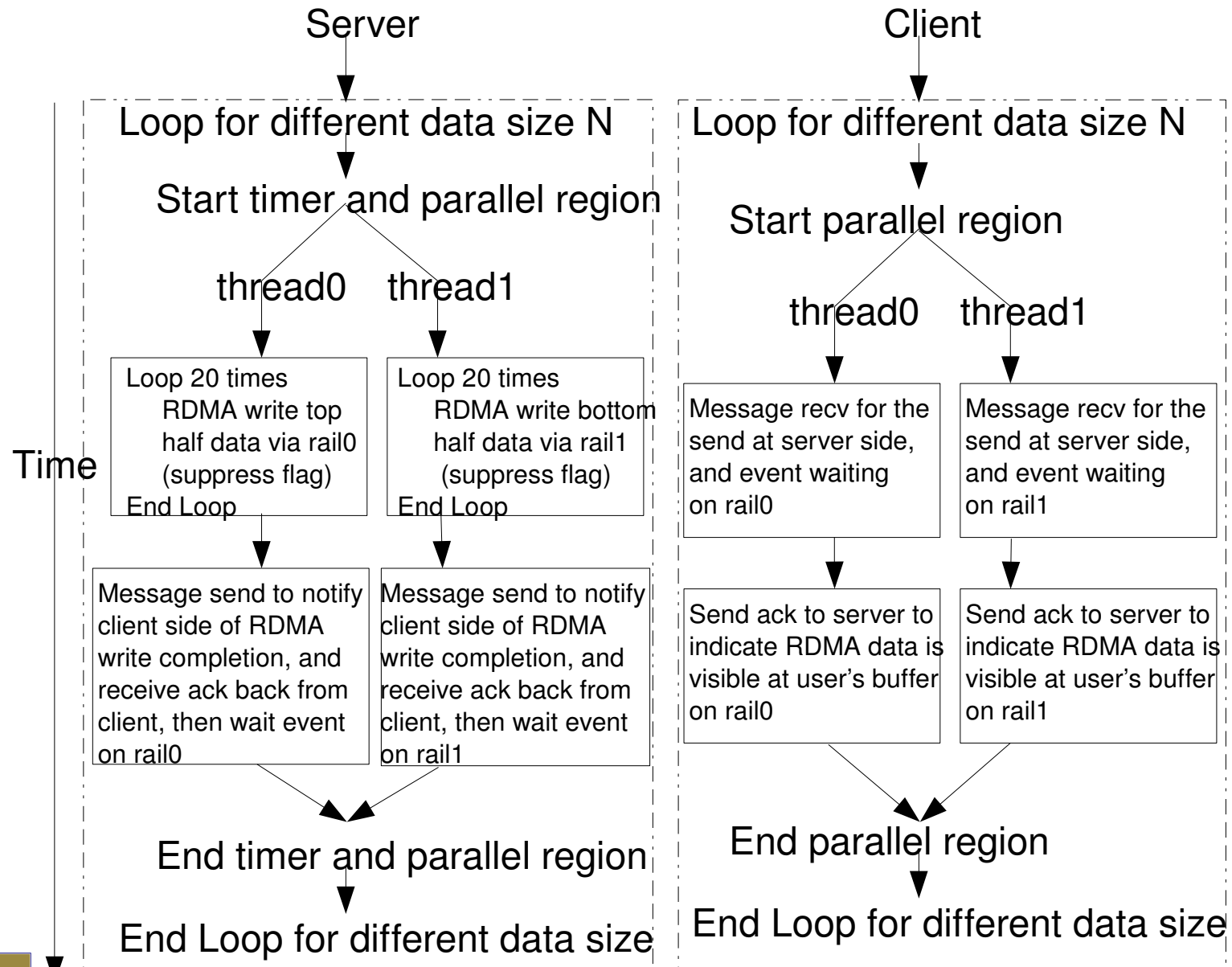
$$B_a = \min(N_r B_{IB}, B_{HCA}, B_{host})$$

N_r is number of rails, B_{IB} is IB speed (SDR/DDR/QDR),
 B_{HCA} is HCA PCI speed, B_{host} is host PCI speed

Bandwidth Benchmarks – Non-threaded



Bandwidth Benchmarks – Threaded

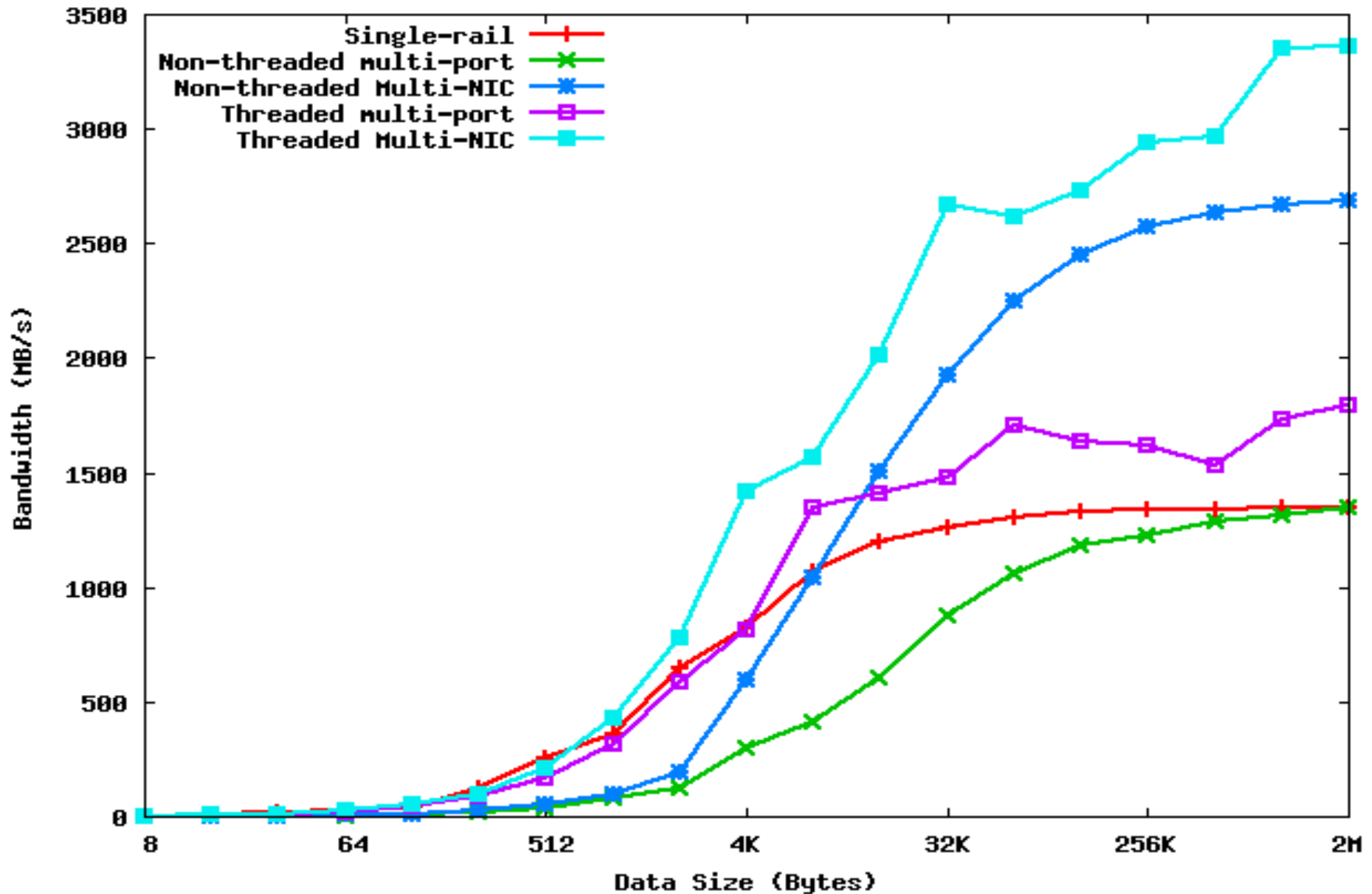


Experimental Setup

- InfiniBand cluster:
 - Four Sun Ultra24 workstations, with Intel Core 2 Quad Q6600 CPU, 4GB DDR2 memory.
 - x16 PCI-e Gen2 slots, $B_{host} = 4\text{GB}/s$
 - Mellanox ConnectX MHGH28-XTC dual-port HCAs.
 - PCI-e 1.1, $B_{HCA} = 2\text{GB}/s$
 - IB 4x DDR, $B_{IB} = 2\text{GB}/s$
 - Bottleneck for Multi-HCA is 4GB/s
 - Bottleneck for Multi-port is 2GB/s

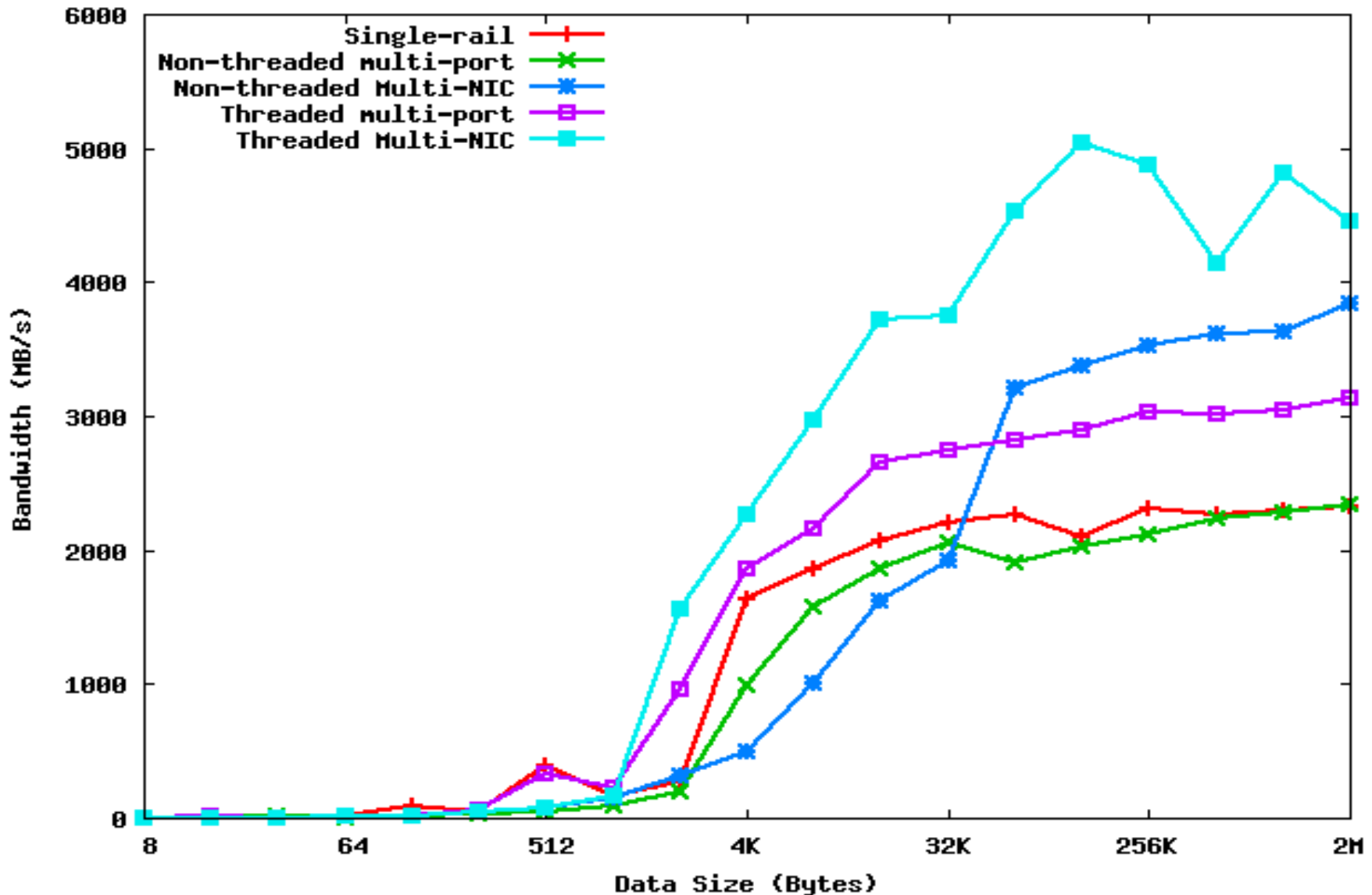
Performance – Uni-Directional Bandwidth

Uni-Directional RDMA Write Multirail Bandwidth



Performance – Bi-Directional Bandwidth

Bi-Directional RDMA Write Multirail Bandwidth



Results Discussion

- Uni-directional bandwidth
 - Threaded approach:
 - ~33% improvement using multi-port
 - ~148% improvement using multi-HCA
 - Non-threaded approach:
 - No improvement using multi-port
 - 90% improvement using multi-HCA
- Bi-directional bandwidth
 - A similar pattern of improvement as uni-directional bandwidth

Conclusion

- Utilizing a portable communication library does improve significant bandwidth over multirail networks.
- The best way to achieve the bandwidth improvement is to use the threaded approach over a multi-HCA configured network.

Thank You!



Acknowledgement:

This work is funded by Australian Research Council Grant LP0669726, ANU CECS Faculty Research Grant, Intel Corporation, and Sun Microsystems.