rockport.

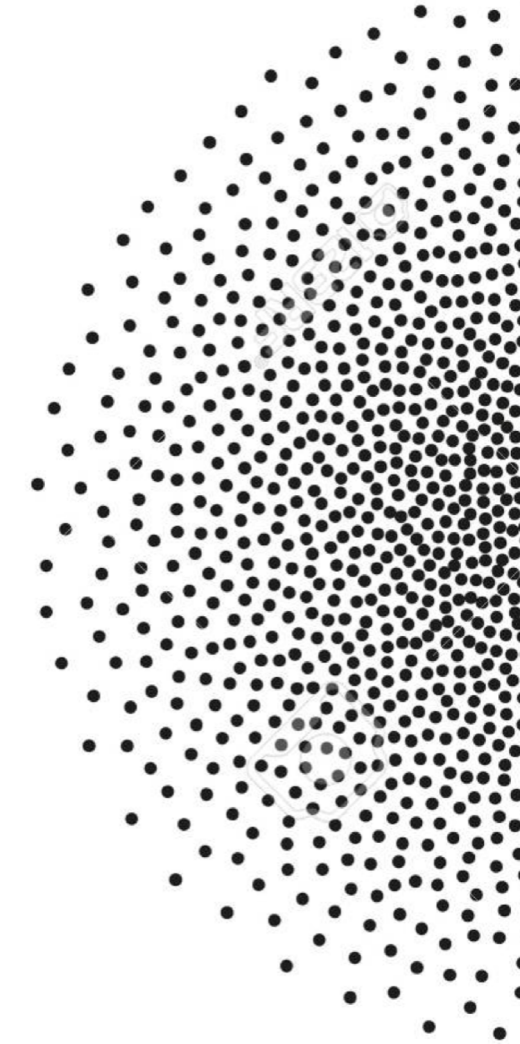
Upcoming MVAPICH2 Design Enhancements on the Rockport Switchless Network

9th Annual MVAPICH User Group (MUG) Meeting

Matthew Williams, Rockport Networks, mwilliams@rockportnetworks.com
Tuesday, August 24th, 2021

Agenda

- Rockport Architecture Review
- Upcoming Rockport Firmware and Mvapich2 Design Enhancements
- Critical Message Latency Performance Benchmark Results
- Questions



rockport.

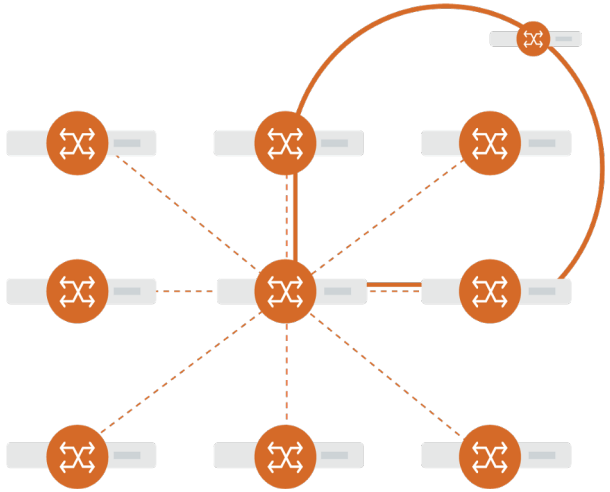
Simplify the Network

Rethinking Network Performance at Scale for HPC Environments

Rockport has reimagined performance networks with an embeddable switchless architecture that delivers the performance at scale needed for HPC, AI, and HPDA.

By distributing the network switching function into each device endpoint,
the nodes become the network:

- Direct interconnect
- Distributed routing and control planes
- Linear scaling
- Standard Ethernet-based host interface (RoCEv2 and TCP/UDP)
- No external, centralized switches
- Field-upgradable firmware with rich roadmap
- Supported in the latest MVAPICH2 (2.3.6) library



Rockport Architecture
Self-discovering, self-configuring, self-healing
Very High Path Diversity
Distributed, embedded FLIT switching

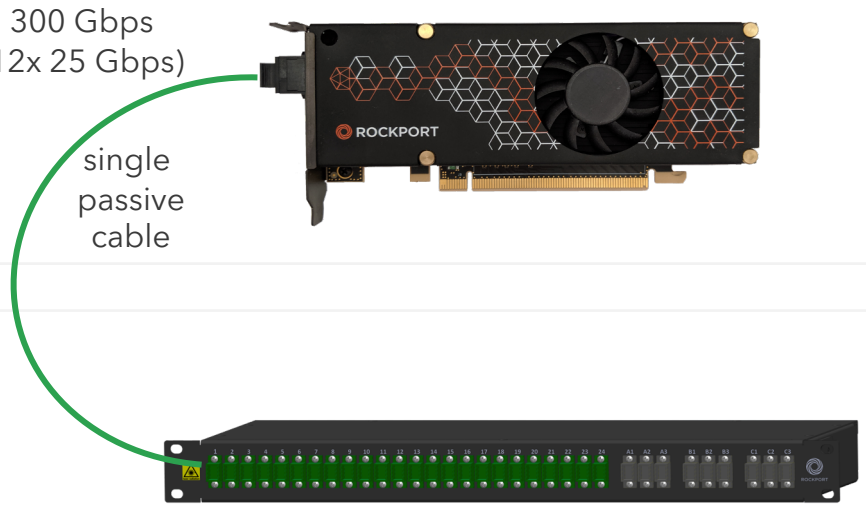
rockport.

Scalable Supercomputer Networking, Simplified

Rockport Switchless Network Solution

Rockport RO6100 Network Card

- World's first Network Card
- Standard Ethernet interface (verbs and sockets)
- Patented FLIT Switching in a field-upgradable FPGA



Rockport SHFL

- Supercomputer networking topologies prewired in box
- Stunningly simple cabling solution
- Completely passive

Rockport Autonomous Network Manager

- Bird's eye view into active network
- Deep insight into network performance on a per-job basis
- Never seen before time travel

rockport.

Rockport Switchless Network

Performance Network Fabric

Topology Discovery

- Self-discovering, self-configuring, self-healing
- Scales in and out easily

Distributed Source Routing

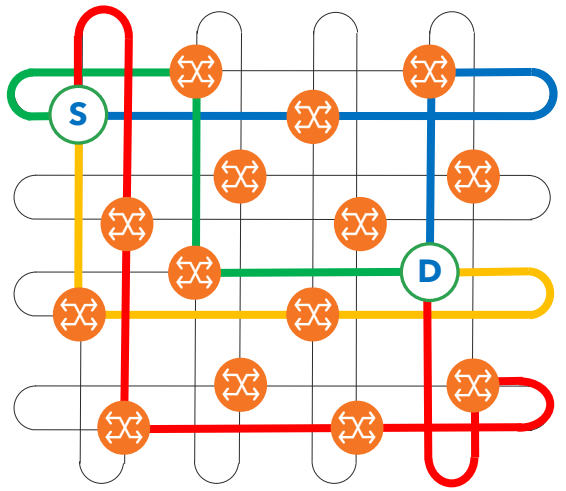
- Rockport distributed Deadlock-Free Routing algorithm (DFR)
  - Deadlock free routing across all topologies (complete or sparse)
  - Paths are physically independent and have no common links
  - Ensures high path diversity
- Traffic spread across all available paths on a per-flow or per-packet basis

Extremely Fast Distributed Switching

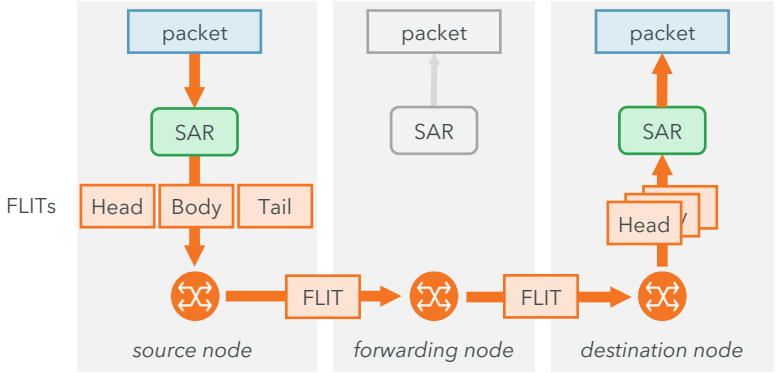
- Packets segmented into small pieces (FLITs)
  - Ensures very low latency performance, even under heavy load
- Embedded FLIT switching forwards FLITs to destination
- Destination reassembles packets

Inherent Performance Advantages

- Predictably low latency at every scale
- Zero congestive loss



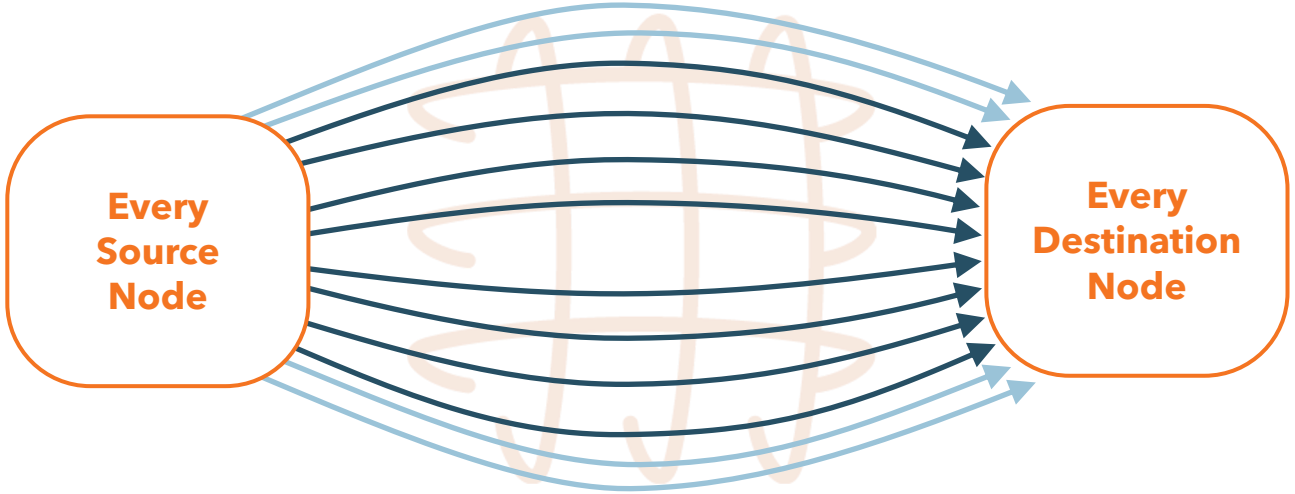
Distributed FLIT Switching



rockport.

Performance Advantages

Very High Path Diversity



- High path diversity is a very important element of network design
- Rockport nodes distribute packets across the 8 optimal of 12 source routes to each destination to:
  - Distribute the network load across the topology
  - Avoid multiple congested paths through adaptive routing
  - Immediately react in hardware to local or network issues

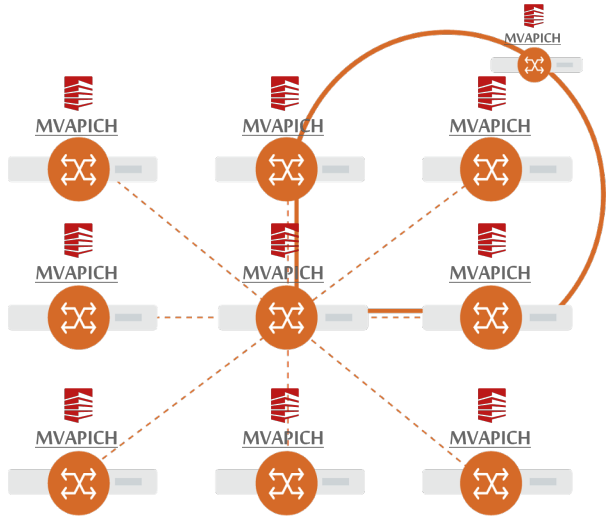
rockport.

High Performance MPI Solution

Rockport Partnership with OSU

MVAPICH

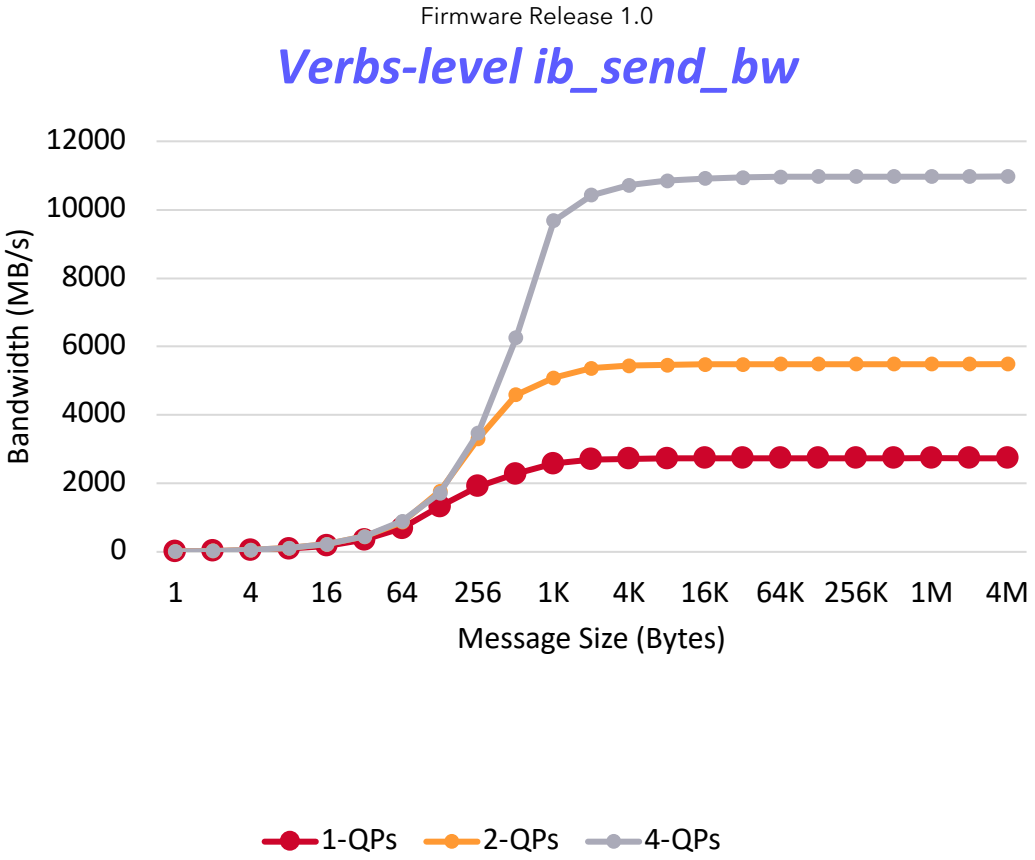
- Rockport has partnered with OSU to ensure that our switchless network solution delivers a high performance MPI solution
- Rockport is supported in the latest MVAPICH2 (2.3.6) library
- The OSU team is adding new capabilities to the MVAPICH2 library to take advantage of some unique Rockport architectural benefits
  - Using multiple queue pairs for large messages to leverage high path diversity
  - Identifying critical latency-sensitive MPI messages so that **unmodified MPI code** has consistently excellent workload performance even under heavy noisy-neighbor congestion
  - Targeted for MVAPICH2 2.3.7
  - Additional enhancements under development



Rockport Architecture
Directly Connected Nodes
Distributed Switching
Very High Path Diversity

rockport.

Multiple Queue Pair Designs in MVAPICH2 for Rockport NW

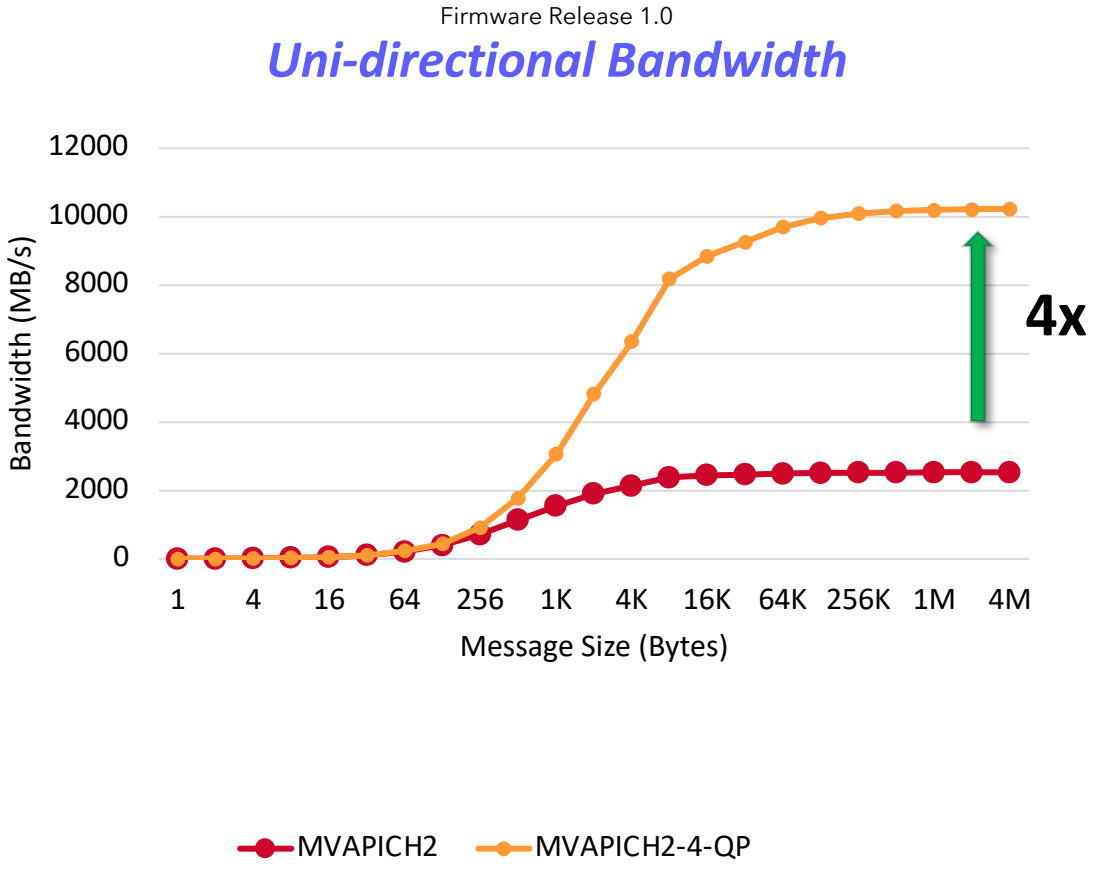


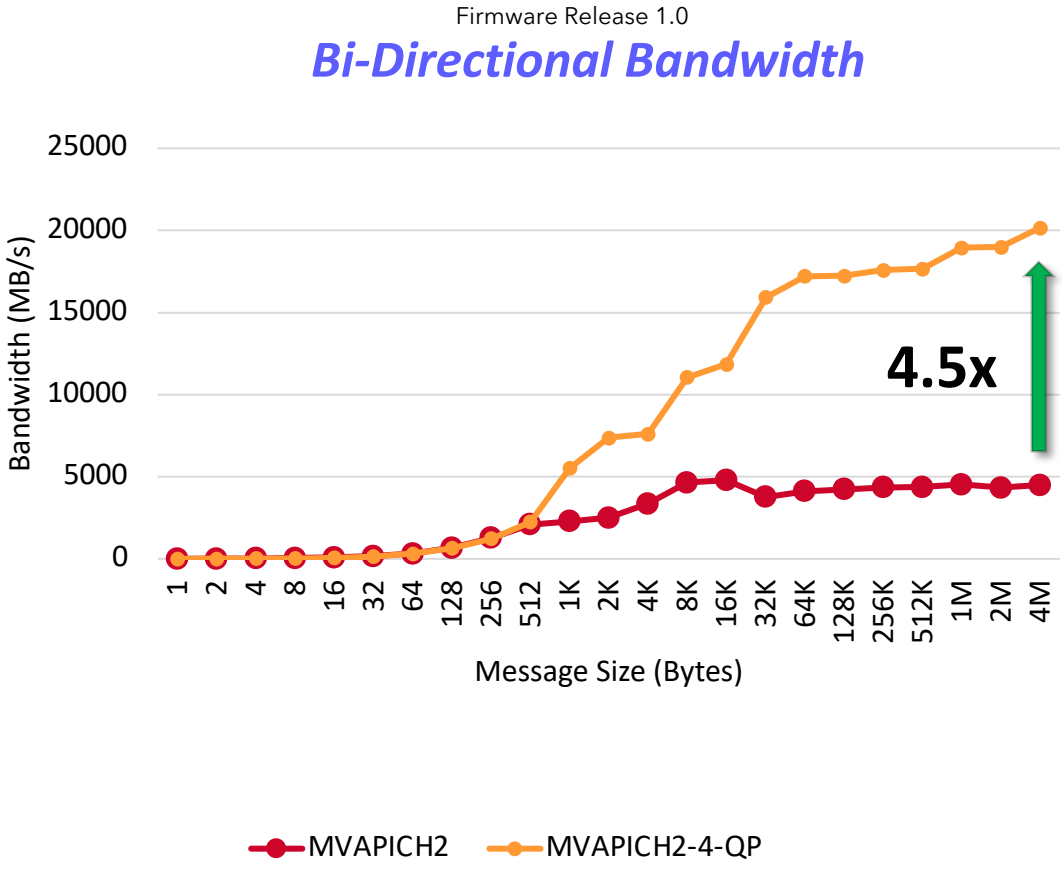
- Verbs-level one directional bandwidth increases with more QPs
  - RO6100 Network Card uses multiple parallel 25 Gbps paths
  - Peaks with four QPs
- Multi-QP design in MVAPICH2
  - Use one QP for message sizes less than the Eager threshold
    - Avoids unnecessary overheads with more QPs
  - Pick from set of QPs in a round-robin fashion for rendezvous message transfers

Experimental Setup

Specification type	System value
CPU	Intel Xeon Platinum 8280
Clock Speed	2.7GHZ
No. of Sockets	2
Cores per Socket	28
RAM	192GB
HCA	Rockport RO6100 Network Card Firmware Version 1.0
RoCE Version	RoCEv2
Benchmarks	OSU Microbenchmarks(OMB) v5.7.1

Inter-node point-to-point Bandwidth

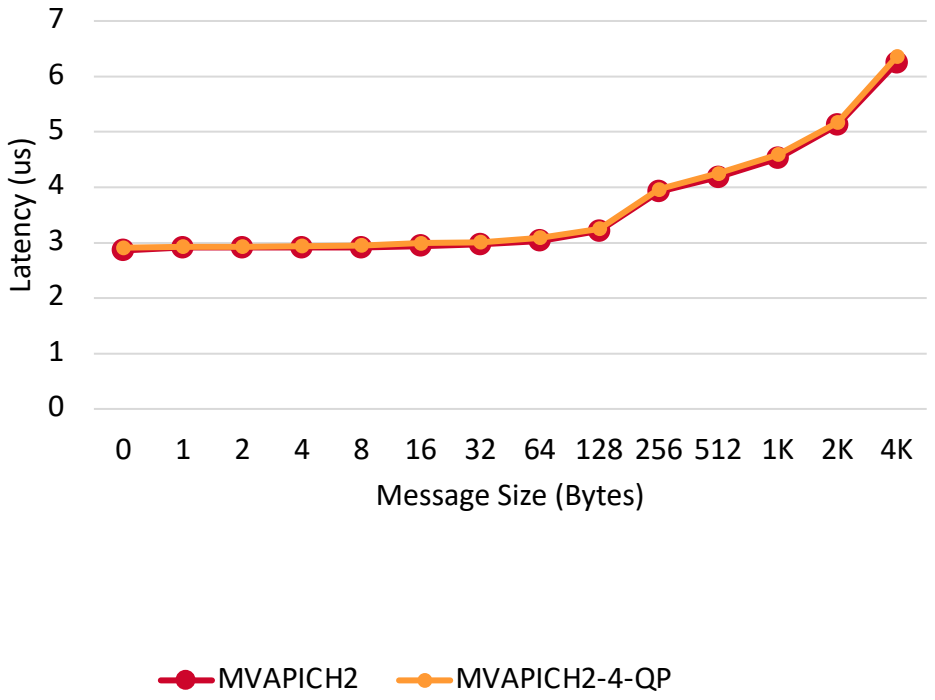




Inter-node point-to-point Latency

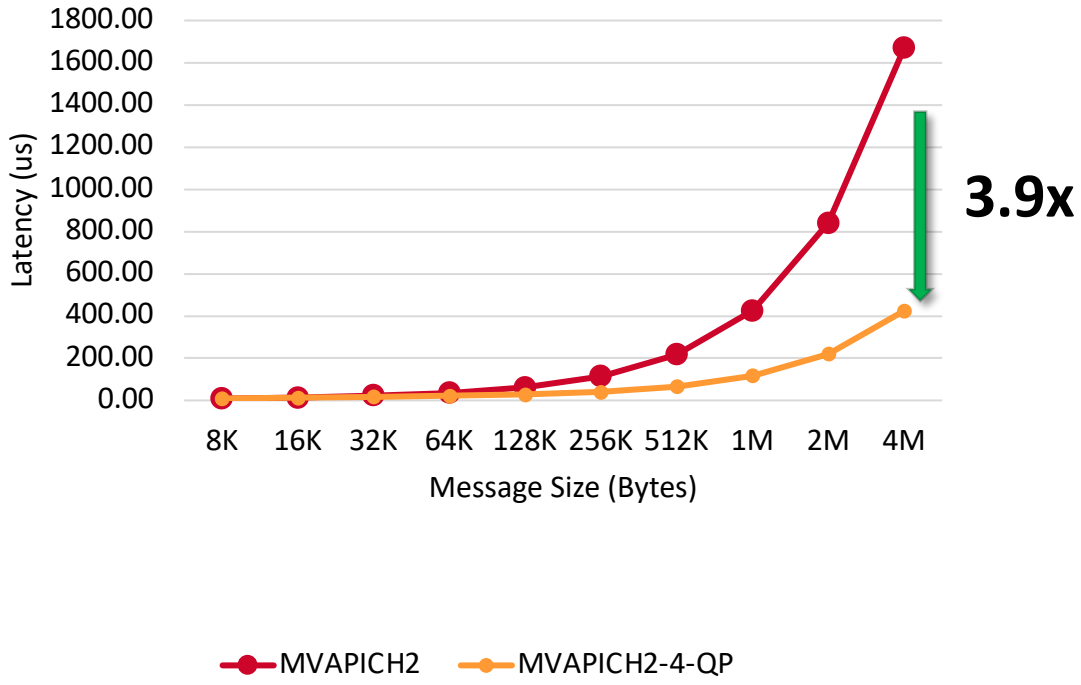
Firmware Release 1.0

Small message Latency



Firmware Release 1.0

Medium/Large message Latency

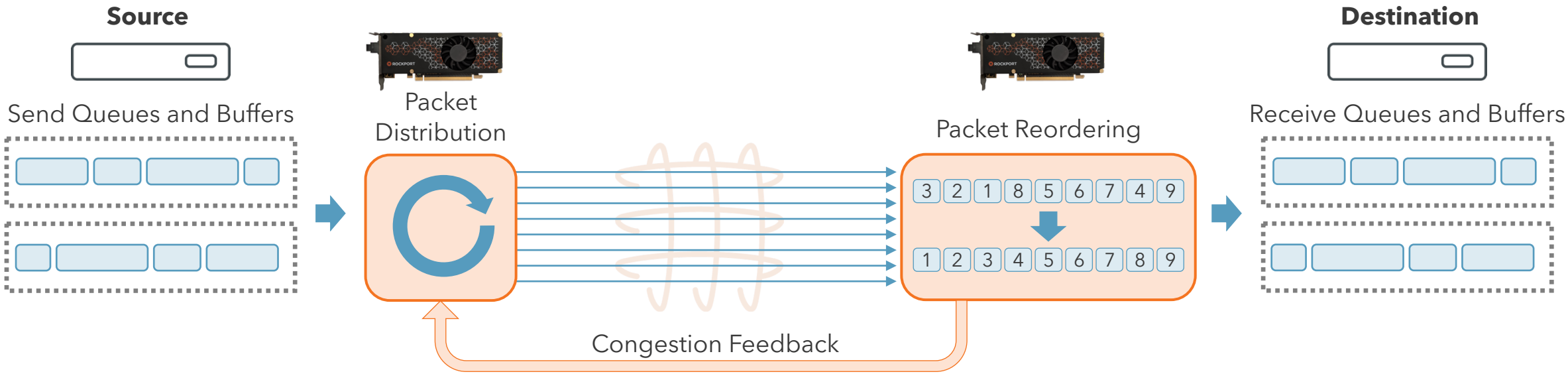


Fall 2021
Firmware Enhancements

rockport.

Fall 2021 Firmware Enhancements

Rockport Adaptive Multi-Path Flows With End-to-End Path Monitoring

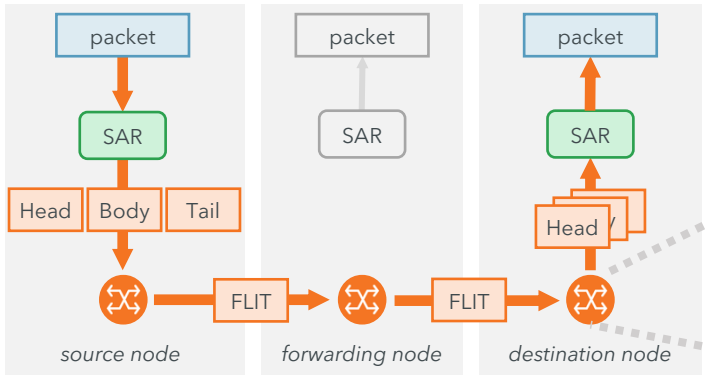


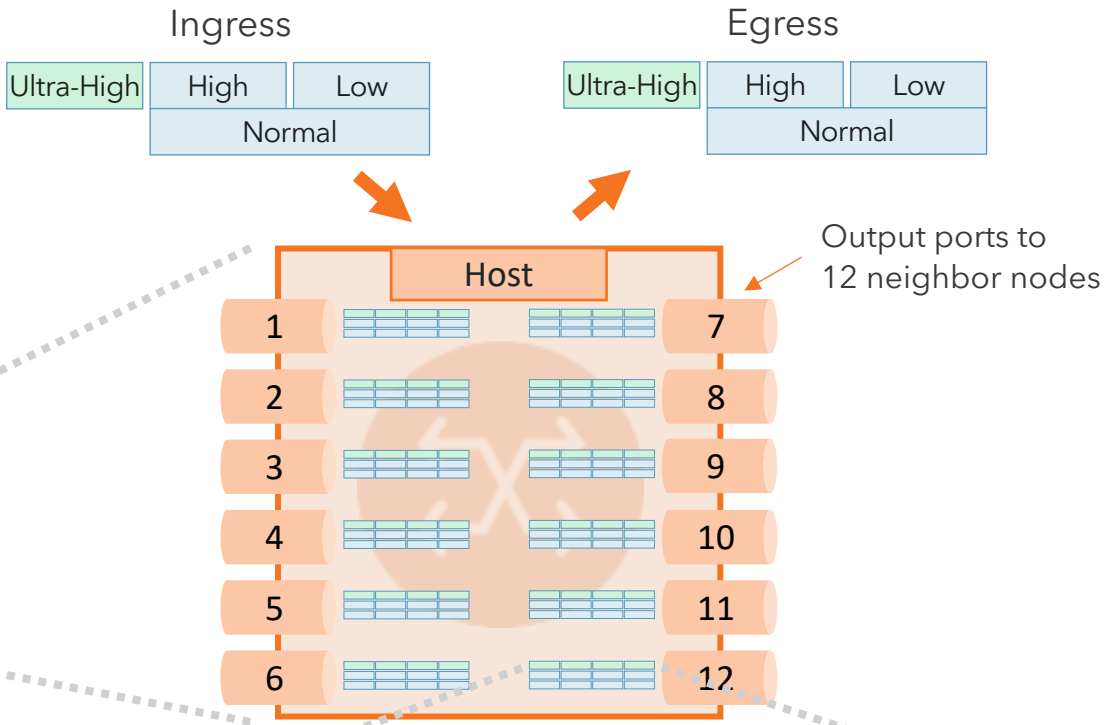
- Using Multi-Path Flows, RDMA datagrams between a source and destination node are adaptively distributed across multiple paths on a per-packet basis
- The real-time end-to-end congestion of each path is used to select the optimum paths on a continual basis
- Packets are delivered in order, supporting all RDMA transports and use cases

rockport.

Fall 2021 Firmware Enhancements

FLIT Switch Details





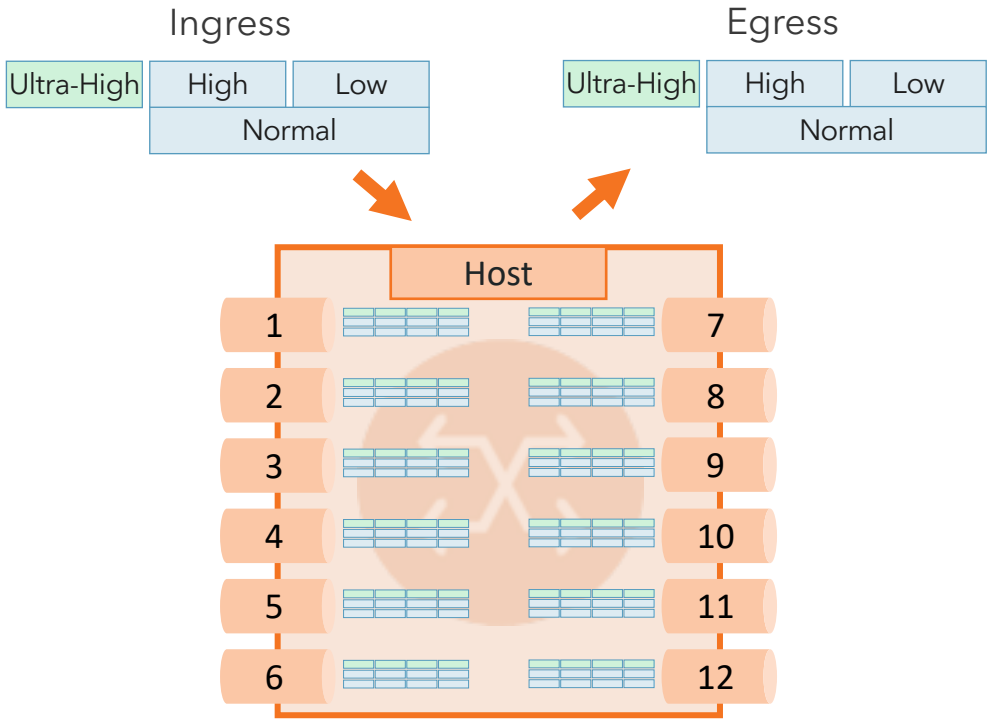
12 VCs per Port

Layer 1	Layer 2	Layer 3	Layer 4
Ultra-High	Ultra-High	Ultra-High	Ultra-High
Normal 1	Normal 1	Normal 1	Normal 1
Normal 2	Normal 2	Normal 2	Normal 2

rockport.

Fall 2021 Firmware Enhancements

Ultra-High Priority



12 VCs per Port

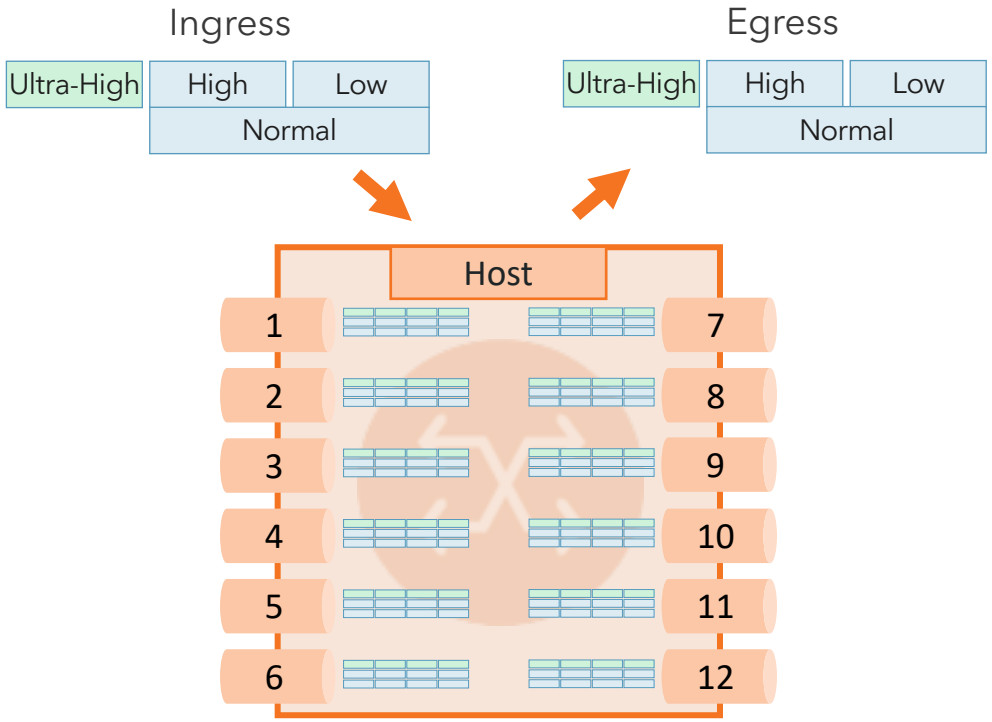
Layer 1	Layer 2	Layer 3	Layer 4
Ultra-High	Ultra-High	Ultra-High	Ultra-High
Normal 1	Normal 1	Normal 1	Normal 1
Normal 2	Normal 2	Normal 2	Normal 2

- Ingress traffic is classified as Ultra-High, High or Low priority
  - Based on either layer 2 or layer 3 tagging
  - All three classes are respected at ingress and egress
  - High and Low are combined into a Normal class in forwarding/transit nodes
- Ultra-High traffic is always serviced first
  - Ensures that latency-sensitive traffic is immune to congestion
  - Average of 25 ns additional latency per hop even under heavy congestion (50 ns maximum)
  - **MVAPICH2 will automatically recognize and tag critical MPI messages for Ultra-High Priority**
    - **Unmodified MPI code will have consistently excellent workload performance even under heavy noisy-neighbor congestion**

rockport.

Fall 2021 Firmware Enhancements

Deadlock-Free Routing



12 VCs per Port

Layer 1	Layer 2	Layer 3	Layer 4
Ultra-High	Ultra-High	Ultra-High	Ultra-High
Normal 1	Normal 1	Normal 1	Normal 1
Normal 2	Normal 2	Normal 2	Normal 2

- Rockport's Deadlock-Free Routing algorithm (DFR) ensures deadlock free routing across all topologies
  - Operates in a distributed fashion
  - Ensures high path diversity
  - Complete or sparse topologies with or without link or node failures
  - Patent-pending
- Each port's 12 virtual channels (VCs) are grouped into 4 layers of 3 VCs each (1x Ultra-High and 2x Normal)
- Generated source routes include a VC layer along with the set of egress ports to reach the destination
  - Only VCs in the assigned layer will be used for forwarding traffic
  - Ensures cycle-free forwarding

rockport.

Critical MPI Message Identification

- The OSU team is adding intelligence to the MVAPICH2 library to take advantage of some unique Rockport architectural benefits
  - Using multiple queue pairs for large messages to leverage high path diversity
  - Identifying critical latency-sensitive MPI messages so that **unmodified MPI code** has consistently excellent workload performance even under heavy noisy-neighbor congestion

- Operators can set a maximum size threshold for critical messages
  - Can be set on a per-workload basis at run-time

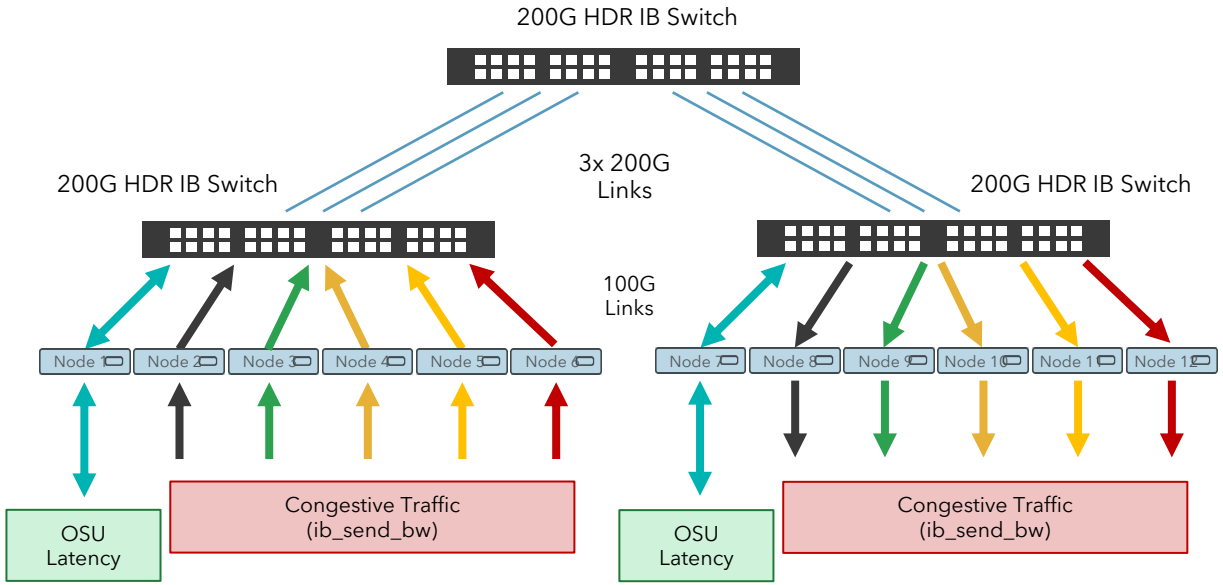
Benchmark Results

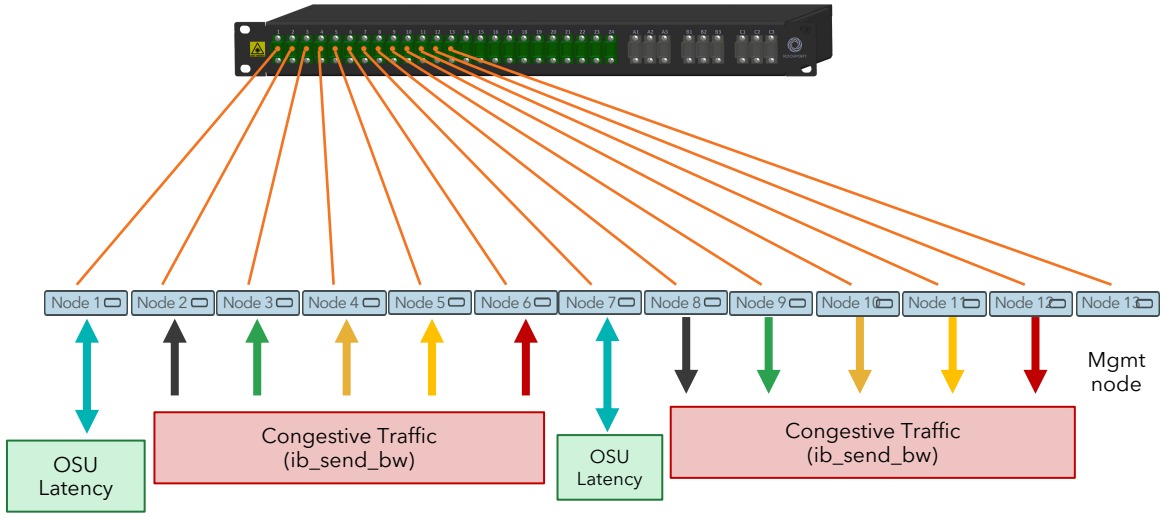
Network Performance Testing Under Load

- Typical network benchmarks run on unloaded networks and only provide a baseline, best-case view into the performance of the network
  - Unloaded = network dedicated to benchmark with no competing network traffic
- These baseline results are not useful to predict the performance of the network in a multi-workload production environment as they do not include competing, noisy neighbor traffic
  - We regularly hear from our customers and partners on how the performance of their existing production networks is not what they expected or require
- To accurately predict how well a network will perform in production, network benchmarks must be run with additional, competing loads on the network
- Traffic generators like ib_send_bw and iperf are useful tools to generate these competing loads in controlled environments

rockport.

Critical Message Performance Benchmark





InfiniBand without Oversubscription

OSU Latency benchmark between Nodes 1 and 7 in two scenarios
Unloaded: No other traffic in network
Loaded: 5x ib_send_bw

Rockport

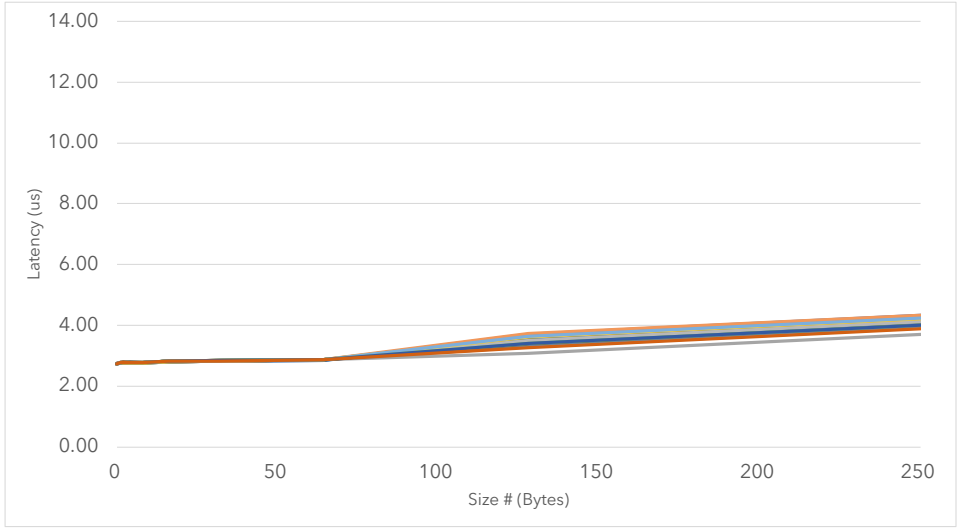
OSU Latency benchmark between Nodes 1 and 7 in two scenarios
Unloaded: No other traffic in network
Loaded: 5x ib_send_bw -q4

rockport.

Rockport and MVAPICH2 Performance Enhancements

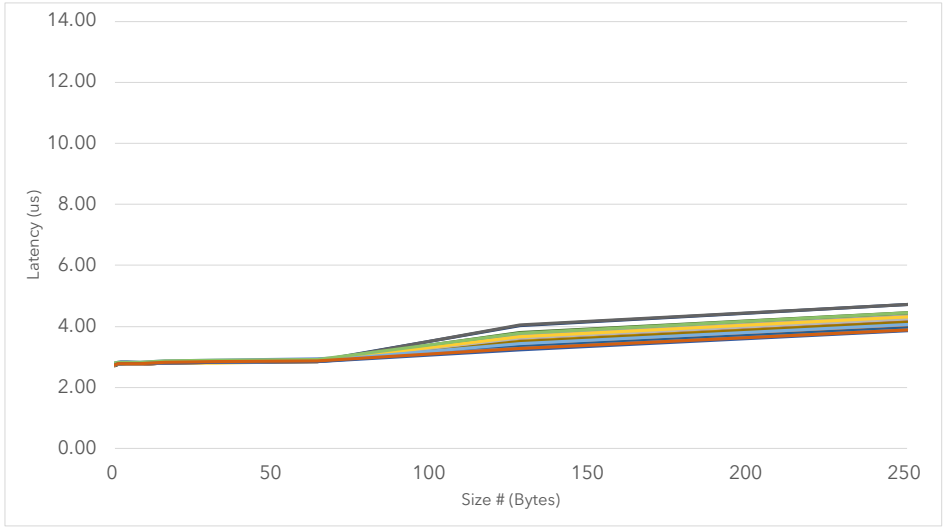
Critical Messages Immune to Heavy Network Congestion

OSU Unloaded Latency



Message Size	Average Latency Impact Under Heavy Load
0	< 0.1%
1	< 0.1%
2	< 0.1%
4	< 0.1%
8	< 0.1%
16	< 0.1%
32	< 0.1%
64	< 0.1%
128	2.9%
256	2.9%

OSU Loaded Latency



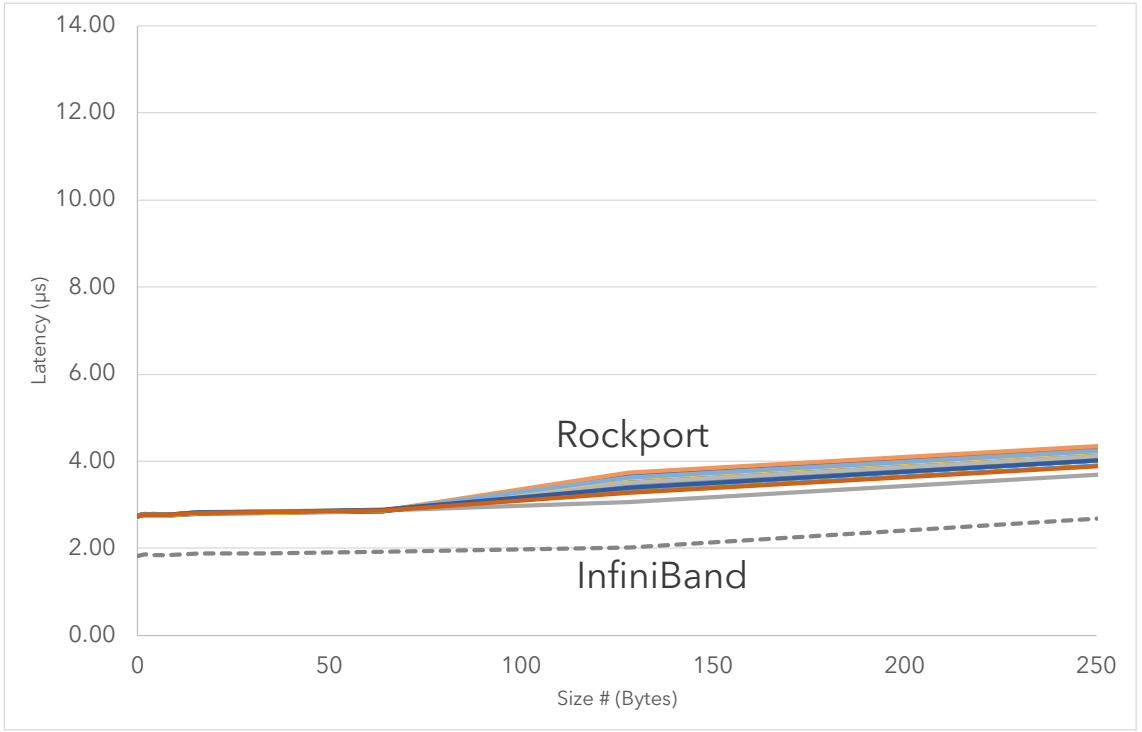
- MVAPICH2 configured at run-time to limit latency sensitive message size to 64 bytes or less
- With Rockport's Ultra-High Priority, the latency impact of heavy congestion is less than 0.1%
- Low Latency under Load, Predictable Performance

Graphs show the results of 20 runs of the OSU latency benchmark in unloaded and loaded conditions

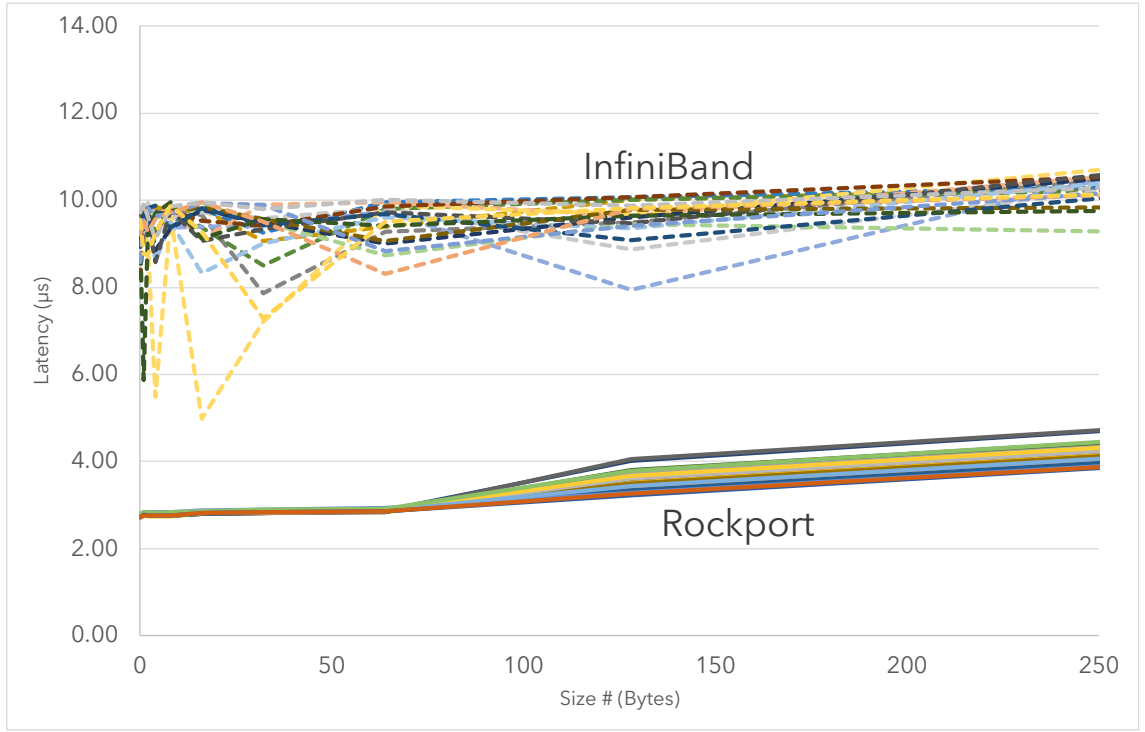
rockport.

Unloaded and Loaded Performance Test Setup vs InfiniBand

OSU Unloaded Latency



OSU Loaded Latency



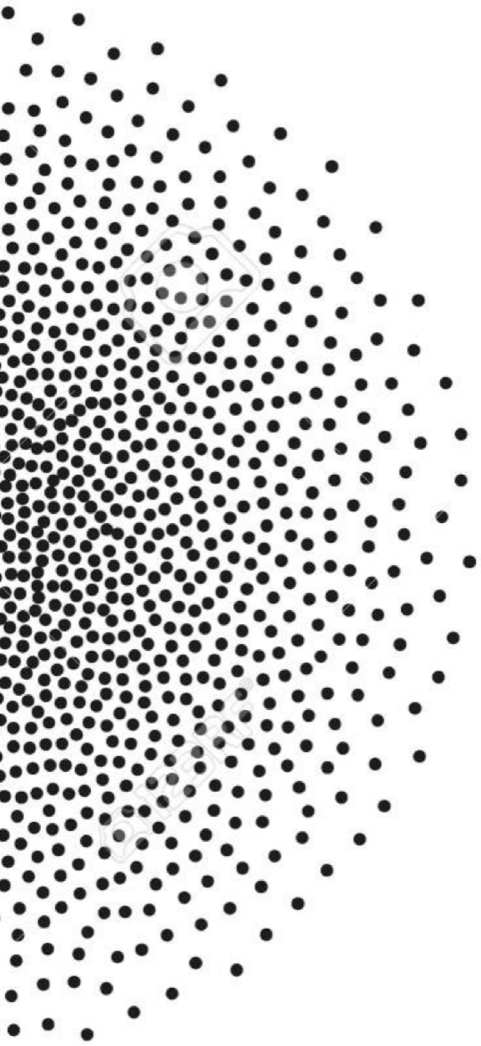
Low Latency under Load, Predictable Performance

Graphs show the results of 20 runs of the OSU latency benchmark in unloaded and loaded conditions

rockport.

Summary

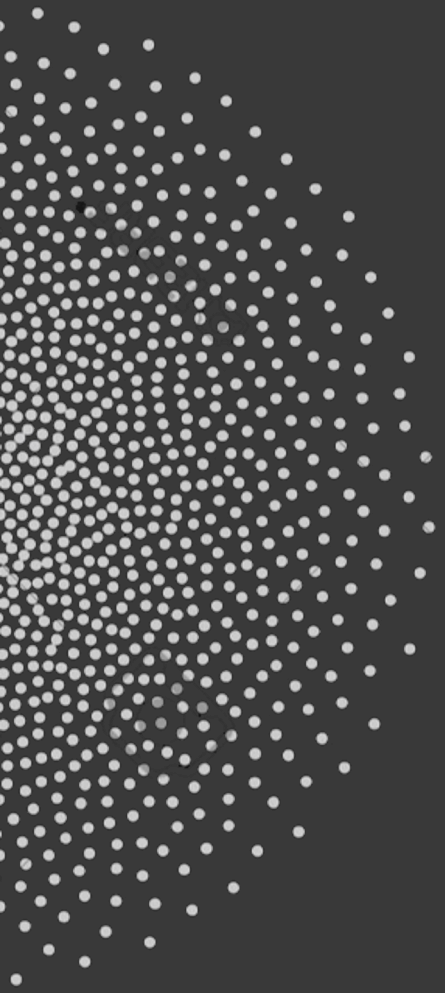
New switchless direct interconnect
MVAPICH2 design enhancements for Rockport
Critical messages immune to congestion



Thank You.
Questions?

To learn more about addressing congestion:
rockportnetworks.com/MUG

rockport.



rockport.