rockport.

Addressing Network Congestion with Rockport Networks

9th Annual MVAPICH User Group (MUG) Meeting

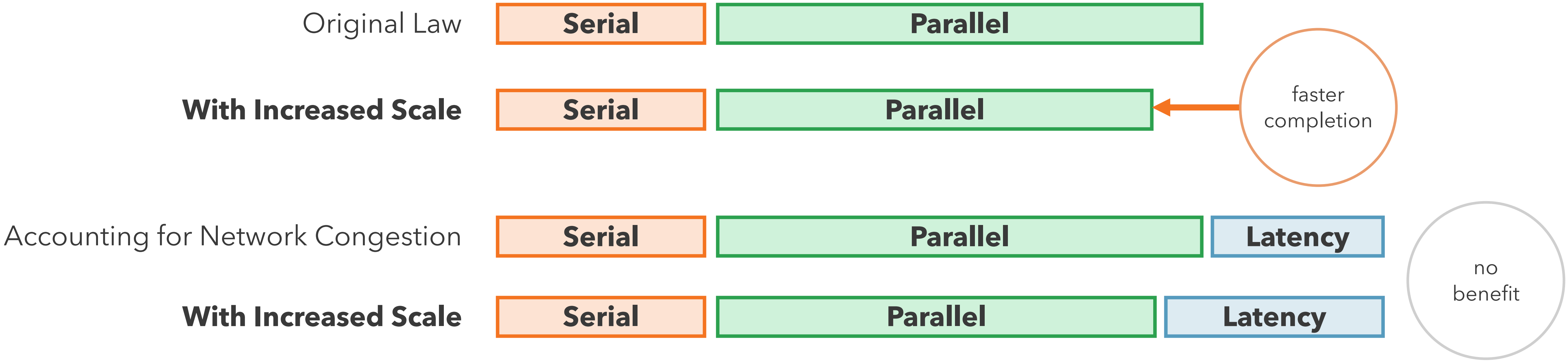
Matthew Williams, Rockport Networks, mwilliams@rockportnetworks.com
Monday, August 23rd, 2021

Agenda

- The Impact of Network Congestion
- The Rockport Architecture
- Rockport Benchmark Results – Release 1.0.1
- Questions

The Impact of Network Congestion

Amdahl's Law and How Congestion Limits Scale



- Amdahl's original law shows that there's a limit to how much parallelizing a task will improve completion times
- When network congestion is present, a third component "latency" needs to be added that is dependent on:
    - The level of network congestion
    - The amount of interprocess communication
    - The ability of the network to control latency under load
- Increasing workload scale will expand the amount of interprocess communication, limiting the performance gains of increased parallelization if network latency is not controlled

rockport.

The Broad Impacts of Network Congestion

Degradation of Workload Performance

Creates high tail latency

+

Extended and unpredictable workloads

+

Longer wait times for results

+

Workload scale is limited

Reduction in Workload Capacity

Longer to complete, fewer workloads can be run

+

Job queues get longer

+

Longer time to start

Unnecessary Cluster Inefficiencies

Idle Resources (i.e. CPU)

+

Workload costs unnecessarily increased

+

Less “work” can be done

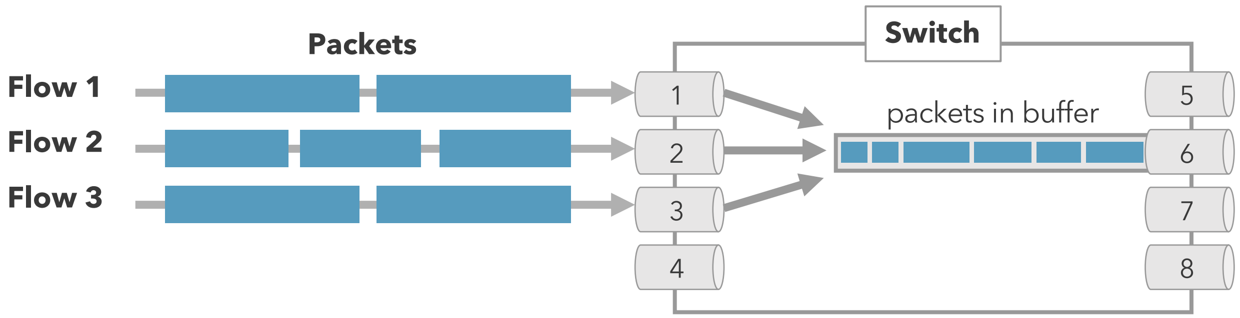
+

Reduces cluster ROI

You're Only As Good as Your Short Timeframe Response

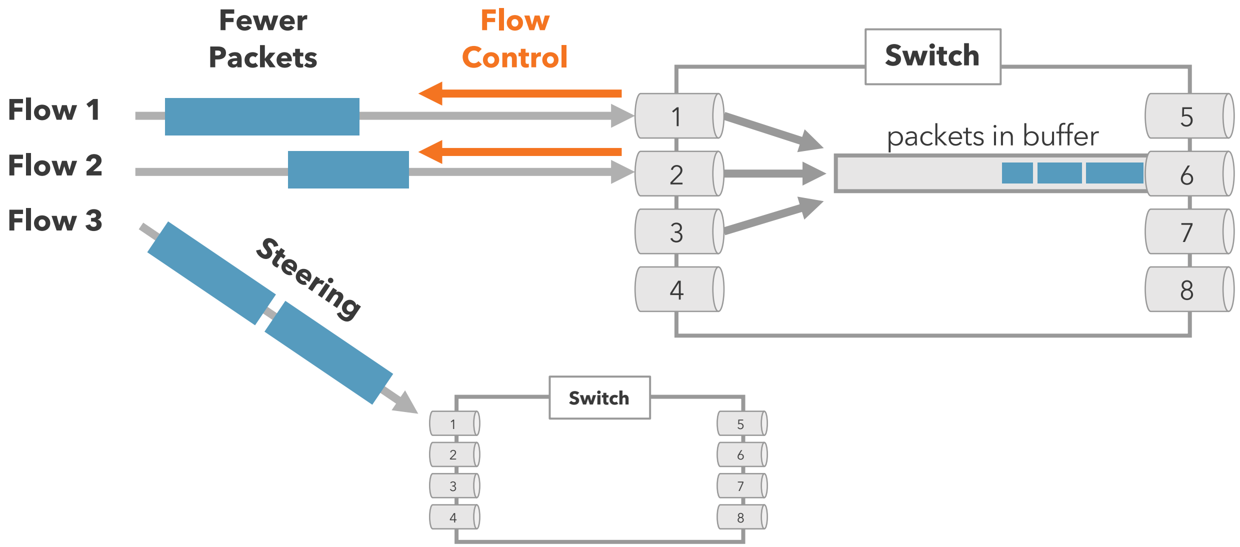
All Congestion Starts Out as Short Timeframe Congestion

Short Timeframe



- It is common for multiple flows to converge on the same switch output port
- At small timescales, the switch's only option is to buffer the excess traffic
- This leads to spikes in latency, driving high tail latency and extended workload completion times

Long Timeframe



- If the contention for the switch port lasts long enough, the network can react by:
  - Using flow control to slow flows
  - Steering flows away from the congestion (adaptive routing)

rockport.

The Rockport Architecture

rockport.

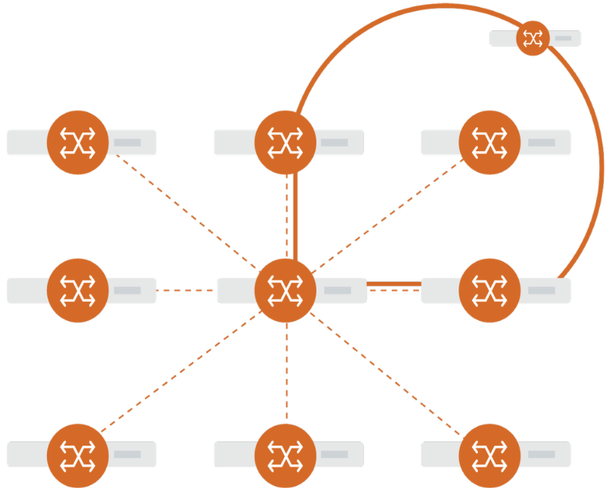
Simplify the Network

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
- Standard Ethernet-based host interface (RoCEv2 and TCP/UDP)
- Distributed routing and control planes
- Linear scaling
- 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
Distributed, embedded FLIT switching
Very High Path Diversity

rockport.

Scalable Supercomputer Networking, Simplified

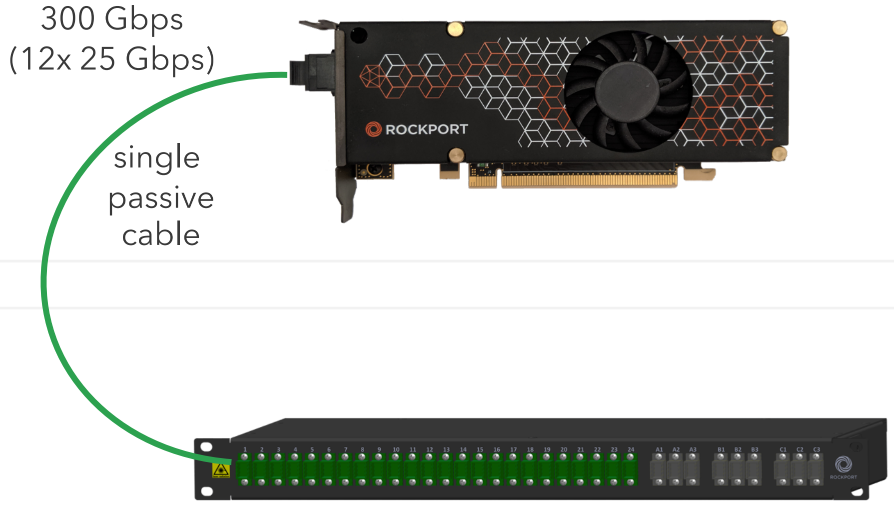
Rockport Switchless Network Solution

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

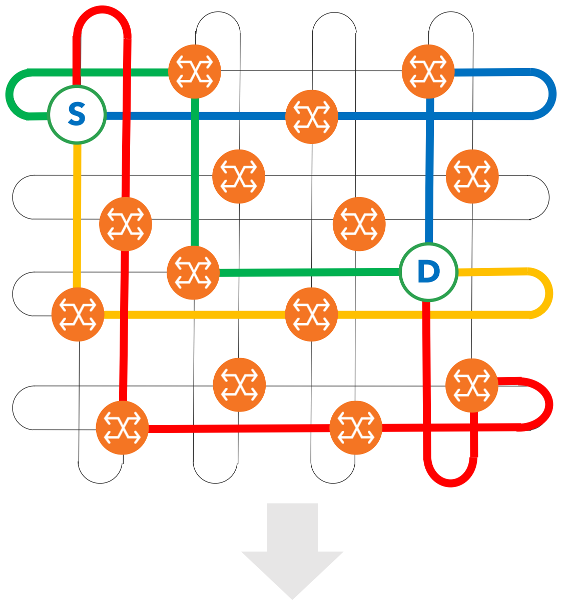
- Rockport distributed Deadlock-Free Routing algorithm (DFR)
  - Deadlock free routing across all topologies (complete or sparse)
  - Paths are physically independent and have no common links
  - Ensures high path diversity
- Traffic spread across all available paths on a per-flow or per-packet basis

Extremely Fast Distributed Switching

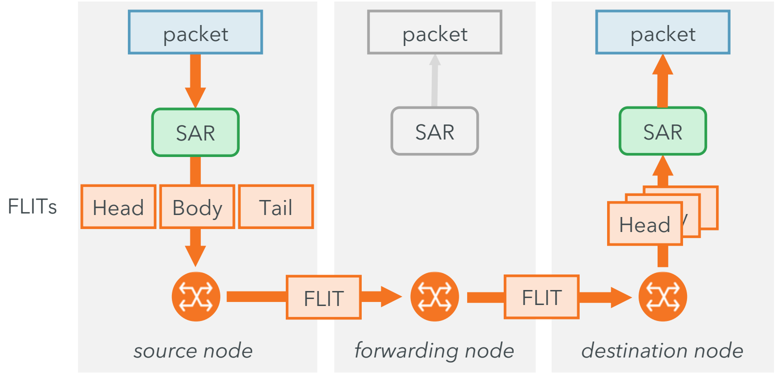
- Packets segmented into small pieces (FLITs)
  - Ensures very low latency performance, even under heavy load
- Embedded FLIT switching forwards FLITs to destination
- Destination reassembles packets

Inherent Performance Advantages

- Predictably low latency at every scale
- Zero congestive loss



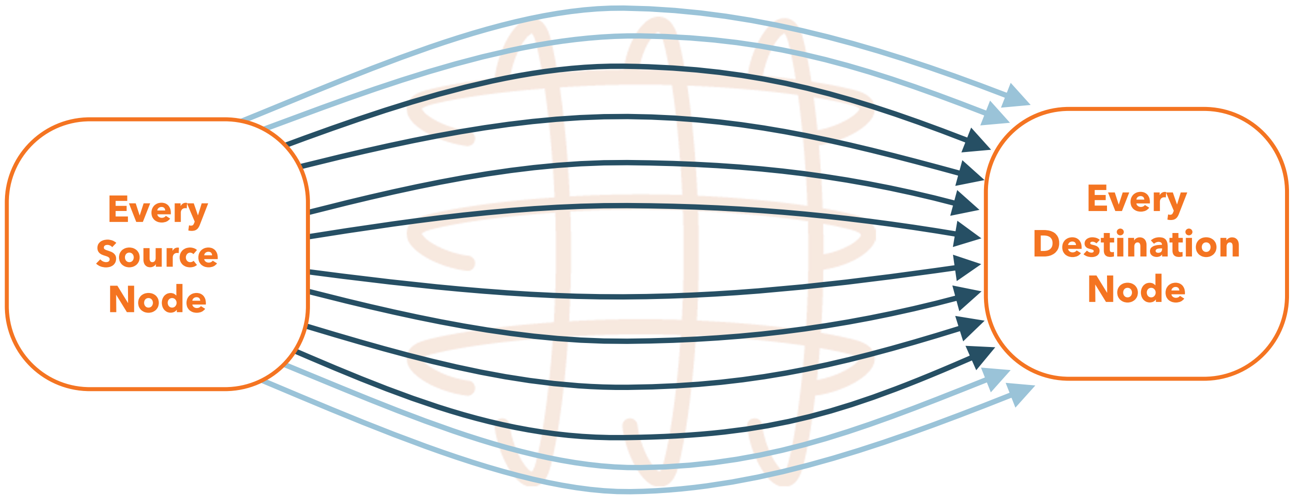
Distributed FLIT Switching



rockport.

Performance Advantages

Very High Path Diversity

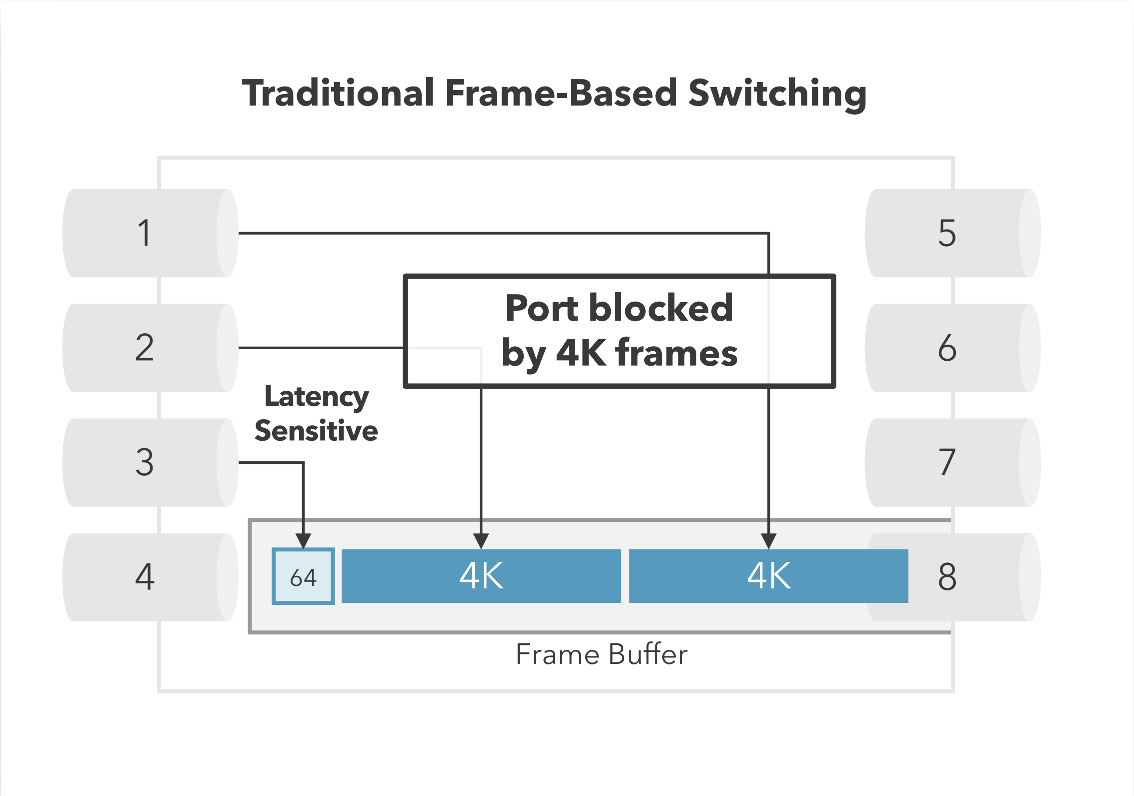


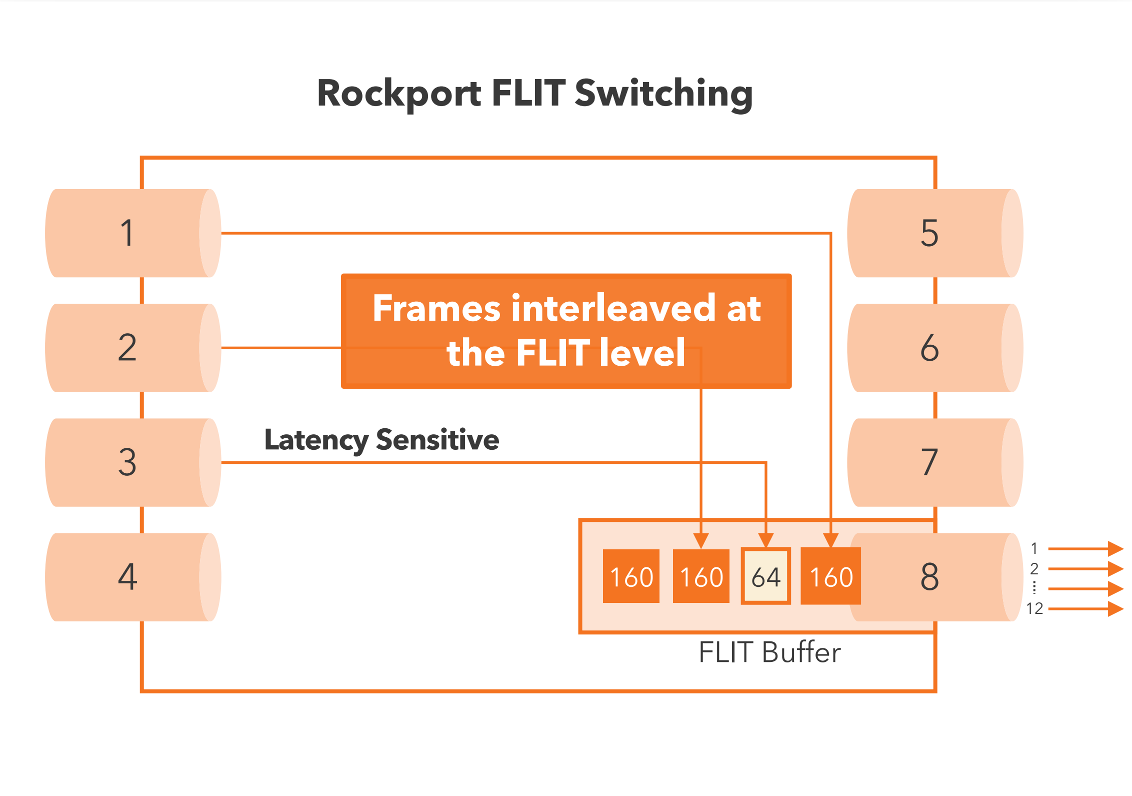
- High path diversity is a very important element of network design
- Rockport nodes distribute packets across the 8 optimal of 12 source routes to each destination to:
  - Distribute the network load across the topology
  - Avoid multiple congested paths through adaptive routing
  - Immediately react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

rockport.

Rockport Distributed FLIT Switching

Short and Long Timeframe Latency Advantages





Average blocking time for Latency Sensitive Packets is only 25 ns (50 ns max)

rockport.

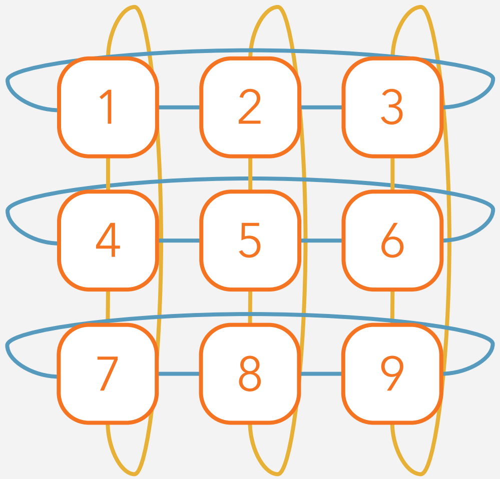
Performance Architecture

Initial Target Topologies

Rockport's initial target topologies are based on the 6D torus with some key enhancements

- Supports sparse, unbalanced topologies with easy scale-in/scale-out
- Distributed deadlock-free routing with high path diversity
  - Even with failed links or nodes
- Simplified, modular wiring approach
- Distributed operations: self-discovering, self-configuring, self-healing

e.g. 2D Torus

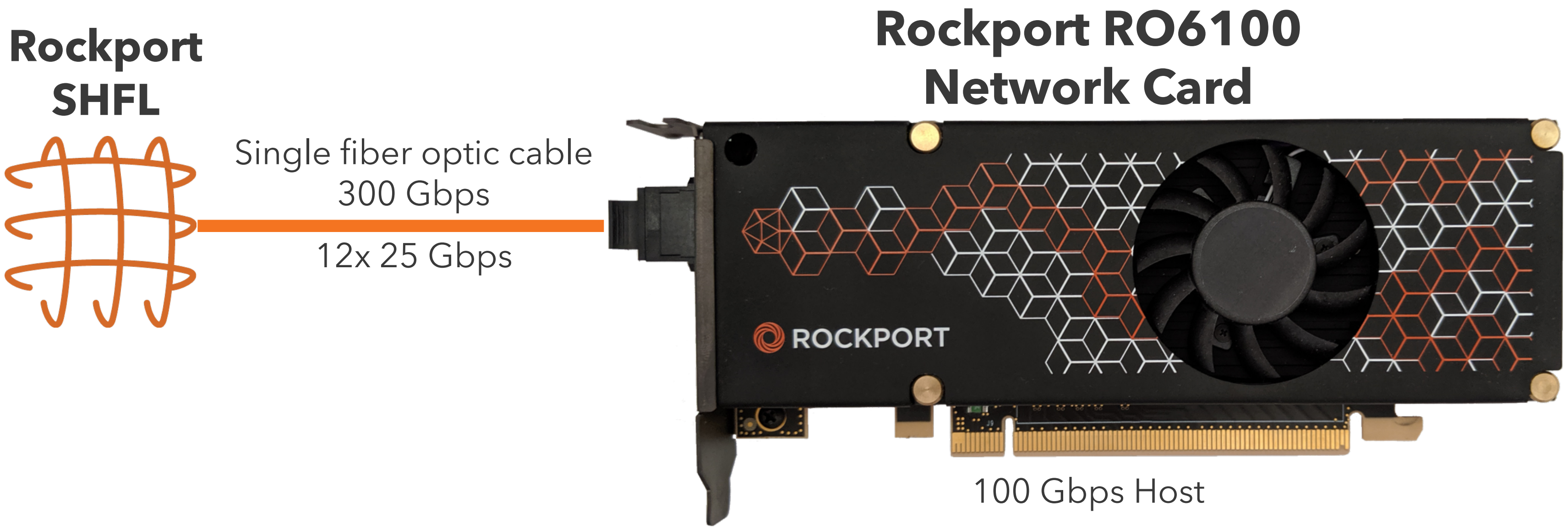


rockport.

Rockport Solution Components

rockport.

Rockport RO6100 Network Card

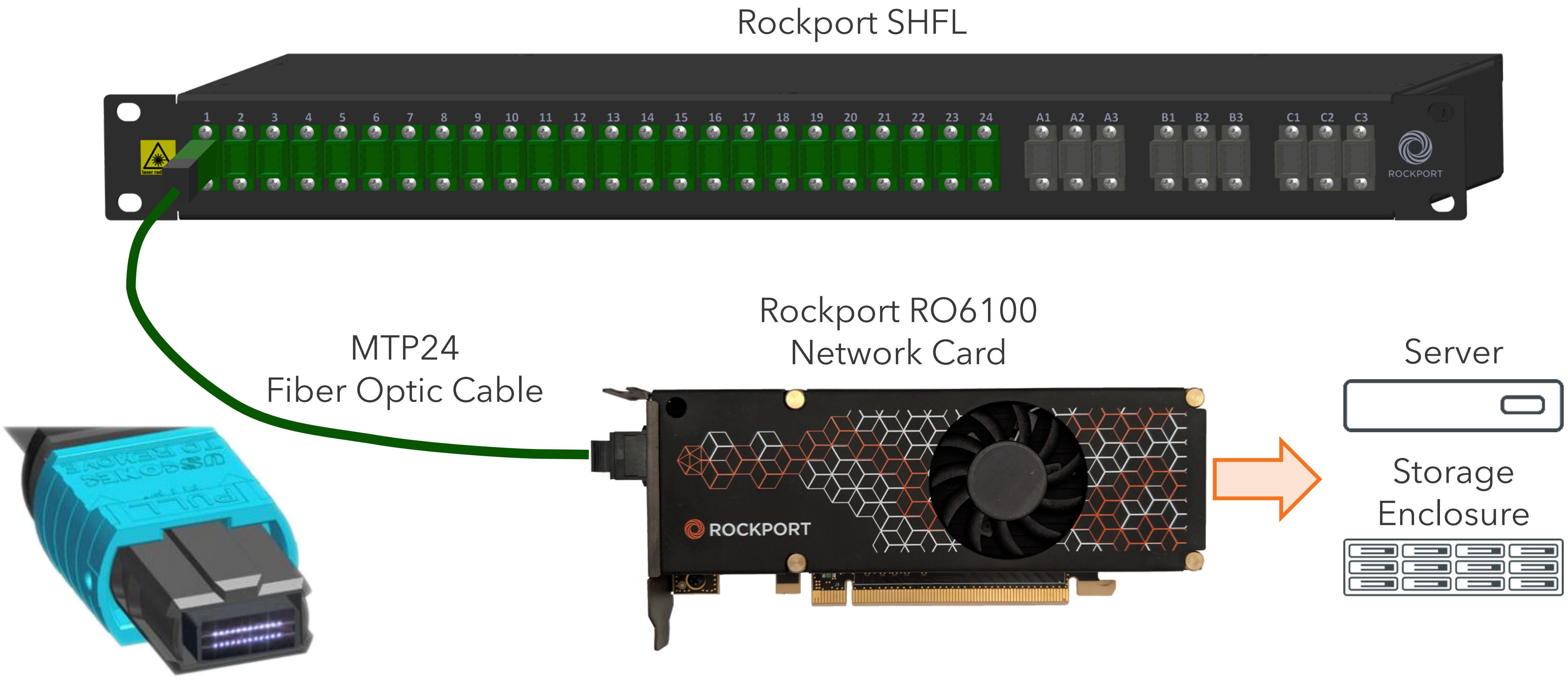


- Standard in-box drivers across Linux, VMware, Windows
- Appears to the operating system to be an industry standard Ethernet NIC
  - Sockets (TCP/UDP) and verbs (RoCE) API support
- 300 Gbps (12x 25 Gbps) network links in a single fiber optic cable
- 100 Gbps host bandwidth
- All Rockport Networks functions (dataplane, control plane, etc.) performed by embedded hardware

rockport.

Rockport Simplicity

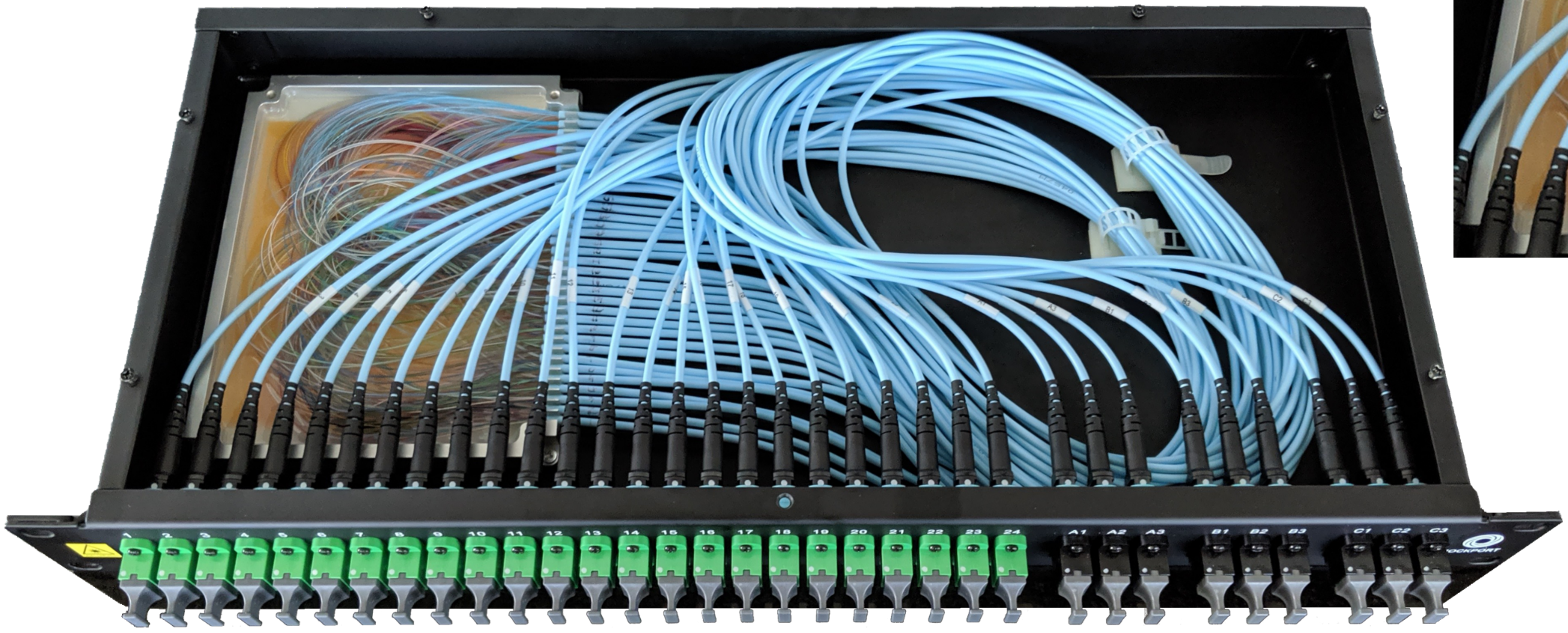
Simple Deployment

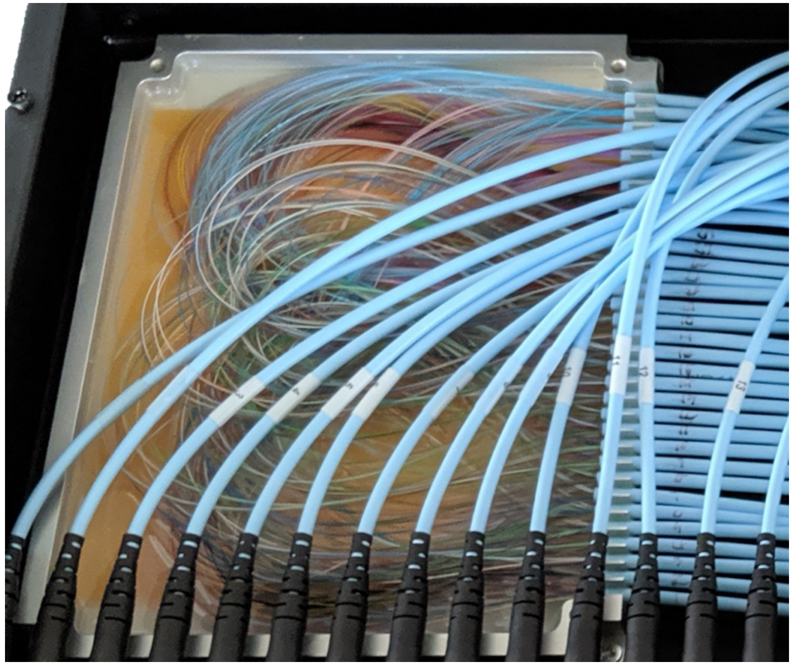


Rockport Simplicity

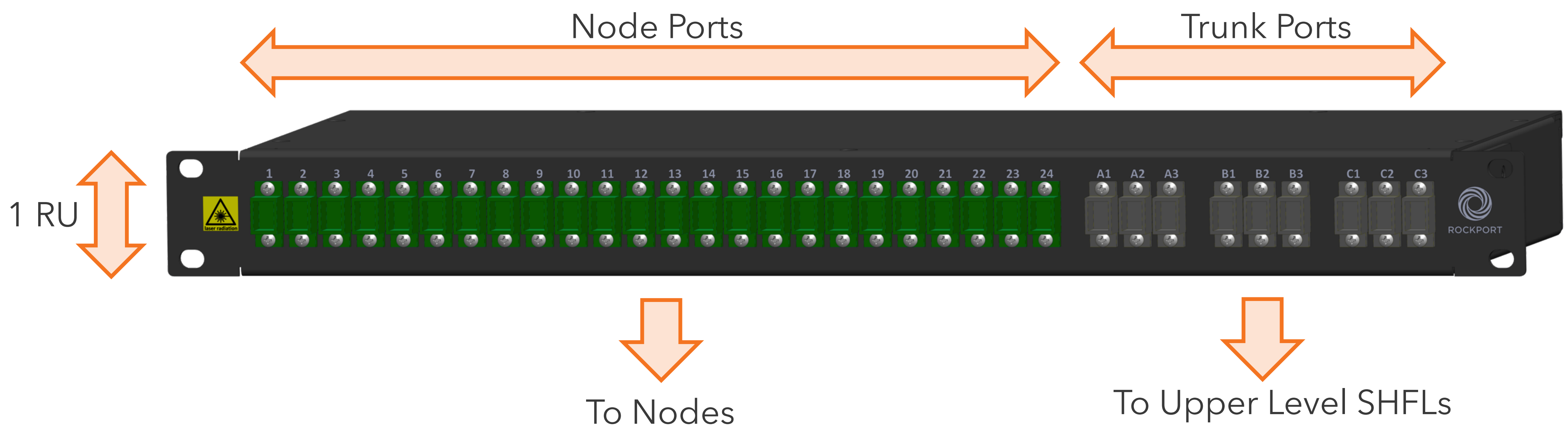
Rockport SHFL

- Topology complexity hidden from end users
- Single fibre optic cable to any node
- Modular system to fit different rack configurations





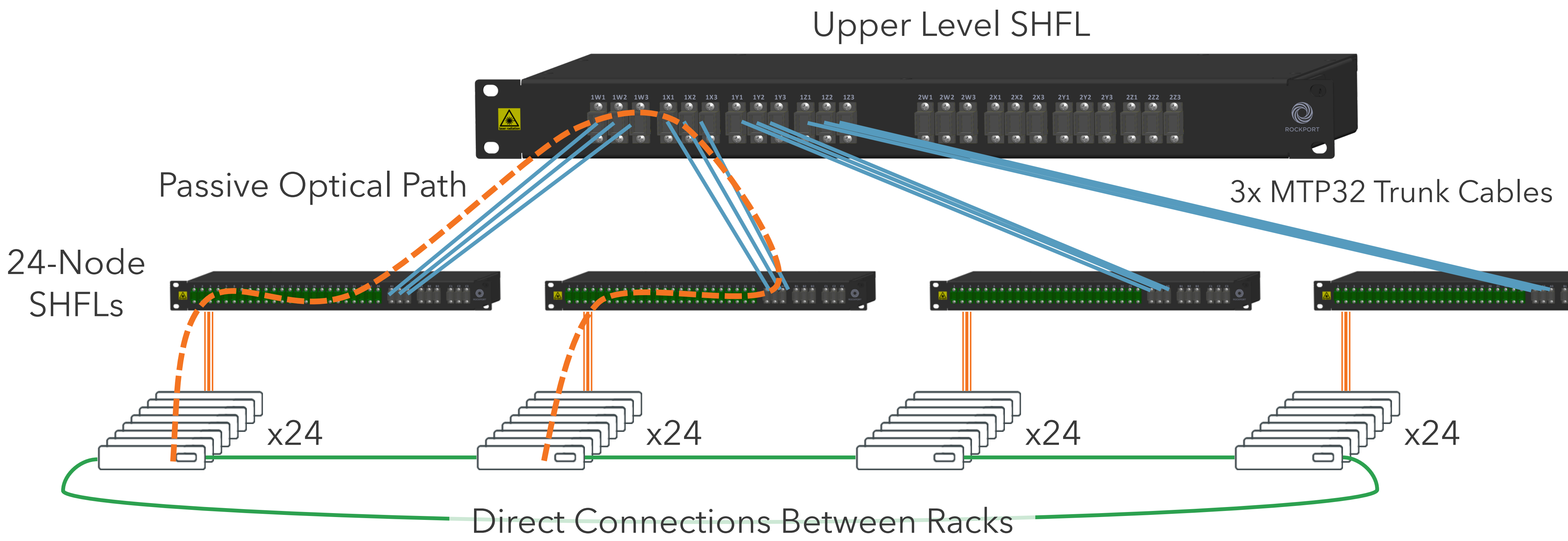
Rockport 24-Node SHFL



- Topology complexity hidden from end users
- Single fibre optic cable to any node
- Modular system to fit different rack configurations

rockport.

Example 96-node Deployment



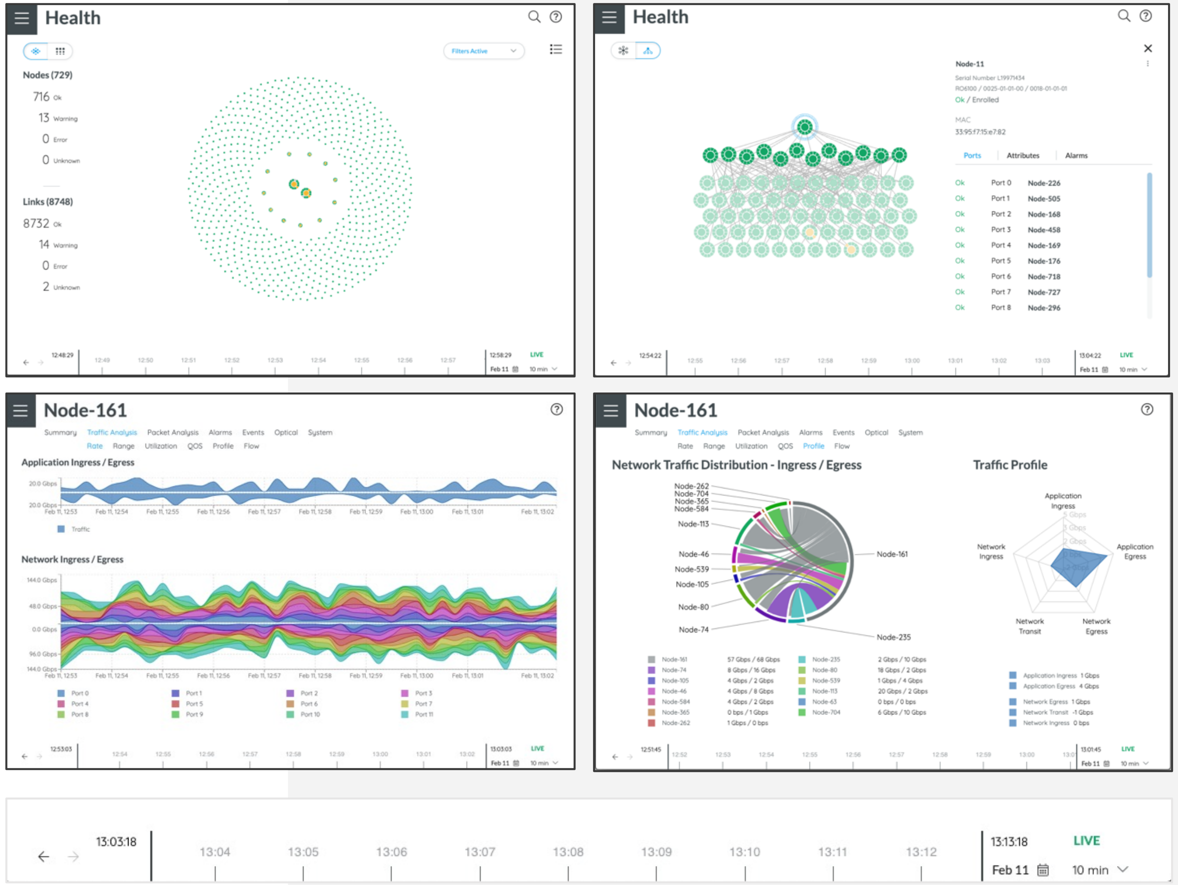
- Three fiber optic cables connect each 24-node SHFL to the Upper Level SHFL
- Passive Upper Level SHFL creates rings between nodes attached to different 24-node SHFLs
- **Direct inter-rack connectivity removes requirement to place workloads in the same rack**
  - **No locality restrictions for workload placement**

rockport.

Autonomous Network Manager

Management Simplicity

- Configure, manage, and troubleshoot the Rockport Switchless Network
- Intuitive user interface, visualizations and single dashboard approach to provide real-time health and performance monitoring
- RESTful APIs to retrieve reporting, monitoring, and management data with easy integrate with existing monitoring tools
- Temporal database
  - 7 days of full metrics storage
- SNMP traps
- Scalable architecture
- Secure design



rockport.

Benchmarks

Release 1.0.1

rockport.

Benchmark Results

Network Performance Testing Under Load

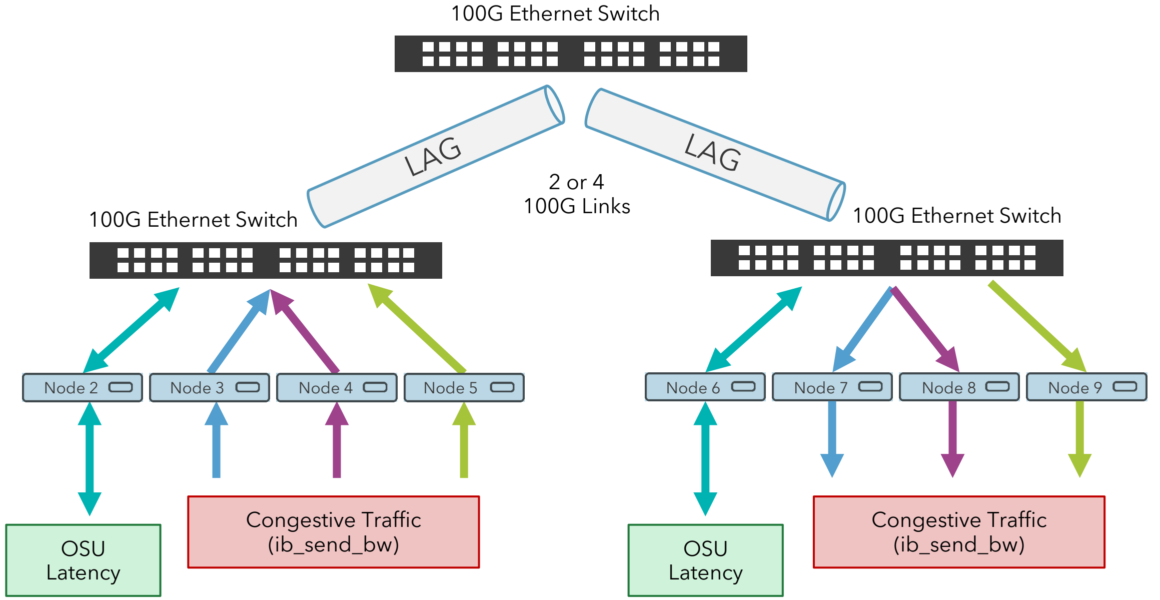
- Typical network benchmarks run on unloaded networks and only provide a baseline, best-case view into the performance of the network
  - Unloaded = network dedicated to benchmark with no competing network traffic
- These baseline results are not useful to predict the performance of the network in a 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

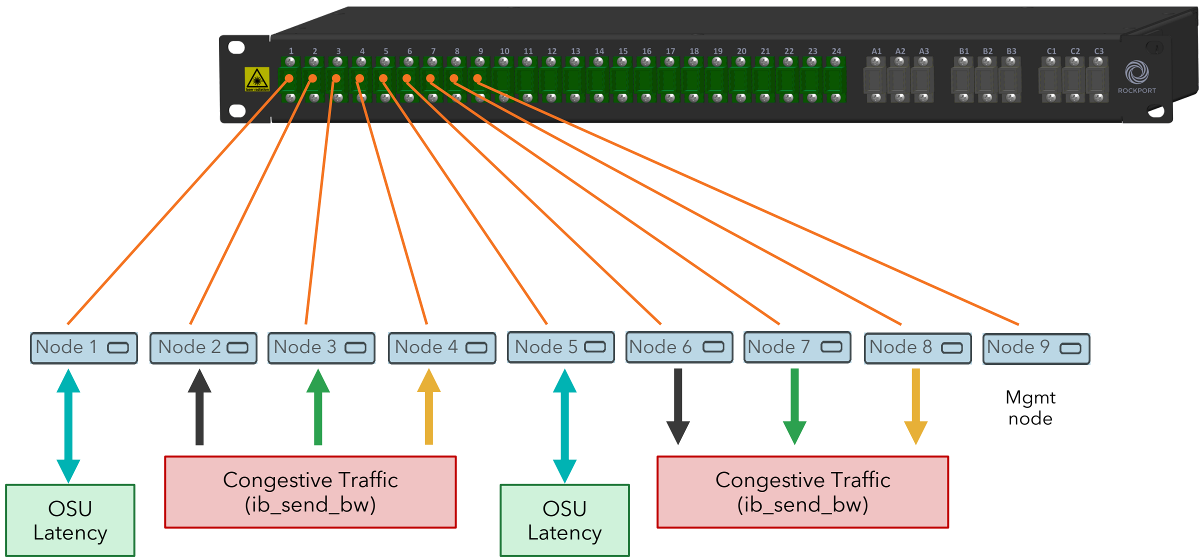
OSU Unloaded and Loaded Latency vs Traditional Ethernet

rockport.

Benchmark Results

Test Setup with Traditional Ethernet





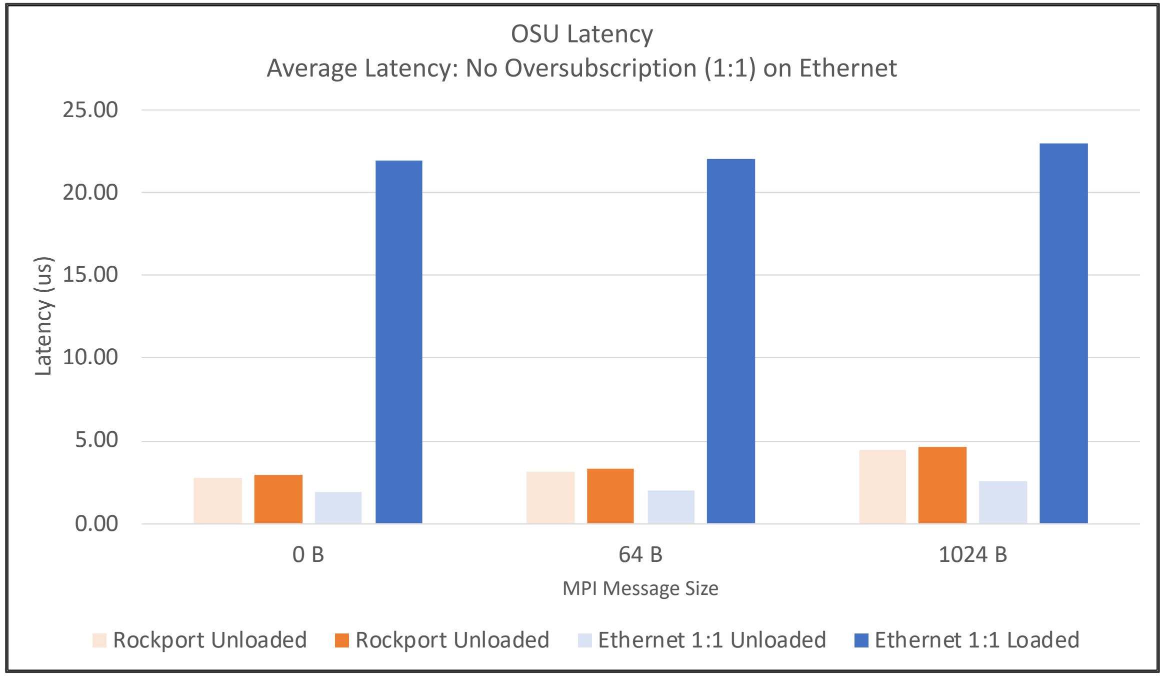
Traditional Ethernet

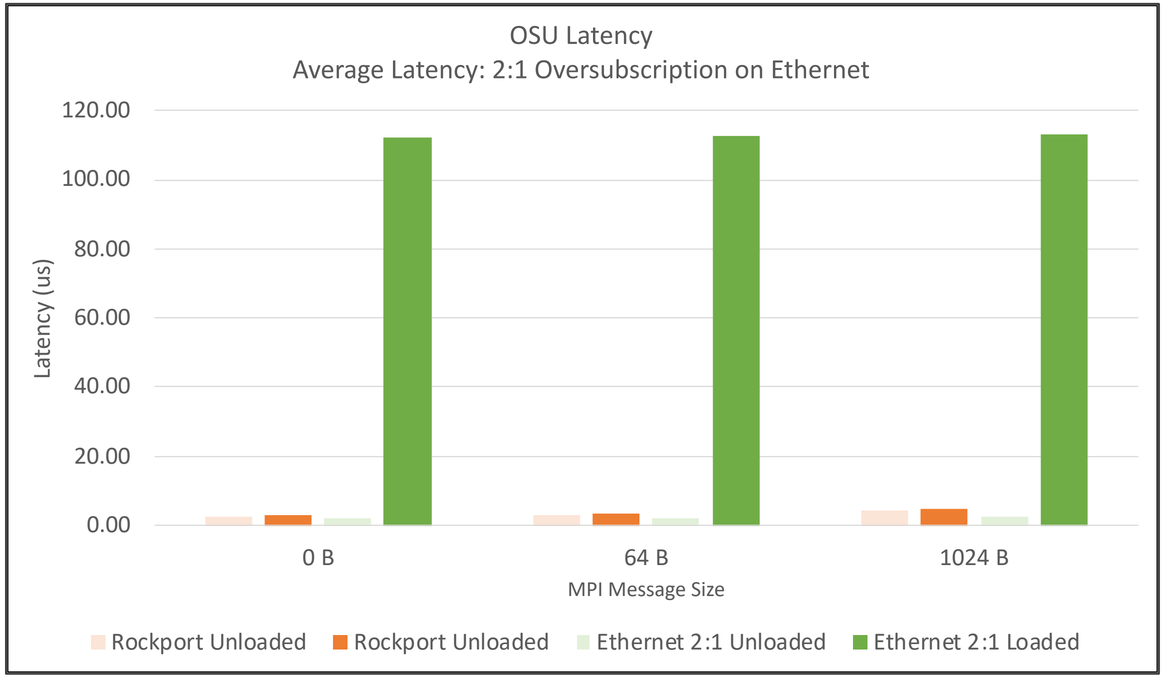
OSU Latency benchmark between Nodes 2 and 6 in four scenarios	
No Oversubscription (4 Uplinks/Leaf)	2:1 Oversubscription (2 Uplinks/Leaf)
Unloaded: No other traffic in network	Unloaded: No other traffic in network
Loaded: 3 x ib_send_bw	Loaded: 3 x ib_send_bw

Rockport Switchless Networking

OSU Latency benchmark between Nodes 2 and 6 in two scenarios
Unloaded: No other traffic in network
Loaded: 3 x ib_send_bw -q4

Unloaded and Loaded Latency Results vs Traditional Ethernet





Low Latency under Load, Predictable Performance

rockport.

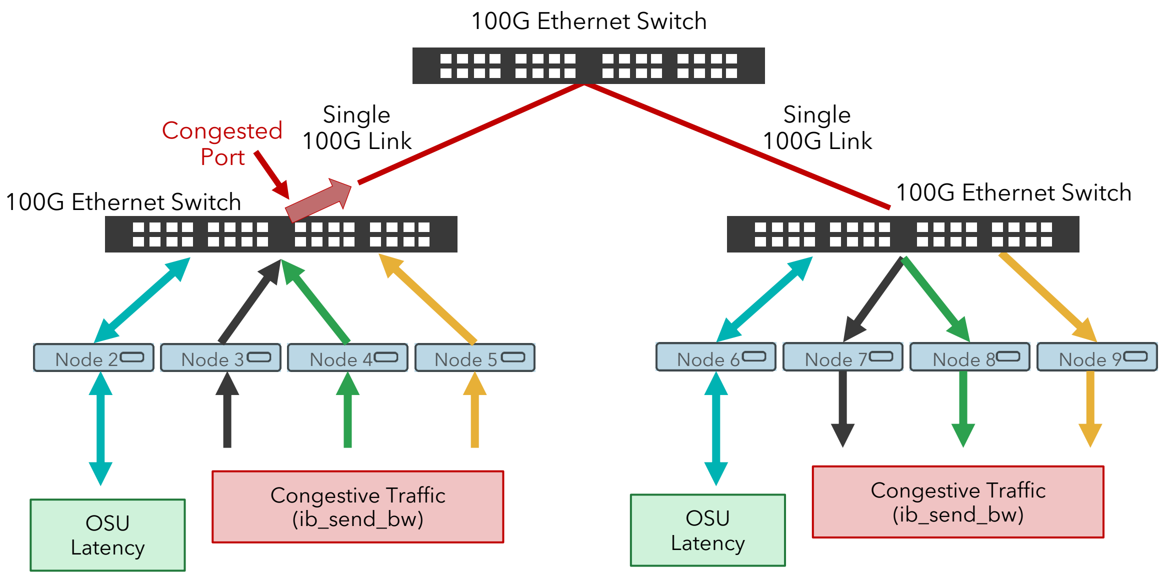
Restricted Path Testing vs Traditional Ethernet

rockport.

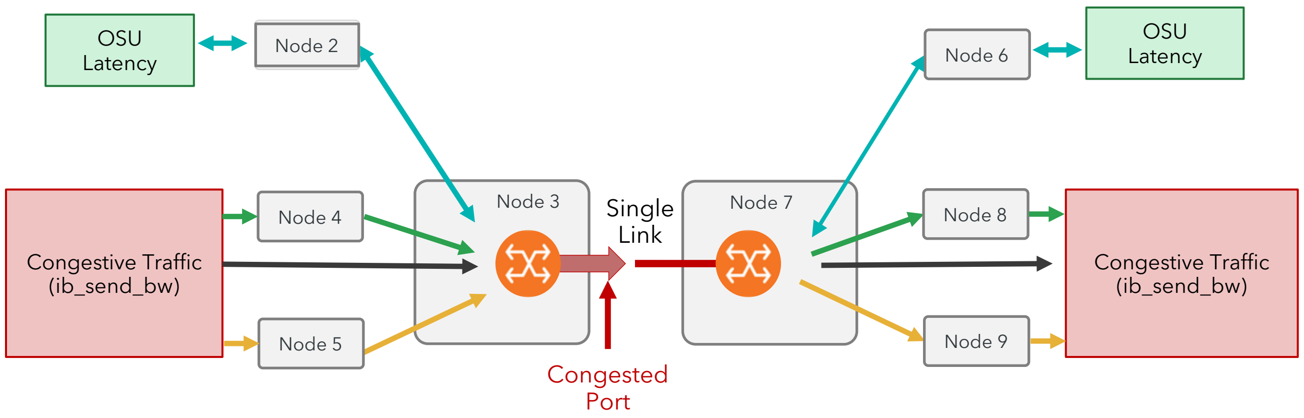
Benchmark Results

Restricted Networks Paths Under Load vs Traditional Ethernet

Artificially restrict each environment to a single path and add congestive traffic



Rockport network links disabled to force all traffic over a single link



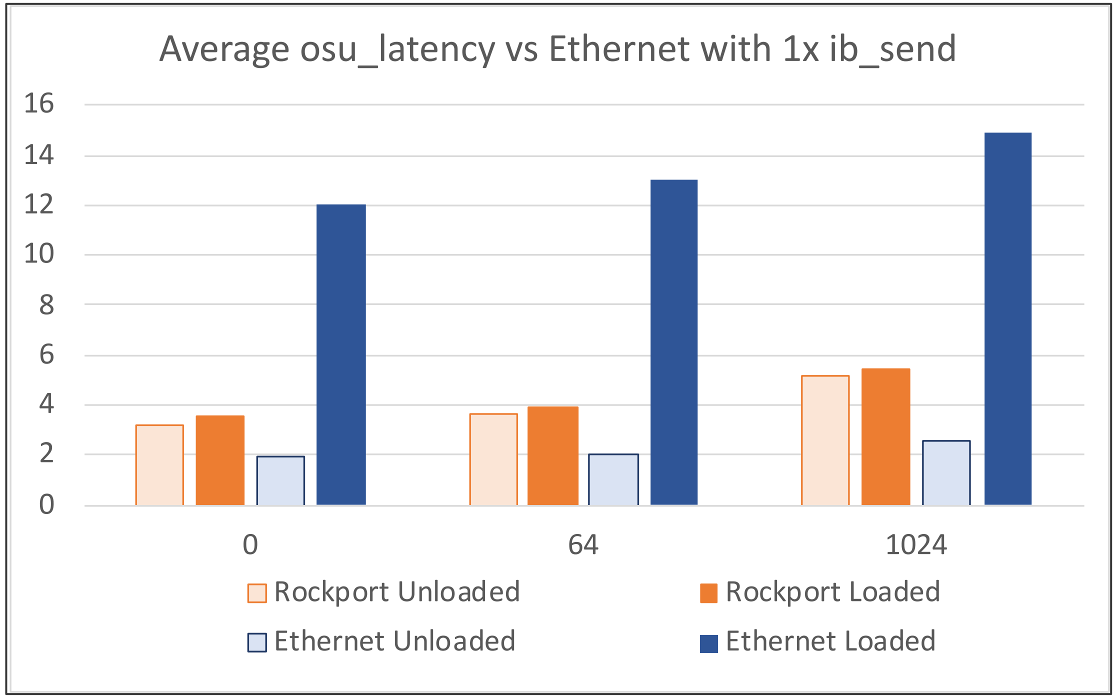
OSU Latency benchmark between 2 nodes with 3 sets of network conditions:

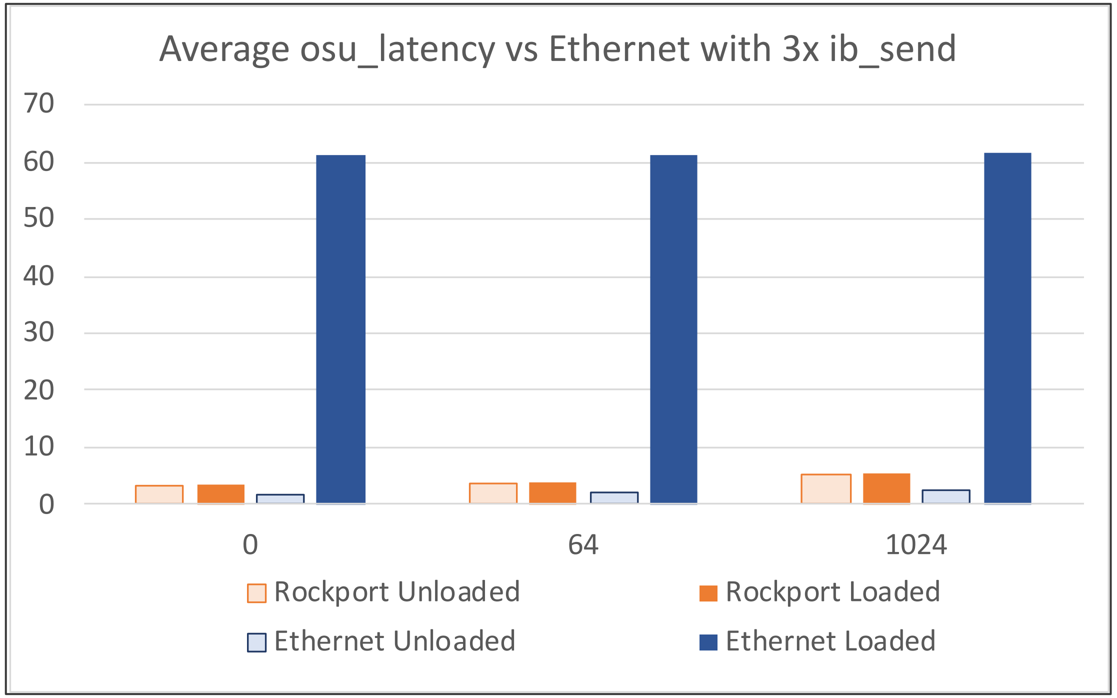
1. No other traffic on the network (unloaded)
2. With IB Send between one pair of nodes across the single link
3. With IB Send between three pairs of nodes across the single link

rockport.

Benchmark Results – Current Release

Restricted Path Unloaded and Loaded Performance vs Traditional Ethernet





Low Latency under Load, Predictable Performance

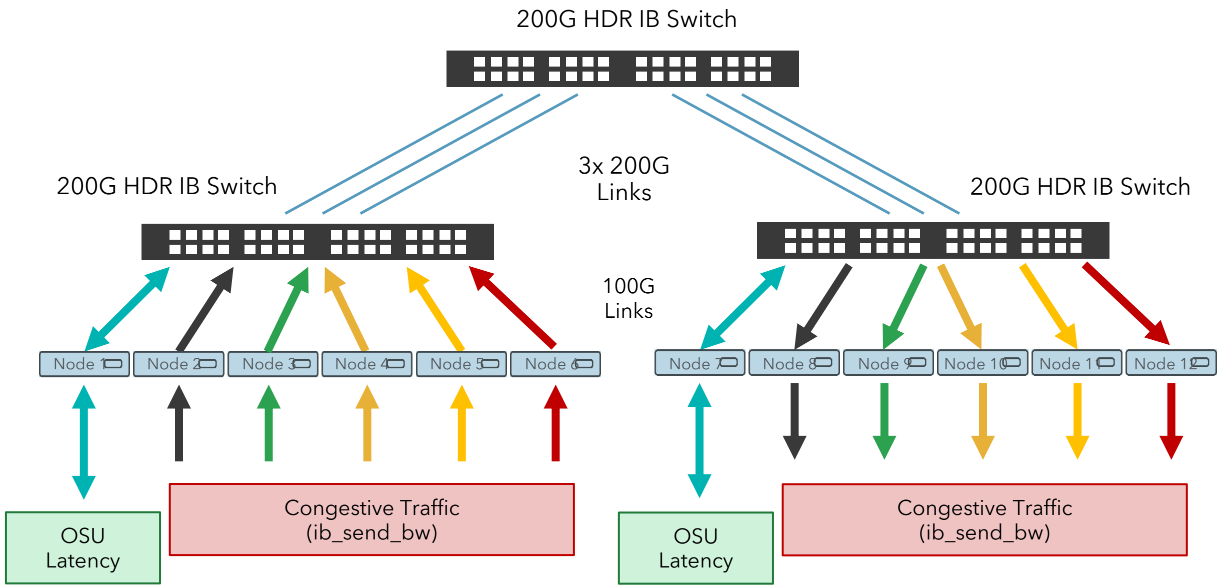
rockport.

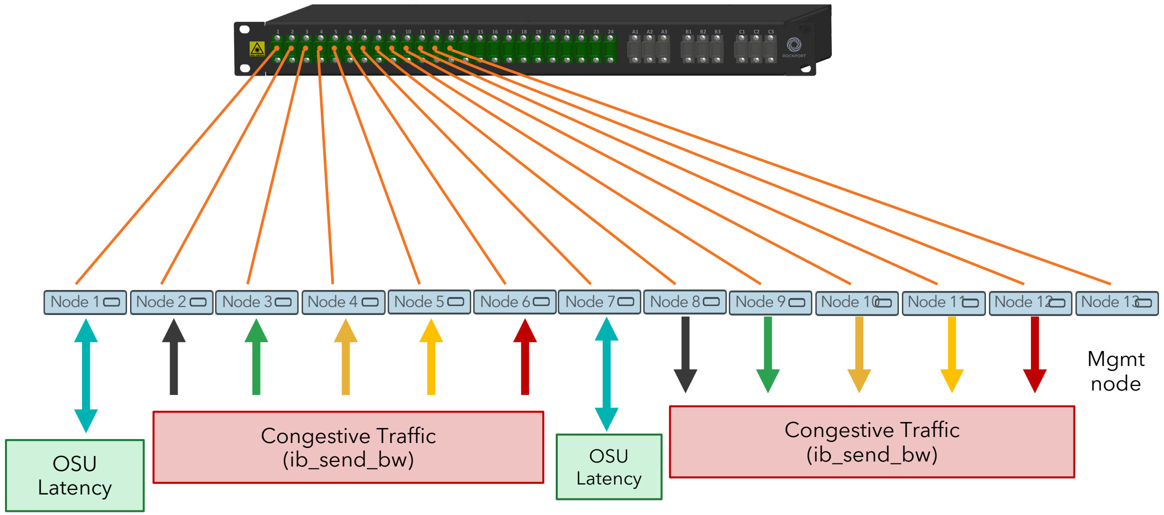
OSU Unloaded and Loaded Latency vs InfiniBand

rockport.

Benchmark Results

Test Setup vs InfiniBand





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

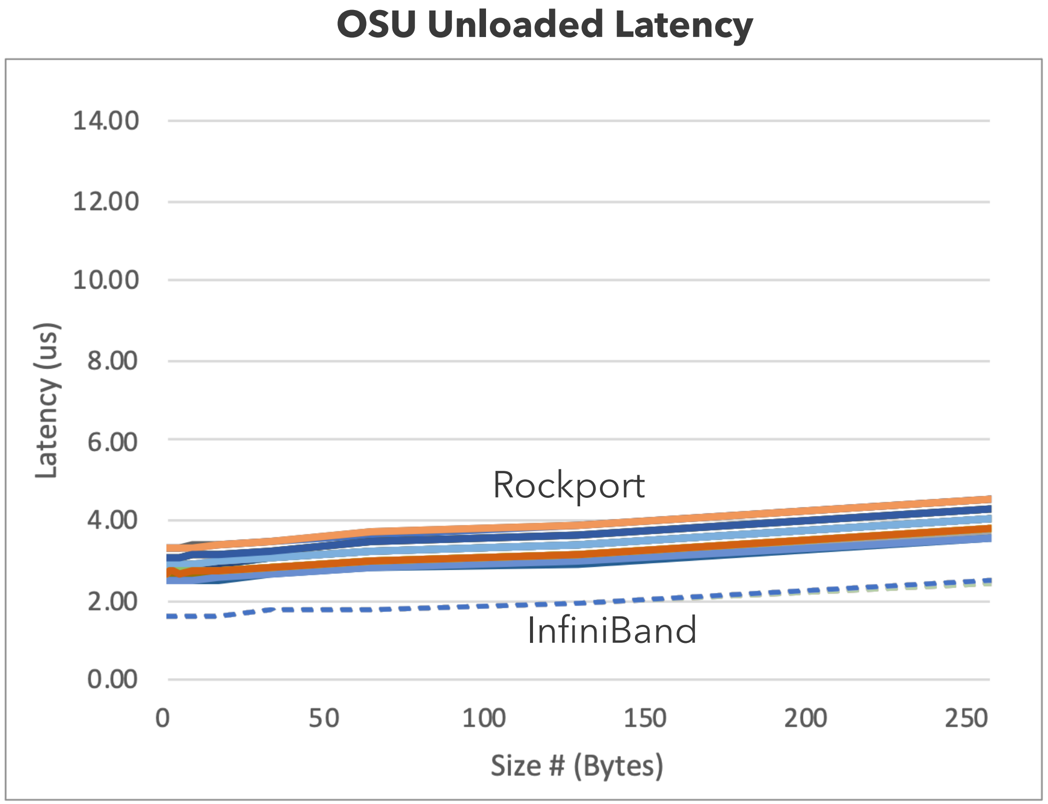
Rockport

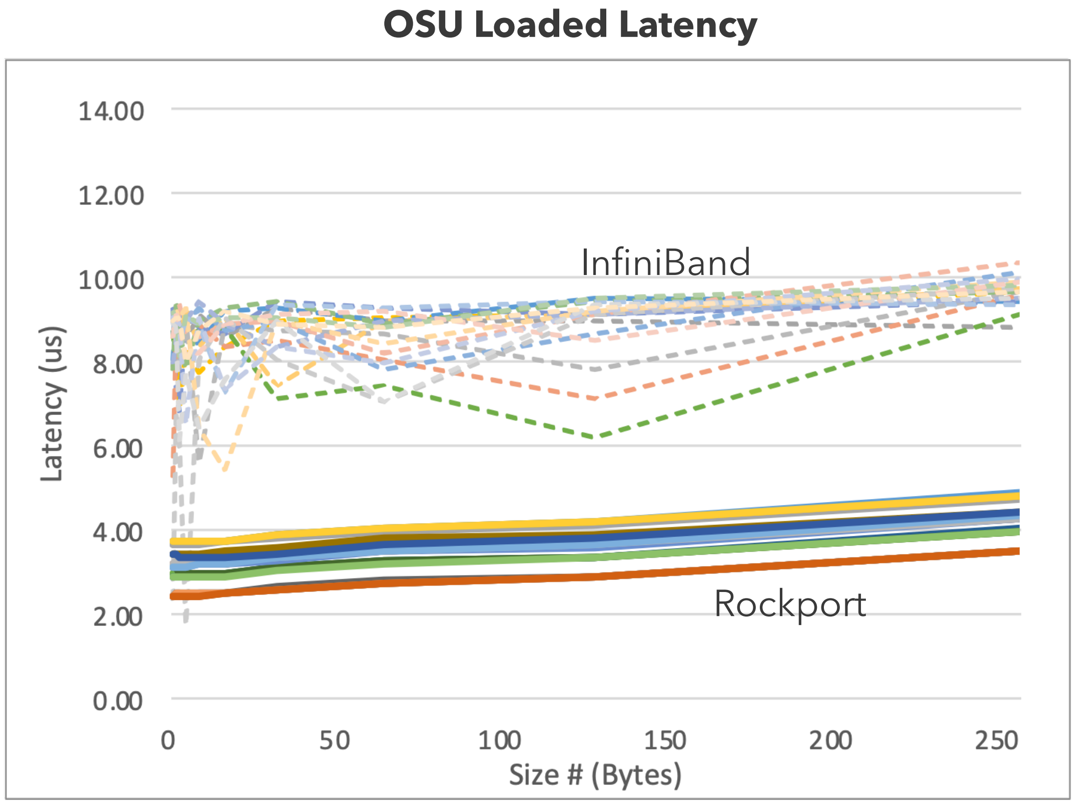
OSU Latency benchmark between Nodes 1 and 7 in two scenarios
Unloaded: No other traffic in network
Loaded: 5x ib_send_bw -q4

rockport.

Benchmark Results – Current Release

Unloaded and Loaded Performance Test Setup vs InfiniBand





Low Latency under Load, Predictable Performance

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

rockport.

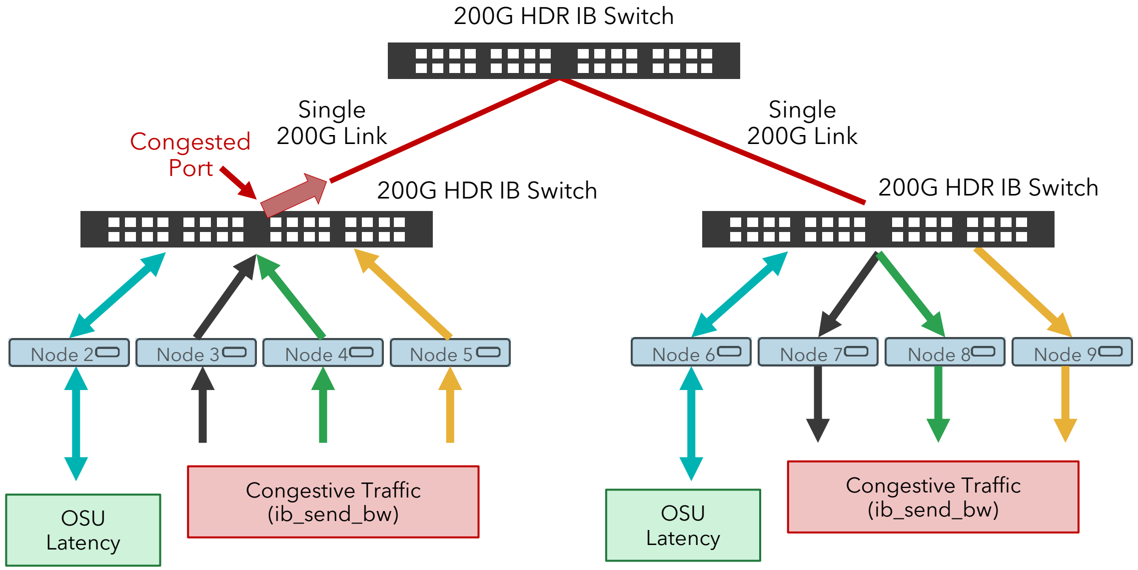
Restricted Path Testing vs InfiniBand

rockport.

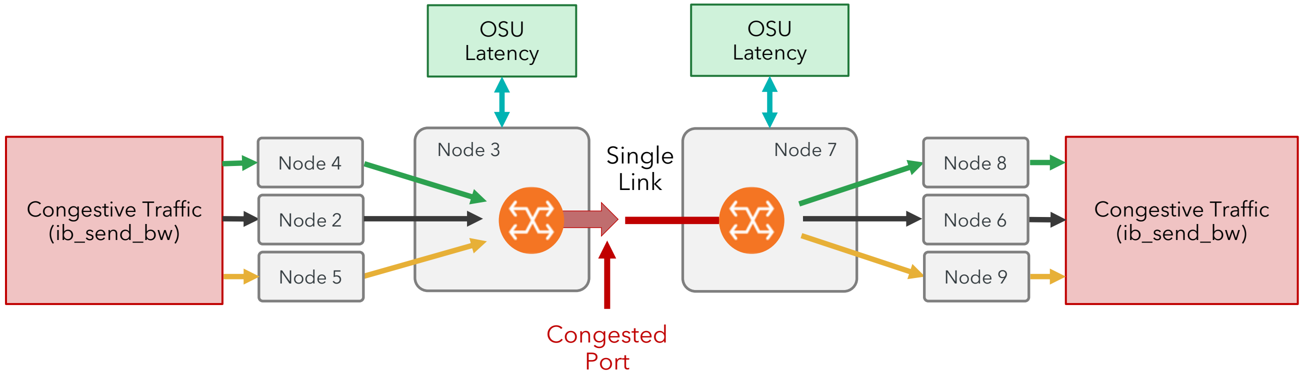
Benchmark Results

Restricted Networks Paths Under Load vs InfiniBand

Artificially restrict each environment to a single path and add congestive traffic



Rockport network links disabled to force all traffic over a single link



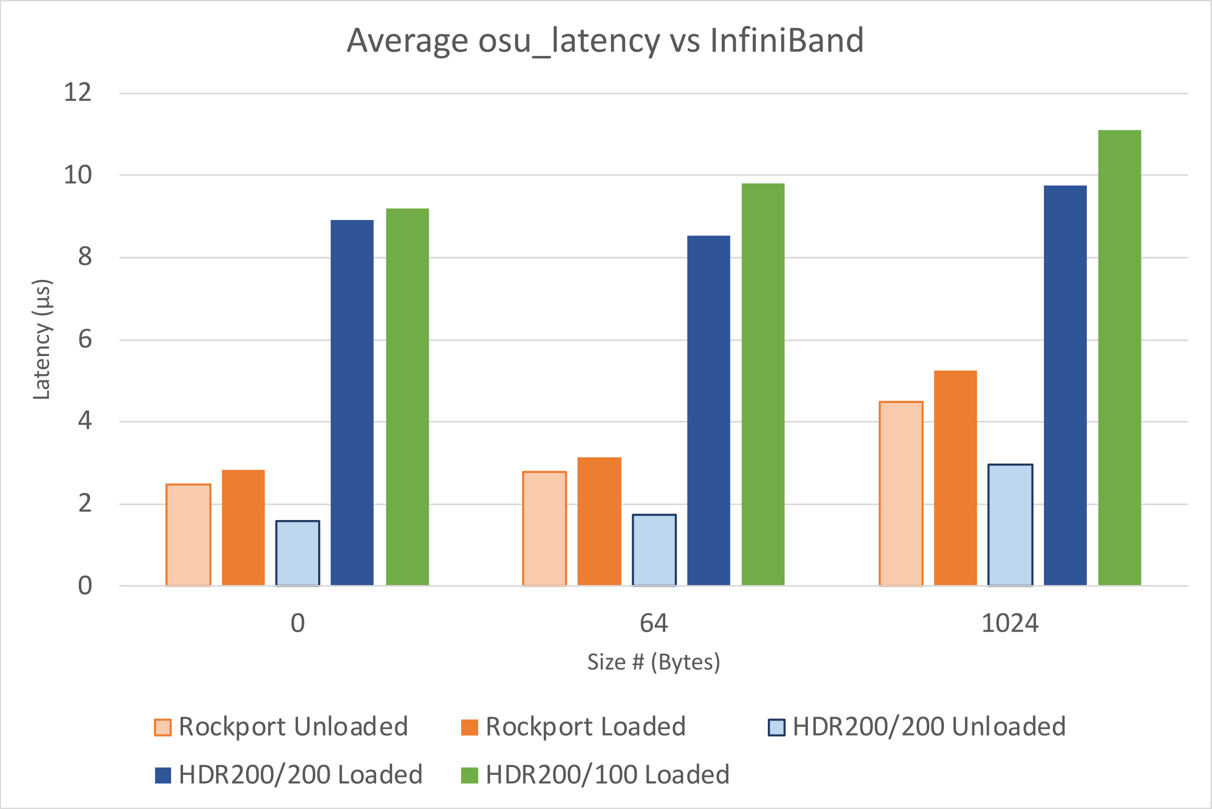
OSU Latency benchmark between 2 nodes with 2 sets of network conditions:

1. No other traffic on the network (unloaded) with HDR200 host links
2. With IB Send between three pairs of nodes across the single link with HDR200 host links
3. With IB Send between three pairs of nodes across the single link with HDR100 host links

rockport.

Benchmark Results – Current Release

Restricted Path Unloaded and Loaded Performance vs InfiniBand



Low Latency under Load, Predictable Performance

rockport.

GPCNeT

rockport.

Benchmark Results

GPCNeT

- Congestive network performance benchmarks
  - 2019 paper – Argonne National Lab, Lawrence Berkeley National Lab, Cray

- Latency and bandwidth performance tests using a canary workload
  - First set of runs on an unloaded network
  - Second set of runs on a loaded network with 4 unique congestion patterns

- The Congestion Impact is the ratio of loaded and unloaded latency performance
  - i.e. how much worse does a network perform under load
  - A congestion impact of 1.0x is ideal as it means that there is no difference in measured performance between unloaded and loaded networks

$$\text{Congestion Impact} = \frac{\text{loaded performance}}{\text{unloaded performance}}$$

rockport.

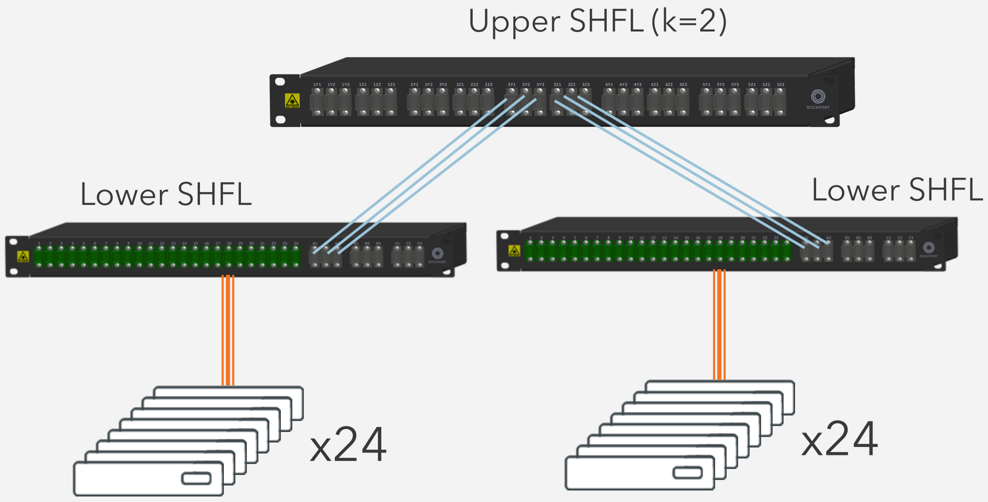
Benchmark Results – Current Release

GPCNeT System Configuration

Cluster Details

- **48 server cluster**
  - Dual-socket AMD EPYC7302 16-core @3.0 GHz
  - 1536 total cores
- **OpenMPI 4.10**
- **GPCnet 1.2**

Network Configuration



- **Two 24-node Lower SHFLs (LS24T) connected via a k=2 Upper SHFL**
- **Topology based on a 4x3x2x2 torus**

rockport.

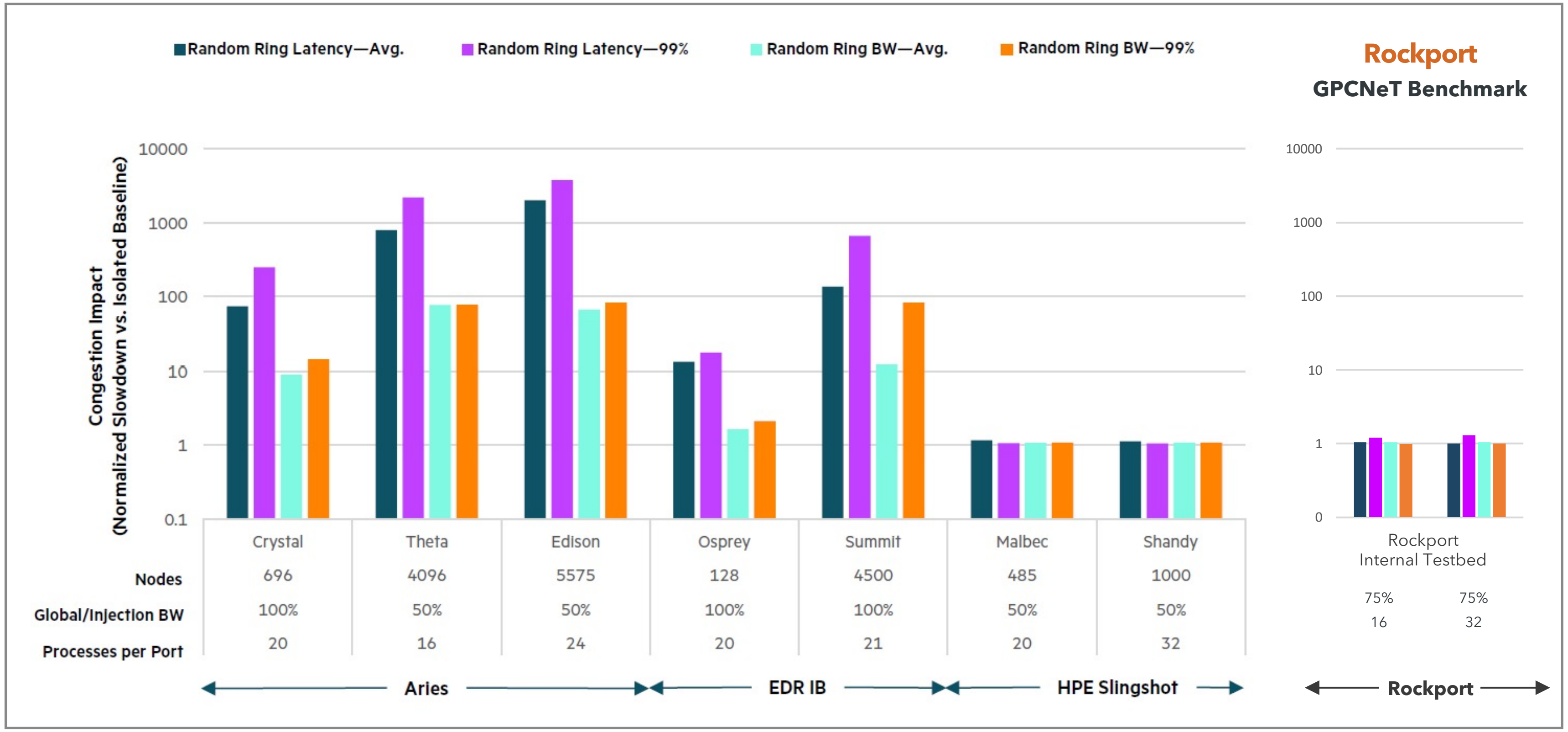
Benchmark Results – Current Release

Rockport GPCNeT Results

	RR Latency	RR Bandwidth	All Reduce
Average	1.0x	1.0x	1.0x
99%	1.3x	1.0x	1.2x

- Very strong demonstration of Rockport's latency consistency under load
- Larger scale testing underway

Note: Charts are Logarithmic scale - base 10 and scaled equally.



GPCNeT: Designing a Benchmark Suite for Inducing and Measuring Contention in HPC Networks: https://escholarship.org/uc/item/17m1r82n
Measuring Network Performance to Better Manage It https://psnow.ext.hpe.com/doc/a50002193enw?jumpid=in_lit-psnow-red

rockport.

Summary

The real cost of congestion being exposed
New switchless direct interconnect
Buyers looking for loaded measures to gauge performance

Don’t miss our MUG talk on Tuesday, August 24 @ 2pm EDT:

Upcoming MVAPICH2 Design Enhancements on the Rockport Switchless Network

rockport.

Thank You.
Questions?

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

rockport.

rockport.