rockport.

Rockport Networking Solutions for Large-Scale HPC and AI systems

Higher Performance by Design

Matthew Williams, CTO
mwilliams@rockportnetworks.com

ABOUT ROCKPORT NETWORKS

Simplify the Network

> Rockport has re-imagined the network with an embeddable switchless architecture that delivers the performance needed for enterprise, AI/ML, and HPC workloads

Global Coverage

NA & EMEA Centers of Excellence

20+ Switchless Networks

85+ Issued Patents

rockport.

The Network is the Problem

Today's High-Performance Networks are considered the single biggest issue known to throttle compute, storage, and application performance - especially at scale

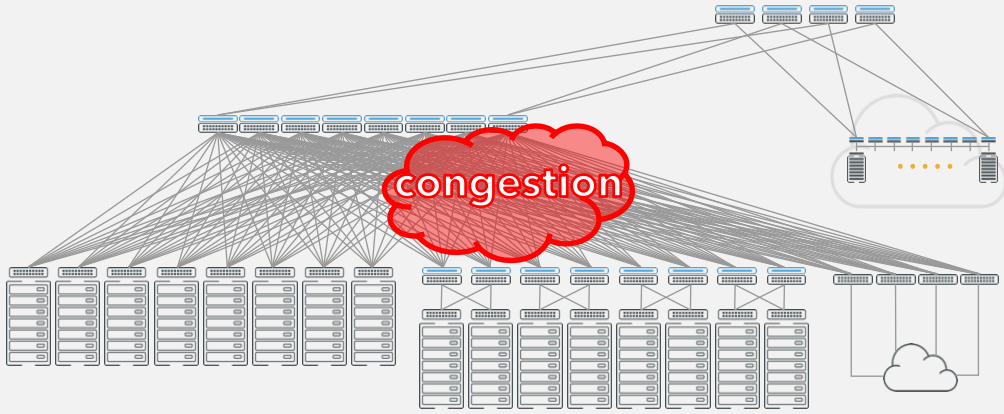
Traditional spine and leaf architectures can dilute effective bandwidth by up to **68%**

The network can degrade workload performance between **20-30%**

Congestion leaves expensive compute and storage resources sitting **idle**

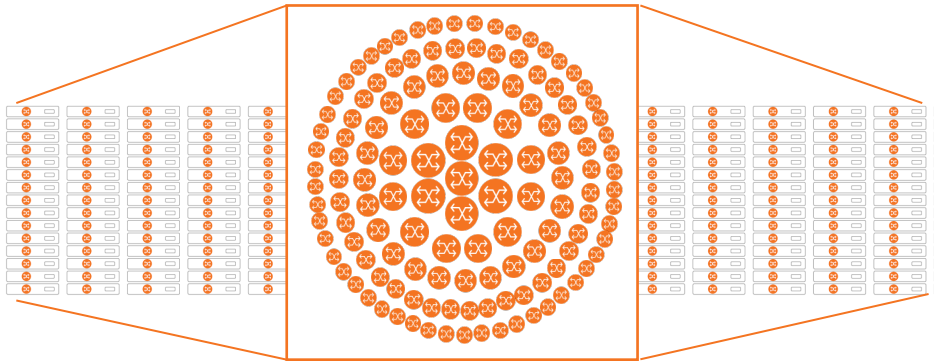
Moving the Network to the Endpoints

Traditional
High Performance Spine-and-Leaf



1-2 network ports per node
Nodes connected to external switches
Layers and layers of switches

Rockport
Switchless Network



12 network ports per node
Direct connections between nodes
NO EXTERNAL SWITCHES

rockport.

Simplify the Network

High Performance by Design

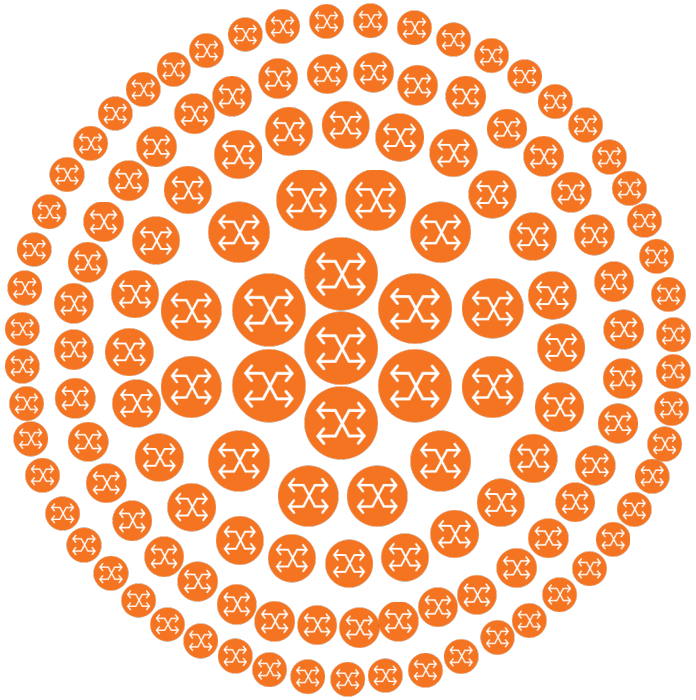
By removing the switch and distributing the network switching function into each device endpoint, **the nodes become the network:**

- Standard Ethernet-based host interface (RDMA/RoCEv2 and TCP/UDP)
- Supports all MPI libraries and Parallel Programming Models
- Per-packet adaptive routing
- Advanced congestion controls
- Critical messages immune to congestion
- Linear, elastic scaling

Rockport provides an embeddable switchless architecture that delivers a **300 Gbps Fabric** enabling the performance at the scale needed for HPC, HPDA, and AI/ML/DL.

Rockport Switchless Architecture

Directly Connected Nodes - Distributed Switching - Very High Path Diversity



PCIe for Servers and Storage Enclosures

Storage Devices and Enclosures Computational Storage

Integrated - Other Devices (FPGAs, ASICs)

rockport.

Main Solution Components

Rockport Switchless Network

Rockport NC1225 Network Card

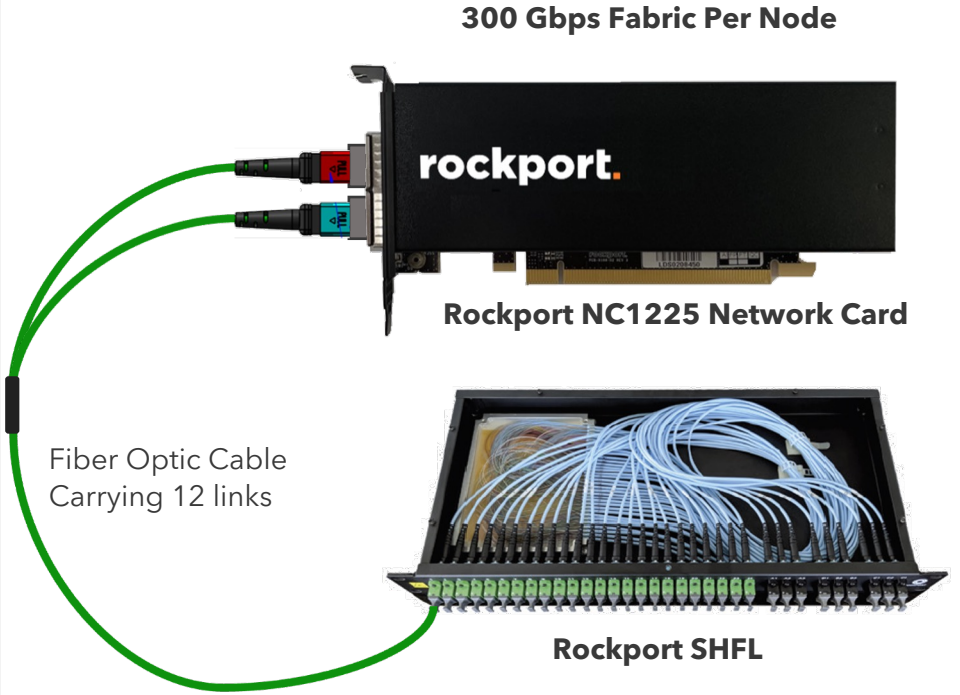
- World's first Network Card – 300 Gbps Fabric Per Node
- Standard High Performance Ethernet interface (verbs and sockets)
- Runs rNOS (Rockport Network Operating System) in a field-upgradable FPGA

Rockport SHFL

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

Rockport Autonomous Network Manager (ANM)

- World's first autonomous direct connect network manager
- Bird's eye view into active network
- Deep insight into network performance on a per-job basis



Rockport Autonomous Network Manager (ANM)

rockport.

Smarter Network Fabric

Performance rNOS Fabric

Topology Discovery

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

Distributed Source Routing

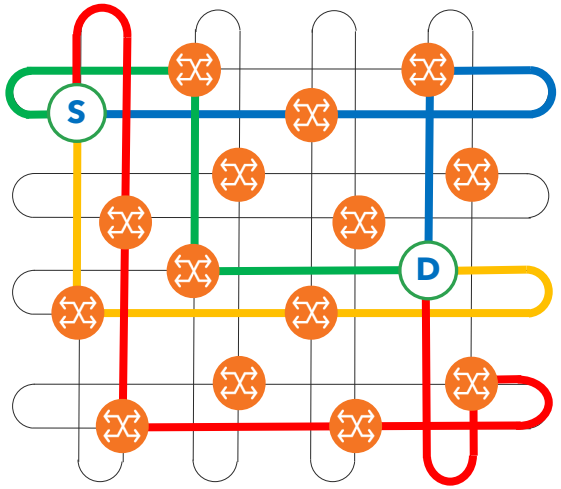
- Rockport distributed Deadlock-Free Routing algorithm (DFR)
  - Deadlock free routing across all topologies
  - Paths are physically independent and have no common links
  - Ensures high path diversity
- Traffic spread across all available paths on a per-flow or per-packet basis

Extremely Fast Distributed Switching

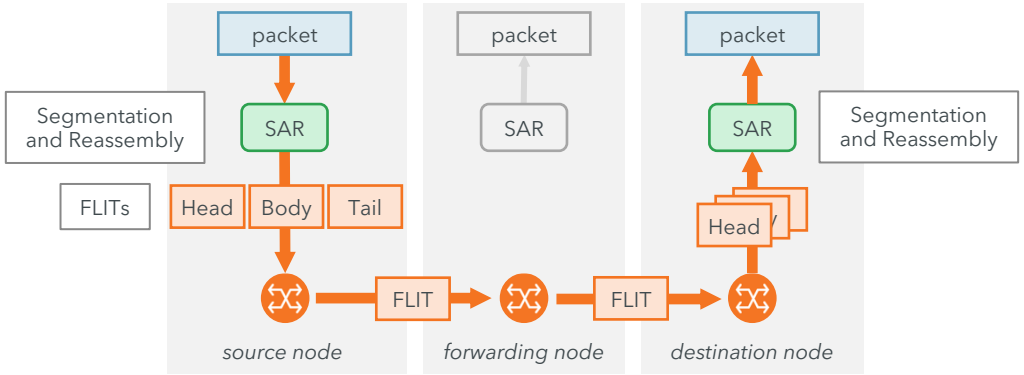
- Packets segmented into small pieces (FLITs)
  - Ensures very low latency, even under heavy load
- Embedded FLIT switching forwards FLITs to destination
- Destination reassembles packets

Inherent Performance Advantages

- Predictably low latency at every scale
- Zero congestive loss



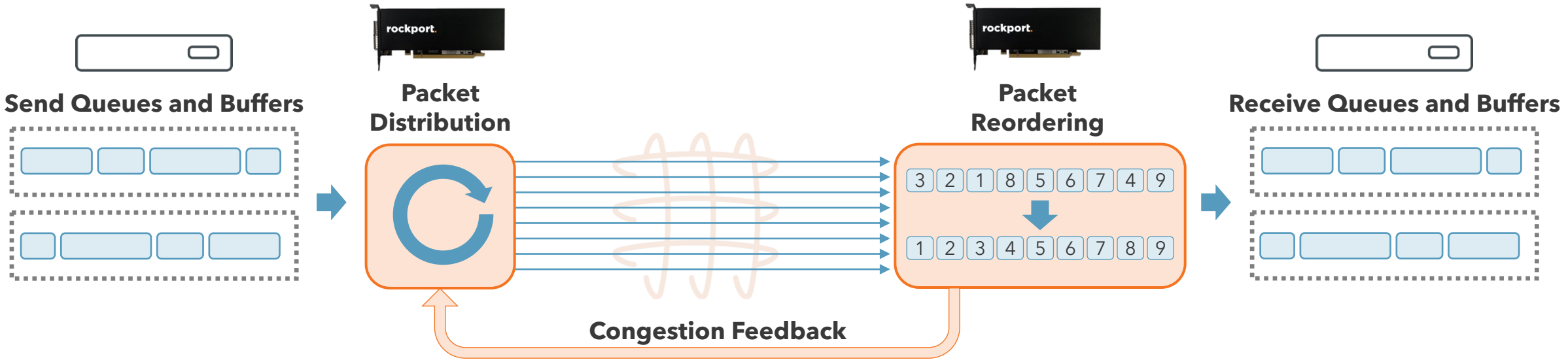
Distributed FLIT Switching



rockport.

Higher Performance by Design

Adaptive Multi-Path Routing with End-to-End Path Monitoring

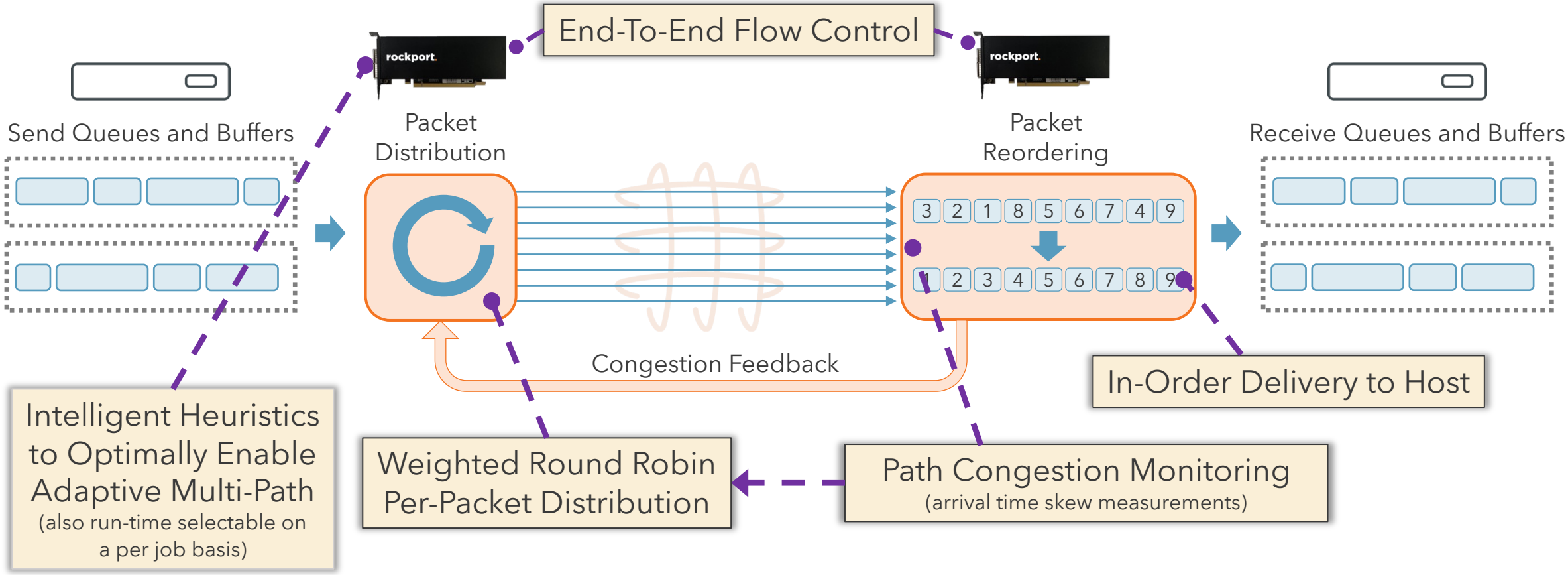


- Traffic adaptively distributed across the 8 optimum physically independent paths **on a per-packet basis**
  - Based on the end-to-end congestion along each path through real-time feedback
- Allows traffic to concurrently make use of the bandwidth of all 8 paths
- Works with all transports (RC/UD/DC) as packet order is guaranteed
- **In Rockport's Exascale solutions, the number of concurrent parallel paths will be 16 or higher**

rockport.

Higher Performance by Design

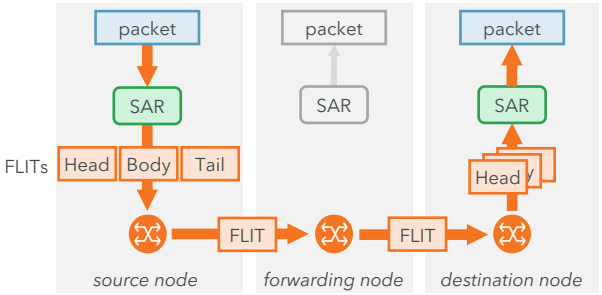
Adaptive Multi-Path Routing with End-to-End Path Monitoring

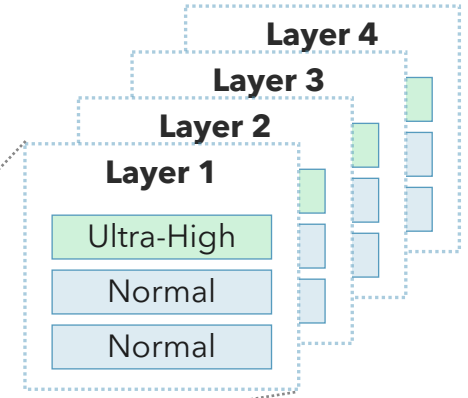


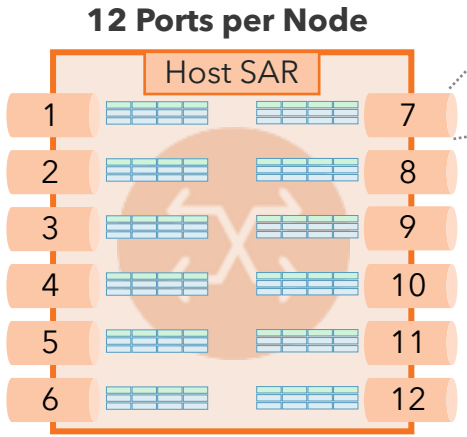
rockport.

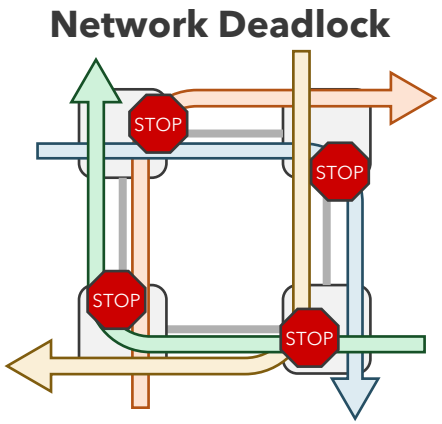
Higher Performance by Design

Deadlock-Free Routing







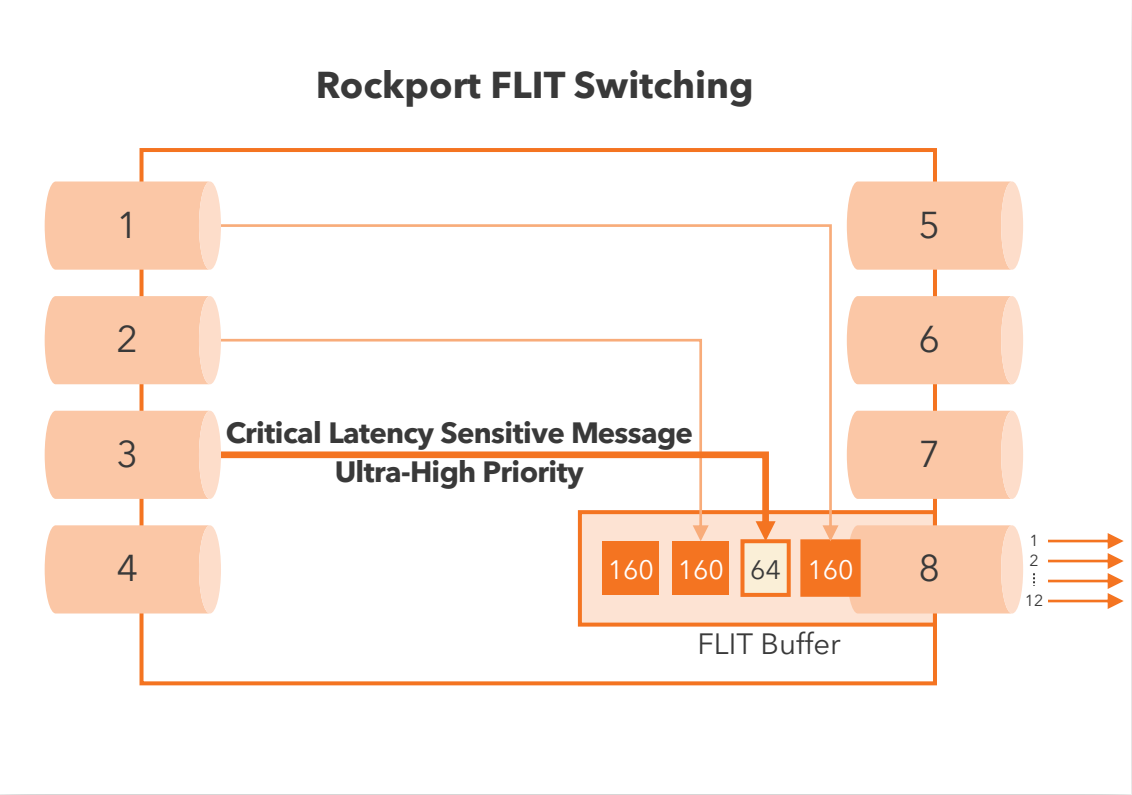


- Rockport's Deadlock-Free Routing algorithm (DFR) ensures deadlock free routing **across all topologies**
  - Ensures high path diversity
  - Complete or sparse topologies, with or without link or node failures
    - **Critical due to likelihood of cluster resource issues at Exascale**
  - Operates in a distributed fashion
  - 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

rockport.

Higher Performance by Design

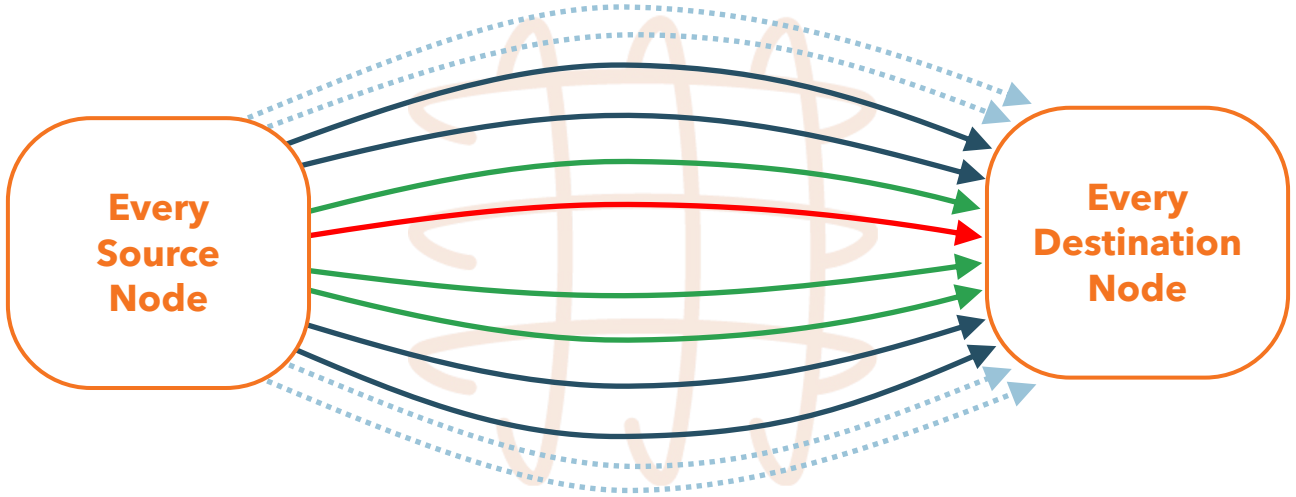
Ultra-High Priority for Critical Message Performance



- Many workloads send critical small messages that are highly sensitive to latency
  - Delays in these messages will slow down workloads
- Rockport's Ultra-High priority ensures that these critical messages are immune to network congestion
  - Ultra-High traffic is always serviced first with an average of only 25 ns additional latency per hop, even under heavy congestion
- Identification of critical messages can be done:
  - Explicitly by MPI libraries and middleware (e.g. Upcoming version of MVAPICH2)
  - Through rNOS automatic detection for all MPI libraries (Summer 2022)

rockport.

Intelligent Path Selection Based on Payload Requirements



Bandwidth sensitive traffic → **Adaptive Multi-Path** flows use all 8 paths (**dark blue** + **green** + **red**)

Latency sensitive traffic → **Ultra High Priority** messages will use the shortest path (**red**)

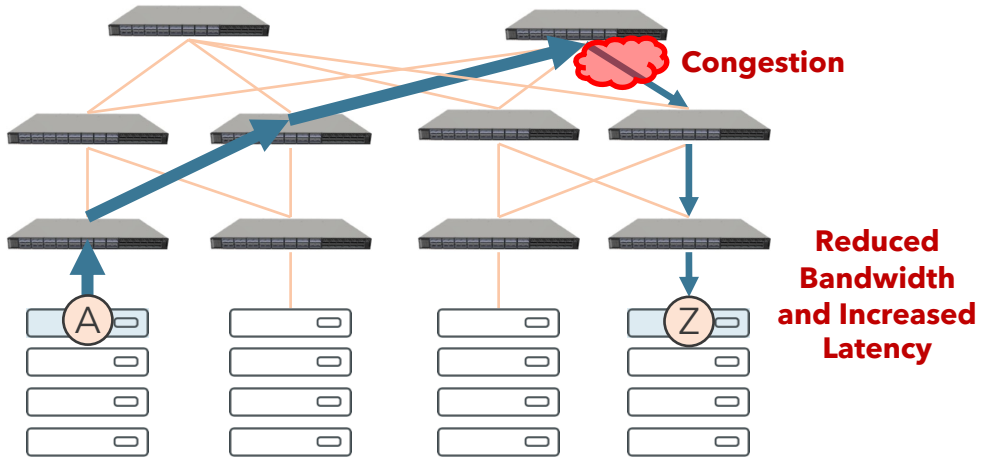
Balanced requirements → All other traffic will use the 4 shortest paths (**green** + **red**)

rockport.

Higher Performance by Design

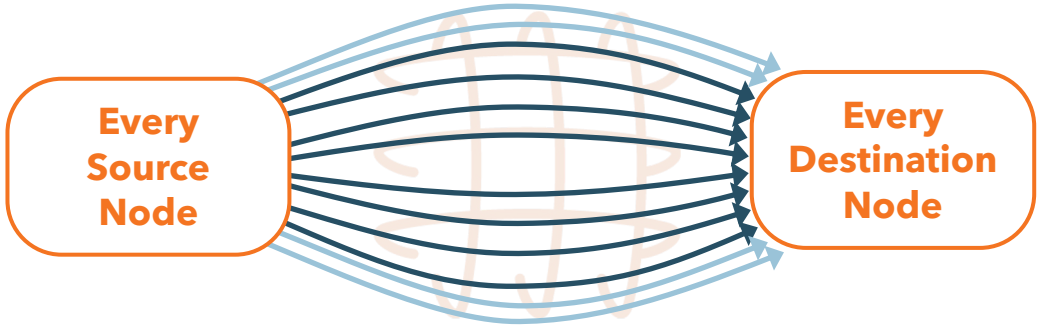
High Path Diversity

Traditional High Performance Networks



- With traditional per-destination routing, a single path is used for all traffic to the same destination
  - A single congested link can dramatically degrade performance for all traffic
  - As cluster scales increase, network layers and job counts also increase, making congestion along all paths almost a certainty

Rockport Switchless Network



Rockport's high path diversity ensures higher and more consistent performance

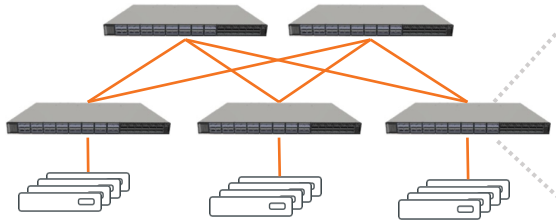
- Rockport nodes distribute packets across 8 optimal source routes to each destination to:
  - Distribute the network load across the topology to avoid creating a hotspot
  - Avoid any congested paths through adaptive routing
  - Immediately react in hardware in case of local or remote link issues

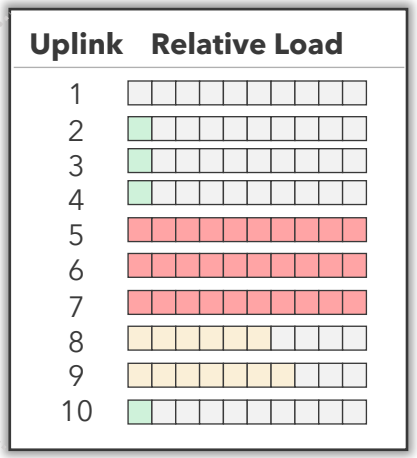
rockport.

Higher Performance by Design

Multi-Path Routing and Effective Bandwidth

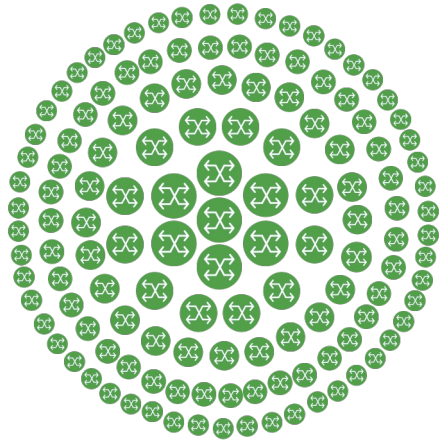
Traditional High Performance Networks





- Traditional high performance network technologies may use per-destination forwarding tables on each switch
  - Each destination node is assigned an uplink
- When jobs want to communicate to remote nodes, there is a high likelihood that route collisions occur due to statistical clumping
  - Flows are not evenly distributed across the uplinks
- This results in ineffective use of the network capacity, high latency and poor bandwidth performance

Rockport Switchless Network



Load evenly spread across network

- Rockport distributes network traffic on a per-packet and per-flow basis to ensure efficient use of network capacity
- Uses the 8 optimal physically independent paths between each source-destination pair
- Adaptively adjusts path loading based on real-time end-to-end, full path congestion information

rockport.

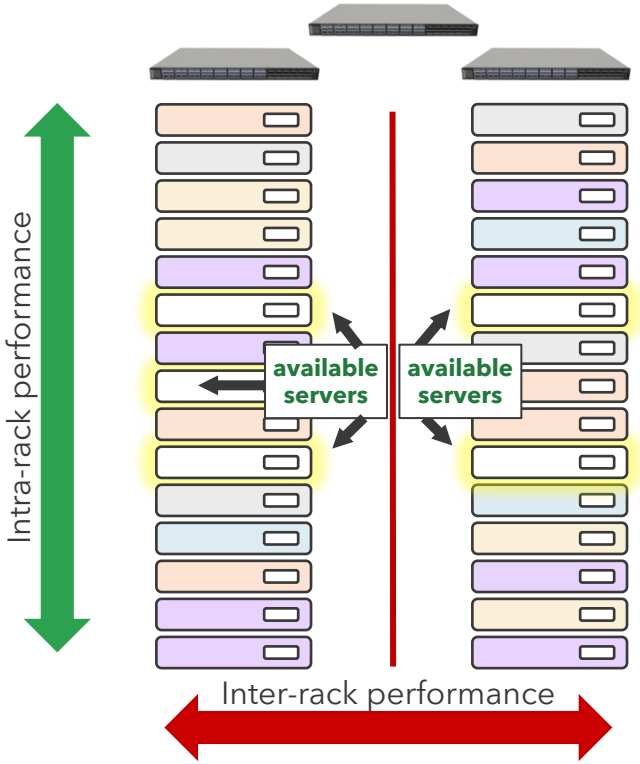
High Performance by Design

Fragmented Resource Advantages in Multi-Workload Clusters

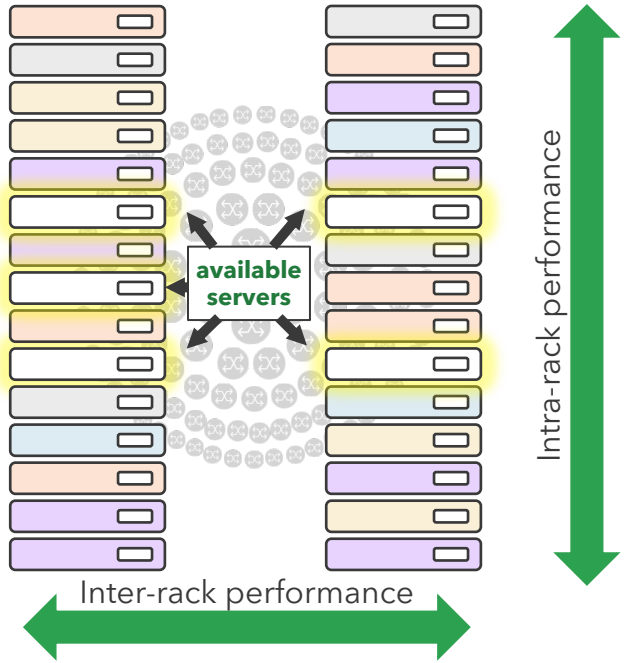
By removing the switch and distributing the network switching function into each device endpoint,
intra and inter rack performance is exceptional for multi-workload clusters

- In multi-workload clusters, resources become fragmented
- New jobs often need to be spread across multiple racks
  - e.g. 5-server workload requires servers across different racks
- Inter-rack performance becomes the critical, limiting factor for workload performance
  - This is especially true at Exascale

Traditional Switches
Poor inter-rack performance due to extra congested switch hops



Rockport Switchless Network
Excellent intra and inter rack performance due to direct connections between racks

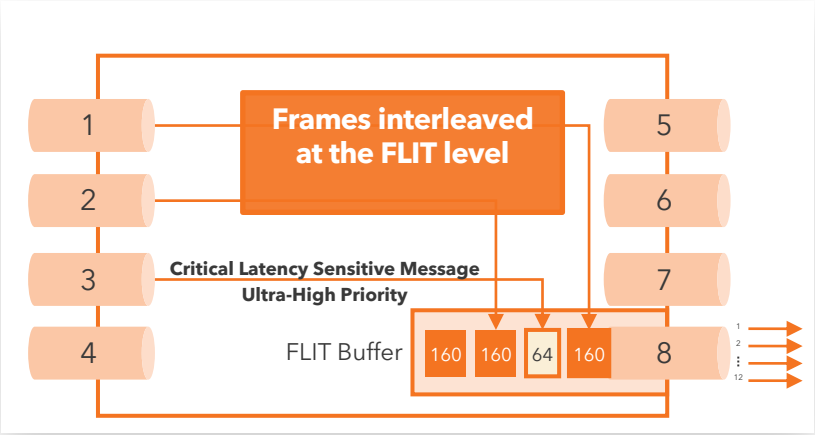


rockport.

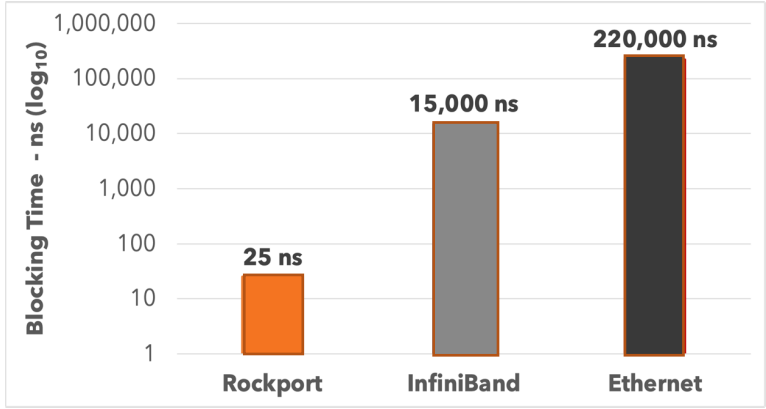
Smarter Network Fabric – Higher Performance by Design

Performance Advantages Under Load

Extreme Switching
Disaggregated FLITs ensures lowest possible latency

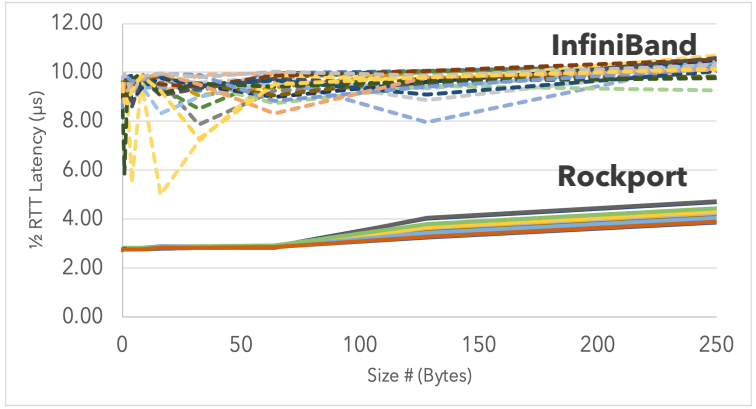


Optimal Velocity
Insignificant blocking time compared to traditional options



lower is better

Predictable Under Load
Predictable low latency performance at every scale



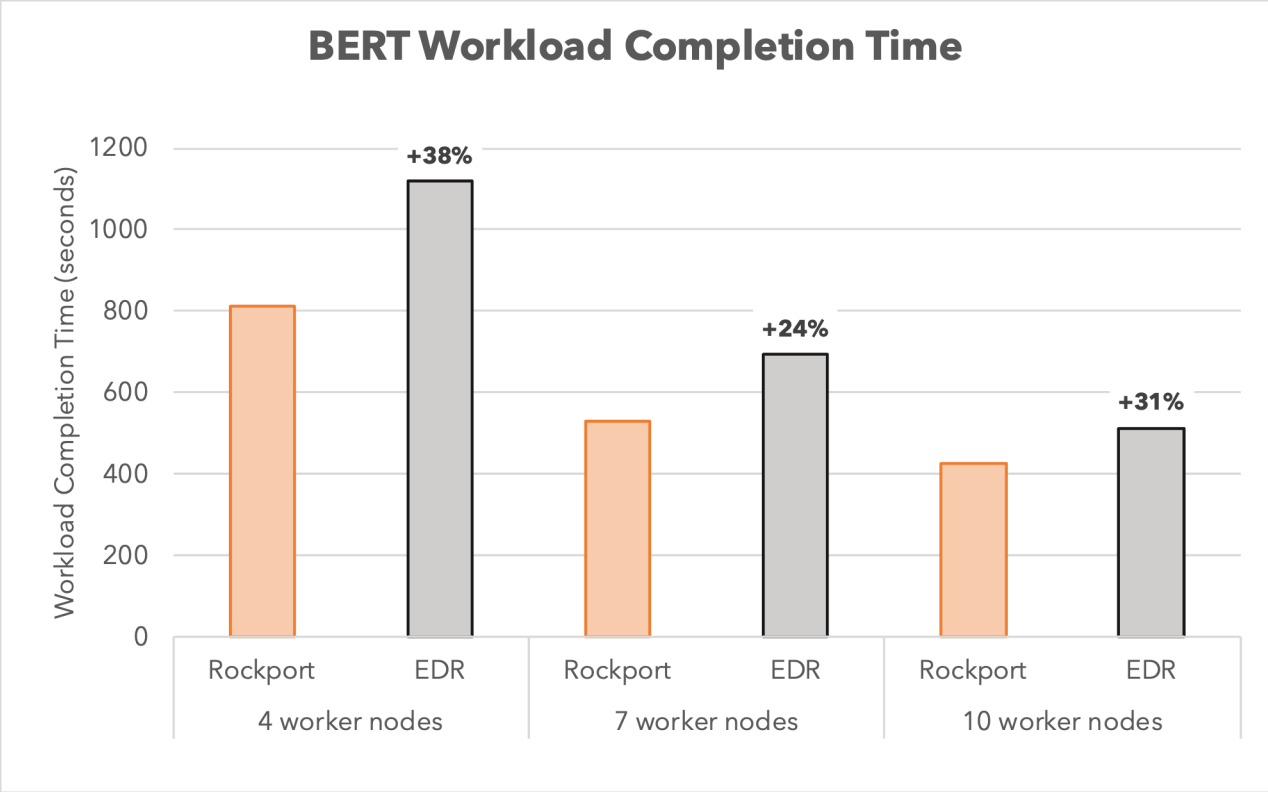
lower is better

Customer Workload Results

rockport.

Customer Workload Results

BERT AI/ML Workload Results – Lower is Better



The customer's BERT AI/ML workload completed up to 38% faster on Rockport

BERT Workload Testing

- Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. – Wikipedia
- 24% to 38% BERT workload completion time improvement by Rockport driven by path diversity advantages

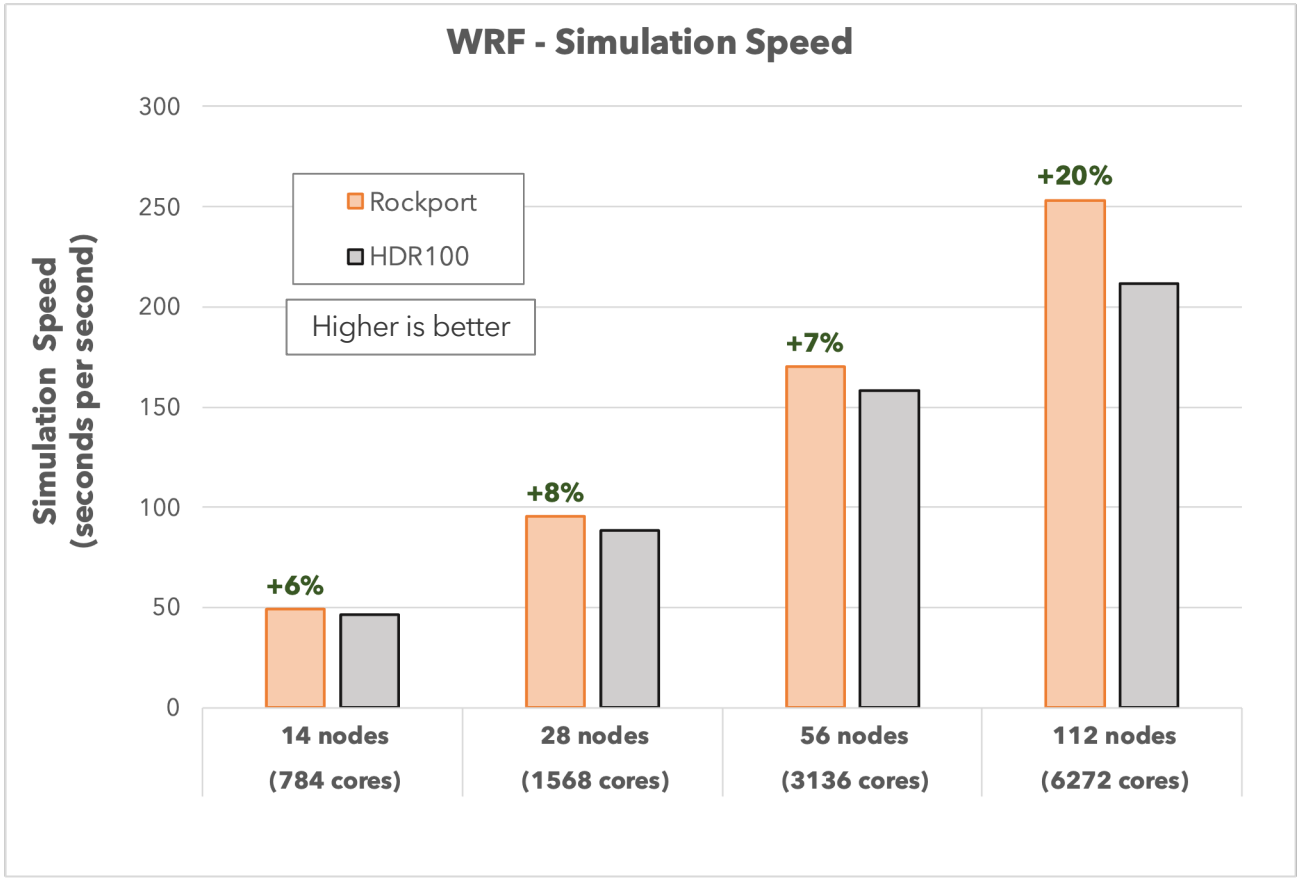
Cluster Configuration

- 10 nodes, each with
  - 1x AMD EPYC 7V12 64-Core Processor
  - 2x RTX A600 GPUs
- Nodes connected across 3 InfiniBand switches with **undersubscribed** spine

rockport.

Customer Workload Results

Climate Science – WRF



- WRF 3.9.1.1 with CONUS dataset
- 3 hour simulation time from 2005-06-04_06:00:00 to 2005-06-04_09:00:00

- Weather Research and Forecasting (WRF) (open-source numerical weather prediction modeling system) has scaling limitation in Traditional High Performance Networks
- Simulations run on Frontera at TACC during April 2022 Texascale Days event
  - (dedicated set of nodes without "noisy neighbor" traffic)
- WRF runs 20% faster on the Rockport network at 6272 cores
- **Demonstrates Rockport's performance advantages, especially at larger scales**

rockport.

"Based on the results and our first experience with Rockport's switchless architecture we were confident in our choice to improve our exascale modeling performance – all supported by the right economics."

Dr. Alastair Basden,
DiRAC/Durham University

Durham University

COSMA

Durham University Selects Rockport Switchless Network for COSMA7 Exascale Workload Modeling; Breakthrough Performance and Resource Utilization

Advances in high-performance networking is changing how data center clusters are built, measured, and expected to perform overall

News Highlights

- A leading UK HPC research facility chooses 232 nodes for upgrade to a new Rockport switchless network architecture based on some codes seeing up to 28% performance improvement over InfiniBand
- By conquering congestion – a workload killer – Durham is able to speed research results with more predictability, better resource utilization and economic efficiency
- Rockport Joins ExCALIBUR's Technology Provider Panel

Rockport Switchless
High-Performance Network

Summary

Accelerated and Predictable Workloads

A powerful alternative over Traditional High Performance IB or Ethernet Networks

Superior Resource Utilization and TCO

Effective and efficient compute, storage and networking, especially at scale

Simplified, Smarter, and Green

High Performance without the headaches, ready for the next generation

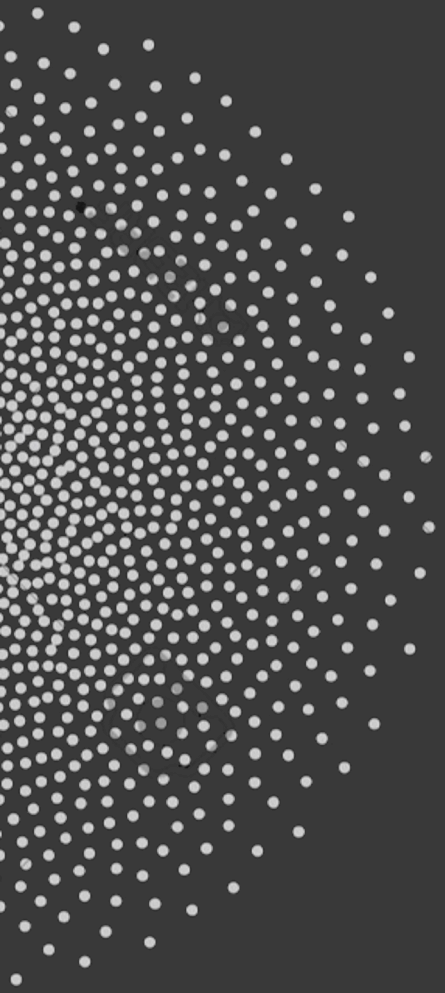
Thank You.
Questions?

Headquarters

600 – 515 Legget Drive
Kanata Research Technology Park
Floors 5,6,7

Contact

Matthew Williams
mwilliams@rockportnetworks.com



rockport.