CORNELIS® NETWORKS

Reimagining the Scale-Out Architecture for HPC and AI

Matthew Williams
Field CTO

HOT INTERCONNECTS

Cornelis in a Nutshell

Inventors of **Omni-Path**, the highest-performance inter-node interconnect, created at **Intel** and based on **QLogic** and **Cray** supercomputer technologies.

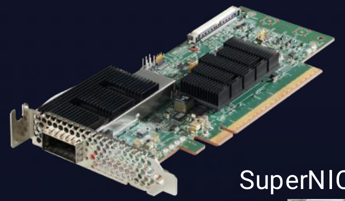
230+ people with **deep HPC expertise** and a **maniacal focus on application performance**, who understand that the **network** is the **dominant bottleneck** for HPC and AI.

Proudly behind the network solution with the **lowest latency**, **highest small message rate** and **fastest application performance.**

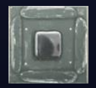
CN5000: Highest-Performance End-to-End Networking Solution

The World's First Lossless Zero-Congestion Scale-Out Network

400G SuperNICs



SuperNIC ASIC



Complete 400G Switch Portfolio

48-Port
Switch

SWITCH ASIC

Up to 576-Port
Director Switch

Optimized Open-Source Host and Management Software

OPENFABRICS ALLIANCE

The Most Advanced Architecture

2.5x Higher Message Rates

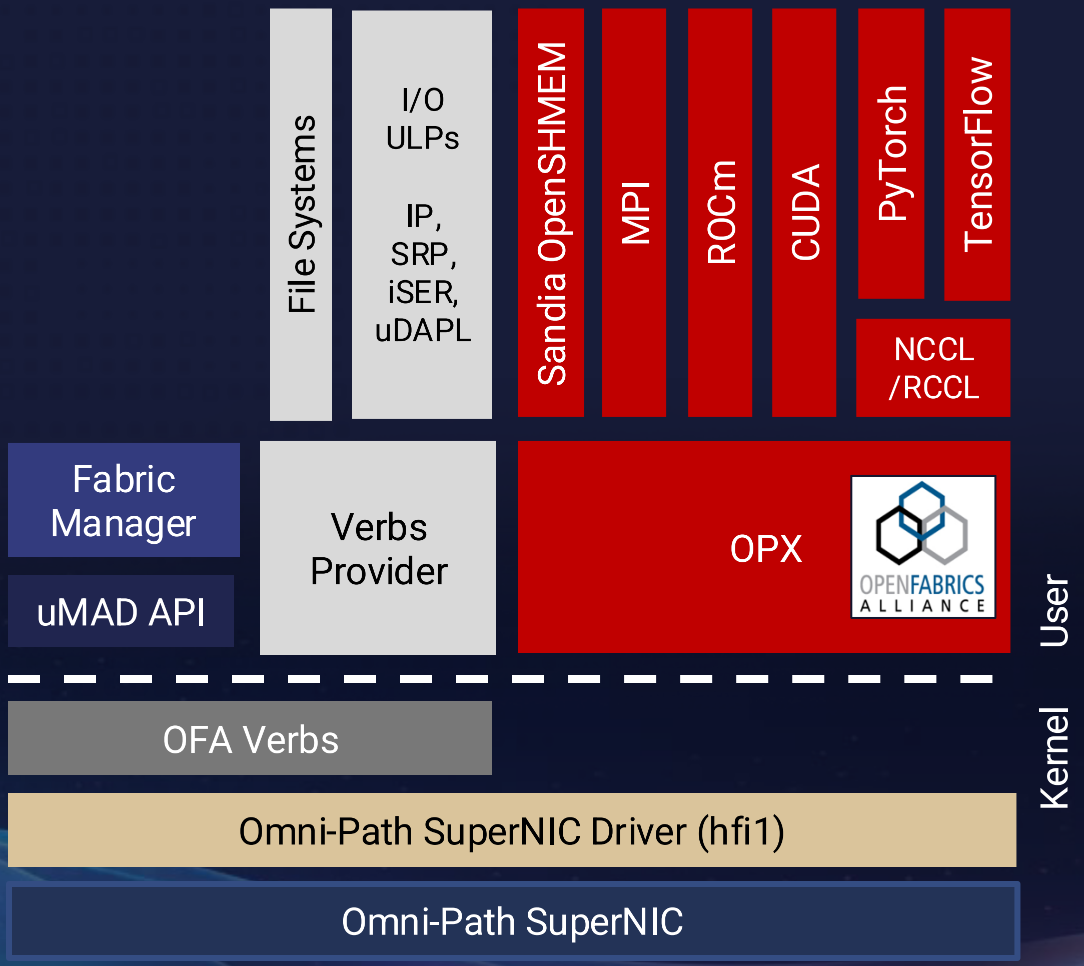
34% Lower Latency

The Performance Leader

Up to 45% Higher HPC Application Performance

Performance claims based on CN5000 relative to NDR 400G

Omni-Path OFI Libfabrics Software Stack



Optimized Performance

Tight Software-Hardware Integration and Co-Design

Seamless Integration

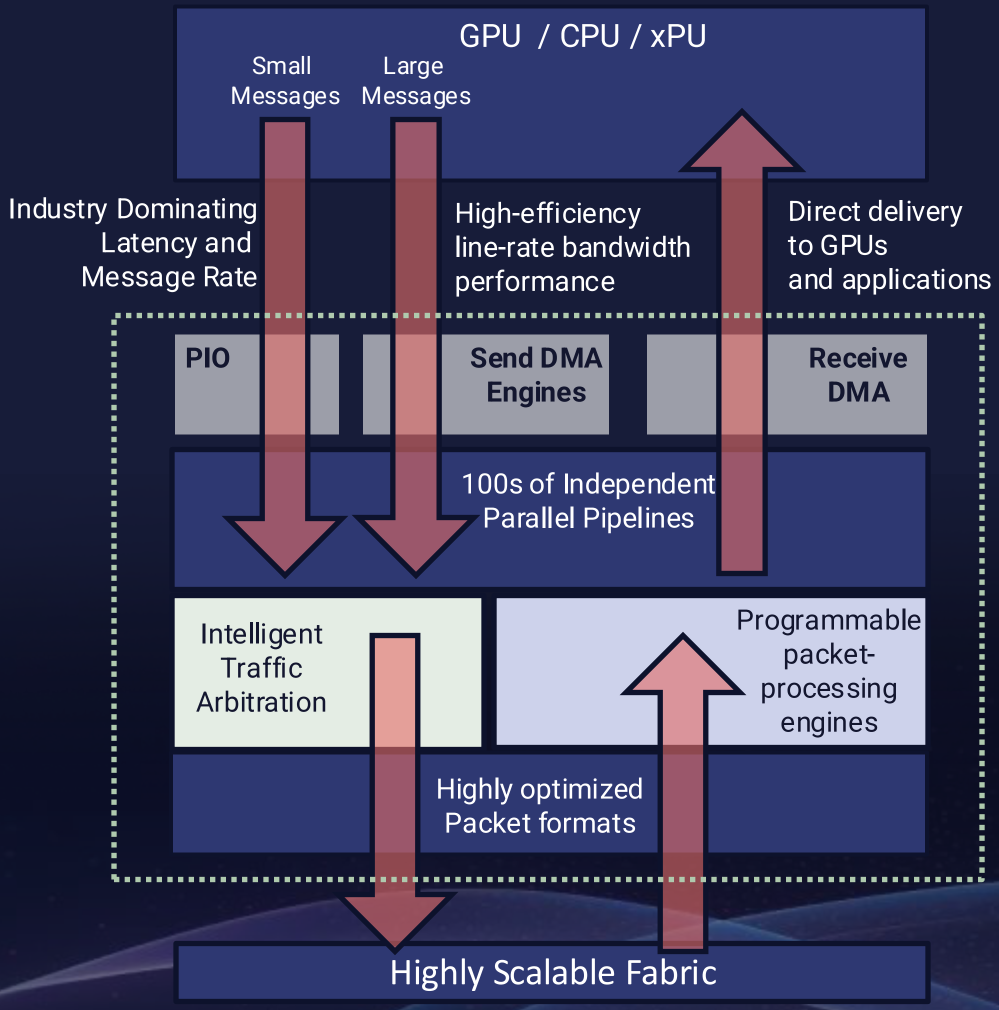
Adoption of Open Standards and Full Stack Support

Trusted Deployment

Open-Source Libfabric Provider and Upstreamed Kernel Driver

Omni-Path Performance Architecture

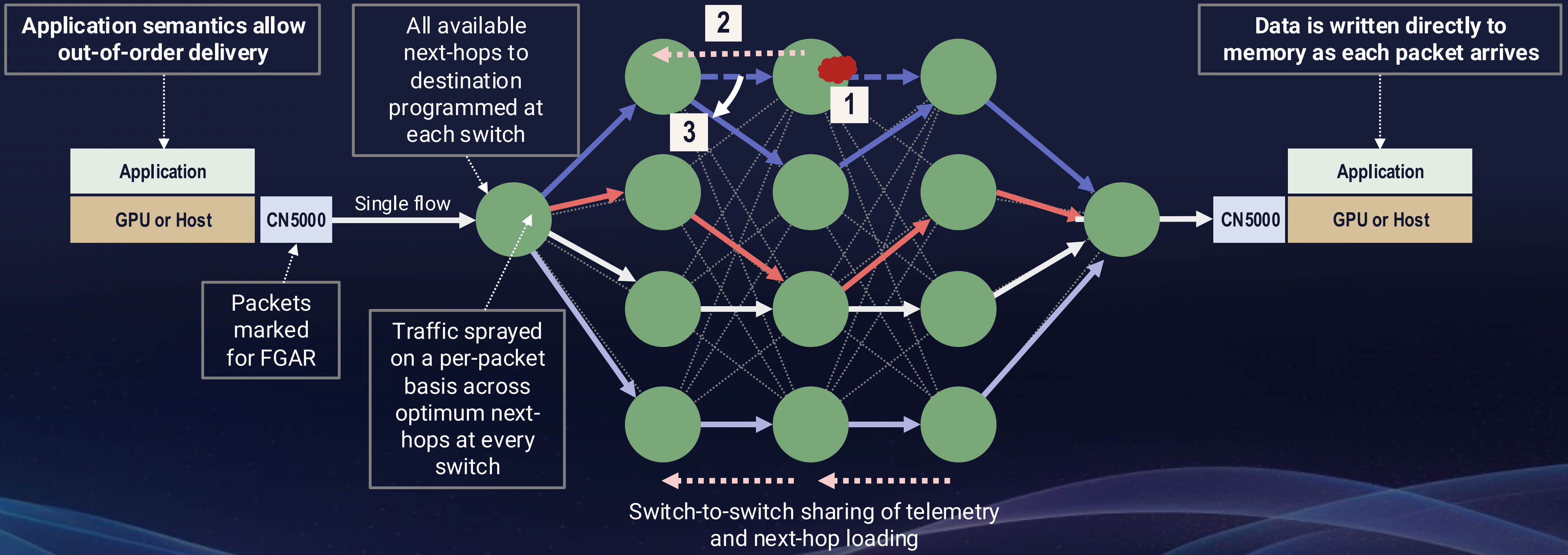
- Application performance is the Cornelis North Star
- Our open-source OPX libfabric provider is the primary software framework
- Each process (e.g MPI rank) is assigned 1 or more of the 100s of independent parallel pipelines (contexts)
- Small messages are sent directly from each process to the SuperNIC
  - Programmed I/O (PIO)
  - **Sub-microsecond 1-hop MPI latency**
  - **Up to 2.5x NDR message rate**
- Large messages and data transfers leverage the 16 Send DMA (SDMA) engines
- Received data is placed directly into host memory
  - Application buffers for rendezvous
  - Ring buffer for eager



Industry-leading MPI message rate, latency and bandwidth ramp

Fine-Grained Adaptive Routing (FGAR)

1. Heavy load on switch ports
2. Congestion information shared with neighbor switches
3. New set of optimum next-hops selected based on local and remote congestion



Consistent bandwidth performance for AI and storage applications

Hierarchical LIDs (HLID)

- Local Identifiers (LIDs) are the addresses used within an Omni-Path network
- The CN5000 can use 24-bit Hierarchical LIDs (HLIDs) to support a wide range of network topologies across a wide range of network scales
- Depending on the topology of the network, the HLID is broken into multiple sub-fields
  - Flexible definitions and sub-field sizes through the Fabric Manager
- These sub-fields can be thought of as coordinates that identify SuperNIC locations within the topology
- The Cornelis Fabric Manager calculates routes that optimize traversal between sets of coordinates
  - E.g. To move to VRG 7 from node (6,1,2), the next hop from a switch is programmed to be from a set of 8 egress ports (p_1, p_2, …, p_8)
  - Highly efficient route tables
    -> 500K nodes in a flat layer 2 network

Example HLID Sub-Fields

24-bit HLID

Topology	Sub-fields			
Tree	7b Pod	1b Tier	5b Switch in Pod	6b Endpoint Within Switch
Dragonfly & Megafly	10b Virtual Router Group (VRG)	6b Switch Within VRG	6b Endpoint Within Switch	
3D Hyper-X	4b S3 Coordinate	4b S2 Coordinate	4b S1 Coordinate	5b Endpoint Within Switch

Optimized topologies for each use case

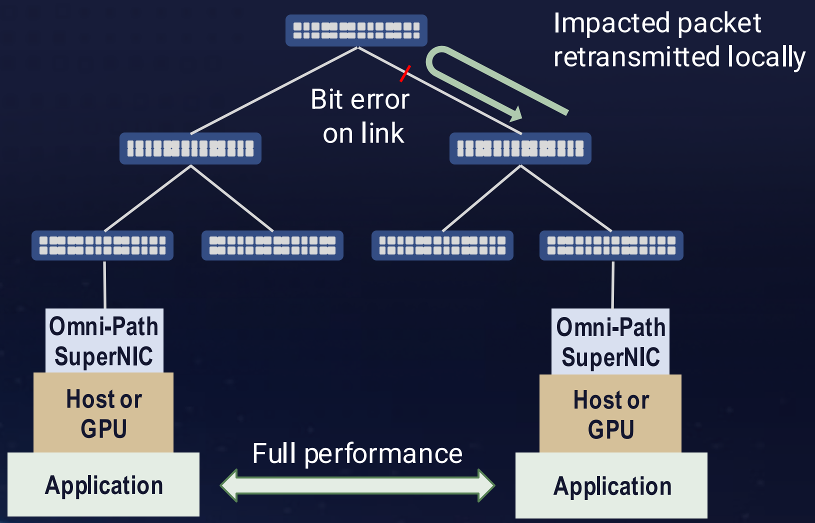
Link Level Retry

10^{-12} BER @ 400 Gbps
= 1 bit error every 2.5 seconds on each link

CORNELIS NETWORKS

Omni-Path

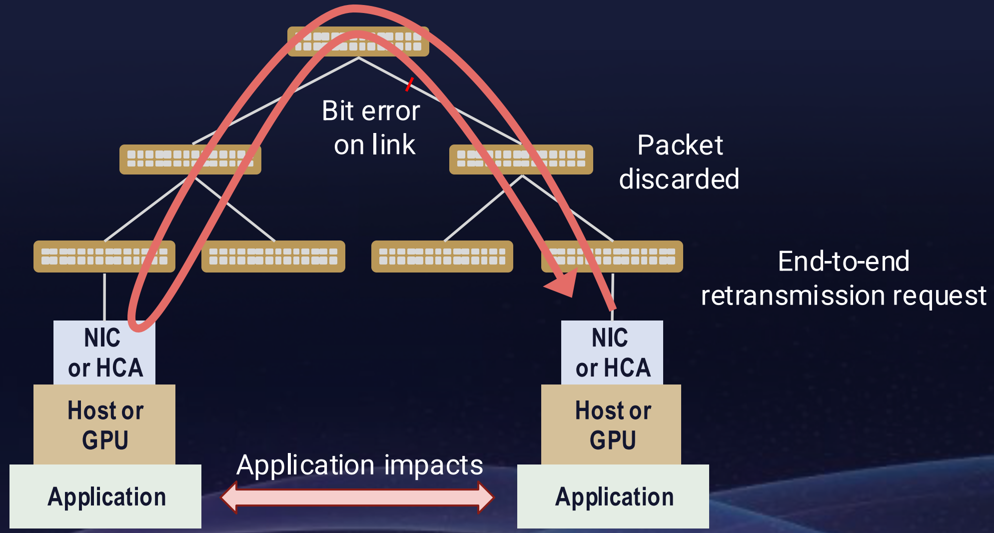
- Switches keep copies of transmitted packets until acknowledged by next hop
- Bit errors handled by rapid retransmission (~2x link time-of-flight)
- Application is unaware of any issue and continues to run at full rate
- **No packets dropped due to bit errors**



No packets dropped due to Bit Error

Other Networks

- Far end must request a retransmission
  - Full RTT spike in latency while waiting for retransmission
  - Go-back-N approach results in inefficient network loading
- A retransmission request may be interpreted as loss due to congestion
  - Network stack slows down traffic to mitigate false congestion
- **Application impacted by latency spike and stack slow-down**



Bit Error results in dropped packets and end-to-end retries

Maximum application performance in real-world environments

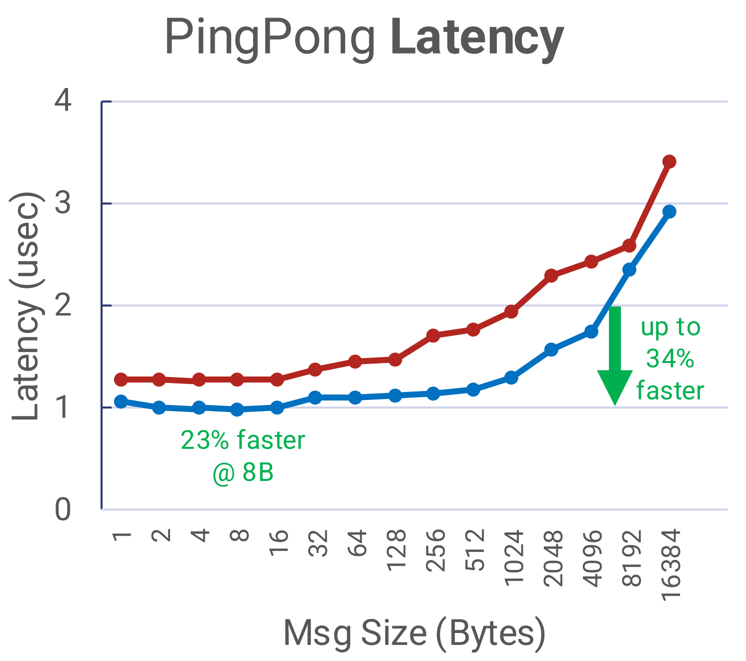
Industry Leading Network Performance

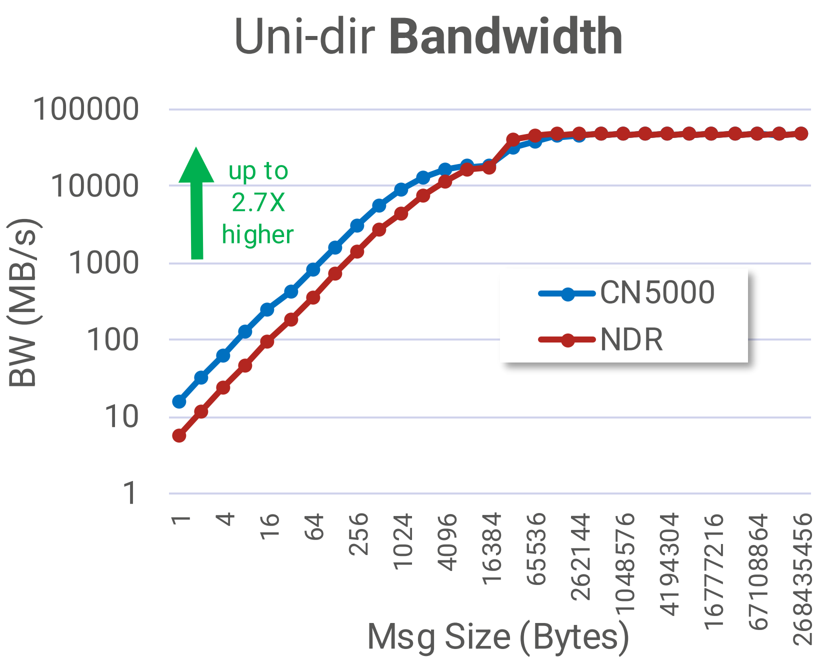
Expect further performance upside with future performance tuning and scaling

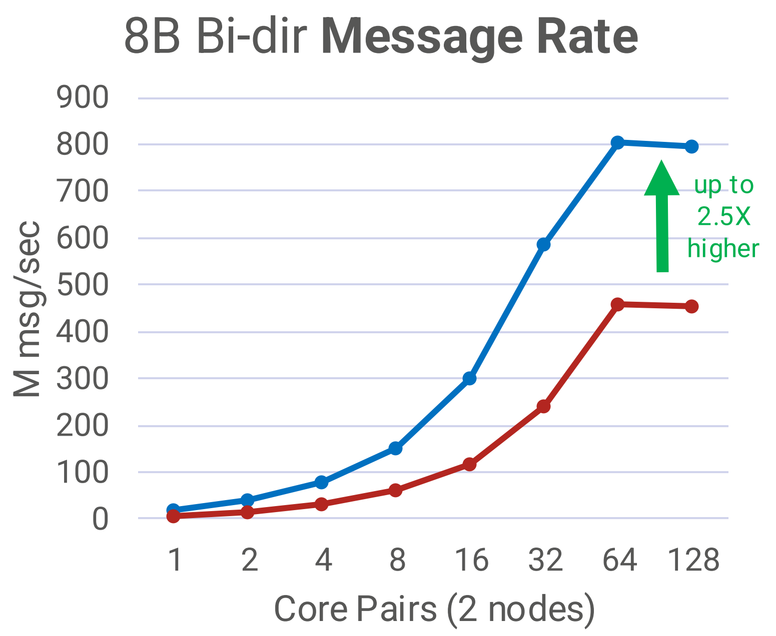
CN5000 vs NDR 400G – AMD EPYC 9755 (Turin)

Leadership Latency, Small-Message Bandwidth, and Message Rate

Delivering Enhanced Application Scaling & Addressing The Broad Range of Application Sensitivity

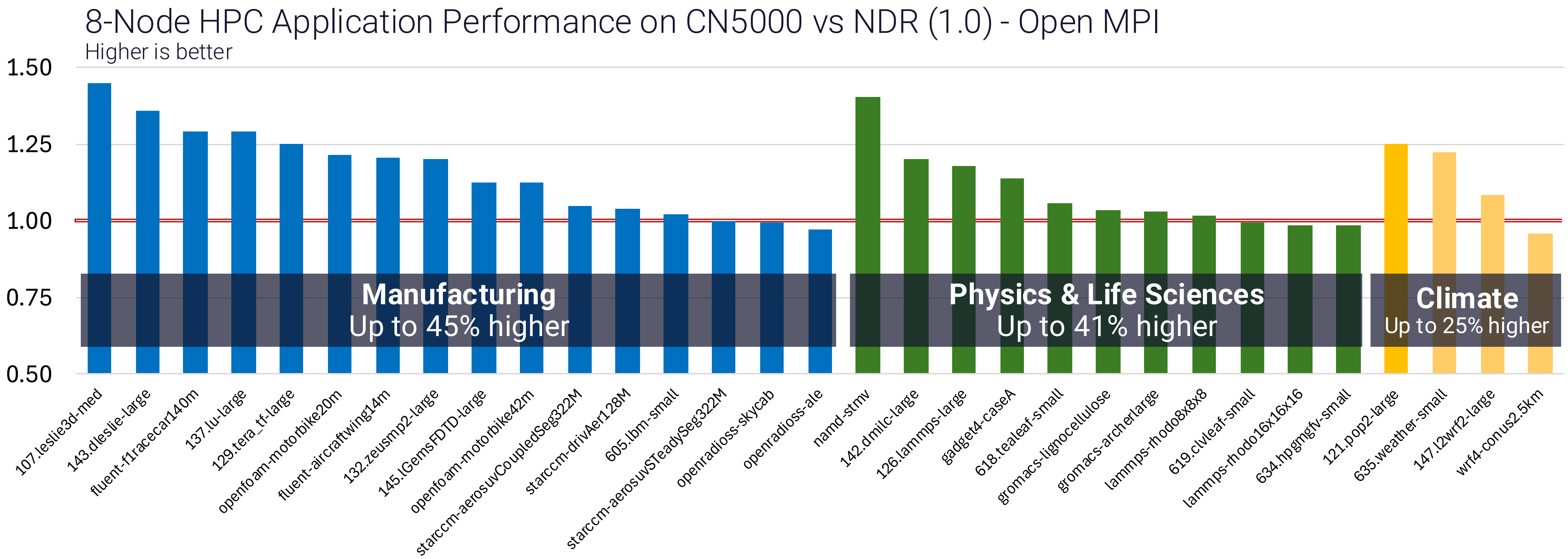






Tests performed on 2 socket AMD Eng Sample: 100-000001535-03. Turbo enabled with acpi-cpufreq driver. Rocky Linux 9.5 (Blue Onyx). 5.14.0-503.33.1.el9_5.x86_64 kernel. 24x32GB, 768 GB total, Memory Speed: 5600 MT/s. Cornelis Omni-Path Express Suite (OPXS) 12.0.0.0.17. SuperNIC driver parameters: num_user_contexts=0,128 num_vls=4 num_sdma=8 sdma_threshold=16 pad_sdma_desc=16 sdma_align=2. Intel MPI 2021.15, Intel(R) MPI Benchmarks 2021.9. NVIDIA NDR InfiniBand: Mellanox Technologies MT2910 Family [ConnectX-7]. MQM9700-NS2F Quantum 2 switch. 2M passive copper cables. UCX as packaged in hpcx-v2.23.

HPC Application Performance - AMD Turin



Achieve Up To 45% Higher HPC Performance; Differentiated Performance Across Segments

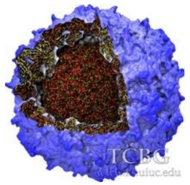
Expect further performance upside with future performance tuning and scaling

Tests performed on 2 socket AMD Eng Sample: 100-000001535-03. Turbo enabled with acpi-cpufreq driver. Rocky Linux 9.5 (Blue Onyx). 5.14.0-503.33.1.el9_5.x86_64 kernel. 24x32GB, 768 GB total, Memory Speed: 5600 MT/s. Cornelis Omni-Path Express Suite (OPXS) 12.0.0.0.17. SuperNIC driver parameters: num_user_contexts=0,128 num_vls=4 num_sdma=8 sdma_threshold=16 pad_sdma_desc=16 sdma_align=2. Open MPI 5.0.6. NVIDIA NDR InfiniBand: Mellanox Technologies MT2910 Family [ConnectX-7]. MQM9700-NS2F Quantum 2 switch. 2M passive copper cables. UCX and Open MPI 4.1.7 as packaged in hpcx-v2.23. Applcation specific detail available upon request.

CORNELIS NETWORKS

NAMD-stmv Scaling Performance – AMD Turin

Molecular Dynamics application



Satellite Tobacco Mosaic Virus Benchmark

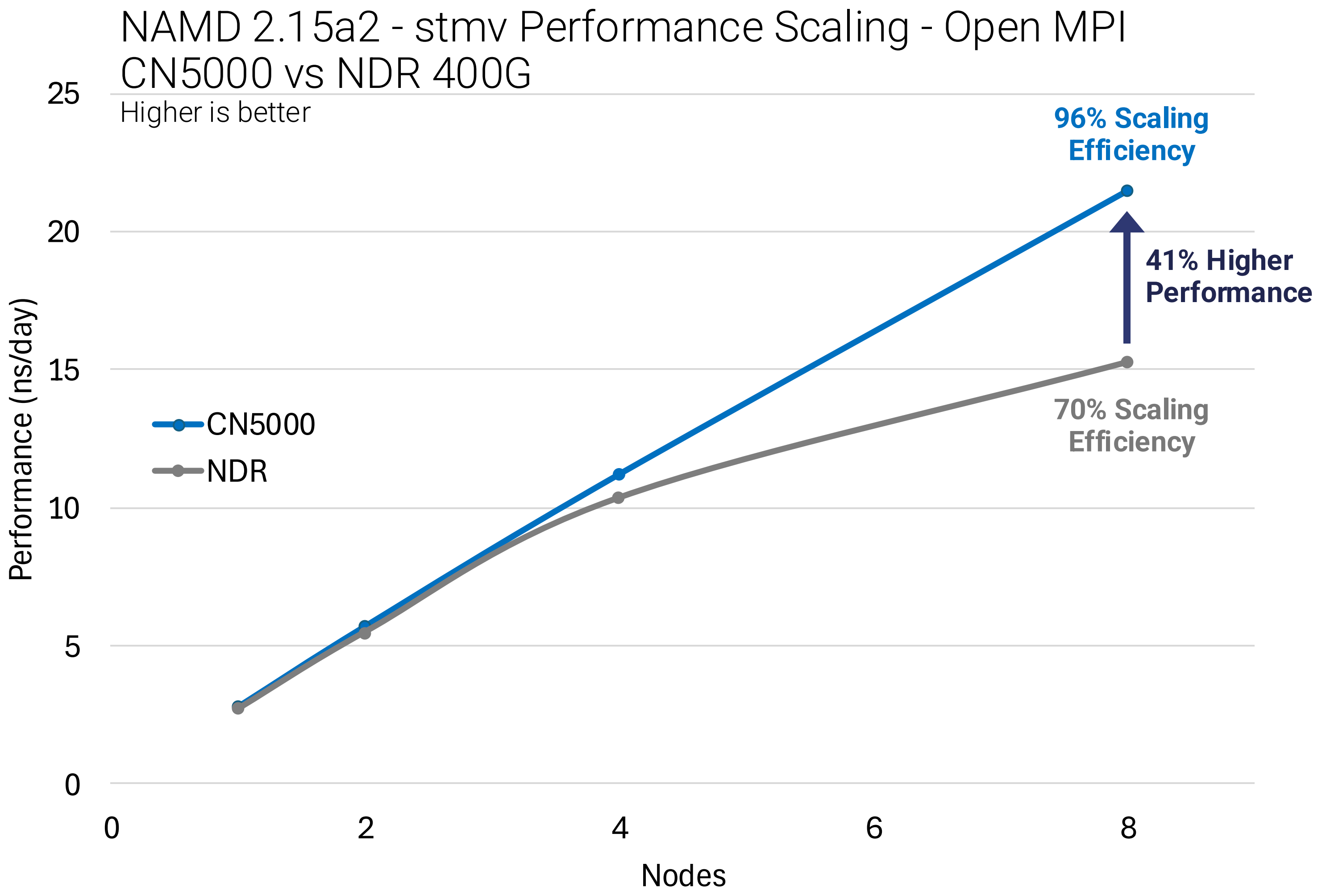
Delivering Performance At Scale

41% higher performance at 8N

Delivering More Efficient Scaling

Achieving 96% scaling efficiency

CN5000 delivers predictable scaling, growing workload performance as infrastructure scales



NAMD 2.15a2. Tests performed on 2 socket AMD Eng Sample: 100-000001535-03. Turbo enabled with acpi-cpufreq driver. Rocky Linux 9.5 (Blue Onyx). 5.14.0-503.33.1.el9_5.x86_64 kernel. 24x32GB, 768 GB total, Memory Speed: 5600 MT/s. Cornelis Omni-Path Express Suite (OPXS) 12.0.0.0.17. HFI driver parameters: num_user_contexts=0,128 num_vls=4 num_sdma=8 sdma_threshold=16 pad_sdma_desc=16 sdma_align=2. Open MPI 5.0.6. NVIDIA NDR InfiniBand: Mellanox Technologies MT2910 Family [ConnectX-7]. MQM9700-NS2F Quantum 2 switch. 2M passive copper cables. UCX and Open MPI 4.1.7 as packaged in hpcx-v2.23. Applcation specific detail available upon request.

Summary

Omni-Path Architecture delivers proven scalability and performance

Leading application performance at all scales for HPC & AI workloads

Maximizes efficiency of GPU and CPU based clusters

Thank You