CORNELIS® NETWORKS

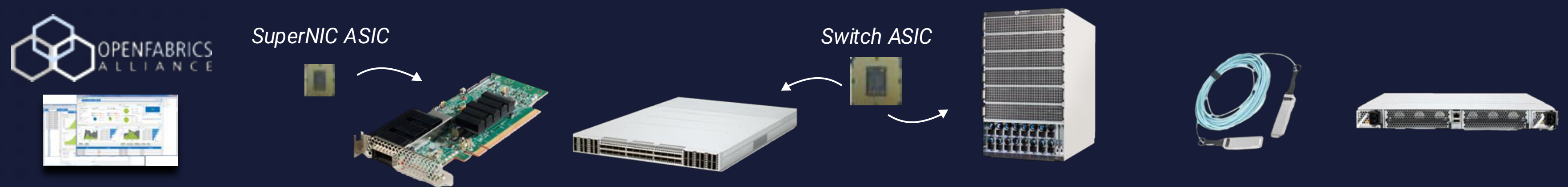
# Cornelis Networking Solution Deep Dive

Matthew Williams, Field CTO

April 2025

# Cornelis Networks: End-to-End High-Performance Network Solutions

## Catalyzing the next wave of AI and HPC innovation through a portfolio of differentiated HW and SW IP



Host and Management Software | SuperNICs | Edge Switches | Director Class Switches | Copper and Optical Cables | Gateways

### AI/HPC-Focused Host Interface

- Purpose-built accelerated SuperNICs supporting highly-optimized OFI host stack implementation

### High-Bandwidth, Low-Diameter Topologies

- High-radix, high-bandwidth switch building blocks supporting tree and advanced topologies

### Low Latency at Any Utilization / Scale

- Minimum fall-through latency coupled with advanced traffic management

### Advanced Adaptive Routing and QoS

- Multi-tenant security and QoS via dynamic virtual fabrics and fine-grained adaptive routing

### State-of-the-Art Congestion Control

- Comprehensive network telemetry synthesized to pace senders and modulate path selection

### Leading Enterprise and Government Agency Customers

ASTON MARTIN · Audi · BMW · IARPA · NNSA National Nuclear Security Administration · Department of Defense · U.S. DEPARTMENT OF ENERGY · VOLVO · VW

# Cornelis Technology Partners

## Multi-Vendor Enablement across all functions

| | |
|---|---|
| Compute | intel<br>NVIDIA<br>AMD<br>Custom ASICs |
| OEMs | SUPERMICRO<br>DELL Technologies<br>Lenovo<br>PENGUIN SOLUTIONS<br>Hewlett Packard Enterprise |
| Storage | VDURA<br>BeeGFS<br>ddn<br>IBM<br>Lenovo |
| Software | NVIDIA CUDA<br>AMD ROCm<br>OPENFABRICS ALLIANCE |
| Interconnect | Ultra Ethernet Consortium<br>UALink |

## Strong Software Ecosystem Interoperability

CORNELIS NETWORKS + ISV/OSV Partner Technologies

ANSYS

Altair

CANONICAL

CONVERGENT SCIENCE

Red Hat

SIEMENS

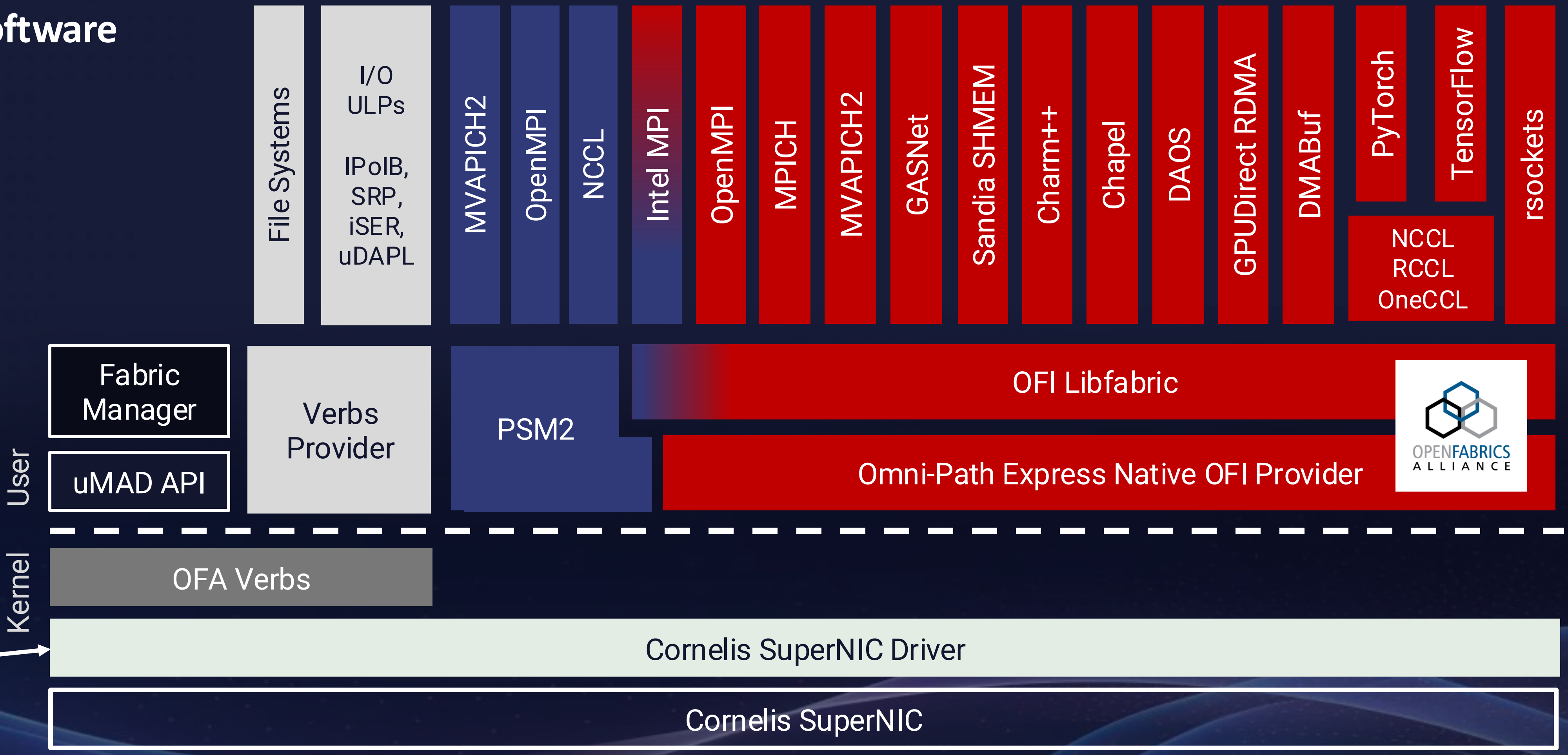
OpenFOAM

CiQ

openSUSE

# Omni-Path Express Host Software Stack

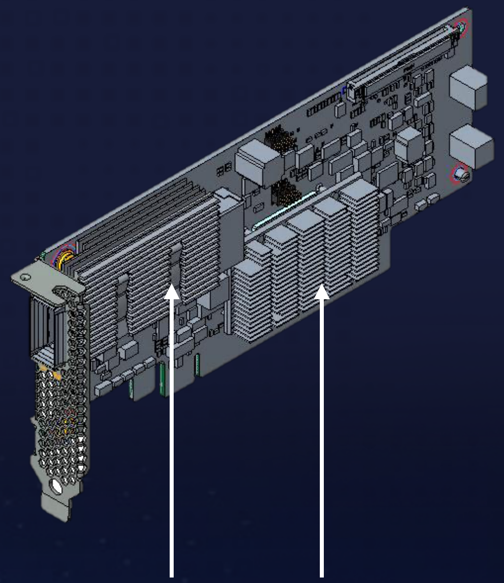
**Fully open-sourced messaging software stack**



Upstreamed to mainline kernel

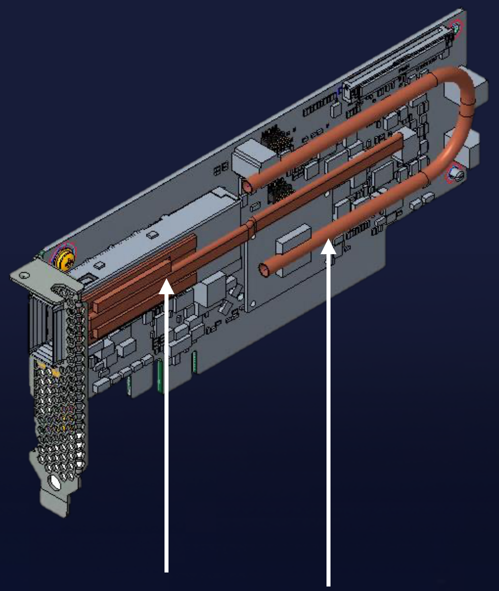
# CN5000 SuperNIC

### Air-Cooled SuperNIC



Heatsinks

### Liquid-Cooled SuperNIC



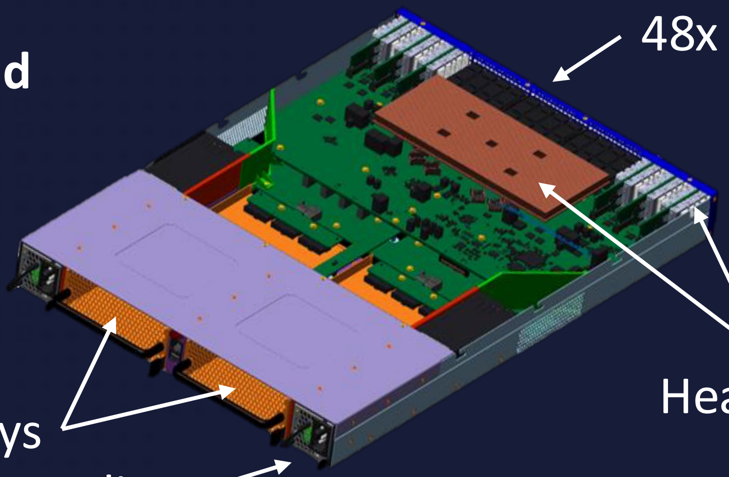
Heat pipes to a server liquid cooling infrastructure

## CN5000 SuperNIC

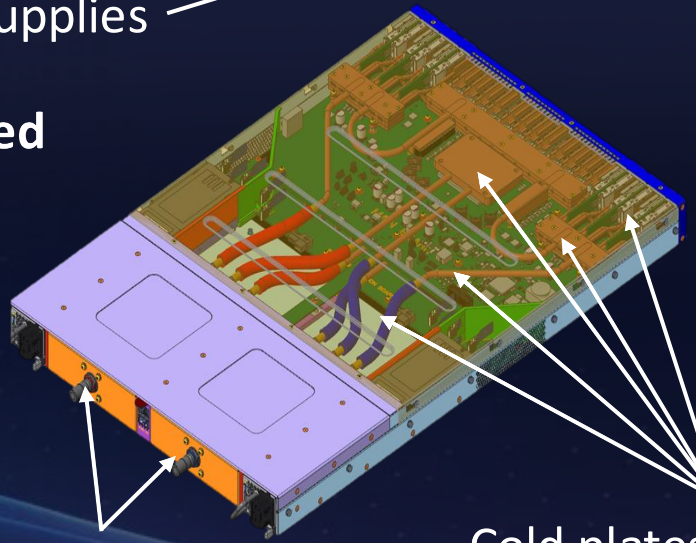
- Host interface – PCIe Gen5 x16
- Fabric ports:
  - Single port – 4x100G via QSFP
  - Dual port – 2x 4x100G via QSFP-DD
- Low profile PCIe
- Power consumption (w/o optics): 17-20 W
- Cooling options:
  - Air cooling (heatsinks on ASIC and I/O)
  - Liquid cooling – heat pipes from ASIC and I/O connector to a server cold plate
  - Liquid cooling heat pipe from ASIC to server cold plate – copper cables only

# CN5000 Edge Switch

**Air-Cooled**



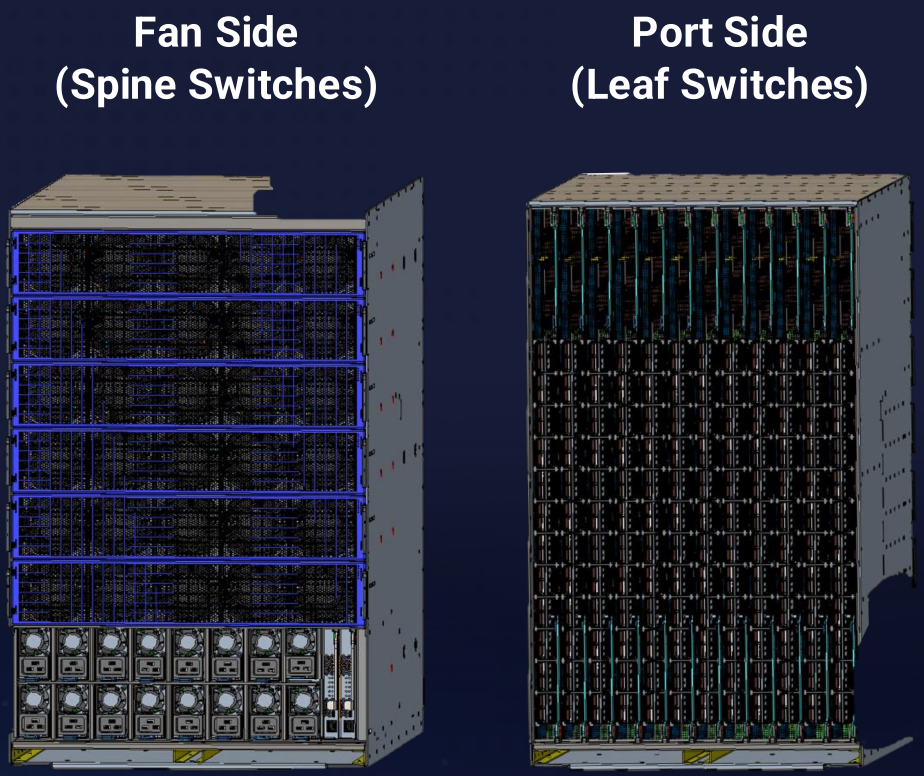
**Liquid-Cooled**



## 48x 400G port Edge Switch

- 1U, 19” rack mount chassis
- 48x 400G QSFP ports
- Redundant hot-plug power supplies & fan trays
- Integrated OpenBMC-based management
- Power with 48x 7.5W AOC:
  - Air cooled – 1100W
  - Liquid cooled – 850W
- Cooling options:
  - Air cooling:  Fan-to-Port and Port-to-Fan airflow
  - Liquid cooling:  cold plates on ASIC, I/O, and voltage regulators
  - PS cooling:  240/277 VAC PS (pluggable), air cooled

# CN5000 Director Class Switch



Cost, power, and rack space optimized for spine-and-leaf topologies

**576x 400G port Director Class Switch**

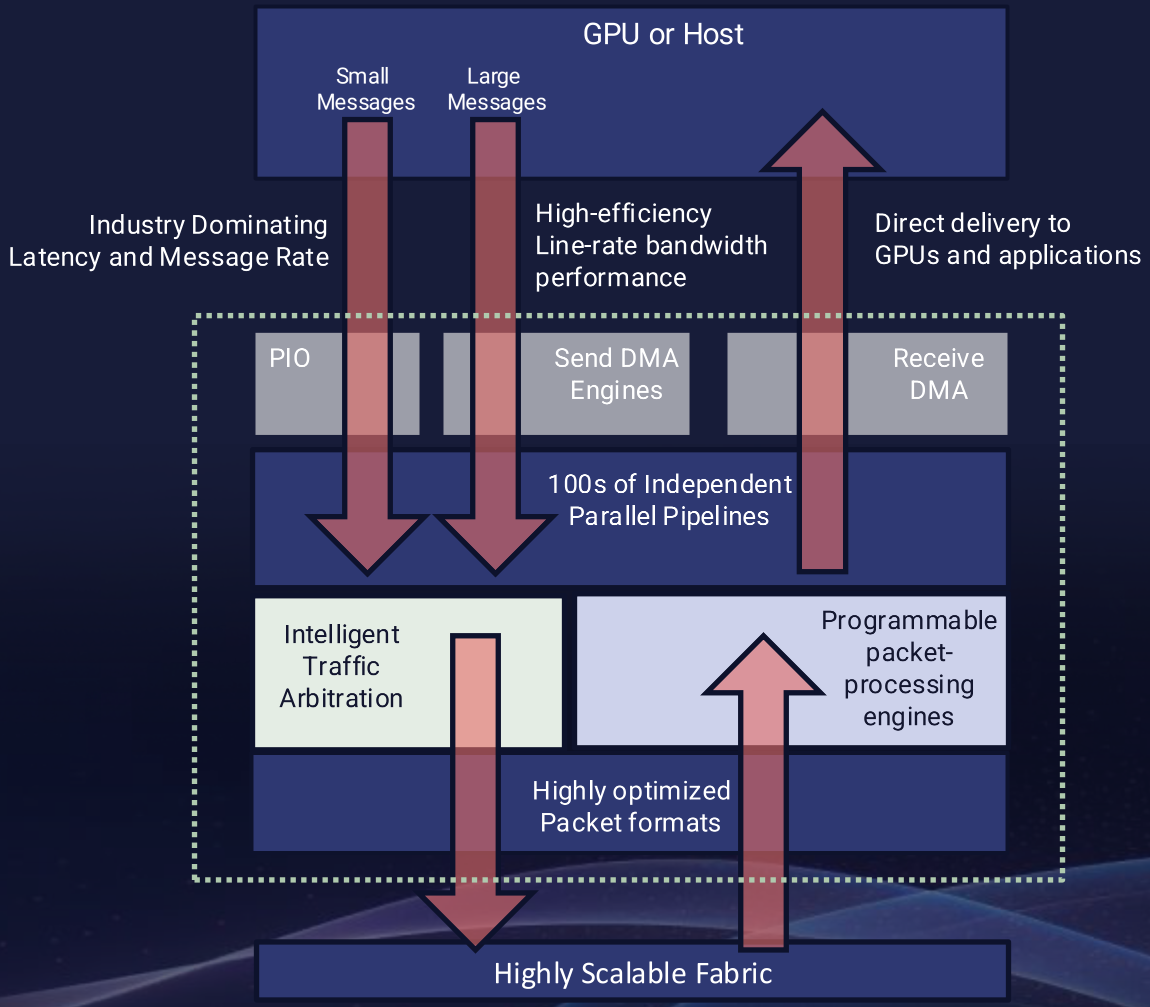
- 17U, 19” rack mount chassis
- Orthogonal interconnect between Leaf and Spine modules
  - No backplane
  - **Eliminates spine-leaf optical cables**
- 12x 48-port 400G port Leaf Switch modules
- 6x Spine Switch modules
- 2x Management modules 1+1 redundant
- All modules hot pluggable: Leaf, Spine, Mgmt. Module, Power Supply, Fan Tray
- Redundant modules: Mgmt., Power Supply, Fan Tray
- Power – 20 kW (including optics)
- Cooling options:
  - Air cooling – Port-to-Fan airflow
  - Liquid cooling: cold plates on ASIC, I/O, voltage regulators
  - PS cooling - 240/277 VAC PS (pluggable) – air cooled

# CN5000 Architecture

CORNELIS™ NETWORKS

# CN5000 SuperNIC Architecture

- Application performance is the Cornelis North Star
- Libfabric is primary software framework
- Each process (e.g MPI rank) is assigned 1 or more 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
  - 40% to 100% higher message rate than NDR
- Large messages and data transfers leverage the Send DMA (SDMA) engines
  - Highest performance choice for messages ≥ 32 kB
- Received data is placed directly into host memory
  - Application buffers for rendezvous
  - Ring buffer for eager
- Uses opaque addressing vs virtual memory addresses
  - No need to exchange memory address regions before initiating transfers

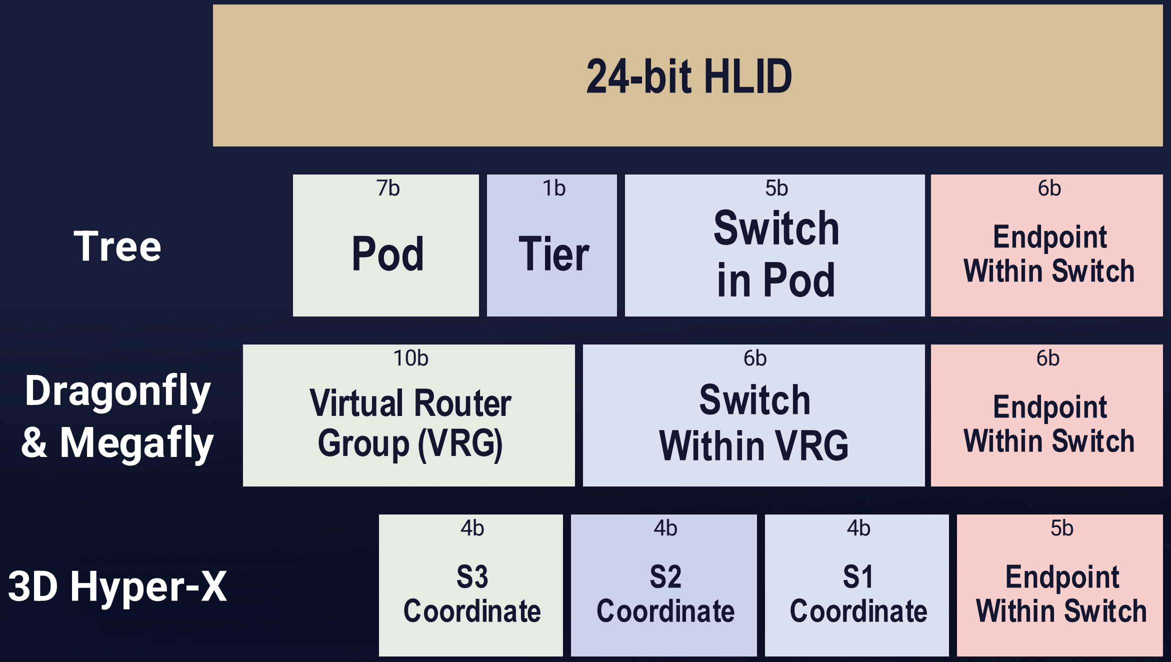


GPU or Host
Small Messages
Large Messages
Industry Dominating Latency and Message Rate
High-efficiency Line-rate bandwidth performance
Direct delivery to GPUs and applications
PIO
Send DMA Engines
Receive DMA
100s of Independent Parallel Pipelines
Intelligent Traffic Arbitration
Programmable packet-processing engines
Highly optimized Packet formats
Highly Scalable Fabric

# 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 scales and network topologies
- 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 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 -> 250K nodes

## Example HLID Sub-Fields



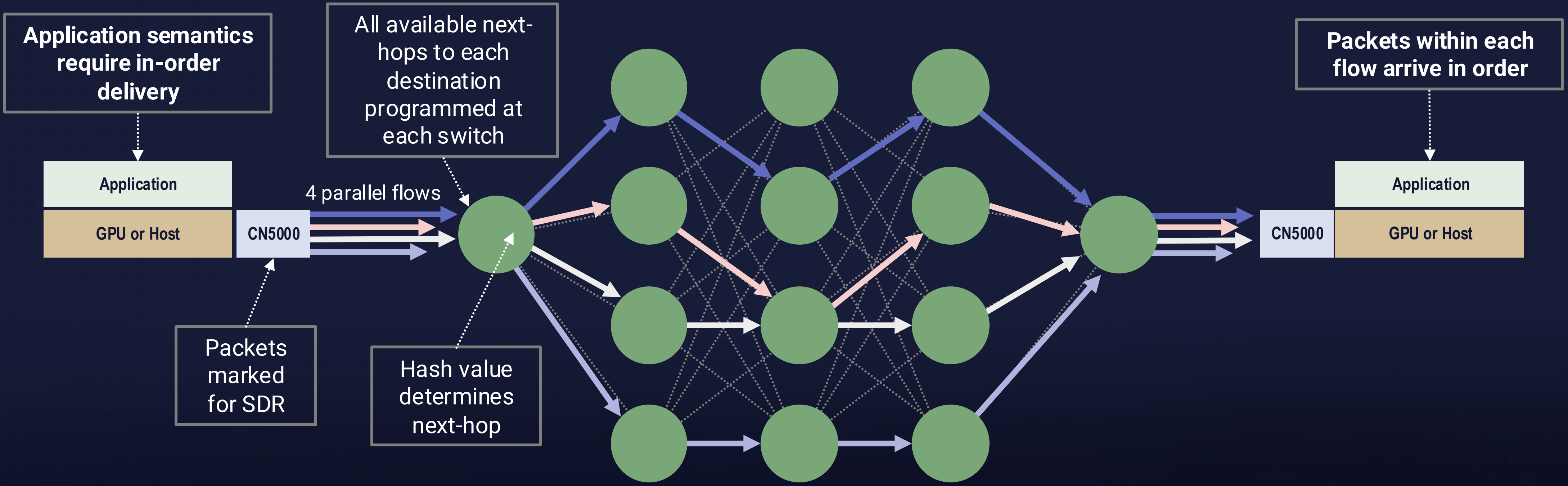
24-bit HLID

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

# Static Dispersive Routing (SDR)



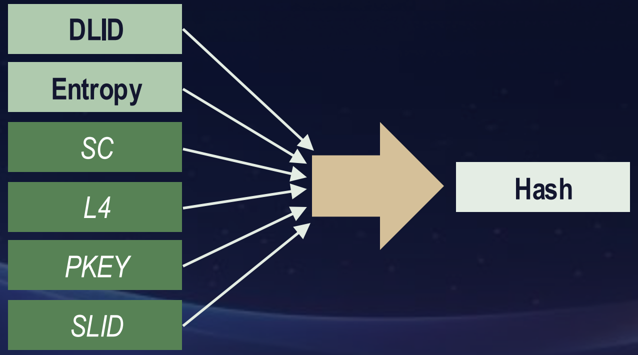
**Application semantics require in-order delivery**

All available next-hops to each destination programmed at each switch

**Packets within each flow arrive in order**

Packets marked for SDR

Hash value determines next-hop



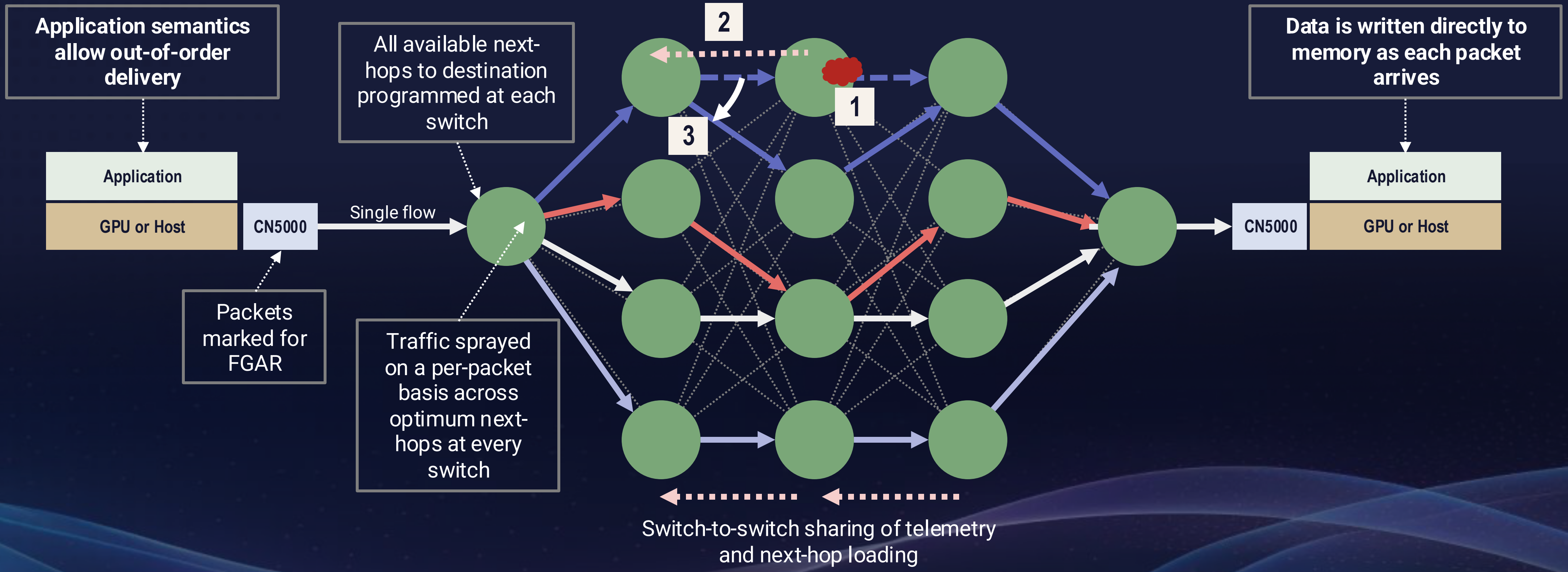
**Programmable Hash Function**

DLID – Destination LID
Entropy – **Software-controlled** field to identify related packets (e.g. flows with ordering requirements)

***Optional***
SC – Service Channel, combination of Virtual Lane and Traffic Class
L4 – Transport Mode
PKEY – Partition Key
SLID – Source LID

# 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



**Application semantics allow out-of-order delivery**

Packets marked for FGAR

All available next-hops to destination programmed at each switch

Traffic sprayed on a per-packet basis across optimum next-hops at every switch

Switch-to-switch sharing of telemetry and next-hop loading

**Data is written directly to memory as each packet arrives**

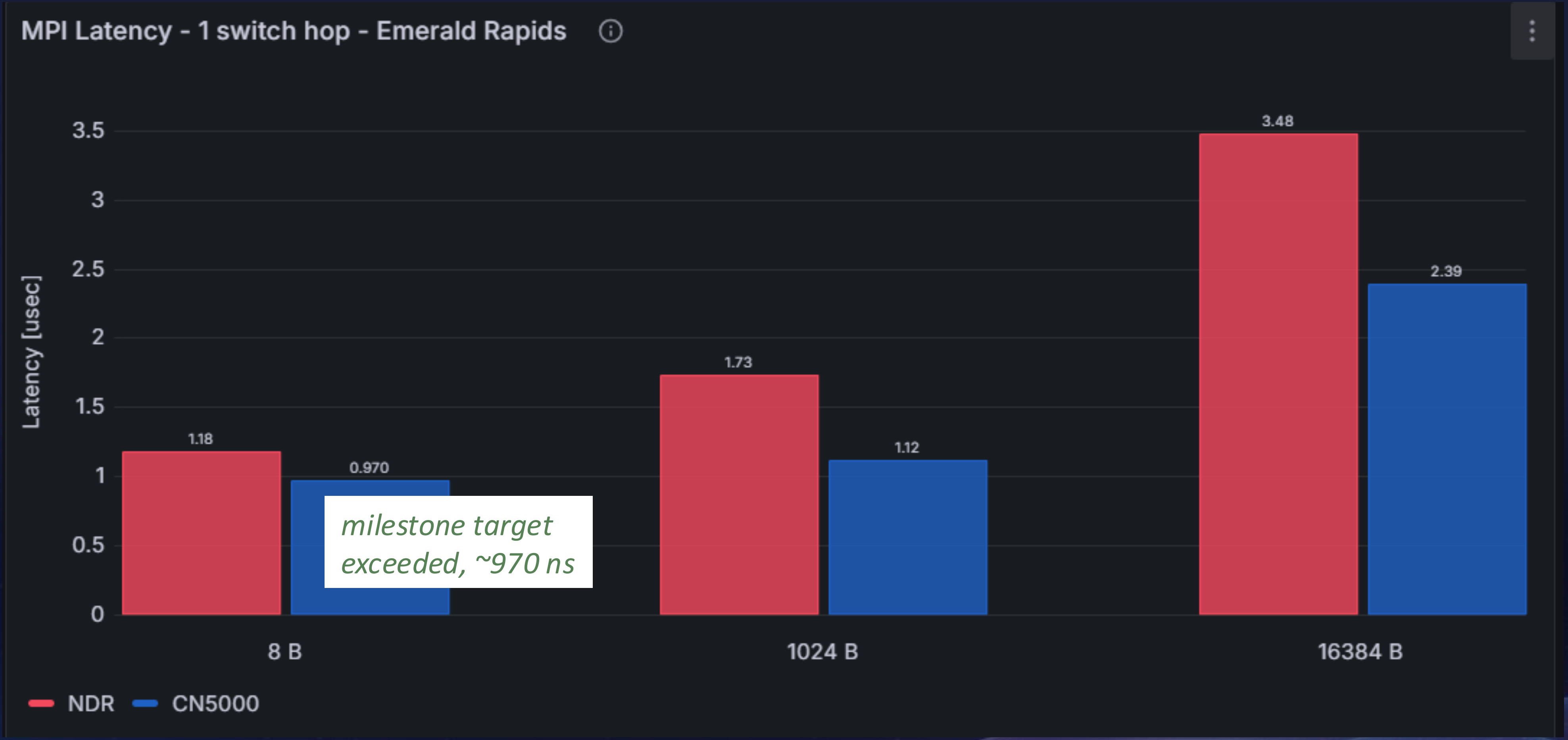
# CN5000 Performance Sneak Peek

CORNELIS™ NETWORKS

NOT FINAL PERFORMANCE.  RESULTS REPRESENT INITIAL OUT-OF-BOX MEASUREMENTS.

# CN5000: Sub-Microsecond MPI Latency on Emerald Rapids

**Testing as of 3/19/2025**


MPI Latency - 1 switch hop - Emerald Rapids

*milestone target exceeded, ~970 ns*

Better

Tests performed on 2 socket INTEL(R) XEON(R) PLATINUM 8568Y+. Intel(R) Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with acpi-cpufreq driver. Red Hat Enterprise Linux 9.2 (Plow). 6.5.0-rc1.upstream_v42+ kernel. 16x32GB, 512 GB total, Memory Speed: 5600 MT/s. Intel MPI 2021.14
Cornelis CN5000: Cornelis Omni-Path Express Suite (OPXS) 12.0.0P0.14, 1M copper cables, I_MPI_OFI_LIBRARY_INTERNAL=0 FI_PROVIDER=opx
NVIDIA NDR InfiniBand, hpcx-v2.22-gcc-doca_ofed-redhat9-cuda12-x86_64, 2M copper cables, FI_PROVIDER=mlx

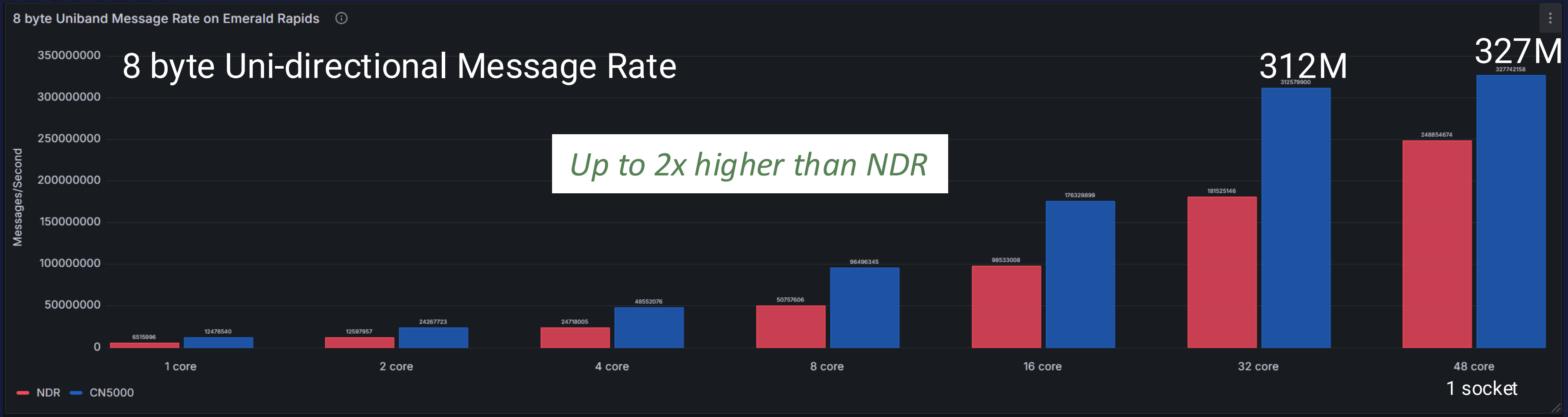
NOT FINAL PERFORMANCE. RESULTS REPRESENT INITIAL OUT-OF-BOX MEASUREMENTS.

# CN5000: Up to 2x Message Rate vs NDR on Emerald Rapids

Testing as of 3/19/2025

higher is better

Better



8 byte Uniband Message Rate on Emerald Rapids

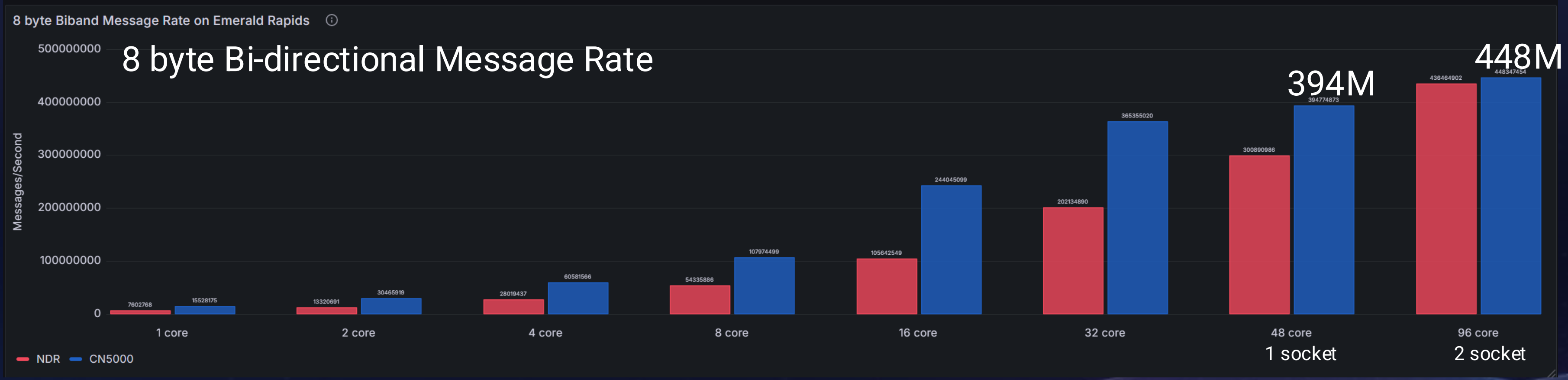
8 byte Uni-directional Message Rate

Up to 2x higher than NDR

312M

327M

1 socket



8 byte Biband Message Rate on Emerald Rapids

8 byte Bi-directional Message Rate

394M

448M

1 socket

2 socket

## Out-of-box, CN5000 is up to ~2x more performant than NVIDIA NDR

**Scaling and SW tuning in progress**

Tests performed on 2 socket INTEL(R) XEON(R) PLATINUM 8568Y+. Intel(R) Hyper-Threading Technology enabled. Intel(R) Turbo Boost Technology enabled with acpi-cpufreq driver. Red Hat Enterprise Linux 9.2 (Plow). 6.5.0-rc1.upstream_v42+ kernel. 16x32GB, 512 GB total, Memory Speed: 5600 MT/s. Intel MPI 2021.14
Cornelis CN5000: Cornelis Omni-Path Express Suite (OPXS) 12.0.0P0.14, 1M copper cables, I_MPI_OFI_LIBRARY_INTERNAL=0 FI_PROVIDER=opx
NVIDIA NDR InfiniBand, hpcx-v2.22-gcc-doca_ofed-redhat9-cuda12-x86_64, 2M copper cables, FI_PROVIDER=mlx

# Cornelis Networks Roadmap for AI and HPC

| Launch Year | 2025 | 2026 | 2027 |
|---|---|---|---|
| Product | 400G / PCIe Gen 5<br>CN5000 | 800G / PCIe Gen 6<br>CN6000 | 1600G / PCIe Gen 7<br>CN7000 |
| Status | Validation Testing | Implementation | Architecture |
| Protocol | Omni-Path | Omni-Path/Ethernet | Omni-Path/Ultra Ethernet |
| Features Overview | **Advanced Performance**<br>Delivering enhanced throughput and intelligent optimization for scalable, future-ready networks<br>• 400G OPA SuperNIC<br>• 48p OPA ToR/Edge Switches<br>• 576p OPA Director Switches<br>• 250K nodes in a single subnet<br>• First Customer Shipment: 1H25 | **Converged Connectivity**<br>Offering seamless bandwidth and flexibility with multi-protocol support for next-gen heterogeneous deployments<br>• 800G SuperNIC: OPA & RoCEv2<br>• ToR/Edge Switches<br>• Director Switches | **Universal Scale-Out Networking**<br>Providing revolutionary lossless interconnect for unparalleled performance and multi-vendor compatibility<br>• 1600G SuperNIC: OPA, RoCEv2 & Ultra Ethernet<br>• Fabric and app offloads<br>• ToR/Edge Switches<br>• Director Switches |

Leverages Proven CN5000 Architecture

# Thank you

For more information: **https://cornelisnetworks.com**

CORNELIS™ NETWORKS