

Congestion-Free Networking: Breaking the AI + HPC Bottleneck

White Paper

Executive Summary

Artificial Intelligence (AI) and High-Performance Computing (HPC) workloads are increasingly constrained, not by raw bandwidth, but by the unpredictable slowdowns caused by congestion and its resulting long-tail latencies. Even brief spikes in congestion can cascade into stalled GPU clusters, wasted resources, extended training times, and poor inference performance. Traditional interconnects like Ethernet with RDMA over Converged Ethernet (RoCEv2) and InfiniBand attempt to react to congestion after it has already taken hold. They rely on feedback loops that operate on round-trip timescales and mistake bit errors for congestion. This reactive approach leaves AI training and inference clusters vulnerable to performance collapse just when predictable throughput is most critical.

The Cornelis CN5000 Omni-Path is designed from the ground up to prevent congestion from forming and to neutralize it instantly if it appears. Omni-Path achieves congestion-free AI and HPC networking through:

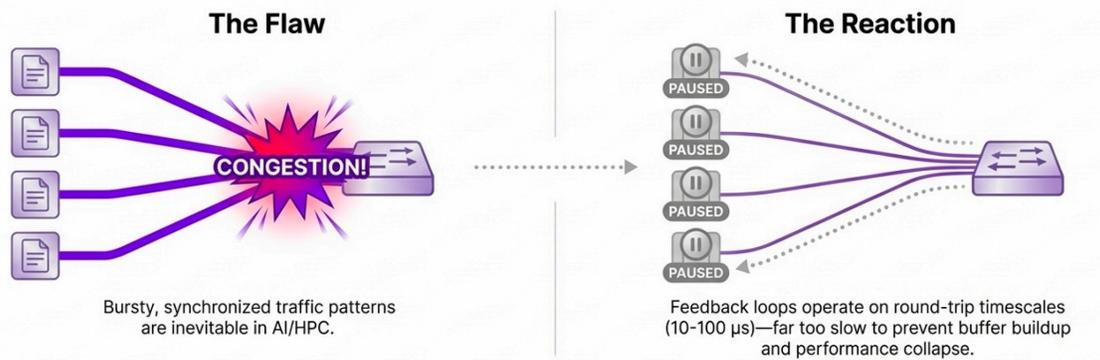
- Fine-grained adaptive routing (FGAR) which dynamically selects optimal paths on a per-packet basis within the network itself
- Incast-aware flow control that detects and mitigates many-to-one traffic overloads at the receiver
- Credit-based lossless transport that ensures reliable delivery without the pitfalls of Priority Flow Control (PFC)
- Link-level replay (LLR) for immediate local error correction

Overall, the Omni-Path fabric holds tail latency flat as the system scales, keeps GPUs and CPUs continuously productive, and delivers predictable performance across workloads of various scales.

Why Congestion Matters in AI and HPC

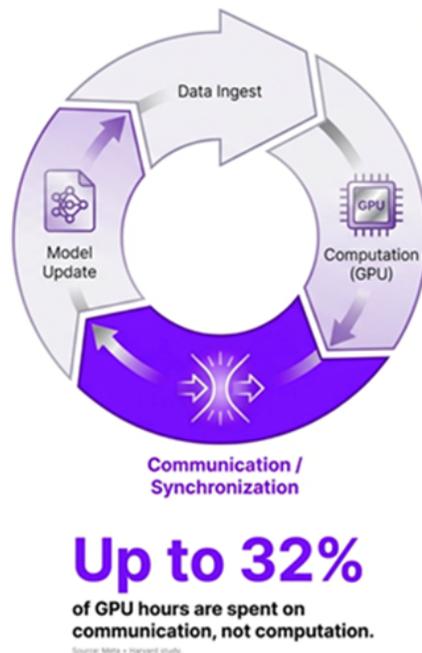
Congestion in distributed systems occurs when multiple flows contend for the same link or switch port output (refer to the following figure). This can cause buffer buildup, head-of-line (HoL) blocking, and delayed packet delivery. At a technical level, this manifests as queuing delays where packets are buffered excessively, leading to increased latency variance. This reactive approach leads to cascading failures including HoL blocking, congestion spreading, and massive spikes in tail latency.

Figure 1. Legacy networks allow congestion to form and then attempt to mitigate it



For AI training clusters, the impact of congestion is magnified. Synchronous collectives, such as AllReduce operations, where all nodes must exchange and reduce data (for example, gradients in neural network training), are gated by the slowest participant. A handful of delayed messages can stall thousands of GPUs (refer to the following figure).

Figure 2. The AI training loop and wasted capacity due to network bottlenecks



In HPC solvers, where tightly coupled timesteps must advance in lockstep, congestion creates imbalances that ripple through simulations. This often results in synchronization barriers that amplify small delays into significant runtime extensions.

Studies of production in deep learning platforms reveal that GPUs are frequently underutilized, with average utilization below 50%.¹ Much of this wasted capacity traces back to communication stalls, from checkpointing large models involving terabytes of data transfer, exchanging gradients across GPUs in distributed training frameworks like PyTorch or TensorFlow, to waiting on data transfers in workloads such as molecular dynamics simulations. In other words, the real bottleneck in AI and HPC today is not compute power, it is the network where unpredictable latency distributions undermine the determinism required for efficient scaling.

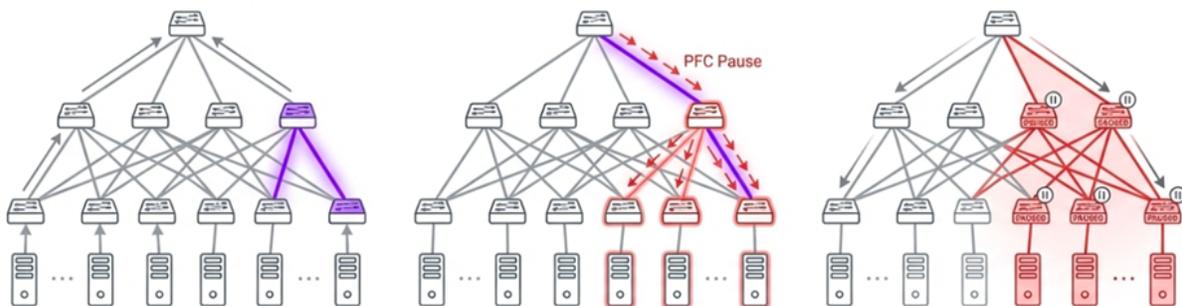
Why Legacy Approaches Fall Short

Ethernet with RoCEv2

Reactive networks, such as Ethernet with RoCEv2, operate on a model that implicitly accepts congestion as an inevitability. Ethernet fabrics with RoCEv2 rely on Explicit Congestion Notification (ECN) marking in switches and congestion notification packets to throttle senders retroactively. To maintain lossless delivery, they often employ Priority Flow Control (PFC) that pauses traffic on specific priority classes when Switch or SuperNIC buffers are near capacity. While functional at small scales, these mechanisms introduce well-documented side effects:

- HoL blocking
- Congestion spreading
- Congestion trees (refer to the following figure)
- Deadlocks in multi-path scenarios

Figure 3. Reliance on ECN and PFC to manage traffic leads to HoL blocking and congestion spreading



¹ [MAD-Max Beyond Single-Node: Enabling Large Machine Learning Model Acceleration on Distributed Systems.](#)

Even worse, the congestion-control loop operates at end-host round-trip times (RTTs) that are typically 10 to 100 microseconds. This is far too slow to prevent queues from bloating during sudden incast events—common in AI where thousands of nodes send data to one aggregator. Benchmarks can show that under incast loads, RoCEv2 can experience latency spikes exceeding 100x nominal values, leading to throughput collapse.

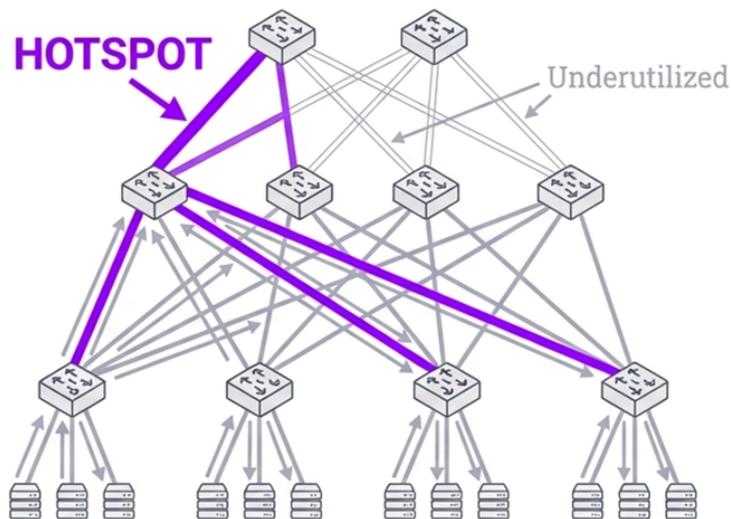
InfiniBand

Unlike RoCEv2, that relies heavily on Ethernet mechanisms such as ECN marking, PFC, and host-driven rate throttling to react to congestion, InfiniBand implements congestion control directly within the fabric. However, while InfiniBand moves congestion signaling into the fabric, it remains fundamentally reactive, responding to congestion after queues have already formed.

InfiniBand uses a combination of forward explicit congestion notification and backward explicit congestion notification marking, congestion notifications, and deterministic routing per destination, where paths are fixed based on local identifier hashing. Adaptive routing options exist in some vendor implementations; however, in practice, they often suffer from fairness issues where some flows dominate paths and still depend on delayed feedback loops that react after congestion builds.

The deterministic and inflexible nature of path selection can also concentrate flows and create hot spots. This is especially true in fat-tree or dragonfly topologies, leading to uneven load distribution and reduced bisection bandwidth utilization. Moreover, InfiniBand's reliance on end-to-end retransmissions for errors exacerbates these issues under the bit-error rates typical to real-world operational environments.

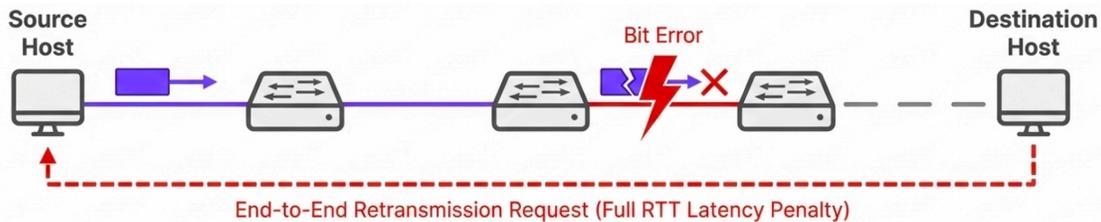
Figure 4. Deterministic routing per destination concentrates flows creating hotspots and uneven load distribution



Bit Errors Masquerading as Congestion

At today's link speeds (200–800 Gbps), random bit errors are inevitable. A bit error rate (BER) of 10^{-12} translates into roughly one error every 2.5 seconds at 400 Gbps, when following Poisson distribution models. Bit errors cause corrupted and lost packets that, in legacy RoCEv2 and InfiniBand networks, require an end-to-end retransmission request producing a full round trip latency penalty (refer to the following figure).

Figure 5. Latency spikes due to end-to-end retransmission requests



Often the packet loss is misinterpreted as signs of congestion, triggering congestion responses that throttle healthy traffic and cause avoidable slowdowns. In RoCEv2, this often leads to unnecessary rate reductions via data center quantized congestion notification; while in both RoCEv2 and InfiniBand, it invokes go-back-N retransmissions that amplifies latency tails.

Defining Zero Congestion

Zero Congestion is not a claim of absolute perfection but an architectural design point. It is a network fabric built so that congestion is statistically rare, short-lived, and fully contained. The goal is to keep latency distributions flat even under heavy load. This ensures that p99 and p99.9 latencies remain steady as clusters scale, typically within 10 to 20% of median latency as opposed to 10x spikes in legacy systems. Achieving this requires:

- **Avoidance** through fine-grained adaptive routing that spreads traffic evenly and reroutes around emerging hotspots using real-time fabric telemetry
- **Containment** through instant, in-fabric suppression of overload before it propagates into congestion trees, leveraging switch-level intelligence
- **Loss elimination** by using credit-based flow control and local link-level replay, preventing both packet drops and false congestion signals from bit errors
- **Freedom to reroute** by lifting in-order constraints, allowing packets to always take the best available path without reassembly overhead at the endpoint

Figure 6. Zero Congestion architectural design pillars



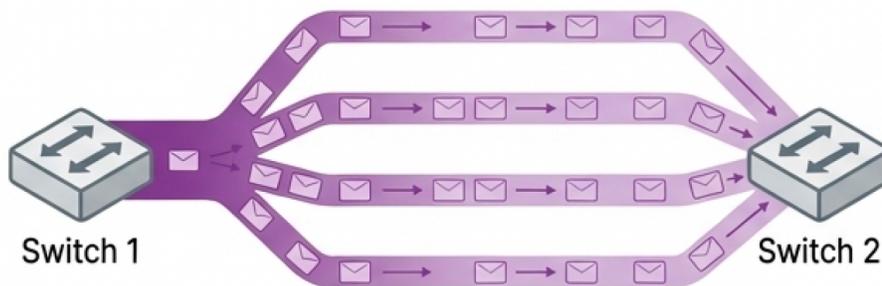
How Cornelis Omni-Path Delivers Zero Congestion

Fine-Grained Adaptive Routing

Unlike Ethernet's equal-cost multi-path hashing that assigns flows to paths statically, often leading to collisions or InfiniBand's deterministic but fixed and inflexible routing, Omni-Path employs per-packet, fabric-coordinated adaptive routing. This allows the CN5000 Omni-Path 400G end-to-end solution to continuously disperse flows across all available paths, also known as packet spraying.

In topologies like fat-trees or dragonflies, FGAR can avoid the building of hot spots and prevent unfair load distribution. Adaptive routing decisions are made at sub-RTT timescales (nanoseconds) within the Omni-Path network itself using a combination of switch-local metrics like buffer occupancy and link utilization along with real-time congestion telemetry shared by neighbor switches. This eliminates the delay of end-host signaling. In practice, FGAR can improve bandwidth utilization significantly in unbalanced workloads, as it incrementally adjusts paths based on intelligent algorithms that weigh alternatives dynamically.

Figure 7. Adaptive routing achieves packet spraying while avoiding hotspots and enabling full utilization



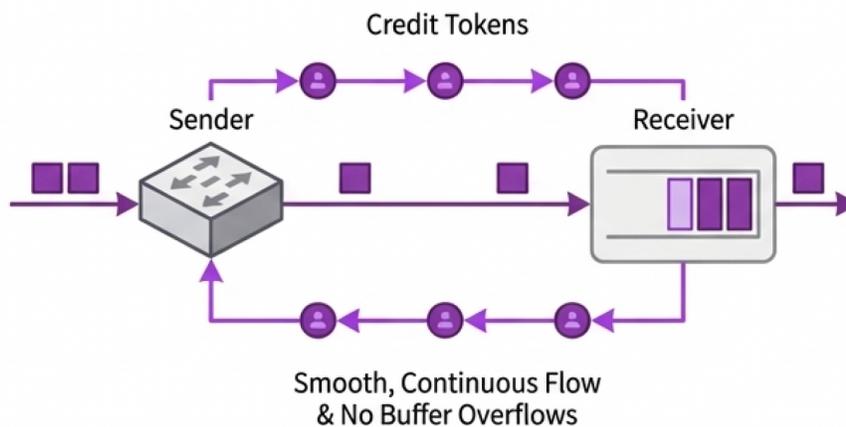
Incast-Aware Flow Control

CN5000 Omni-Path detects many-to-one overloads at the receiver, common in AI all-to-all operations or HPC reductions and directly signals upstream switches to quell the burst before queues build uncontrollably. This is implemented using enhanced congestion notifications that pace senders proactively, using telemetry to modulate injection rates. This is critical in AI workloads, where thousands of workers may simultaneously transmit gradients or checkpoints to a single node, potentially overwhelming buffers. With incast-aware flow control, Omni-Path prevents the collapse that plagues Ethernet fabrics under similar conditions, maintaining throughput even at massive incast ratios.

Credit-Based, Lossless Transport

By using a credit-based system, where receivers grant credits representing available buffer space, Omni-Path maintains lossless operation without relying on PFC. Credits are exchanged at the link level, ensuring senders only transmit when space is guaranteed. This removes the risk of HoL blocking, congestion spreading, and deadlocks while still ensuring reliable packet delivery even at scale. This contrasts with PFC's class-based pauses, offering finer granularity, lower latency, and more reliable packet delivery.

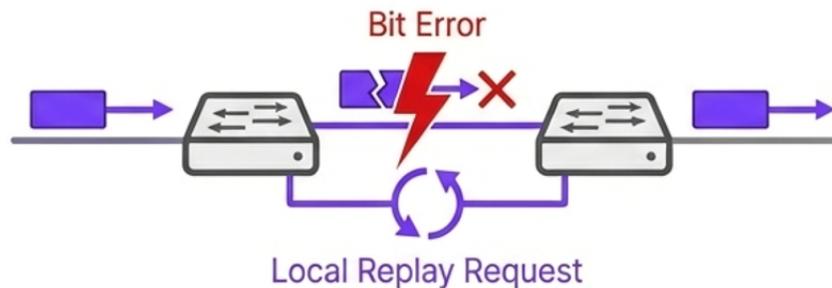
Figure 8. Lossless and guaranteed delivery using buffer credits



Link-Level Retry

CN5000 Omni-Path implements local error correction at each hop using LLR. Transient bit errors are detected using 14-bit CRCs and replayed immediately between switches, often with zero added latency, avoiding the end-to-end retransmissions that plague other fabrics. Combined with multiple levels of forward error correction for higher BER tolerance, this ensures stable throughput at high speeds and prevents random errors from triggering unnecessary congestion responses. Omni-Path dynamic lane scaling further enhances resilience by adjusting active lanes on degraded links.

Figure 9. Link Level replay avoids the latency costs of end-to-end retransmissions



Hardware and Software Stack

The CN5000 portfolio is the physical embodiment of the Omni-Path architecture, providing a complete, end-to-end 400G networking solution.

- The CN5000 SuperNIC is a high-performance network adapter designed to connect servers and accelerators to the Omni-Path fabric.
- The CN5000 Switch provides 48×400G ports with more than 19.2 billion packets per second throughput and sub-microsecond MPI latency.
- The CN5000 Director Class Switch supports up to 576 ports.

SuperNIC and Switch architecture integrate FGAR, incast-aware flow control, credit-based transport, and LLR natively in silicon. On the software side, Omni-Path leverages an open libfabric provider, has full legacy verbs compatibility, a scalable fabric manager for topology discovery and routing optimization, and the OPX Software for seamless integration with MPI libraries like MVAPICH2 or Open MPI, and GPU libraries such as CUDA, ROCm, NCCL, and RCCL. All the hardware and software work together to ensure interoperability in HPC and AI environments.

Why Cornelis is the Right Choice for AI and HPC

AI jobs, such as large language models with billion-parameter scales and enterprise re-training/inferencing; HPC solvers, such as finite element analysis; and large-scale checkpointing workloads all demand a fabric that keeps GPU and CPU resources fully utilized.



With a goal of >90% efficiency at scale, Cornelis Omni-Path delivers exactly that. By combining advanced congestion avoidance and containment mechanisms, lossless transport without PFC, and local replay for error handling, Omni-Path attempts to hold tail latency flat as clusters scale to hundreds of thousands of endpoints. The result is predictable performance, higher parallel efficiency (up to 2.5x message rates versus competitors), and faster time-to-solution. Benchmarks show this with up to 45% lower latency and 2x higher application performance.

Where Ethernet and InfiniBand struggle with tuning sensitivity, congestion trees, and reactive loops, Omni-Path is proactive, adaptive, and resilient. It is not just another high-speed interconnect; it is a network deliberately engineered for the zero-congestion era of AI and HPC.

Conclusion

The future of AI and HPC depends on keeping accelerators busy and predictable. Congestion is the hidden tax on every cluster, sapping utilization and inflating runtimes. With the CN5000 Omni-Path 400G end-to-end networking solution, Cornelis has built the first interconnect designed to eliminate this tax. Through architectural innovations such as fine-grained adaptive routing, incast-aware flow control, credit-based transport, and link-level replay, Omni-Path enables clusters of any size to run with consistent, congestion-free performance. It is the network AI and HPC have been waiting for.

Legal Disclaimer

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Cornelis Networks products described herein. You agree to grant Cornelis Networks a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All product plans and roadmaps are subject to change without notice.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cornelis Networks technologies may require enabled hardware, software, or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

Cornelis, Cornelis Networks, Omni-Path, Omni-Path Express, and the Cornelis Networks logo belong to Cornelis Networks, Inc. Other names and brands may be claimed as the property of others.

Copyright © 2026, Cornelis Networks, Inc. All rights reserved.