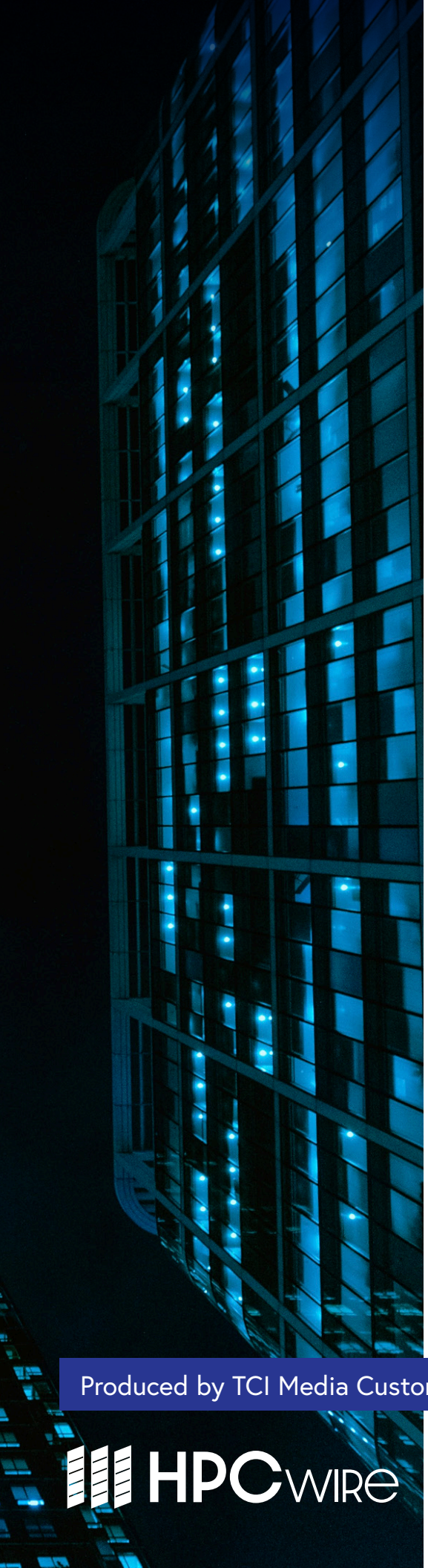

IMPROVING THE ECONOMICS OF LARGE-SCALE AI

Breaking the AI Networking Bottleneck

Produced by TCI Media Custom Publishing in conjunction with:

HPCwire

CORNELIS™ NETWORKS

TCI MEDIA
CUSTOM PUBLISHING

AI growth is crushing traditional networks – and driving up costs. Cornelis Networks' CN5000 family offers a congestion management solution that redefines network efficiency for large-scale AI.

Network infrastructure is pushed to its limits as it struggles to perform at the scale required to accommodate the explosion of AI growth. The rapid proliferation of AI models demands the movement of enormous volumes of data for training and deployment, which places a substantial burden on legacy and even modern network systems.

Though the workloads and network patterns differ across training and inference, the network's burden remains constant. The training and fine-tuning phases are bandwidth-intensive, while inferencing is latency-intensive.

Training and fine-tuning demand persistent, high-throughput access to massive datasets—often distributed across storage systems or cloud zones—even though computing is typically centralized. Data orchestration platforms may replicate remote datasets locally to accelerate training, but large-scale transfers and accelerator synchronization still stress the network, especially during initial loading and parallel processing. These demands create severe traffic congestion in AI pipelines, whether from high-throughput model training or sustained inference delivery across large user bases.

Bottlenecks, delays, and increased operational costs adversely affect the return on investment for AI initiatives, especially in large-scale training regimes where data flows need to be continuous and coordinated. The high bandwidth nature of these flows magnifies these issues, exacerbated in latency-sensitive inference tasks like real-time analytics or streaming decision-making, where microsecond delays considerably worsen performance.

Organizations are now searching for adaptive, flexible networking strategies that can evolve and scale alongside rapidly advancing AI technologies. This includes investing in AI-focused high-performance network solutions that ensure consistent AI application performance, even under heavy load from multiple concurrent applications.

The division between the AI goal and the underlying framework is profound—with networking being the starkest example. A Network World survey indicated that 42% of IT managers feel their infrastructure cannot fully support AI workloads. This poses a strategic challenge: when networks lag behind the rapid pace of AI development and deployment, decision-making slows, time-to-insight suffers, and competitiveness in agile markets is put at risk.

TCI MEDIA
CUSTOM PUBLISHING

Economic and Performance Implications of Network Bottlenecks

AI is showing business leaders that network infrastructure represents far more than a technical challenge. It's an economic and market challenge, too. With each newly released AI model or capability, infrastructure and energy costs are pushed higher.

As AI becomes more deeply embedded in business operations, the financial and operational risks associated with underperforming networks are growing rapidly. AI-driven workloads require massive data throughput, low latency, and high availability. Congestion, latency, and synchronization failures prevent those requirements from being fully realized.

Traditional enterprise networks were not built to deliver these characteristics at such a scale, nor were they designed to continue scaling alongside evolving AI models with no performance degradation. That's a serious issue because network bottlenecks do not just impact performance; they directly affect business agility, decision-making speed, and the ability to generate returns on AI investments. Further, costs increase dramatically from missed SLAs, extended training time, or underutilized compute.

Each new release of AI software or a model increases infrastructure and energy spending, impacting the company's finances. The last few years have seen model sizes roughly double in six-month intervals, increasing the need for compute power and inter-node communication bandwidth necessary to train and operate them effectively.

During training, networks must provide dataset access and continuous dynamic synchronization across distributed GPUs… …fetching gradients, weights, and activations with extremely low lag every instant. Training stalling occurs without a fabric that can scale and provide low latency for these rigid communication patterns—real-time AI analytics become unresponsive, AI services perform below expectations, users get frustrated, and SLAs become compromised.

Stockholder and market demand are putting pressure on organizations to improve network performance. As organizations rush to deploy AI, network limitations are becoming a silent drag on project timelines, operational budgets, and revenue potential. Business leaders are caught in the squeeze.

Resource Underutilization

Even the most sophisticated AI accelerators and computational resources cannot overcome network bottlenecks. These accelerators are built to execute matrix multiplications and tensor operations at massive scale and speed, but they depend on fast, consistent data movement to reach their theoretical performance potential. Without a high-bandwidth, low-latency networking environment to match their throughput capabilities, these systems become underutilized, idling as they wait for data to arrive or synchronize across distributed nodes. This mismatch between compute power and network performance drastically limits the reach and efficiency of large-scale AI workflows—including training, fine-tuning, and inference.

The impact of these network limitations isn't just technical—it's deeply strategic. Stockholder and market demand are putting pressure on organizations to improve network performance. As organizations rush to deploy AI, network limitations are becoming a silent drag on project timelines, operational budgets, and revenue potential. Business leaders are caught in the squeeze.

Network-induced performance bottlenecks drain resources and delay time-critical outcomes across industries. For example, AI-enabled diagnostics or real-time image analysis lags in healthcare may postpone essential interventions in a timely decision-making workflow. In logistics, AI delays can interfere with routing optimization, and in financial systems, even minimal delays can affect transaction execution. Ultimately, subpar network performance undermines AI. Organizations need tailored network architecture built from the ground up to address these challenges to eliminate congestion, optimize workflows, and seamlessly scale to address AI requirements.

Current Networking Bottlenecks

AI systems face critical networking bottlenecks that create both technical and economic challenges. Host and GPU inefficiency undermines performance as AI models scale to trillions of parameters across thousands of accelerators, resulting in expensive compute resource underutilization. Complex congestion management becomes increasingly difficult at an unprecedented scale, while persistent packet loss and tail latency issues create performance inconsistencies that compound throughout the system. Addressing these networking constraints represents a strategic imperative that can yield significant competitive advantages, reduce computational overhead, and accelerate time-to-insight.

The industry increasingly recognizes that AI workloads require fundamentally new approaches to networking. While industry efforts like the Ultra Ethernet Consortium work toward future standards, the CN5000 delivers those capabilities today.

Cornelis CN5000: Solving AI Networking Bottlenecks

Current networking architectures face challenges with host efficiency, packet loss, and tail latency—critical bottlenecks that extend AI model training times *and delay inference and fine-tuning delivery*.

The industry increasingly recognizes that AI workloads require fundamentally new approaches to networking. While industry efforts like the Ultra Ethernet Consortium work toward future standards, the CN5000 delivers those capabilities today.

Leading this transformation, the Cornelis CN5000 eliminates bottlenecks and unlocks the full potential of AI infrastructure through advanced congestion management and intelligent flow control. By integrating these innovations into a holistic, scalable solution, the CN5000 delivers unmatched performance across the full AI lifecycle — *from training and fine-tuning to inference* — operational efficiency and seamless integration for both training and inference workflows.

The CN5000 open, performance-enhanced software stack accelerates deployment and simplifies integration, delivering faster time-to-value for AI and HPC environments.

Built around the three core pillars of modern AI infrastructure—Performance, TCO, and Scalability — CN5000 delivers:

Performance Optimization

- Maximum throughput across distributed workloads by eliminating network congestion and minimizing tail latency
- Peak accelerator utilization through high-speed, deterministic data movement
- Accelerated collective communications for faster training of large AI models
- Real-time inference performance with industry-leading low latency
- Faster job completion times for large language model (LLM) training and deployment

Optimal Total Cost of Ownership (TCO)

- Lower power consumption per compute job, directly reducing operational expenses
- Intelligent workload placement and network tuning through deep telemetry and analytics
- Greater network efficiency and reduced infrastructure waste

Scalability & Interoperability

- Consistent performance through guaranteed reliability and fine-grained QoS across every network hop
- Seamless integration with open-standards-based software and infrastructure
- Support for scaling to 400G and beyond via a future-proof architecture

CN5000 achieves these outcomes through a unique combination of adaptive routing, advanced congestion control, and fine-grained Quality of Service mechanisms. This approach ensures optimal performance even under the most demanding workloads, with telemetry engines delivering unmatched visibility and control over traffic flow and congestion hotspots.

The CN5000 open, performance-enhanced software stack accelerates deployment and simplifies integration, delivering faster time-to-value for AI and HPC environments. In short, CN5000 powers end-to-end infrastructure efficiency — supporting the speed, scale, and reliability that modern enterprises demand.

Unlock the Full Potential of Your AI Infrastructure

Don't let networking bottlenecks limit your AI Infrastructure ROI. The CN5000 delivers deterministic performance, advanced congestion control, and scalable fabric integration — ready for today's LLMs and tomorrow's AI demands.

Contact Cornelis Networks for a technical deep dive or POC evaluation: sales@cornelisnetworks.com.