Smarter Interconnects for Power-Constrained AI

White Paper by GigaIO

CONTENTS

GigaIO's AI fabric vs. RoCE Ethernet: A Comprehensive Performance Analysis for AI Workloads

Interconnect technology in AI infrastructure represents one of the most important but often neglected aspects of machine learning performance. This comprehensive benchmark study reveals striking performance improvements of GigaIO's AI fabric technology (FabreX) over RDMA over Converged Ethernet (RoCE), the improvements can be observed in raw throughput metrics and in application benchmarks. Through rigorous testing across diverse AI workloads, we present compelling evidence that the right interconnect choice can dramatically impact training, fine-tuning, and inference performance, particularly for cutting-edge large language models and distributed computing scenarios.

The tests demonstrate that GigaIO's FabreX interconnect offers a significant improvement over RoCE Ethernet for AI workloads — delivering much higher performance, lower power consumption, reduced infrastructure costs, and much simpler deployment. Across training, fine-tuning, and inference workloads, using both AMD and NVIDIA GPUs, **FabreX consistently outperforms RoCE.** For organizations facing constraints in power, budget, or engineering time, FabreX facilitates faster time-to-value and more scalable AI deployments.

We specifically tested a range in multiple dimensions — older and slower GPUs and the very latest and fastest, different vendors (AMD and NVIDIA), and different formats — some that fit into PCIe slots and others that use the latest inter-GPU interconnects. Across the testing studies we implemented ethernet and FabreX setups that have the same theoretical bandwidth to focus on the differences of the interconnects. We used the same operating systems, frameworks, application software, GPUs, and varied the interconnect. Across the multiple dimensions, GigaIO's FabreX simply performed faster and better.

The Importance of Interconnect Technology in AI Infrastructure

When building high-performance AI infrastructure, much attention focuses on GPU selection and storage configurations. Interconnect technology that binds these components often receives little thought (because there have been so few alternatives) despite its profound impact on overall system performance. Our benchmark study directly addresses this knowledge gap by comparing two leading interconnect approaches: GigaIO's PCIe-based AI fabric and RoCE Ethernet's network-oriented architecture.

As AI models continue growing in size and complexity, the efficiency of data movement between computational units increasingly determines practical performance limits and power usage. The bandwidth and latency characteristics of interconnect technologies create seamless scaling or frustrating bottlenecks. Our analysis reveals that despite the same

theoretical maximum bandwidth between the tested GigaIO's AI fabric and RoCE configurations, their real-world performance differs substantially across various AI workloads.

This benchmark provides actionable intelligence for infrastructure architects and ML engineers facing crucial design decisions. By quantifying performance differences across specific AI workloads, we empower teams to make informed choices aligned with their particular use cases rather than relying on generic specifications or anecdotal evidence. The findings illuminate clear patterns of relative advantage that should inform strategic infrastructure planning for organizations deploying modern AI systems.

Understanding GigaIO's AI fabric and RoCE Interconnect Technologies

GigaIO AI fabric Architecture

GigaIO's AI fabric represents a fundamentally different approach to system interconnection by extending PCIe as a system-wide fabric. Unlike traditional network technologies that require protocol translation and additional software layers, GigaIO's AI fabric maintains native PCIe communication throughout the system. This architecture allows direct memory access between components with minimal protocol overhead, offering lower latency for critical AI workloads.

The core innovation of GigaIO's AI fabric lies in its ability to maintain the simplicity and efficiency of internal PCIe communication while extending these benefits across multiple chassis. This approach aligns particularly well with the communication patterns of distributed AI workloads, where frequent synchronization of large data volumes between GPUs drives overall system performance. The PCIe-native approach eliminates several layers of networking overhead present in traditional solutions, reducing inter-GPU communication.

In our benchmark configuration, we implemented GigaIO's AI fabric as an internal PCIe Gen4 x16 connection, providing 256 Gbit/sec bandwidth. This configuration represents a typical deployment scenario for organizations seeking to maximize performance for AI workloads while maintaining reasonable infrastructure costs and complexity.

RoCE Ethernet

RDMA over Converged Ethernet (RoCE) represents a classic network-oriented approach to high-performance interconnects. By implementing Remote Direct Memory Access capabilities over standard Ethernet, RoCE attempts to combine the ubiquity and familiarity of Ethernet with the performance benefits of RDMA. RoCE's primary advantage lies in its compatibility with widely deployed Ethernet infrastructure and networking expertise. Organizations with substantial investments in Ethernet networking can leverage RoCE to improve performance

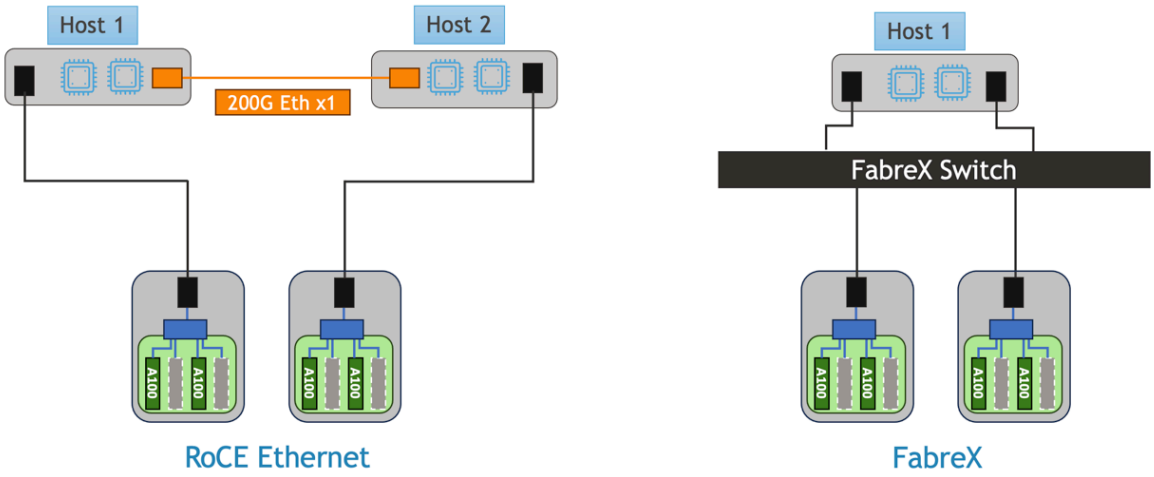
without completely abandoning their existing infrastructure or expertise. However, this approach necessarily includes additional protocol layers and potential bottlenecks compared to PCIe-native solutions like GigaIO's AI fabric.

RoCE's compatibility advantages come with tradeoffs in complexity and potential performance limitations. While modern RoCE implementations have dramatically improved over previous generations, they still require traversing the networking stack, which introduces additional latency and processing overhead compared to direct PCIe communication. Our testing sought to quantify these differences in real-world AI workloads rather than theoretical scenarios.

NVIDIA A100-80GB Training, Fine-Tuning, and Inference Performance Analysis

Test Methodology and Environment

Our testing utilized a consistent hardware environment to isolate the impact of interconnect technology. The test platform featured four NVIDIA A100-80GB GPUs, representing a common high-performance computing configuration for AI workloads. This multi-GPU setup allows for evaluation of both single-GPU performance and multi-GPU scaling efficiency across different interconnect configurations.



NVIDIA A100 GPU Test Configuration

The methodology involved creating two otherwise identical system configurations, differing only in the interconnect technology. The first configuration utilized GigaIO's AI fabric as an internal PCIe Gen4 x16 connection between nodes, while the second replaced this with a 200Gb Ethernet connection implementing RoCE.

This controlled test environment confirms that the measured performance differences represent genuine technological distinctions rather than testing artifacts or configuration anomalies.

Benchmark Selection and Workload Characteristics

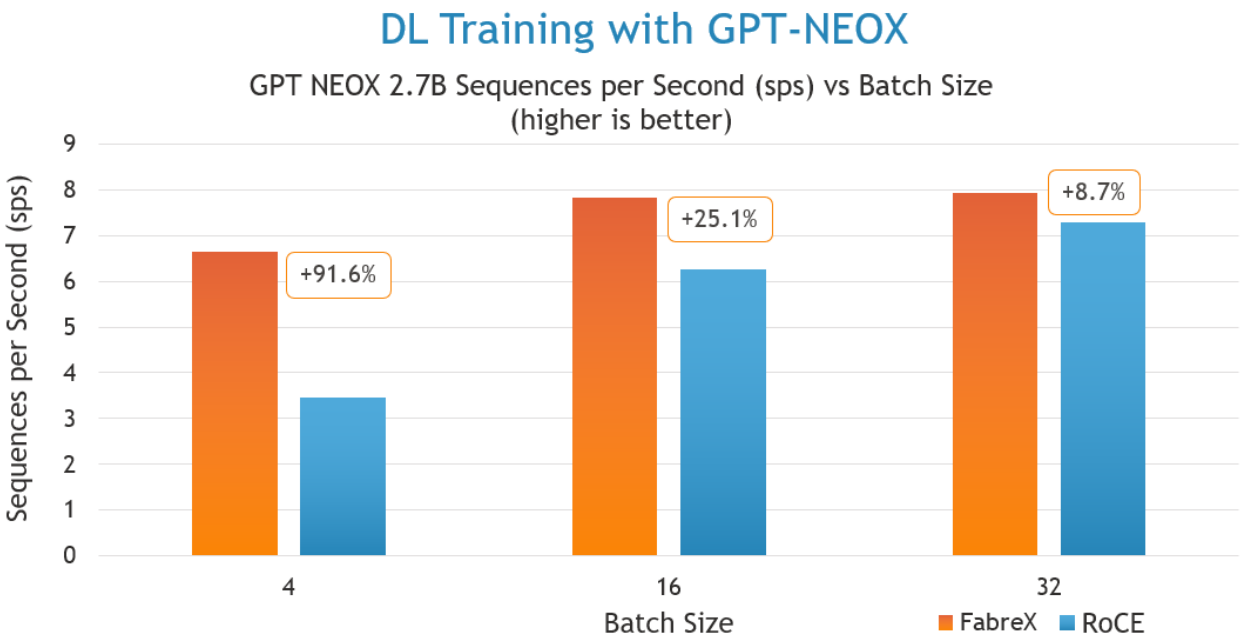
We selected benchmark workloads representing real-world AI applications rather than synthetic tests, ensuring results with direct practical relevance. For training performance, we utilized GPT-NeoX, an open-source implementation of large language models similar to GPT-3. Two model sizes were evaluated: GPT-NeoX 2.7B and 1.3B, allowing assessment of scaling properties across different model complexities.

For inference testing, we chose SGLang, a framework specialized for the inference of large language models. This benchmark measured two critical metrics: median time-to-first token (TTFT) and requests per second. Additionally, we tested the Llama 3.2-90B Vision Instruct model, one of the most significant multimodal models currently available for inference deployments.

These workloads were selected to represent the range of diverse communication patterns and computational characteristics found in modern AI applications. Training workloads feature frequent all-reduce operations for gradient synchronization, while inference workloads emphasize low-latency response and sustained throughput. Including smaller models and massive models requiring distribution across multiple GPUs provides insights across deployment scenarios.

Small-Batch Training Advantage

One of the most striking findings from our benchmark was GigaIO's AI fabric's significant performance advantage in small-batch training scenarios, which are typical of LLM fine-tuning. When training the GPT-NeoX 2.7B model, GigaIO's AI fabric consistently demonstrated higher sequences per second (sps) metrics compared to RoCE Ethernet, with the performance gap widening dramatically as the batch size decreased.



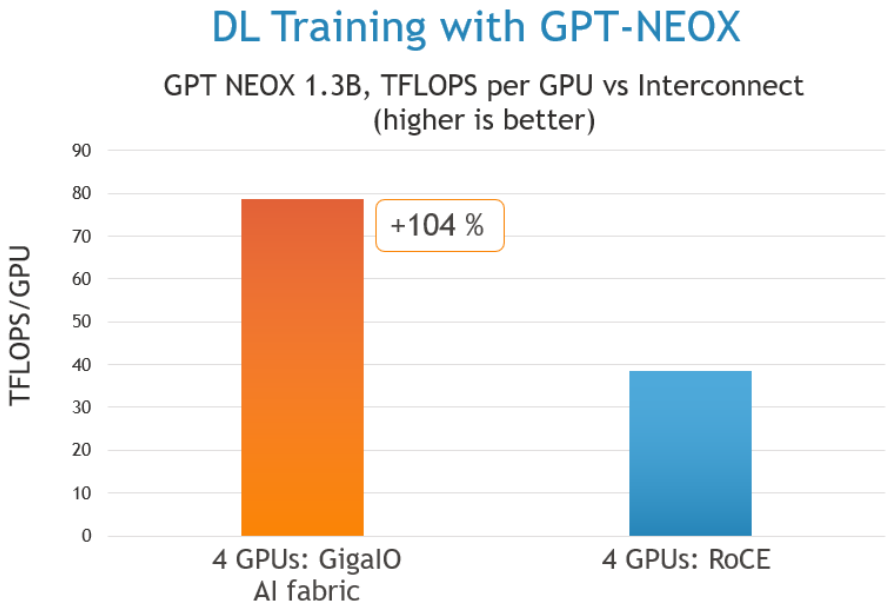
Small Batch Training and Fine-Tuning

This advantage for small batch training has significant practical implications. Small batch sizes are often used during model fine-tuning phases, when working with limited datasets, or when GPU memory is a constrain for larger models. Maintaining high performance even with small batch sizes can substantially reduce development cycle times, particularly during the iterative optimization phase of AI model development. Organizations focused on the rapid development and deployment of specialized models may find this advantage particularly compelling.

The performance gap stems from GigaIO's AI fabric's lower latency for the frequent communication events during distributed training. With smaller batch sizes, the ratio of computation to communication shifts toward more frequent synchronization, making efficient communication increasingly critical to overall performance. GigaIO's AI fabric's PCIe-native architecture provides substantial benefits in precisely these communication-intensive scenarios.

Scaling Efficiency Across Multiple GPUs

The benchmark also evaluated how efficiently performance scales as more GPUs are added to the training process. For the GPT-NeoX 1.3B model, we measured TFLOPS per GPU across 4 GPUs, revealing GigaIO's AI fabric's superior scaling efficiency. GigaIO's AI fabric maintained higher TFLOPS per GPU than RoCE Ethernet, indicating better utilization of available computational resources.



Scaling Efficiency Across GigaIO FabreX vs. RoCE

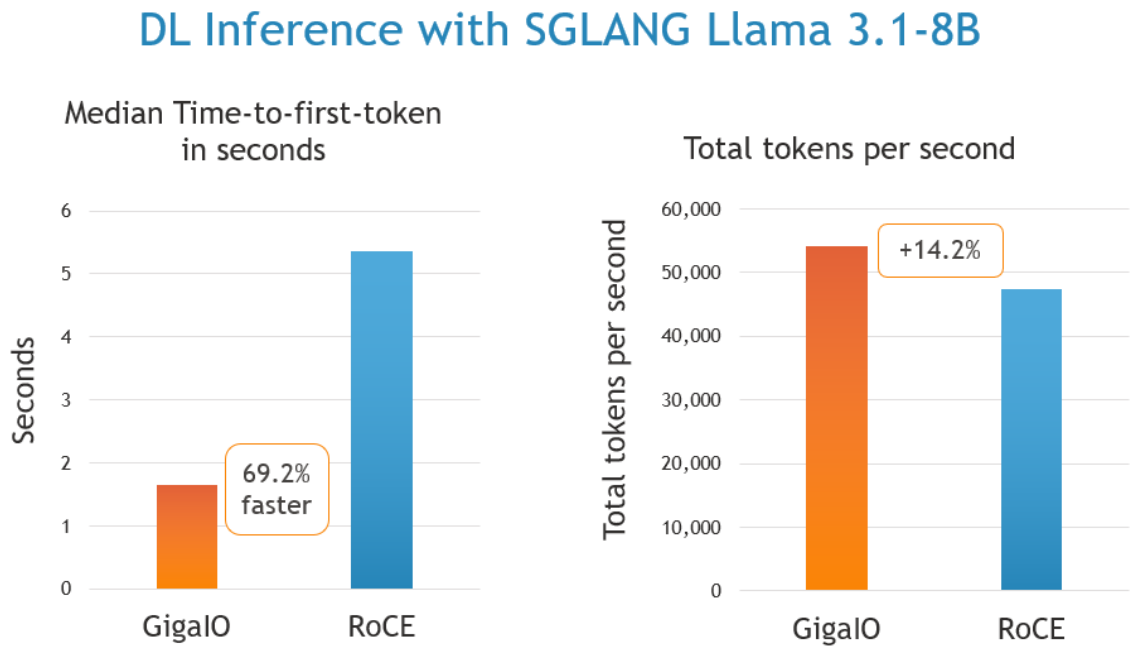
This scaling advantage becomes increasingly important as organizations deploy larger clusters for distributed training of state-of-the-art models. GigaIO's AI fabric's ability to minimize performance degradation translates directly to faster training times and better resource utilization for large-scale distributed training.

The observed scaling advantage results from reduced communication latency and overhead during the frequent all-reduce operations required for gradient synchronization in distributed training. As more GPUs participate in training, the complexity and volume of these synchronization operations increase substantially. Because of its inherently lower latency, GigaIO's AI fabric architecture is better suited to handling this increasing communication complexity without introducing significant performance penalties.

Inference Performance: Responsiveness and Throughput

Time-to-First Token Under Load

For interactive AI applications, time-to-first token (TTFT) is a critical user experience metric that measures how quickly the system generates a response after receiving a query. Our testing revealed that GigaIO's AI fabric delivers substantially better TTFT performance compared to RoCE Ethernet, particularly when the server is under load, which represents typical production deployment scenarios.



COMPARATIVE INFERENCE PERFORMANCE

This latency advantage translates directly to improved user experience for interactive AI applications. Reducing the response time from 5.3 seconds to 1.8 seconds in conversational AI systems can dramatically improve user engagement and satisfaction. The performance gap became increasingly pronounced under higher server loads, suggesting that GigaIO's AI fabric provides more consistent performance in production environments where multiple requests from multiple users are processed simultaneously.

The observed TTFT advantage stems from GigaIO's AI fabric's reduced protocol overhead. When a model is distributed across multiple GPUs, generating the first token requires coordination between these GPUs, with any communication latency directly impacting response time. GigaIO's AI fabric's architecture minimizes these coordination delays.

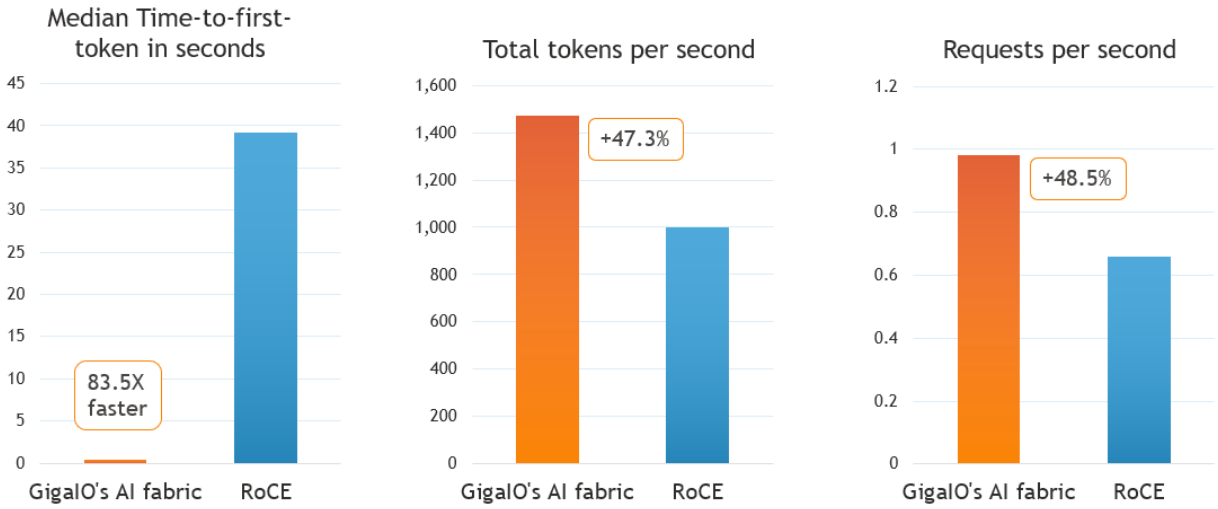
Throughput Advantages for Model Parallelism

Beyond initial response time, sustained throughput (measured in tokens per second or requests per second) determines the overall capacity of inference systems and the ability of the system to support more simultaneous users. Our benchmarks showed that GigaIO's AI fabric delivered higher throughput than RoCE Ethernet, with the advantage becoming particularly significant when model parallelism was required due to model size.

This throughput advantage has significant implications for deploying very large models that exceed the memory capacity of individual GPUs. As models like Llama 3.2-90B Vision Instruct require distribution across multiple GPUs, communication efficiency between these GPUs increasingly dictates overall system performance. GigaIO's AI fabric showed notably higher requests per second than RoCE when running these large, distributed models. GigaIO's

superior requests per second performance shows that RoCE systems need 35% more equipment, adding considerable cost and dramatically more power usage (and power) to handle the same number of concurrent users compared to GigaIO.

Multi-GPU DL Inference with SGLANG Llama 3.2-90B Vision Instruct



Multi-GPU Inference Performance

These accelerators often provide exceptional computational efficiency but require effective communication infrastructure when deploying larger models. GigaIO's AI fabric's superior performance for model parallelism suggests it may be particularly well-suited for these emerging specialized hardware ecosystems.

The inferencing benchmarks demonstrate two important benefits of using FabreX:

1) FabreX delivers substantially better time-to-first-token performance. This gap cannot be closed using an Ethernet RoCE interconnect.
2) With FabreX, users can deliver the desired simultaneous number of user metrics using 35-40% less hardware. Both systems can achieve similar simultaneous user metrics, but the system cost and power usage will be 35-40% higher using Ethernet RoCE interconnect.

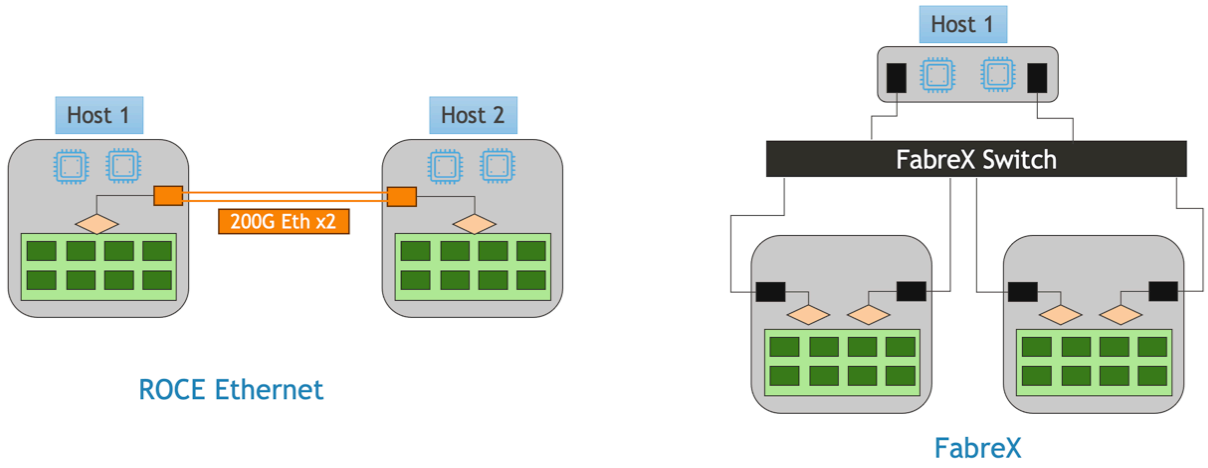
AMD MI300X Multi-Node Distributed Training Performance

GPT-NeoX Training Throughput on a 16×AMD MI300X Cluster

To further validate the interconnect impact, we performed distributed training of a GPT-NeoX language model across a cluster of 16 AMD Instinct MI300X accelerators under both GigaIO and RoCE Ethernet fabrics. The GigaIO-enabled cluster consistently achieved higher training throughput (measured in sequences per second) across all tested batch sizes.

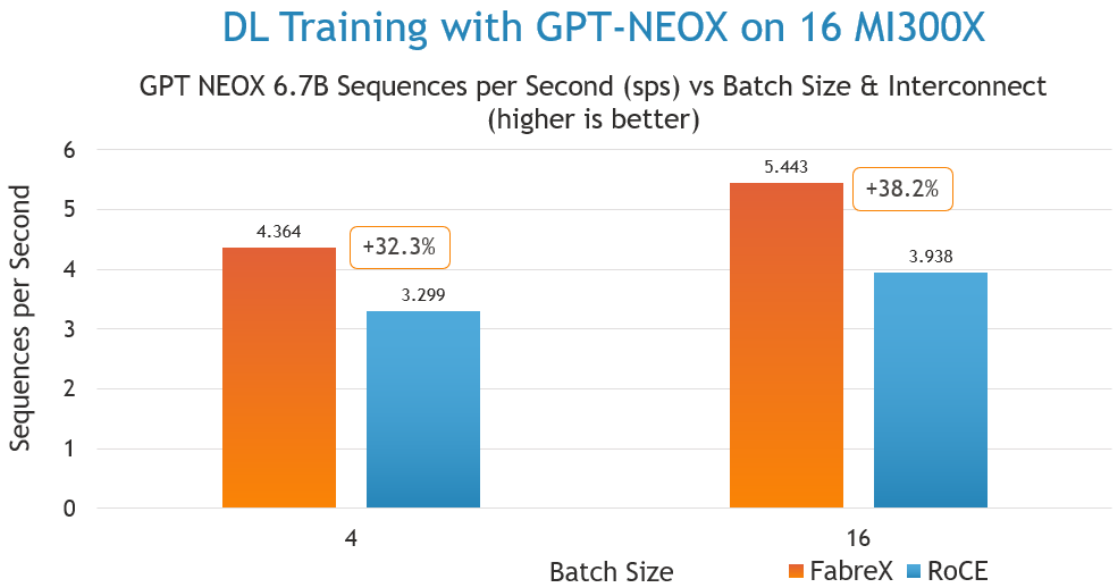
Test Configuration and Methodology

The extended testing utilized a configuration of 16 AMD MI300X accelerators, representing a substantial computational cluster common in production AI environments. Similar to our previous testing methodology, we maintained consistent hardware environments across test cases, varying only the interconnect technology to isolate its specific impact on performance metrics.



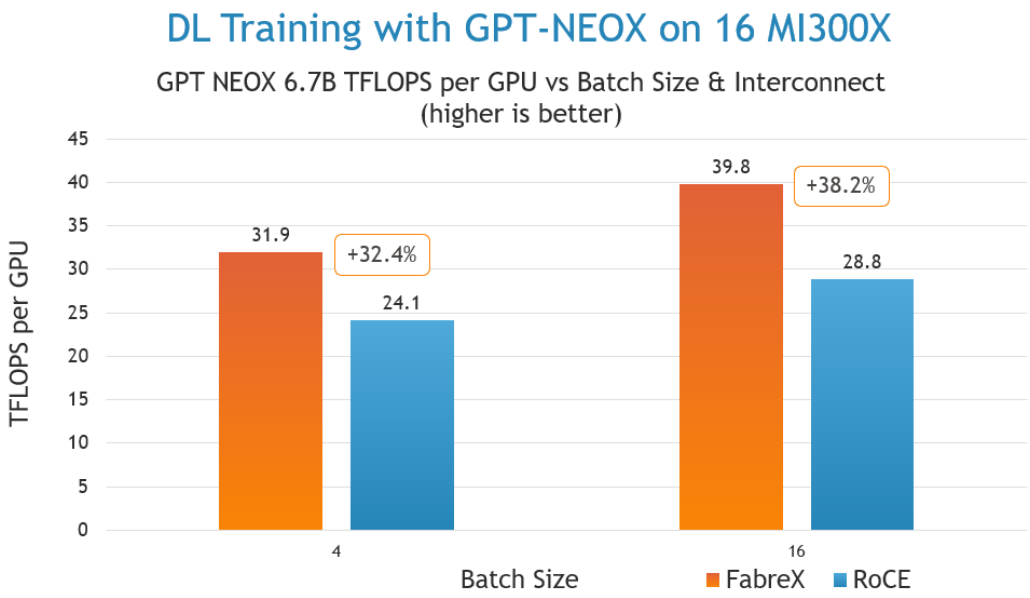
AMD MI300X Test Setup

The test system was configured with AMD MI300X accelerators connected via XGMI (AMD's high-speed GPU interconnect) within node boundaries, while inter-node communication occurred through either GigaIO's fabric or RoCE Ethernet. This configuration allows evaluation of how efficiently each interconnect technology can scale beyond the boundaries of individual compute nodes, a critical factor for large-scale AI deployments.



GPT NEOX 6.7B Sequences per Second (sps) vs Batch Size & Interconnect

At a representative batch size of 16, GigaIO delivered roughly 5.44 sequences per second versus about 3.94 sps over RoCE – an approximately 38% improvement in raw throughput. This performance gap persisted at both small and large batch settings, indicating that GigaIO provides superior communication efficiency regardless of workload sizing. Such uniform gains across batch sizes reinforce earlier observations: the low-latency, high-bandwidth GigaIO fabric keeps GPUs better fed with data and synchronized, whereas the Ethernet-based approach incurs communication overhead that hampers overall throughput in multi-node AI training.

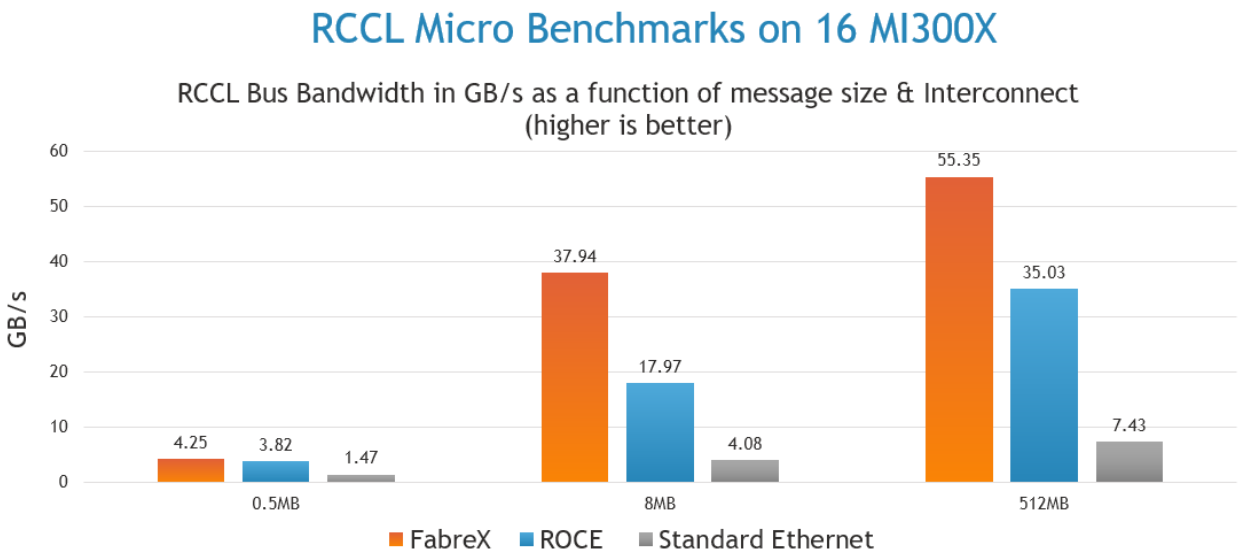


GPT NEOX 6.7B TFLOPS PER GPU VS BATCH SIZE & INTERCONNECT

In addition to higher throughput, the MI300X cluster tests highlighted better GPU utilization under GigaIO. Each MI300X accelerator sustained a higher fraction of its peak compute capability when connected via GigaIO, as evidenced by greater achieved TFLOPS per GPU during training. In one scenario with a ~6.7B parameter GPT-NeoX model, the GigaIO configuration drove each GPU to ~40 TFLOPS of sustained throughput, whereas the RoCE Ethernet run plateaued at a lower per-GPU TFLOPS level of ~29 TFLOPS. This suggests that the superior latency and bandwidth characteristics of GigaIO translate into more effective scaling of compute power – the GPUs spend more time doing useful work and less time waiting on data transfers. In practical terms, even at the scale of 16 GPUs, GigaIO's PCIe-native fabric maintained its performance edge, enabling faster model training convergence compared to the Ethernet alternative.

RCCL Collective Communication Bandwidth Results

We also conducted low-level communication benchmarks using RCCL (AMD's collective communication library, analogous to NCCL) to measure the raw data exchange capabilities of the two interconnects. The results were strongly in favor of GigaIO. In all test cases, GigaIO provided significantly higher inter-GPU bandwidth for collective operations (such as all-reduce) compared to RoCE Ethernet.



RCCL Micro Benchmarks Across 16 MI300X GPUs

Measured bus bandwidth on GigaIO reached into the tens of GB/s, approaching the theoretical limits of the PCIe Gen4 fabric, whereas RoCE Ethernet achieved substantially lower throughput. Notably, a standard non-RDMA Ethernet configuration delivered very weak performance – essentially an order of magnitude lower bandwidth than GigaIO, underscoring that conventional Ethernet networking (without RDMA) is ill-suited for GPU communication. Even with RDMA enabled, the RoCE Ethernet fell far short of GigaIO's communication efficiency. **These microbenchmark outcomes illustrate the fundamental advantage of a PCIe-native fabric: it moves data between GPUs with minimal overhead, whereas even optimized Ethernet-based approaches introduce latency and protocol costs that severely limit effective bandwidth**. This discrepancy at the communication layer directly correlates with the observed training throughput differences, as distributed deep learning workloads heavily depend on fast gradient synchronization. In short, the RCCL tests confirm that GigaIO's networking technology can greatly outperform RoCE in raw data exchange performance, providing a stronger foundation for scaling AI workloads across many nodes.

Deployment and Time-to-Results Considerations

Beyond pure performance metrics, the 16×MI300X study revealed a clear difference in ease of deployment and optimization between the two interconnect solutions. The GigaIO cluster was straightforward to set up – the GPUs across nodes were seamlessly discovered and operated as if on a single unified PCIe backbone, requiring minimal manual tuning to reach optimal performance. In contrast, achieving even the lower performance attained by RoCE Ethernet demanded careful configuration of RDMA networking, driver optimizations, and troubleshooting of network bottlenecks. Despite this effort, RoCE still could not match GigaIO's throughput.

This highlights that GigaIO not only delivers superior speed, but also does so with far less complexity, allowing teams to bring a multi-GPU system online and fully optimized in a shorter time frame. As Greg Diamos, Co-founder and CTO of Lamini, puts it: “GigaIO means less time messing with infrastructure and faster time to running and optimizing LLMs.” This real-world validation underscores how GigaIO enables AI teams to focus on model development and performance tuning rather than infrastructure firefighting.

The ease of deployment and reduced need for low-level performance tuning translate into faster time-to-product and time-to-revenue for organizations using GigaIO. In practical terms, an AI infrastructure built on GigaIO can start delivering high training and fine-tuning performance out-of-the-box, whereas an Ethernet-based solution might undergo extended tuning cycles and still operate at a performance deficit.

Overall, the new MI300X multi-node results strongly reinforce GigaIO’s value proposition. The data from this 16-GPU study confirms that GigaIO’s advantages are not tied to a specific workload or hardware generation but rather are fundamental improvements in distributed AI performance. Even with cutting-edge MI300X GPUs, the GigaIO fabric enabled higher throughput, better scaling efficiency, and more robust communication bandwidth than the best-effort RoCE Ethernet setup. Equally important, these gains were achieved with less complexity, underscoring a shorter path from deployment to productive use. GigaIO’s AI fabric consistently provides a superior solution compared to RoCE Ethernet. The 16 GPU MI300X benchmarks serve as compelling evidence that the right interconnect choice — in this case, GigaIO — can dramatically accelerate AI workloads.

Implications for AI Infrastructure and Conclusion

The Primary Constraint in AI is Power at the Facility

As AI workloads continue to demand ever-increasing computational power, many facilities face a hard ceiling not in rack space or compute density, but in available power and cooling capacity. In this environment, GigaIO’s FabreX offers a significant strategic advantage by enabling organizations to **achieve target performance levels with fewer GPUs**. Thanks to superior interconnect efficiency and lower communication overhead, each GPU in a FabreX-enabled system performs more useful work, reducing the total number of accelerators required for a given workload. Fewer GPUs mean lower total power consumption, fewer cooling challenges, and reduced strain on infrastructure. For data centers facing power budget constraints, GigaIO provides a powerful solution: **maximize AI output per watt**, scale effectively within fixed energy limits, and extend the operational lifespan of existing facilities.

Ease of Deployment

Beyond raw performance metrics, our benchmark study revealed that GigaIO’s FabreX was easier and quicker to deploy and optimize compared to RoCE Ethernet. This deployment

advantage translates to faster time-to-product and potentially earlier revenue generation for organizations implementing AI infrastructure.

The deployment efficiency stems from FabreX's more integrated approach to system-wide memory access, which reduces the complexity of configuration and optimization compared to traditional networking approaches. For organizations with limited specialized networking expertise, this reduced complexity can accelerate infrastructure deployment and minimize debugging time for performance issues.

Cost-Performance Considerations

While our analysis primarily focused on performance, the results strongly suggest compelling opportunities for cost-performance optimization. GigaIO's FabreX delivers superior or comparable performance to RoCE Ethernet. First, FabreX's architecture eliminates the need for networking hardware typically required in RoCE-based systems, such as high-speed Ethernet switches, NICs, and cabling. With FabreX, communication occurs over a native PCIe fabric, significantly reducing the number of devices and layers involved in data movement. This architectural simplicity not only lowers capital expenditure (CapEx) but also reduces power and cooling requirements, driving down operating costs (OpEx).

Second, because FabreX enables higher performance per GPU, organizations may achieve their AI workload goals using **fewer GPUs** and **fewer servers**. For example, in inference workloads, FabreX demonstrated up to 35% higher throughput than RoCE. That means **a FabreX-based system can serve the same number of users with 30–40% less hardware**, directly translating to lower total cost of ownership.

Lastly, FabreX's ease of deployment and optimization reduces the engineering time and effort required to bring a system online and achieve peak performance. This shorter path to full utilization cuts time-to-value and allows organizations to see returns on their infrastructure investment sooner.

In summary, FabreX's cost advantages stem not from being a lower-spec alternative, but from being a **more efficient, more performant architecture that does more with less**—less hardware, less power, and less complexity.

These benchmark studies provide compelling evidence that interconnect technology significantly impacts AI workload performance. GigaIO's FabreX demonstrated consistent advantages over RoCE Ethernet across diverse scenarios. The performance gap was particularly pronounced for small batch training, inference under load, and large model deployment requiring model parallelism—all increasingly common in modern AI workflows.

These findings indicate that interconnect technology warrants careful consideration for organizations building or expanding their AI infrastructure instead of simply relying on familiar

networking methods. The right choice depends on specific workload characteristics, but GigaIO's FabreX is particularly beneficial for advanced AI applications.

As AI models continue growing in size and complexity, the efficiency of interconnect technology becomes increasingly crucial for maintaining practical performance. This benchmark study provides a foundation for understanding these performance implications while suggesting valuable directions for future optimization of AI infrastructure deployments.

v1.04 04242025

About GigaIO

GigaIO redefines scalable AI infrastructure, seamlessly bridging from edge to core with a dynamic, open platform built for every accelerator. Reduce power draw with GigaIO's SuperNODE, the world's most powerful and energy-efficient scale-up AI computing platform. Run AI jobs anywhere with Gryf, the world's first suitcase-sized AI supercomputer that brings datacenter-class computing power directly to the edge. Both are easy to deploy and manage, utilizing GigaIO's patented AI fabric that provides ultra-low latency and direct memory-to-memory communication between GPUs for near-perfect scaling for AI workloads.

GigaIO's open architecture supports NVIDIA and non-NVIDIA GPUs and inference cards in its comprehensive ecosystem of accelerator manufacturers, AI/HPC software providers, and storage vendors. GigaIO's solutions provide superior performance for LLM inference, RAG, and fine-tuning, as well as for engineering and scientific workloads.

Visit www.gigaio.com or follow on Twitter (X) and LinkedIn.