Smarter Interconnects for Power-Constrained AI

GigaIO performed a series of **AI training, fine-tuning, and inference benchmarks** highlighting the transformative impact of interconnect technologies on AI infrastructure. These benchmarks demonstrate the performance, cost, and power efficiency of GigaIO's AI fabric compared to RDMA over Converged Ethernet (RoCE). In all of our testing, we varied only the interconnect so that the differences the interconnect makes could be explored. We used the same GPUs, servers, operating systems, and application software and varied ONLY the interconnect.

Practical Implications

Power Efficiency

GigaIO's AI fabric enables organizations to achieve target performance with fewer GPUs and lower power consumption. Its PCIe-native design eliminates the need for additional networking hardware like NICs and Ethernet switches, further reducing energy use.

Cost Savings

The efficiency of GigaIO's AI fabric translates into lower total cost of ownership:

- Fewer GPUs and servers are required for equivalent performance.
- Reduced infrastructure costs due to simpler hardware configurations.

Ease of Deployment

GigaIO's AI fabric simplifies system setup with seamless GPU discovery and minimal tuning requirements. In contrast, RoCE demands extensive configuration and troubleshooting to achieve suboptimal performance.

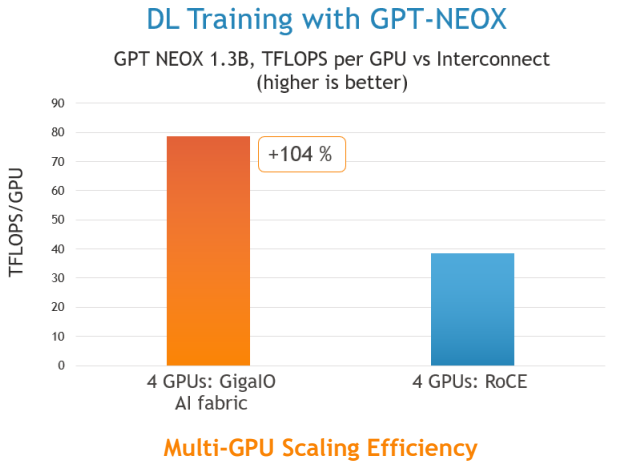
Why You Should Care

For organizations facing power constraints or seeking to optimize AI infrastructure costs, GigaIO's AI fabric offers a compelling alternative to traditional Ethernet-based interconnects. It enables faster time-to-value and more scalable AI deployments by delivering superior performance while consuming less power. As AI workloads grow increasingly demanding, choosing the right interconnect technology is no longer optional — it's critical for staying competitive.

Benchmark Results

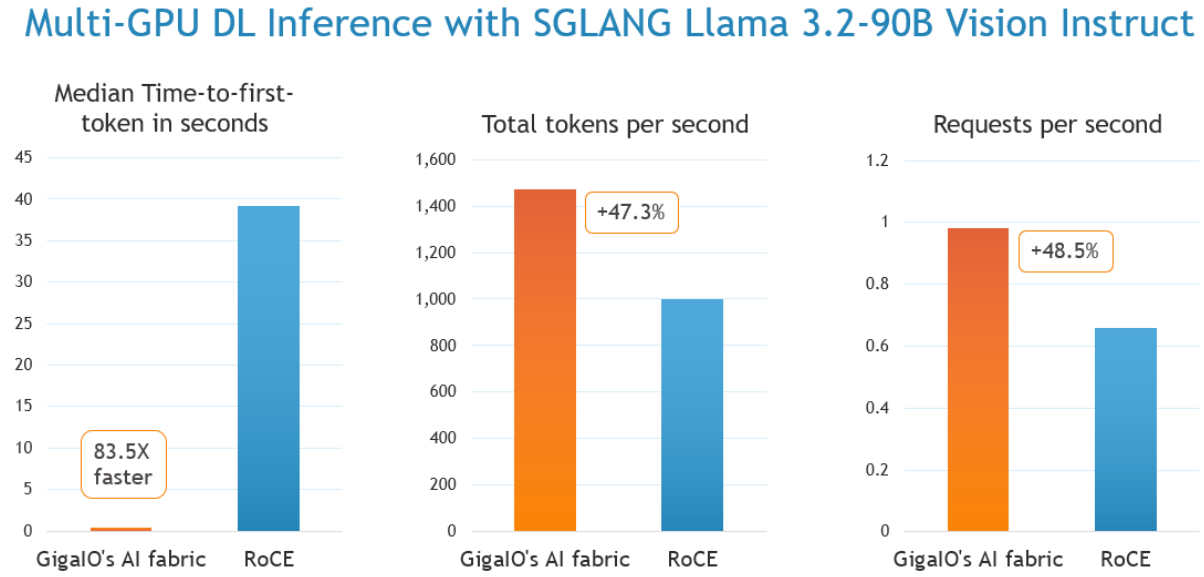
Training and Fine-tuning

GigaIO's AI fabric achieves better GPU utilization in multi-GPU setups, with **104% higher throughput** in distributed training scenarios compared with RoCE.



Inference

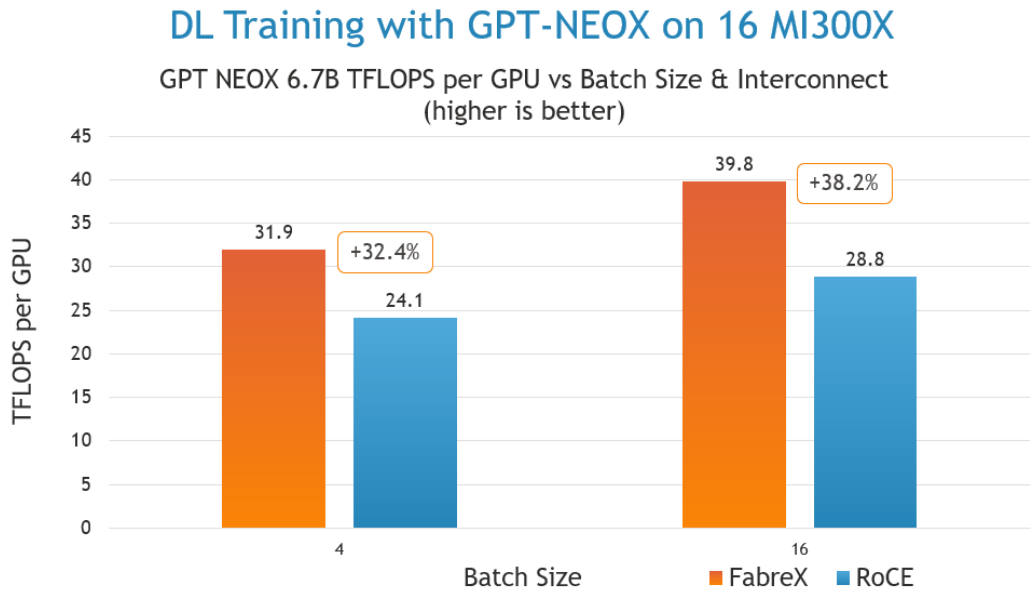
- **Time-to-First Token (TTFT):** GigaIO's AI fabric reduces TTFT by **83.5 times**, significantly improving responsiveness for interactive AI applications like chatbots.
- For the large model Llama 3.2-90B Vision Instruct, GigaIO's AI fabric achieves **47.3% higher throughput** and handles the same user load with **30-40% less hardware** than RoCE.



Multi-GPU Inference Performance

Distributed Training

In a 16-GPU AMD MI300X cluster, GigaIO's AI fabric delivered **38% higher training throughput** and superior GPU utilization, enabling faster convergence on large-scale models.



GPT NEOX 6.7B TFLOPS per GPU vs Batch Size & Interconnect