GIGAIO

# Empowering every accelerator to lead the AI revolution

# SCALE-UP AI PLATFORMS WITH INNOVATIVE MEMORY FABRIC TECHNOLOGY

GigaIO's two primary platforms

## SuperNODE

The world's most powerful and energy efficient scale-up AI inference platform

## Gryf

World's first carry-on suitcase-sized AI supercomputer bringing datacenter-class computing power directly to the edge

# WHY SCALE-UP MATTERS FOR AI INFERENCE

## AI inference at scale is bottlenecked not by compute, but by data movement latency

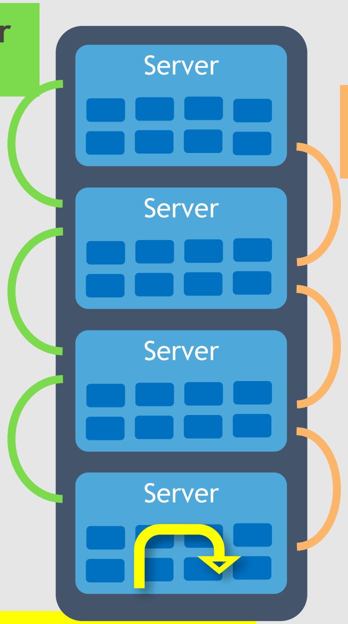
- Inference is about latency = user experience, response time
- When models are distributed across servers using traditional scale-out networks, **latency increases and GPUs sit idle waiting for data**
- **Wasting GPU cycles** due to inter-server communication overhead
- **Higher TCO** from multiple OS instances, software licenses, and complex setup and management

# COMMUNICATION ALTERNATIVES FOR INFERENCE

## Conventional Networking of Servers

**Server-to-Server Interconnect**

Ethernet
IB



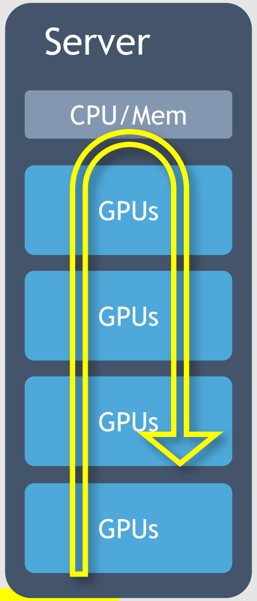
**GPU-to-GPU Interconnect**

Ethernet/IB
NVLink/Switch

**Intra-Server Interconnect**

PCIe
NVLink
Infinity Fabric

## SuperNODE



**Intra-Server Interconnect**

PCIe
NVLink
Infinity Fabric

# GigaIO's **competitive advantage** lies in its ability to achieve higher performance and power efficiency at lower price points

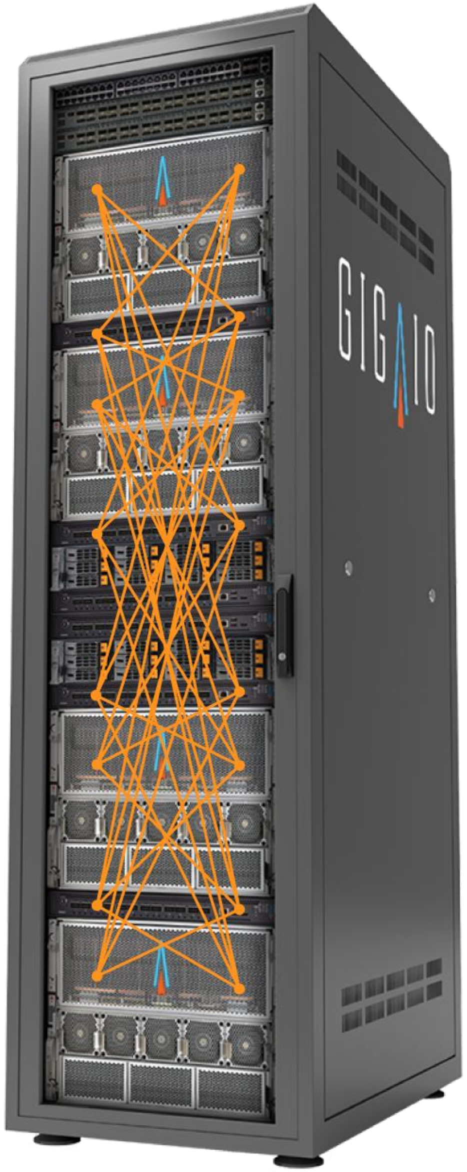
| | FOUR 8–GPU SERVERS<br>WITH ETHERNET | NVIDIA<br>NVL72 | GIGAIO<br>SuperNODE |
|---|---|---|---|
| ǦẩYảẮ ẶḥȉẩŴȉẩŴẫẊḻȉẫYŴ | ✓ | ✗ | ✓ |
| ĒȉĉYảḶẽ | 10,000ns | 9,000ns | 330ns |
| ẠȉẩŴẻ ẩŴẫ | 400Gb / 800Gb | 900Gb | 512Gb |
| ḤẽãĉYả ẶẬāãċ | Moderate | Very Expensive | Moderate |
| Ḡāè YẩẶẬYẩểẫYả Yảċ | 43,800W | 130,000W | 32,870W |

# GigaIO SuperNODE™

## Delivers true single-server scale-up performance and efficiency

- Connects dozens of accelerators to a single node server – homogenous or heterogenous
- Accelerator agnostic: Brand - NVIDIA, AMD, Inference ASIC (d-Matrix, Tenstorrent…..); Type - GPU, FPGA, ASIC
- Form factor agnostic: OAM, SXM, PCIe
- No network and inter-server overhead, lowest latency in the industry – 330ns end to end
- Increases utilization and performance, while decreasing cost and power consumption
- Simplest AI infrastructure — one OS instance, one driver stack with full container, PyTorch and TensorFlow support

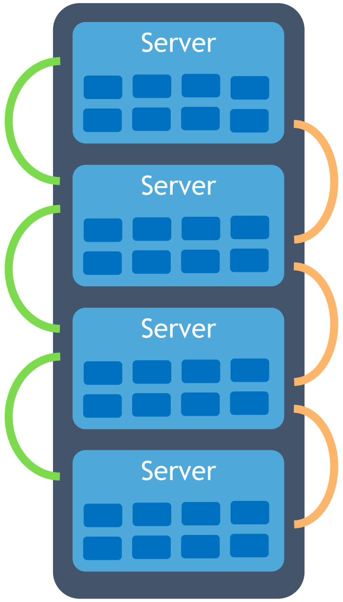
*"That's ONE server"*



FabreX AI Fabric

# INFERENCE AND FINE-TUNING TESTS

**Conventional Scale-Out**
**4 servers connected with Ethernet, dedicated rail connections for GPUs**

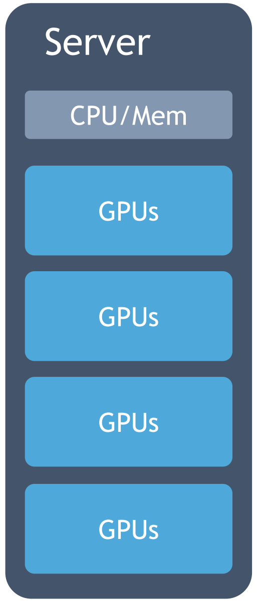


*SMC servers; AMD MI300 with Infinity Fabric;*
*400Gb Ethernet RoCE*

Same GPUs
Same # of GPUs
Same processor
Same memory
Same storage
Same OS
Same frameworks
Same application

Different interconnect
Different NW topology

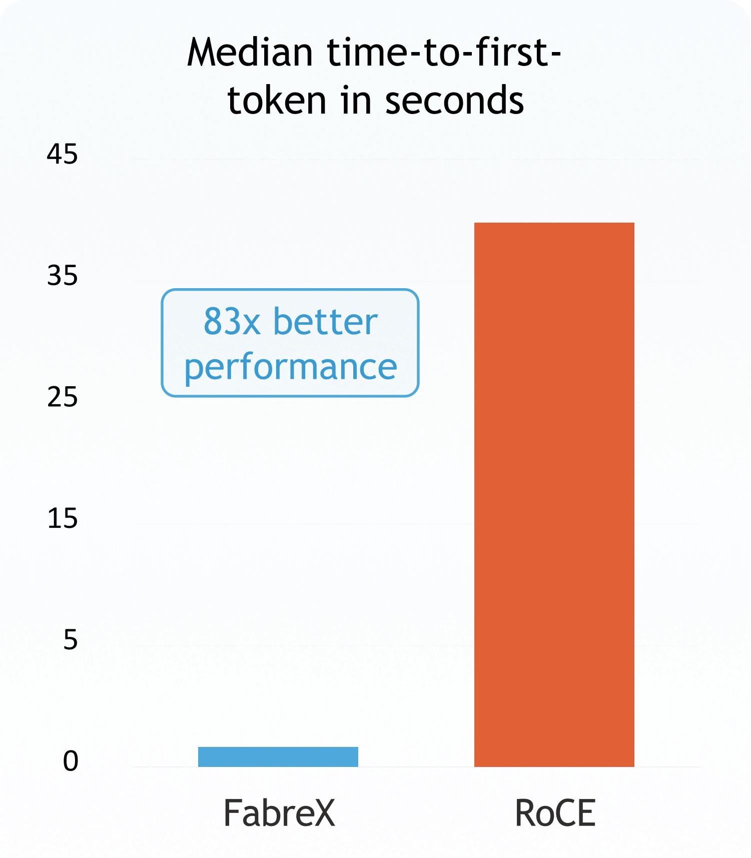
**FabreX Scale-Up**
**32 GPUs connected with rail optimized PCIe fabric**



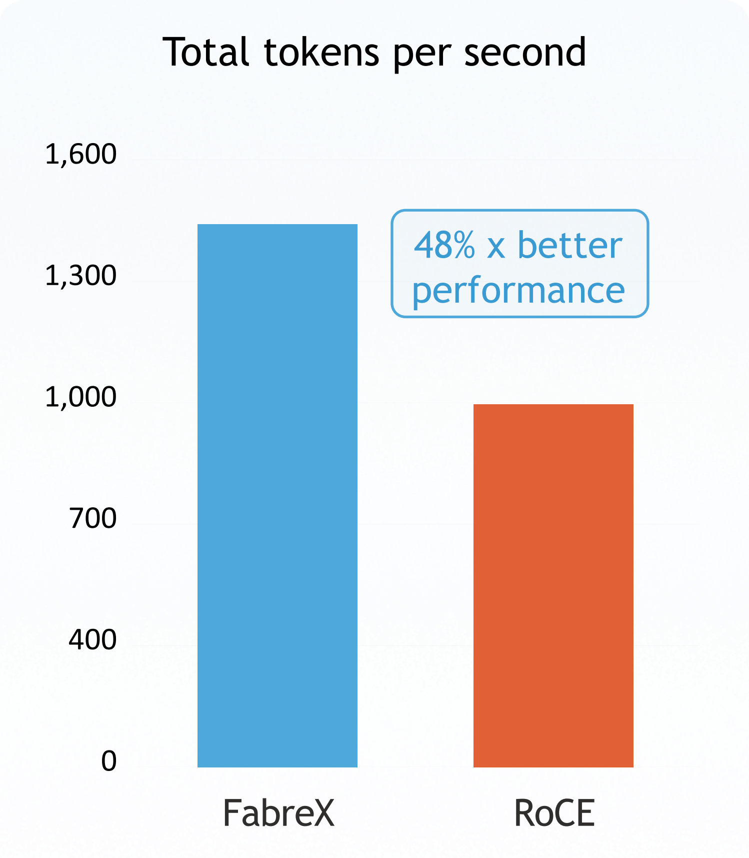
*256Gb PCIe Gen4; incorporating Infinity Fabric*

# AI INFERENCE PERFORMANCE BENCHMARKS

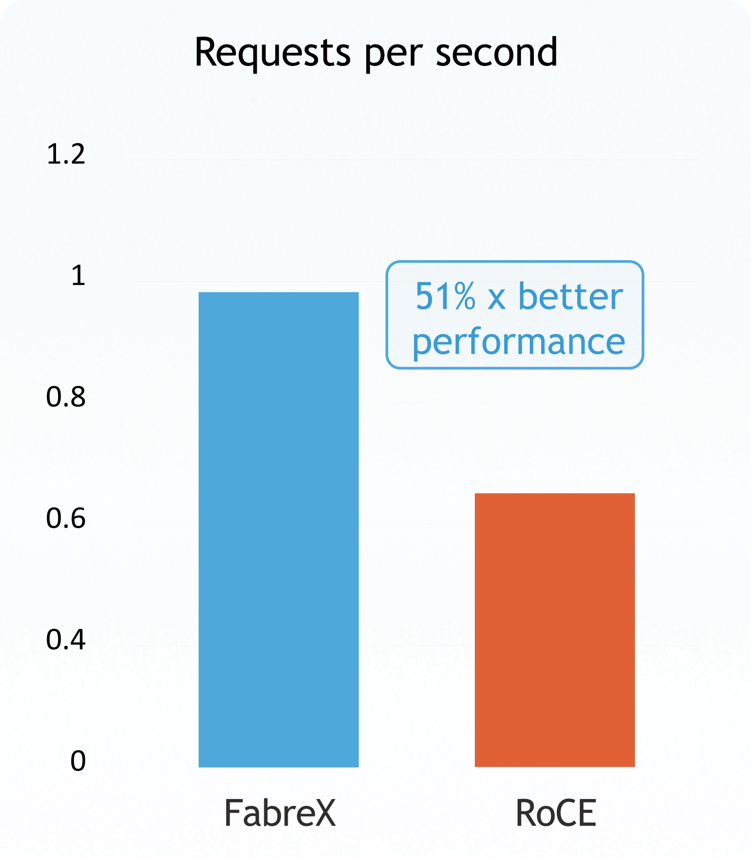
## Multi-GPU DL Inference with SGLANG Llama 3.2-90B Vision Instruct (large model)



Smaller is better



Larger is better

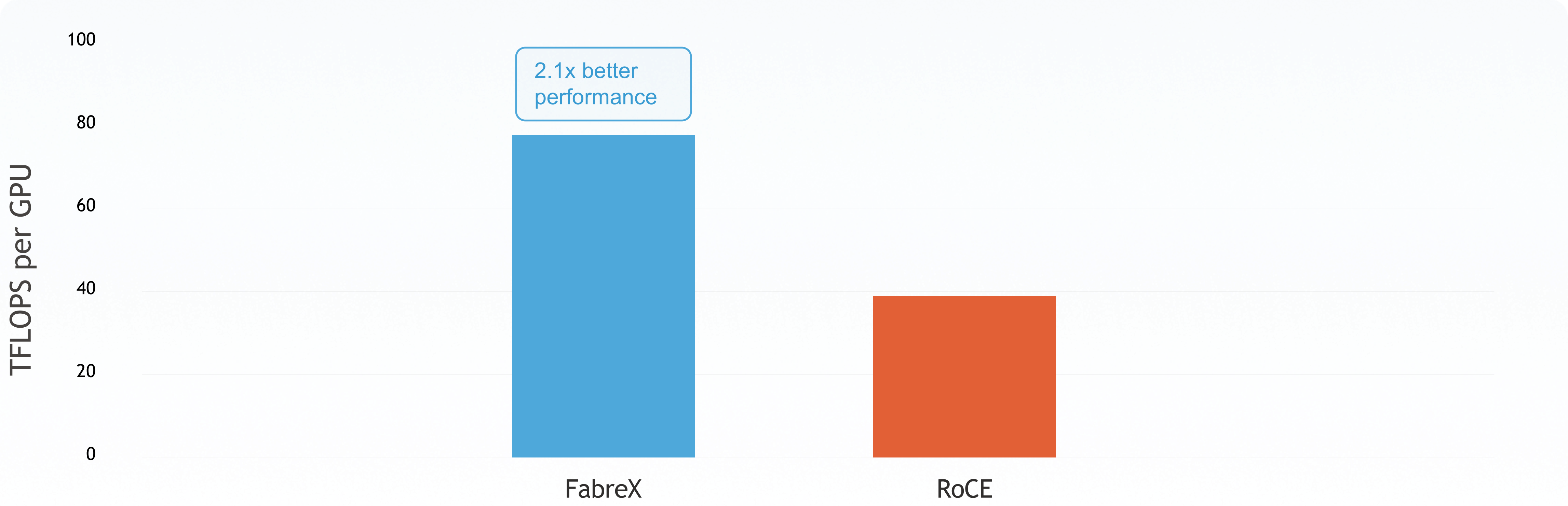


Larger is better

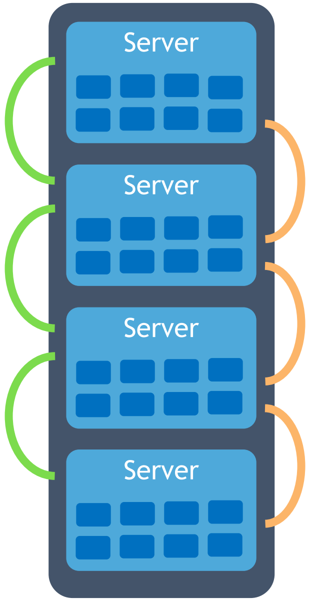
# AI FINE TUNING PERFORMANCE BENCHMARKS

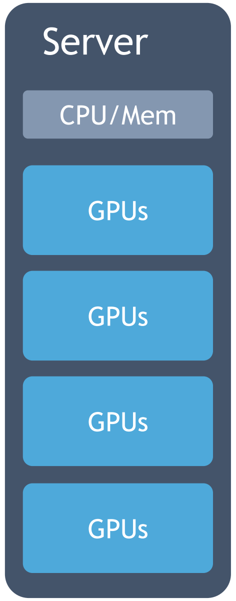
## Fine Tuning with GPT-NEOX

GPT NEOX 1.3B, TFLOPS per GPU vs #GPUs and Interconnect – Higher is Better



# TOKENOMICS INFERENCE TESTS





## Conventional Scale-Out

- Tokens/watt = 0.41
- Tokens/dollar = 0.0077

## FabreX Scale-Up

- Tokens/watt: 0.74 +80%
- Tokens/dollar: 0.0114 +50%

GIGAIO

# Empowering every accelerator to lead the AI revolution