MLPerf Inference Study with SuperNODE

TECH BRIEF

The GigaIO SuperNODE™ is the only single-node server with 32 accelerators and up to 1 petabyte of storage. The power of all these accelerators seamlessly connected by GigaIO's AI memory fabric, FabreX™, can now be harnessed to drastically speed up time to results. SuperNODE can connect up to 32 AMD or NVIDIA accelerators. Workloads run faster because they utilize the power of each accelerator as if it was located in a single rack-scale server rather than in multiple constrained server nodes connected by a congested, complex network. All resources communicate over a single dynamic memory fabric.

Prior to SuperNODE, the only way to harness 32 GPUs would have required four servers with eight GPUs apiece, or 8x 4-GPU servers, connected using messaging protocols. While FabreX can also natively operate with the same messaging protocols, SuperNODE introduces a fourth option with all GPUs directly benefiting from native PCIe performance when communicating with either the processors or with other GPUs.

MLPERF INFERENCE

Inference is the key challenge in deploying AI at scale. The economic viability of AI hinges on TCO-efficient inference. Additionally, low latency and a strong user experience are critical for successfully integrating AI into business applications.

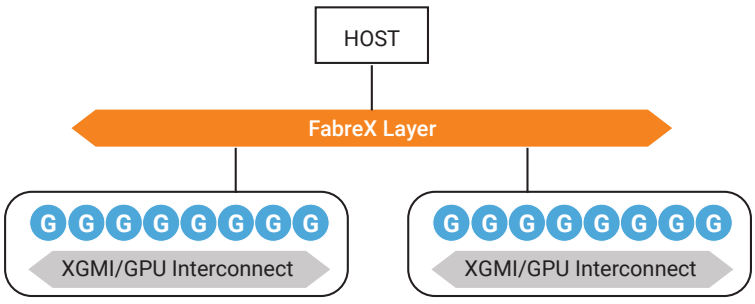
The MLPerf Inference Benchmark, developed by the ML Commons organization, is a widely respected suite designed to assess the performance of Machine Learning (ML) models in real-world scenarios. MLPerf aims to offer standardized, fair, and transparent evaluations to help hardware and software developers, researchers, and organizations understand the performance of ML models across different platforms.

Llama 2-70B is employed to test high-performance hardware for large-scale generative AI tasks, evaluating various hardware configurations' capacity to manage complex inference workloads. The offline version of the benchmark focuses on measuring the number of tokens generated per second.

RUNNING MLPERF

For this test, we used a SuperNODE with two MI300X systems, totaling 16 GPUs, which can be attached to a single host. The process of connecting accelerators to a host system is commonly known as composing the system. The specific topology used in this test is outlined below:

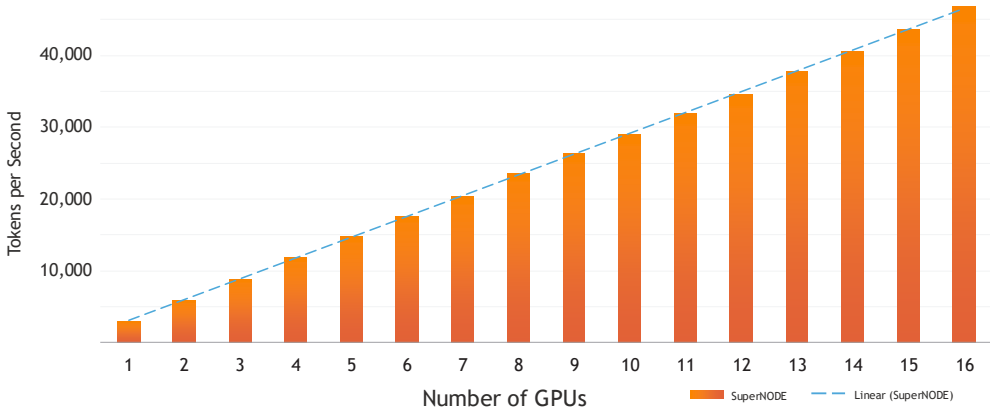
In our composed SuperNODE setup, we achieved slightly higher single-node performance compared to a traditional bare-metal configuration, where the OS and accelerators run in the same chassis. This result suggests that system composition does not degrade performance; on the contrary, it may offer a minor performance boost (see the "Context to Competition" graph to the below).



Our test results demonstrate near-linear scaling across all 16 GPUs, facilitated by the low-latency, high throughput FabreX interconnect. This performance scaling enables our customers to deploy AI workloads at scale with optimized TCO.

MLPERF INFERENCE: DATACENTER BENCHMARK

Highest Single-node Token Per Sec Recorded To Date



Scaling Llama2-70b Inference on 16 GPU SuperNODE

Unverified benchmark results not officially submitted to or verified by MLCommons Association.

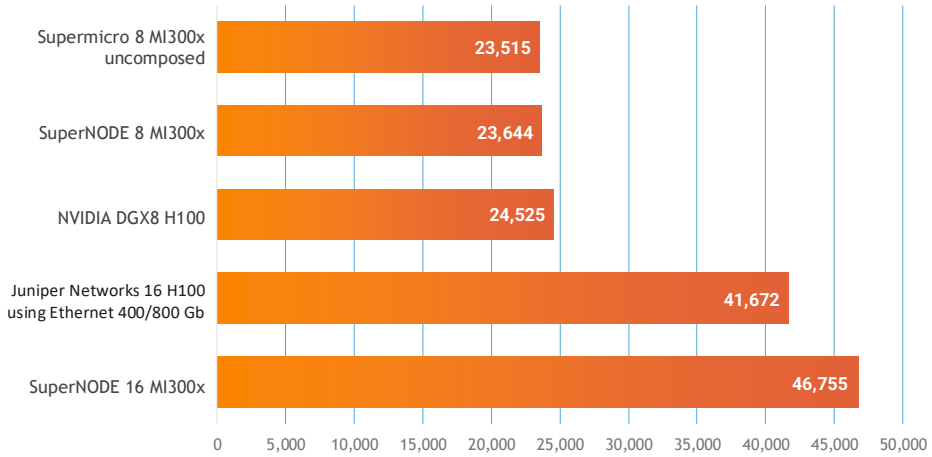
TEST RESULTS IN CONTEXT TO COMPETITION

As previously noted, we achieve near-linear scaling, an already impressive outcome that becomes even more remarkable when compared to our competition. There are limited publicly available results for the Llama2-70B benchmark with 16 GPUs in the latest MLPerf Inference, making direct comparisons scarce.

When we compare our 8x MI300X setup to NVIDIA's DGX system with 8x H100, we observe similar performance. However, once we scale beyond 8 GPUs and the competition requires internode networks, the advantages of our SuperNODE powered by FabreX become clear. We can generate 12.2% more tokens per second than our competitors using 16 GPUs. A closer look at the scaling of SuperNODE vs its competition shows a scaling advantage of 16.4%

This performance edge, coupled with the advantage of using the same software stack as a single node, allows for efficient management of 16 GPUs. SuperNODE with FabreX is a high performance and low TCO solution.

TOKENS PER SECOND FOR DIFFERENT SYSTEMS



Performance in context to competition, Juniper and NVIDIA numbers from MLPerf Inference Data Center v4.1 llama2-70b-99 offline