GIGAIO

FOR IMMEDIATE RELEASE

GigaIO Introduces the First Ever 32 GPU Single-Node Supercomputer for Generative AI and Accelerated Computing

GigaIO SuperNODE^{TM} with FabreX^{TM} dynamic memory fabric delivers unprecedented compute capabilities with AMD Instinct^{TM} accelerators for low-power, accelerated computing.

Carlsbad, California, July 13, 2023 – GigaIO, the leading provider of workload-defined infrastructure for AI and technical computing, recently announced that it successfully configured 32 AMD Instinct MI210 accelerators to a single-node server utilizing the company's transformative FabreX ultra-low latency PCIe memory fabric. Available today, the 32-GPU engineered solution, called SuperNODE, offers a simplified system capable of scaling multiple accelerator technologies such as GPUs and FPGAs without the latency, cost, and power overhead required for multi-CPU systems.

As large language model applications demand even more GPU performance, technologies that reduce the number of required node-to-accelerator data communications are crucial to providing necessary compute power at improved infrastructure TCO.

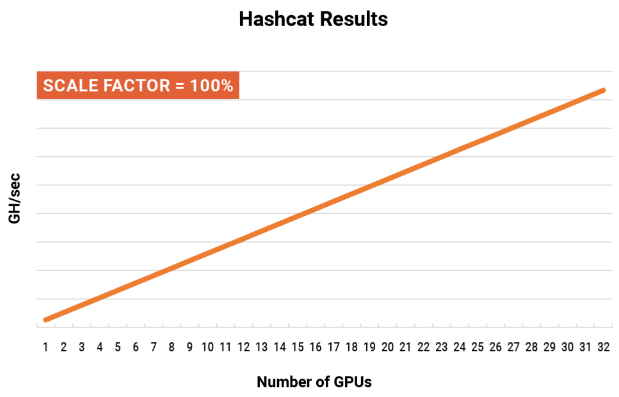
"As AI workloads become more broadly adopted, systems that offer the ability to harness the compute power of multiple GPUs and better manage data saturation at ultra-low latency are essential," said Mark Nossokoff, Research Director, Hyperion Research. "And as large language model applications drive demand for more GPU performance, technologies that work to minimize node-to-accelerator traffic are better positioned to provide the necessary performance for a robust AI infrastructure."

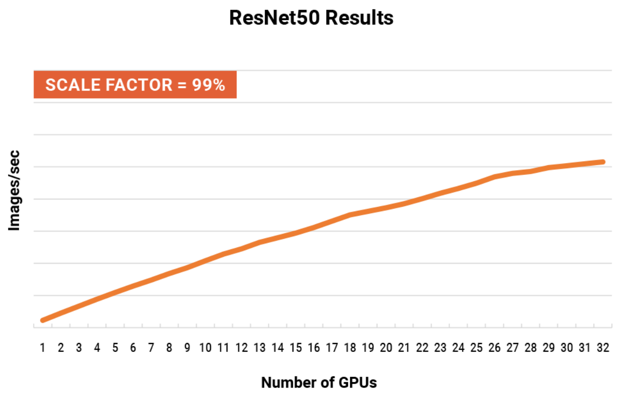
"AMD collaborates with startup innovators like GigaIO in order to bring unique solutions to the evolving workload demands of AI and HPC," said Andrew Dieckmann, corporate vice president and general manager, Data Center and Accelerated Processing, AMD. "The SuperNODE system created by GigaIO and powered by AMD Instinct accelerators offers compelling TCO for both traditional HPC and generative AI workloads."

GigaIO's SuperNODE system was tested with 32 AMD Instinct MI210 accelerators on a Supermicro 1U server powered by dual 3rd Gen AMD EPYC^{TM} processors.

- *Hashcat:* Workloads that utilize GPUs independently, such as HashCat, scale perfectly linearly all the way to the 32 GPUs tested.
- *Resnet50:* For workloads that utilize GPU Direct RDMA or peer-to-peer, such as Resnet50, the scale factor is slightly reduced as the GPU count rises. There is a one percent degradation per GPU, and at 32 GPUs, the overall scale factor is 70 percent.

These results demonstrate significantly improved scalability compared to the legacy alternative of scaling the number of GPUs using MPI to communicate between multiple nodes. When testing a multi-node model, GPU scalability is reduced to 50 percent or less. The following charts show two real-world examples of these two use cases:





More testing results can be found here.

“This testing shows the enormous value of using GigaIO’s SuperNODE to get all the benefits of composability, without any of the hassles,” said Alan Benjamin, CEO & President, GigaIO. AMD and GigaIO engineered the entire hardware and software stack of the SuperNODE up to and including the TensorFlow and PyTorch libraries so that applications “just run” without any software changes. “Customers can scale GPU performance without the overhead of multiple servers using our FabreX software, and get unprecedented flexibility. When a large job needs results fast, 32 GPUs can be deployed on a single compute node simply and efficiently, with leadership low latency and power usage. Those same accelerators can then be easily and quickly reallocated to other servers, thus optimizing their utilization. Let the job define your system, and not the other way around,” added Benjamin.

About GigaIO
GigaIO provides workload-defined infrastructure through its dynamic memory fabric, FabreX, which seamlessly composes rack-scale resources and integrates natively into industry-standard tools. FabreX lets customers build impossible servers for AI and technical computing — from storage to accelerators to memory — at a fraction of cloud TCO, by optimizing the utilization and efficiency of their existing hardware, allowing them to run more workloads faster at lower cost through higher utilization of resources and more agile deployment. Visit www.gigaio.com, or follow on Twitter and LinkedIn.

Contact
Danica Yatko
760-487-8395
danica@xandmarketing.com

AMD, the AMD Arrow logo, EPYC, AMD Instinct, and combinations thereof, are trademarks of Advanced Micro Devices, Inc.