FOR IMMEDIATE RELEASE

GigaIO Doubles GPU Performance at a 30% Cost Savings with Intel Sapphire Rapids

San Diego, California, January 19, 2023 – GigaIO, the leading provider of workload-defined infrastructure without compromise for HPC + AI workflows, today announced its certification of Intel Sapphire Rapids with a benchmark of 106% improvement in GPU utilization over equivalently configured NVIDIA DGX systems connected via InfiniBand. The system was composed from 16 NVIDIA A100 GPUs in a direct, peer-to-peer configuration along with a single Intel Sapphire Rapids server, more than doubling GPU utilization. Using energy-efficient accelerator pooling appliances in this configuration cost 30% less than putting 16 GPUs in dedicated servers.

"GPUs are what drive performance today, but GPUs are trapped in servers, limiting flexibility and throttling utilization to as low as 15%," said Alan Benjamin, CEO of GigaIO. "Different workloads need different GPU resources and configurations, and since GPUs can represent up to 80% of the total system cost, this is a significant waste of expensive resources. Underutilized GPUs needlessly waste energy and drive up operating costs. GigaIO's engineered solutions free GPUs from confining servers."

A complete Sapphire Rapids-based computing platform for accelerating HPC and AI workloads can be provisioned from GigaIO's GigaPod engineered solutions platform. Whether an application needs large numbers of GPUs on a single server, or runs on many servers using just a few GPUs, composability can provide the exact configuration required to meet dynamically changing mixed workload requirements, providing cloud-like agility without cloud costs.

GigaIO's composable technology provides the immediate benefit of doubling performance with the same equipment by eliminating stranded and underutilized resources, as well as the longer-term benefit of being able to easily upgrade or add individual servers, storage, and accelerators via plug-n-play at the component level. This architecture allows every primary subsystem to operate on its own upgrade cycle, providing for a system's total cost to be optimized over its entire lifecycle.

This doubling in GPU performance is due to GigaIO's Fabrex™, the industry's first universal dynamic memory fabric. FabreX disaggregates CPUs, GPUs, storage, memory, and other processing resources that are trapped in servers into a single, unified, rack-scale system connected by a high-speed, low-latency memory fabric. This sub-microsecond fabric offers 10x better latency than InfiniBand and 3x better latency than NVLink.

FabreX transforms all resources connected to fabric into unique memory resources, using the same PCI Express (PCIe) and CXL technology used within industry-standard servers to provide the lowest possible

latency and highest possible bandwidth performance. What makes FabreX unique is its use of Direct Memory Access (DMA) to transfer data between the memory of the various processing elements connected to the fabric, exactly as if the components were plugged directly into the server motherboard.

FabreX can be accessed using industry-standard cluster managers, workload scheduling managers, and DevOps scripting tools, due to GigaIO's native integrations with leading cluster scheduling and management software via Open Standard Redfish APIs. NVIDIA Bright Cluster Manager is bundled with every GigaIO system, enabling users to compose discrete computational elements into highly functional and reconfigurable systems that allow multiple workflows to co-exist on the same infrastructure.

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 HPC + AI workflows — 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