Test Results for GigaIO SuperNODE™

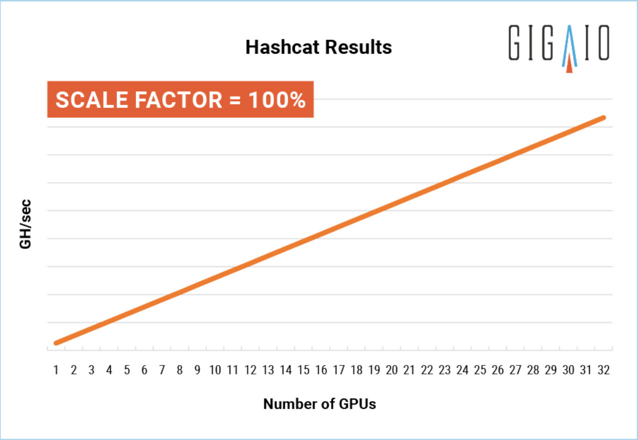
GigaIO's SuperNODE™ system was tested with 32 AMD Instinct™ MI210 GPUs on a 1U server with dual AMD EPYC™ "Milan" processors connected over GigaIO FabreX™.

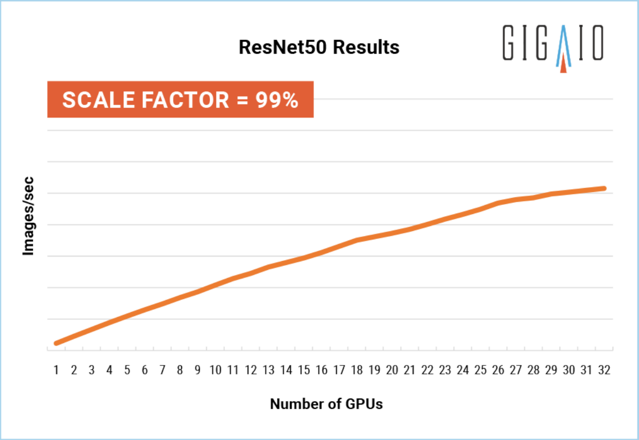
- *Hashcat:* Workloads that utilize GPUs independently, such as Hashcat, scale perfectly linearly all the way to the 32 GPUs tested.

- *ResNet50:* For workloads that utilize GPU Direct RDMA or peer-to-peer, such as, 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:





Democratizing Access to AI and HPC's Most Expensive Resources

The alternative off-the-shelf systems offering SuperNODE's accelerator-centric performance are impractical, if not prohibitive, for most organizations.

- *SuperNODE drastically reduces AI costs:* A 32-GPU deployment with a standard 4-GPU-to-server configuration would require a total of eight servers, at an average cost of $25,000 apiece ($175,000) — not including the cost of the GPUs. Eliminating additional per-node software licensing costs results in additional savings.

- *SuperNODE delivers significant savings on power consumption and rack space:* Eliminating seven servers saves approximately 7KW, with additional power savings in associated networking equipment — all while increasing system performance. Compared to 4-GPU servers, SuperNODE delivers a 30% reduction in physical rack space (23U vs. 32U).

- *SuperNODE keeps code simple:* An eight-server, 32-GPU system would require significant rewrites of higher order application code in order to scale data operations, further adding complexity and cost to deployment.

- *SuperNODE shortens time-to-results:* Eliminating the need to connect multiple servers via legacy networks using MPI protocol, and replacing them with native intra-server peer-to-peer capabilities, delivers significantly better performance.

- *SuperNODE provides the ultimate in flexibility:* When workloads only need a few GPUs, or when several users need accelerated computing, SuperNODE GPUs can easily be distributed across several servers.