GIGAIO Data Sheet

GigaIO Gen4 Accelerator Pooling Appliance

– Fully Managed Disaggregated Compute Accelerator Appliance for AI/DL, HPC, Visualization and Data Science Applications

HIGHLIGHTS

All slots PCIe Gen4
Up to 256Gb/sec bandwidth

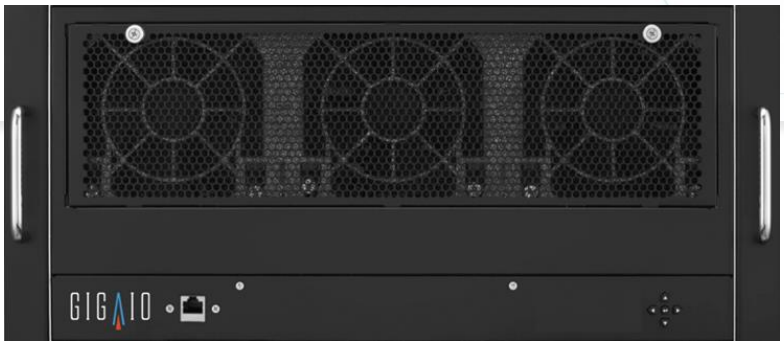
Compact
5U 19” rack-mount

Up to 10 PCIe boards
Including NVIDIA A-100

Independent Power Supply
Control on all slots

Full Redfish® API support
For device control

Outstanding in-box BMC
For serviceability and control



The GigaIO™ Gen4 Accelerator Pooling Appliance is the industry’s highest performing PCIe accelerator appliance fully supporting PCIe Gen4 with up to 1Tb/sec bandwidth into and out of the box.

It delivers advanced provisioning and monitoring of compute accelerators and next-gen NVMe assets, supporting up to 8 double-width PCIe Gen 4.0 x16 accelerator cards and 2 PCIe Gen 4.0 x16 low-profile slots.

This flexible expansion platform enables users to add any PCIe Gen 4.0 application accelerators, including GPUs (and the new NVIDIA A100), FPGAs, IPUs, thin-NVMe-servers and specialty AI chips.

Composable Disaggregated Infrastructure by dynamically provisioning and scaling PCIe devices

GigaIO FabreX fabric creates shared pools of PCIe devices in disaggregated GPU chassis connected by standard PCIe cables. Users can dynamically provision GPUs when needed and return them to the common pool when finished.

Intelligence for resource optimal utilization and failure prediction

GigaIO’s advanced Redfish® API interface provides complete centralized analytics such as GPU utilization and performance, as well as continuous monitoring for any faults for faster problem resolution and improved reliability.

Designed from the ground up for enterprise data center deployment

In addition to continuous monitoring and control of the accelerators, should a device fail, it is essential to provide fast, easy and intuitive serviceability. With service indicators internal and external, quick and sure identification of problem devices is assured. Coupled with independent card power control and redundant power supplies and fans, this appliance is the perfect answer for disaggregating essential accelerators but maintaining tight control and easy serviceability.

Specifications

Enclosure	5U; 219 (H) X 435 (W) X 466.3 (D)mm, ~8.6 (H) x 17.12 (W) x 18.35 (D) inches
Cables	Connect to the FabreX Adapter Card with Copper or Active Optical Cable
Cooling	6 redundant 12038 fan, hot-swap
Power	Two (2) 2000W PSU; Hot Swap 2+2 redundant (optional)
BMC/mCPU	Aspeed AST2500
Status Display	16 x2 matrix LCD Display status and key information
PCIe Device Slots	Eight (8) PCIe 4.0 x16 double width FHFL Two (2) PCIe 4.0 x16 Low-Profile slots (can be configured as host interfaces)
Host Interface – Standard Mode	Two (2) PCIe 4.0 x16 (Four mini-SAS HD connectors) Need HBA (host bus adapter) cards on server hosts Mini-SAS HD cables (SFF-8644)
Operating Environment	Operating Temperature: 0°C ~ 35°C (50°F ~ 95°F) Non-operating Temperature: -20°C to 70°C (-4°F to 158°F) Operating Relative Humidity: 10% to 90% (non-condensing) Non-operating Relative Humidity: 0% to 95% (non-condensing)
Agency Compliance	Agency Certifications (testing pending): - AS/NZS CISPR32, class A - EN 55032, class A - EN 55024 - EN 61000-3-2 - EN 61000-3-3 - FCC Part15B, class A - RoHS Directive 2011/65/EU

Software Features

Features	GPU composability Device peer-to-peer (GPU P2P) PCIe port configuration Performance and error monitoring
Management Features	Redfish®, RESTful API
Standard Features	Device peer to peer (GPU P2P) Performance and error monitoring
Advanced Features	GPU dynamic provisioning Device surprise add and remove PCIe port configuration
System Management	FabreX management center Real-time GPU cluster topology Dashboard for GPU utilization, performance and more Predictive health monitoring Role-based authentication and access control