GIGAIO

Engineered Solution Brief

GigaPod™

The Open Solution to Speed and Simplify AI and HPC Workloads

HIGHLIGHTS

BREATHTAKING PERFORMANCE
Lowest possible latency and highest effective bandwidth

COMPELLING ECONOMICS
Lower TCO, higher resource utilization, and less complexity, power, cooling, CAPEX, and OPEX

OUTSTANDING SIMPLICITY AND EASY MANAGEMENT
Turnkey, fully tested using robust FabreX topologies

EXTREME FLEXIBILITY
True disaggregation with dynamic composability

LEADERSHIP EFFICIENCY
Strips away conversion and overhead, eliminating bounce buffers with 100% PCI Express interconnect

NO SUPPLY CHAIN CONSTRAINTS
Immediate availability and no networking vendor dependencies

The GigaIO GigaPod is a flexible solution for heterogeneous compute designed to dynamically accommodate different types of workloads on a single hardware architecture.

GigaPod delivers the industry's lowest possible latency and highest effective bandwidth by integrating compute and GPU acceleration I/O into a single-system dynamic fabric, using standard PCI Express (PCIe) technology. It is a complete, enduring disaggregated infrastructure solution that facilitates dynamic workflows and is simple to deploy and manage.

By eliminating stranded and underutilized resources, the GigaPod family of solutions decreases the Total Cost of Ownership (TCO) of on-prem and hybrid infrastructure. GigaPod delivers the agility of the cloud by dynamically allocating the right resources at the right time for each workload, extending access across a broader set of users.

Workloads run faster, because they are utilizing resources as if they were in a single rack-scale server rather than in multiple constrained server nodes connected by a congested, complex network. All servers, GPUs, and accelerators communicate over one dynamic memory fabric.

A Dynamic Memory-Centric Fabric

At the heart of the GigaPod family is FabreX™, our universal dynamic memory fabric. With FabreX, you can effortlessly connect new memory, network, and storage products — along with a multitude of accelerators — to your choice of AMD processors, either directly attached, or via server configurations like NVMe-oF and GDR (GPU Direct RDMA).

FabreX disaggregates computing, GPUs, storage, and other resource I/O into pools connected by a single-system fabric, using industry-standard PCI Express (PCIe) technology. PCIe is used internally in every server on the market today. GigaIO

breaks the server barrier, enabling components previously internal to the box to be connected at similar latency and bandwidth, but outside the server enclosure.

In addition, FabreX enables server-to-server communication across PCIe and makes cluster-scale computers possible. Individual servers can now directly access the system memories of all other servers in the fabric, for the industry's first true in-memory network. This new architecture enables a unified, software-defined, and memory-centric composable infrastructure.

Breathtaking Performance

The GigaPod platform running FabreX delivers both the lowest possible latency AND the highest effective bandwidth. Latency from one server to another is less than 300ns for true PCIe performance across the entire cluster, scaling up to 128 Gb/s bandwidth.

Outstanding Simplicity and Easy Management

Accessing the entire power of a GigaPod is as simple as using the same cluster management and workload scheduling managers as you always have. In order to support workflow-defined infrastructure, all GigaPods include NVIDIA AI Enterprise Essentials to compose compute infrastructures that can be configured and scaled in minutes to handle the most demanding AI and HPC workloads. NVIDIA Base Command Manager provides complete dynamic infrastructure management, including schedulers and container management. All hardware and software are preconfigured, tested, and validated.

Extreme Flexibility Through Composability

The GigaPod platform can compose an unprecedented variety of resources, such as accelerators (including GPUs, DPUs, TPUs, FPGAs and SoCs), storage devices like NVMe or PCIe native storage, and other I/O resources connected to compute nodes. In addition to device-to-node, FabreX is unique in enabling node-to-node and device-to-device communication across the same high performance PCIe memory fabric. GigaPods span multiple servers, while GigaClusters span multiple racks, in order to scale up single-server systems and scale out multi-server systems, all unified via the FabreX dynamic memory fabric.

An Enduring Infrastructure Built for Change

Gone are the days of having to make massive and costly changes to large numbers of servers in order to incorporate new technology. Upgrade or add compute, GPUs, and application accelerators that plug-n-play with your environment at the component level. Every major subsystem can now operate on its own upgrade cycle. The total cost of the system is optimized, as FabreX drives much higher utilization of its resources.

www.gigaio.com

How it Works

The magic in FabreX, and the reason it is succeeding where others have failed, is its ability to use PCIe as a routable PCI network enabling all server-to-server communication, rather than just connecting resources to a single processor using a PCI tree. FabreX is the only PCIe-based routable network in the market, and uses Non-Transparent Bridging (NTB) to shatter the boundary between two PCIe hierarchical domains.
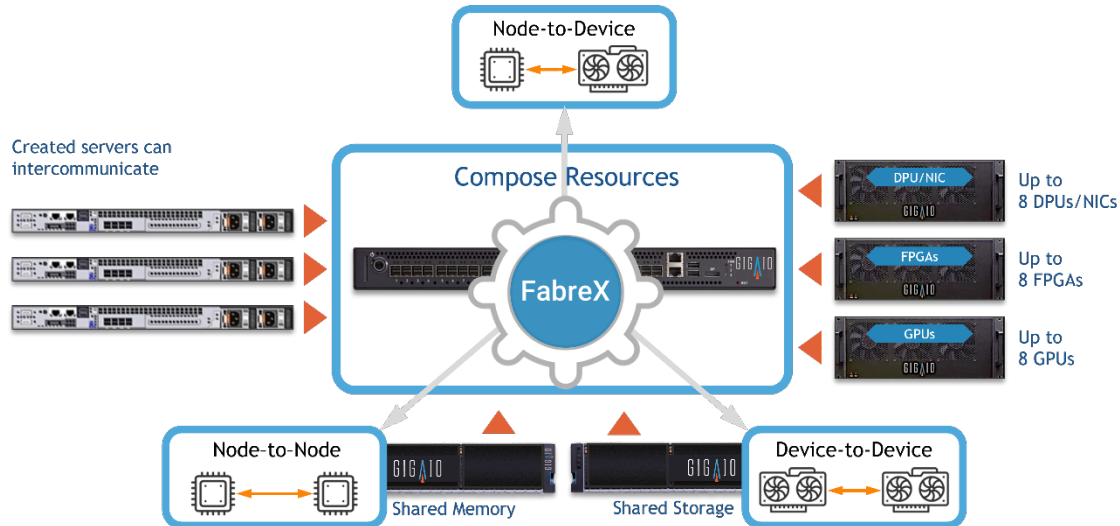
At its core, FabreX incorporates a construct for a memory address-based router with a built-in intelligence unit hosted in the switches that comprise the FabreX interconnect network. This routing mechanism virtualizes all hardware resources consisting of processor ecosystems and I/O devices as memory resources within a 64-bit Virtual Address Space. Communication between these resources consists exclusively of using the memory semantics of 'Memory Read' and 'Memory Write'. This means you can compose servers and CPUs in the exact same way as you would compose end points (GPUs, FPGAs, ASICs, Smart NICs — anything with a PCIe connection).

GigaPod Capabilities Included

| Accelerator Composability/Disaggregation	Storage Composability/Disaggregation
- Each host can access individual GPU devices	- Each host can access individual NVMe-oF storage devices
Management - Supports Redfish® RESTful APIs; real-time cluster topology; and utilization and performance dashboards - Predictive health and telemetry monitoring - Role-based authentication and access control	**Bundled Software** - FabreX Fabric Manager - Preloaded Linux & libraries including NVMe-oF and GDR - NVIDIA AI Enterprise Essentials featuring Base Command Manager for Composable Cluster Management and HA Storage

GigaPods are Delivered as Fully Engineered Solutions

The FabreX GigaPod Family comes pre-configured with FabreX composable switches, network cards, cables, management software and expansion chassis — everything you need to easily deploy the world's most advanced and flexible accelerated computing platform.

GigaPod
The perfect entry point and preconfigured for easy deployment.
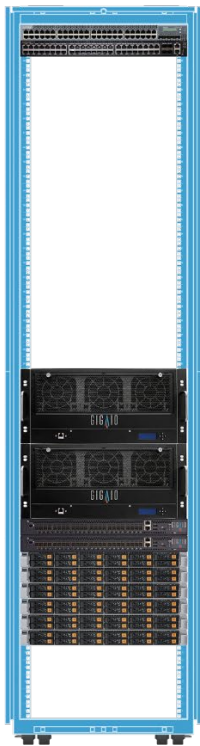
SUPPORTS UP TO:

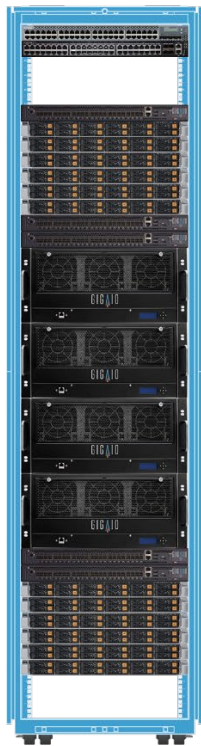**640 CPU Cores**

4.6TB Memory

307TB Storage

16 Accelerators

- 1 GigaPod kit — includes FabreX™ composable switches, network adapter cards, and cables
- Up to 2 GigaIO Accelerator Pooling Appliances — each up to 8 mix-and-match GPUs
- Up to 2 storage servers — each up to 154TB
- Up to 4 compute nodes — pick the processor — dual AMD EPYC™ 7713, 75F3, or 7543 — each with up to 1TB memory and 128 cores
- NVIDIA AI Enterprise Essentials

GigaCluster
Expand your system to increase compute and storage performance.

SUPPORTS UP TO:
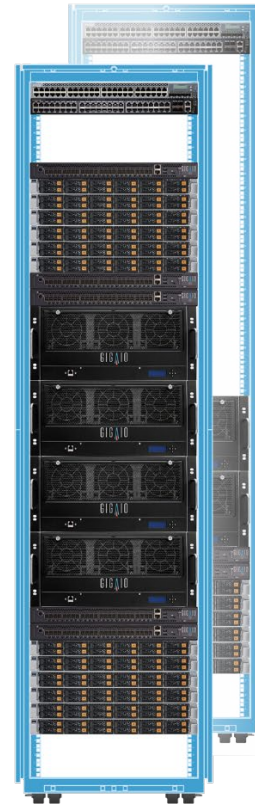
1,280 CPU Cores

9.2TB Memory

614TB Storage

32 Accelerators

- 1 GigaCluster kit — includes FabreX™ composable switches, network cards, cables, and expansion pack
- Up to 4 GigaIO Accelerator Pooling Appliances — each up to 8 mix-and-match GPUs
- Up to 4 storage servers — each up to 154TB
- Up to 8 compute nodes — pick the processor — dual AMD EPYC™ 7713, 75F3, or 7543 — each with up to 1TB memory and 128 cores
- NVIDIA AI Enterprise Essentials

3Pod GigaCluster
When you need the ultimate in compute power.

SUPPORTS UP TO:

1,920 CPU Cores

13.8TB Memory

922TB Storage

48 Accelerators

- 3Pod GigaCluster kit — includes FabreX™ composable switches, network adapter cards, cables, and expansion pack
- Up to 6 GigaIO Accelerator Pooling Appliances — each up to 8 mix-and-match GPUs
- Up to 6 storage servers — each up to 154TB
- Up to 12 compute nodes — pick the processor — dual AMD EPYC™ 7713, 75F3, or 7543 — each with up to 1TB memory and 128 cores
- NVIDIA AI Enterprise Essentials

Learn more at gigaio.com/gigapod

www.gigaio.com