GigaPod

An Engineered Solution for Dynamic Computing That Maximizes GPU Utilization — Speeding and Simplifying HPC and AI Workloads

HIGHLIGHTS

BREATHTAKING PERFORMANCE
Lowest latency and highest bandwidth

COMPELLING ECONOMICS
Higher resource utilization, less complexity, power, cooling, CAPEX, OPEX

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

EXTREME FLEXIBILITY
True disaggregation with dynamic composability

LEADERSHIP EFFICIENCY
Strips away conversion and overhead with 100% PCI Express interconnect

NO SUPPLY CHAIN CONSTRAINTS
Immediate availability and no networking vendor dependencies

The GigaIO GigaPod delivers the industry's lowest latency and the highest effective bandwidth by integrating compute and GPU acceleration I/O into a single-system dynamic fabric, using standard PCI Express (PCIe) technology.

The GigaIO GigaPod is a flexible solution for heterogeneous compute designed to dynamically accommodate different types of workloads on a single hardware architecture. 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 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 to a workload at the right time, extending access across a broader set of users.

Workloads run faster, because they are utilizing resources as if they were in one rack-scale server – not multiple constrained server nodes connected by a congested, complex network. All servers, GPU's and accelerators communicate over one dynamic memory fabric.

A Dynamic Memory-Centric Fabric

At the heart of the GigaPod family is FabreX™, our dynamic memory fabric. With FabreX you can effortlessly connect new memory, network and storage products and 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 to enable the 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, with direct memory access by an individual server to system memories of all other servers in the fabric, for the industry's first in-memory network. This new architecture enables a unified, software-defined, and memory-centric composable infrastructure.

Breathtaking Performance

The FabreX GigaPod platform delivers the industry's lowest latency AND the highest effective bandwidth. Latency from one server to another is less than 300ns – true PCIe performance across the entire cluster. The FabreX Gen4 implementation scales up to 512 bidirectional Gbits/sec bandwidth.

Outstanding Simplicity and Easy Management

Accessing all the power of a GigaPod is as simple as using the same cluster management and workload scheduling managers as you always have. In order to provide a complete turnkey solution to support its workflow-defined infrastructure, all GigaPods include NVIDIA Bright Cluster Manager software to compose compute infrastructures that can be configured and scaled in minutes to handle the most demanding AI and HPC workloads. NVIDIA Bright Cluster Manager coordinates 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 or storage devices, such as NVMe, 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 all across the same high performance PCIe memory fabric. GigaPods span multiple servers while GigaClusters span multiple racks to scale up single-server systems and scale out multi-server systems, all unified via the FabreX 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 to incorporate new technology. Upgrade or add compute, GPU and application accelerators at the component level that plug-n-play with your environment. Every major subsystem can now operate on its own upgrade cycle. And the total cost of the system is optimized as FabreX drives much higher utilization of the resources.

www.gigaio.com

How it Works

The magic in FabreX, and the reason it is succeeding where others have failed, is the ability to use PCIe as a routable PCI network enabling all server-to-server communication, not 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.
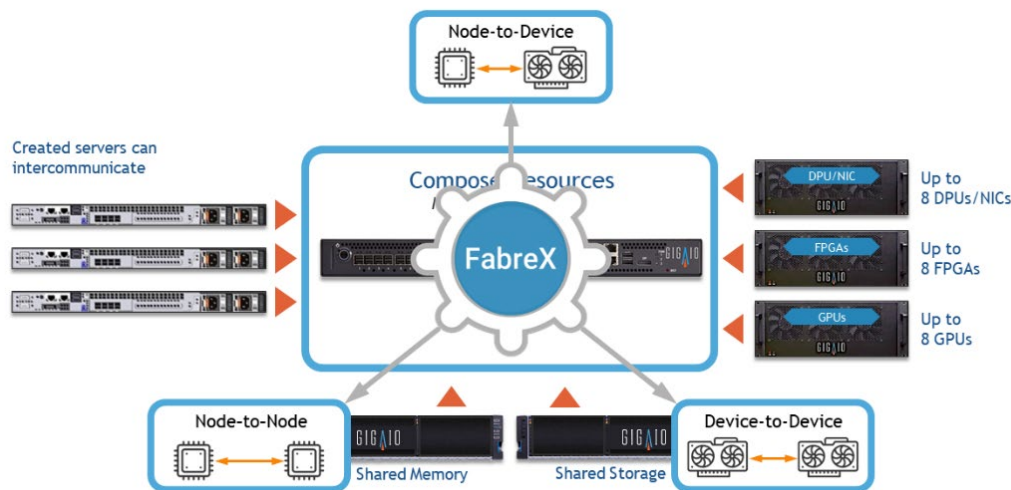
At its core, FabreX incorporates a construct for a memory address-based router with a built-in intelligence unit hosted in the switches which comprise the FabreX interconnect network. This routing mechanism virtualizes all hardware resources consisting of processor ecosystems and I/O devices as memory resources within a 64-bit Virtual Address Space. Communication between these resources consists of using exclusively memory semantics of 'Memory Read' and 'Memory Write'. This means you can compose servers and CPUs exactly the same as you compose end-points (GPUs, FPGAs, ASICs, Smart NICs – anything with a PCIe connection).

Included GigaPod Capabilities

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 API
  Real-time cluster topology
  Utilization and Performance Dashboards
- Predictive health & telemetry monitoring
- Role-based authentication and access control
	Bundled Software
	- FabreX Fabric Manager
	- Preloaded Linux & libraries including NVMe-oF and GDR
	- Composable cloud management platform with Bright Cluster Management supporting High-Availability (HA) Storage and Bright for Data Science

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:

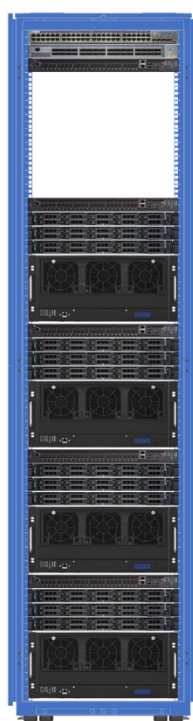**640 CPU Cores**

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
- Bright Cluster Manager supporting High Availability (HA) storage and Bright for Data Science

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
- Bright Cluster Manager supporting High Availability (HA) storage and Bright for Data Science

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
- Bright Cluster Manager supporting High Availability (HA) storage and Bright for Data Science

www.gigaio.com