GigaIO FabreX & NVIDIA Base Command Manager

SOLUTION BRIEF

A complete solution for building and managing composable clustered infrastructure using GigaIO FabreX™ and NVIDIA Base Command Manager.

Composable infrastructure is the latest advance in increasing data center efficiency. First there was converged infrastructure, and following that, Hyper-Converged Infrastructure (HCI) became mainstream in enterprise data centers. HCI provides greater compatibility and cost savings and streamlines data center management, but at a cost. Converged infrastructure locks a company into a single vendor.

The need to extract maximum value from the ever-growing mountains of data being collected from many different sources is driving increased use of Artificial Intelligence (AI). In turn, the processing power required to parse this data is driving the increased use of accelerators such as GPUs, FPGAs, Vector Engines, and DPUs in a variety of configurations that many traditional data centers have difficulty accommodating. Enter Composable Infrastructure.

Composable infrastructure is an emerging technology that eliminates the restrictions imposed by traditional, static server architectures. It delivers the agility and flexibility of cloud computing while lowering total cost of ownership.

The Challenge

AI is driving the demand for heterogeneous compute. Different workloads need radically different configurations, making it extremely difficult for data center managers to predict and anticipate their users' future needs. Traditional on-premise data centers face difficulty provisioning for these new workloads, since they have fixed configurations and can be difficult to scale, resulting in poor resource utilization. These limitations have driven many workloads into the cloud, where they can find more flexible configurations, and scale on the fly. So what's the problem?

While moving workloads to the cloud gets around some of the limitations of on-premises resources, it also limits the choice of accelerators to what a given cloud vendor offers. And while scaling is quick and easy, the cost of running workloads in the cloud can be difficult to control. Here is a comparison of considerations for deploying workloads on-premises versus in the cloud.

Public Cloud			Traditional On-Premises
Flexible Configurations	▲		**Fixed** Configurations
Limited Accelerator Combinations		▲	**Choice of** Components
Scale on the Fly	▲		**Difficult** to Scale
Difficult to Control **Costs**			**Poor** Resource Utilization
Non-deterministic		▲	**Deterministic**
Optimized for **Agility**	▲	▲	Optimized for **Performance**

GigaIO and NVIDIA created an end-to-end solution for composable rack-scale computing and made it easy to deploy and manage.

GigaIO FabreX

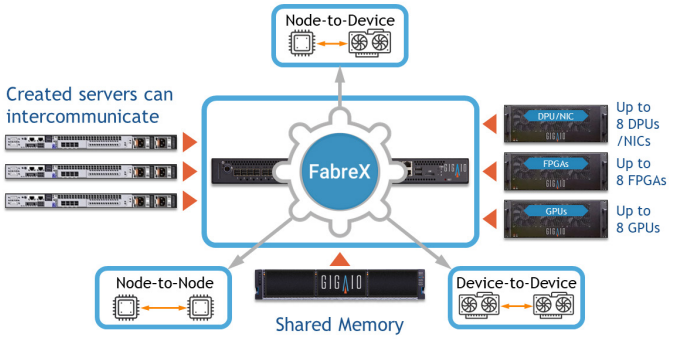
With GigaIO's FabreX, an entire rack of resources can be treated like a collection of disaggregated puzzle pieces that can be re-assembled on the fly based on the specific requirements of each workload.

NVIDIA Base Command Manager

NVIDIA Base Command Manager lets you quickly build and manage heterogeneous high-performance Linux clusters that host HPC, machine learning, and analytics applications that span from core-to-edge-to-cloud.

How It Works

The GigaIO and NVIDIA solution combines NVIDIA Base Command Manager's ability to build and manage clusters with GigaIO FabreX's ability to disaggregate and compose servers, accelerators, storage and memory together within a seamless memory fabric.



Create "impossible servers™": The GigaIO and NVIDIA solution lets you configure "impossible" servers to handle those tricky workloads. Need 16 or even 32 GPUs in one server? Yes, you can do that with the GigaIO and NVIDIA solution.

Deploy scale in minutes: NVIDIA's auto-scaling creates a dynamic, multi-purpose infrastructure. GigaIO's FabreX extends that agility through to each hardware element in a rack by enabling the creation of composable GigaPods and GigaClusters with cascaded and interlinked switches, as well as SuperNODEs.

Maximize resource utilization: Some workloads require servers with a particular set of resources, be that memory or a specific number and type of accelerators. That can leave perfectly good servers and the resources they contain idle while jobs are running. With this GigaIO and NVIDIA solution, all individual available resources can be put to use, optimizing resource utilization and minimizing cost.

Heterogeneous compute: Release your data center from vendor lock-in. The GigaIO and NVIDIA solution works with any server, accelerator, or workload manager you're likely to need. Mix and match different types of GPUs seamlessly within a single workload to optimize your time to results.

Easy to deploy and maintain: NVIDIA Base Command Manager reduces the time, skills, effort, risk, and complexity of building clustered infrastructure. It automates the entire process of provisioning clustered servers from bare metal, and images them with all of the software they need to run their intended workloads. Once deployed, NVIDIA Base Command Manager automatically propagates updates without disrupting users. Similarly, once the FabreX software and hardware solution is installed, through the magic of software-defined hardware the administrator never needs to recable or touch the physical hardware infrastructure.

Integrated build, management & monitoring: NVIDIA Base Command Manager monitors your entire cluster, providing detailed insight into resource usage, and quickly isolating the source of failures and performance problems.

The GigaIO and NVIDIA solution lets you handle more workloads while maximizing resource utilization, minimizing cost, and managing everything from an enterprise-class user interface.

Learn more at gigaio.com/nvidia-bcm