Cloud-Class™ Composable Disaggregated Infrastructure

White paper by GigaIO Networks and Microchip Technology

CONTENTS

This document is for informational purposes only and may contain typographical errors and technical inaccuracies. The content is provided as is, without express or implied warranties of any kind.

What is Composable Disaggregated Infrastructure?

The nascent category of Composable Disaggregated Infrastructure (CDI), sometimes also referred to as SDI (Software-Defined Infrastructure) is generating increasing interest in the data center. Composable infrastructure seeks to disaggregate compute, storage, and networking resources into shared resource pools that can be available for on-demand allocation (i.e., "composable").

CDI is a data center paradigm with a software layer that abstracts a hardware layer consisting of IT resources across a virtual fabric and organizes them into logical resource pools, not dissimilar to a Hyper-Converged Infrastructure (HCI) cluster. But rather than having a linked-up series of self-contained nodes of compute, storage, network and memory, as in HCI, composable infrastructure is a rack-scale solution like Converged Infrastructure (CI). Unlike CI, however, composable infrastructure is, at its core, an Infrastructure-as-Code delivery model with an open and extensible unifying API layer that communicates and controls hardware resources. GigaIO uses open standard RedFish® APIs (instead of vendor-proprietary APIs as other CDI vendors do) enabling programmatic and template-driven control and automation so admins can rapidly configure and reconfigure disaggregated IT resources to meet the requirements of specific workloads and DevOps. This flexibility and speed are, in part, the reason many analysts expect the composable infrastructure market will take off over the next few years.

According to Frost & Sullivan, "*A Composable Infrastructure enables you to manage your infrastructure resources (physical, virtual, general-purpose, application-optimized, on-premises, and cloud) to deliver a better mix of performance, security, scalability, and cost for your workloads. It's as if a child's set of Lego bricks came with the ability to replicate blocks as needed and programmed instructions to configure those blocks into a ninja temple today and a working race car tomorrow.*"

The appeal of Composable Infrastructure is obvious: it promises to make data center resources as scalable and readily available as cloud services. The technology delivers fluid pools of networking, storage, and compute resources that can be composed and re-composed as needed. This on-the-fly capacity, which is based on what the software applications require at any given time, means enterprises can minimize over-provisioning and stranded assets to   maximize efficiency. Many industry experts consider it the logical next phase of Hyper-Converged Infrastructure (HCI). Enterprises who need to unite computing, accelerators and storage resources are no longer limited to just converged and hyper-converged infrastructure (CI/HCI) frameworks: they can now utilize a system that marries the best of both worlds — CDI.

Trends Driving CDI Adoption

The emergence of AI with its avalanche of data has spurred the growth of purpose-built compute nodes beyond the traditional X86 CPU: GPUs, FPGAs, ASICs, all designed to optimize various types of compute operations. Data streams are far wider and deeper than ever, with workflows so varied they may combine different infrastructure set ups within a single workflow. Yesterday's breakthrough was CPUs combined with GPUs. Today, a single workload may require different types of accelerators (i.e., V100s for compute and RTXs for visualization, maybe FPGAs for encryption), all within the same workflow.
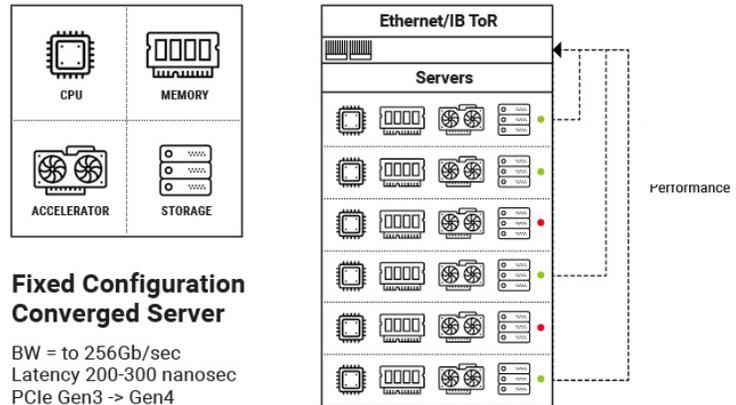
Figure 1 –Rack Architecture Today

At the same time, the server as the foundational unit of a data center is becoming obsolete. Microservices, containers, and VMs are all but abstracting the physical server. This has led to a heterogeneous computing environment where the old silos based on server provisioning make less and less sense. The promise of CDI is to render all resources fluid and available to be composed on the fly, as the workload requires. Industry analysts agree this is the future:

> *"Software-defined infrastructure has become the leading option for enterprise executives who want to cut costs, simplify IT management, and eliminate platform silos in the datacenter,"* said Chris Kanthan, research manager, Infrastructure Systems, Platforms, and Technologies group at IDC. *"Software-defined infrastructure plays a key role in the digital transformation journey of enterprises by providing extraordinary agility, scalability, and competitive advantage."* (IDC's Worldwide Software-Defined Infrastructure Taxonomy, 2019)

> *"I do think it's the future of enterprise computing,"* said Patrick Moorhead, president and principal analyst at Moor Insights and Strategy is quoted as saying in sdxCentral.

> *"We think that in the long run, disaggregated but composable compute and storage are going to be the architecture of choice for future platforms – this is what the hyperscalers and cloud builders have created – and this does not fit into the IDC definitions. Disaggregation means smashing the motherboard in the system (logically speaking) and using protocols and interconnects to link CPUs, memory, memory, and storage to each other; composable means essentially creating systems from these components on the fly to run specific workloads, something a lot more physical than a logical partition."* Timothy Prickett Morgan, The Next Platform.

The Key Benefits of Disaggregation

Even if your organization does not anticipate a need to change infrastructure on the fly because you have predictable workloads which essentially diminish CDI's composability advantage, you might still benefit from the ability to disaggregate your rack resources from the servers where they are trapped.

The Key to Eliminating Stranded Resources is Disaggregation

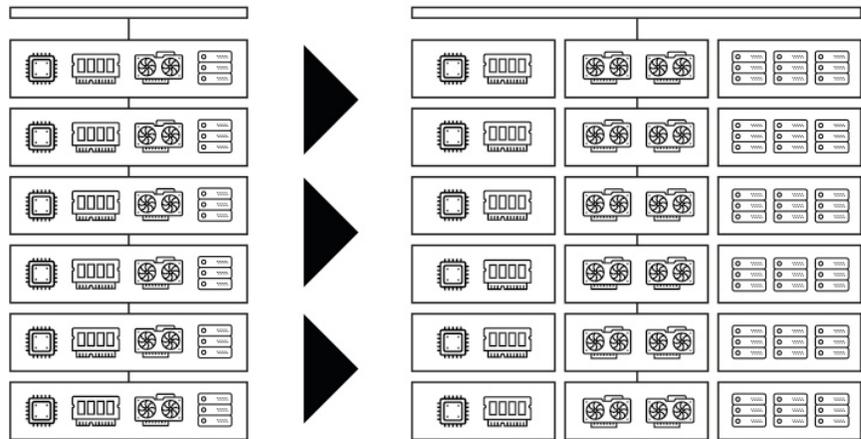Enabling servers to use any device at any time

Figure 2 –Eliminating Stranded Resources

Following are some of those benefits:

- Extend the product life cycle of your most expensive resources. When accelerators or SSDs are no longer trapped within a server, upgrading or replacing an individual component no longer implies an involved H/W upgrade procedure, or worse, throwing the entire server out;

- Scale your resources independently. You might have plenty of computing power but need more storage, or more GPUs, without necessarily needing more CPUs;

- Pay as you grow. One thing you know for sure is that your users will have intermittent peaks in demand over the next planning horizon, but instead of overprovisioning to accommodate the peaks, add just the resources you need, when you need them;

- Decouple purchasing decisions. A disaggregated rack allows you to refresh existing resources as needed - for example, add the latest GPU model in an existing cluster;

- Modify your CPU to GPU ratio as needed. Different workloads may need different combinations of compute resources;

- Give your users access to different types of GPUs and accelerators. For example, for compute heavy tasks you might want to give them access to a set of NVIDIA V100s, but for visualization a different model (like the NVIDIA RTX8000) might be a better fit;

- Manage your software license expenditures more efficiently. Instead of paying for another core and another license when all you need is more accelerators, just add GPUs to your existing CPU license;

- Improve serviceability. Certain kinds of accelerators may fail more often than CPUs, and with a disaggregated GPU bank in a JBOG (Just a Bunch of GPUs) appliance, there is no need to open up the servers;

- Optimize CapEx by matching the depreciation schedule of assets to their expected lifespan: give your users the option to upgrade sooner to components like GPUs and FPGAs by taking them out of the servers and giving them a shorter depreciation timeline;

- Finally, keep it simple. By moving the GPUs and FPGAs, which use a lot of power and generate lots of heat, to an enclosure outside the server, you get back the advantages of the robustness and simplicity of an optimally configured workhorse server.

How is GigaIO's Implementation of CDI Unique?

Most vendors in the composable infrastructure space treat data center resources as two conceptual blocks: storage/not storage. Working with the limitations of incumbent networking technologies, they disaggregate and recompose storage resources, but none of the other resources in the rack: CPUs, GPUs, FPGAs, and memory. But why stop at storage? GigaIO's FabreX™ network enables data centers to disaggregate and recompose at the finest bare-metal granularity in the rack, turning the rack itself into the server.

- Our platform is open, and based on RedFish APIs, not another proprietary software package you need to pay license fees for and administer as another pane of glass. The RedFish standard is designed to deliver simple and secure management for converged, hybrid IT and the Software Defined Data Center (SDDC).

- Several enterprise-class bare-metal composability and private cloud software certified options are available from a variety of vendors depending on your needs – including advanced security and provisioning features, and access control capabilities;

- Using these interfaces, it is now possible to compose down to the resource level – not just at the server level. You can reach inside the rack and orchestrate the resources as it makes sense for a given workflow;

- We are completely hardware and software agnostic: use your favorite server or accelerator vendor, storage, or container or Hyper Converged Infrastructure supplier;

- There is no performance penalty or composability tax with FabreX, because we are the only native PCIe fabric throughout the rack, *including node to node*. The ultra-low latency inherent in PCIe remains throughout the disaggregation and re-aggregation process, we never need to switch to InfiniBand or Ethernet within the rack, *including to communicate server to server*. As a result, the full end-to-end latency remains below a microsecond.

The Technology Undergirding GigaIO's Composable Infrastructure Platform

For almost 30 years, PCI, PCI-X and the PCI SIG's latest incarnation, PCIe, have become the ubiquitous interconnect for compute and storage inside every manner of data center appliance. One of the reasons for this is that PCIe enjoys wide support across all popular OSs, hypervisors, CPU architectures, NVMe storage, GPUs and even

next-generation AI ASICs. The other reason is that the standard has constantly evolved to support ever increasing bandwidth and functionality. In fact, PCIe has been so successful that new emerging standards such as Compute Express Link (CXL) and Gen-Z have leveraged it and built their features on top of PCIe.

The team at GigaIO realized that the ubiquitous nature of PCIe coupled with its low latency performance was a very attractive baseline on which to construct a rack-level network. However, using PCIe in this manner does create significant challenges since the standard was not created for the rack-level use-case. GigaIO has developed its FabreX network based on IP that is enshrined in 6 fundamental patents. FabreX allows the creation of a true network natively running on PCIe without requiring the use of another network like Ethernet or Infiniband. This means that applications running on servers can communicate with each other using TCP/IP, NVMe-oF, MPI and Open MPI all natively over PCIe.

Additionally, GPUs can communicate directly with each other while traversing host PCIe domains using technologies like Nvidia's GPU Direct RDMA (GDR) which enable direct GPU to GPU communication. FabreX's optimized data mover essentially converts what are typically complex RDMA operations into much simpler PCIe DMAs thereby removing expensive bounce buffers from the data path. This approach inherently lowers the memory requirement in the system, and in the case of NVMe-oF and GDR also reduces latency for transactions.

Microchip's industry leading Switchtec PCIe switches form the foundation of FabreX's hardware layer. The latest version of these devices offers connectivity with up to 256Gb/s unidirectional bandwidth per link. FabreX's server-to-server communication software leverages the NTB functionality in the switch which enables hosts to traverse PCIe root complex domains to access PCIe endpoints.

To facilitate Infrastructure-as-Code, GigaIO has adopted industry standard Redfish APIs for managing all appliances in the system (i.e., both network switches and resource pool enclosures). The advantage of this approach is that FabreX can be easily adopted into existing bare-metal or private cloud software without a massive overhaul of existing workflows. GigaIO's switch utilizes its Redfish API to serve advanced features such as auto-discovery of rack resources – a first for composable infrastructure.

To deliver true enterprise-grade infrastructure, data center operators are clamoring for greater visibility into thermals and in-service performance metrics for devices such as GPUs or next-gen NVMe SSDs. Again, these kinds of features are implemented via Redfish APIs in fully managed enclosures thereby enabling operators to meet stringent availability and downtime requirements on their CDI systems.

GigaIO's FabreX, made possible by Microchip's silicon, is designed to be compatible with future generations of PCIe (Gen 5 and Gen 6). As silicon and platforms for these data rates become available in the industry, FabreX can adopt them with minimal effort. Additionally, emerging standards such as CXL and GenZ will also be adopted as transport layers into FabreX to enable features such as cache coherency across a FabreX cluster.

GigaIO's FabreX enables capital and operating expense savings of between 30-50% depending on use-case and is appropriate for a wide variety of workloads in HPC, AI/ML/DL, genomics, computational chemistry, quantitative research, visualization, at an enterprise-wide or departmental scale.

About GigaIO

GigaIO has invented the first truly composable cloud-class software-defined infrastructure network, empowering users to accelerate workloads on-demand, using industry-standard PCI Express technology. The company's patented network technology optimizes cluster and rack system performance, and greatly reduces total cost of ownership. With the innovative GigaIO FabreX™ open architecture, data centers can scale up or scale out the performance of their systems, enabling their existing investment to flex as workloads and business change over time. For more information, contact the GigaIO team at info@gigaio.com or visit www.gigaio.com.

About Microchip Technology

Microchip Technology Inc. is a leading provider of smart, connected and secure embedded control solutions. Its easy-to-use development tools and comprehensive product portfolio enable customers to create optimal designs which reduce risk while lowering total system cost and time to market. The company's solutions serve more than 120,000 customers across the industrial, automotive, consumer, aerospace and defense, communications and computing markets. Headquartered in Chandler, Arizona, Microchip offers outstanding technical support along with dependable delivery and quality. For more information, visit the Microchip website at www.microchip.com.