GIGAIO

The Future of Composability with

White paper by GigaIO

CONTENTS

December 2023 | Rev 3.0

Introduction: The Advent of Heterogeneous Computing

GigaIO today delivers the most comprehensive implementation of composable disaggregated infrastructure (CDI) through its universal dynamic fabric, called FabreX™. Our software-defined composable architecture empowers users to pool and deploy previously static hardware. As a pioneer in full disaggregation and composition, the company has long been committed to incorporate the memory coherence benefits of CXL™ (Compute Express Link) in its next generation PCIe architecture, as the last step in full component disaggregation. As a matter of fact, GigaIO was one of the first companies to become a contributing member of the CXL Consortium and continues to participate and support the effort through significant involvement in the working groups.

The promise of full CDI is to **reduce time-to-insight** by ushering in the era of the hybrid data center with cloud-like agility and flexibility. GigaIO's software-first composable platform enables otherwise impossible configurations such as SuperNODE with 32 GPUs in a single node. FabreX increases resource utilization and lowers cost of ownership by allowing data center managers to compose individual components as needed and on-the-fly to adapt to changing workflows, and to burst to the cloud as needed, as organizations pivot to container-native architectures.

With the end of Moore's law, the compute function is splintering and has enjoyed a real renaissance of new forms and functions. Specialized functions that used to be entirely hosted within the CPU are being moved to more specialized chips to continue the drive for ever more performant systems. A veritable "Cambrian explosion", in the words of NVIDIA, is occurring today in both compute models and accelerator hardware, fueling new compute solutions.

The emergence of AI, in particular, fed by the vast increases in data being collected, is fundamentally reshaping the "three pillars of computing — compute, applications, and memory".

This revolution (machine learning and deep learning) necessitates far more computing resources than are available with CPUs alone, namely specialized compute accelerators such as general purpose graphics processing units (GPUs), Field-Programmable Gate Arrays (FPGAs), and tensor processing units (TPUs) with thousands of purpose-dedicated cores. And it is not just AI that is now benefitting from these new compute capabilities, but all compute-intensive tasks. For example, high performance computing modeling and simulation such as computational fluid dynamics (CFD), which previously used dozens of vector processors (Cray super computers), can now instead utilize thousands of GPU floating point cores.

As processing moves beyond the CPU to include GPUs, why stop at GPUs? Many data center applications that have evolved beyond CPUs, and are now relying on GPUs, will incorporate TPUs, but also FPGAs and application-specific integrated circuit (ASIC) accelerators, DPUs (Data Processing Units) and IPUs (Intelligence processing Units) to continue to drive performance improvements for the majority of their computation. Hence the data center is entering **the age of heterogeneous computing.** Now the challenge becomes how to seamlessly connect these devices together with the latency, bandwidth, and congestion control that next-gen applications require.

GigaIO's Solution: A Universal Dynamic Fabric

Several options exist to connect these heterogeneous resources. For example, NVLink and NVSwitch provide excellent intra-server interconnect speed, but they can only connect components inside the server, and only NVIDIA-branded GPUs. PCIe switching solutions can connect servers to accelerators or storage via PCIe, but server to server communication requires paying a composition penalty through the "store-and-forward" networks like InfiniBand or Ethernet. In contrast, FabreX is completely hardware, brand, and software agnostic and can connect any resource to any other over native PCIe, including server to server, server to device, and device to device, all while mixing and matching component brands. FabreX is the only solution on the market today that enables the use of any or all of these heterogenous resources **at native-device performance.**

Today there are several "composable" solutions on the market, which can be described as belonging to one of these categories:

- "Storage/not storage": the vast majority of so-called composable options only disaggregate and compose storage (SSDs and HDDs), but not accelerators (HPE Synergy, Dell EMC, Western Digital Open Flex, Cisco UCS)
- Accelerators as long as they are part of a server, and/or their own brand (NVIDIA, HPE OneView)
- "Fan-out": Disaggregated accelerators of various types and brands with select server types (Liqid, Dolphin)
- Disaggregated components including servers composed over native PCIe

FabreX is akin to the universal remote of TVs of yore, replacing the individual components' remotes, because it is the only data center fabric that requires no translation or performance hit since it composes components in their "native language", the one they already "speak" inside the server sheet metal.

This is possible because FabreX eliminates bounce buffers, and for most communication avoids costly (in terms of time) trips through the root ports and operating system kernels.

Past Barriers

Early barriers to composability have included **the lack of suitably fast point-to-point networks.** Prior to GigaIO, only store-and-forward communication networks such as InfiniBand or Ethernet had been available. I/O was internal only and restricted to the motherboard. The computing ecosystem had no high-speed switched fabrics available for rack-level **native** I/O device communication.

Another key factor was **the lack of coherent sharing** between host memory and device memory in the same address space, that is, memory coherence could not be guaranteed. This long-standing issue was illustrated through the memory coherence problem, which states that a memory is coherent if the value returned by a read operation is always the same as the value written by the most recent write operation to the same address[1].

The last barrier to disaggregation has been **the sheet metal boundary.** Modern applications demand up to 64 devices (e.g., accelerators) and beyond per system, but the majority of servers provide only a limited number of available device slots, usually only 2 to 4. Historically, external enclosures had been available for storage devices only. There were no enclosures for memory pools or large numbers of accelerators, and no capability for device reassignment. In addition, the historical computing ecosystem provided only complex operating system–level "device sharing", such as other servers.

The increasing demand for more compute and storage devices, and for access to larger memory spaces, driven by the rise of machine learning and deep learning systems, underscores the necessity to make these devices sharable. The industry needs fully composable disaggregated rack-scale computing that is easily re-composable on the fly, and that enables broad resource availability at native, device-level sharing, without sacrificing performance.

Composability without a composability tax.

Previous Attempts to Achieve the Goal

Previous attempts to achieve fully composable disaggregated infrastructure include:

- The early development of hot-plug and plug-and-play device movement/replacement (mostly an OS feature);
- The advent of InfiniBand for server and device enclosure communication, device sharing, and a (fan-out) I/O network;
- And PCIe switching solutions.

[1] Li K, Hudak P: Memory coherence in shared virtual memory systems. ACM Trans Computer Systems 7(4):321359, 1989.

Hot-plug and plug-and-play (aka “plug-and-pray”) were designed mainly for storage devices and do not address 100% disaggregation and composition throughout the rack.

InfiniBand is a non-native (i.e., a higher-level protocol not used natively by all devices in the rack) communications protocol and requires protocol translation — “the composition tax” — before communication with the devices in the rack. InfiniBand also lacks rack-level independent device discovery and memory coherence.

Existing PCIe switching solutions before GigaIO were not full fabrics (they are fan-out only), and did not support server-server cluster communication/composition or GDR cross-server GPU communication. Further, they do not provide rack-level device discovery of unassigned devices (e.g., unassigned servers) and do not have visibility into the multiple device trees in the rack. Lastly, they do not provide rack-level device discovery of unassigned components, nor provide memory coherency.

Major pain points today in the move to heterogeneous computing are resource utilization and stranded resources. For example, GPUs and NVMe storage are typically the most expensive discrete resources in the datacenter. But prior to composable infrastructure, these resources were held captive within the server, not readily sharable outside that sheet metal. Since servers do not provide enough device slots (PCIe slots) to establish sufficiently large resource pools, such as memory or GPU pools, increasing the numbers of resources like GPUs required for a given workload meant increasing the numbers of servers.

Hyperscaler datacenters such as AWS and Azure, and AI-specialized cloud service providers such as CoreWeave and TensorWave, are especially sensitive to resource utilization, as it is their de facto manufacturing floor, and they depend on the ability to scale architectures appropriately for increased demand. These datacenters need the flexibility of workload-specific custom composition for cost-effective sharing of these expensive resources. GigaIO’s technology for composable disaggregated infrastructure is a perfect answer to the challenges presented by applications, computation, and memory evolution, by enabling resource sharing and broad availability of otherwise stranded resources.

The Holy Grail: Memory

The final missing element and the holy grail in the composable infrastructure play has been memory. The closest available today has been GigaIO composing Optane Persistent Memory, but with Optane being discontinued by Intel, and without CXL, there is no memory coherency, and transfers operate over the I/O bus, not the memory bus. Some customers have

implemented Network Attached Disaggregated Memory schemes across InfiniBand with response times of 2-4 microseconds, 100x slower than DRAM memory. Others have used FabreX to improve the performance of this disaggregated memory to about 1 microsecond. Much better, but still a considerable performance hit to memory — and without coherency.

Memory itself has evolved, transforming from L1/L2/L3 cache (tier 1) with dynamic random access memory (DRAM; tier 2), to include tier 3 SCM and 3DXP (3D-crosspoint) as persistent memory, and, to tier 4 SCM memory used as large 3DXP nonvolatile memory express (NVMe) solid state drives (SSDs). But so far it has eluded attempts to disaggregate it from the CPU or the GPU and pool it as a common resource addressable from any compute device. That is where CXL comes in, and why GigaIO has been so keen to support its development.

Another consideration about memory in accelerated computing is the importance of large aggregated GPU memory space, as ever increasing AI models need to be held in memory. In this case we are describing the combined GPU memory or "VRAM" that exists in one single memory space. While CXL is not expected to impact that aspect of memory, it is worth keeping in mind for accelerated computing.

The Ideal Composability Solution

Fully composable disaggregated infrastructure requires 100% disaggregation, re-aggregation capability without a performance penalty, a memory-centric I/O communication network, and enterprise-class dynamic device discovery, composition, and control. And the ideal solution includes a fully integrated hardware and software stack where applications require zero code change to take advantage of composability. The ideal architecture includes the ability to disaggregate and compose all server elements, including CPU modules, with CPUs of various types, with 1st and 2nd tier memory; enclosures of 3rd and 4th tier SCM memory pools; enclosures of accelerators of various types, such as GPUs (JBOGs), FPGAs (JBOFs), and ASICs; and enclosures of fully disaggregated storage devices of various tiers including Optane/3DXP, 3D NAND, hard disk drives, and archive hard disk drives.

In that environment, the central processing unit becomes **the central coordinating unit** for each composable system in the rack. The choice of CPU/server is no more or less important than the choice of other elements in the composed system, and the specific type of CPU/server, along with the types of memory, accelerators, and storage, are all selected based on the workload to be assigned.

Complete auto discovery requires device-level communication with all disaggregated device types in the rack, on the same fabric, including servers. Other partial (fan-out) I/O networks are incapable of full-rack server discovery and communication.

The complete composability solution also requires an open approach to enterprise-class rack and cluster management via standard APIs, such as Redfish™, to enable dynamic composition by any and all of the emerging composition management tools such as NVIDIA Base Command Manager (formerly Bright Cluster Manager) Slurm, Grid Engine, VSphere and the hyperscalers' own VM software, to avoid yet another pane of glass.

For AI today, 85% of the applications run under Pytorch or TensorFlow, and the ideal solution includes a fully integrated hardware and software stack where applications require zero code change to take advantage of composability.

How CXL Breaks the Logjam

The new Compute Express Link (CXL) standard for CPU and I/O device communication will finally enable the full composability of memory and the development of memory-coherent I/O networking. GigaIO's high-speed switched I/O fabric, based on PCIe and CXL standards, enables unprecedented low-latency communication. This low-latency communication, coupled with the memory-device cohesion and coherency of CXL coming with PCIe Gen5, enables 100% rack-level disaggregation and composition at hyperconverged performance.

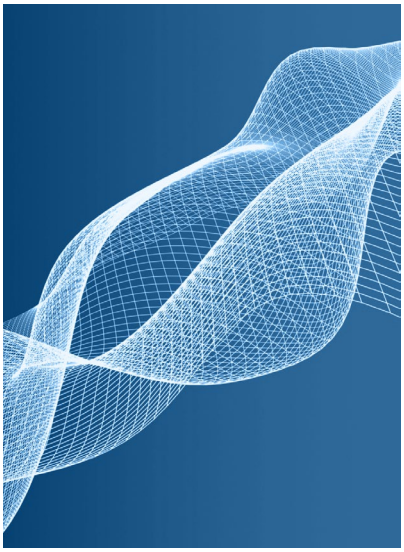
One hundred percent disaggregation requires rack-level device discovery and identification of 100% of the devices (servers, memory pools, accelerators, and storage devices), whether already composed or as yet unassigned. This can only be accomplished using the CPU-native and device-native I/O interfaces, which are based on CXL/PCIe. **Ethernet or InfiniBand are simply not capable of supporting discovery, disaggregation, and composition at this level of granularity.** GigaIO FabreX with CXL is the only solution which will provide the device-native communication, latency, and memory-device coherency across the rack for full performance composition.

Current State of the Art

GigaIO FabreX provides 100% disaggregation with full granularity of composition. This includes CPUs on multiple servers, multiple memory servers, multiple pooling appliances of 8–16 accelerators each (GPUs, FPGAs, ASICs), and multiple pooling appliances of dozens of storage devices. As evidence GigaIO is leading the technology race, SuperNODE, the first 32-accelerator to a single node supercomputer, was recently awarded "Best AI Product or Technology" at SC23 by the editors of HPCWire.

With CXL, GigaIO FabreX will support memory pooling appliances as well (e.g., JBOMs, or "Just a Bunch Of Memory"). All resources are contained within a single pod of racks and may be assigned among various composed systems. With GigaIO FabreX, the entire communication and composition is performed with the device-native I/O interface, including server-server, a first in the industry, and protected by our IP.

> This has been made possible by GigaIO's strong expertise in PCIe. As stated by 451Research, *"GigaIO's technologists have a history of developing PCIe fabrics that goes back many years. They have seen what works and what doesn't. The experience gives them a significant advantage this time around."*[2]



The GigaIO FabreX communication network infrastructure is deployed via software, link cards, and switches. The high-speed switched fabric is currently based on PCIe Gen4 and will soon be transitioned to PCIe Gen5 with CXL. At present, the GigaIO FabreX CPU-native, device-native communication infrastructure provides I/O device communication, NVME-oF and GPU GDR (GPU-oF) communication, and rack-level inter-server, inter-process network communication (MPI and TCP/IP). CXL-based coherent memory sharing and messaging will be implemented in FabreX. Only minor engineering upgrades of existing FabreX infrastructure are required to support the transport of CXL coherency messaging. **No invention or redesign is required.** And we anticipate all versions of our existing network software will naturally evolve and support CXL.

GigaIO FabreX provides dynamic disaggregation and composition for all device types right now, with composition of any and all disaggregated device types, auto discovery of both assigned and unassigned devices, visibility into and communication with multiple device trees, and enterprise-class remote composition and management APIs via standard Redfish.

Next Steps

In order to complete CXL deployment, adding memory-coherent I/O communication network infrastructure will involve implementing the 3 CXL multiplexed subprotocols, that is, CXL.io, CXL.memory, and CXL.cache.

CXL.io provides support for PCIe, and facilitates processes such as device discovery, link negotiation, interrupts, I/O messaging, etc. CXL.memory and CXL.cache provide support for device, memory pool, and host memory coherency messaging over the same data link and physical network layers as used by PCIe Gen5. GigaIO completed its transition from PCIe Gen3 to Gen4 in early 2020. Planning and support for PCIe Gen5, with CXL messaging, is already underway, with an expected completion date in 2024. Actual deployment will be driven by

[2] 451Research "Coverage Initiation: GigaIO uses PCIe fabric as basis for composable infrastructure" — January 2021

availability of silicon and speed of CXL adoption and support by the entire ecosystem (CPU, operating systems and endpoint device manufacturers).

GigaIO FabreX with CXL will initially support vertical coherency domains. A host and all its composed devices, such as accelerators, memory pools, and storage devices, will share CXL memory and cache coherency within a single composed system. Thus, multiple vertical coherency domains, each within their own composed systems, will be available within a pod of racks. Horizontal coherency among a cluster of servers for NUMA-style (Non Uniform Memory Access) memory sharing will come later.

The benefits of GigaIO FabreX with CXL and PCIe Gen5 will include a double bandwidth advantage, increasing from 256Gb/s (Gen4) to 512 Gb/s, full duplex, for each x16 link, and significantly lower disaggregated I/O communication latency. Overall, composed disaggregated systems will operate with the cohesion and performance of a hyperconverged system, but much greater resource utilization (or efficiency). And the FabreX CXL I/O network will extend application-transparent memory and cache coherency to composed devices throughout the disaggregated system.

Required Ecosystem Evolution

The benefits of FabreX with CXL will require the evolution of CPUs, operating systems, servers, and endpoint devices in the ecosystem to implement the CXL standard. CPU developers such as Intel, AMD, IBM, and ARM will need to support CXL.io (PCIe Gen5), CXL.memory, and CXL.cache in terms of coherency messaging and coherency policy management. Server developers such as HPE, Dell, and Supermicro will need to support CXL.io (PCIe Gen5), CXL.memory, CXL.cache, and provide coherency policy management for DRAM. They will also need to provide 21st century BIOSes with support for much larger numbers of enumerated devices, well beyond the current limits of the servers' sheet metal. Accelerator developers such as NVIDIA, AMD and Intel, and storage developers such as Micron, Intel, WD, and Samsung will need to support CXL.io (PCIe Gen5), CXL.memory, and CXL.cache in terms of coherency messaging and coherency policy management. And the operating systems will need to include support for "far" memory such as CXL.

For GigaIO FabreX, the upgrade from PCIe Gen4 to PCIe Gen5 will require very similar effort as was involved in going from FabreX Gen3 to Gen4. Support for CXL.cache and CXL.memory messaging will be developed, but for the FabreX I/O network, **the coherency policy management is already there.** FabreX CXL need only ensure that the coherency messages are routed to their intended destinations. In this respect, there is far less work for GigaIO than for CPU and device developers, to upgrade to the CXL standard.

The holy grail is within sight, and GigaIO is leading the quest.

About GigaIO

Headquartered in Carlsbad, California, GigaIO democratizes AI and HPC architectures by delivering the elasticity of the cloud at a fraction of the TCO (Total Cost of Ownership). With its universal dynamic infrastructure fabric, FabreX™, and its innovative open architecture using industry-standard PCI Express/soon CXL technology, GigaIO breaks the constraints of the server box, liberating resources to shorten time to results. 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 info@gigaio.com or visit www.gigaio.com. Follow GigaIO on Twitter (X) and LinkedIn.