GIGAIO

POSITION PAPER

The GigaIO™ FabreX™ Network

Position Paper

Composable Disaggregated Infrastructure with CXL/PCIe Gen5

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.

June 2020| Rev 1.0

Introduction

GigaIO Networks' intellectual property has enabled the company to deliver the fullest implementation of composable disaggregated infrastructure (CDI) through its high-performance interconnect network, FabreX. 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 Gen5 architecture. As a matter of fact, GigaIO was one of the first companies to become a full 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 increase resource utilization and lower cost of ownership by allowing data center managers to compose all individual resources as needed and on-the-fly to adapt to changing workflows. 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 in compute models today that continue to fuel new compute solutions. The emergence of AI, in particular, that is 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), 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. 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 datacenter applications that have evolved beyond CPUs, and are now relying on GPUs, will incorporate

TPUs, and field-programmable gate array (FPGAs) and application specific integrated circuit (ASIC) accelerators to continue to drive performance improvements, for the majority of their computation. Hence the data center is entering the age of heterogeneous computing.

It should be noted that 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. For example, NVLink and NVSwitch provide excellent intra-server interconnect speed, but they can only connect intra server, and only with NVIDIA-branded GPUs. Liqid can connect servers to accelerators or storage via PCIe, but server to server communication requires paying a composing penalty through InfiniBand or Ethernet. In contrast, FabreX is completely hardware and software agnostic and can connect any resource to any other over PCIe, including server to server, of any brand.

The missing element and the holy grail in the composable infrastructure play has been memory. Memory itself has also 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.

A major problem today in the movement to heterogeneous computing are concerns regarding resource utilization and stranded resources. For example, GPUs -such as NVIDIA V100s -, and SCM memory/storage pools - such as Intel Optane and Micron 3DXP-, are typically the most expensive discrete resources in the datacenter. But prior to CDI, these resources were held captive within the server and were not readily sharable outside the server. 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 requires increasing the numbers of servers. Hyperscaler datacenters such as AWS and Azure 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. Hyperscaler 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 utilization and broad availability of otherwise stranded resources.

Past Barriers

Early barriers to composable disaggregated infrastructure (CDI) included a lack of suitably fast I/O networks. Prior to GigaIO, only inter-process 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, and only internal bus interconnects had been implemented. Importantly, there was no 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 sheet metal boundary was yet another barrier to disaggregation. 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 in Penguin systems. Importantly, the increasing demand for more compute and storage devices, driven by the rise of machine learning and deep learning systems,

[1]Li K, Hudak P: Memory coherence in shared virtual memory systems. ACM Trans Computer Systems 7(4):321-359, 1989.

highlighted 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.

The Goal

Fully composable disaggregated infrastructure requires 100% disaggregation, re-aggregation capability without a performance penalty, a memory-coherent I/O communication network, and enterprise-class dynamic device discovery, composition, and control. The ideal architecture includes the ability to disaggregate and compose servers, or (in the future) 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.

The new Compute Express Link (CXL) standard for CPU and I/O device communication will finally enable the development of memory-coherent I/O networking. GigaIO's high-speed switched I/O fabric, based on PCIe 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 CDI.

The ideal CDI also requires enterprise-class remote management via standard APIs, such as Redfish, to enable dynamic composition by any and all of the emerging composition management software systems such as Bright, Slurm Grid Engine, VSphere and the hyperscalers' own VM software. 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.

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, Dolphin "device sharing", and the Liqid partial (fan-out) I/O network. 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 before –"the composition tax"- communication with the devices in the rack. InfiniBand does not provide rack-level independent device discovery and does not provide memory coherence. Dolphin and Liqid are not full fabrics (they are fan-out only), and do not support server-server cluster communication/composition or GDR cross-server GPU communication. Further, Dolphin does not provide rack-level device discovery of unassigned devices (e.g., unassigned servers) and does not have visibility into the multiple device trees in the rack. Liqid does not provide rack-level device discovery of unassigned devices. Neither Dolphin nor Liqid provides memory coherency.

Current State of the Art

GigaIO FabreX provides 100% disaggregation with full granularity of composition. The 100% disaggregation provided by GigaIO FabreX includes CPUs on multiple servers, multiple memory servers, multiple enclosures of 8–16 accelerators each (GPUs, FPGAs, ASICs), and multiple enclosures of dozens of storage devices. With CXL, GigaIO FabreX will support memory enclosures as well (e.g., JBOMs). 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.

The GigaIO FabreX communication network infrastructure is deployed via 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.

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

Regarding CXL/PCIe Gen5 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 Gen5, 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 early in the year. Planning and support for PCIe Gen5, with CXL messaging, is already underway, with an expected completion date of Q3 2021, assuming CXL chips are available in Q2. Actual deployment will be driven by availability of silicon and speed of CXL adoption and support by the entire ecosystem (CPU 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 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. 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, 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. GPU developers such as NVIDIA and AMD 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.

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, no coherency policy management will be required. 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.

About GigaIO

GigaIO was established in 2016 by networking industry veterans with decades of domain expertise in communications, data centers, high-performance computing, open source, and infrastructure management. The company is headquartered in Carlsbad, CA, and home to more than 30 staff members, most of whom are engineers with advanced degrees and more than 400 years of combined industry experience.

GigaIO has invented the first truly composable software-defined infrastructure network, empowering users to accelerate large-scale workloads on-demand, using industry-standard 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™ 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.

