GIGAIO

PRIMER

FabreX PCIe network fabric: a primer

CONTENTS

A BIT OF HISTORY **3**
PAST ATTEMPTS HAVE LARGELY FAILED **4**
GIGAIO'S INNOVATION **4**
SOLVING THE BIOS LIMITATIONS **5**
THE PATH FORWARD TO CXL **5**
CONCLUSION **6**
ABOUT GIGAIO **7**

February 2021 | Rev 1.0

At GigaIO we are staunch believers that the best *network* today for HPC workflows is PCIe, and that the future belongs to *network fabrics* based on PCIe and its future incarnation in Compute Express Link (CXL). This primer outlines how GigaIO's technology has successfully transformed what used to be a bus protocol – PCI – limited to usage within the server chassis, into a fully featured network fabric outside the box, and uniquely prepared for CXL.

In this paper we will refer to PCIe but many of the underlying technology described here are also applicable to CXL.

FabreX Fabric Manager is the only *routable PCIe network* solution on the market today, with the industry's lowest latency and highest performance not only from a server to its resources, but also from a server to other servers throughout the rack, and beyond. How is that possible?

Let's dive into what PCIe is, and what makes our implementation so ingenious.

A bit of history

The Peripheral Component Interconnect Express (PCIe) technology was first introduced in 2003 and originated as an interconnect for processors to communicate with I/O devices. It is a point-to-point communication protocol unlike its predecessor PCI, which was a bus topology with I/O devices hanging off it. Unlike PCI, which is a parallel bus with its own address, data and control signals, PCIe is a serial bus with all of these communication signals embedded in one serial link.

An attribute common to both PCIe and PCI is the ability to make all peripheral components, i.e. I/O devices, operate as Plug-n-Play. This is important because it means any PCIe device can theoretically be used and swapped in and out. This was done through establishing a hierarchy of devices, with one source under which all I/O devices lie, referred to as a PCIe tree. The process of discovering all of the I/O devices in this tree at server boot time is called enumeration, and is invoked by the BIOS, which is an exclusive piece of code executed by the processor.

Lucky for us, PCIe also introduced an extremely robust associated communication protocol. The industry took the best of all communication protocols prevalent at the time PCIe was specified, and incorporated those features within the protocol.

PCIe was originally conceived of as an interconnect within the server, and today it is the common interface for computers' graphics cards, hard disk drives, host adapters, SSDs, and Ethernet hardware connections. All computer resources thus "speak" PCIe natively.

Past attempts have largely failed

Over the years, since it is such a robust communication protocol, and offers the absolute lowest network latency, many attempts have been made to take PCIe out of the chassis and transform it a network fabric. Most of these attempts failed until GigaIO, because the PCIe hierarchy mentioned above was unique to a processor, and exclusively under its domain; this made it practically impossible to communicate between processors and devices belonging to two different hierarchies.

Another issue was that the BIOS of servers varies widely in their ability to enumerate I/O devices, even within the same server model. As such the ability to build a PCI tree off any given server was, and remains to this day, mostly unpredictable and in many cases made the enumeration impossible beyond what was connected on the motherboard.

This is the reason why all other composable infrastructure vendors either a) sell you a completely proprietary hardware and software solution where they control all the elements in the stack, or, b) produce a short and limited list of supported server models and BIOS versions, and their solution will simply not work with anything not on the list.

GigaIO's innovation

The magic in FabreX, and the reason it is succeeding where others have failed, is fundamentally 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 networking technology in the market, and uses using a scheme called 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, itself requiring a built-in intelligence unit hosted in the switches which comprise the FabreX interconnect network.

This routing mechanism virtualizes all hardware resources comprising 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, 3D-Xpoint storage – anything with a PCIe connection).

Solving the BIOS limitations

And what about the BIOS issues you ask? Well, GigaIO cannot solve the issue of variability between servers, but because we have invented full end-to-end native PCIe communication for server to server connectivity, without resorting to Ethernet or InfiniBand, we are not limited by those constraints in the same way as others have been in the past. This is how we have cracked this nut:

- For storage, we can use NVMe-oF over native PCIe. In that configuration, servers can effectively "borrow" BAR (Base Address Register) space from other servers to enumerate more resources. So, server #1 becomes an NVMe-oF initiator, not using any of its own BAR space, server #2 is the NVMe-oF target, with for example Optane SSDs, and server #1 reaches over FabreX and "sees" the SSDs as its own storage. Server #2 can compose and create volumes to any other server on the network., with negligible latency penalty because all communication is DMA over PCIe, and BIOS limitations on server #1 have been circumvented.

- For GPUs, in a similar fashion, GigaIO can use GDR over PCIe using NCCL rings. Even if the BIOS of a particular server is limited as to how many GPUs it can enumerate, we routinely use NCLL rings of 16 GPUs across several servers, and in some cases up to 32 GPUs. With AMD or Intel GPUs, which don't force the developer to go through the kernel instead of p2p as NVIDIA does, there is no software overhead and de facto the BIOS limitations are nullified. Even with NVIDIA GPUs, depending on the application, staying with PCIe to run the NCCL rings generally delivers similar performance across servers as if the GPUs were direct attached.

While other vendors will tout they can do the same, the big difference is they will have to introduce another network in the rack, to run NVMe-oF or GDR over Ethernet or InfiniBand, in order to go server to server. This doubles the number of HBA cards in each server, doubles the number of switches, and introduces latency: more money for less performance, or more pithily, as we like to say: ***"pay more, get less".***

The path forward to CXL

Our own network fabric, FabreX™, currently uses PCIe as its data plane, but the management plane, which is what constructs the network, will stay the same as we move the data plane to CXL. The interface into the management plane is industry standard Redfish APIs, which offers a seamless integration path for system and software developers to integrate the new benefits and features that CXL will bring as a data plane. Our technology enables us to uniquely mix the two under a single management plane, making the transition

from one to the other over time far simpler for the ecosystem and accelerating CXL adoption.

The new 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 disaggregation and device pooling promised in composable disaggregated infrastructure (CDI).

The future of application- and workflow-defined infrastructure 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.

Conclusion

The reason others failed in building PCIe-based networking where GigaIO is succeeding, is that they took a bottoms-up approach to solving this issue, whereas GigaIO took a top-down approach. GigaIO is uniquely positioned to take advantage of and thrive under the new CXL deployment, because our management plane will stay the same and facilitate broad adoption by the software and developer ecosystem.

For more details about CXL see the position paper "Composable Disaggregated Infrastructure with CXL/PCIe Gen 5"

For more details about our implementation of disaggregated infrastructure, read the primer on "Rack-Scale Composable Infrastructure"

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.

