The GigaIO™ FabreX™ Network

New Frontiers in NVMe-oF

CONTENTS

TABLE OF FIGURES

Introducing GigaIO FabreX

GigaIO FabreX is a fundamentally new network architecture that integrates computing, storage and communication I/O into a single-system cluster fabric, using industry standard PCI Express technology. GigaIO FabreX enables server-to-server communication across PCIe and makes true cluster scale networking possible. The result is direct memory access by an individual server to system memories of all other servers in the cluster, creating the industry's first in-memory network. The GigaIO FabreX architecture enables a hyper-performance network with a unified, software-defined, composable infrastructure.

GigaIO FabreX provides multiple benefits for composable hardware infrastructure, whether in the cloud or in private data centers, including (1) native PCIe communication between hundreds of heterogeneous computing, storage and accelerator. Additionally, FabreX delivers the industry's lowest latency and highest effective bandwidth together with increased flexibility and increased utilization, thus reducing Capex and Opex. GigaIO FabreX minimizes latency by (1) eliminating the entire translation layer required by other networks and (2) enabling direct memory access and sharing of all connected processors and memory with point-to-point connections between any two devices, thus improving overall system performance, throughput and reduced time to solution. In addition. FabreX will dramatically improve utilization and reduce total cost of ownership.

NVMe Overview

NVMe (nonvolatile memory express) is a scalable host controller interface designed to address the requirements of client and enterprise systems that utilize PCIe (peripheral component interconnect express) based solid state storage.[1] Escalating demand for storage solutions mandates that storage transports must provide enhanced overall performance. These escalating demands mirror increases in big data volume and big data velocity, and the quantity and complexity of devices participating in the Internet of

---

[1]NVM Express Organization: NVM Express Overview — http://nvmexpress.org/wp-content/uploads/NVMe_Overview.pdf

Things. The NVMe specification was developed to allow solid state drives (SSDs) to perform to their maximum potential, by eliminating dependencies on legacy storage controllers. The NVMe interface specification enables fast access for direct-attached SSDs. Instead of using an interface such as SATA (serial advanced technology attachment), NVMe devices communicate with the system CPU using high-speed PCIe connections. NVMe provides support for up to 64K input/output (I/O) queues, with each I/O queue supporting up to 64K commands. As well, the NVMe interface provides support for multiple namespaces and includes robust error reporting and management capabilities. NVMe is a much more efficient protocol than SAS (serial attached small computer system interface) or SATA. Overall, the NVMe command set is scalable for future applications and avoids burdening the device with legacy support requirements.

NVMe-oF Introduction

NVMe-oF (NVMe over fabrics) is a network protocol utilized for communication over a network between multiple hosts and various storage subsystems (Figure 1). NVMe-oF uses NVMe as the storage protocol with the goal of extending NVMe onto fabrics such as TCP, Fibre Channel and RDMA (remote direct memory access) technologies including RoCE, iWARP and Infiniband®. Consequently, the benefits of NVMe storage such as low latency are available regardless of the fabric type, making possible large-scale sharing of storage devices.
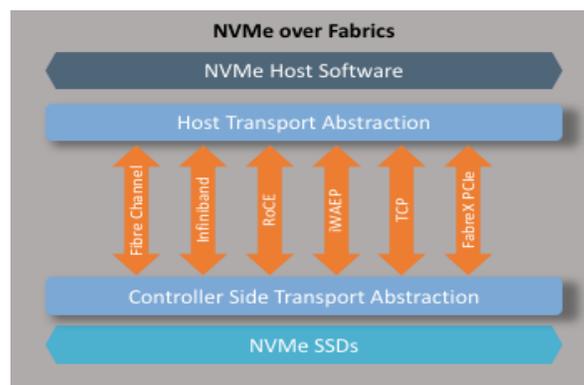
Figure 1 -- NVME-oF topology.

NVMe-oF defines protocol handling over these different fabrics. Fabrics are network protocols with support for reliable connectivity, such as RDMA, Fibre Channel, TCP,

and now GigaIO's FabreX. NVMe-oF eliminates almost all processing that occurs with NVMe command handling within existing data center networks. NVMe-oF communications can occur at speeds that approach the speed of directly attached NVMe communications. Consequently, the NVMe-oF protocol is much faster than the iSCSI (internet small computer system interface) protocol.

NVMe-oF Scaling and Performance

NVMe-oF enables scaling out to numerous NVMe devices and extending the distance within a data center over which NVMe devices and subsystems may be accessed.[2] Importantly, the goal of NVMe-oF is to provide NAS (Network Attached Storage) connectivity to NVMe devices at performance levels comparable to DAS (Direct Attached Storage). The NVMe standard improves access times by several orders of magnitude compared to implementations using SCSI and SATA devices.

NVMe-oF is designed to take advantage of NVMe efficiency and extend it over fabrics. Overall, CPU overhead of the I/O stack is reduced, and implementation of NVMe is relatively straightforward in a variety of fabric environments. Additionally, the inherent parallelism of NVMe I/O queues makes NVMe capable of implementation in very large fabrics. Regarding transport mapping, commands and responses in a local NVMe implementation are mapped to shared memory in a host over the PCIe interface. The NVMe-oF standard allows transport binding for supported networks to utilize both shared memory and message-based communication between host and target systems. GigaIO FabreX is a PCIe-based network and therefore utilizes only shared memory communications, resulting in the lowest latency and highest throughput compared to the other NVMe-oF offerings.

NVMe-oF maintains NVMe architecture and software consistency across fabric designs, providing the benefits of NVMe regardless of fabric type or the type of nonvolatile

---

[2]NVM Express Organization: NVM Express Over Fabrics Overview — http://nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf

memory utilized by the storage target. It is anticipated that there will be full integration of NVME-oF into the NVMe specification in 2020.

GigaIO's FabreX NVMe-oF Implementation

The NVMe standard allows a high level of parallelism of commands being handled over multiple I/O queues. The GigaIO FabreX implementation of NVMe-oF extends this parallelism over the network, enabling direct data exchange between nodes and remote system memory with zero copy, and the lowest latency and highest throughput. Thus, NVMe over FabreX extends the low-latency, efficient NVMe block storage protocol over fabrics to provide large-scale sharing of storage over distances — FabreX NVMe-oF delivers NAS access at DAS speeds.

Transport binding specifications describe how data are exchanged between the host and target subsystem and establish protocols for connectivity and negotiation (Figure 2).
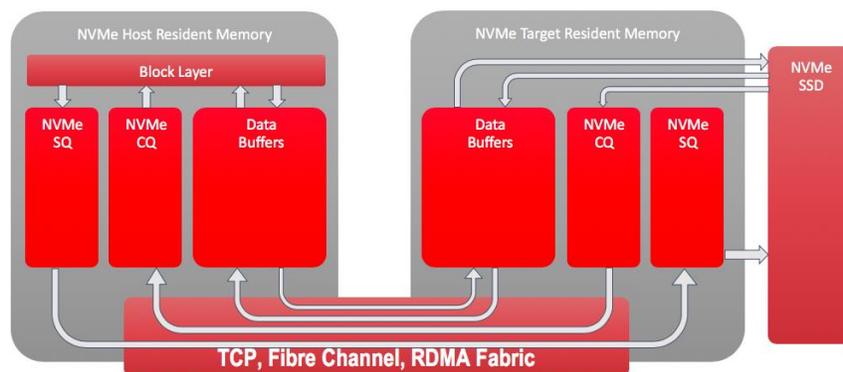

*Figure 2 -- N*VMe-oF with Current Fabrics

GigaIO has developed NVMe-oF for the FabreX network, defining and implementing the NVMe-oF transport binding for GigaIO FabreX. The GigaIO FabreX implementation fully supports the principles and requirements of the fabric NVMe-oF specification. By incorporating transport support, GigaIO FabreX enables faster communication between NVMe hosts and NVMe target subsystems. A unique feature of GigaIO FabreX and NVMe-oF is zero copy data movement between local host memory and remote NVMe drives. No other fabric provides this critical attribute. Zero copy data movement results in the industry's lowest latency and highest throughput. In addition, CPU resource

utilization is reduced, freeing the CPU for other activities. Benefits include handling more loads on lower-power devices and obtaining the same performance on lower-priced systems, thus reducing total cost of ownership. Additionally, with GigaIO FabreX there is no protocol loss or fabric-specific bandwidth loss, as the entire implementation utilizes PCIe. Thus, GigaIO FabreX instantiates a low latency, high bandwidth network, and supports resource-hungry loads on systems that use NVMe-oF. By providing the lowest latency and highest throughput, GigaIO FabreX outperforms all other fabric types and all other transports that are included in the current NVMe-oF specification (Figure 3).
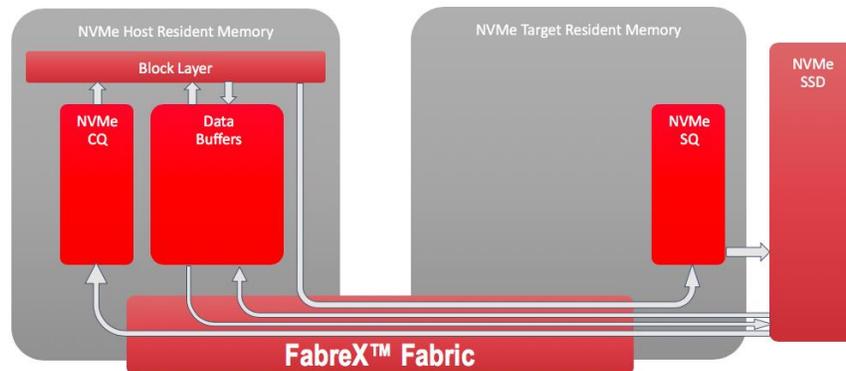
Figure 3 -- *NVMe-oF GigaIO FabreX Implementation*

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 15 years of industry experience. GigaIO develops innovative, high-performance interconnect network for computing clusters, with the objective of accelerating large-scale workloads on-demand, using industry-standard technology. GigaIO FabreX eliminates conversion layers, maximizes throughput, and enables data centers to run at full efficiency and obtain outstanding performance. The company's patented network technology facilitates development of broad and deep network architecture. GigaIO's extreme connectivity for high-end computing delivers optimized resource utilization and reduced total cost of ownership. For more information contact the GigaIO team at info@gigaio.com or visit www.gigaio.com.