TECHNICAL NOTE

The GigaIO™ FabreX™ Network

Technical Note

FabreX Rack Scale Gateway

CONTENTS

TABLE OF FIGURES

June 2019| Rev 1.0

Overview

The term Rack Scale System conjures up different images in different people. However, as the name suggests, it is essentially an embodiment of a computing system with all of the standard resources constituting a typical compute ecosystem housed within a single Rack. The functionality provided by the Rack in its totality is called ‘Cluster’ in today’s parlance. However, there can potentially be multiple Clusters in a single Rack with each owning its own set of hardware resources.

Purpose

The output of a Rack Scale System is either designed to produce the final outcome of a specific application or provide a construct of data to be shared with one or more Rack Scale Systems to get the final outcome in a collaborative manner. Incidentally, the number of Racks in this context can be as many as practical.

The trend towards building efficiency in the IT infrastructure world is to integrate Clusters of computing islands to form a synergistic work flow. These Clusters are usually resource rich and finely tuned to deliver optimal performance. This situation in many scenarios involves a collection of Racks working in unison interconnected via a highly efficient and flexible communication link.

The Basic Architecture of Rack Scale System

Figure 1 depicts an example of a Rack Scale System made up of a variety of hardware resources. The Compute Servers are not restricted to x86 based platforms and alternately can be FPGA, ASICs with embedded SoC, and a variety of other Computing engines.

The GPUs shown can, instead of being traditional devices, be AI processing elements espousing serial or parallel computing architecture.

The NVMe-oF Target servers provide the ability to feed the Application Servers with vast amounts of data from NVMe Drives under its purview.

Data stored in NVMe Drives can be extracted by the GPU servers on demand to constantly feed the slew of GPU cores for them to process.

In Rack Scale architecture there is potentially the need to have multiple Compute Servers iteratively processing data in a serial manner whereby output of Compute or GPU server is fed to another compute or GPU server or in a loop fashion to iteratively process the data till the desired final outcome is achieved.

One point to note, the descriptions of the individual tasks mentioned are primarily defined by the specific application that needs to be executed. This in turn will define the resources required and the resulting architecture of tying them together. This composition is carried out by the orchestration software of the FabreX solution offering.

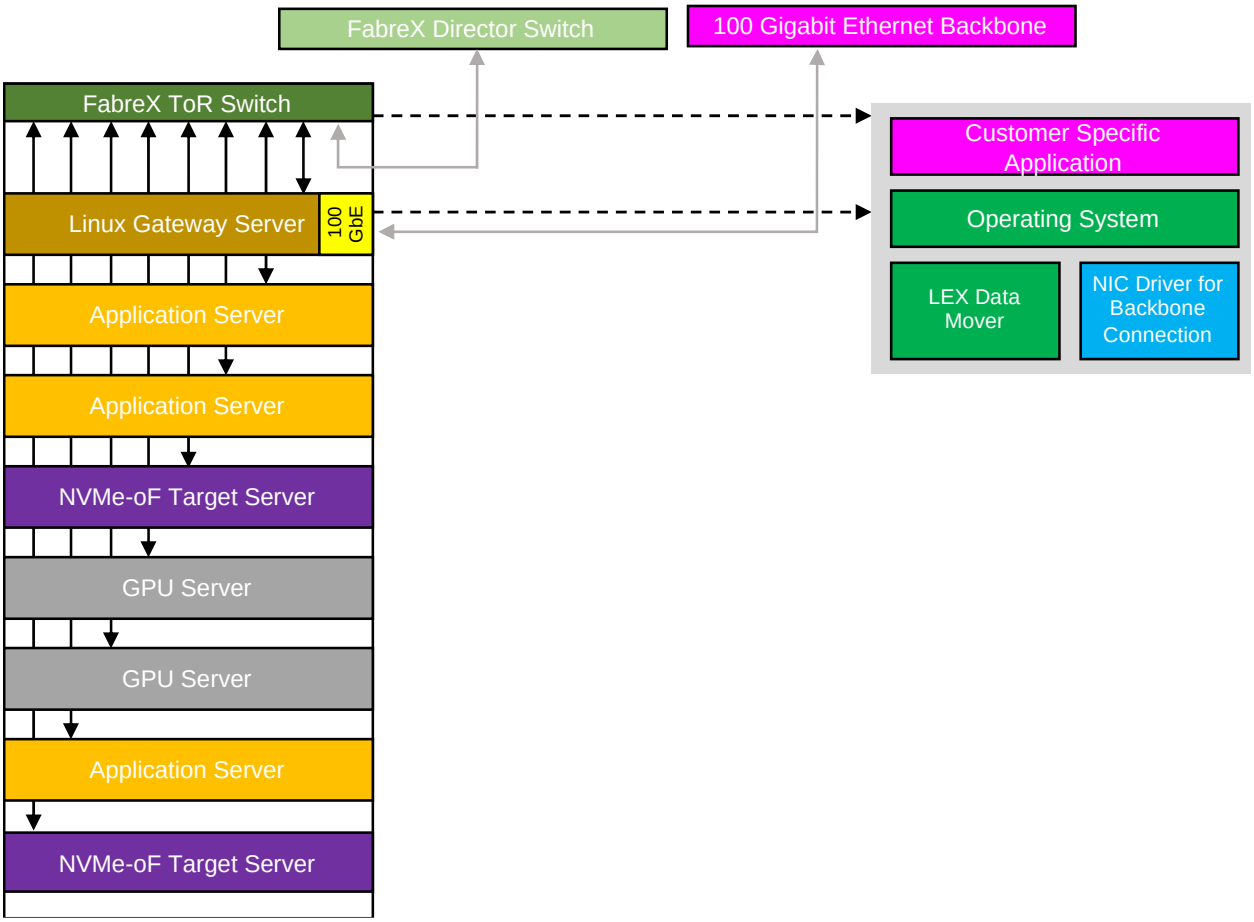

Figure 1 – Rack Scale System Example

All of the resources shown in the Rack of **Figure 1** are attached to each other with the FabreX fabric via the Top of Rack (ToR) FabreX switch providing the lowest latency and the highest bandwidth of interconnect paths.

The Requirement of Multiple Racks

In many applications a single Rack Scale System may not suffice to derive the final desired result of an application job. In these scenarios it may require multiple Racks to share the workload.

In this case, one Rack Scale System does some specific job whose output needs to be passed on to the next Rack Scale System to perform some of the complementary operations towards the goal of deriving the final result.

However, in multiple Rack Scale Systems it is very seldom the case where all of the respective resources within the individual Racks are configured to be attached as nodes to one homogeneous network fabric. Moreover, this type of configuration would surely fall outside the scope of a Rack Scale System.

Figure 2 shows multiple Racks interconnected via FabreX. This will undoubtedly be the most efficient mode of attaching multiple Racks constituting its respective Cluster function. This will allow for the lowest latency and highest bandwidth of the interconnect path between Racks.

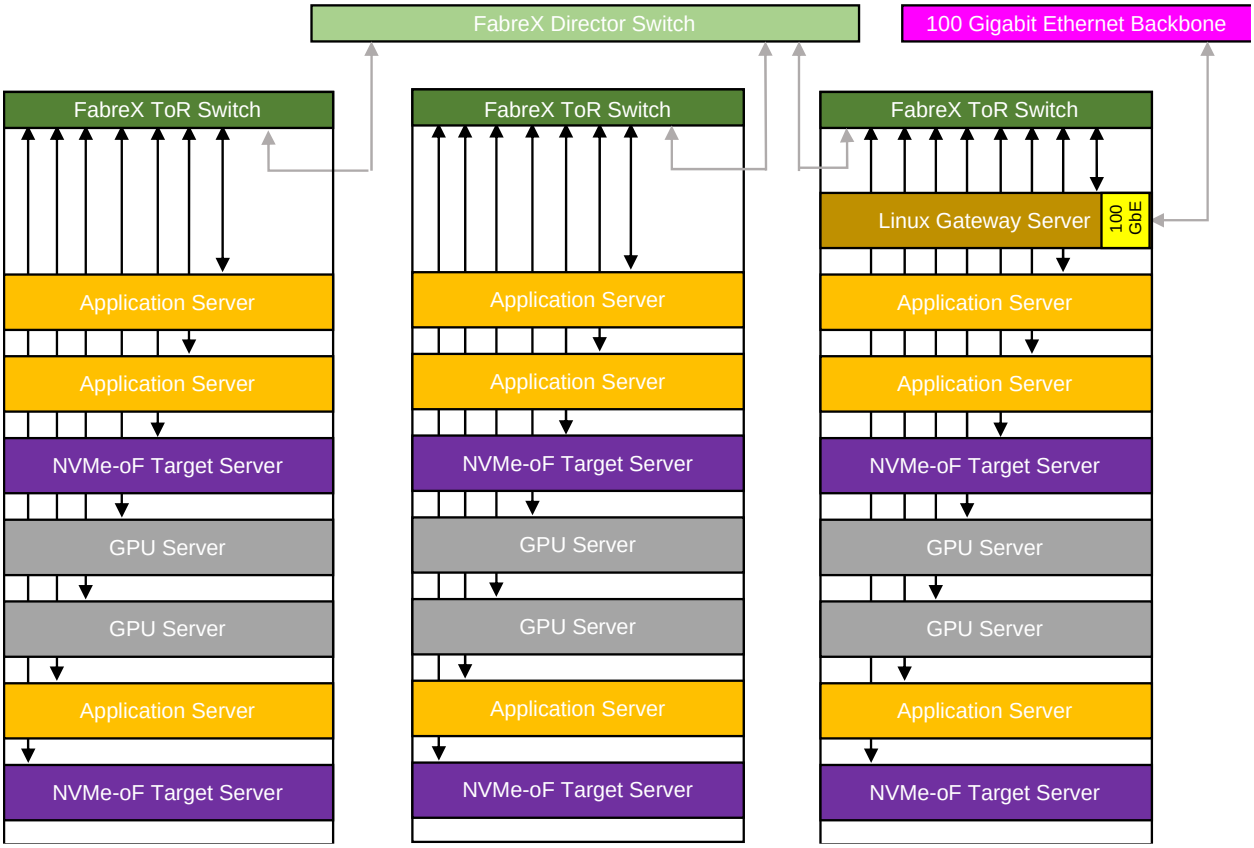

Figure 2 – Multiple Rack Connected via FabreX

However, the data constructs may be somewhat different when using FabreX as the interconnect between Racks since the Cluster boundaries get somewhat blurred; for instance, an integrated 3 Racks can serve as one Cluster as compared to individual Racks constituting a Cluster.

This situation is more likely the case for FabreX since its architecture allows for resources attached to it being interpreted as respective memory resources in a homogeneous flat 64-bit Virtual Address Space. Consequently, the position of these resources as per its physical location in a Rack becomes somewhat irrelevant.

This figure shows an optional resource, namely a Gateway Server, installed in one of the Racks to provide the connectivity between the Racks and the legacy network backbone.

Figure 3 depicts an alternate way of providing interconnect between the Clusters using the legacy network.

This depicts a Gateway Server installed in every Rack to provide the connection function between the FabreX within the Rack and the legacy network. However, in many applications a single Cluster function embedded in one Rack would suffice to address the user's application needs and consequently not require this configuration.

In contrast to the previous example of FabreX Director Switch providing interconnect between Racks the functions of a Gateway Server for providing the bridging function is more defined for all of the reasons already mentioned.

In this configuration it is imperative every Rack has a Gateway Server installed to provide the connection function between FabreX and the legacy network.

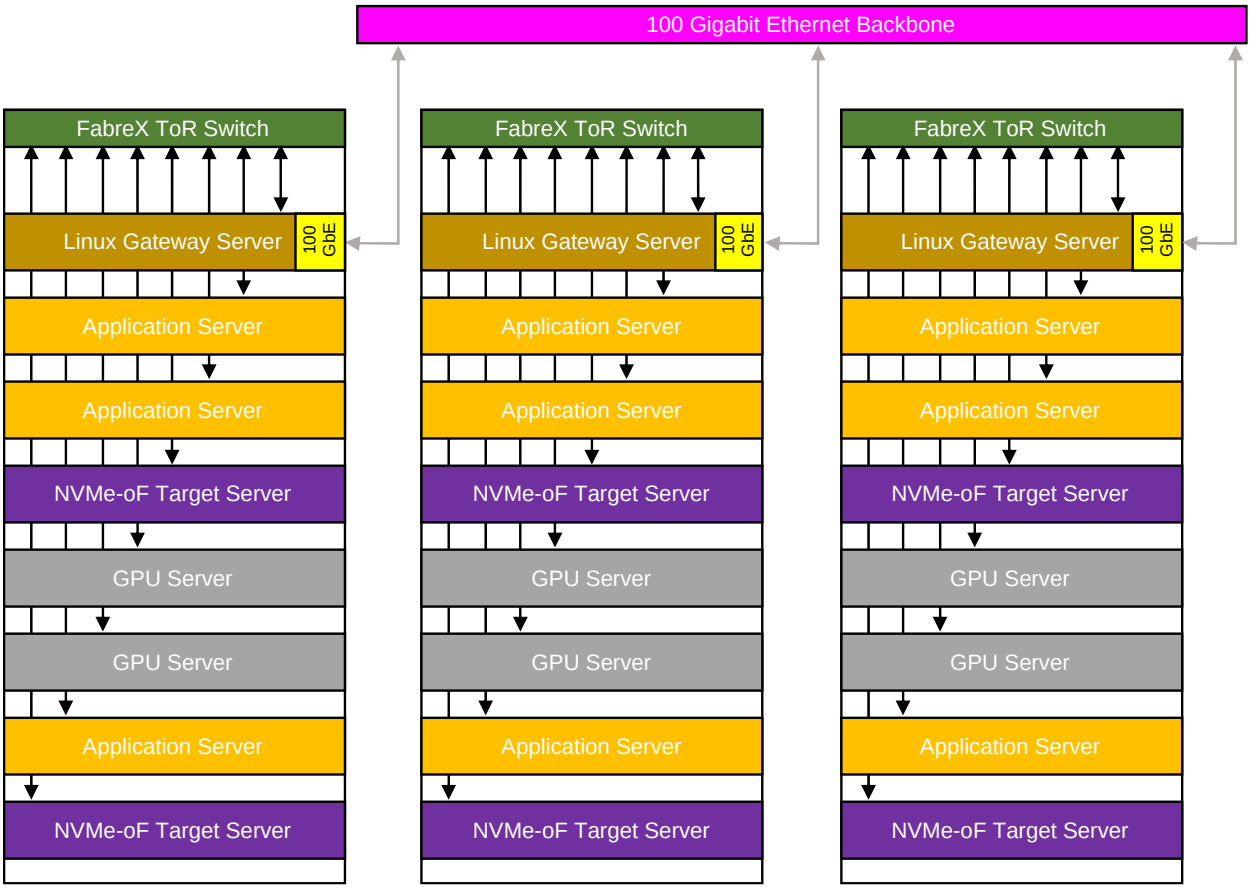

Figure 3 – Connectivity to Legacy Network Example

The Gateway Server

Figure 4 is a conceptual diagram showing the data paths between the Application Servers and the buffer used as the holding space for the data to be transferred to the legacy network of choice.

It is envisioned there exists a proprietary rule-based protocol and a resultant data structure implemented by the Application Servers to communicate with other Clusters running on other Racks and/or to some other entity attached to the legacy network backbone.

However, regardless of its final destination, these data constructs and protocols or some flavor of it is already prevalent in the Application Servers of customers' Racks since this is part and parcel of their overall system architecture.

The Gateway Server in this case will transfer the data resident in the holding buffer by packetizing the data in the legacy network's frame structure and send it to the desired Rack or entity attached to the legacy network.

It is envisioned the metadata associated with the data to be transported has the information on destination node attached to the legacy network.

A similar sequence of steps take place when a Gateway Server receives data from another Rack destined for a specific Application Server in another Rack.

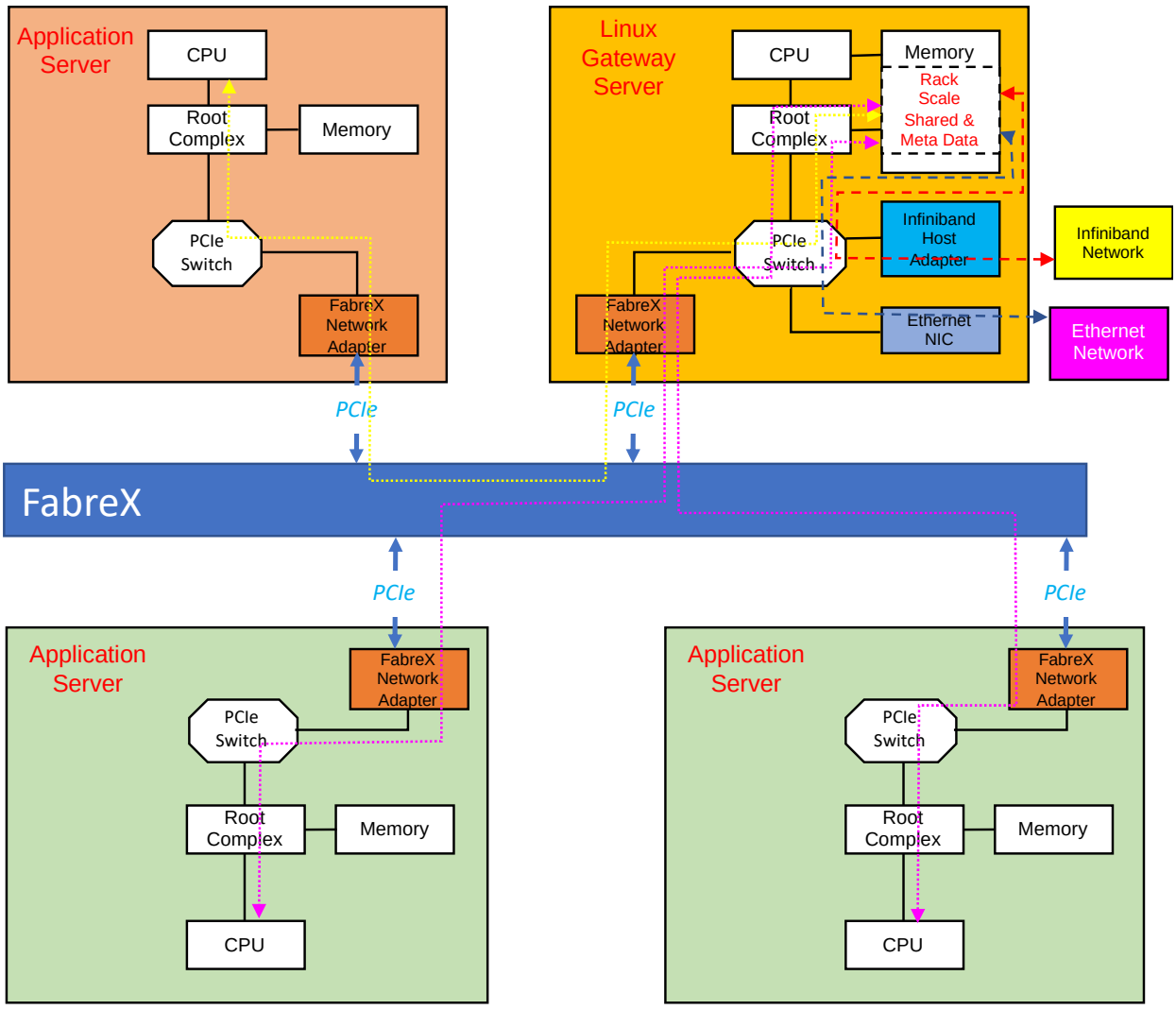

Figure 4 – Data Paths Between Application Servers and Legacy Networks

GigaIO Service

GigaIO in conjunction with the customer can provide the driver for the Application Servers to access the holding buffer space of the Gateway Server.

In addition to this, GigaIO can provide the driver of the NIC and/or HBA to transfer the data to the legacy network from the holding buffer in the Gateway server.

GigaIO can standardize the driver of the Gateway Server and document the API to interface to the holding buffer. This way, the driver in the Application Servers can be designed to interface to this standard API.

This approach will confine the software task for a new customer to only developing the driver for the customer's Application Server to interface to a standardized GigaIO's Gateway Server.

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.