1 Introduction

With today's demanding backplane requirements, the era of GbE as the de-facto backplane interconnect is coming to a close. As such, a number of interconnect technologies are vying to replace GbE with the top contenders being 10 Gigabit Ethernet (10GbE), Infiniband (IB) and PCI Express (PCIe). Though a clear winner has not yet emerged as the ideal backplane interconnect, PCIe, with its advanced capabilities, makes a strong case for becoming the backplane interconnect solution.

Over the last decade, PCIe has evolved from functioning merely as the transport for single host to IO device communication model wherein one host manages a set of IO devices. Today, PCIe can easily support an efficient host to host communication model as well as other configurations that include IO resource sharing across multiple hosts. Such features lead to a significant cost and complexity reduction. In addition, mainstream processor companies like Intel have been integrating PCIe, not just in their chipsets, but as an integral part of the core silicon. With such inherent advantages over that of 10GbE or IB, we believe that PCIe can indeed fill the mantle of being an ideal backplane interconnect.

2 Backplane Choices

A fundamental backplane requirement is obviously the need for high performing fabric – in the form of throughput (10Gbps+) and low latency (< 5µs). Said interconnect must also support backplane distances not only for deploying bladed environments (e.g. blade servers) but also for cabling across multiple blade chassis or potentially supporting rack mounted servers.

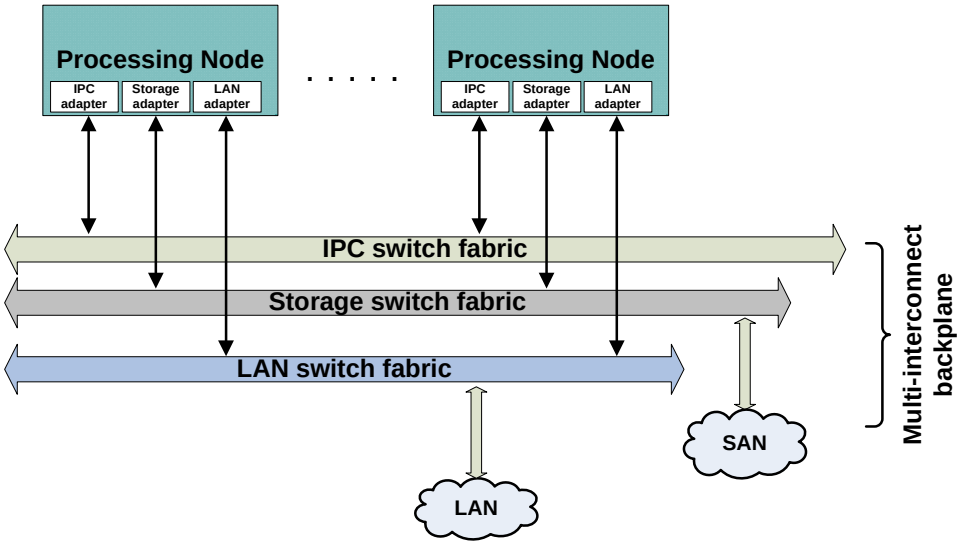

Figure 1. Traditional backplane for supporting IPC, LAN and SAN connectivity.

From a functionality point of view, the backplane must support inter-processor communication (IPC) as well as access to an external local area network (LAN) and a storage area network (SAN). Traditional approaches use three different IO interfaces on each server to accomplish this. Consequently, three different backplane interconnects are required for supporting the IPC, LAN and SAN communication model in the backplane.

Figure 1 shows a traditional backplane where a server uses a GbE interface for LAN connectivity, a Fiber channel (FC) card for SAN connectivity and 10GigE or IB based card for IPC connectivity. Clearly, this is not the optimal nor is it the preferred model from both a cost perspective and complexity level. The need for a unified backplane that supports all three types of traffic wherein each server connected to the backplane uses a single IO interface instead of three is obvious.

There are three worthy candidates ready to make the claim for a unified backplane: PCI Express, 10GbE and Infiniband. Each of the three technologies provides features and enhancements designed to support the unified backplane model. But as it will become apparent in the subsequent sections, PCIe has the edge over 10GbE and Ib.

2.1 10Gigabit Ethernet

Ethernet is used now for both LAN as well as IPC. However, Ethernet is not a viable candidate for carrying storage traffic due to its intrinsic tendency to drop packets in the wake of congestion. Data loss and corruption in storage systems is simply not an option. To work around it, upper layer protocols, such as TCP/IP, must be used in order to provide a reliable communication over Ethernet notwithstanding the inherent overheads associated with them.

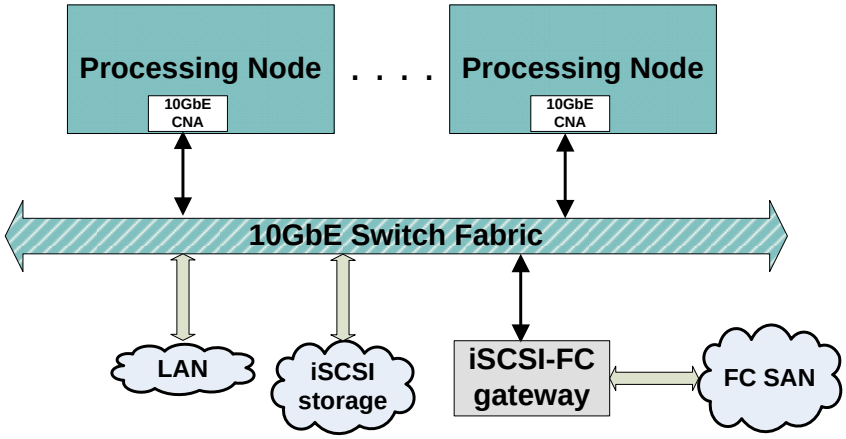

Figure 2. Supporting IPC, LAN and SAN with 10GbE

From a storage perspective, the iSCSI layer built atop TCP/IP provides a safe mechanism for transferring SCSI storage commands and data over the unreliable Ethernet fabric. The complexities of the iSCSI protocol impose the use of a dedicated iSCSI adapter at the server. The purpose of the iSCSI adapter is primarily for handling the storage communication that occurs in the Ethernet fabric which is a contradiction to what a unified server IO interface should be. In addition, the use of iSCSI as a storage transport allows servers to interface with only iSCSI-based storage arrays. Support for other storage infrastructure, based on Fiber Channel (FC) for example, forces the use of an iSCSI-to-FC gateway to terminate the iSCSI protocol and transfer the underlying SCSI commands using the native FC protocol (See Figure 2). The use of gateways has its own drawbacks including low performance due to the need for TCP/IP protocol termination/translation and the increase in cost due to additional hardware.

Recent efforts in the Ethernet community have been made in order to improve the reliability of Ethernet by incorporating a flow control mechanism for preventing packet drops. Converged Enhanced Ethernet (CEE), as it has been named, obviates the need for using the high overhead TCP/IP protocol and also allows the FC protocol to be run directly over Ethernet (Fiber Channel over Ethernet or FCoE). But as with the iSCSI approach, the full benefits of FCoE can be realized only when the protocol is offloaded to hardware rather than done in software. Indeed, the new and upcoming converged network adapters (CNA) provide such hardware capabilities. These CNAs provide dual capabilities and function as both an FCoE adapter for transporting FC packets and a traditional 10GigE adapter for transporting TCP/IP packets.

However, the CEE and the FCoE models are still in its infancy and its wide-spread adoption is debatable. Furthermore, most of the underlying storage infrastructure do not provide a native FCoE frontend but are based on traditional FC instead. To fully accommodate the many storage systems currently deployed using FC in this model, a new type of FCoE switches that can handle both FCoE as well as TCP/IP traffic (See Figure 3) needs to be developed. Overall, it is unclear how the CEE approach which claims to extend Ethernet and yet requires significant modification to the legacy Ethernet ecosystem could be considered any different from other specialized interconnect such as Infiniband.

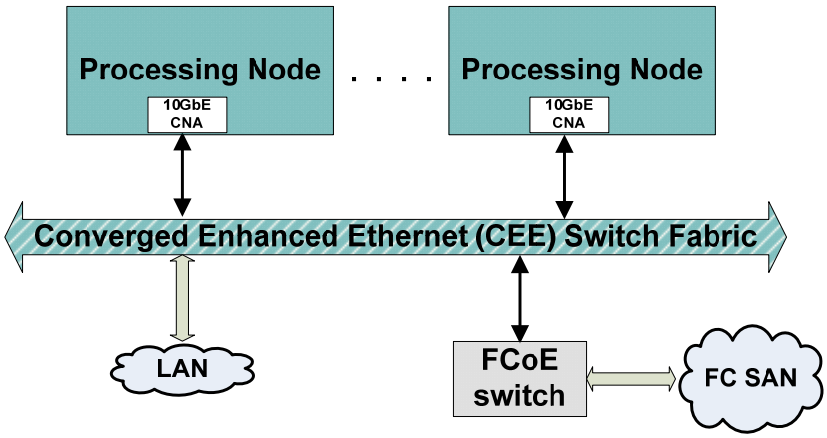

Figure 3. Supporting IPC, LAN and SAN with Converged Enhanced Ethernet (CEE)

2.2 Infiniband

The strengths of Infiniband (IB) lie in its support for high throughput and low latency form of communication. Infiniband is suitable for IPC and is indeed a commonly deployed interconnect in the arena of high performance computing (HPC). However, when the need for LAN or SAN connectivity to an IB fabric is taken into consideration, there are some disadvantages. For LAN connectivity, servers must use the TCP/IP over IB protocol (IPoIB) and must go through an IPoIB gateway – one that serves as a bridge between IB and the LAN.

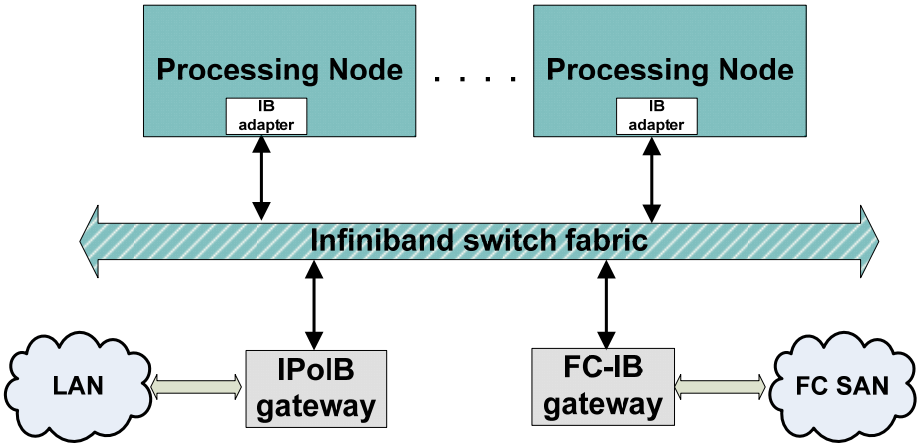

Figure 4. Supporting IPC, LAN and SAN with IB

As is the case with LAN connectivity, a gateway is required for providing SAN connectivity. When servers need to connect to FC-based storage, a gateway in the form of a FC-to-IB bridge must be used (See Figure 4). Note that the iSCSI model described for Ethernet earlier also requires an iSCSI-to-FC gateway (Figure 2). But the availability of iSCSI-attached storage implies that iSCSI-to-FC gateways are not required for native iSCSI-based storage deployments. In contrast, the near absence of native Infiniband attached storage implies that a gateway is needed with most certainty for Infiniband based storage deployments. Unlike iSCSI storage deployments, the deployments for storage using Infiniband have been rather minimal.

2.3 PCI Express

PCI Express (PCIe) is no longer restricted as the communication interface between the CPU and IO subsystem within the confines of a single server platform. Since its introduction, PCIe has evolved considerably and adopted innovative features which can now support other advanced usage models. This includes support for virtualization wherein a particular IO adapter can be shared by multiple virtual machines (VMs) running on a single server. In this model, each VM gets access to its own set of hardware resources on the adapter. In addition to supporting efficient virtualization in the context of a single host, additional enhancements to PCIe allow the IO adapter to be disaggregated from the servers such that a pool of IO adapters can be shared across VMs running on different hosts.

Figure 5 illustrates a virtualization environment in which multiple processing nodes share a single 10GbE and a FC adapter with the PCIe switch fabric providing the requisite support for IO sharing. Note that the role of the PCIe extender in the processing node can be as simple as extending the underlying PCIe interconnect wiring and signaling to outside of the processing node or it can also provide additional features for host isolation. The extension allows the node to access an external PCIe switch fabric and eventually the IO adapters (10GbE or FC). The extender plays a passive role and when considered in terms of complexity and cost, is an order of magnitude lower than that of a 10GbE or IB adapter. The underlying switch fabric provides the necessary functionality to isolate the different processor address domains. Alternatively, such functionality can also be incorporated into the PCIe extender without unduly increasing its complexity. Regardless of whether the switch or the extender provides such a functionality, the overall effect is that the IO adapters that are used for LAN or SAN connectivity continue to be an integral part of the processing node's IO subsystem.

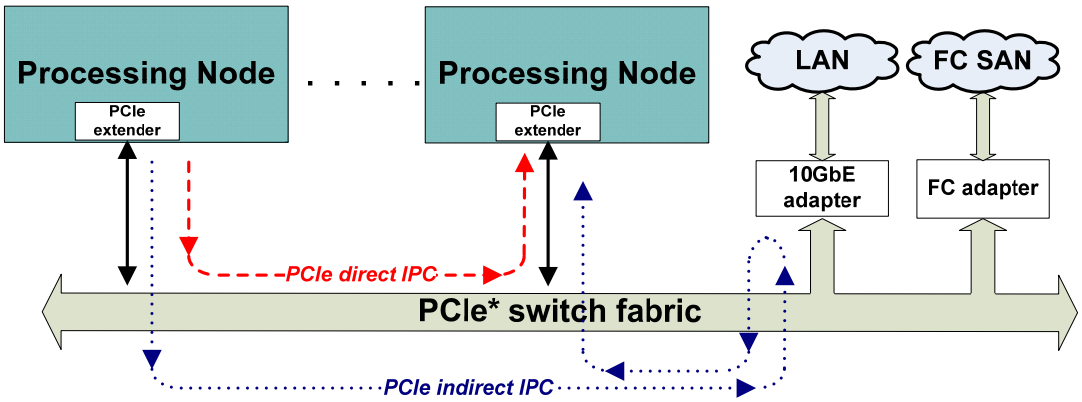

Figure 5. Supporting IPC, LAN and SAN with PCI Express.

In addition to LAN and SAN connectivity, the underlying PCIe switch fabric can also provide IPC connectivity. PCIe offers two options for supporting IPC. The first option is to use the shared IO model and make use of an IO adapter (e.g. 10GbE) that supports IPC. For instance, in Figure 5, the shared 10GbE adapter appears private to each processing node and hence has its own Ethernet MAC and IP address associated with it. From a processing node point of view, communication to another node is similar to that of using a dedicated IPC adapter except that in this case, the dedicated adapter resides remotely. The communication path in this model is illustrated using dotted lines in Figure 5. Data from the originating node traverses the PCIe switch fabric and gets "reflected" back by the shared network adapter card so as to reach the target node.

Alternatively, the nodes can communicate directly via the PCIe switch fabric without going through the 10GbE adapter. This mode of communication is shown with a dashed line in Figure 5. As mentioned earlier, the PCIe extender or the switch fabric can provide additional features to isolate the different processing address domains and can also incorporate additional mechanisms in the form of doorbells or scratchpad registers that permit a very efficient IPC mechanism to be supported.

To this point, the focus has been centered on the capabilities of each of the three interconnect technologies in the backplane. It is clear that PCIe holds architectural advantages over Ethernet and Infiniband. The following sections will focus on two key components -overall performance and deployment cost.

3 Performance

Providing low latency and high bandwidth at the hardware level is a good foundation for a high performance system. The interconnect must also provide the applications an efficient interface to fully exploit the underlying hardware. In this section, consider two orthogonal mechanisms towards achieving high performance: hardware capabilities and usage model for applications.

3.1 Hardware Capabilities

10GbE offers an evolutionary path from gigabit Ethernet (GbE) in terms of higher throughput with minimal disruption to existing GbE installations. As a matter of fact, a large number of servers are shipping with 10GbE capability. And with 40GbE and 100GbE in the roadmap, Ethernet is a viable candidate for future backplane deployments. However, bandwidth is not considered the sole performance metric designers take into consideration. Ethernet fails in two important areas – latency and jitter.

Ethernet is inherently unreliable and can drop packets in the wake of congestion which often results in high and unpredictable latencies. This is not only unacceptable for IPC where many latency sensitive messages occur between hosts but also for SAN communication where data loss is unacceptable. As previously mentioned, CEE definitions which allows Ethernet to be inherently reliable is under consideration. CEE implements a mechanism that supports prioritization of different types of traffic as well as the ability to convey congestion information for each traffic flow. Such enhancements allow the end points to limit the data rate until congestion clears and can result in a simplified software stack with ultimately lower latencies. At this point however, it is still unclear whether or not the improved latencies can rival those offered by Infiniband or PCIe (<2µs).

In addition, consider how a converged network adapter (CNA) for CEE or an Infiniband adapter communicates with the server CPU/memory subsystem. It is with most certainty that such adapters will use PCIe as the communication interface to the host. For Ethernet CNAs that do not support bandwidths beyond 10Gbps, the PCIe interface offers enough bandwidth to sustain line rates. But in the case of IB, a QDR port can sustain a raw data rate of 40Gbps. Even though this is the equivalent to what a 4-lane PCIe 2.0 interface can support, the overheads of the PCIe results in the realizable IB bandwidth to drop below 40Gbps. For instance, a dual ported IB adapter offering 40 Gbps for each port will not be able to sustain a throughput of 80Gbps even when using an 8-lane PCIe 2.0 interface. The next generation of IB will use as EDR (eight data rate) version that would further push the maximum bandwidth to 80Gbps per port – and as with the QDR scenario, the EDR adapter's performance will be affected by PCIe. With PCIe holding the key to the bandwidth performance of IB, it is obvious that the bandwidth benefits could be derived by directly using PCIe. From a latency perspective, PCIe adds very little latency – so the impact on the CNA or IB adapter can be considered minimal. Nevertheless, as with the bandwidth, PCIe places a minimum threshold on the best latencies that either CNA or IB could hope to achieve.

From a raw numbers perspective, PCIe has an extremely low end-to-end latency (<1µs). The new PCIe 3.0 standard also supports higher throughput of 8Gbps per lane. Hence, a 16-lane (x16) PCIe interface can support an impressive128Gbps.

From the perspective of the two end points involved in the communication, Infiniband and Ethernet adapters serve as bridges to PCIe. So, rather than terminating PCIe inside the system and using a different protocol (IB or Ethernet) for communication, it is advantageous to extend PCIe outside the system so as to realize its full latency and bandwidth potential and benefit from direct read/write of remote memory. This is a feature that only PCIe offers and is the base for the application usage models described in the next section.

3.2 Application Usage Model

In general, there are two ways an application running on a server can transfer data from or to a remote server. The first of the two approaches is remote direct memory access (*RDMA)*. This is the oft-quoted paradigm for supporting efficient data transfers. The RDMA model enables an application on one server to directly place or access data on the memory of a remote server. This is in contrast to the second approach of using *messaging* wherein an application writes to an anonymous (common) buffer on the remote node from which the data is eventually copied to the target application buffer. By eliminating data copies, RDMA enables low latency and high bandwidth transfers to be achieved.

Infiniband has native support for both RDMA as well as the messaging model. Ethernet, by default, supports only the traditional messaging model. However, there are two different approaches to support RDMA over Ethernet. One approach is to provide support for RDMA at the TCP/IP protocol level (also known as iWARP). For the iWARP approach to result in any meaningful performance improvement, it is necessary for the complete protocol to be offloaded to the underlying Ethernet adapter. The other approach for supporting RDMA over Ethernet builds upon the lossless Ethernet (CEE) model. And rather than using TCP/IP, this model runs RDMA directly atop Ethernet. Unlike iWARP, this approach to doing RDMA has not been standardized yet. Furthermore, compared to Infiniband, RDMA over Ethernet is not commonly used. But regardless of whether Infiniband or Ethernet is used, in the context of IPC communication across a backplane where most of the IPC communication would involve small- and medium-sized messages, the direct addressing mode as opposed to using RDMA would be considered more efficient.

PCIe indeed supports this unique mechanism to directly access data in remote memory and does so by using non-transparent bridging (NTB). Support for NTB can be incorporated either within the PCI extender or at the switch. In the NTB approach, part of the local address space of each server is mapped to a remote server's address space using a translation mechanism. Whenever a CPU generates a read or a write operation targeting that part of the local address space, the NTB consumes the incoming transaction and automatically generates an outgoing transaction on the remote server. This transaction contains an address that falls within the remote server's address space. From a sender's perspective, the operation targets its own address space and from a receiver's perspective, the operation originates from within its own address space. The implication is that the CPU can directly write data from its first-level cache to a remote memory location. Contrast this to the rather onerous RDMA approach of flushing the data to physical memory, setting up the local DMA engine and finally initiating the RDMA transfer. Moreover, the NTB approach in PCIe also offers other IPC-centric features such as scratchpad registers, doorbells etc. to aid in generating short remote messages and interrupts.

Essentially, the PCIe direct read/write addressing paradigm provides a more efficient mode of IPC when compared to the RDMA or messaging models supported by Infiniband or Ethernet. At the same time, it has been argued that the PCIe mode of direct access is not a secure mode of operation and could likely result in memory corruption by wayward applications. For IPC that occurs between servers within the confines of the backplane, this is unlikely. Nonetheless, the NTB approach can be further augmented with protection bits to ensure that all remote accesses are validated.

4 Deployment Cost

The level of IO scalability and sharing afforded by PCIe allows for an efficient solution while its mass adoption results in a significant cost reduction. This section expands upon the merits of IO virtualization as well as the adoption of PCIe into mainstream processor technology.

4.1 IO Virtualization

PCIe offers a simplified solution by allowing all IO adapters (whether it be 10GbE or FC) to be completely moved outside the server. With a PCIe switch fabric providing virtualization support, each of the adapter can be shared across multiple servers and at the same time provide each server with a logical adapter (Section 2). The servers (or the VMs on each server) continue to have direct access to its own set of hardware resources on the shared adapter. The resulting virtualization allows for better scalability wherein the IO and the servers can be scaled independent of each other.

IO virtualization avoids over-provisioning the servers or the IO resources leading to cost and power reduction. For example, from a high-availability point of view, each server must be provided two adapters. But by disaggregating the IO adapters, one could potentially use one additional adapter as a redundant spare for a group of servers. So, for a group of N servers, this leads to a reduction from 2N adapters to N+1 adapters. And as such, PCIe based IO virtualization will continue to support the CEE or other IO adapters and at the same time, enable efficient use of expensive IO adapters by removing the one-to-one association between the server and the IO resource (See Figure 5).

4.2 Mainstream Technology

Yet another important factor which determines the cost of deployment in the long run is the level of integration of the technology by the processor vendors. Mainstream processor vendors like Intel have long made PCIe part of their chip sets. But the level of integration has reached a point that PCIe has now become an integral part of the processor die and resides alongside the memory controller and processing cores. Such a level of integration eliminates the need for discrete components as well as offers wide flexibility in the deployment of PCIe.

Ethernet is not as tightly integrated as PCIe but nonetheless it has been offered by processor vendors like Intel as part of their chipset. Though mostly available as a dedicated PCIe adapter card, 10GbE is on its way to getting integrated into the chipset. At the same time, it is unlikely that Ethernet will become part of the processor core in the near future.

And finally, at the far end of the spectrum is Infiniband. Most of its deployments are in the form of PCIe-based adapters which connect to the server's PCIe slots. Infiniband has not been integrated to any chipset. With Infiniband far removed from the processing logic, there is a need for discrete components to interface with the rest of the system. Also, with one company being the sole supplier of Inifinband based silicon, the costs associated with deploying it will continue to be higher than that of Ethernet and PCIe.

5 Summary

Unlike other interconnects, PCIe provides a unique cost/performance value proposition by providing an affordable solution for low-end deployments while at the same time offering exceptional performance for high-end applications.

The key features that distinguish PCIe are:

- High performance in terms of bandwidth and latency
- Support for an efficient IPC paradigm through use of non-transparent shared memory
- IO virtualization in a multi-host environment wherein PCIe does not supplant traditional IO resources but allow them to be used efficiently
- Integration in mainstream processor chips