PCI Express®-over-Optics in Embedded Applications: A Guide to Easy Implementation

White Paper

April 2024

COPYRIGHTS, TRADEMARKS AND PATENTS

Terms of Use

Use of this publication is limited to viewing the pages for evaluation or purchase. No permission is granted to the user to copy, print, distribute, transmit, display in public, or modify the contents of this document in any way.

Disclaimer

The information in this publication may change without notice. All materials published here are “As Is” and without implied or express warranties. Samtec, Inc. does not warrant that this publication will be without error, or that defects will be corrected. Samtec, Inc. makes every effort to present our customers an excellent and useful publication, but we do not warrant or represent the use of the materials here in terms of their accuracy, reliability or otherwise. Therefore, you agree that all access and use of this publication’s content is at your own risk.

Updated Documentation

Please visit www.samtec.com to get access to the latest documentation.

NEITHER SAMTEC, INC. NOR ANY PARTY INVOLVED IN CREATING, PRODUCING, OR DELIVERING THIS PUBLICATION SHALL BE LIABLE FOR ANY DIRECT, INCIDENTAL, CONSEQUENTIAL, INDIRECT, OR PUNITIVE DAMAGES ARISING OUT OF YOUR ACCESS, USE OR INABILITY TO ACCESS OR USE THIS PUBLICATION, OR ANY ERRORS OR OMISSIONS IN ITS CONTENT.

Abstract

This paper examines the use of PCIe® for backplane expansion using fiber optics, including its applications and limitations for high-performance embedded systems, SoCs, and microcontrollers. We focus on the hardware and architectural options available to system designers when using PCIe as a high-performance connectivity solution for peripheral expansion. In addition, we include examples of flexible configurations, which are particularly important when several host computers are interconnected.

Authors

Jean-Frédéric Gauvin Jean Frédéric Gauvin is a sales manager at Dolphin Interconnect Solutions, bringing over 15 years of extensive experience in the technology industry and embedded systems. His journey in the tech realm commenced as an engineer at Airbus Industries, where he honed his skills in navigating the intricacies of cutting-edge technologies and transitioned into a consultancy role, specializing in embedded software development. This pivotal career move deepened his understanding of the relationship between technology and business.

With a profound appreciation for the transformative power of technology, Jean Frédéric has merged his technical expertise with business acumen. At Dolphin Interconnect Solutions, he leverages his background to assist clients in navigating the complex landscape of cutting-edge technology. Passionate and forward-thinking, Jean Frédéric is a staunch advocate for the adoption of PCIe as the interconnection solution for embedded systems. He firmly believes that PCIe holds the key to unlocking unprecedented levels of efficiency and performance.

Matthew Burns develops go-to-market strategies for Samtec's Silicon-to-Silicon™ solutions. Over the course of 20+ years, he has been a leader in design, technical sales and marketing in the telecommunications, medical and electronic components industries. Mr. Burns holds a B.S. in Electrical Engineering from Penn State University.

Integration leads to Innovation

Samtec is structured like no other company in the interconnect industry: we work in a fully integrated capacity that enables true collaboration. The result is innovative solutions and effective strategies supporting optimization of the entire signal channel.

For more information contact SIG@samtec.com

Introduction

Engineers designing embedded systems typically focus on optimizing capital equipment usage while enhancing overall performance. This includes particular attention to reliability and long-lifetimes for these high-performance solutions, particularly if the systems will be deployed in challenging environments where troubleshooting and part replacement can be difficult.

In many applications, the real-time aspect is a central consideration, and the system must be able to process a high volume of data with extremely low latency. Engineers use specialized hardware solutions such as FPGAs, ASICs, GPUs, and advanced storage solutions to meet these demands effectively. However, when building large systems where the interconnection between I/O or processor must be over several meters, ensuring low latency, high bandwidth, real-time aspect, and signal integrity becomes a major concern.

Traditionally perceived as a chip-to-chip, single-host interconnect technology, PCIe (PCI Express)-over-fiber is making inroads into switch fabrics, challenging, and potentially replacing previous interconnect technologies in embedded systems. Moreover, PCIe-over-fiber unlocks new possibilities for remote PCIe expansion. In this topology, the host or server can be positioned up to 100 meters away from the I/O, while still harnessing the full benefits of low latency and high bandwidth associated with PCIe.

Keeping these considerations in mind, this paper explains how to expand PCIe peripherals to external chassis using optical fiber. It also covers how to build an embedded computer cluster topology that enables adaptable configurations.

Flexible PCI Express Configurations

PCI Express is a serial connection that operates more like a network than a bus. The PCIe protocol demonstrates exceptional robustness, incorporating a control mechanism that ensures zero (0) packet loss due to acknowledging every packet at every hop. Limitations in traditional communication buses, such as packet loss or bandwidth, have brought the PCIe communication protocol forward as a viable option in embedded systems, especially military or automotive applications that have no tolerance for packet loss. With the PCIe protocol, failure to deliver a transaction layer packet is a major malfunction likely caused by hardware failure.

The PCIe interface uses the concepts of a root complex and peripheral devices. The root complex is almost always at the processor end, and it is the responsibility of the root complex to scan the PCIe bus and find all connected downstream devices. This process, called enumeration, queries each device found on the PCIe bus, determines interface parameters such as memory window requirements of each endpoint device,

and assigns resources from within the root complex's memory map to satisfy these requirements. After enumeration, the root device can access all connected PCIe peripherals with simple read/write operations within its local address space.

The PCIe interface does not allow multiple root complexes to exist on a single PCIe bus because each root complex has its own enumeration process and root complexes do not share a common address space. This limitation means two processors, each with independent root complexes, cannot connect directly with each other.

In the embedded system computing environment, diverse applications often require an expansion of the PCIe bus to accommodate additional PCIe slots. This need has given rise to creative solutions, one of which involves the use of PCIe for chassis expansion in non-transparent mode for multiple host configurations.

Single Host Configuration

Figure 1 shows one way to connect 16 peripherals using PCIe over optical fiber with passive backplanes (such as the Dolphin backplane IBP-G4X16-5 shown). In this example, the host is equipped with an MXH945 card (Dolphin ICS) where each of the four lanes of fiber is connected to the target board (MXH948) in the backplane using Samtec FireFly™ Micro Flyover System™ optical cable assemblies (Figure 2) and standard MPO-based optical fiber cables.

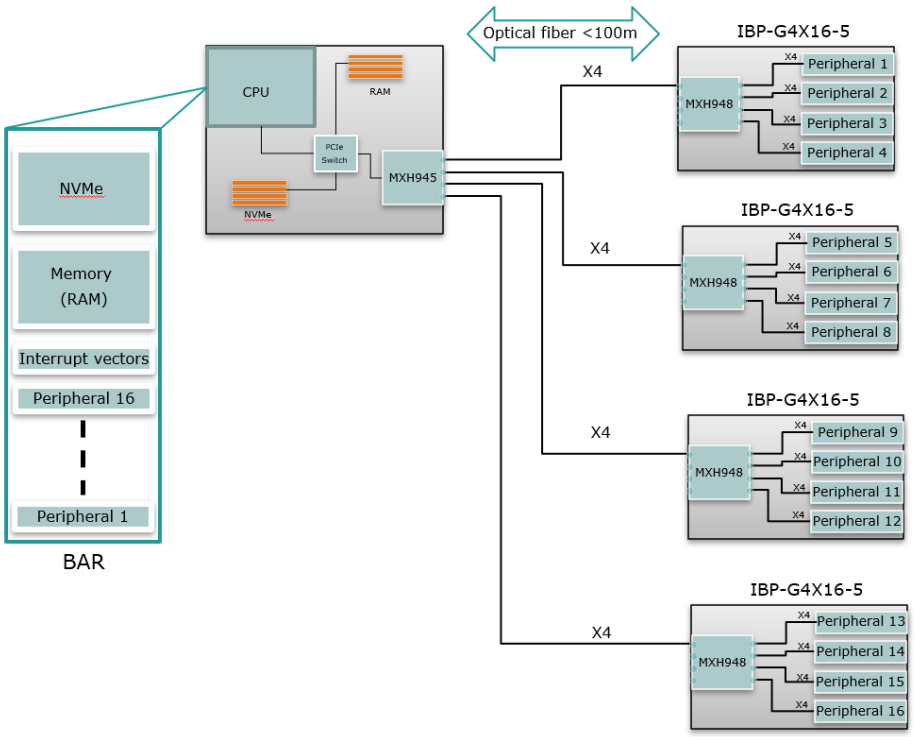

Figure 1: Single-host configuration showing expansion in transparent mode.

The 16x PCIe bus originating from the target board is then split into four 4x PCIe 4.0 links. This type of configuration supports any expansion board compliant with PCIe standards.

This system uses products selected to deliver an industry-leading miniature footprint. Greater density and closer proximity to the IC enables chip-to-chip, board-to-board, on-board, and system-to-system connectivity. The layout in Figure 1, for example, features performance of up to 32 Gbps over 100 m of fiber in a footprint of 406 mm². Close proximity to the ASIC also enables lower drive voltages and pre-emphasis resulting in reduced power consumption.

The miniature size of Samtec's high-density optical transceivers (Figure 2) allows them to be easily designed into the downstream system, ultimately making these systems smaller. Highly flexible, small diameter, industry-standard optical patch cords provide connection to the control system.

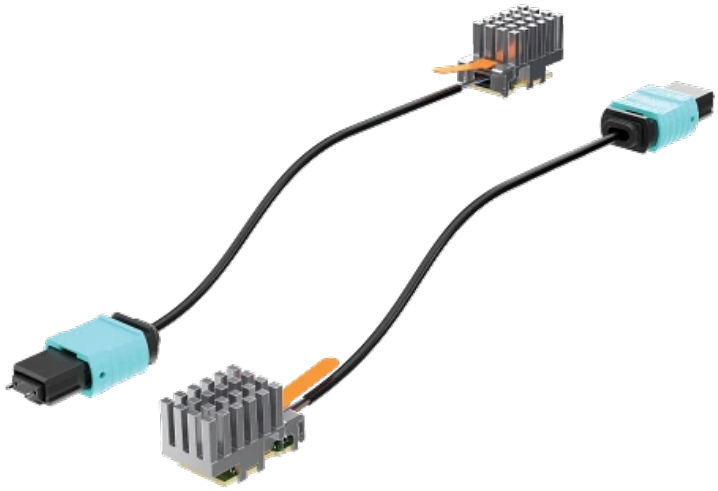

Figure 2: Samtec PCIe®-Over Fiber, FireFly™ Optical Cable Assembly

At boot-up, the delegated I/O (peripheral) functions in this application are registered in the device tree of the server CPU using PCIe, and they are directly addressable from the host application through its device driver.

It is important to note that different architectures and systems may require a tailored PCIe transparent expansion architecture that is designed to meet the specific demands of an application. As technology advances, the adaptability and configurability of PCIe for chassis expansion in transparent mode offers a method that enables diverse computing environments and an easy way to expand existing systems.

However, in a transparent bridge architecture like the one in Figure 1, two hosts cannot coexist on the same fabric. The PCIe transparent architecture does not allow them to share the same root complex. Use cases with multiple hosts need to use non-transparent bridging.

Multiple Host Configurations

To solve the processor-to-processor connection problem, some PCIe switches can configure a port as a non-transparent bridge (NTB). When the root node finds a port configured as non-transparent (NT), it does not look for devices past the NT port, and the enumeration scan does not continue through the NT port. This effectively isolates the PCIe bus on each side of an NT port (see Figure 3).

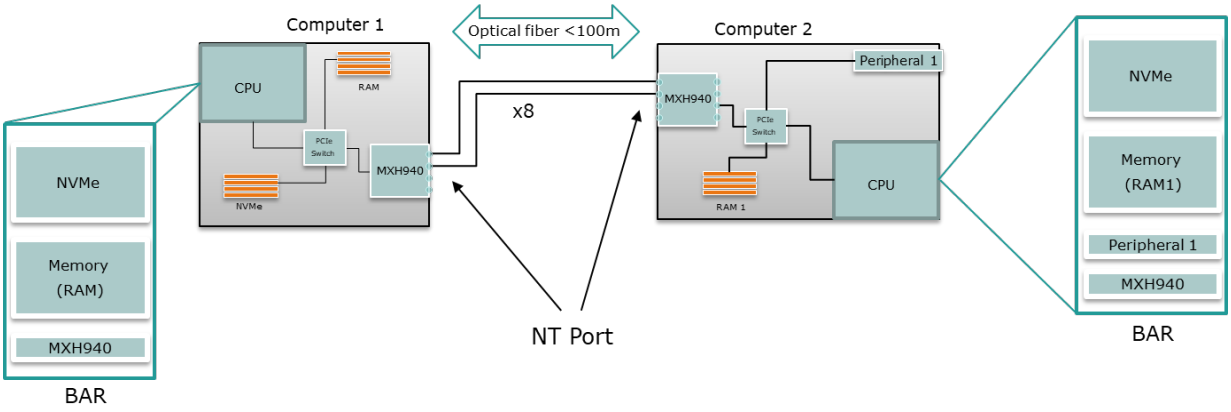

Figure 3: Using a non-transparent bridge configuration allows the coexistence of two hosts.

The NT port has a mapping mechanism allowing the two processors to communicate (Figure 3). When CPU 1 enumerates the NT port, the NT port requests an address range within CPU 1's address space. Similarly, when CPU 2 enumerates the NT port, it also assigns an address range within CPU 2's address space. NTB software sets a reachable memory region in the remote server. The NT port provides an address translation mechanism, so data written from computer 1 is written safely into a designated addressable memory buffer on computer 2.

This non-transparent port mechanism allows the two processors, or root complexes, to communicate with one another. The setup and configuration of such a system can be complex as in-depth knowledge about the PCIe NTB implementation is required. Alternatively, NTB setup can be simplified using an application development environment, such as Dolphin eXpressWare technology, which supports Dolphin's SuperSockets, IPoPCIe, SmartIO and SISCI API software. Designed to support hardware features such as direct memory access (DMA) and multi-core processing, this application development environment includes reflective memory or multi-cast, peer-to-peer communication, and Dolphin smart I/O technology.

Embedded Clusters

The example in Figure 3 showed a two-host configuration connected by optical fiber, but different processing workloads can have highly variable demands for processing power and IO resources. A physical hardware system may have limited resources, but it may be desirable to scale up to allocate more resources and release them on demand.

Dynamic scaling based on current workload requirements leads to more efficient use of the available physical resources (see Figure 4).

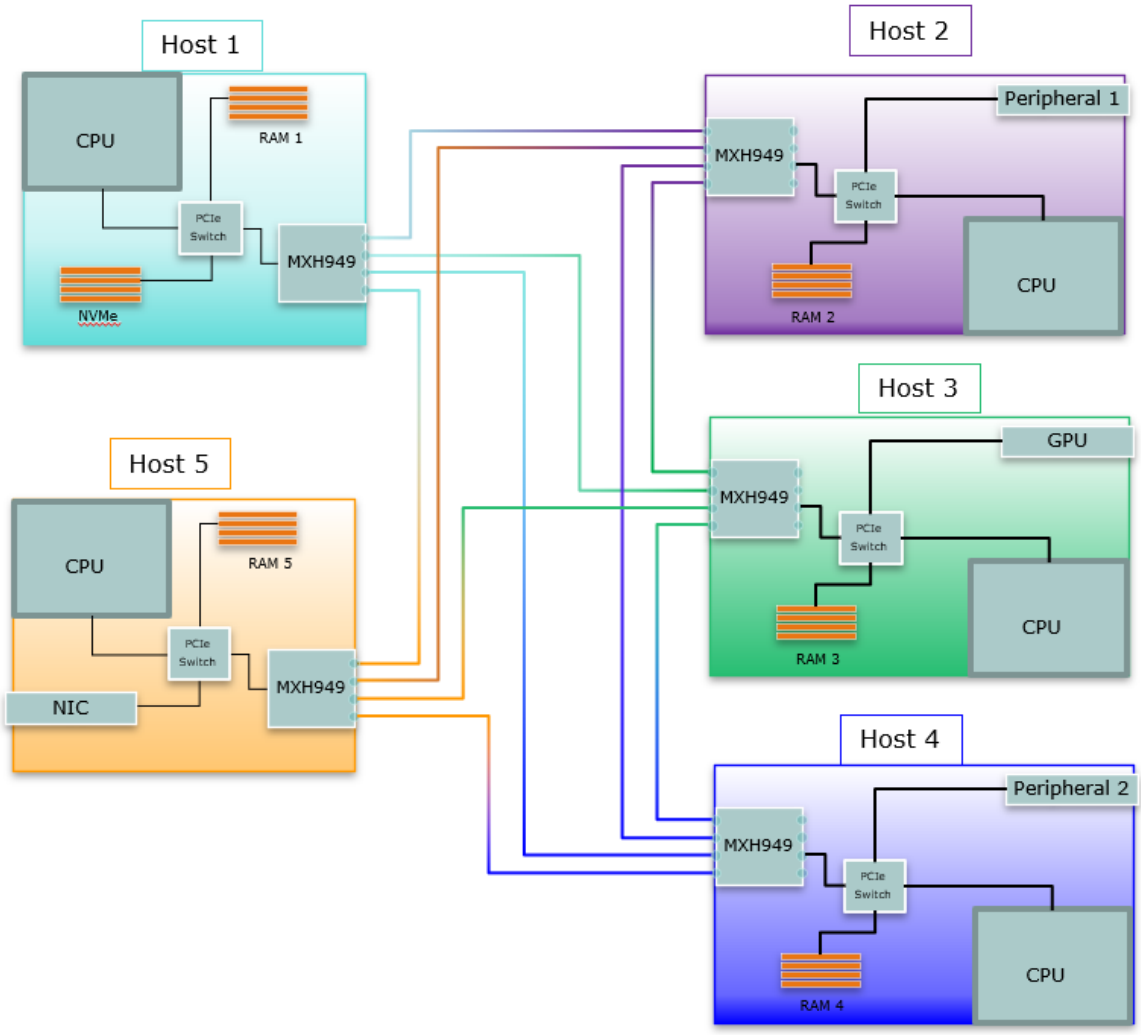

Figure 4: Using a non-transparent bridge configuration allows dynamic scaling across remote devices to address variable workloads.

In the example in Figure 4, a local CPU can perform memory operations on a remote device, such as reading from or writing to registers. Conversely, it is also possible to map local resources for a remote device, allowing it to write MSI interrupts and access the local system's memory across the NTB. Through PCIe, RDMA, RPIO, and remote interrupt, a cluster of embedded systems can use the fabrics to disaggregate the processing needs to several SoCs.

Device Lending is a mechanism for decoupling devices from the hosts they physically reside in. This process of temporarily "borrowing" remote devices and "lending" away local devices is illustrated in Figure 4. To the local system, the remote device appears to

be dynamically hot-added, allowing local applications and device drivers to use the device transparently, without being aware that the device is remote.

During periods of peak processing demand, the system can efficiently borrow resources from underutilized computers, mitigating the need for excess capacity. And, during periods of lower demand, surplus resources can be redistributed, ensuring optimal resource utilization across the entire cluster.

This concept provides a cost-saving, flexible infrastructure as compared to a traditional model where each computer system is dimensioned to handle its maximum load. Instead, the architecture is designed to facilitate the borrowing and sharing of resources among computers within the cluster. This eliminates the need for overprovisioning and introduces a responsive system capable of adapting to varying workloads.

This approach enables cost savings as expensive resources such as NPU or GPU can be shared between computers. The dynamic nature of the embedded cluster addresses the evolving needs of modern embedded computing environments.

Conclusion

PCIe-over-optics systems offer designers flexibility and high performance. PCIe-over-optics interconnects can enable high data throughput, coherency, and low latency in data center, edge infrastructure, AI/ML, and embedded applications. Running these embedded applications over PCIe fabric can provide the capability of a small high-performance computer while reducing hardware costs (because memory duplication is not needed, CPUs can borrow resources from other systems, and GPUs or NICs can be landed between CPUs).

Point-to-point, point-to-multipoint and next-gen system architectures like the ones featured in this paper can make embedded applications future proof while also being backwards compatible with existing systems.

Resources

[1] Samtec FireFly™ Product Information.

[2] FireFly™ Application Design Guide Samtec.

[3] Dolphin eXpressWare Software for PCI Express

[4] Dolphin High Performance PCI Express Products