Building a Low-End to Mid-Range Router with PCI Express® Switches

White Paper

Notes

By Kwok Kong

Introduction

PCI buses have been commonly used in low end routers to connect CPUs and network adapter cards (or line cards). The CPU performs packet forwarding, handles routing protocols to manage the routing table, and runs management software to control the operation of the router. A PCI bus-based router is shown in Figure 1. A PCI bridge forwards transactions between the CPU's local bus and the PCI bus. Packets received by a network adapter card are usually transferred by DMA to the CPU's memory subsystem using the PCI bus. The CPU processes these packets and forwards them to the destination network adapter card using the PCI bus.

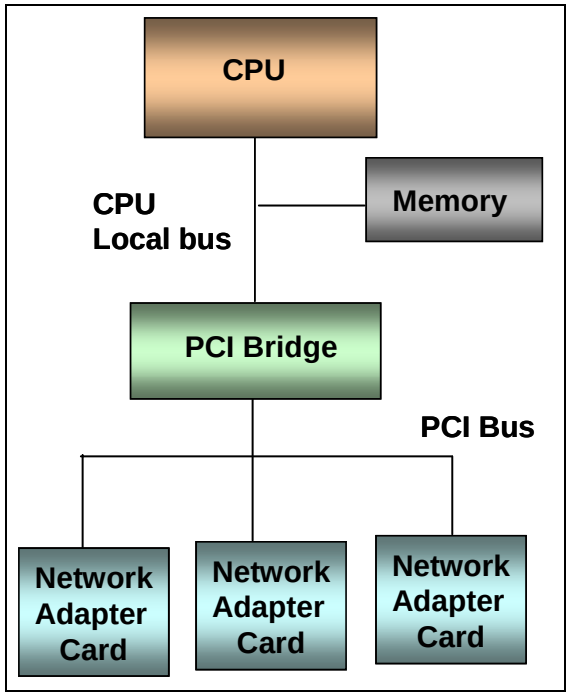

Figure 1 PCI-based Low End Router

PCI is a multi-drop, parallel bus implementation that is approaching its practical limits of performance. Today's CPUs and network interfaces are demanding much higher bandwidth than PCI can deliver. PCI Express (PCIe®) was introduced several years ago to be the next generation interconnect technology to replace PCI. The migration from PCI to PCIe in the desktop and server markets is occurring at an accelerating pace.

PCIe is a bidirectional serial interface and point-to-point technology using a 2.5 Gbps signaling rate per differential pair (lane). It offers significant performance improvements over PCI. Load and store operations are split transactions. The bandwidth is scalable via higher serial speed and additional lanes. It is a low cost technology since it leverages the high volume in the desktop and server markets. The serial interface and point-to-point technology make PCIe more suitable than PCI for backplane and cabled interconnect applications.

Notes

PCIe can be used to build low cost, high performance low-end routers as well as mid-range routers. A low-end router can support system bandwidths up to a few hundred Mbps. A mid-range router can support system bandwidths up to 10 Gbps.

System Architecture

Figure 2 illustrates a chassis-based router system architecture. This system has one slot for the CPU module and 4 line card slots for network interface cards. Each card has a point-to-point connection to the CPU module. Systems can be configured with up to 8 line card slots; the number of slots depends on system requirements and the needs of the target market. In this example, PCIe is used as the backplane interconnect. A multi-ported PCIe switch is integrated in the CPU module such that a separate switch module is not required.

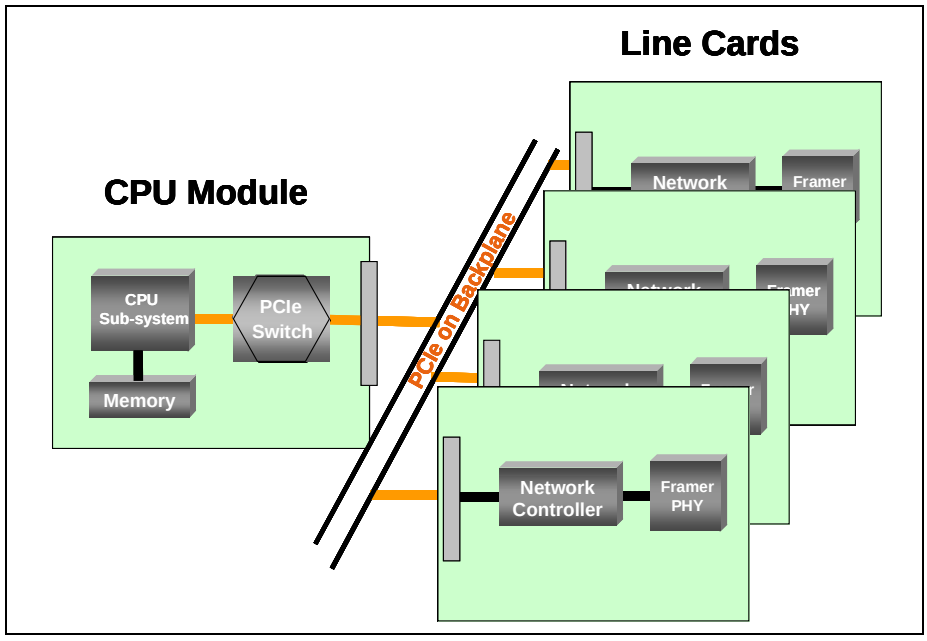

Figure 2 System Diagram of a PCIe Based Router

The line cards are initialized and configured by the CPU module during system startup. All packets received by a line card are sent to the CPU module which handles all packet processing and forwarding decisions. Once the packet forwarding decision is made, the CPU module sends the packet to the destination line card which transmits the packet to its network interface. The operation is very similar to a PCI Bus-based router which is shown in Figure 1.

The system architecture diagram is shown in Figure 3. A x8 2.5 Gbps PCIe interface is able to provide a raw data rate of 16 Gbps, which is sufficient bandwidth to compensate for the PCIe overhead, resulting in an effective user data rate of well over 10 Gbps.

A multi-ported PCIe switch, integrated in the CPU module, provides line card connection via the backplane. A x4 2.5 Gbps PCIe interface to each one of the line card slots can deliver an effective user data rate of over 5 Gbps. A line card does not need to use all the x4 PCIe interfaces when its network interface speed is much less than 5 Gbps. For example, a T1/E1, T3/E3, or OC-12 ATM network interface line card only needs a x1 PCIe interface. However, a x4 interface will allow the scaling of bandwidth to meet future needs.

If the backplane is designed to support a per lane signaling rate of 5 Gbps, the system can be upgraded in the future to use a PCIe switch that supports a 5 Gbps serial interface. This would effectively double system throughput. To upgrade, it is first necessary to upgrade the CPU module. The older line cards can still be supported, but running at 2.5 Gbps on each serial lane as the protocol auto-negotiates to the 2.5 Gbps rate. Then, the newer line cards that support a 5 Gbps serial lane can be plugged into the system to take advantage of the upgraded CPU module. This allows line cards to be gradually upgraded to 5 Gbps on a “as needed” basis.

Notes

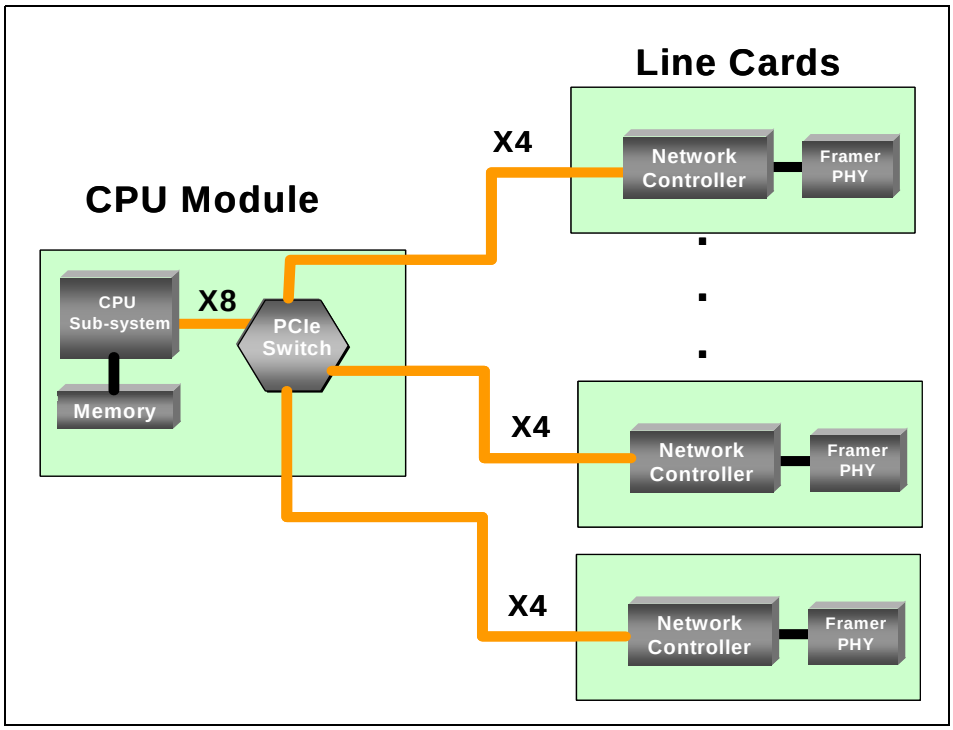

Figure 3 System Architecture Diagram

Redundancy

A mid-range router often requires some level of redundancy. A dual-CPU topology has both an active and a standby CPU. An example of this topology is given in Figure 4. A Mux/DeMux device such as the IDT PS421 Multiplexer/Demultiplexer Switch connects the Upstream Port (UP) of the PCIe switch to either CPU. The Mux/DeMux may also be integrated into the PCIe switch. The CPU that is connected to the UP of the PCIe switch becomes the active CPU. During normal operation, the active CPU is in control of the system and performs all routing functions. The standby CPU is not connected to the system. There is an out-of-band connection (not shown) between the CPUs, allowing heart beat and checkpoint messages to be sent periodically from the active to the standby CPU. The standby CPU monitors the state of the active CPU.

The standby CPU takes over as the active CPU when a managed switchover is requested or when the standby CPU detects a failure in the active CPU. A managed switchover is one that is initiated by the user for scheduled maintenance or software upgrades.

Managed Switchover Procedure

- Active CPU disables all the line cards. Line cards stop accessing CPU memory.
- The multiplexer/demultiplexer on each line card is switched to connect to the standby CPU.
- After the switchover, the standby CPU becomes the active CPU.
- The CPU re-initializes the PCIe switch and all the line cards. The system is then ready.

Failure Switchover Procedure

- Standby CPU monitors the state of the active CPU through heart beat and checkpoint messages.
- When there is a failure in the active CPU, the standby CPU takes over as the active CPU. The newly activated CPU configures the multiplexer/demultiplexer on all line cards to switch connections to itself.
- The active CPU re-initializes the PCIe switch and all the line cards. The system is then ready.

Notes

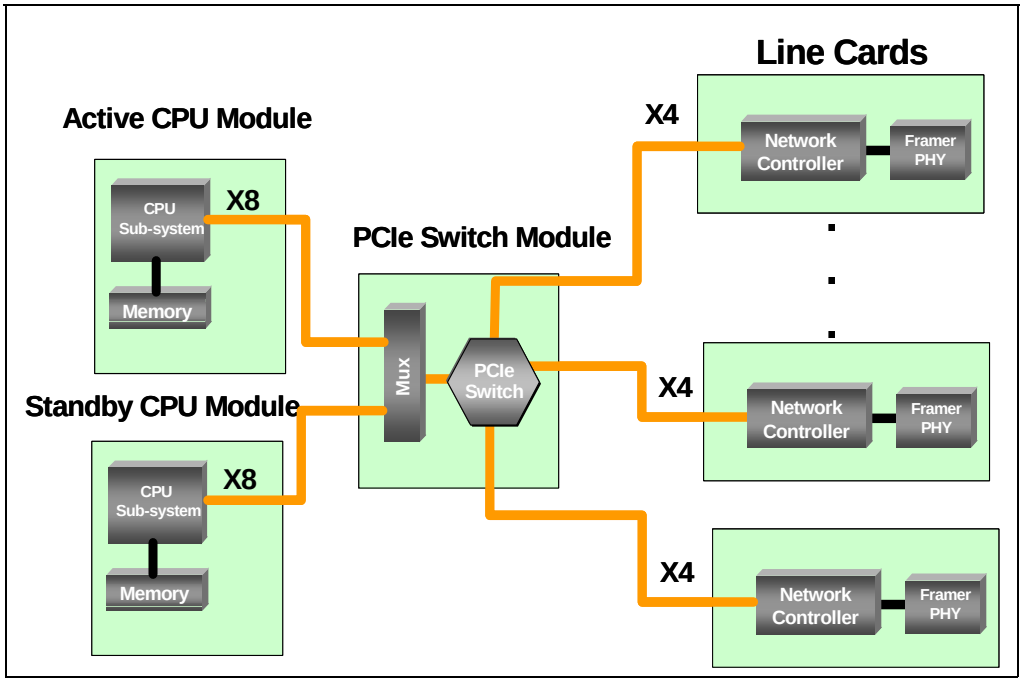

Figure 4 Dual-CPU Topology

The switch is still a single point of failure in the dual-CPU topology. For a fully redundant system, a dual-star topology may be deployed as shown in Figure 5. In this topology, an additional PCIe switch is added to the CPU module (the module now includes a PCI switch in addition to the CPU), and a multiplexer/demultiplexer is added to the line card. Each line card connects to both PCIe switches but only one connection is active.

During normal operation, all line cards are connected to the active CPU module which is in control of the system. The standby CPU module remains in standby mode and is not connected to the line cards. Similar to the dual-CPU system, there is an out-of-band connection (not shown) between the CPUs, allowing heart beat and checkpoint messages to be sent periodically from the active to the standby CPU. The standby CPU monitors the state of the active CPU.

The switchover procedure is similar for the dual CPUs and the dual-star topology.

Managed Switchover Procedure

- Active CPU disables all the line cards. Line cards stop accessing the CPU memory.
- The multiplexer/demultiplexer on each line card is switched to connect to the standby CPU.
- After the switchover, the standby CPU becomes the active CPU.
- The CPU re-initializes the PCIe switch and all the line cards. The system is ready.

Failure Switchover Procedure

- Standby CPU monitors the state of the active CPU through heart beat and checkpoint messages.
- When there is a failure in the active CPU, the standby CPU takes over as the active CPU. The newly activated CPU configures the multiplexer/demultiplexer on all line cards to switch connections to itself.
- The active CPU re-initializes the PCIe switch and all the line cards. The system is then ready.

Notes

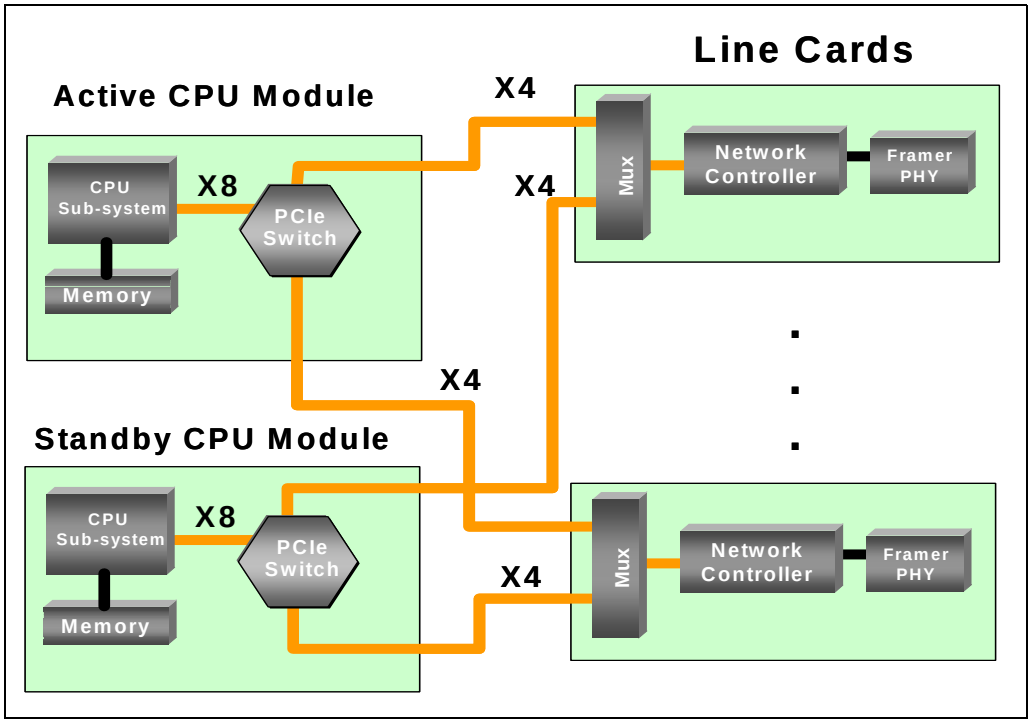

Figure 5 Dual-star Topology

Summary

The PCI bus has been widely used for the past 10 years in a low-end to mid-range routers to connect network adapter cards. However, today's CPUs and network interfaces demand much higher bandwidth than PCI can deliver. Especially in the desktop and server market segments, the migration from PCI to PCIe is accelerating. Now, it is time for low-end routers to also migrate from PCI to PCIe. Because PCIe supports much higher bandwidth than PCI, it can serve the mid-range router market as well. Both low-end and mid-range routers can be designed and built based on the same architecture, CPU module, backplane, and line cards. Only the number of slots in the chassis defines the difference between a low-end and mid-range router.

Redundancy is important in a mid-range router system. Two topologies, dual-CPU and Dual-star, have been identified as providing different levels of redundancy.

PCIe is supported in many market segments, including desktop, server, communication, and embedded devices. This guarantees high volume and hence low cost. PCIe is software-compatible with existing PCI-based software, allowing for a smooth migration from older PCI systems to higher-bandwidth PCIe systems in the future. PCIe provides scalable performance via additional lanes and higher serial speeds. Finally, for low-end and mid-range router systems, PCIe is the natural choice when upgrading from PCI because of its backplane connectivity.