

BY RICHARD A QUINNELL • CONTRIBUTING TECHNICAL EDITOR

PCI EXPRESS **CONTENDS** FOR COMMUNICATIONS ROLE

SCALABLE PERFORMANCE AND LOW COST ARE MAKING PCI EXPRESS ATTRACTIVE FOR COMMUNICATIONS-DEVICE DESIGNS, CHALLENGING PROPRIETARY-BUS STRUCTURES FOR NEXT-GENERATION DESIGN WINS.

Replacing PCI as a peripheral bus in general-purpose computers, PCIe (PCI Express) is now seeking a role in communications. It offers the raw performance for communications and has significant cost advantages over today's popular proprietary buses. Its legacy link to the PC may limit its communications success, however, unless proponents can solve critical shortfalls in its architecture.

Decades after it surfaced, the IBM PC is still having a ripple effect throughout the electronics industry. The PC's immense popularity and resulting production volumes have made PC-centric technology both inexpensive and widespread. These advantages, in turn, have made the technology appealing to a variety of other applications. The PC's bus structures have spun off a number of derivatives, including PC/104, PXI, and CompactPCI, which make the PC's processor, peripheral devices, and software elements available to non-PC applications.



PCIe is the most recent version of the PC's peripheral-bus structure to begin finding its way into other applications. As with the earlier PCI and AT buses, PCIe is generating interest because it allows embedded-computing developers to use the proven, powerful, low-cost, and widely available technology that arises from desktop computing. Unlike the previous buses, however, PCIe offers performance levels that match the processor's—with room to grow. At the same time, PCIe retains software compatibility with the PCI bus, preserving the cost advantages that previous generations of PC technology have enjoyed (see sidebar "PCI Express basics").

This combination of high performance and low cost has caught the attention of the communications industry. Traditionally, communication developers have used proprietary-bus structures to handle their highest performance needs. Cost and time-to-market pressures, however, have made proprietary approaches increasingly unattractive.

Still, the home that PCIe will find in the communications market is not yet certain. Communications devices have a broad range of needs, varying with their position in the network hierarchy. This hierarchy spans multiple levels with differing mixes of control and data-handling requirements (Figure 1).

VARYING NEEDS

The transport tier lies at the high-data-rate end. This tier provides long-distance data transport over high-capacity channels. Statically configured transport-tier devices do not interact significantly with the data they are transporting. As a result, these devices are not strong candidates for PCIe. Devices start to become aware of the data they handle at the core tier, although their interaction with the data is still limited. Core devices prioritize their functions based on the tags and labels added to the data in the lower tiers. Typically, core devices use an architecture that separates the control of the data from the handling of the data. The control plane manages the tables that direct the data plane in routing the data through the system. Whereas PCIe may have a role in the control plane of core-tier devices, the data plane tends to employ proprietary-bus structures and protocols to minimize the packet-routing

AT A GLANCE

- ▶ PCIe (PCI Express) replaces the PC's parallel PCI bus with a switched serial bus that offers software compatibility.
- ▶ The cost and availability of such PC hardware make it attractive to other applications.
- ▶ The scalable bandwidth performance of PCIe gives it the raw performance to match many communications-system requirements.
- ▶ PCIe allows only a single master processor, making it awkward for implementing redundancy and high-availability design.

overhead and achieve high bandwidth efficiency.

High-end edge-tier devices often have similar architectures to but lower performance demands and greater data interaction than core devices. These tier devices provide service-aware QOS (quality-of-

service) enforcement and traffic management. Lower end edge devices are typically aggregation nodes that help maintain a relatively even amount of traffic to and from the high-end edge devices to maximize edge bandwidth-usage efficiency. At this tier, the performance of PCIe begins to more clearly match the needs of both control and data planes.

The strongest match, however, occurs at the access tier, at which users connect to the network through a service provider. At this tier, requirements are diverse because the access tier must handle a variety of protocols, including data, voice, and multimedia, in its interaction with the end user. In addition, physical considerations, such as the distance to the user, affect the requirements. Designs for this equipment show considerable diversity in their number of ports, node capacity, and redundancy strategies. Designers can choose from multiple architecture options, including separated control and data planes and merged traffic, in these designs.

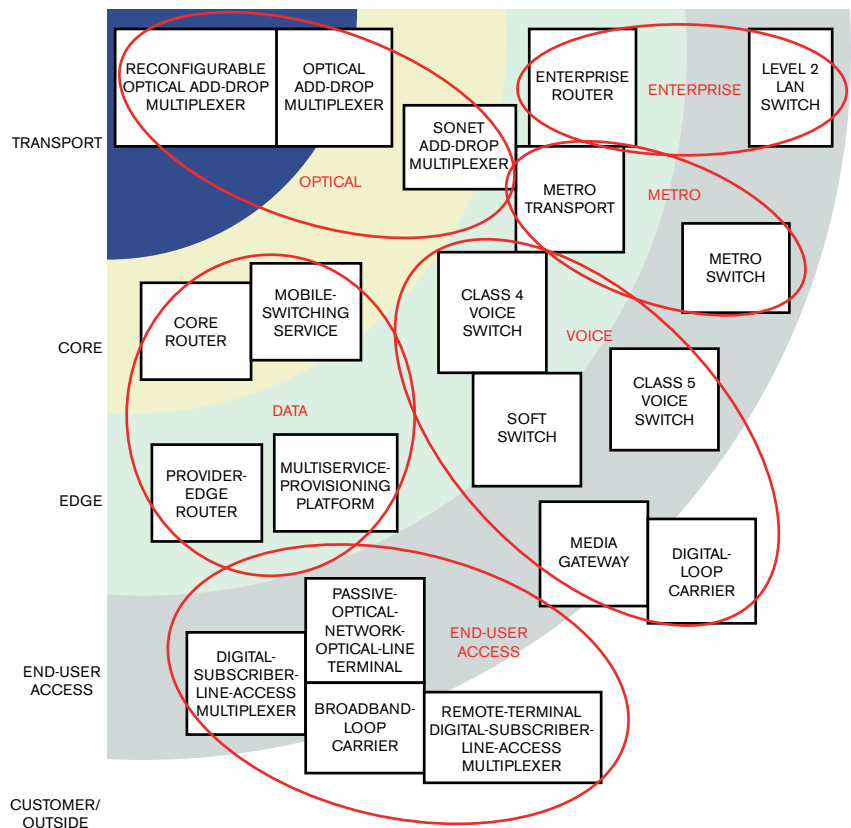


Figure 1 Devices for the communications market, depending on the tier in which they operate, vary in their bus needs and the applicability of PCI Express (courtesy IDT).

PCI EXPRESS BASICS

Parallel buses become harder to implement as clock speeds increase and skew between clock and signal lines becomes an ever-larger percentage of the clock cycle. As a result, the PCI bus has languished at a modest clock rate of 266 MHz while processor clock speeds have topped 1 GHz. To eliminate this mismatch and the system bottleneck it causes, the PCI-SIG (PCI Special Interest Group) developed PCIe (PCI Express). The group's goal was to bypass the bandwidth limitations of a parallel bus and maintain application-software compatibility with PCI.

A PCIe bus comprises a set of parallel "lanes" that make a point-to-point connection between two nodes, such as the CPU and a peripheral controller. A multiport switch at the center of a star topology allows the point-to-point PCIe connections to replicate the many possible paths of a traditional parallel bus (Figure

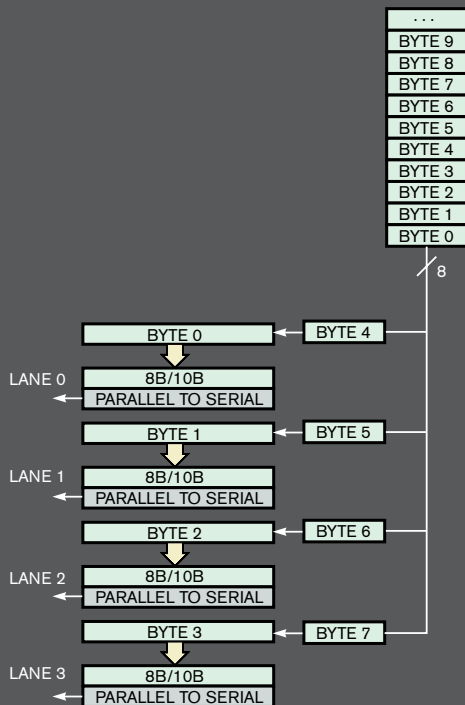


Figure A Physical-layer hardware in PCI Express automatically stripes successive data bytes across whatever lanes are available and recombines them in the proper order at the receiving end.

A). With the appropriate design, the switch can also allow two or more paths to operate concurrently as long as they do not share the same ports.

The lanes in PCIe are serial links that operate with 8b/10b encoding at a 2.5-GHz clock rate. Each lane has a forward and a return channel and uses differential signaling for a total of four wires per lane. To provide for bandwidth scaling, PCIe allows the connection between nodes to comprise one, two, four, eight, 12, 16, or 32 lanes. A PCIe connection can thus offer a raw data rate ranging from 250 Mbytes/sec to 8 Gbytes/sec. A pending upgrade to the PCIe specification will increase the allowable serial clock rate to 5 GHz, doubling the bandwidth capacity of a given lane configuration.

Designs need not use devices with matched lane counts. During power-up initialization or following a plug-and-play board insertion, nodes negotiate the lane width of the connections they will use. Thus, a device with a 16-lane interface can communicate at its full bandwidth with 16- and 32-lane devices and at decreasing bandwidths with devices having fewer lanes.

Maintaining application-software compatibility with the parallel PCI bus while operating over a variable-width serial bus requires several modifications at the lower levels of the communications model. At the physical layer, for instance, a PCIe link includes hardware that stripes data bytes across the available lanes for transmitting and reassembles them at the receiving end. The serial-data packets contain header information that allows the reassembly of packets in the correct order even if they arrive at different times, eliminating the effects of skew between lanes.

The link layer handles the detection of and response to transmission errors in the serial links. Each packet sent to the physical layer includes a packet sequence number and a CRC (cyclic-redundancy-check) character for error detection. If a transmission error occurs, the link-layer hardware automatically resends the damaged packets.

The transaction layer interacts with the system software to convert a software agent's memory-mapped read and write transactions that target the PCI bus into command and data packets that pass to the PCIe link layer. Each packet includes a unique identifier that associates it with a given transaction. This identifier allows the layers to route the outgoing transmission to the appropriate node and to route responses to the appropriate software agent.

The transaction layer supports memory, I/O, and configuration spaces within the system and can handle both 32-bit and extended 64-bit addressing. These capabilities make PCIe able to fully mimic the load-store architecture and flat memory space of a PCI bus, so none of the higher software levels in the system need alteration. Applications, operating systems, and hardware drivers developed for PCI all work unaltered with PCIe hardware.

In addition to the memory-mapped transactions over the parallel bus, the PCI bus includes sideband signals, such as interrupts, power management, and reset. PCIe handles these functions by incorporating them into a message space. The PCIe-interface hardware converts such sideband signals into data links along with the command and data transactions. The sideband signals are reinterpreted as control lines at the node. Thus, PCIe provides "virtual wires" to replace the interrupts and other control lines of PCI.

The structure of PCIe does more than simply mimic the PCI bus to legacy software, however. It also offers new features that new software can exploit. One such feature is the ability to assign attributes such as "relaxed-ordering" or "priority" to packets. The system can use these attributes when managing the switch and resolving contention between nodes for I/O resources. Thus, PCIe can support the QOS (quality-of-service) features that communications applications such as VOIP (voice over Internet Protocol) require.



With separate control and data planes, a communications device has two sets of bus requirements. The control plane handles access to control registers and counters and the movement of data blocks into and from table memory. This situation means that the bus traffic typically flows between a central housekeeping CPU and individual data-handling nodes. This type of traffic is a good match with the memory-mapped-addressing structure that PCIe inherited from the PC architecture. As a result, PCIe is a strong candidate for the use of control-plane buses within communications devices. Similarly, PCIe matches the needs of merged control and data planes. Its high performance supports the necessary data rates, and the structure maps well to the need of the control plane to access the data headers before initiating data movement.

The data plane has a different set of requirements, however. Its primary need is to move data at high speed from any input port to any output port that the device provides. A traditional multidrop parallel bus, such as PCI, is inefficient at meeting this need. Although multidrop buses allow cross-connectivity, only one pair of devices can communicate at a time. PCIe, however, uses switches in its datapaths. These switches can be nonblocking—that is, able to connect multiple pairs of ports at the same time (Figure 2). This ability means that PCIe potentially offers a suitable data-flow structure for handling data-plane requirements. PCIe fares less well with some other architectural requirements of data planes, however. One of the most significant is the need for redundancy.

THERE CAN BE ONLY ONE

Communications systems, particularly at the higher tiers, need extremely high reliability. The complete failure of a node can bring the network to a halt. At best, a node failure represents lost revenue and a set of angry customers. As a result, communications systems require redundant designs that allow multiple CPUs to operate concurrently.

PCIe allows only a single point of control. In a PC, the central processor initializes and controls all other elements in the system, forming a single system

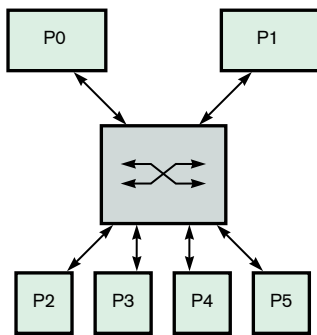


Figure 2 Switches in PCI Express, configured as two-by-eight lanes (top) or four-by-four lanes (bottom), connect pairs of ports together and, if nonblocking, can support two or more simultaneous connections, such as P1 to P3 and P2 to P5, as long as there is no port contention (courtesy NEC).

complex. It does not provide for multiple independent processors to have access to resources within the complex. If one of the other system elements is also a processor, that second processor must be a slave to the central processor; it cannot initiate any transactions to peripheral devices on its own. Similarly, PCIe does not support having multiple processor complexes share access to system resources.

Recognizing this limitation in PCIe, the PCI-SIG (PCI Special Interest

Group) that governs the PCIe standard is actively working on a solution. Its approach, IOV (I/O virtualization), allows multiple processors and processor complexes to share peripherals and other system endpoints. The virtualization will be available at two levels. The peripheral or endpoint itself provides the first level: single-root virtualization. At this level, the endpoint provides its resources, including interrupts and direct-memory access, independently to each processor. In second-level IOV, the endpoint and the switch have mechanisms that allow multiple processor complexes to share a common endpoint resource.

IOV is still under development, however. Developers seeking to add system redundancy within the current PCIe specifications can choose one of two approaches. One is to use multiplexers that connect the primary- and the backup-system elements to the PCIe switches in a dual-star topology (Figure 3). The other approach is to use a nontransparent switch, such as those from IDT, Intel, and PLX Technology.

MULTIPROCESSING NEEDS

A nontransparent switch takes packets coming from a processor complex on one side and converts the addressing elements in the header to map the packets to the processor complex on the other side. During power-up, a PCIe

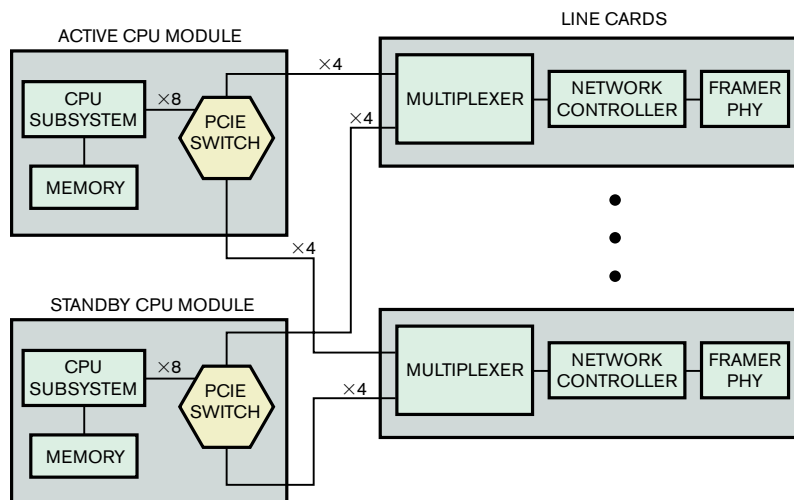


Figure 3 Implementing high reliability with the current version of PCI Express, which allows only one master processor, may require the use of multiplexers to isolate redundant processors (courtesy IDT).

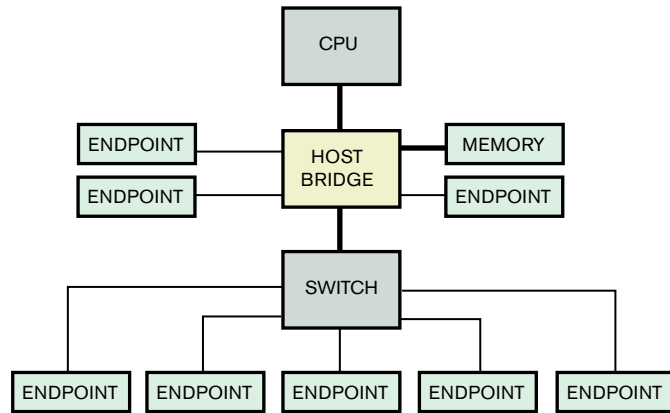


root processor initializes and enumerates the resources in its complex so that it can communicate with them through memory accesses. If two root processors were sharing the same bus, the two would generate conflicting address maps. By providing address translation, the switch effectively hides the existence of one root CPU from the other and allows each root CPU to use the address mapping that it generated to access the resources. The switch also resolves contention for resources. The major drawback of the nontransparent bridge is that no standard for its implementation exists. Thus, each vendor's product has its own unique software impact.

The limitations of PCIe in a data-plane application make it a weak contender for replacing proprietary fabrics in this aspect of communications designs. Yet, the economic factors that prompted interest in PCIe have begun to initiate a replacement of proprietary fabrics. The replacement is typically switched Ethernet.

Ethernet also represents a technology that enjoys lower cost based on high volume. Further, it offers a standardized way of implementing communications from multiple hosts to shared endpoints. It also has substantial data-handling capacity; 10-Gbps switch devices from Fulcrum Microsystems are on the market. A host of products provides bridges from Ethernet to other communications protocols, simplifying the design of a multiprotocol system. These attributes have begun prompting communications-system vendors to begin designing access-tier systems, such as DSLAMs (digital-subscriber-line-access multiplexers) using PCIe for the control plane and switched Ethernet for the data plane.

Supporters of PCIe point out, however, that the economics and performance road map of PCIe may eventually push Ethernet, as well as proprietary fabrics, off the data plane. The high volume of Ethernet production is currently in the 1-Gbps devices. PCIe now offers eight- and 16-lane devices, offering data rates as high as 40 Gbps, which vendors developed for the graphics needs of PCs. PCIe supporters see that technology quickly becoming available for switches, bridges, and additional endpoint peripherals.



PCI Express replaces the parallel PCI bus with switched serial links containing one or more lanes (courtesy Intel).

Still, these products represent only a potential. Currently, only a few PCIe devices on the market do not specifically target PC and graphics applications. These products include mostly switches and bridges. Companies such as IDT, NEC, Texas Instruments, and PLX Technology are offering both transparent and nontransparent switches ranging from two-port, four-lane devices to eight-port, 48-lane devices. PCIe bridges are also available from these companies, as well as from AMCC, Intel, and others. The bridges include both upstream and downstream connections between PCIe and PCI, PCI-X, and Ethernet.

Opportunity also exists on the custom-design front. FPGA devices from Lattice Semiconductor are available with PCIe interface cores from Northwest Logic and PHY (physical)-layer components from Genesys Logic. NEC also offers PCIe cores, including controller and PHY cores. Even the EDA vendors have started getting into the custom-PCIe market, with Cadence offering both design IP (intellectual property) and verification tools for PCIe designs.

With the current emphasis in the market on PCs and graphics applications, the use of PCIe in communications designs is still at its earliest stages. The potential is there, however, in both architecture and performance, to fill many niches in communications control. The efforts of the PCI-SIG and individual companies to address the architectural mismatches that still exist between the processor-centric PC and data-centric needs of communications systems will only improve the situation. PCIe may not dominate communications-system designs, but it could well become a strong player in that market. **EDN**

AUTHOR'S BIOGRAPHY

Contributing Technical Editor Richard A Quinnell has been covering technology for more than 15 years after an equally long career as an embedded-system-design engineer.

FOR MORE INFORMATION

- | | |
|---|---|
| AMCC
www.amcc.com | Lattice Semiconductor
www.latticesemi.com |
| Cadence
www.cadence.com | NEC Electronics America
www.necel.com |
| Fulcrum Microsystems
www.fulcrummicro.com | Northwest Logic
www.nwlogic.com |
| Genesys Logic
www.genesyslogic.com | PCI-SIG
www.pci-sig.com |
| IDT
www.idt.com | PLX Technology
www.plxtech.com |
| Intel
www.intel.com | Texas Instruments
www.ti.com |

MORE AT EDN.COM

For more on PCI Express, visit www.edn.com/article/CA601851.

For Contributing Technical Editor Richard Quinnell's recent article "Clash of the wireless-USB standards," go to www.edn.com/article/CA6363903.