HP local I/O strategy for ProLiant servers

technology brief

Abstract

This paper is intended to clarify HP local I/O strategy for ProLiant servers. HP server I/O priorities reflect the demands of IT managers. HP supports I/O technologies that protect customer investments and extend I/O performance, flexibility, and reliability. This paper describes today's parallel bus architecture and explains why the industry is beginning a transition to PCI Express. It provides details on PCI Express technology and information to help customers determine which I/O technology to use.

Introduction

There is constant pressure on the Peripheral Component Interconnect (PCI) bus to meet the I/O bandwidth demands of data traffic traveling between increasingly powerful CPUs and I/O devices. The PCI standard has continued to meet the needs of CPUs and I/O devices by increasing performance while maintaining backward compatibility. Historically, server local I/O performance has doubled almost every two years with a new phase of the PCI standard.

With the advent of faster and more complex I/O devices, PCI-X 1.0, which defined device speeds of PCI-X 66 and PCI-X 133, was introduced to extend the performance and RAS (reliability, availability, and serviceability) features of PCI technology in servers. PCI-X 1.0 leveraged the wide acceptance of the PCI bus and doubled the maximum bandwidth to 1066 MB/s with a 64-bit, 133-MHz bus. In 2002, the PCI SIG extended the PCI-X 1.0 specification with PCI-X 2.0. The PCI-X 2.0 specification is a smarter, more robust I/O technology that doubles and quadruples PCI-X bandwidth with two additional device speeds: PCI-X 266 and PCI-X 533.

In 2002, the PCI-SIG also introduced a new physical implementation of PCI, called PCI Express, which was originally optimized for desktop applications. With anticipated bandwidths in excess of 4 GB/s, PCI Express delivers the performance required for next-generation 10-Gb Ethernet adapters.

This paper begins with a summary of today's parallel bus system and then describes PCI Express technology. This paper also provides deployment and timeline information to help customers develop a transition plan.

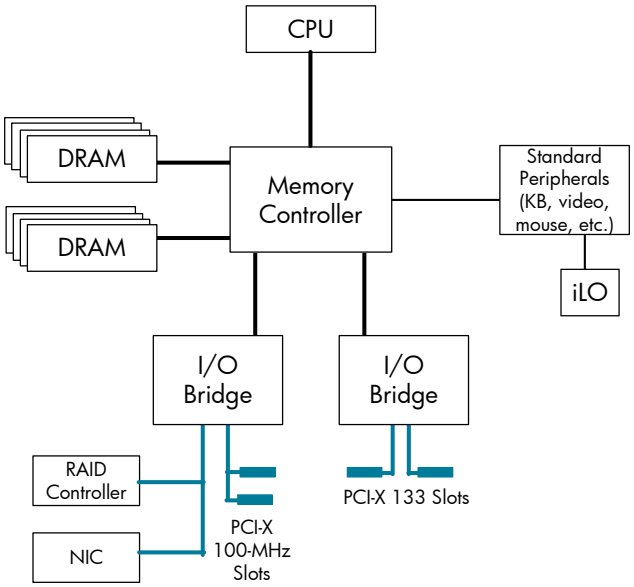
Today's parallel bus system

Today's parallel bus system (PCI and PCI-X) uses bit-parallel, bi-directional, multi-drop connections running at relatively low frequency (Figure 1). PCI and PCI-X send 32 or 64 data bits plus control signals over the parallel bus. The number of pins required corresponds to the width of the bus: a 32-bit bus requires 50 to 60 pins (data plus control) and a 64-bit bus requires approximately 100 pins.

The number of devices sharing one bus shrinks as the speed of the bus increases. For example:

- 33-MHz PCI supports 5 or 6 slots.
- 66-MHz PCI-X supports 4 slots.
- 100-MHz PCI-X supports 2 slots.
- 133-MHz and higher PCI-X supports 1 slot.

Figure 1. Typical parallel bus architecture



The PCI and PCI-X specifications maintain full forward and backward compatibility from conventional 3.3-V, 33-MHz PCI to PCI-X 533. In other words, existing conventional 3.3-V, 33-MHz PCI, conventional 66-MHz PCI, PCI-X 66, PCI-X 133, PCI-X 266, and PCI-X 533 add-in cards will operate in any PCI (3.3 V) or PCI-X system. (PCI-X 133 and 66-MHz PCI systems support universal add-in cards and 3.3-V PCI cards; however, they do not support 5-V-only PCI cards, which operate only at 33 MHz.)

Today's parallel bus system is highly cost effective with its small silicon footprint and low-frequency design rules. However, economics are changing. Backward compatibility becomes increasingly expensive as speeds increase. Higher bus speeds require physically shorter connections, which means fewer potential slots. And as Moore's Law dictates, silicon cost drops over time much faster than package and pin count. As a result, the industry is beginning to phase in PCI Express technology, which will prove to be a more cost-effective solution to provide the bandwidth required for future peripheral devices.

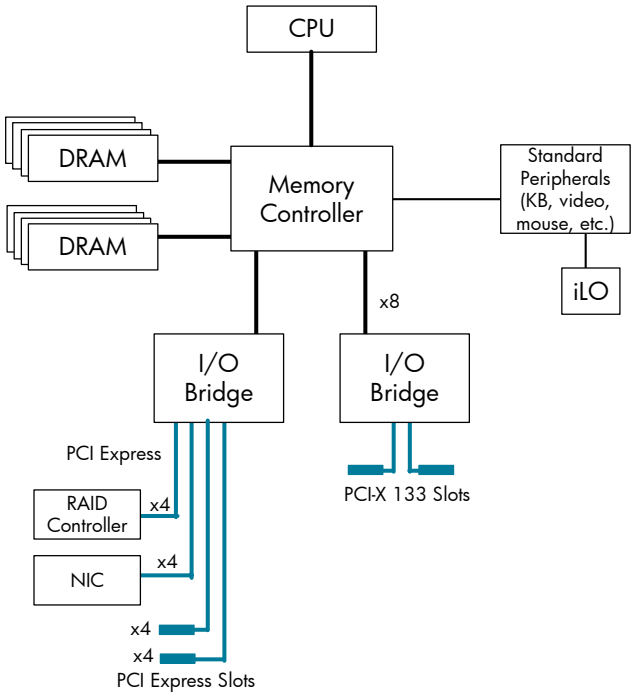
PCI Express technology

PCI Express 1.0 has a signaling rate of 2.5 Gb/s per direction per lane, for a combined receive and transmit bandwidth of 500 MB/s. Multiple lanes can be combined to form higher bandwidth links. For example, four lanes can be combined to form a 2-GB/s link, designated "x4" (pronounced "by four"). Future PCI Express Gen2 is expected to use a signaling rate at least twice as fast and to add features such as mandatory end-to-end cyclic redundancy check (CRC) and larger, more efficient block sizes.

Server topology

The PCI Express architecture (Figure 2) provides point-to-point connections between devices. PCI Express sends the data serially, one bit after the other, over each link rather than sending the data in parallel, one bit beside the other, as in PCI-X. Therefore, PCI Express allows use of fewer pins.

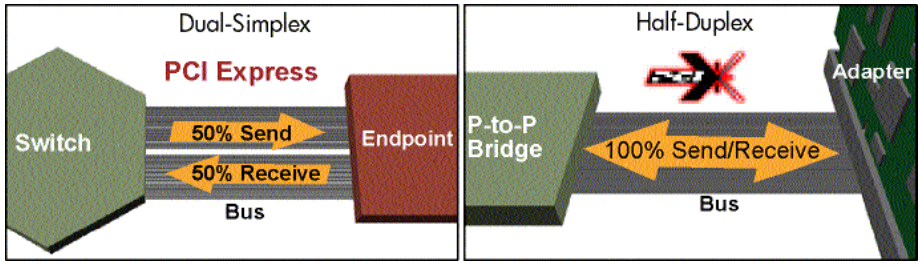
Figure 2. PCI Express architecture



Dual-simplex lanes

A PCI Express serial link consists of one or more dual-simplex lanes. Each lane contains two pairs of wire conductors (a send pair and a receive pair) to transmit data at the signaling rate in both directions simultaneously. The maximum bandwidth of dual-simplex buses is often specified as the sum of the transmit and receive channels (Figure 3 left). In contrast, PCI-X uses a half-duplex scheme where the full bus bandwidth is used either to transmit or to receive data (Figure 3 right).

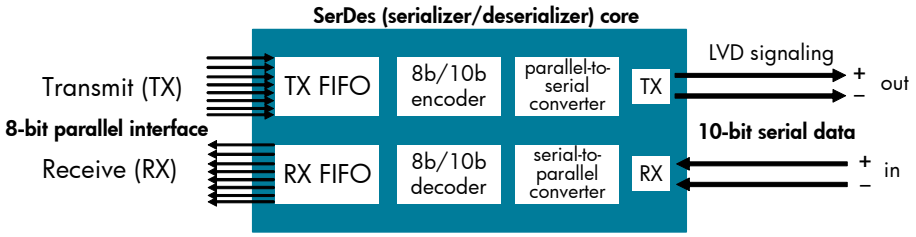
Figure 3. Dual-simplex communication and half-duplex communication



8b/10b encoding

Serial communication requires a device to convert parallel data into a serial bit stream and vice versa. This device, called a serializer/deserializer (SerDes), contains a parallel digital interface, First-In-First-Out (FIFO) caches, 8 bit/10 bit (8b/10b) encoder and decoder, a parallel-to-serial converter, and a serial-to-parallel converter (see Figure 4). The 8b/10b encoder converts each 8-bit data byte to a 10-bit transmission symbol, which enables clocking information to be encoded into the data stream. Although this adds about 20 percent embedded overhead to the data stream, it is the most common data signaling method when the signaling rates exceed 1 GHz.

Figure 4. The SerDes core integrates 8b/10b coding and decoding logic.



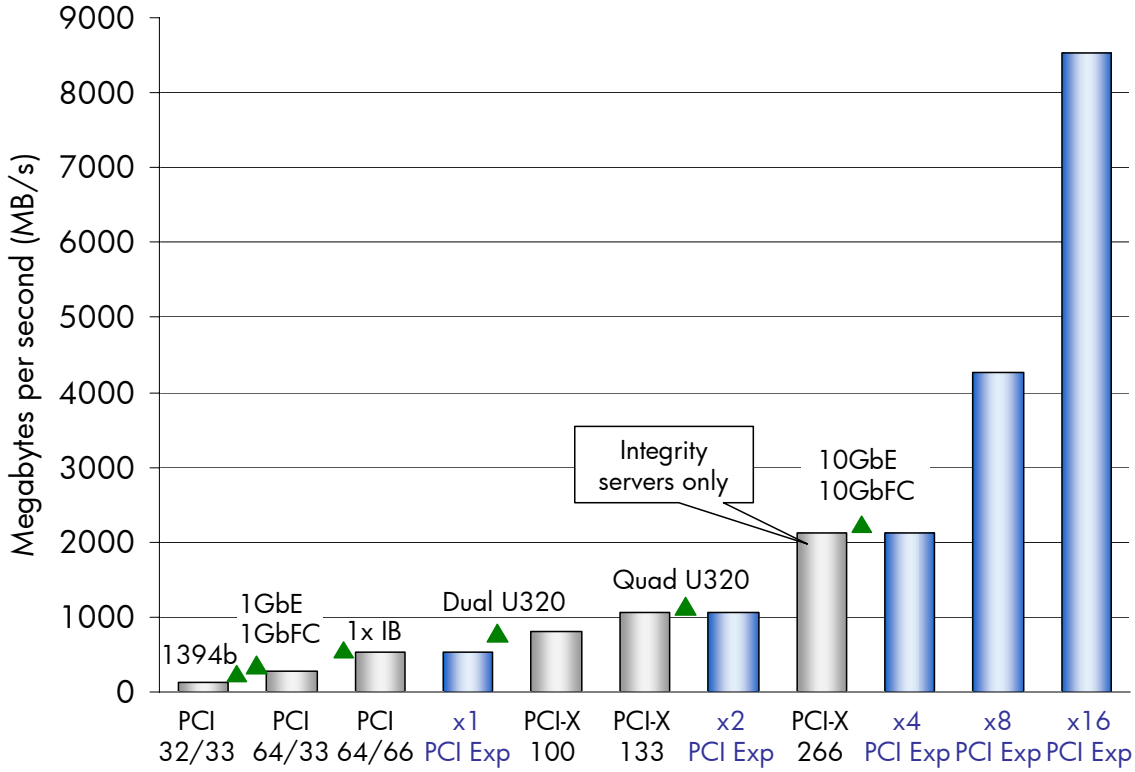
Performance

As previously mentioned, the bandwidth of each PCI Express link can be linearly scaled by increasing the number of lanes. PCI Express Gen1 has a signaling rate of 2.5 Gb/s per lane in each direction, resulting in a unidirectional bandwidth of 250 MB/s after accounting for 8b/10b encoding overhead. Combining eight lanes provides a total unidirectional bandwidth of 2 GB/s, which would be sufficient to support the full wire speed of 10-Gb/s technologies such as 10-Gb Ethernet (about 1.25 GB/s, unidirectionally).

PCI Express Gen2 will increase the signaling rate from 2.5 Gb/s to at least 5 Gb/s. At a signaling rate of 5 Gb/s, PCI Express Gen2 will have a unidirectional bandwidth of 500 MB/s per lane (after factoring encoding overhead). With this bandwidth, the PCI Express Gen2 x4 link will support a 10-Gb Ethernet device, and a x8 link will support the high-speed server I/O peripherals that are expected to exist through the end of the decade.

Figure 5 illustrates the bandwidths of PCI, PCI-X, and PCI Express, which will provide sufficient I/O bandwidth migration for server I/O growth requirements. PCI Express bandwidth is shown using the sum of the receive and transmit bandwidths.

Figure 5. I/O bandwidth of PCI, PCI-X, and PCI Express



Backward compatibility

PCI Express is compatible with PCI at the software layers (operating system, applications, and drivers). New software is required to take advantage of new and extended PCI Express architecture features. The PCI Express configuration uses standard PCI plug and play mechanisms and PCI Hot Plug mechanisms. PCI Express is not compatible with PCI at the board or connector level.

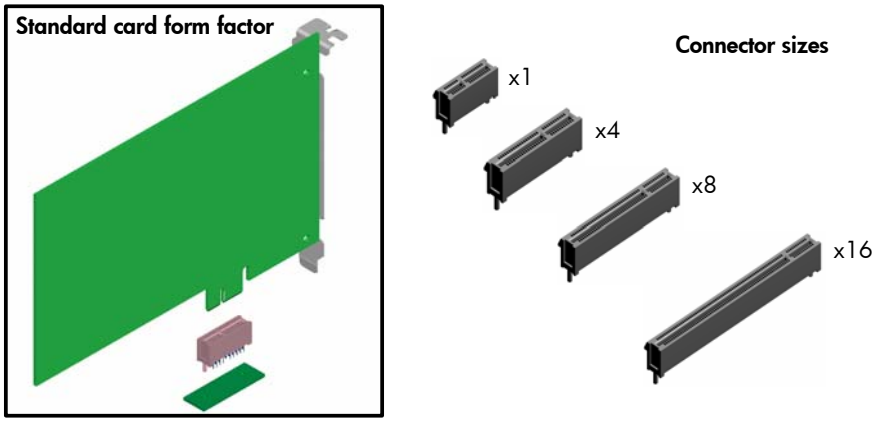
Form factors

The PCI Express specification includes a range of form factors and connector sizes. Connector sizes include the following:

- x1 connector for desktop I/O
- x4 and x8 connectors for server I/O
- x16 connector for desktop graphics

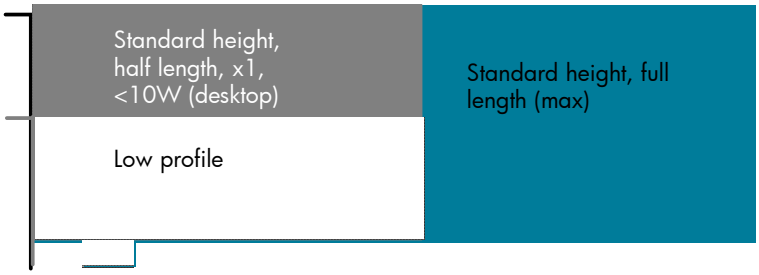
The standard card form factor is similar to the PCI card. Figure 6 illustrates the standard card form factor and connector sizes.

Figure 6. Standard card form factor and connector sizes



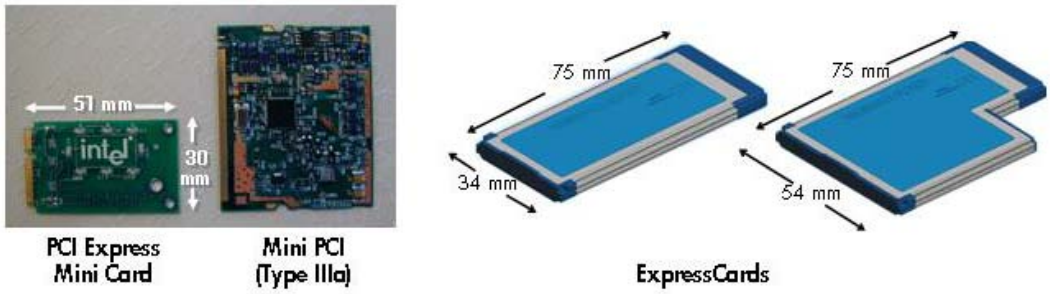
The standard PCI Express card will be the standard height, full length, and same size as today's PCI cards. However, the desktop form factor provides additional options. Some desktop cards with a x1 connector will be standard height but a maximum of half length. These cards are limited to a maximum power of 10 W. A low-profile card, which will be half length, half height, and limited to 10 W, will also be available. Figure 7 illustrates the standard card sizes.

Figure 7. Standard card sizes



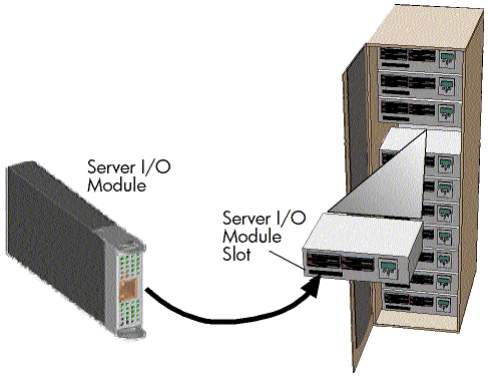
The PCI Express specification also defines two other PCI-like form factors. A PCI Express Mini Card, similar to the Mini PCI card will include a x1 connector for use in portable products. The ExpressCard is similar to PCMCIA cards and will include a x1 PCI Express connector and a USB 2.0 connector for use in desktop and portable products. (See Figure 8.)

Figure 8. PCI Express Mini Card (left) and Express Card (right) form factors



In addition to these evolutionary PCI-like form factors, the PCI Express specification will define a revolutionary server I/O module (SIOM). The SIOM has the potential to add customer value because it allows inserting and removing cards from outside the server chassis. The SIOM is mechanically rugged and it simplifies insertion and removal by providing guide rails with a blind-mate connector at the end. It also protects the card with a physical shell and provides well-defined airflow characteristics. However, because of the radical mechanical differences between SIOM and standard PCI and PCI Express cards, it is unlikely that SIOM will be implemented in standard server designs. The most likely application for SIOM will be in enclosures specifically designed for modular I/O expansion, such as an enclosure full of SIOM cards (Figure 9).

Figure 9. Server I/O module



Card interoperability

The PCI Express specification provides interoperability features. For example, up-plugging (plugging "small cards" into "big slots") is allowed. However, down-plugging (plugging a "big card" into a "small slot") is physically prevented. Down-shifting is allowed in one case so that a x8 connector can be electrically wired (or "plumbed") with a x4 link. Table 1 identifies the possible interoperability scenarios.

Table 1. PCI Express add-in card interoperability

	x1 slot	x4 slot	x8 plumbed x4 slot	x8 slot	x16 slot
x1 card	Required	Required	Required	Required	Required
x4 card	Not allowed	Required	Required	Allowed	Allowed
x8 card	Not allowed	Not allowed	Allowed	Required	Allowed
x16 card	Not allowed	Not allowed	Not allowed	Not allowed	Required

ProLiant server transition strategy

HP is committed to ensuring a smooth transition between I/O standards for customers. HP is beginning to introduce PCI Express technology into next-generation ProLiant servers to coexist with PCI-X 133. This will allow ProLiant server customers to install new higher bandwidth cards (for example, dual 4-Gb Fibre Channel, x4 InfiniBand, and 10-Gb Ethernet) using PCI Express technology beginning in 2004. At the same time, the continued commitment from HP to PCI-X will allow ProLiant server customers to connect to their existing I/O cards. HP will begin phasing PCI Express into Smart Array, ProLiant NIC, and Fibre Channel SAN cards as the technology matures.

PCI Express slots first became available in select ProLiant servers that were introduced in the second half of 2004. ProLiant ML servers have a mix of PCI-X and PCI Express slots, while ProLiant DL servers have optional riser cards to provide different combinations of PCI-X and PCI Express slots.

Customers who require peripheral bandwidth in excess of 1 GB/s will likely find those peripherals implemented with PCI Express technology. It is expected that most device manufacturers will use PCI Express technology for higher bandwidth peripheral devices such as 10-Gb Ethernet and 10-Gb Fibre Channel. For peripherals that require bandwidth less than 1 GB/s, PCI and PCI-X will remain viable technologies and continue to provide backward compatibility. It is anticipated that device vendors will manufacture peripherals with bandwidth requirements in the 1 GB/s range in two versions, providing options for customers to choose either PCI Express or PCI-X technology.

Conclusion

As the customer requirements for higher bandwidth in the I/O subsystem continue to increase, the industry is beginning a transition to PCI Express technology to provide that bandwidth. HP will implement PCI Express in ProLiant servers while maintaining a commitment to PCI-X technology. PCI Express and PCI-X technology will coexist in ProLiant servers, giving customers flexibility in their choice of technology and providing a path for transition to higher bandwidth peripherals.

HP has been a leader in the development and implementation of industry-standard I/O technology and continues to be an active member of the PCI Special Interest Group. HP is committed to delivering industry leading products that meet customer requirements for flexibility, performance, and investment protection. HP demonstrates that commitment by providing a clear transition path to PCI Express technology.

For more information

For more information on PCI Express or PCI-X, visit the PCI SIG home page at http://www.pcisig.com/home. For more information about HP ProLiant servers, visit http://h18004.www1.hp.com/products/servers/platforms/.

Call to action

To help us better understand and meet your needs for ISS technology information, please send comments about this paper to: TechCom@HP.com.

TC040901TB, 09/2004