TECHNOLOGY BRIEF

September 1999

Compaq Computer Corporation

Prepared by
ISSD Technology Communications

CONTENTS

PCI-X: An Evolution of the PCI Bus

PCI-X is an evolutionary bus architecture based on the prevalent PCI bus. Compaq led a special workgroup to develop PCI-X as an enhancement to the conventional PCI bus. PCI-X is an industry-standard I/O interconnect that exceeds a raw bandwidth of 1 gigabyte per second. The 64-bit, 133-MHz interconnect protocol achieves this performance through the use of a register-to-register design and new protocol enhancements such as the attribute phase and split transactions that allow more efficient use of the bus. PCI-X technology is backward compatible with conventional PCI systems at the system, device driver, and the adapter level. Conventional PCI adapters will operate in PCI-X systems, and vice versa; however, when a PCI-X adapter is placed on a conventional PCI bus, it is limited to conventional PCI speeds. PCI-X is poised to become the next standard I/O bus interconnect inside servers, and is designed to integrate smoothly with more complex architectures such as switched-fabric I/O architectures.

Please direct comments regarding this communication to the ISSD Technology Communications Group at this Internet address: TechCom@compaq.com

COMPAQ

NOTICE

PCI-X: An Evolution of the PCI Bus
First Edition February 1999
Second Edition September 1999
TC990903TB

NEED FOR A FASTER I/O BUS

Business-critical applications continue to demand greater bandwidth and shorter response times from the input/output (I/O) subsystems of enterprise computers. As faster and more complex I/O devices such as Gigabit Ethernet, Fibre Channel, Ultra3 SCSI, and multi-port network interface controllers (NICs) appear on the market, conventional PCI bus technology operating at 33 MHz frequently becomes a performance bottleneck.

At the same time, the computing model is demanding more scalability, availability, and reliability as the enterprise becomes more distributed. To fulfill these needs, high-bandwidth architectures such as system area networks are emerging. System area networks are used to connect distributed resources such as clustered servers, I/O, and storage. Using the PCI bus as a backbone interconnect for these high-bandwidth system architectures will stretch the PCI bus to its limit.

Optimal system performance requires a balance between the processor-to-memory subsystem and the I/O subsystem. Since the introduction of the PCI bus in 1992, the internal clock frequencies of processors have increased dramatically, from less than 100 MHz to more than 500 MHz; and published roadmaps forecast frequencies that are greater than 1 gigahertz (GHz). The processor-to-memory bus (front-side bus) is now 64 bits wide and operates in the range of 100 to 333 MHz for X86 and Alpha processors. Although the 64-bit, 66-MHz PCI extension exists, the PCI frequency remains at 33 MHz for the *vast majority* of systems and adapters (Figure 1).

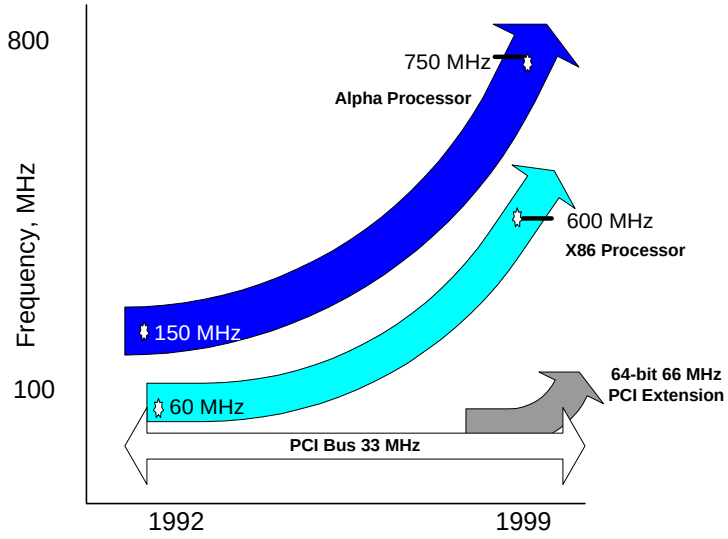

Figure 1: Schematic representation of difference in frequency between processors and the I/O bus.

To break this I/O bottleneck, both system and I/O adapter designers are migrating their designs to the 64-bit, 33-MHz PCI bus, which provides a peak bandwidth of 266 megabytes per second (MB/s). While the *PCI Local Bus Specification Revision 2.2* allows a 66-MHz PCI bus, the specification has many technical design challenges that have slowed its implementation. Even when system designers overcome these challenges, the 66-MHz PCI bus at its peak bandwidth of 533 MB/s (Table 1) is not adequate for long-term needs such as multi-port NICs with Gigabit Ethernet. For example, a four-port Gigabit Ethernet NIC, with each port capable of 1 gigabit per second, or 125 MB/s, would overwhelm the 64-bit, 66-MHz PCI bus bandwidth by using essentially all available bandwidth.

Table 1: Bandwidth of bus in PCI and PCI-X modes

Specification	Bus Width (bits)	Bus Frequency (MHz)	Maximum Peak Bandwidth (MB/s)
PCI 2.2	32	33.3	133
PCI 2.2	64	33.3	266
PCI 2.2	64	66.6	533
PCI-X 1.0	**64**	**133.3**	**1066**

Note: Bus frequency is commonly referred to as an integer rather than the actual decimal frequency. For example, 33 MHz rather than 33.3 MHz. This convention is used throughout this technology brief.

WHAT PCI-X IS

PCI-X technology leverages the wide acceptance of the PCI bus and provides an evolutionary I/O upgrade to conventional PCI. PCI-X technology increases bus capacity to more than eight times the conventional PCI bus bandwidth — from 133 MB/s with the 32-bit, 33-MHz PCI bus to 1066 MB/s with the 64-bit, 133-MHz PCI-X bus. It enhances the PCI protocol to develop an industry-standard interconnect that exceeds a raw bandwidth of 1 gigabyte per second (GB/s) and will meet upcoming bandwidth needs of enterprise computing systems. PCI-X provides backward compatibility with the PCI bus at both the adapter and system level.

Compaq, Hewlett-Packard, and IBM collaborated to increase the I/O performance level of industry-standard servers in enterprise applications and presented a nearly complete specification to the PCI Special Interest Group (SIG) for standardization. Compaq chaired the SIG workgroup that finalized the PCI-X specification and has garnered support for PCI-X from chipset suppliers and independent hardware vendors within the PC industry.

PCI-X TECHNOLOGY OVERVIEW

PCI-X enables the design of systems and devices that can operate at bus frequencies of up to 133 MHz using a 64-bit or 32-bit bus width, a significant improvement in performance beyond that of conventional PCI systems. The performance improvements are a result of two primary differences between conventional PCI and PCI-X: higher clock frequencies made possible by the register-to-register protocol and new protocol enhancements such as the attribute phase and split transactions.

Conventional PCI Protocol

The conventional PCI bus uses an immediate protocol rather than a register-to-register protocol. With a conventional PCI device, the following steps occur when the device switches a control signal (Figure 2):

1. On the rising clock edge, the device switches the signal to a high or low state onto the PCI bus.
2. The signal propagates across the bus (propagation delay).
3. During the same clock cycle, the receiving device decodes the signal to determine whether the signal is for the receiving device and to determine if it must respond by switching one of its outputs.
4. The receiving device responds immediately, that is, in the next clock cycle.

With a 33-MHz clock frequency, the time allocated for the decode logic is 7 nanoseconds (ns) of the total 30-ns clock cycle time. At 33 MHz, this is sufficient time for the receiving device to

respond on the next rising clock edge. However, as the bus frequency is doubled to 66 MHz (with a clock cycle time of 15 ns), the number of nanoseconds available to perform this logic is reduced to 3 ns. These time constraints of the conventional PCI specification make it difficult to design a conventional PCI bus or adapter for 66 MHz.

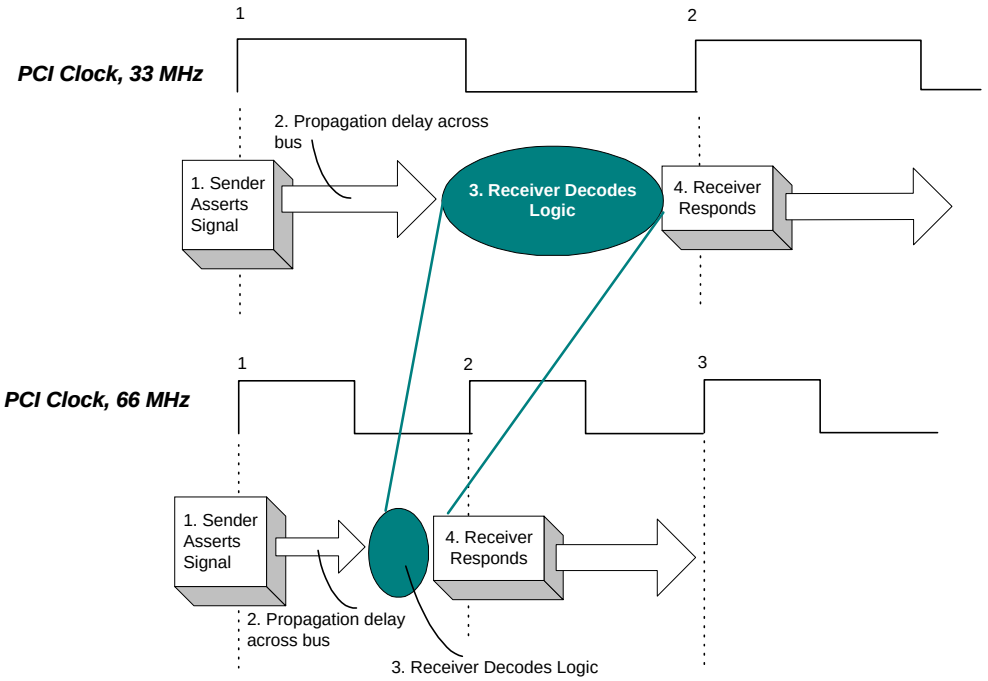

Figure 2: With conventional PCI, the time available to decode a transaction decreases as the bus frequency increases from 33 MHz to 66 MHz.

Register-to-Register Protocol

With the PCI-X register-to-register protocol, the following steps occur (Figure 3):

1. On the rising clock edge, the device switches the signal to a high or low state onto the PCI-X bus.
2. The signal propagates across the bus.
3. The signal is sent to a register, or *flip-flop*, that holds the signal state until the next clock cycle.
4. The receiving device has a full clock cycle to decode the signal and determine the appropriate response.
5. The receiving device responds two full clock cycles after the sending device first switched the signal.

Flip-flop: A digital logic circuit that switches between two states, depending on its inputs. It can be considered as single-bit memory storage.

PCI-X eases the timing constraints by allowing an entire clock cycle for the decode logic to occur. The net difference is that PCI-X transactions generally require one clock cycle more than conventional PCI transactions. A write transaction that completes in nine clock cycles for conventional PCI will complete in ten clock cycles for PCI-X.[1]

[1] This assumes a transaction with six data phases, with no initiator or target wait states, and identical target response timing for both PCI and PCI-X.

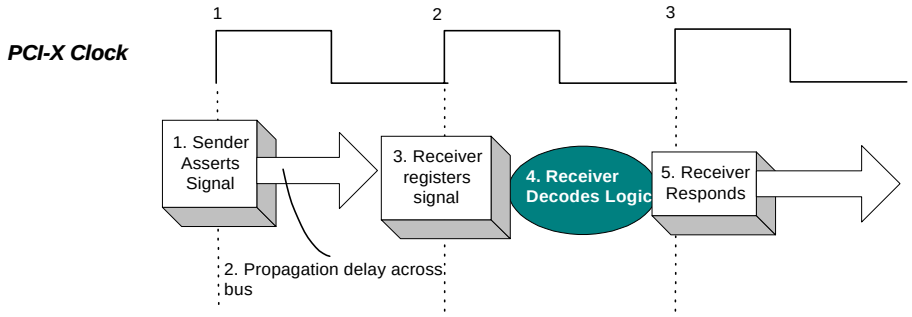

Figure 3: The PCI-X protocol allows an entire clock cycle for the decode logic to occur.

With the timing constraints reduced, it is much easier to design adapters and systems to operate at 66 MHz and greater. System designers can benefit from the eased timing constraints by choosing maximum performance with a single PCI-X slot at 133 MHz or maximum connectivity with additional slots on the PCI-X bus.

If a system designer chooses to increase the bus frequency, the actual time required to complete the transaction is greatly reduced, even though an additional clock cycle has been added (Figure 4). A transaction that takes nine cycles at 33 MHz will finish in 270 ns, while a PCI-X transaction that takes ten cycles at 133 MHz will finish in 75 ns, a *72 percent improvement* in transaction time.

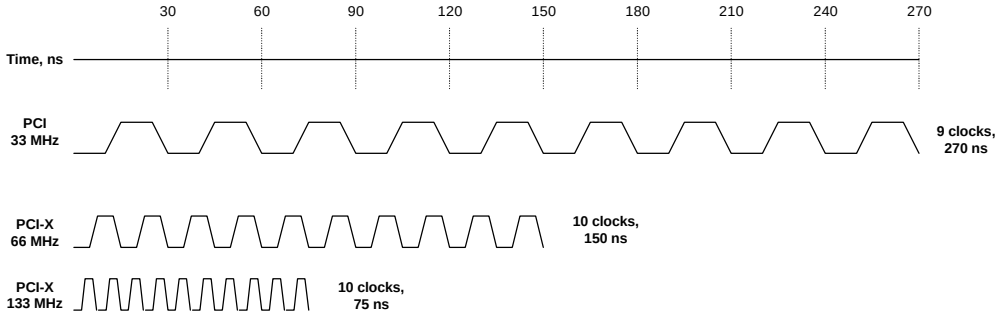

Figure 4: Actual time difference between PCI and PCI-X protocols for a typical write transaction.

Or, if a system designer chooses to keep the frequency at 66 MHz, additional slots can be added to the bus segment. Because more time is budgeted for a signal to propagate from one device to another, the signal can traverse a longer path across multiple slots.

ENHANCEMENTS FOR BUS EFFICIENCY

In addition to the ability to increase the bus frequency, the PCI-X protocol incorporates these best-of-class technologies to improve bus efficiency:

- Attribute Phase
- Split transaction support
- Optimized wait states
- Standard block size movements

Attribute Phase

The PCI-X protocol includes a new transaction phase called the attribute phase. The attribute phase uses a 36-bit attribute field that describes bus transactions in more detail than the conventional PCI

specification allows. It follows immediately after the address phase and contains several bit assignments that include information about the size of the transaction, ordering of transactions, cache snooping requirements, and the identity of the transaction initiator. The following enhancements are included within the attribute phase:

- Relaxed ordering
- Non-cache-coherent transactions
- Transaction byte count
- Sequence number

Relaxed Ordering

With the conventional PCI protocol, the host-to-PCI bridge or PCI-to-PCI bridge handles requests from multiple PCI devices in the order in which they are received. The bridge must perform transactions in this order because it cannot identify which transaction came from which device.

The attribute field within the PCI-X protocol includes a relaxed ordering bit. If the device driver or controlling software sets this bit, the transaction is permitted to pass previously posted transactions from other devices. The bridge can rearrange the transactions in the most efficient manner, depending on which PCI device or system memory port is available.

The use of relaxed ordering can improve efficiency in many applications. Relaxed ordering is especially important in applications such as audio or video streaming, where a delay in information would cause a noticeable interruption.

Non-Cache-Coherent Transactions

One of the design challenges for a multiprocessor system is maintaining a consistent view of memory during a transaction between the processors and the I/O subsystem. This is typically referred to as a cache coherent transaction. Because data is shared among multiple caches on the processor bus and main memory, there is the possibility that copies of the data may be inconsistent. When a device on the PCI bus writes or reads to main memory, the processor must perform a snoop operation to make sure the data does not exist in the processor cache memory. Every snoop cycle adds traffic to the bus and potentially limits the performance of the system.

PCI-X allows non-cache-coherent transactions in the form of a “don’t-snoop” bit within the attribute field. If the device driver or controlling software supports non-cache-coherent transactions, it can set the don’t-snoop bit on a transaction-by-transaction basis to improve performance. The device driver configures its device to initiate PCI-X transactions with this bit set and the PCI-X device informs the system cache controllers that they don’t have to query the processor, thereby reducing the volume of traffic on the processor bus.

Transaction Byte Count

With conventional PCI protocols, the bridge (host-to-PCI or PCI-to-PCI) fetches a default number of cache lines (typically one or two) for every data request. Because the bridge has no way of knowing how much data will be requested, it uses the default number of cache lines.

Sequence: One or more transactions associated with a single logical transfer.

With PCI-X, the bridge knows exactly how much data to fetch because the byte count is included in the attribute field. Each PCI-X transaction in a *sequence* identifies the total number of bytes remaining to be read or written in its associated sequence. This enables more efficient buffer management schemes in the bridge as well as more efficient utilization of the bus and other system resources.

Sequence Number

The sequence number uniquely identifies transactions that are part of the same sequence. It identifies the initiator and which bus segment the initiator resides on, among other things. The sequence number is used to increase efficiency in buffer-management algorithms.

Split Transaction Support

Conventional PCI protocol supports delayed transactions. With a delayed transaction, the device requesting data must poll the target to determine when the request has been completed and its data is available. With a split transaction as supported in PCI-X, the device requesting the data sends a signal to the target. The target device informs the requester that it has accepted the request. The requester is free to process other information until the target device initiates a new transaction and sends the data to the requester. Thus, split transactions enable more efficient use of the bus.

Optimized Wait States

Conventional PCI devices often add extra clock cycles, or wait states, into their transactions. The wait states are added to "stall" the bus if the PCI device is not ready to proceed with the transaction. This can slow bus throughput dramatically.

PCI-X eliminates the use of wait states, except for initial target latency. When a PCI-X device does not have data to transfer, it will remove itself from the bus so that another device can use the bus bandwidth. This provides more efficient use of bus and memory resources.

Standard Block Size Movements

With PCI-X, adapters and bridges (host-to-PCI-X and PCI-X to PCI-X) are permitted to disconnect transactions only on naturally aligned 128-byte boundaries. This encourages longer bursts and enables more efficient use of cache-line-based resources such as the processor bus and main memory. It also facilitates a more pipelined architecture within PCI-X devices.

PARITY ERROR HANDLING

In most conventional PCI systems, a data parity error that occurs on the PCI bus will cause a non-recoverable error in the operating system (OS) and the system must be rebooted. PCI-X adapters have an increased range of error-handling options should a data parity error occur. Depending on the ability of the device driver and OS, adapters may be able to:

- Reschedule and repeat the faulty transaction.
- Notify the user of the faulty transaction.
- Reinitialize the adapter and continue operating.
- Take the adapter offline before a catastrophic failure occurs.
- Report a fatal error to the OS so that the OS can shut down as a last resort.

BACKWARD COMPATIBILITY AND INTEROPERABILITY

PCI-X is backward compatible with the existing PCI bus at the adapter, device driver, and system level. A PCI-X adapter can operate in a conventional PCI system, and vice-versa.

The PCI-X protocol is compatible with existing device drivers and operating systems. If the adapter and system both support PCI-X, the new PCI-X adapter can immediately take advantage of the higher operating frequencies of a PCI-X system. By using the existing driver and device-programming model, a designer can decrease time to market and give customers the benefit of

higher performance bus bandwidth. However, to take full advantage of the PCI-X protocol enhancements such as relaxed ordering and non-cache-coherent transactions, changes are required to the device drivers and the OS.

Adapters

The PCI-X specification requires adapters to operate at conventional 33-MHz PCI frequency and mode when installed in conventional PCI systems. A PCI-X adapter may optionally operate at conventional 66-MHz PCI frequency. Conversely, if conventional PCI devices are installed on a PCI-X bus, the bus clock remains at a frequency acceptable to the conventional adapter, and other adapters on that bus segment are restricted to using conventional PCI protocol. This high degree of backward compatibility offers system vendors, adapter vendors, and customers the ability to choose how and when they will implement PCI-X to their best advantage.

The PCI-X specification defines two frequency design points for PCI-X adapters: PCI-X 66 MHz and PCI-X 133 MHz. *When operating in PCI-X mode*, PCI-X 66-MHz adapters have actual clock frequencies ranging from 50 to 66 MHz, and PCI-X 133-MHz adapters have clock frequencies ranging from 50 to 133 MHz.

Data Bus Width bits	Frequency MHz
32	66 or 133
64	66 or 133

Like conventional PCI adapters, PCI-X adapters can implement a 64-bit interface or a 32-bit interface. The two possible frequencies and two data width interfaces give a total of four possibilities for adapter design. This flexibility built into the PCI-X specification allows adapter designers to meet customer needs with varying combinations of high performance, ease of design, and cost.

PCI-X adapters are designed for a 3.3V I/O signaling level. Optionally, they can be designed as universal adapters, which operate at either 3.3V or 5V I/O signaling[2].

Systems

The operating mode and frequency of the PCI-X bus depend on the types of adapters installed on the bus (only PCI-X adapters or a mixture of PCI-X and conventional PCI adapters) and on the number of adapters installed on the bus.

Systems with Only PCI-X Adapters

If a PCI-X bus segment includes *only* PCI-X devices, the bus operates in PCI-X mode.

A PCI-X system automatically adjusts the bus frequency to match the frequency of the slowest adapter on that bus segment. For example, if the bus includes a 66-MHz PCI-X adapter, the maximum clock frequency of the bus segment is 66 MHz. PCI-X supports up to 256 bus segments, and each segment is initialized separately so that different operating frequencies can be used.

As with conventional PCI, system designers can optimize a PCI-X system for particular I/O bandwidth needs. To support more adapter slots on the PCI-X bus, system designers can lower the bus frequency to a minimum of 50 MHz[3]. For example, a PCI-X bus segment with two expansion slots would typically operate at 100 MHz, rather than at the full 133 MHz. A PCI-X bus segment with four expansion slots would typically operate at 66 MHz (Figure 5). Conventional PCI is generally limited to two slots per bus segment operating at 66 MHz.

[2] Universal adapters can operate at either 3.3V or 5V I/O signaling levels. The adapter must be keyed so that it can fit into either the 3.3V PCI connector or the 5V PCI connector. Note that the I/O signaling levels are independent of the power levels. For example, a 3.3V I/O adapter can operate at 5V power, and vice versa. These conditions are defined in the PCI Specification 2.2.

[3] PCI-X has a minimum operating frequency of 50 MHz, unlike conventional PCI that has a minimum operating frequency of 33 MHz.

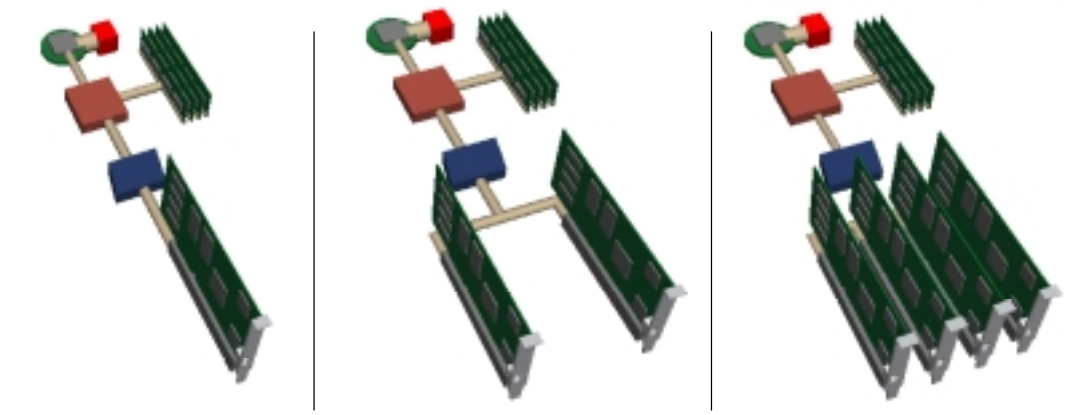

Figure 5: System designers can choose between maximum performance with a single PCI-X slot or maximum connectivity with more PCI-X slots per bridge.

Systems with Conventional PCI and PCI-X Adapters

A PCI-X system automatically switches between conventional PCI and PCI-X mode, depending on the type of adapters installed on the bus segment. If a PCI-X bus segment includes *any* conventional PCI adapters, that segment must operate in conventional PCI mode. Just as with conventional PCI, the operating frequency is adjusted to match that of the slowest device on the bus. If the system includes a 33-MHz adapter, the bus must operate at 33 MHz. If only conventional 66-MHz devices are present, a PCI bus optionally operates in conventional 66-MHz or 33-MHz mode. Table 2 shows several possible combinations of system and adapter frequencies.

Table 2: Maximum operating frequency of adapters, depending on adapter and system design.

	Bus Frequency MHz	Conventional PCI adapters			PCI-X adapters	
		33 MHz (5 V)	33 MHz (3.3V or Universal)	66 MHz (3.3V or Universal)	66 MHz (3.3V or Universal)	133 MHz (3.3V or Universal)
PCI system	**33**	33	33	33	33	33
	66		33	66	33 or 66	33 or 66
PCI-X system	**66**		33	33 or 66	66	66
	100		33	33 or 66	66	100
	133		33	33 or 66	66	133

Note: The shaded cells denote PCI-X areas of operation.

Compatibility with PCI Hot Plug

As the leading developer of both the PCI Hot Plug Specification[4] and PCI-X Specification, Compaq has ensured that the two specifications are compatible. From the initial stages of the specification, PCI-X has been carefully designed to incorporate PCI Hot Plug abilities. Hardware and software needs were considered so that the two technologies can work together seamlessly. Designers who want to implement PCI-X in their hot-plug systems will need to ensure that their hot-plug controllers and hot-plug system drivers meet requirements of the PCI-X specification.

[4] PCI Hot-Plug Specification 1.0, available from the PCI SIG at http://www.pcisig.com/

PERFORMANCE DATA

Even if a PCI-X system operates at the same 66-MHz frequency as a conventional PCI bus, the PCI-X bus enhancements can provide a significant performance improvement. Compaq analyzed the PCI-X protocol efficiency by comparing a model of a host-to-PCI bridge to a model of a host-to-PCI-X bridge, both operating at 64 bits and 66 MHz. The model of the PCI-X bridge added the PCI-X interface but left the queue manager, the memory controller, and the processor bus interface identical to those of the conventional PCI bridge. Figures 6 and 7 show the theoretical maximum and measured throughput as a function of the number of outstanding transactions. Using a read block size typical of existing SCSI controllers or NICs (512 bytes), the *PCI-X protocol improved performance by up to 14 percent* over the conventional PCI protocol (Figure 6).

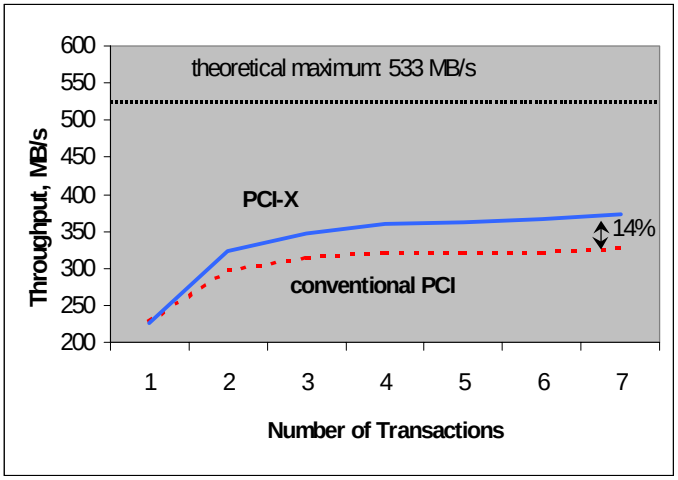

Figure 6: System-level comparison of PCI and PCI-X buses, both using 512-byte reads on a 64-bit, 66-MHz bus.

Using a much larger block size typical of Gigabit Ethernet or Ultra3 SCSI controllers (4 KB), the *PCI-X protocol improved performance by up to 34 percent* over the conventional PCI protocol (Figure 7).

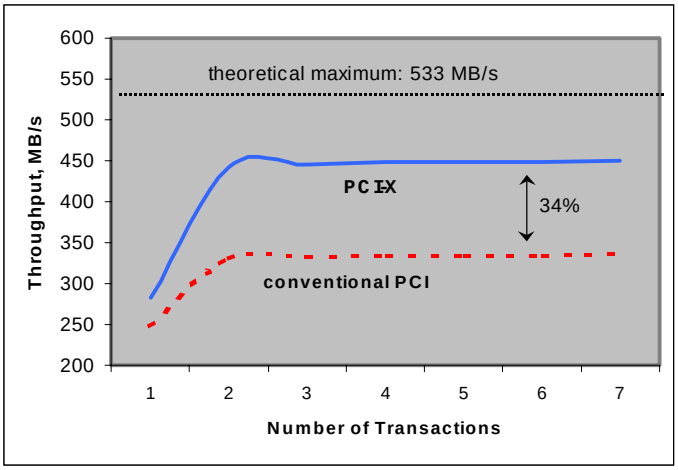

Figure 7: System-level comparison of PCI and PCI-X buses, both using 4-KB reads on a 64-bit, 66-MHz bus.

Figures 6 and 7 illustrate the performance improvements that are possible due to the increased efficiency of the PCI-X protocol. With the PCI-X register-to-register protocol easing the timing

constraints, the bus frequency can be increased above 66 MHz to allow even more throughput. Of course, actual customer performance will depend on the individual system configuration.

NEXT STEPS WITH PCI-X

Compaq led the PCI-X workgroup to develop the PCI-X Specification, which was approved by the PCI SIG in September 1999. The PCI-X Specification is available as an addendum to the PCI Local Bus 2.2 Specification[5]. Now that the specification is complete, Compaq is developing solutions at the adapter and system level to help ensure that PCI-X moves quickly and easily from a specification to shipping product.

Adapter Implementation

Compaq is working with independent hardware vendors to enable PCI-X technology in adapters and chipsets. The Compaq enablement program[6] includes development tools, information on test and measurement equipment for adapter designers, and tools for synthesis.

In addition, Compaq is working with a leading supplier of electronic design automation solutions to provide the industry with a streamlined PCI-X test environment. Because of these development efforts by Compaq, companies will be able to bring products to market quickly and effectively, and customers will have the greatest flexibility when they upgrade their PCI adapters.

System Implementation

Customer needs will set the pace of the transition to PCI-X. The features and requirements of PCI-X were specifically chosen to make this transition easy. The interoperability requirements between PCI-X and conventional PCI enable new designs to transition immediately to PCI-X and still remain compatible with the huge installed base of systems and adapters.

The PCI-X specification is designed to meet the crucial I/O demands for high-end, enterprise servers today. Compaq plans to implement the PCI-X bus in industry-standard X86 servers in mid-2000. The PCI-X bus will quickly move to other systems, including high-performance *Compaq AlphaServer* and *Compaq NonStop Himalaya* systems. As the market requires it, PCI-X will be able to provide the same high I/O bandwidth and ease of design for small-to-medium business servers and desktop units.

PCI-X in the Future

The PCI-X protocol is designed to integrate smoothly with other interface technologies such as switched-fabric I/O architectures. Through enhancements such as split transactions and the transaction byte count, PCI-X improves the I/O efficiency so that interfacing with switched fabric and other architecture types is more feasible. PCI-X systems should easily accommodate such interfaces as their markets develop in the future.

While the PCI-X bus is more than adequate for today's demanding I/O devices, Compaq and its industry partners have designed the PCI-X protocol to be easily extended for the future. Possible protocol extensions have the potential to improve bus performance well beyond 1066 MB/s.

[5] To order the PCI-X specification, see the PCI SIG website at http://www.pcisig.com/

[6] See the Compaq website at http://www.compaq.com/pci-x

SUMMARY

PCI-X updates the PCI protocol to include today's best-of-class technologies so that bandwidths of 1066 MB/s are possible. The PCI-X register-to-register design reduces timing constraints and gives the designers the ability to increase the bus frequency up to 133 MHz. Compaq analysis shows that the increased protocol efficiency of 66-MHz PCI-X improves I/O performance by up to 34 percent when compared to conventional 66-MHz PCI. PCI-X technology will also integrate well with new switched-fabric I/O architectures for enterprise computing systems.

Compaq and other industry leaders have leveraged their collective technology leadership to deliver a breakthrough solution to the problem of I/O limitations in today's enterprise computing systems. Compaq led the PCI-X workgroup to provide a solution that is easy to implement and is highly compatible with existing architectures. It is expected that this high degree of backward compatibility and the ease of designing PCI-X devices will enable the rapid migration of systems and devices to bandwidths in excess of 1 GB/s.