Non-Transparent Bridging Simplified

Multi-Host System and Intelligent I/O Design with PCI Express

Developers have been using non-transparent bridging with the PCI and PCI-X interconnect technologies for years to design multi-host systems and intelligent I/Os. Now they can use a similar non-transparent bridging approach using the PCI Express serial interconnect technology.

The non-transparent bridging (NTB) function enables isolation of two hosts or memory domains yet allows status and data exchange between the two hosts or sub-systems. PLX Technology has pioneered the introduction of NTB in PCI Express architectures by including this feature in its switch and bridge products. PLX NTB design in PCI Express (PCIe) is along the same lines as previous implementations in PCI and PCI-X. This implementation is open and available to other PCI Express developers.

This application note will describe how PCI transparent and non-transparent bridging works. It will also explain how multiple processor domains can be established and address translation performed to enable PCI Express transactions between the two processor domains. Furthermore, it will discuss how NTB enabled PCIe switches can be used for various applications.

How the Transparent Bridge Works

The transparent bridge provides electrical isolation between PCI busses. The host enumerates the system through discovery of bridges and end devices. For transparent bridges (TB), the Configuration Status Register (CSR) with a "Type 1" header informs the processor to keep enumerating beyond this bridge as additional devices lie downstream (as illustrated in Bridge A, B and C in Figure 1).

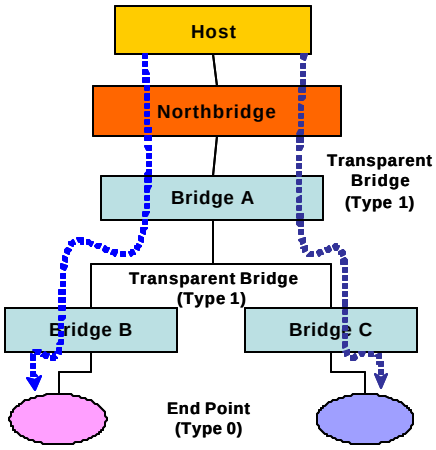

Figure 1. PCI Enumeration

These Bridges with Type 1 headers include CSR registers for primary, secondary and subordinate bus numbers, which, when programmed by the host, define the CSR addresses of all downstream devices.

Endpoint devices have a "Type 0" header in their CSRs to inform the enumerator (BIOS or processor) that no additional devices lie downstream. These CSRs include base address registers (BARs) used to request memory and I/O apertures from the host.

How the NTB Works

In addition to the electrical isolation the NTB adds logical isolation by providing processor domain partitioning and address translation between the memory-mapped spaces of these domains. With the NTB, devices on either side of the bridge are not visible from the other side, but a path is provided for data transfer and status exchange between the processor domains.

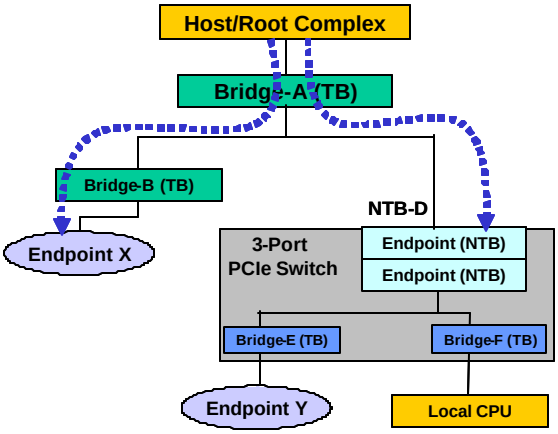

Figure 2. PCI Enumeration with NTB

In this example a system host will enumerate through Bridges A and B (both transparent) on the left branch of the figure 2 until it reaches the endpoint X. On the right side branch, the host will stop enumeration at Bridge D (NTB-D). Similarly, the local CPU will enumerate through Bridge E and F (both virtual bridges within the switch) and discover the endpoint Y, but will not attempt to discover elements beyond Bridge D. This will result in two memory domains.

Address Translation

In the non-transparent bridging environment, PCI Express systems need to translate addresses that cross from one memory space to the other. Each NTB port has two sets of BARs, one for the primary side and other for the secondary side. BARs are used to define address translating windows into the memory space on the other side of the NTB and allow the transactions to be mapped to the local memory or I/Os. Each BAR has a setup register which defines the size and type of the window and an address translation register. While transparent bridges forward all CSRs based on bus numbers, NTB devices only accept CSR transactions addressed to the device itself. Two such translation techniques are direct-address and lookup-table-based.

Direct Address Translation: In direct address translation the addresses of all transactions are translated by adding an offset to the BAR in which the transaction terminates. Base translation registers within the BARs are used to setup these translations. Figure 3 illustrates this shift from the primary side address map to the secondary address map.

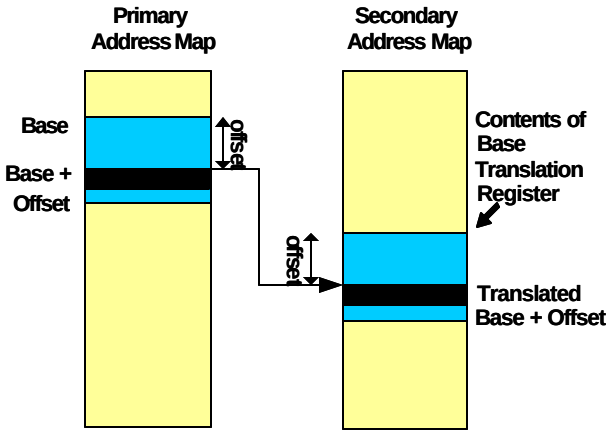

Figure 3. NTB Address Mapping

Lookup Table based Translation: In this scheme, BARs use a special lookup table for address translation of transactions that fall within its window. This approach provides more flexibility in mapping local addresses to host bus addresses as the location of the index field within the address is programmable to adjust window size. The index is used to provide the upper bits for the new memory location.

Inter-processor communication

The non-transparent bridge also allows hosts on each side of the bridge to exchange information about the status through scratchpad registers, doorbell registers, and heartbeat messages.

Scratchpad Registers: These are readable and writeable from both sides of the non-transparent

bridge. The number of scratchpad registers may vary across different implementations. They can pass control and status information between the primary and secondary bus devices, or they can be generic read/write registers.

Doorbell Registers: These registers are used to send interrupts from one side of the non-transparent bridge to the other. These are software controlled interrupt request registers with associated masking registers for each interface on the non-transparent bridge. These registers can be accessed from the primary or the secondary interface of the bridge

Heartbeat Messages: The heartbeat messages are sent from the primary to the secondary host to indicate that it is still alive. The secondary host monitors the state of the primary host and takes appropriate action upon detection of the failure. The doorbell registers, discussed above, may be used for heartbeat messages. Failure of the primary host is declared when the secondary host fails to receive a certain number of the regularly scheduled heartbeat messages.

PLX Technology has pioneered the implementation of the non-transparent bridge concept in PCI Express products to complement the capability of this architecture. The non-transparent bridge provides a powerful feature for designers that want to implement dual host, dual fabric, fail-over and load sharing capability to their systems. The NTB implementation in PCIe by PLX is along the same lines as the NTB use in PCI and PCI-X. PLX has made its implementation open and available to the industry resulting in a broad acceptance of the NTB concept. The ExpressLane switches and bridges from PLX include non-transparent bridging function. The following section illustrates common usage models for NTB.

Intelligent Adapter Card

The ExpressLane PEX 8114 supports **non-transparency** feature. Figure 4 illustrates a host system using an intelligent adapter card.

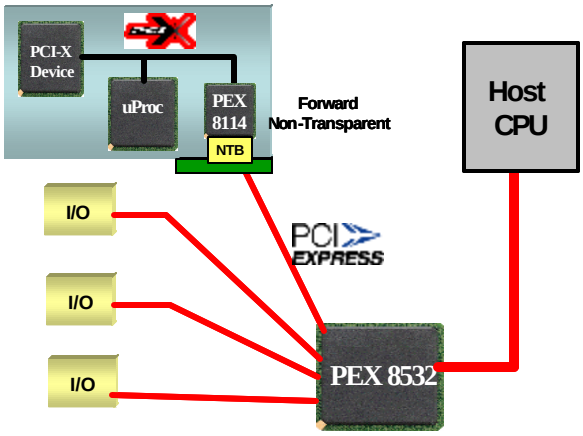

Figure 4. Intelligent Adapter

In this figure, the CPU on the adapter card is isolated from the host CPU. The PEX 8114 bridge non-transparent port allows the two CPUs to be isolated but communicate through the registers discussed earlier.

Dual Host Environment

The NTB function can be used in **dual host**, **host failover**, and **load-sharing** applications. Figure 5 illustrates how two Host CPUs can be isolated using the PEX 8532 NTB feature. Additionally, more PEX8000 switches can be used if multiple CPUs need to be isolated, as shown with the PEX 8516 on the intelligent I/O adapter.

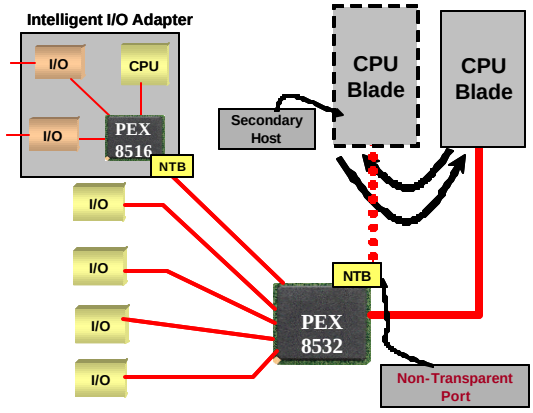

Figure 5. Dual Host Application

In the configuration shown above the secondary CPU monitors the status/heartbeat

of the primary CPU through the NTB registers. In the event the primary CPU fails the secondary CPU will promote itself to primary and isolate the failing CPU.

Multi-Host Systems

In modern storage systems multiple hosts/CPUs are deployed. These CPUs access end-points that may not be in their address domain. The designers are challenged with isolating these CPUs so that they do not interfere with each other and allow them to access all end-points without losing track of their transactions. This concept is also known as Virtual I/O (VIO).

PLX's implementation of the NTB function in PEX 8532 and PEX 8516 enables designers to isolate these CPUs with multiple non-transparent bridge devices while allowing them to communicate to all the end-points. A generic usage model with PEX 8532 and PEX 8516 is illustrated in figure 6. During the enumeration cycle each end-point will have an association with a specific CPU. However, in normal operation, NTB address translation capability will allow all the CPUs to communicate with all end-points.

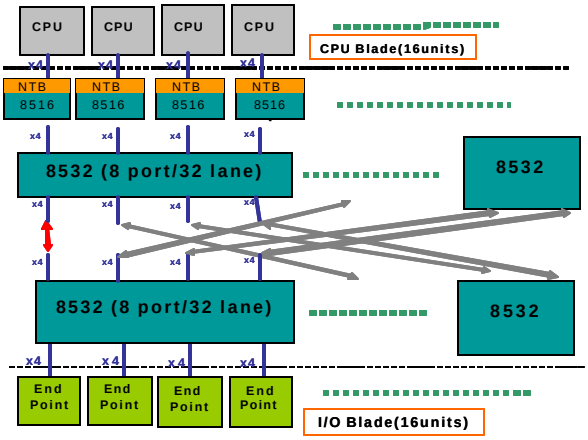

Figure 6. Multi Host System

Summary

The NTB functionality has been used in PCI architectures and it is available for PCI Express based systems. PLX introduced the concept of the NTB in PCI Express and developed products with NTB support. The PCIe implementation of the NTB is similar to what was implemented in PCI. PLX ExpressLane PCI Express switches and bridges, with NTB support, allow a wide range of use from a simple Intelligent Adapter implementation to a complex multi-host system with Virtual I/O capability.

PLX Technology, Inc.
870 Maude Ave.
Sunnyvale, CA 94085 USA
Tel: 1-800-759-3735
Tel: 1-408-774-9060
Fax: 1-408-774-2169
Email: info@plxtech.com
Web Site: www.plxtech.com