PCI Express and Non-Transparent Bridging support High Availability

By Akber Kazmi

In the last ten years, Internet technology has dramatically extended the reach of businesses and consumers around the world. Business dependency on the Internet infrastructure, along with computer systems and their around-the-clock availability, has accordingly increased. Large amounts of data stored on distributed storage devices are being accessed by users spread around the globe, using many different applications while demanding 24/7 service availability.

In the early days of computing, tape backups at the end of the day were sufficient to guarantee data availability. Today's high availability challenges require new approaches. Industry has now adopted solutions that breakdown the data availability task into smaller sub-tasks.

Well defined components

In order to support high availability, all components involved in a particular service or application must be understood. At the highest level are the software elements that deal with the availability of the applications, and the data utilized by these applications.

Another set of elements consists of the hardware that provides the means to store the data, run the applications, and connect them together (such as servers, routers, hosts, switches, and gateways). The hardware that provides this infrastructure may consist of individual modules (for example I/O blade, fabric, and controller) interconnected together to form a system or cluster of systems. Furthermore, the robustness and reliability of each component within a module play a key role in defining the availability of the system.

High Availability and PCI Express

This article will discuss general concepts of High Availability (HA), and then focus on the use of PCI Express technology to support High Availability of the hardware elements used in most common applications.

It will also explore the use of Non-Transparent Bridging (NTB) as one of the key enablers of PCI Express usage in this application.

Furthermore, it will highlight the use of Quality-Of-Service (QOS) features of the PCI Express technology that enables traffic prioritization to guarantee availability of the system resources for high-priority applications and data.

What is High Availability?

A frequent measure and term used for high availability or service availability is the expected amount of time, measured as a percent, that a service or equipment is available to serve the user/application. The 99.999 percent (5-*nines*) availability standard specifies 24/7 service with a maximum of five minutes of downtime in a year. A typical desktop without 5-nines reliability may tolerate nine hours per year of unavailability, but a carrier-class switch or server would require support for 5-nines availability.

Generally, service availability depends heavily on the fault tolerance of the system, including hardware redundancy. The software components of the system use the redundant hardware to enable the service and application availability.

Redundancy

In complex systems such as routers, storage systems, and application servers, multiple intelligent devices are employed to perform various tasks. It is important that these devices interact with each other without causing resource contentions, or access conflicts. The task of connecting these devices becomes much more challenging when they are expected to support redundancy.

As described in the *Providing Open Architecture High Availability Solutions* white paper published by the HA Forum, system redundancy can be divided into the following three categories, or classes:

- Structural redundancy: This involves modifying the information being exchanged or transferred, such as adding Error Checking Codes (ECC) to verify that the information transferred from one device to another was received without loss or errors.
- Temporal redundancy: This involves use of time or system bandwidth to achieve availability, such as using

handshakes or NAK/ACK messages to confirm or acknowledge reception of error-free data.

- Spatial redundancy: This involves availability of more resources than needed in a normal mode of operation, such as hot spares, and spare link/connections to system resources. This may involve 1+1, 1+N, N+M redundancy, and load sharing.

In today's complex systems, support for high availability is as challenging as ever. Economic challenges do not allow for the development of custom solutions. COTS (Commercial Off-The-Shelf) solutions must be used to be cost competitive. Vendors are forming alliances and industry groups to create an ecosystem, where they can rely on solutions that are based on standard implementations with off-the-shelf availability.

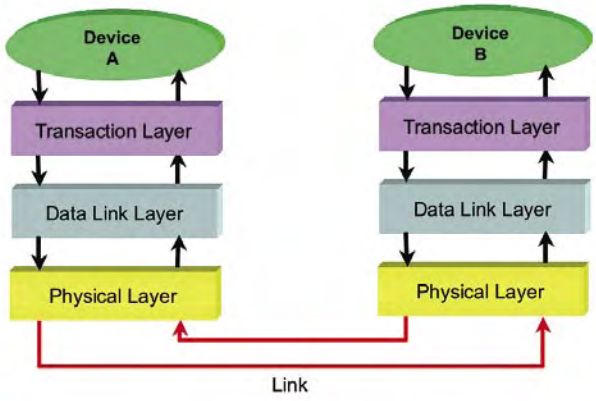

Figure 1

PCI Express technology

PCI Express technology is an emerging interconnect standard. PCI Express is suitable for chip-to-chip, board-to-board, backplane, and box-to-box interconnect for high-performance systems through PCI Express cables. It is an evolution of the PCI standard and is fully backward-compatible with the PCI software structure.

PCI Express is based on a layered architecture, which takes advantage of developments in high-speed serial communication technology. The protocol stack provides three layers:

- Physical layer: Consists of an Low-Voltage Differential Signaling (LVDS) high-speed serial interface specified for 2.5GHz signaling with 8B/10B encoding and AC-coupled differential signaling. A set of LVDS pairs is called a *lane*, and PCI Express allows lane combinations to form bigger and wider ports, such as x1, x2 and so on up to x32. The physical interfaces support hot-plugging.
- Data link layer: Supports packet exchange between neighboring PCI Express entities with data integrity and sequence check, along with packet acknowledgments and flow control.
- Transaction layer: Translates data read/write requests from a host or an end device and optionally provides transaction layer (end-to-end) packet integrity check (CRC-32).

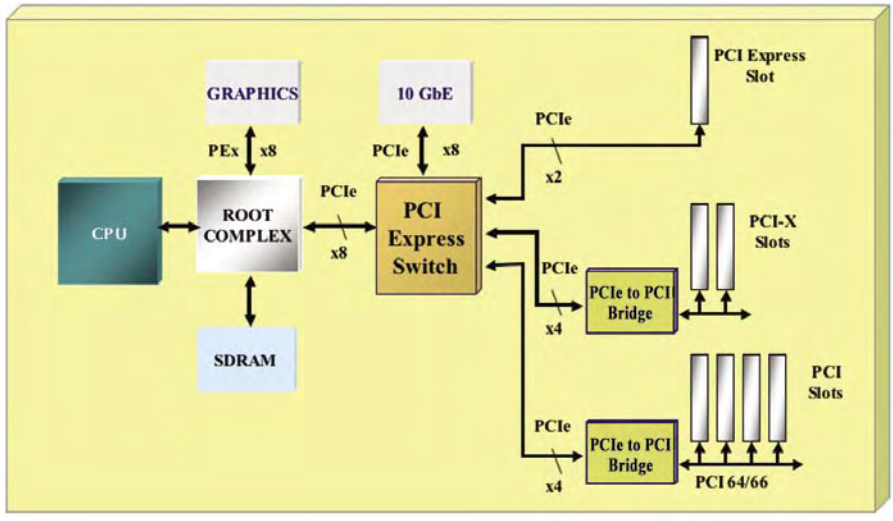

Figure 2

The PCI Express protocol stack is shown in Figure 1. In addition to a well-defined robust protocol stack, PCI Express supports QOS through the use of eight Traffic Classes (TC), eight Virtual Channels (VC), mapping of TCs on VCs, and VC arbitration schemes.

The PCI-SIG, the standards body responsible for the PCI Express specification, has also developed a specification to enable bridging of the PCI/PCI-X bus to the PCI Express serial interface. This would allow many existing applications that use PCI/PCI-X to smoothly migrate to PCI Express. Figure 2 represents a generic use of the PCI Express components such as a root complex, a switch, bridges, and a native PCI Express device.

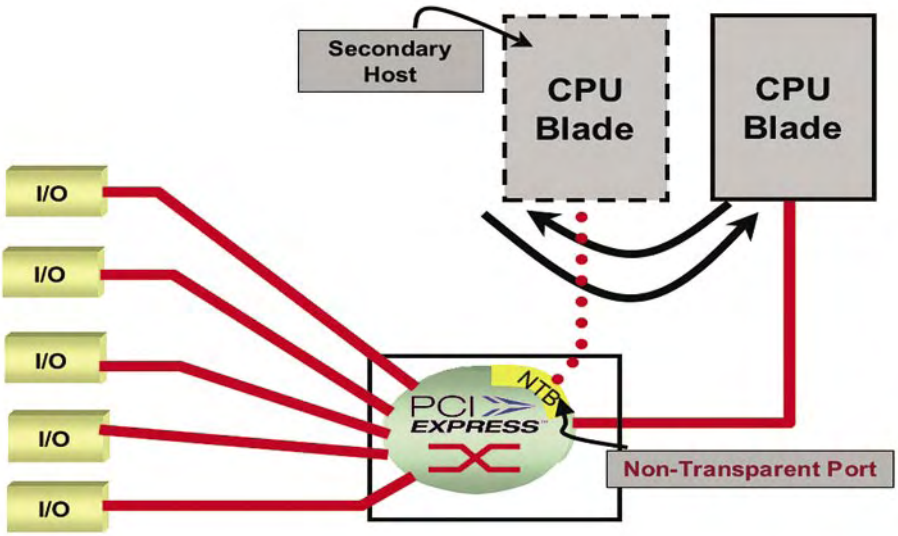

Figure 3

Non-Transparent Bridging

As with PCI and PCI-X, PCI Express was developed with the objective to maintain a host-centric architecture. However, users have developed ways to support PCI Express platforms in a multi-host environment using the non-transparent bridging function, which PCI and PCI-X have used for years.

A non-transparent bridge is functionally similar to a transparent bridge, with the exception that there is an intelligent device or processor on both sides of the bridge, each with its own independent address domain. The host on one side of the bridge will not have the visibility of the complete memory or I/O space on the other side. Each processor considers the other side of the bridge as an endpoint, and maps it into its own memory space as such. The use of NTB in a CPU redundancy (active/standby) application is shown in Figure 3.

In the NTB environment, PCI Express switches translate addresses that cross from one memory space to the other. The NTB also allows hosts on each side of the bridge to exchange information about their status through scratchpad registers, doorbell registers, and heartbeat messages. Scratchpad and doorbell registers are readable from both sides of the bridge, and can be accessed as memory or I/O. Doorbell registers provide a mechanism to generate software controlled interrupts and heartbeat exchanges.

PCI Express in High Availability systems

PCI Express technology complemented by NTB provides valuable features that support high availability applications. PCI Express technology supports the three classes of system redundancy discussed earlier.

Structural redundancy

Structural redundancy involves changing the structure of data moving between entities of a system to support integrity of data exchange. In most communication protocols a checksum, or error-detection code, is appended to a packet or block of data being exchanged. PCI Express provides two levels of structural redundancy as shown in Figure 4:

- Data link layer: 32-bit LCRC (Link Cyclic Redundancy Check) between two neighboring devices.
- Transaction layer: 32-bit ECRC (End-to-end Cyclic Redundancy Check) between two devices separated by one or more entities.

Support of data scrambling and encoding of the embedded clock in the data stream adds another level of structural redundancy for links that use PCI Express as the interconnect technology.

PCI Express allows bandwidth scaling by use of multiple lanes to establish wider data paths. The specification allows wider ports to scale down in order to provide service at a reduced level in case of a failure of one or more lanes. For example, if a port is providing an x4 (2.5Gbps x4) path between two entities and one or more lanes fail, the port would automatically scale down to x2 or x1 to sustain some level of service or availability.

Temporal redundancy

Temporal redundancy involves a handshake mechanism where a receiving entity sends acknowledgement of error-free data reception or error messages if data corruption or loss is detected. The PCI Express specification provides mechanisms for the receiving entity to generate positive/negative acknowledgements to inform the sender about the integrity of the data packets received.

Congestion avoidance is a common mechanism used in communication networks to handle overloading of the system. In addition to providing acknowledgement for packets, PCI Express components provide information to the sender about the availability of buffer space reserved for specific flows (virtual channels). This mechanism, called *credit-based flow control*, can be used to implement efficient use of buffer resources and to avoid the congestion by back-pressuring the source of the data.

Another method of avoiding congestion is to dedicate system resources such as bandwidth for high-priority flows (applications). In a PCI Express system, mapping of traffic classes to virtual channels allows the host to dedicate port bandwidth for high-priority traffic. This enables efficient management of system resources, and implementation of quality of service based on the priority of traffic and associated virtual channel.

Spatial redundancy

Spatial redundancy is one of the most talked about subjects in regards to implementing high availability. In this class, one or more elements of the system are duplicated with one or more backup elements providing a similar function. This allows assigning redundant resources to a particular task, or the creation of multiple paths between two end devices, thus guarding the system against a single point of failure.

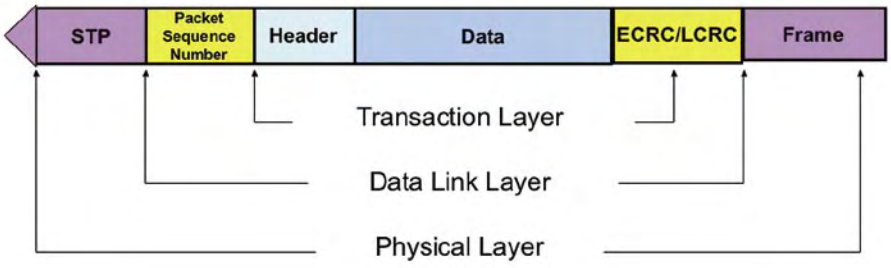

Figure 4

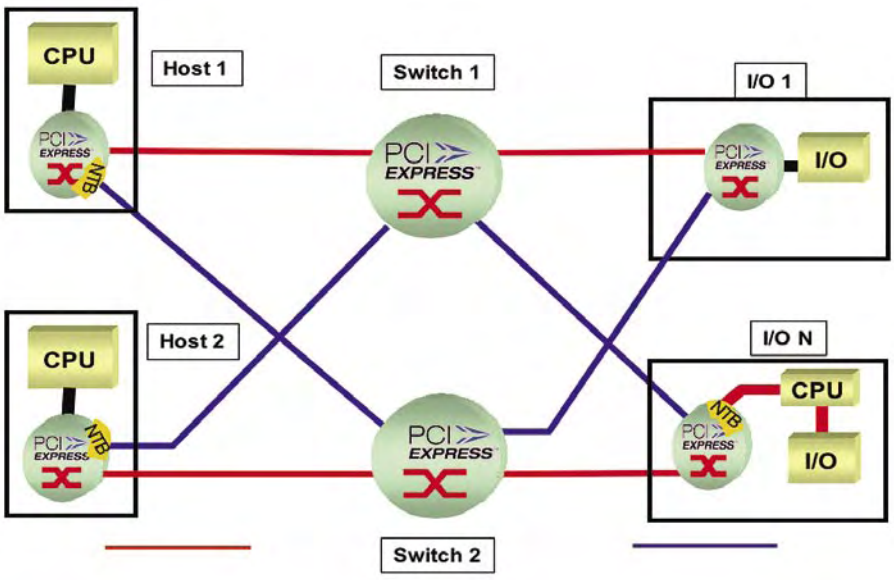

Figure 5

Although PCI Express does not provide a specific solution to address spatial redundancy, innovative developers and users of PCI and PCI Express technology have invented ways to overcome this challenge. Companies such as PLX have created an NTB implementation for the PCI Express applications, which is similar to that of PCI usage models. As discussed earlier, NTB allows two processor or memory domains to be isolated from each other, and yet facilitate limited access for one processor to the other processor's domain (memory & I/Os).

The NTB function in the PCI Express switches would allow the system designers to use multiple CPUs, redundant switch-fabric modules, and intelligent I/O subsystems in a single system. The NTB feature can also be used to create a star, dual-star, or meshed switch fabric with PCI Express switching components. Of course, PCI/PCI-X-to-PCI Express bridges with the NTB function can be used to implement redundancy for the PCI-based modules. An example of NTB function use in a high availability application where a redundant host and switch fabric are required is shown in Figure 5.

This example shows two active CPUs (hosts) communicating with a number of I/O modules through two active switch fabric modules operating in a dual-star configuration. Each CPU module has an active link (shown in red) to one of the switch-fabric modules and a backup (shown in blue) link to the other switch fabric. This configuration provides redundancy for CPU modules, switch fabrics, I/O modules, and the links connecting them. The I/O modules may contain standard I/Os or embedded CPUs isolated with the aid of NTB ports on the switch. The usage model illustrated here can be modified to support 1+N host redundancy or meshed switch fabric.

Summary

High availability is a crucial element of global 24/7 service availability for the Internet, and while there are many elements that make overall service availability possible, hardware redundancy plays a key role. PCI Express technology is the emerging solution for chip-to-chip, board-to-board, backplane, and system interconnect. Complemented by non-transparent bridging, PCI Express offers robust architecture and rich features for hardware redundancy that allow for the development of high availability systems with commercially available PCI Express switch and bridge components. ECD

Akber Kazmi *is Senior Marketing Manager at PLX Technology, and is responsible for the PCI Express switch product line. Kazmi has more than 15 years of marketing experience, with an emphasis on the communications market. Kazmi holds an MSEE from the University of Cincinnati, and an MBA from Golden Gate University.*

For more information, contact Akber at:

PLX Technology

870 Maude Ave.

Sunnyvale, CA 95085

Tel: 408-328-3500

Fax: 408-774-2169

E-mail: akazmi@plxtech.com

Website: www.plxtech.com