Introduction to the PES32NT24G2 PCI Express® Switch Features

Application Note AN-713

Notes

By Kwok Kong

Introduction

The 89HPES32NT24G2 (PES32NT24G2) is a member of the IDT family of PCI Express® switching solutions. The PES32NT24G2 is a 32-lane, 24-port system interconnect switch optimized for PCI Express Gen2 packet switching in high-performance applications supporting multiple simultaneous peer-to-peer traffic flows. Target applications include multi-host or intelligent I/O based systems where inter-domain communication is required, such as servers, storage, communications, and embedded systems.

With Non-Transparent Bridging (NTB) functionality and innovative Switch Partitioning feature, the PES32NT24G2 allows true multi-host or multi-processor communications in a single device. Integrated DMA controllers enable high-performance system design by off-loading data transfer operations across memories from the processors. The PES32NT24G2 supports the PCIe optional features of Access Control Services (ACS) and Alternative Routing ID (ARI).This application note provides a high level introduction to the innovative and unique features of the PES32NT24G2 and provides some application examples.

Port Configurability

The block diagram of PES32NT24G2 is shown in Figure 1. There are four stacks in the device. Stacks 0 and 1 can be configured as a single x8 port, two x4 ports, four x2 ports or any combination. Stacks 2 and 3 can be configured as a single x8 port, two x4 ports, four x2 ports, eight x1 ports or any combination. Each port supports link width negotiation, automatic lane reversal, and crosslink.

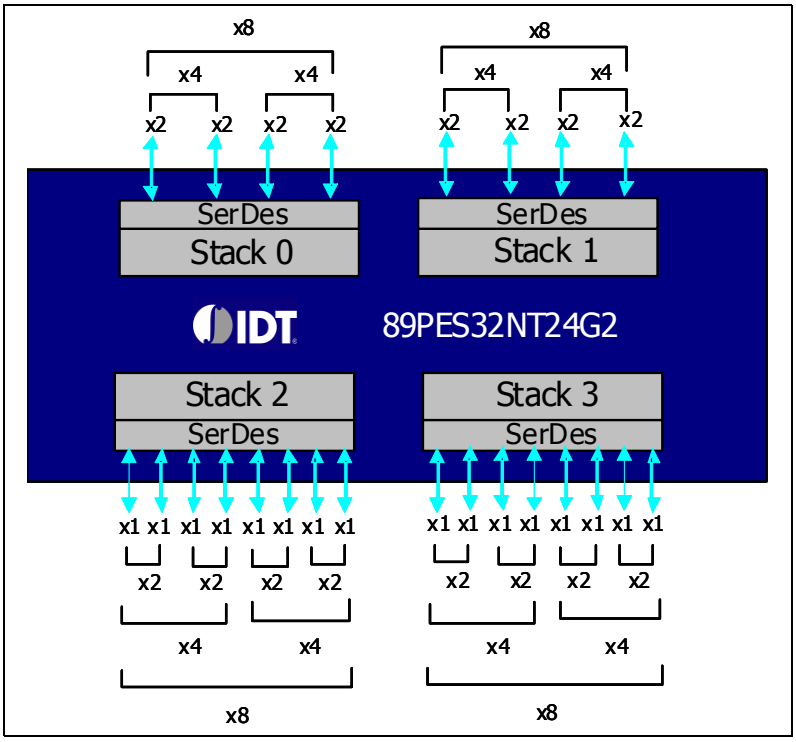

Figure 1 Block Diagram

Notes

Some legacy single partition port configuration examples are shown in Figure 2. In all these configurations, there is only a single upstream port connecting to the root complex.

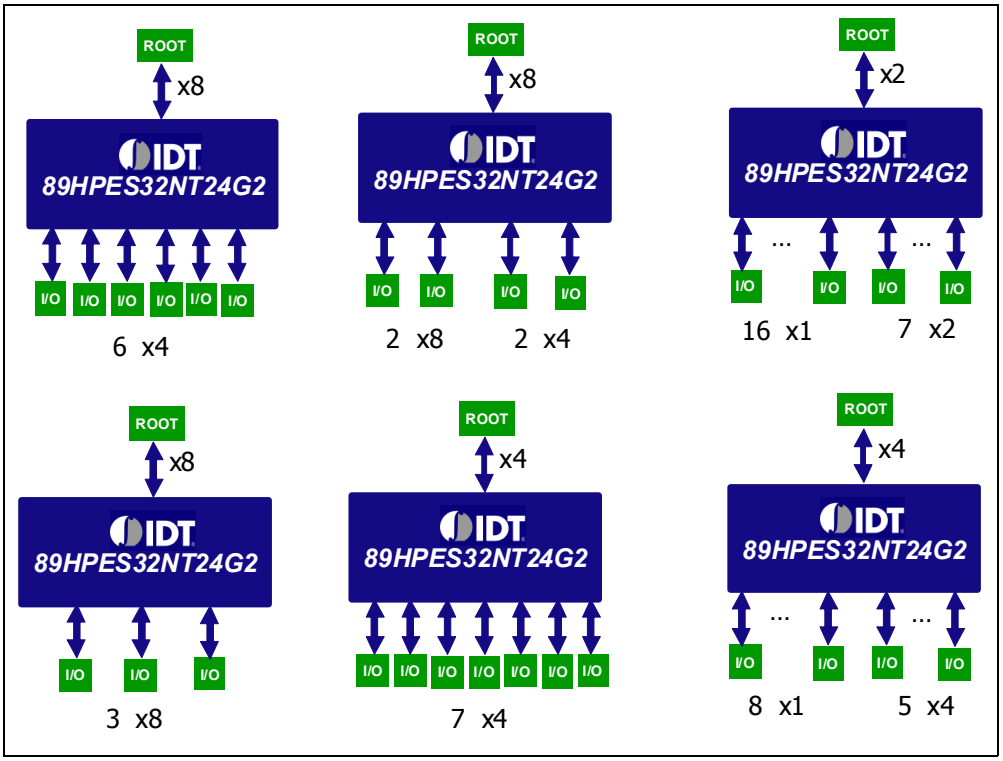

Figure 2 Port Configuration Examples

Switch Partitioning

Switch partitioning is an innovative feature that allows a switch to be statically or dynamically reconfigured into multiple independent logical switches within a single physical device. The PES32NT24G2 supports up to eight switch partitions. Any port can be an upstream or downstream port and any root can have zero, one, or more downstream ports associated with its partition.The partition configuration can be done statically using EEPROM or dynamically by writing into the switch configuration registers.

Switch partitioning enables a number of applications, provides unique benefits and allows product differentiation.

The most direct application of partitioning is to replace multiple discrete physical PCIe switches with a single partitioned PES32NT24G2 switch. Such a replacement shrinks the total cost of ownership by reducing power consumption, decreasing board space, and lowering system interconnect cost. An example of this application is shown in Figure 3. In this example, a single PES32NT24G2 switch is partitioned into four independent logical PCIe switches.

Notes

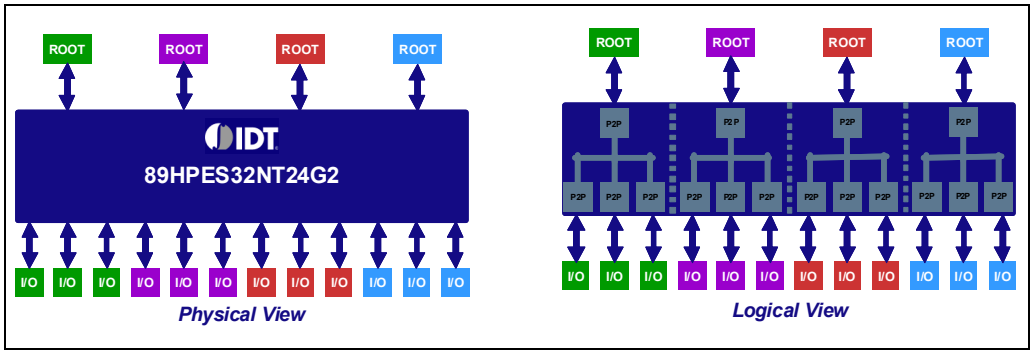

Figure 3 Switch Partitioning

Dynamic switch partitioning can be utilized to perform I/O bandwidth balancing to optimize overall system throughput. A multi-root system, such as in bladed systems, may have unbalanced traffic density across its I/O cards. System bandwidth balancing can be performed by dynamically re-allocating low-traffic or idle I/Os to heavy traffic density partitions from the software application layer. A basic system reconfiguration example is shown in Figure 4. Global I/O resources have been redistributed to offload the leftmost and rightmost roots.

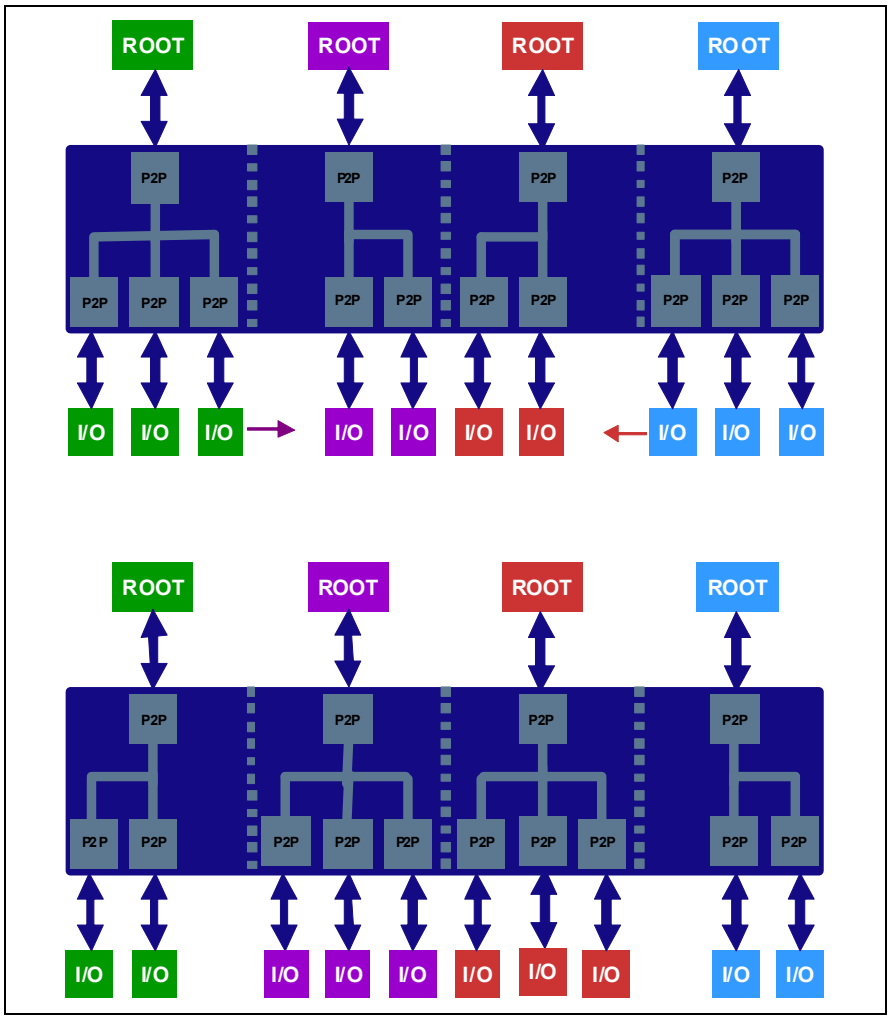

Figure 4 Dynamic Redistribution of I/Os

Notes

Non-Transparent Bridge (NTB)

A PCIe domain consists of a single memory address space, I/O address space, and ID address space. There is only a single root complex in a PCIe domain. There is a single PCIe domain in a switch partition. A NTB port is used to bridge two PCIe domains (or switch partitions), allowing inter-domain communication.

The PES32NT24G2 supports up to 8 NTB functions. Each NTB function may be enabled to co-exist with the upstream PCI-PCI bridge or as an NTB port. There is a virtual NTB interconnect within the switch to allow communication among all the NTB ports. The virtual NTB interconnect is invisible to the PCIe hierarchy. Figure 5 shows an example of a possible NTB configuration which has 5 switch partitions and 4 NTB functions. The switch partition without an NTB port is isolated from the other 4 partitions. The 4 NTB functions are all connected via a virtual NTB interconnect to allow inter-domain communication. Another possible NTB configuration is shown in Figure 6. In the example, there are 8 NTB ports.

Each NTB function within a port appears as a PCIe endpoint. The main function of NTB is to forward PCIe packets across PCIe domains (or switch partitions). The memory address and the device IDs are translated by NTB after a packet is forwarded. PES32NT24G2 supports up to 64 device ID translation using the NTB mapping table which implies that up to 64 PCIe devices in a system may communicate with each other across NTBs. The NTB mapping table is global and shared by all NTB ports. Each NTB port supports six 32-bits or three 64-bits address mapping windows. Direct address and lookup table translation are supported to provide more flexibility in address translation. Each NTB port supports a 32-bit doorbell register that may be used for inter-domain signaling between PCIe domains. Four message registers are supported by each NTB port to allow mailbox message passing between PCIe domains.

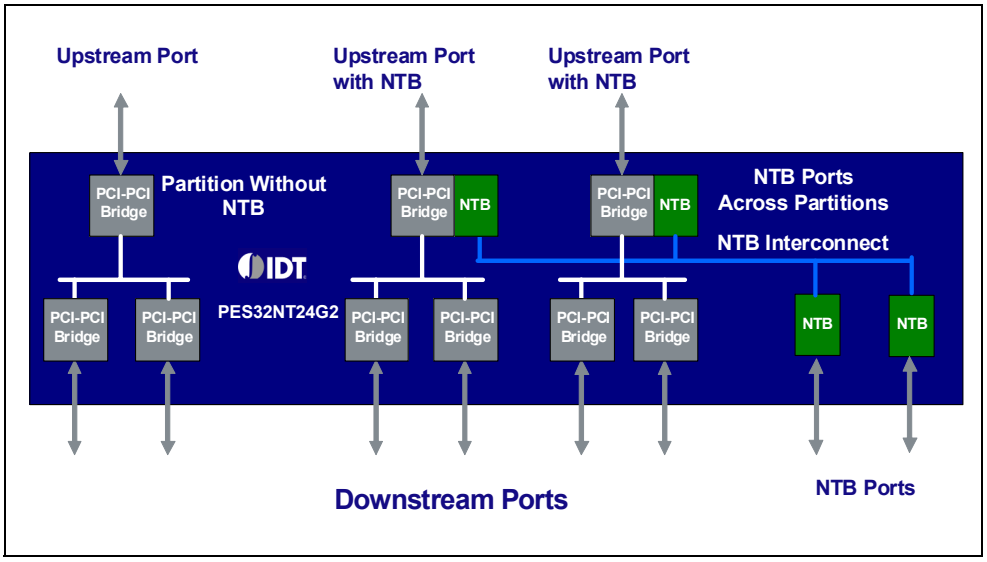

Figure 5 Possible Configuration of NTB ports

Notes

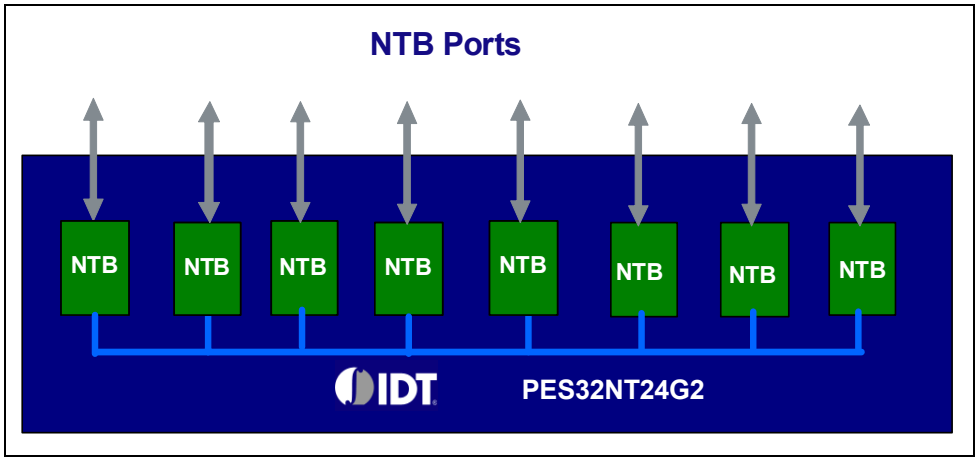

Figure 6 NTB on all Ports

DMA Controller

The DMA engines are used to transfer large amounts of data between any endpoints and the root complex within a PCIe system. It offloads the CPU from moving data around which results in more CPU cycles to perform data processing and manipulation, thereby increasing system performance. The PES32NT24G2 supports two DMA engines, each of which has two channels.

A DMA engine appears as a PCIe endpoint function in a switch partition's upstream port, as shown in Figure 7. Within an upstream port, a DMA function may co-exist with a PCI-PCI bridge function, an NTB function, or both. There can be only one DMA engine within a partition. Therefore, up to two partitions may be configured to support 2 DMA engines.

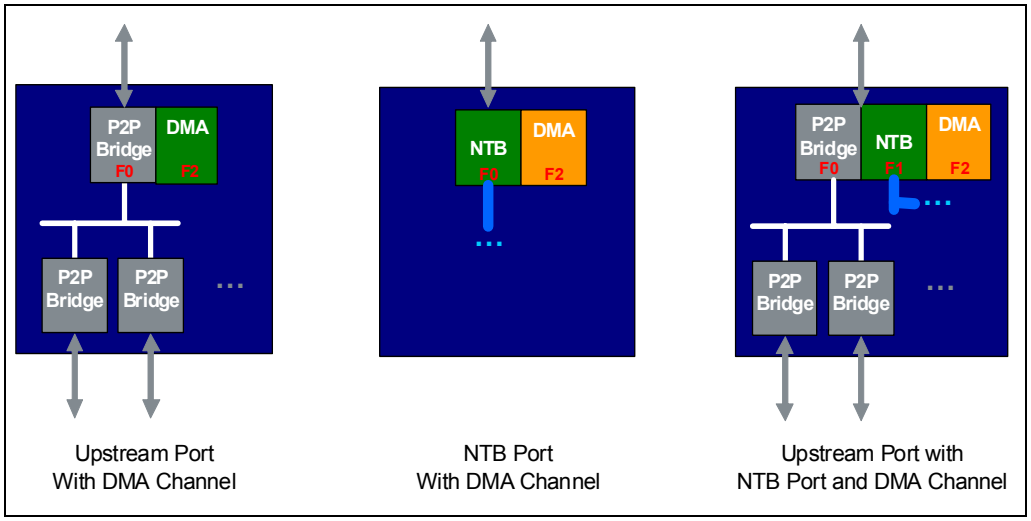

Figure 7 Possible DMA Configurations

A DMA engine operates by reading descriptors from system memory, copying data length bytes of user data from the source address to the destination address as outlined in the descriptor, and writing status information to the descriptor. Complex data movement, such as scatter/gather, may be implemented by linking descriptors together to form a descriptor list. The descriptors may be resident anywhere in the system memory address: above the upstream port, below the downstream port, or in another partition.

Notes

The DMA source and destination address may be located anywhere in the system memory address without any limitations between the two memory regions. If the destination address is a PCIe multicast address, the data is multicasted to multiple destinations according to the multicast rules. The DMA engine can copy data between the following two memory regions:

- from an upstream port to the same upstream port
- from an upstream port to any downstream port
- from a downstream port to the same downstream port
- from a downstream port to a different downstream port
- any port within a partition to any port in a different partition via a NTB port

Examples of the DMA data paths within a partition and across partitions via NTB ports are shown in Figure 8.

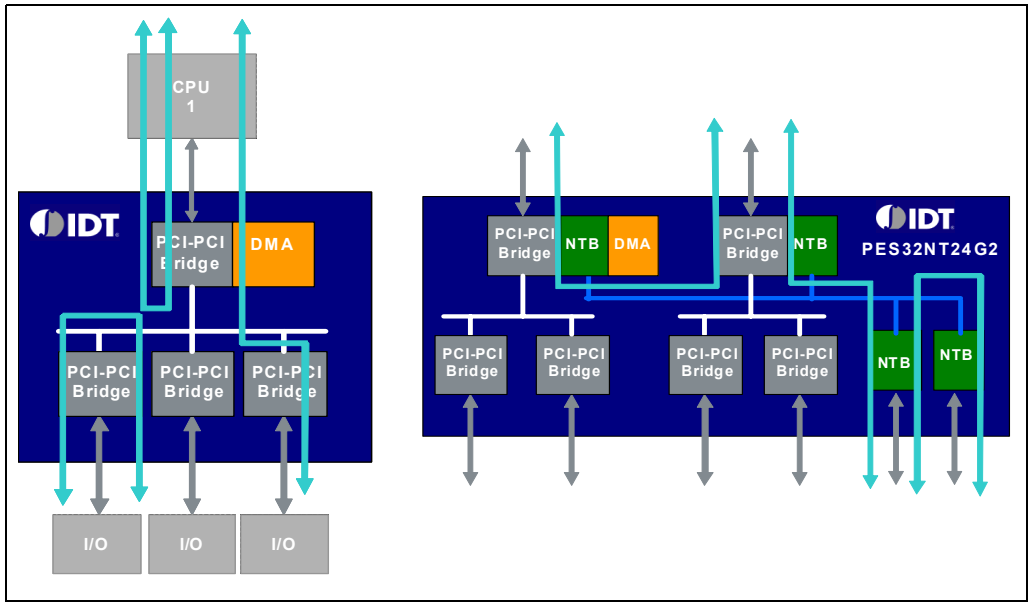

Figure 8 Possible DMA Data Paths

Multicast

Without multicast support in a switch, when a CPU needs to send data (e.g. an updated route table) to multiple I/O destinations, the CPU must send the same data to each of the intended destinations, one at a time. The multicast feature provides the ability to replicate data from a sender to multiple destinations simultaneously, off-loading CPU cycles and creating better link utilization. Adding multicast support to the switch substantially improves overall system performance for those applications that require data replication to multiple destinations, such as storage mirroring, redundant applications, and table updating in communication systems.

The PES32NT24G2 supports the PCI-SIG Multicast ECN. The device supports up to 64 multicast groups and multicast across NTBs. The multicast data path examples are shown in Figure 9. In a single partition transparent mode, the CPU sends a single data packet to the switch, the switch replicates the data packets and sends it to the destination port. In a multi-partition mode, the switch replicates the data packet and sends it to the destination port within the same partition and across partitions via the NTB interconnect.

Notes

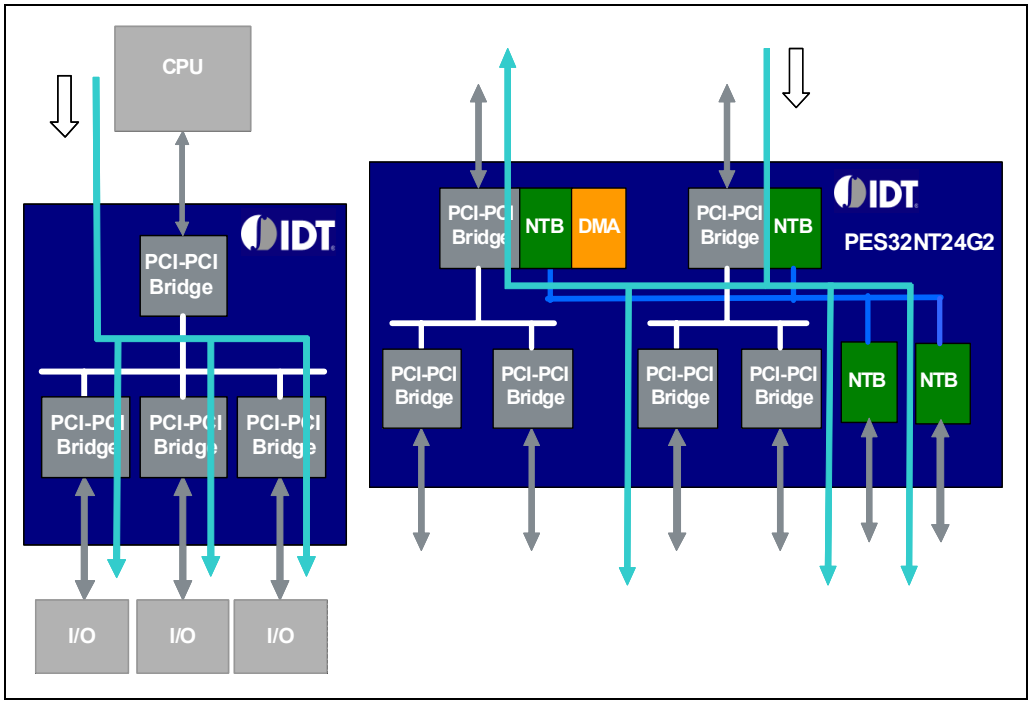

Figure 9 Possible Multicast Data Path

Failover

The PES32NT24G2 supports a flexible failover mechanism. It allows the construction of a dual-host failover system to provide high reliability, availability, and serviceability in case of a CPU failure. Figure 10 shows an example of a dual-host failover system. Under normal operating conditions, the primary CPU is active and serves as the root complex of all the downstream linecards. The secondary CPU is on standby and is ready to take over in case the primary CPU fails or is taken down for routine maintenance.

The primary and the secondary CPUs may use the NTB link between them to exchange management information such as status update and heart beat. When a decision is made to initiate a failover, either CPU can trigger a failover. When failover is completed, the secondary CPU becomes the active root complex of all the downstream linecards

A failover may be initiated by software, by an external signal pin, or by the watchdog timer. A software initiated failover is triggered by writing to a Failover Trigger register. The Failover Trigger register may be written from the primary CPU, secondary CPU, or any of the linecards. A failover may also be triggered by a change in the state of an external signal pin. The watchdog timer initiated failure is triggered upon the expiration of the watchdog timer. The watchdog counter is typically initialized with a certain timeout value by software. The counter is decremented regularly. A failover is triggered when the counter reaches a value of zero. Software that is running in the active CPU updates the watchdog counter with the initial value before it timeouts. If there is a failure in the active CPU, the counter is not updated. When the value of the counter is decremented to 0, the watchdog timer expires and a failover is initiated automatically by the switch.

Notes

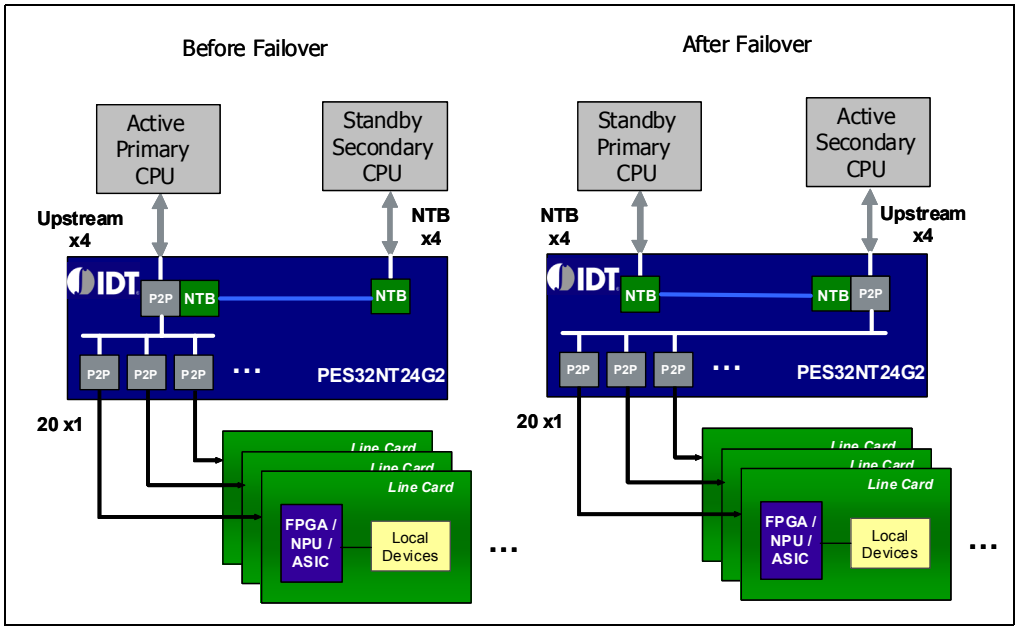

Figure 10 Dual-Host Failover System

Event Signaling

There is a need to signal the occurrence of global switch events in a multi-host system. There is also a need to allow communication channels for a host to notify the occurrence of switch events to other hosts. The PES32NT24G2 supports the monitoring and signalling of multiple events via an interrupt.

The following global events may be monitored by any host in a system:

- Port link up / down event
- Port AER error
- A fundamental or hot reset in a partition
- Failover mode change initiated and completed

When any one of these events occurs, an event is signalled to those hosts that have the event monitor enabled for this particular event. The hosts may start the error recovery procedure upon receiving the event.

There is also a communication channel to allow a host to send events to other hosts which are in different partitions. A general purpose 32-bit register may be used to pass a 32-bit data as part of the event.

Request Metering

Request metering may be used to reduce congestion caused by static rate mismatch in the PES32NT24G2 PCIe switch. A static rate mismatch is a mismatch in the bandwidth capacity of an input and output port. An example of a static mismatch is shown in Figure 11. In this example, there are two endpoints issuing memory read requests to a root. Endpoint A has a x1 link to the switch, while Endpoint B and the root have a x8 link to the switch.

When both Endpoints A and B are making read requests to the root at a high rate, the root returns completion data to the switch at a rate of 8 times the rate of completion data that is forwarded from the switch to Endpoint A. This causes congestion in the switch. The input buffer of the switch's upstream port and the output buffer of the switch's downstream port connected to port A are filled with completion data to Endpoint A, thus blocking completion data from being sent to Endpoint B. The end result is a degradation in overall system performance.

Notes

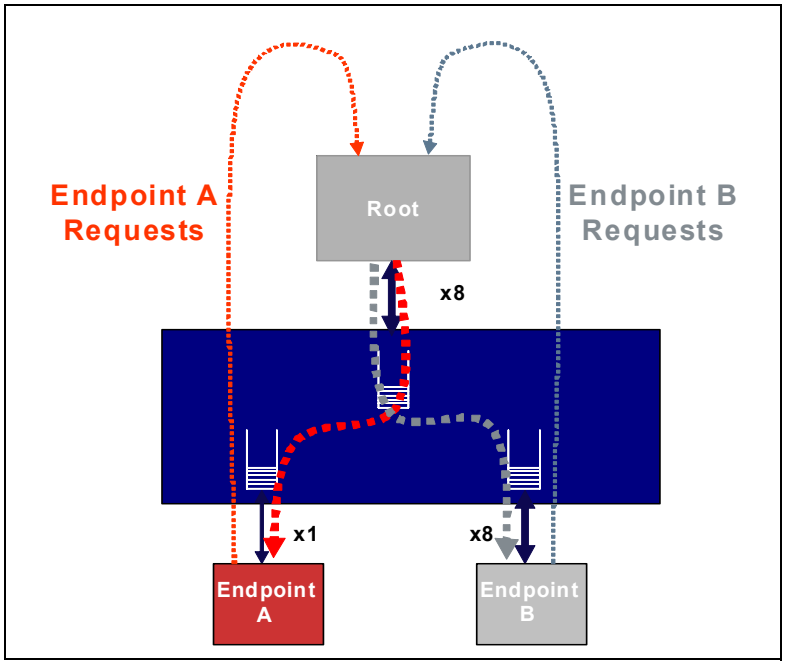

Figure 11 Static Rate Mismatch

Request Metering controls the rate of memory read requests that may be issued by a port to match the completion data rate that is expected to be returned, thereby avoiding or reducing congestion. This results in a much higher system performance. Figure 12 shows the operation of Request Metering. This example shows that an endpoint makes three back-to-back memory requests. Without Request Metering, the switch forwards all the memory requests to the destination port without any delay. The completion data for all three requests will be returned back-to-back, thus likely causing congestion in the switch. With Request Metering, the switch delays forwarding these memory requests to match the expected time it takes to return the completion data to the requestor. Switch congestion is avoided.

The PES32NT24G2 supports Request Metering on all its ports. The Request Metering implementation makes a number of assumptions to estimate the time it takes to return the completion data to the requestor. Those assumptions may not be true for all systems. The switch allows Request Metering parameters to be fine tuned to achieve optimum performance.

Notes

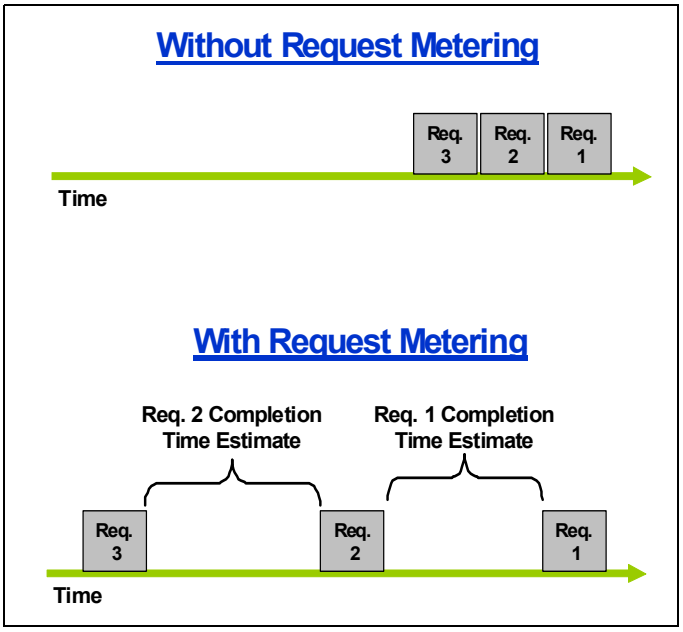

Figure 12 Request Metering Operation

Clocking

The PES32NT24G2 supports very flexible clocking modes. Each port supports two clocking modes: global clocked and local port clocked. In global clocked mode, the port uses the switch's main reference clock. In local port mode, the port uses a dedicated reference clock. Local port clocking allows independent Spread Spectrum Clocking (SSC) on each port. Without this feature, a system with SSC requires that the switch and all devices attached to it operate with a common clock. With local port clocking, a port and the device attached to it may operate with an independent reference clock source, which may have SSC if desired. This provides greater in-board design and system-level configuration across backplanes or cable. This is important in a multi-host system since each host can run its own clock with SSC enabled. The SSC option reduces radiated emissions.

Figure 13 shows an example of supporting two different independent clock domains with SSC enabled using a single PES32NT24G2 switch. There are two partitions and each root complex within its own partition uses its own clock with SSC enabled. The local port clock mode is used on the port which connects to its root complex on the upstream port.

Notes

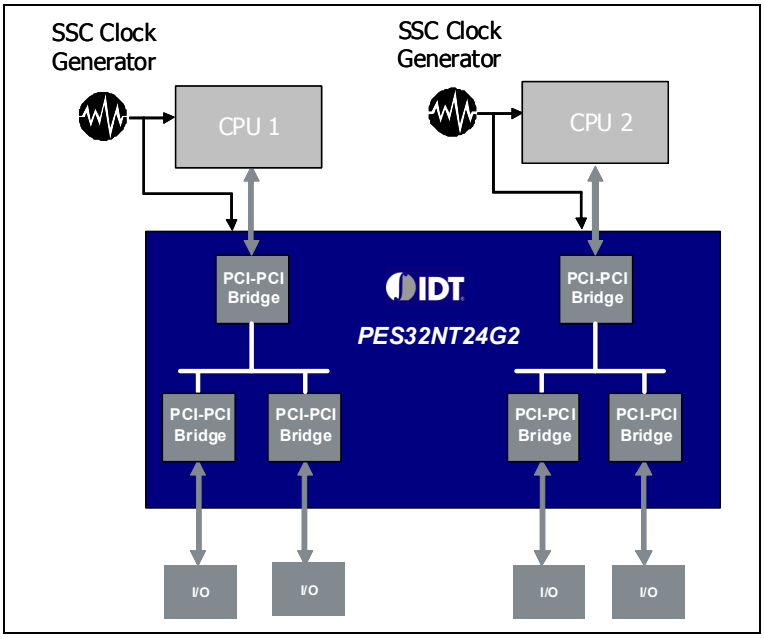

Figure 13 Multiple SSC Clock Domains

Running multiple clock domains with SSC enabled is not limited to a single switch. The local port clock mode also allows multiple independent clock domains across multiple switches. Figure 14 shows a common topology of using NTB to connect two CPUs using two PES32NT24G2 switches. Each CPU runs from its own clock with SCC enabled.

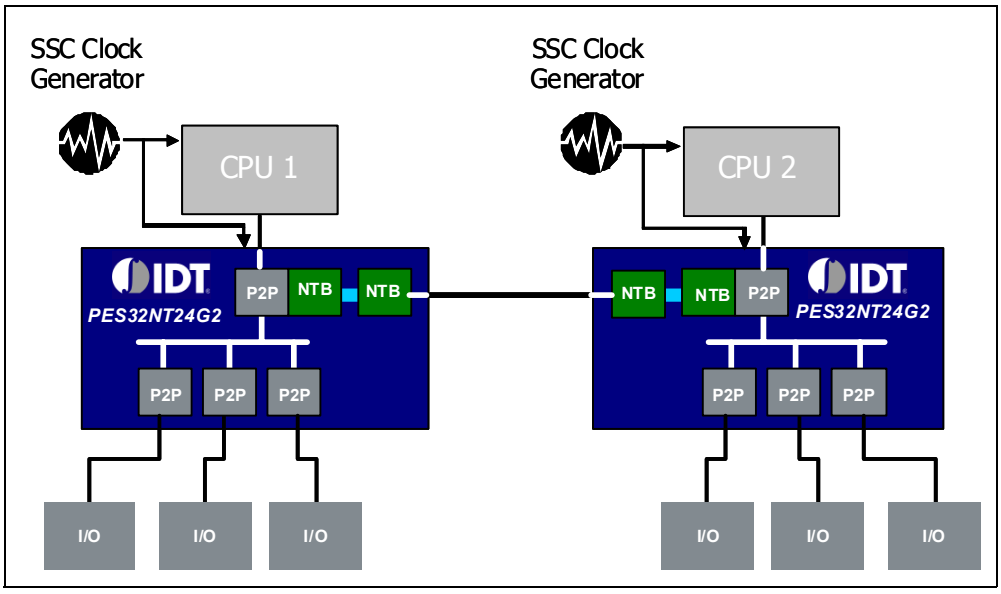

Figure 14 Multiple SSC Clock Domains Across Multiple Switches

Temperature Sensor

The PES32NT24G2 contains an on-chip temperature sensor with three programmable temperature thresholds and a temperature history capability. An alarm is generated when the temperature is above or below one of the three programmable thresholds. The maximum high and minimum low temperature history is automatically recorded in a register for software access. The temperature sensor has a read-out in the range of 0 to 127.5 degree Celsius in 0.5 degree increments.

Notes

The on-chip temperature sensor enables system control and monitoring to enable a number power-saving and system reliability benefits. System software can be used to monitor the on-chip die temperature to turn on/off cooling fans. When temperature falls below the low threshold temperature, cooling fans can be turned off to save power. Cooling fans are turned back on when temperature rises above the mid threshold temperature. Additionally, in the event of a failure within the system, such as a broken fan, the temperature rises to the high threshold temperature which is set to be the upper operating temperature of the device. System software can shutdown the system gracefully in order to avoid permanent damage to the device or system.

Summary

The PES32NT24G2 provides many unique and innovative features to support a wide variety of applications such as servers, storage, communications, and embedded systems. The switch ports are highly configurable, allowing the PES32NT24G2 to be quickly designed into multiple system configurations for different applications, thereby reducing overall development time and cost. The switch partitioning feature allows multiple logical virtual switches to be created. A single PES32NT24G2 switch can replace multiple physical PCIe switches. The total cost of ownership is reduced by lower power consumption, less board space requirements, and lower system interconnect cost.

Up to eight NTB ports are supported to allow multi-host systems to be built. The NTB function on a port can be enabled or disabled dynamically to support a highly redundant system. The integrated DMA engine moves data in the PCIe address space, offloading the CPU from moving data and improving system performance. The multicast feature provides the ability to replicate data from a sender to multiple destinations simultaneously, off-loading CPU cycles and creating better link utilization. The failover feature allows a dual-host active/standby system to be built, ensuring reliability and availability in case of a host failure. In a multi-host system, each host can use event signaling to monitor the status of any port in the switch and initiate error recovery procedures.

Request metering may be used to reduce congestion, thereby improving overall system performance. Each host in a multi-host system may run from its own local clock with SSC enabled. Multiple clock domains support is important in such systems. The PES32NT24G2 supports very flexible clocking modes. Ports may operate with an independent SSC, providing greater in-board design and system-level configuration across backplanes or cable. The on-chip temperature sensor allows the device temperature to be monitored in real time. The system may be shut down gracefully when the temperature exceeds the upper operating temperature of the device to avoid permanent damage to the system or device.

In addition to having the standard PCIe Base Specification 2.1 features, the PES32NT24G2 supports many innovative and unique IDT features to allow high performance, redundant, and lower cost systems to be built. This device is a must-have in every system designer's tool box.

References

Application Note AN-714, DMA in PCIe® Switches, IDT, August 2009.

Application Note AN-715, Clocking Architecture in IDT's PCIe® Gen2 System Interconnect Switch Family, IDT, August 2009.

Application Note AN-716, Building Failover Systems with IDT's PES32NT24G2 PCIe® Switch, IDT, August 2009.