# Using the PES32NT24G2 PCI Express® Switch in Multi-root Compute, Storage, and Communication Systems

*Application Note AN-717*

Notes

By Kwok Kong

## Introduction

The most common usage model for PCI Express (PCIe®) is to connect a single Root Complex to I/O devices and add-in cards in a desktop computer or standalone server. Virtually all the traffic flows from/to the Root Complex to/from the I/O devices. There is very little peer to peer traffic.

However, another model is appearing with increasing frequency: using PCIe as the System Interconnect to take advantage of the high volume and low cost of PCIe technology. 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 an 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.

This paper presents several system architectures of bladed high performance computing platforms including I/O sharing, storage systems, and network switches that use the PES32NT24G2 PCIe switch as the System Interconnect. The traffic pattern for these applications is mainly peer to peer. Redundancy is a must for these systems and a few redundancy topologies are identified.

## Computing

High Performance Computing requires a high-bandwidth and low-latency system interconnect for inter-processor communication. Virtually all the latest x86-based CPUs, PowerPC, and embedded CPUs support PCIe. PCIe offers high bandwidth in the order of a few Gbps to tens of Gbps and low latency in the order of 150 nanoseconds, thereby providing a very cost-effective system interconnect solution for high performance and cost-driven segments of the computing market.

### Multi-Host System

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.

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 1. In this example, a single PES32NT24G2 switch is partitioned into two independent, logical PCIe switches.

## Notes

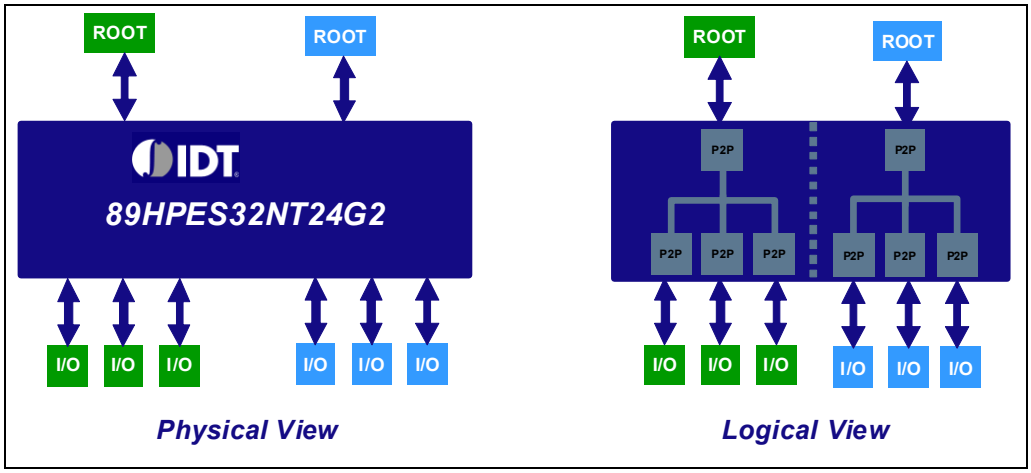


**Figure 1 Multiple Logical Switches**

Dynamic switch partitioning can be utilized to perform I/O bandwidth balancing to optimize overall system throughput. A multi-root system may have unbalanced traffic density across its I/O cards. Using the software application layer, system bandwidth balancing can be performed by dynamically re-allocating heavy traffic I/Os in one partition to another partition with low-traffic I/Os. A basic system reconfiguration example is shown in Figure 2. Global I/O resources have been redistributed to offload the left root.

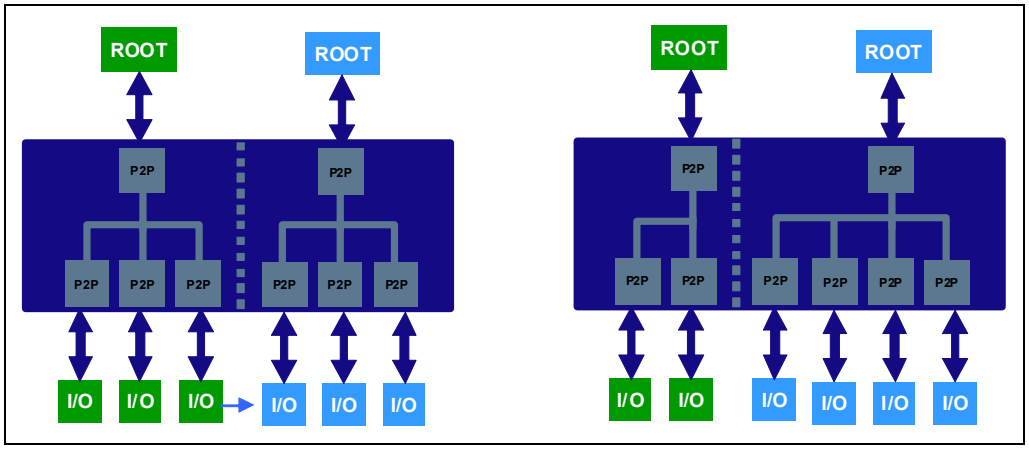


**Figure 2 Dynamic Redistribution of I/Os**

The flexibility of port mapping in switch partitioning allows maximum hardware reuse for multiple variations of a product line configuration to meet the customized needs of end users, saving cost and improving time to market. Figure 3 shows a 2-socket CPU vs. a 4-socket CPU configuration using the same hardware platform with a different switch partitioning setup. When there are 4 CPUs in the system, the PES32NT24G2 is partitioned into two logical switches. The I/O slots in the system are divided evenly and mapped to two independent root complexes. When there are only two CPUs in the system, all the I/O slots are mapped to the single root complex.

## Notes

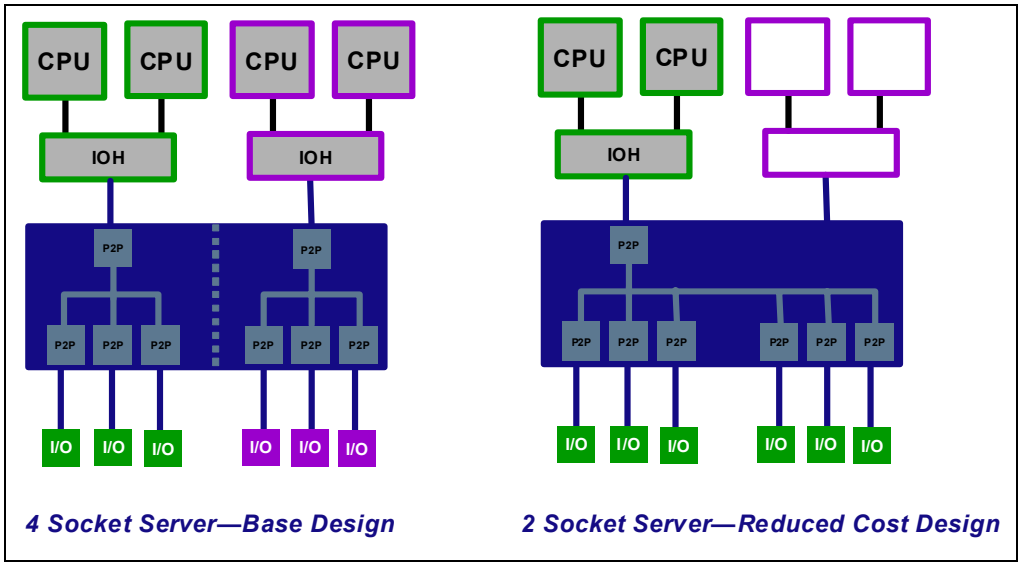


Figure 3 Flexible Slot Mapping

### Bladed Computing Systems

An example of a bladed high-performance computing system is shown in Figure 4. All the compute blades are interconnected using the PES32NT24G2. The compute blade at the top of the switch is the System Root Complex. All other compute blades at the bottom of the switch connect to the NTB ports of the PES32NT24G2. The NTB function isolates the address domains of the local compute blade from the rest of the system. Memory address windows are created to allow each of the compute blades to directly access the memory of all the other compute blades. Large blocks of data can be passed from one compute blade to all other compute blades with high throughput and low latency. In addition to having direct memory access to the other compute blades, each compute blade can communicate with other blades using the NTB communication capability, such as the doorbell, message, and scratchpad registers.

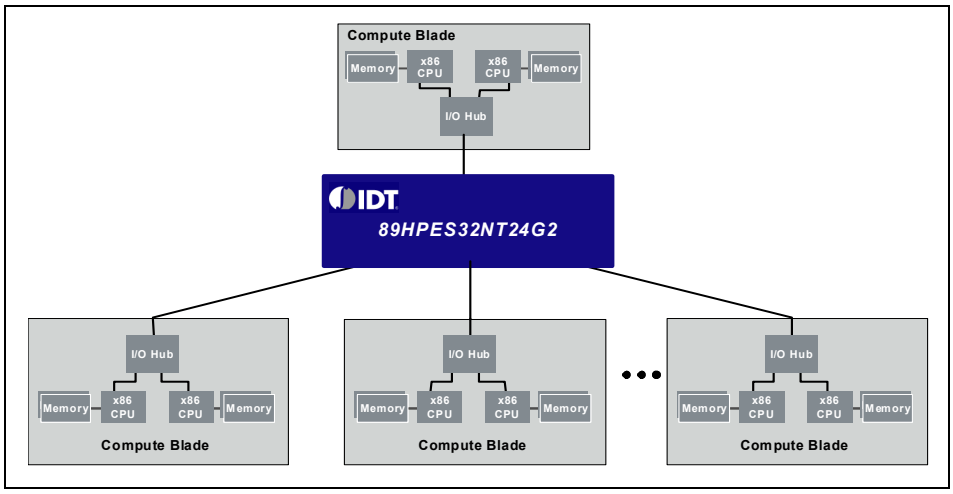


Figure 4 Bladed High Performance Computing System

Bladed compute server system architectures enable equipment providers to produce high-density, modular equipment that can be easily scaled to match user needs. The inherent flexibility and scalability of a bladed architecture supports the expansion or reconfiguration of existing server, storage, and telecommunications infrastructure and supports the rapidly growing and shifting computer services market.

## Notes

Each blade in a bladed system typically has its own dedicated system I/O resources such as the network interface card, Fibre Channel host bus adaptor, and direct attached storage. The hardware resource utilization in a bladed system is usually very low.

In order to improve resource utilization, IO sharing can be used. A bladed server architecture utilizing I/O sharing is shown in Figure 5. The PES32NT24G2 is used to interconnect all the compute blades and I/O blades. The Root Complex is the compute blade that is at the top left corner. All other compute blades connect to the NTB ports of the switch.

The I/O blades are connected to the bottom of the PCIe switch. All the I/O blades are intelligent blades running I/O sharing software and the device driver for the particular devices being supported. The storage blade provides the interface to a Storage Area Network. The Local Storage provides local storage for the compute blades, and the Network I/O blade provides Gigabit Ethernet interfaces.

Virtual device drivers are installed on the compute blades. In this example, a virtual Ethernet device driver, a virtual SATA device driver, and a virtual Fibre Channel device driver are installed in each of the compute blades. When a compute blade needs to send an Ethernet packet to the Ethernet network, the virtual Ethernet device driver forwards the Ethernet packet to the Network I/O blade via the PCIe interface. When the Network I/O blade receives the Ethernet packet, the Ethernet I/O sharing software examines the packet header and forwards the packet to the appropriate external Gigabit Ethernet interface to reach its destination. When a reply is received by the Network I/O blade from its Gigabit Ethernet interface, the I/O sharing software examines the Ethernet packet header and forwards the packet to the destination compute blade. The Gigabit Ethernet interfaces are shared by all the compute blades.

When a compute blade needs to access local storage, the virtual SATA device driver forwards the request to the Local Storage blade. The local storage I/O sharing software sends a request to the local disk on behalf of the compute blade and the result is returned to the requesting compute blade. Local storage is shared by all the compute blades.

The same applies to the storage blade. When a compute blade makes a request to access a remote disk via the Fibre Channel interface, the virtual Fibre Channel device driver forwards the request to the storage blade. The storage I/O sharing software forwards the request to the external Fibre Channel interface. When a response is received on the Fibre Channel interface, that response is forwarded by the storage I/O sharing software to the target compute blade.

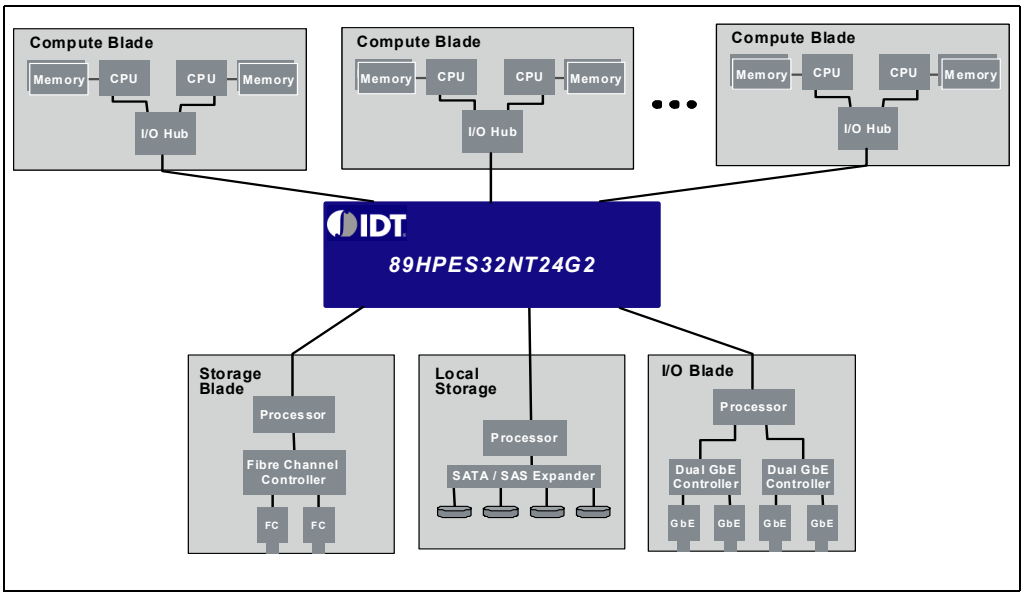


**Figure 5 IO Sharing in Multi-roots Bladed System**

Notes

### ATCA-Based Computing Systems

Advanced Telecom Computing Architecture (ATCA) is a series of industry standard specifications for the next generation of carrier grade communications equipment. PCIe is a standard defined for the backplane high-speed interconnect of the ATCA platform. ATCA provides a standardized platform architecture for carrier-grade telecommunication applications. Embedded high-performance computing platforms have been built using the ATCA platform and using PCIe as the System Interconnect.

An example of an embedded compute blade is shown in Figure 6. Up to 8 Advanced Mezzanine Card (AMC) modules can be put onto an ATCA blade. In this example, 4 AMC compute modules are on a single blade. A PES32NT24G2 is used to interconnect all the AMC modules. The x86 CPU that connects to the top of the switch is the Root Complex. The x86-based AMC modules are connected to the NTB ports of the PCIe switch. The NTB ports isolate the local CPU's address domain from the rest of the system. A system designer can build a complete compute system on a blade with AMC modules.

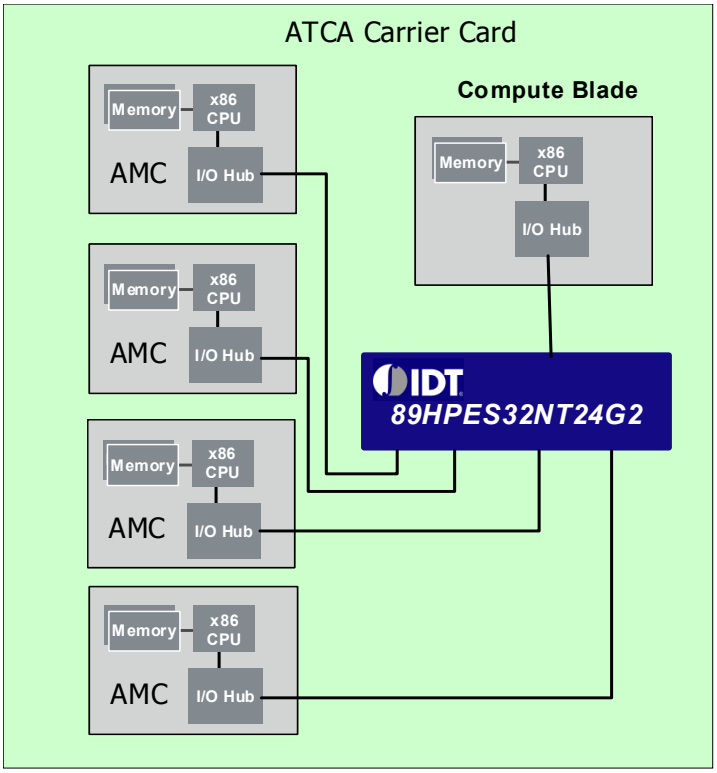


**Figure 6 ATCA Based Embedded Computing**

## Storage System

### Low-End to Midrange Systems

An example of a low to mid-range storage system is shown in Figure 7. This system connects to a Storage Area Network (SAN) via the Host Fibre Channel interface. A 10-Gigabit Ethernet interface may be added to the system to provide Network Attached Storage (NAS) or iSCSI storage. It provides highly reliable and high throughput access to data. Disks may be added to increase the storage capability or bad disks can be swapped out dynamically when the system is up and running to provide a highly available storage system. Two processor controllers in the system provide mutual back-up. Both processor controllers have access to the local disks.

In an active/standby configuration, the active controller has full control over all the local disks while the standby controller has no access at all. If there is a failure in the active controller, the standby controller becomes the active controller and assumes full control over all the local disks.

## Notes

In an active/active configuration, both controllers are active. The disks are divided into two partitions with each controller controlling one partition. Both controllers are actively serving requests from the host. The overall system throughput is much higher compared to the active/standby configuration. If one controller fails, the other controller takes over the rest of the disk partition. The performance of the system degrades gracefully. The active/active configuration is a more common usage model in this application space.

The NTB ports on IDT's PES32NT24G2 are used to connect the two controllers. The controllers can communicate with each other using an NTB communication capability, such as the doorbell, message, and scratchpad registers. Heart beat and checkpoint messages are sent periodically from the active controller to the standby controller. The standby controller monitors the state of the active controller.

The NTB connection is also used to maintain disk cache coherency. When the active controller receives a "disk write" request from a host, the data is stored in the disk cache memory and is written to the disks later to improve system performance. If the controller fails before the data is written from the disk cache memory to the disk, that data is lost. To avoid lost data, the disk cache is mirrored in the other controller. When a controller receives a "disk write" request from a host, the data is stored in the local disk cache memory and a copy of the data is also sent to the other controller's disk cache memory as a backup. The DMA engine in the PES32NT24G2 is used to transfer the cache data across the NTB link, offloading the CPU cycles to focus on serving requests, thereby increasing system performance.

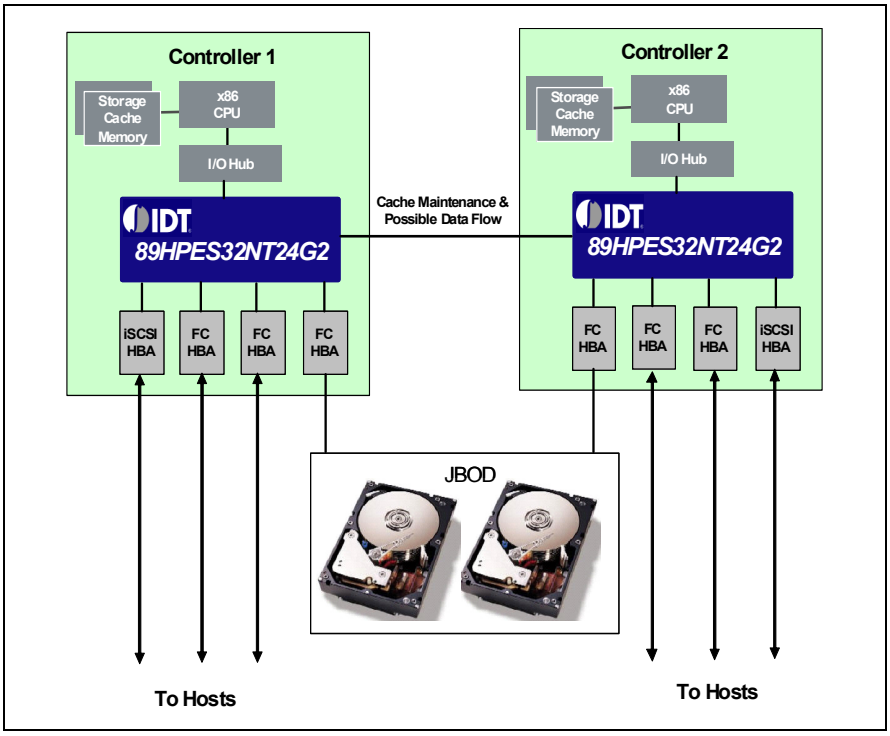


**Figure 7 Low-end to Mid-range Storage System**

An example of a mid-range to high-end storage system is shown in Figure 8. There are multiple storage controllers in such a system. All controllers are active and, at the same time, backing up other controllers. For example, controllers 1 and 2 are working as an active/active pair, controllers 3 and 4 are working as an active/active pair, etc. There are multiple Fibre Channel and iSCSI devices on the storage controller to provide the interfaces to the hosts. A Fibre Channel controller is used to interface to the local disk array via an internal fibre channel switch. There is only one disk array shown in this example, but more Fibre Channel controllers may be added to interface to multiple local disk arrays to scale up the storage capability. The number of storage controllers can be increased or decreased to match the requirements of the transaction load.

Notes

A PES32NT24G2 is used on the storage controller to expand the number of PCIe ports. All the controllers are interconnected using an IDT PES64H16G2 System Interconnect PCIe switch. NTB is enabled on the port which connects a controller to the downstream port of the PES64H16G2 switch. The NTB port isolates the address domains of the storage controllers when the storage processors are x86-based CPUs. The controllers can communicate with each other using the NTB communication capability, such as doorbell, message, and scratchpad registers. The DMA engine in the PES32NT24G2 is used to transfer the cache data among the controllers across the NTB link offloading the CPU cycles.

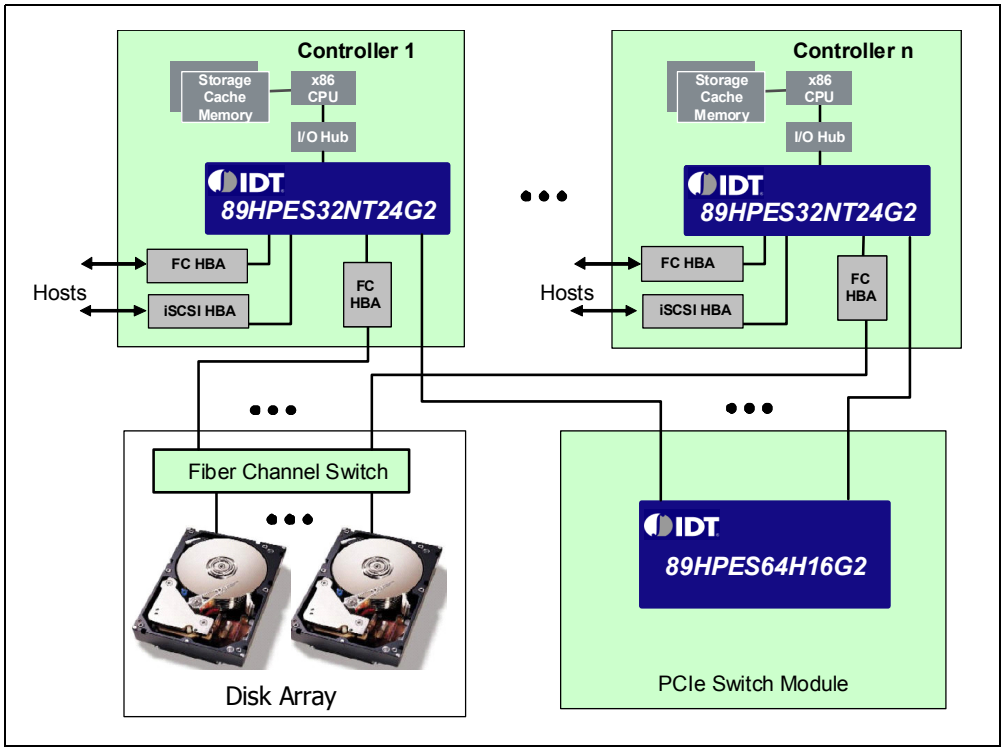


**Figure 8 Mid-range to High-end Storage System**

# Communication System

## Control Plane Application

The PES32NT24G2 may be used in network routers to connect CPU and network adapter cards (or line cards) for control plane traffic. The CPU handles routing protocols to manage the routing table, and runs management software to control the operation of the router. A router using a PES32NT24G2 is shown in Figure 9. The PES32NT24G2 provides a high port count (up to 24 ports) to connect many line cards in a system. The PCIe interconnect is for control plane traffic. There is a separate high bandwidth interconnect to carry network data plane traffic. The data plane in a large communication system typically is proprietary and not shown in Figure 9.

Redundancy is a must in a large communication system. The PES32NT24G2 supports the failover feature to allow an active/standby dual-host redundant system to ensure reliability and availability of the system in case of CPU failure. During normal operation, the primary CPU is the active management CPU. It manages the line cards using the PCIe connections. The secondary CPU is in standby. The primary and secondary CPUs are connected using the internal NTB interconnect. State and checkpoint information can be exchanged between the primary and secondary CPUs via the NTB interconnect. If the primary CPU fails or has to be taken down for routine maintenance, a failover can be initiated by either the primary or secondary CPU, such that the secondary CPU takes over as the active CPU to maintain continuity of operation.

## Notes

The management CPU maintains the routing table for the system and updates the local routing tables in each of the line cards whenever there is a change in the routing table. The CPU sends the same updated table to each line card, one at a time. The PES32NT24G2 supports a multicast feature to replicate data from a sender to multiple destination simultaneously. The CPU can set up the multicast feature in the PES32NT24G2 to multicast the updated route table to all the line cards. The CPU only needs to send a single copy of the updated route table to update all the line cards, offloading CPU cycles and creating better link utilization.

The management CPU periodically gathers network statistics from each line card. The CPU has to perform many "PCIe Memory Read" requests to read the statistics from each of the line cards. Memory Read requests to the PCIe address space are slow and consume many CPU cycles. The PES32NT24G2 supports two DMA engines. The management CPU can set up the DMA engine to "read" the statistics from each line card and "write" the statistics to the local CPU memory. The DMA may also be used for the transfer of the checkpoint information between the active/standby roots. It is much more efficient for the CPU to read from its local memory than from line cards.

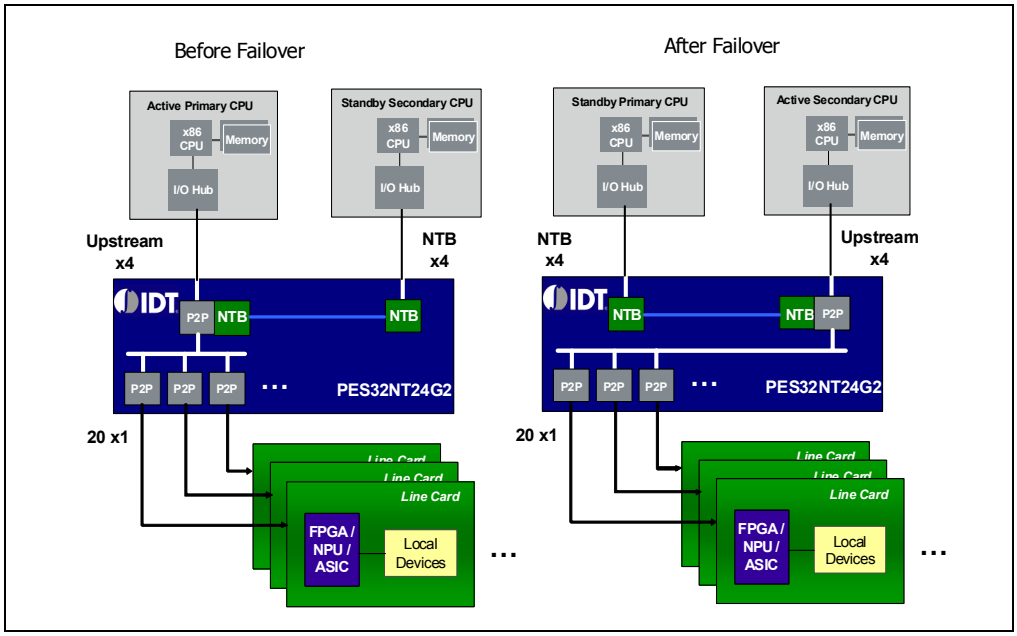


**Figure 9 Redundant Control Plane in Networking System**

### Control and Data Planes Application

A small to mid-range network router can be built using the PES32NT24G2. An example of a network router is shown in Figure 10. In this architecture, the Root Complex is the Route Processor Engine. It handles the control plane traffic such as the routing protocols to manage the routing table and runs management software to control the operation of the router. The packet processing and forwarding functions are distributed to the line cards. All line cards are intelligent and have Packet Processing Engines. NTB is enabled on a port of the PES32NT24G2 when a x86 CPU is used as the Packet Processing Engine.

When a packet is received by a line card, the local Packet Processing Engine processes the packet and forwards the data plane packet to the destination line card directly. The Route Processor Engine is not involved in the packet forwarding. All line cards can forward packets to all other line cards concurrently and autonomously. The processing power of the Packet Processing Engine is required only to handle the bandwidth of the line card where the Packet Processing Engine is resident. This allows the system bandwidth to scale up or down depending on the line card interface bandwidth requirement.

## Notes

The diagram only shows a single PCIe interface to carry both the control and data planes traffic. Optionally, users may physically split the control and data planes traffic by using separate paths. Two PCIe switches are required in this model; one to carry the control plane traffic and the other to carry the data plane traffic.

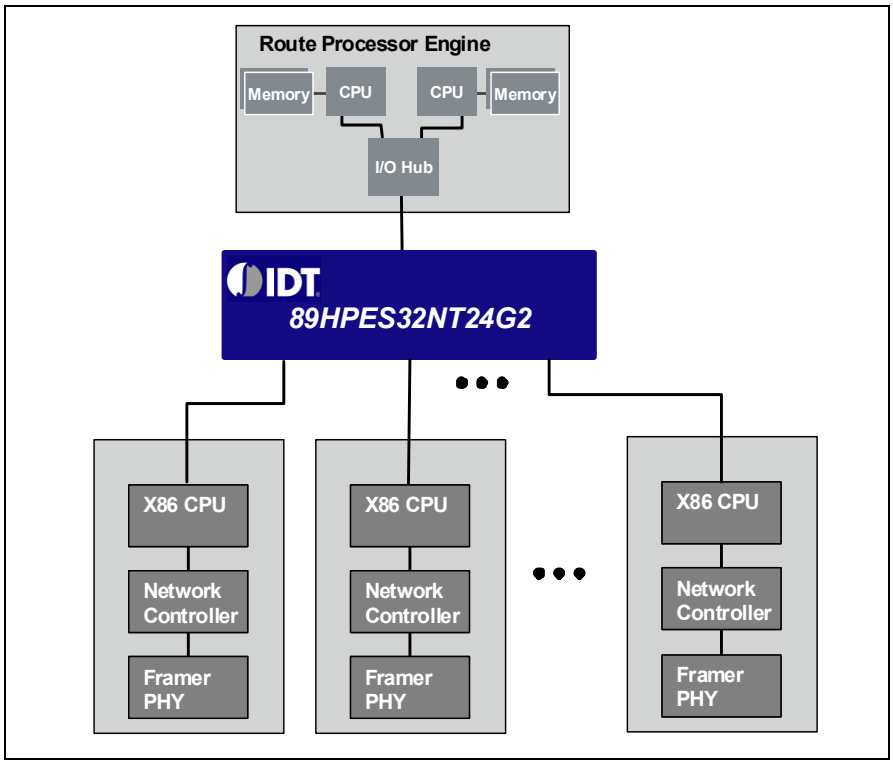


**Figure 10 Mid-range to High-end Network Switch**

# Redundancy

A highly available and reliable application generally requires some level of redundancy. A dual Root Complex topology has both an active and a standby Root Complex. An example of this dual Root Complex topology is shown in Figure 9. The PES32NT24G2 supports the failover feature. The primary Root Complex is the active Root Complex, and the secondary Root Complex is the standby Root Complex. During normal operation, the active Root Complex is in control of the system and is the Root Complex of the system domain. The standby Root Complex connects to the active Root Complex via the NTB interconnect.

The standby Root Complex takes over as the active Root Complex when a managed switchover is requested or the standby Root Complex detects a failure in the active Root Complex. A managed switchover is initiated by the user for scheduled maintenance or software upgrades or in response to some form of demerit checking within the active Root Complex.

The standby Root Complex monitors the state of the active Root Complex through heart beat and checkpoint messages. When the standby Root Complex detects a failure in the active Root Complex, the standby Root Complex configures the PCIe switch to failover. After failover, the standby Root Complex becomes the active Root Complex.

The example shown in Figure 9 describes a software-initiated failover. In addition, the PES32NT24G2 supports the watchdog timer-initiated failover feature. As an alternative method of initiating the failover, when the active Root Complex has not reset the watchdog timer within a certain time period (as configured during initialization), the standby Root Complex then takes over as the active Root Complex. This is expected when a hardware failure or major software failure occurs in the active Root Complex.

## Notes

Figure 11 shows an example of another dual-host system. There are two independent systems connecting to each other via an external NTB port. During normal operation, each CPU manages its own I/O sub-system. CPU 1 exchanges its state information and heart beats with CPU 2 using the NTB port. CPU 2 takes over the CPU 1 system when CPU 1 fails. CPU 2 reconfigures the NTB ports of the PES32NT24G2 dynamically such that its local NTB port becomes a downstream port and the remote NTB port becomes an upstream port with an NTB function. The upstream port that had previously connected to CPU 1 becomes an NTB port. The CPU 1 system becomes part of the PCIe domain of CPU 2.

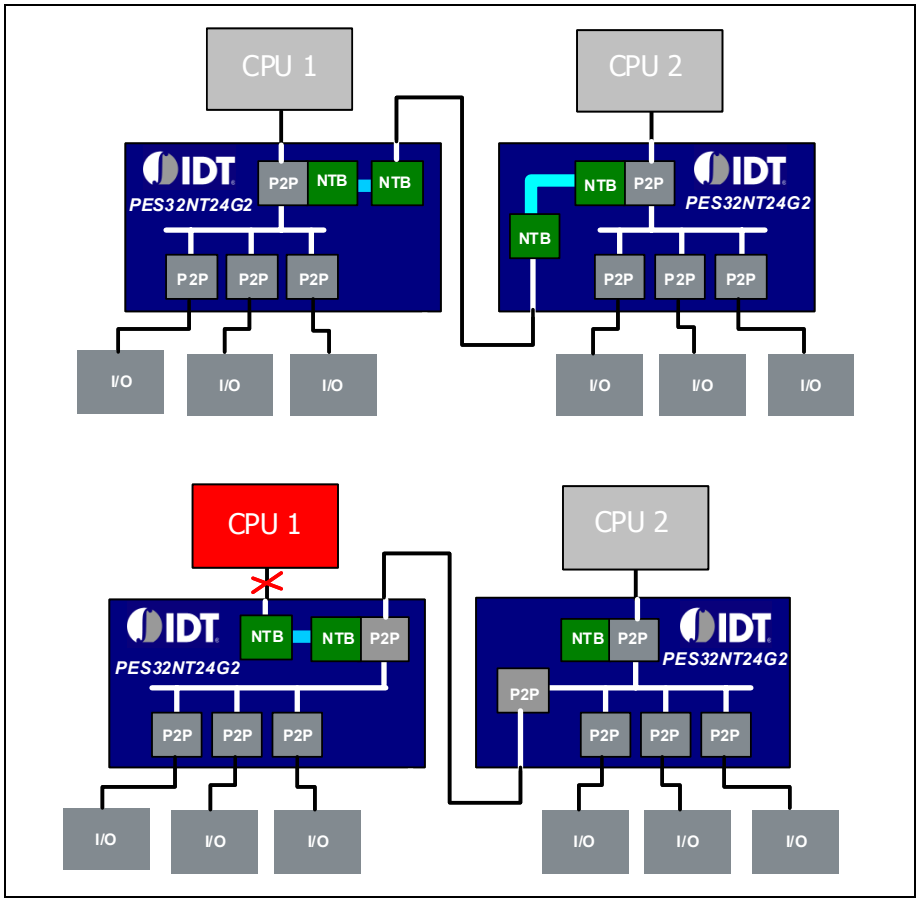


**Figure 11 Dual Root Complex Redundancy**

### Dual-Star Topology

The switch is still a single point of failure in the dual Root Complex topology. For a fully redundant system, a dual-star topology may be deployed as shown in Figure 12. In this topology, a second Root Complex and a System Interconnect PCIe switch are added to provide switch redundancy. The PES32NT24G2 is added to the compute blade to provide the NTB function. Each compute blade connects to two System Interconnect PCIe switches using two NTB ports. Both connections are active in an active/active mode.

Traffic can be sent to either connection. To avoid packet re-ordering problems, a default connection is chosen to send all traffic among the compute blades. When the default connection fails, the other connection is used to send traffic to the other compute blades.

## Notes

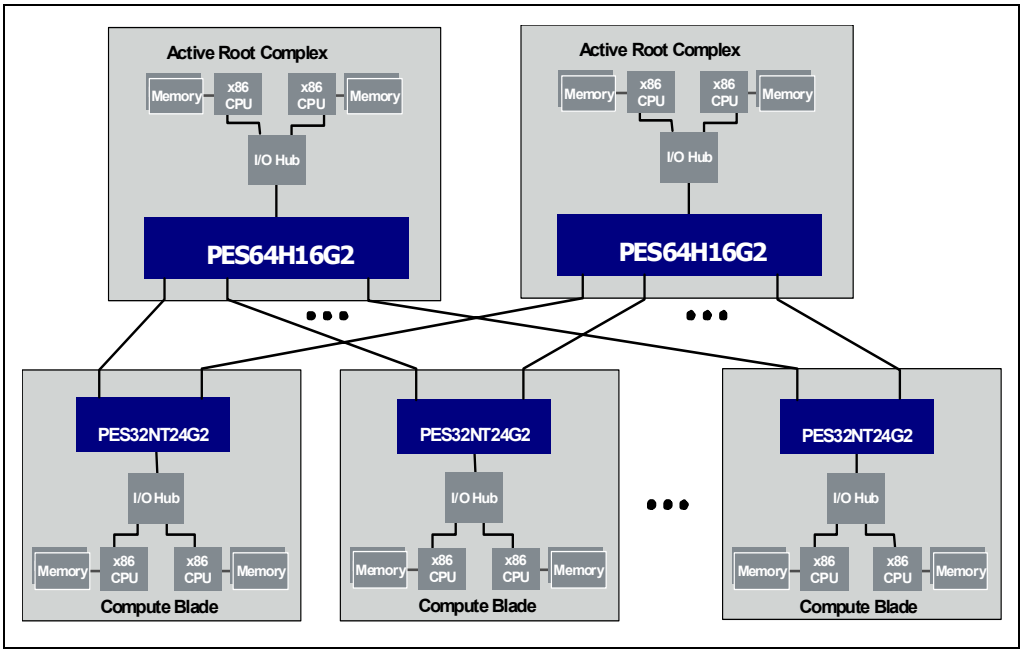


**Figure 12 Dual-Star Redundancy**

The dual-star topology requires two separate switch modules in a system, using up precious slots in a chassis and increasing total system cost. This may be unacceptable in smaller systems. A fully meshed redundant system, as shown in Figure 13, is an alternative to the dual-star topology. The compute blade has direct connection to all the other blades in the system, eliminating the requirement for separate switch modules. NTB is enabled on all the ports in the PES32NT24G2.

A compute blade sends traffic directly to the destination. When a compute blade fails or is removed from the system, there is no need for failover. When the active compute blade detects the failure of another compute blade, the active compute blade removes the failed compute blade from its topology and stops sending traffic to it.

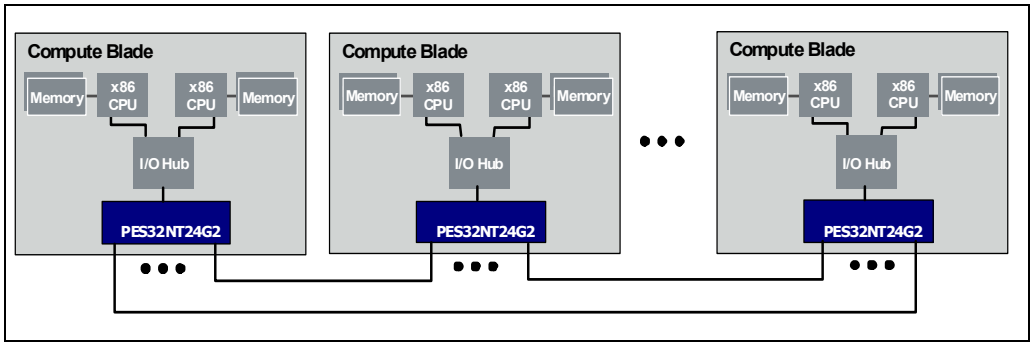


**Figure 13 Fully Meshed Redundancy**

Notes

## Conclusion

PCIe is commonly used to connect a single Root Complex to I/O devices in a desktop computer and a standalone compute server. System vendors are beginning to deploy PCIe as the System Interconnect in multi-host systems and embedded applications to take advantage of the high-volume, low-cost PCIe technology. Several systems that use the PES32NT24G2 PCIe switch have been presented and shown to be viable, cost-effective solutions for the compute, storage, and communication segments.

I/O sharing in a bladed multi-host system can be implemented with a software solution using PCIe switches and I/O devices that are available today. Storage systems and network switches can also make use of the PES32NT24G2 as the System Interconnect. PCIe forms the basis for a viable System Interconnect solution for many non-traditional PCIe usage models. Only a few applications of the PES32NT24G2 have been identified in this paper, but PCIe is gaining momentum in other embedded applications.

Redundancy is important in bladed multi-host, storage, and communication systems. A dual Root Complex topology, a Dual-star topology, and a fully meshed topology provide different levels of redundancy.

The PES32NT24G2 solution delivers the performance, optimal resource utilization, scalability, RAS, and security features that are essential for the successful design and deployment of systems that have been described in this paper. The solution presented in this Application Note was designed with the needs of leading bladed multi-host, storage, and communication system manufacturers in mind, since it is architected to scale with their performance and capability needs well into the future.

## References

PCI Express Base specification Revision 1.1

PCI Express Base specification Revision 2.0

IDT 89HPES32NT24G2 PCI Express® Switch User Manual

IDT 89HPES64H16G2 PCI Express® Switch User Manual

Application Note AN-713, Introduction to the PES32NT24G2 PCI Express® Switch Features, IDT, August 2009.

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.