16-Lane 16-Port PCIe® Gen2 System Interconnect Switch with Non-Transparent Bridging

89HPES16NT16G2
Product Brief

Device Overview

The 89HPES16NT16G2 is a member of the IDT family of PCI Express® switching solutions. The PES16NT16G2 is a 16-lane, 16-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 functionality and innovative Switch Partitioning feature, the PES16NT16G2 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. Each lane is capable of 5 GT/s link speed in both directions and is fully compliant with PCI Express Base Specification 2.1.

Features

- **High Performance Non-Blocking Switch Architecture**
  - 16-lane, 16-port PCIe switch with flexible port configuration
  - Integrated SerDes supports 5.0 GT/s Gen2 and 2.5 GT/s Gen1 operation
  - Delivers up to 16 GBps (128 Gbps) of switching capacity
  - Supports 128 Bytes to 2 KB maximum payload size
  - Low latency cut-through architecture
  - Supports one virtual channel and eight traffic classes
- **Port Configurability**
  - Three stacks
    - One x8 port configurable as:
      - One x8 port
      - Two x4 ports
      - Four x2 ports
      - Eight x1 ports
      - Several combinations of the above lane widths
    - One x4 port configurable as:
      - One x4 port
      - Two x2 ports
      - 4 x1 ports
    - Four x1 ports
  - Automatic per port link width negotiation (x8 → x4 → x2 → x1)
  - Crosslink support
  - Automatic lane reversal
  - Per lane SerDes configuration
    - De-emphasis
    - Receive equalization
    - Drive strength
- **Innovative Switch Partitioning Feature**
  - Supports up to 4 fully independent switch partitions
  - Logically independent switches in the same device
  - Configurable downstream port device numbering
  - Supports dynamic reconfiguration of switch partitions
    - Dynamic port reconfiguration — downstream, upstream, non-transparent bridge
    - Dynamic migration of ports between partitions
    - Movable upstream port within and between switch partitions
- **Non-Transparent Bridging (NTB) Support**
  - Supports up to 4 NT endpoints per switch, each endpoint can communicate with other switch partitions or external PCIe domains or CPUs
  - 6 BARs per NT Endpoint
    - Bar address translation
    - All BARs support 32/64-bit base and limit address translation
    - Two BARs (BAR2 and BAR4) support look-up table based address translation
  - 32 inbound and outbound doorbell registers
  - 4 inbound and outbound message registers
  - Supports up to 64 masters
  - Unlimited number of outstanding transactions
- **Multicast**
  - Compliant with the PCI-SIG multicast
  - Supports 64 multicast groups
  - Supports multicast across non-transparent port
  - Multicast overlay mechanism support
  - ECRC regeneration support
- **Integrated Direct Memory Access (DMA) Controllers**
  - Supports up to 2 DMA upstream ports, each with 2 DMA channels
  - Supports 32-bit and 64-bit memory-to-memory transfers
    - Fly-by translation provides reduced latency and increased performance over buffered approach
    - Supports arbitrary source and destination address alignment
    - Supports intra- as well as inter-partition data transfers using the non-transparent endpoint
  - Supports DMA transfers to multicast groups
  - Linked list descriptor-based operation
  - Flexible addressing modes
    - Linear addressing
    - Constant addressing
- **Quality of Service (QoS)**
  - Port arbitration
    - Round robin
  - Request metering

- IDT proprietary feature that balances bandwidth among switch ports for maximum system throughput

– High performance switch core architecture
  - Combined Input Output Queued (CIOQ) switch architecture with large buffers

◆ **Clocking**
– Supports 100 MHz and 125 MHz reference clock frequencies
– Flexible port clocking modes
  - Common clock
  - Non-common clock
  - Local port clock with SSC (spread spectrum setting) and port reference clock input

◆ **Hot-Plug and Hot Swap**
– Hot-plug controller on all ports
  - Hot-plug supported on all downstream switch ports
– All ports support hot-plug using low-cost external I^2C I/O expanders
– Configurable presence-detect supports card and cable applications
– GPE output pin for hot-plug event notification
  - Enables SCI/SMI generation for legacy operating system support
– Hot-swap capable I/O

◆ **Power Management**
– Supports D0, D3hot and D3 power management states
– Active State Power Management (ASPM)
  - Supports L0, L0s, L1, L2/L3 Ready, and L3 link states
  - Configurable L0s and L1 entry timers allow performance/power-savings tuning
– SerDes power savings
  - Supports low swing / half-swing SerDes operation
  - SerDes associated with unused ports are turned off
  - SerDes associated with unused lanes are placed in a low power state

◆ **Reliability, Availability, and Serviceability (RAS)**
– ECRC support
– AER on all ports
– SECDED ECC protection on all internal RAMs
– End-to-end data path parity protection
– Checksum Serial EEPROM content protected
– Ability to generate an interrupt (INTx or MSI) on link up/down transitions

◆ **Initialization / Configuration**
– Supports Root (BIOS, OS, or driver), Serial EEPROM, or SMBus switch initialization
– Common switch configurations are supported with pin strapping (no external components)
– Supports in-system Serial EEPROM initialization/programming

◆ **On-Die Temperature Sensor**
– Range of 0 to 127.5 degrees Celsius
– Three programmable temperature thresholds with over and under temperature threshold alarms
– Automatic recording of maximum high or minimum low temperature

◆ **9 General Purpose I/O**

◆ **Test and Debug**
– Ability to inject AER errors simplifies in system error handling software validation
– On-chip link activity and status outputs available for several ports
– Per port link activity and status outputs available using external I^2C I/O expander for all remaining ports
– Supports IEEE 1149.6 AC JTAG and IEEE 1149.1 JTAG

◆ **Standards and Compatibility**
– PCI Express Base Specification 2.1 compliant
– Implements the following optional PCI Express features
  - Advanced Error Reporting (AER) on all ports
  - End-to-End CRC (ECRC)
  - Access Control Services (ACS)
  - Device Serial Number Enhanced Capability
  - Sub-System ID and Sub-System Vendor ID Capability
  - Internal Error Reporting
  - Multicast
  - VGA and ISA enable
  - L0s and L1 ASPM
  - ARI

◆ **Power Supplies**
– Requires three power supply voltages (1.0V, 2.5V, and 3.3V)

◆ **Packaged in a 19mm x 19mm 324-ball Flip Chip BGA with 1mm ball spacing**

Block Diagram

Figure 1 illustrates the block architecture of the switch. It contains a flexible and high-performance switch core, 16 full-duplex SerDes capable of delivering PCI Express Gen 2 speeds, up to 16 ports with movable upstream ports, and a number of innovative and unique features that enables product differentiation, cost and power saving, and time-to-market improvement.

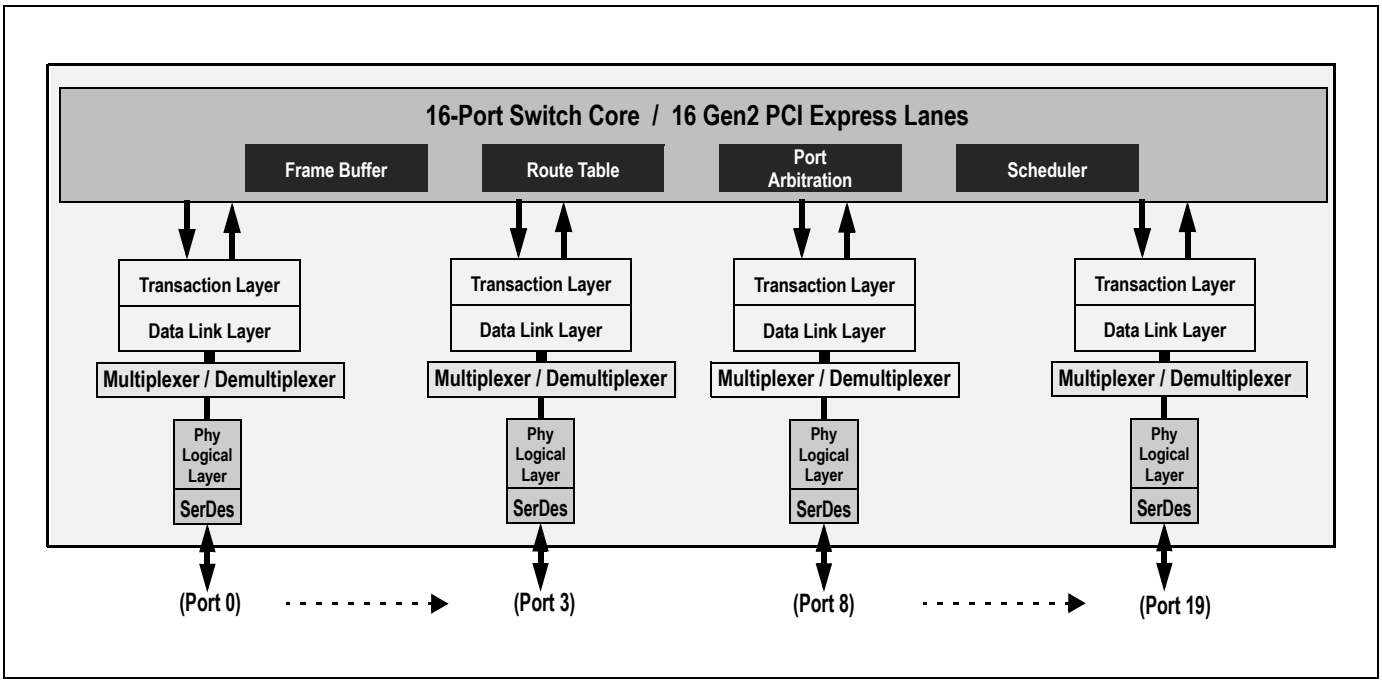

Figure 1 PES16NT16G2 Block Diagram

Applications

PES16NT16G2 is optimized for storage, communications control plane, embedded systems and high-port PCI Express fanout applications. Applications shown here are generic to IDT's Gen2 PCIe switch family.

Storage Applications

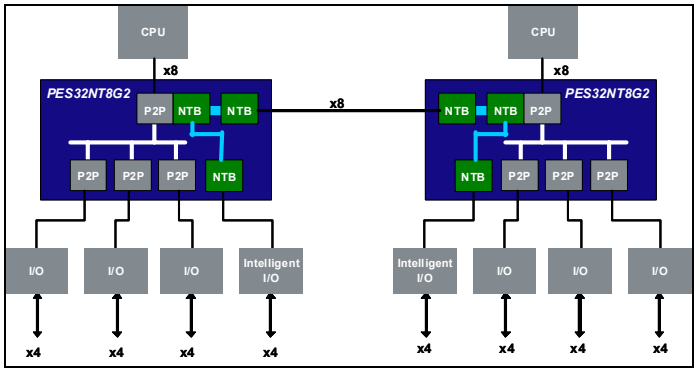

Figure 2 Storage Redundancy Model

Large RAID storage systems, either direct-attached or SAN/NAS attached, are generally built with redundancy. PCI Express switches are used to connect CPUs to I/O controllers and across the redundancy trunks, as illustrated in Figure 2. I/O controllers can be SAS, SATA, Fibre Channel, or some combination of these. NTB functions enable communications across PCI Express domains and allow data exchanges and synchronization across trunks. Additional NTB ports can also be connected to intelligent I/O cards hosting embedded processors.

Communications Control Plane Applications

High port count, dynamic switch partitioning configuration, and multiple NTB ports enable many possible communications control configurations. Some common configurations are depicted below.

High Fanout

A straightforward use of the switch is to fan out to many different line cards as illustrated in Figure 3. Cascading with another Gen 2 PCIe switch will allow more than twenty-three x1 downstream ports.

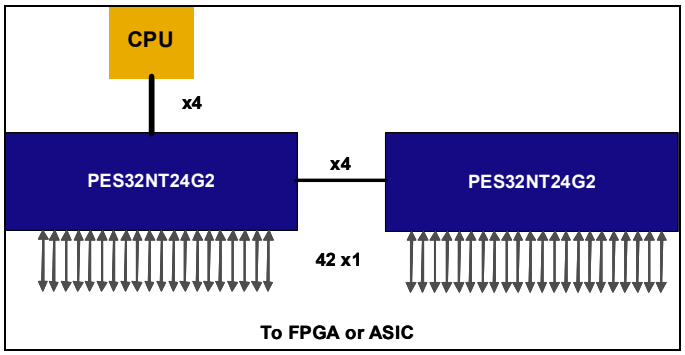

Figure 3 High Fanout Usage Model

Advanced Failover System

Dual host control with failover (shown in Figure 4) ensures reliability and availability of systems in case of CPU failure. NTB across domains allows active-active dual host topology, where both CPUs can access all the downstream line cards. If the primary CPU fails, dynamic switch partitioning capability can allow reallocation of downstream ports to the secondary CPU and maintain continuity of operation.

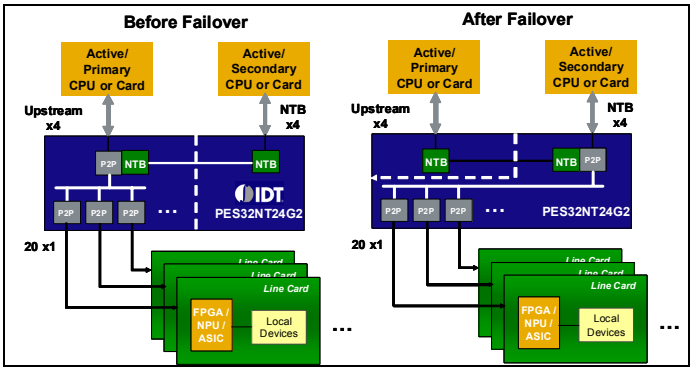

Figure 4 Dual-Host Failover System

Dual-Star Topology

A dual-star topology, shown in Figure 5, ensures that no single node or connection failure will bring down the system. Each I/O or line card has connectivity to one primary CPU and also backup connectivity to a secondary controller via an NTB port. If failure occurs, the NTB port can be reallocated via dynamic reconfiguration of the line card switch while the secondary CPU takes over. A high number of ports will allow the switch to connect to many line cards.

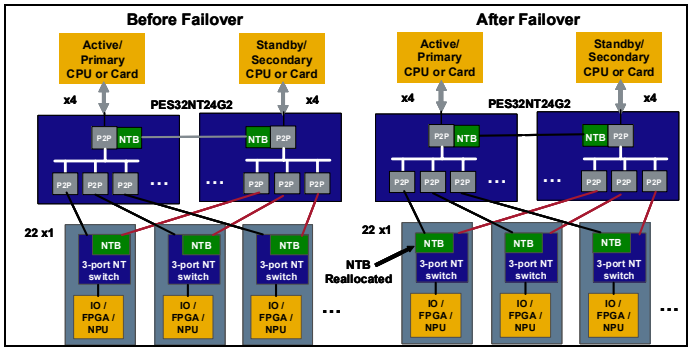

Figure 5 Dual-Star Topology

Feature Descriptions & Benefits

Switch Partitioning

Switch partitioning is an innovative and unique IDT feature that allows a switch to be statically or dynamically reconfigured into multiple independent logical switches within a single physical device. PES16NT16G2 can support up to 8 partitions. Any port can be an upstream port or downstream port and any root can have zero, one, or more downstream ports associated with its partition (see Figure 6). The partition configuration can be done statically or dynamically by writing into the switch configuration registers via configuration EEPROM, I^2C interface, or one of the roots.

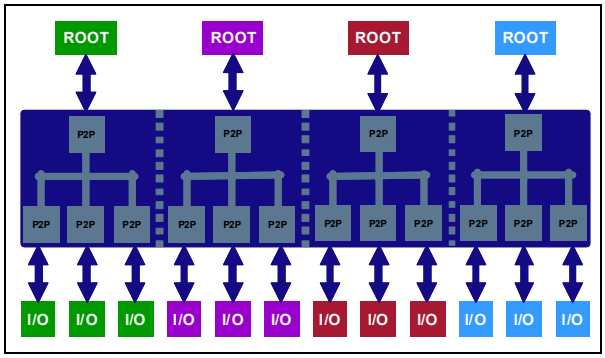

Figure 6 Example of Switch Partitioning Logical View

Switch partitioning enables a number of applications, allowing unique benefits and differentiating value proposition for your products. See Table 1 for a partial list of these benefits.

Application	Benefits
Replacing multiple discrete switches	Saves power, space and cost over multiple discrete PCIe switches
Bandwidth balancing in multi-root multi processor systems	Improved performance through optimal allocation of system resources
Flexible slot mapping	Saves power, space and cost over PCIe signal switch solutions Enables configurations that are not practical using PCIe signal switches
Port failover in high availability systems	Provides greater flexibility than movable upstream port or upstream port failover

Table 1 Switch Partitioning Applications and Benefits

Replacing Multiple Discrete Switches

When switch partitioning is configured with multiple independent PCI Express domains, it can replace multiple discrete PCI Express switches, providing savings in cost, power, and board space.

Bandwidth Balancing

Dynamic switch partitioning can be utilized to perform I/O bandwidth balancing to optimize overall system throughput (Figure 7). 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.

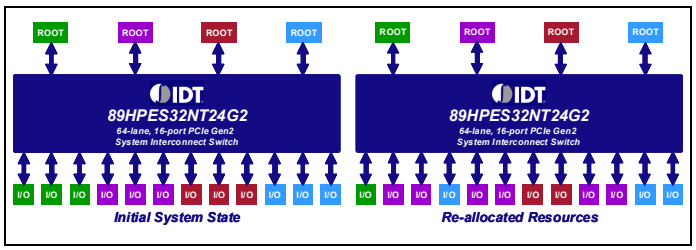

Figure 7 Dynamic Redistribution of I/Os to Optimize System Bandwidth

Flexible Slot Mapping for Hardware Re-Use

The flexibility of port mapping in switch partitioning allows maximum hardware re-use for multiple variants of product line configurations to meet the customized needs of your end customers, saving cost and improving time to market.

Figure 8 below illustrates a 2-socket CPU vs. a 4-socket CPU configuration using the same hardware platform with a different switch partitioning setup.

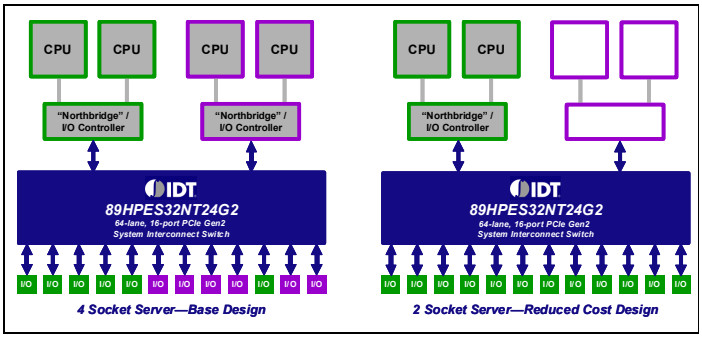

Figure 8 Example of Flexible Slot Mapping

Advanced Failover

Multi-root systems with high availability requirement can take advantage of dynamic switch partitioning by re-allocating downstream ports to a standby/secondary root upon failure (Figure 9). The device provides a built-in automatic failover mechanism by specifying failover configuration registers. Failover can be initiated by software, external signal pins, or by a watchdog timer.

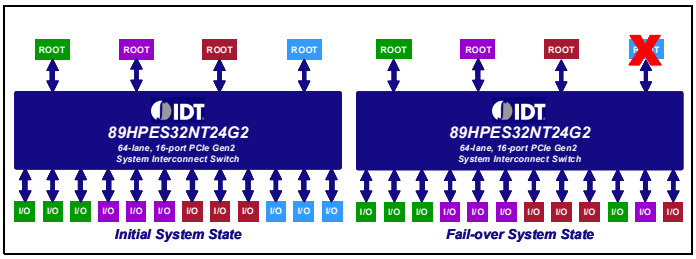

Figure 9 Example of Advanced Port Failover

Non-Transparent Bridging (NTB)

A non-transparent bridge (NTB) is required when two PCI Express domains need to communicate with each other. The main function of the NTB block is to initialize and translate addresses and device IDs to allow data exchange across PCI Express domains. The major functionalities of the NTB block are summarized in Table 2 .

Function	Number	Description
NTB ports	Up to 4	Each device can be configured to have up to 4 NTB functions and can support up to 4 CPUs/roots.
Mapping table entries	Up to 64 for entire device	Each device can have up to 64 masters ID for address and ID translations
Mapping windows	Six 32-bits or three 64-bits	Each NT port has six BARs, where each BAR opening an NT window to another domain
Address translation	Direct-address and lookup table translations	Lookup-table translation maps the BAR address to an entry in a lookup-table memory up to 64, providing more flexibility in address translation
Doorbell registers	32 bits	Doorbell register is used for event signaling between domains, where an outbound doorbell bit sets a corresponding bit at the inbound doorbell in the other domain
Message registers	4 inbound and outbound registers of 32-bits	Message registers allow mailbox message passing between domains -- message placed in the inbound register will be seen at the outbound register at the other domain

Table 2 Non-Transparent Bridge Function Summary

Each switch partition can have its own NTB port and can communicate with other partitions as well as NTB ports that connect to an external domain.

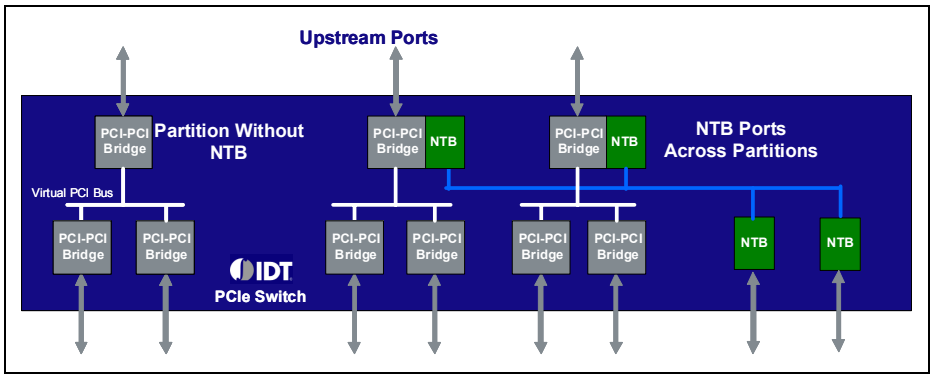

Figure 10 Possible Configuration of NTB Ports

Multicast

Multicast provides the ability to distribute data from one sender to multiple recipients simultaneously, off-loading processor cycles. The PES16NT16G2 supports up to 64 multicast groups and is fully compliant to the PCI-SIG Multicast ECN. This device also supports multicast across switch partitions. Applications include storage mirroring and data or table update in communications control planes.

DMA

The DMA engines are used to transfer large amounts of data to and from system memory, off-loading the CPU cycles to focus on data processing and manipulation, thereby increasing system performance. Along with NTB, the DMA engines can be used in Storage systems, large data or control planes in networking equipment, High-Performance Computing platforms, and other embedded systems.

Isolated Port Clocking With SSC

In addition to the global clock network, the PES16NT16G2 supports independent per-port clocking with SSC. Each port can operate in the common or separate clock configuration. This IDT-unique capability provides greater flexibility in board design and configuration for modular systems and connectivity across backplanes or cables. The SSC option allows EMI reduction to meet FCC standards, where applicable.

Hardware Error Containment

Traditional error handling mechanisms in most systems are handled by software in the root processor. Often, delays in error notification and processing allow errors to spread through the system. A hardware error containment feature blocks the spreading of errors by isolating bad packets and preventing them from leaving the switch, thereby avoiding system contamination. Benefits of hardware error containment include increased robustness in high-availability systems and insurance against a variety of error-handling scenarios in arbritrary CPUs/endpoints.

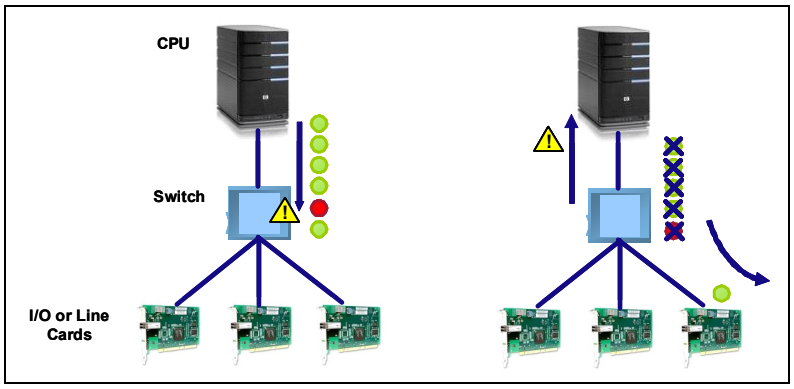

Figure 11 Example of Hardware Error Containment Mechanism

Evaluation Kit and Software Tools

Evaluation kit 89KTPES16NT16G2 will be available with evaluation board, user manual, schematic, and board layout.

The software package will include:

- PCIe Browser: GUI application to read/write to PCIe switch registers
- Switch Partitioning API: Allows dynamic reconfiguration of switch partitioning in Gen2 system interconnect switches
- Multicast API: Allows multicast configuration in Gen2 system interconnect switches
- DMA Driver: Activates the interface to DMA engines in PCIe switch
- System Interconnect software: Connects multiple processors using PCIe as the system interconnect
- Hot Swap Device Driver: Supports the hot insertion/removal of PCIe devices including switches

CORPORATE HEADQUARTERS
6024 Silver Creek Valley Road
San Jose, CA 95138

for SALES:
800-345-7015 or 408-284-8200
fax: 408-284-2775
www.idt.com

for Tech Support:
email: ssdhelp@idt.com
phone: 408-284-8208