Switch Partitioning in PCI Express® Switches

Application Note AN-708

Notes

By Subi Windoro

Introduction

With the continued demand for higher system performance, lower cost, lower power, and scalability, system designers look to PCI Express® switches to support more sophisticated capabilities to enable differentiators in their end products. Multi-CPU topology is commonly deployed in the server, storage, communications, and bladed systems and requires switches to be able to support multi-root capability.

Fortunately, not all switches are created equal. IDT's latest Gen2 system interconnect family implements Switch Partitioning, an innovative and unique feature that enables a single device to be configured into multiple logical and independent partitions.

Switch partitioning enables several applications and usage models. Table 1 summaries these applications and associated benefits.

Applications	Benefits
Replace multiple discrete PCIe switches with one switch	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 – replace PCIe signal switch	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 Applications and Benefits of PCIe Switch Partitioning

PCI Express Multi-Domain Challenges

A PCI Express tree contains one root complex and zero or more endpoints. A root complex is generally a microprocessor that controls its entire domain and has access to the system memory. When more than one endpoint exists in a system, a switch is required to manage the traffic transactions across different nodes per PCI Express protocol. Figure 1 illustrates an example of a PCIe® switch configuration.

Notes

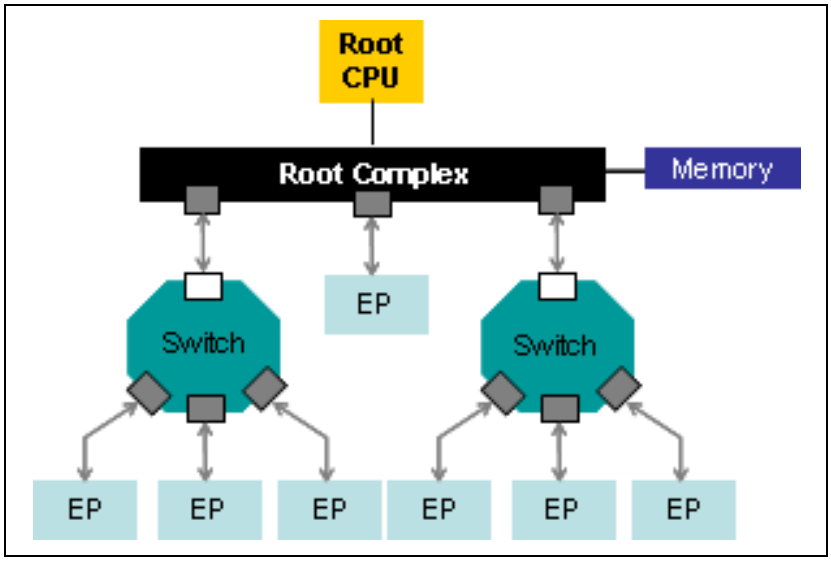

Figure 1 Example of PCIe Configuration with Switches

PCI Express is a highly robust and a very high performance protocol standard. As its adoption has spread to many different application segments, such as servers, communications, storage, and embedded systems, new usage models and system requirements pose new challenges that require innovative solutions.

Challenge #1: Higher performance with reduced cost, power and footprint

Multi-processor systems designed to meet the ever-increasing need for higher performance often come with cost, power, and board space constraints. Multi-root systems generally require multiple PCIe switches. One or more switches may be needed for each PCIe domain.

Challenge #2: Time to market with hardware re-use

Customization of hardware platforms to meet varying end customer needs can be expensive and time consuming. System OEMs and ODMs have consistently communicated a need for hardware re-use to reduce manufacturing expenditures and time to market.

Challenge #3: Reliability, Availability, and Serviceability

High-end servers, storage systems, and control systems for networking equipment have a high level of reliability, availability, and serviceability (RAS) requirements, utilizing redundant or failover topology. For systems that require multiple intelligent processors, however, solutions with PCI Express are not straightforward and generally require multiple switches and complex software implementations. Each processor in a multi-processor system will own its separate domain and cannot share its endpoints unless an address translation bridge is built.

The traditional solution for multi-processor PCI Express systems is to add multiple discrete switches where each switch generally has a fixed configuration of endpoints and roots with limited or no communication among the domains, as shown in Figure 2.

Notes

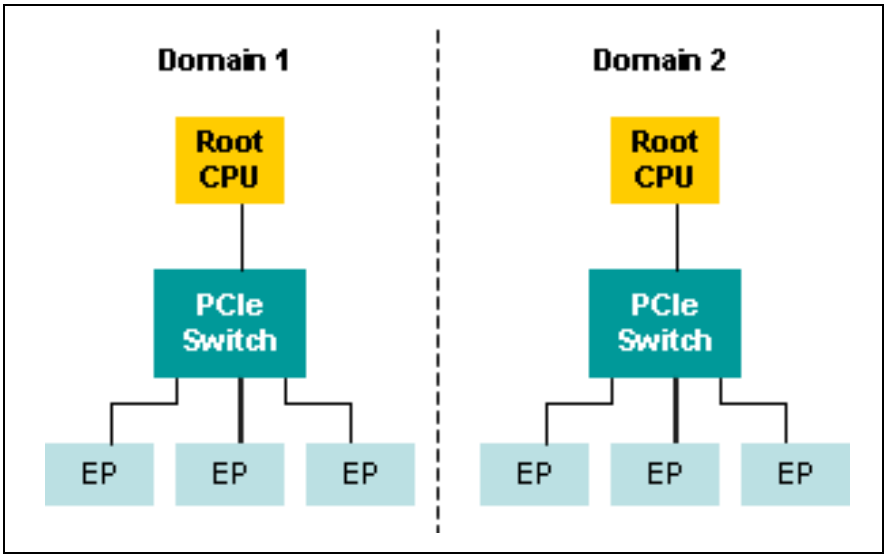

Figure 2 Traditional Multiple PCIe Switches Configuration for Multiple Roots

Switch Partitioning

Switch partitioning is a very simple yet elegant and innovative concept to address multi-root challenges in PCI Express systems. In a nutshell, switch partitioning is the capability to create logical switch partitions within the same device.

IDT's Gen2 system interconnect families of switches support switch partitioning with the following properties:

- Each device can be configured to contain multiple partitions, up to 16 partitions in the largest device
- Each partition is an independent PCI Express domain with one root and zero or more downstream ports
- Flexibility in port configuration where
  - Any port can be configured as an upstream or downstream port
  - Any downstream port can be assigned to any root or partition
- Partition configuration can be done statically via configuration EEPROM or dynamically via in-band commands from one of the roots or the SMBus interface.

Figure 3 shows a switch with 4 independent partitions, each with its root and downstream port assignments.

Notes

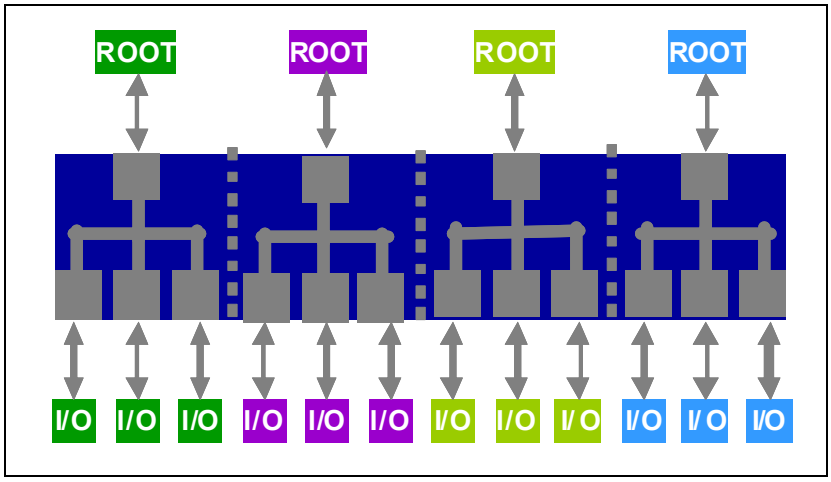

Figure 3 Example of Switch Partitioning

Switch Partition Configuration

Configuring switch partitions is done by writing into the switch port configuration registers to set the switch partition ID, port modes (upstream, downstream, or unattached) and downstream port associations. Configurations of the switch partitioning can be done via configuration EPROM, SMBus, or in-band instructions from one of the upstream roots.

Static configuration is a straightforward use of switch partitioning when the logical partitioning of the root systems and their downstream ports or endpoints are fixed and is not expected to change. This configuration image can be programmed from the configuration EPROM or from one of the master roots upon system initialization.

Dynamic reconfiguration represents the ability to change the mode of operation of any port in a partition while the system and the switch are active. Partition reconfiguration occurs when one or more of the following occurs:

- A downstream port is added to or removed from a partition
- An upstream port is added to or removed from a partition
- The operating mode of the upstream port is modified

Partition reconfiguration must be managed at the software application level to determine the start and completion time, and traffic at the affected ports must be quiesced before the dynamic reconfiguration is started.

When dynamic reconfiguration is taking place, unaffected partitions will not be disturbed and traffic can continue to flow normally through those partitions.

Applications and Benefits

The combination of supporting multiple roots/domains in a single device and dynamic re-configuration allows many different usage models and unique system configurations that bring a number of benefits and differentiating value propositions to the end products. The following section discusses each usage model and its benefits.

1. Replacing Multiple Discrete Switches

The obvious advantage to switch partitioning is replacing multiple discrete switches with a single device. Benefits include smaller footprint, lower BOM cost, and lower power consumption. Refer to Figure 4.

Notes

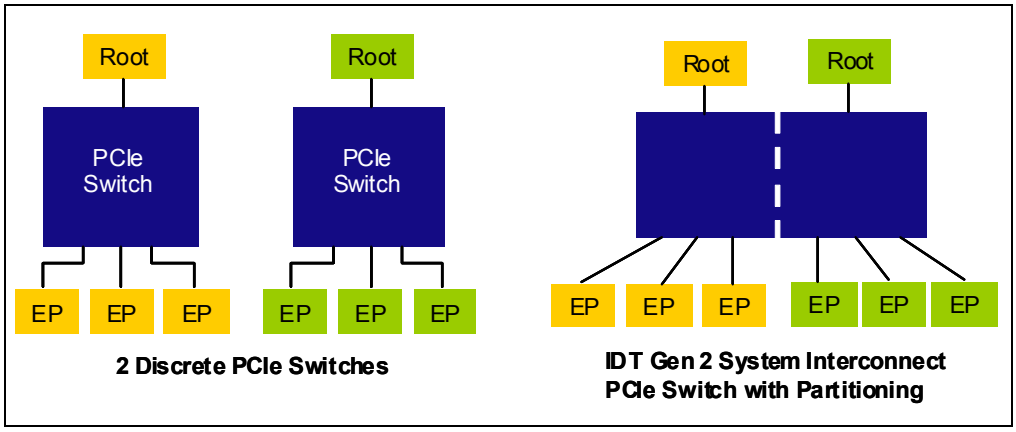

Figure 4 Replacing Multiple Switches with Switch Partitioning

2. Advanced Failover

Many high-reliability systems employ dual-host failover topology, where one CPU acts as the primary CPU or root and another CPU serves as backup or secondary CPU. Such an implementation in traditional switches is very challenging and has several limitations. Recent new switches from some vendors also tried to solve the problem with very complex bridging solutions.

Switch partitioning allows the failover system to seamlessly reconfigure the switch upon failure by reallocating all downstream ports to the backup CPU and then resuming the data traffic. The partitioning reconfiguration process is initiated and executed dynamically by the secondary CPU once failure has been detected. Refer to Figure 5.

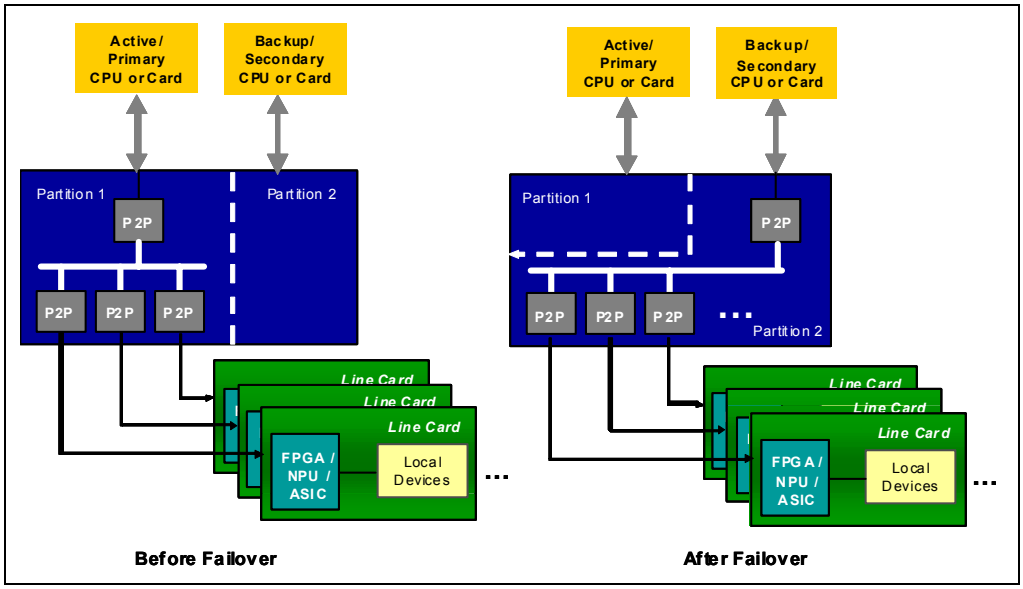

Figure 5 Advanced Failover with Switch Partitioning

Advanced failover topology provides high reliability, availability, and serviceability, while switch partitioning offers lower development costs, lower risk, and lower complexity by providing a simple, straightforward hardware and software solution.

Notes

3. Flexible Slot Mapping / Hardware Re-use

The third usage model of switch partitioning is flexible slot mapping, where a number of different product configurations or part numbers are built from the same hardware platform. Because switch configuration allows any port to be downstream or upstream and any downstream port can be allocated to any partition, the same device and board layout can be programmed as a different product configuration to customize end customer's needs. Refer to Figure 6.

Flexible slot mapping allows high hardware re-use, saving development and product cost and significantly improves time-to-market.

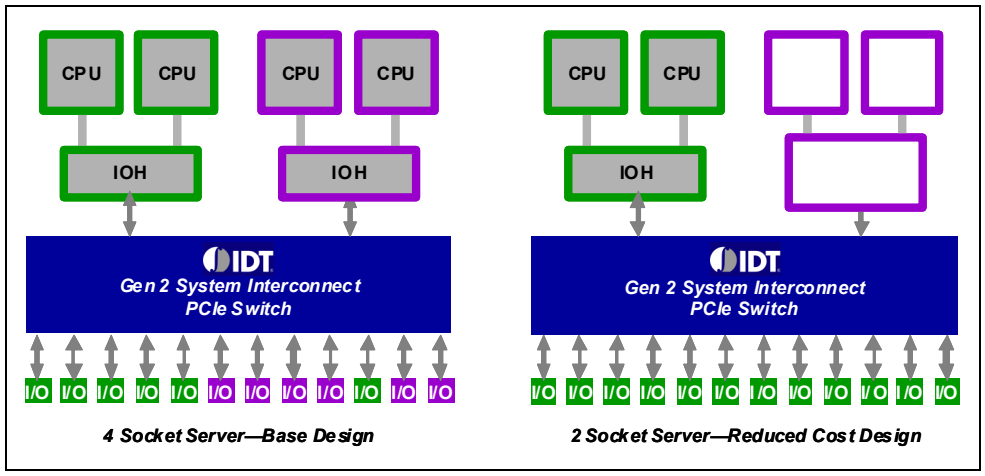

Figure 6 Example of Flexible Slot Mapping with Switch Partitioning

4. Load Balancing

A unique usage model that is only possible with switch partitioning is load balancing. In multi-CPU computing systems, traffic capacity across a number of independent domains may not be equivalent. Load balancing is a process where some or all downstream ports are dynamically re-allocated from an under-utilized or an idle partition to heavy-traffic domains. Refer to Figure 7.

Load balancing with switch partitioning optimizes overall system throughput with a single-switch solution, saving cost and power.

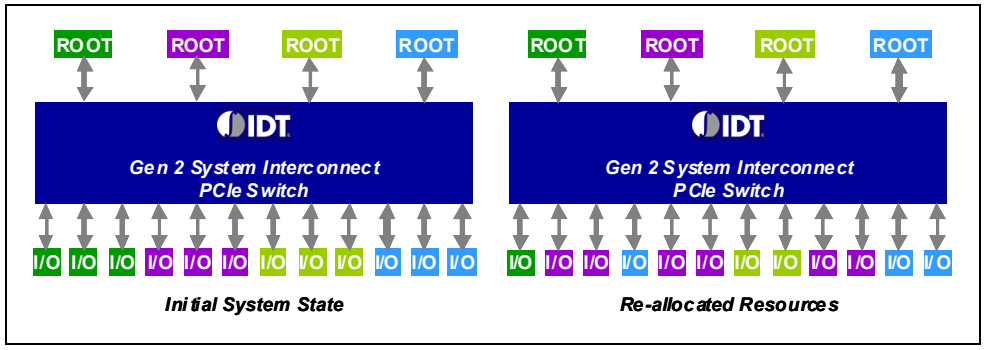

Figure 7 Load Balancing with Switch Partitioning

References

Details of the switch partitioning architecture, usage, and configuration registers can be found in the IDT Gen2 system interconnect user manuals. Please contact your IDT sales representative for additional information.