White Paper

Hot and Surprise Plug Recommendations for Enterprise PCIe Switches in the Data Center (Short Form)

Preliminary

May, 2016

Microsemi

Power Matters.™

Abstract

This short form white paper describes the software and hardware needed to support hot and surprise plug of PCIe Express storage devices with Enterprise Storage PCIe switches in Data Center, Storage, and Server applications. This white paper is meant to be used in conjunction with Intel's white paper #539126 to provide an end-to-end description of hot plug with PCIe switches and NVMe targets. In particular, this paper aims to address the more complex use cases involved with hot-plugging devices behind PCIe switches in transparent configurations, and addresses specific challenges involved with these configurations.

Contents

Preface

Introduction

This document serves as a companion to the Intel white paper #539126: "Hot Plug Recommendations for Enterprise PCIe SSD in Data Center." As a result, the terms outlined in this document regarding the concepts of software-managed hot add/remove and surprise hot add/remove will follow the conventions outlined in the Intel document.

Formatting Conventions

For easier readability, items in ALL CAPS refer to configuration space registers. Items that are colored in ALL CAPS refer to hardware pins.

References

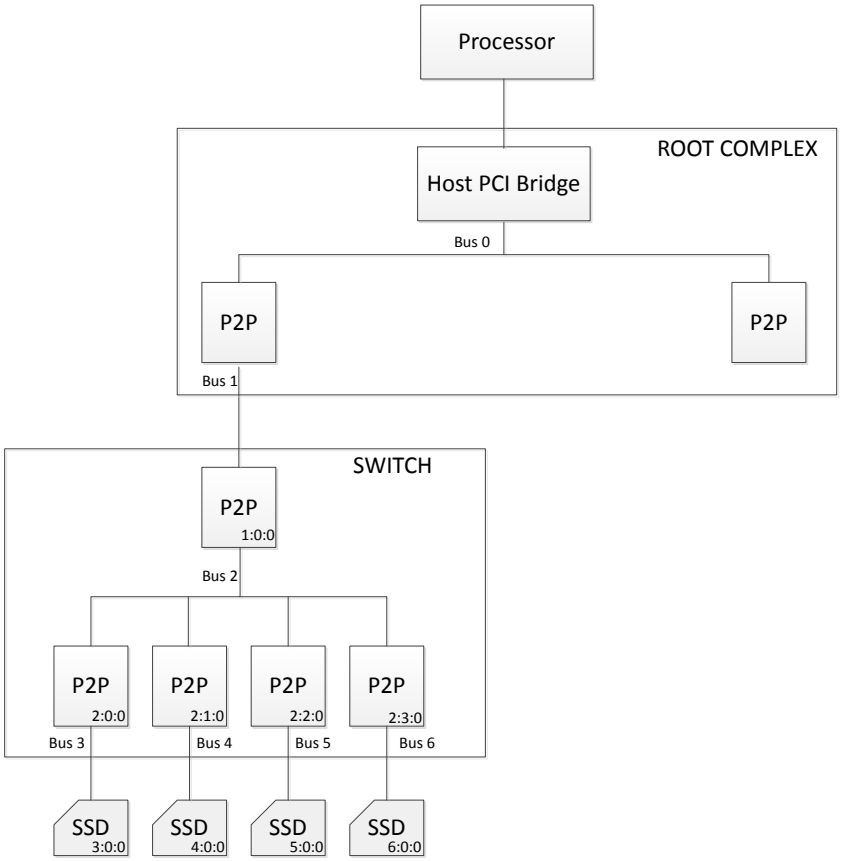
1. *Hot Plug Recommendations for Enterprise PCIe SSDs in the Data Center.* #539126. Rev. 1.0. Intel. October 2013.
2. *PCI Express Base Specification*. Rev. 3.1. PCI-SIG. Oct. 8 2014

Overview

This document covers similar use cases as the Intel white paper, with a focus on the role of a PCIe switch and recommended behaviors required to ensure the highest compatibility level with industry-leading OS and BIOS software as well as SSD hardware/firmware.

Figure 1 illustrates a basic topology consisting of a host/root complex, a switch, and some downstream endpoints.

Figure 1 • Basic Topology



The switch appears as a series of PCI-to-PCI (P2P) bridges. The port connected to the host is referred to as the upstream port (USP) and the P2P associated with that port is called the upstream P2P (US P2P). Similarly, the ports connected to the PCIe SSDs (or other endpoints) are referred to as downstream ports (DSP). The P2Ps connecting to the DSPs are known as downstream P2Ps (DS P2P).

Each upstream and downstream P2P appears as an endpoint in the PCIe topology with its own configuration space register (CSR). The P2P function uses a Type 1 configuration header that is used by the system to discover the existence of the P2P function, configure the operation of the P2P, and check the status of the function.

PCIe hot plug is supported through the P2P endpoint CSR through a combination of status registers (read by the host) and control registers (modified and controlled by the host). Events such as a PRESENCE DETECT CHANGED result in an interrupt (INTx or MSI) to be fired, which, in turn, will cause the host to read the P2P configuration space for more information.

Assumptions

When considering the concept of a managed hot add/remove, and surprise add/remove, we make the following assumptions:

- Managed events are well orchestrated and the outstanding I/Os are stopped/completed prior to the time of the hot plug
- Surprise events are fundamentally not well orchestrated and result in complex error handling

For the context of this white paper:

- Downstream devices will be assumed as being enterprise PCIe NVMe SSDs
- Downstream devices are assumed to be more robust against surprise events such as hot add/remove when compared to upstream host ports
- It is assumed the drives will perform sufficient error handling to put themselves in a recoverable state when they are removed and re-inserted
- All discussions in this document pertain to transparent switch operation

Hardware Recommendations

Hosts, switches, and devices should fully support the hot plug specification as defined in the *PCI Express Base Specification Revision 3.1*.

Root Complex Recommendations

Host CPUs need to support hot plug as per *PCI Express Base Specification Revision 3.1*. For surprise device removal support, the PCIe switch must assist with handling outstanding completions, or a host must support an equivalent mechanism to return outstanding non-posted (NP) completions from a removed device. For surprise switch removal support, a host must support the handling of outstanding NP read completions or a similar mechanism to prevent the host from timing out while waiting on an outstanding NP completion.

OS Recommendations

A host OS with support for INTx or MSI and a hot plug-aware device driver (NVMe in the use cases covered here) are required to support both managed and surprise hot add and removal. Below is a list of supported OS versions as sourced from the Wikipedia topic "MSI":

- Microsoft Windows Longhorn and later
- FreeBSD 6.3 or later
- Solaris Express 6/05
- CentOS 7
- Linux kernels 3.19 or later
- For DPC support, Linux kernels 4.4 or later

The support for hot plug at the NVMe driver level is outside of the scope of this document, but most modern NVMe drivers do support hot plug natively. Some OS out-of-box drivers, however, do not natively support hot plug.

Downstream Port Containment

Downstream Port Containment, or DPC, is a PCIe 3.1 protocol-defined mechanism for detecting uncorrectable errors/messages and shutting downstream ports to prevent the spread of data corruption. DPC is a robust error containment mechanism that provides the opportunity for systems to recover gracefully. DPC can be used to manage asynchronous removal (surprise hot remove) events. A surprise down event is defined as an unexpected link down. Hot resets, for example, are not surprise down events.

Legacy Surprise Down Handling

Legacy PCIe is not adept at handling these kinds of events and, in general, a surprise down will cause the host to crash. The cause of the host crash is related to uncorrectable errors (UE) such as ERR_FATAL messages being returned to the host when an endpoint suddenly disappears from the topology. When hot plug support was added to legacy PCIe, a mechanism was added in configuration space (HOT PLUG SURPRISE) to mask fatal UE messages to the host that occur due to a surprise down error. This prevented the host from crashing during hot remove operations.

Note:
HOT PLUG SURPRISE can only suppress UE errors that occur due to a surprise down error event. This does not mask other errors related to surprise removal.

Current Surprise Down Handling

When PCI-SIG introduced DPC, it became the preferred mechanism for handing surprise hot remove events. DPC senses unmasked UEs and takes error handling steps to log, process, and respond to them. Because DPC triggers on unmasked UEs related to a surprise down error, it is critically important that legacy mechanisms like HOT PLUG SURPRISE do not mask these UEs. This allows DPC to manage the error containment.

Note:
Since DPC is the PCI-SIG preferred mechanism for handing surprise hot remove events, the slot capability register **HOT PLUG SURPRISE is no longer recommended and should be configured with a value of '0'**. For more information, see the PCIe Base Specification 3.1 Implementation Note "Avoid Disable Link and Hot Plug Surprise Use with DPC."

System Preparation

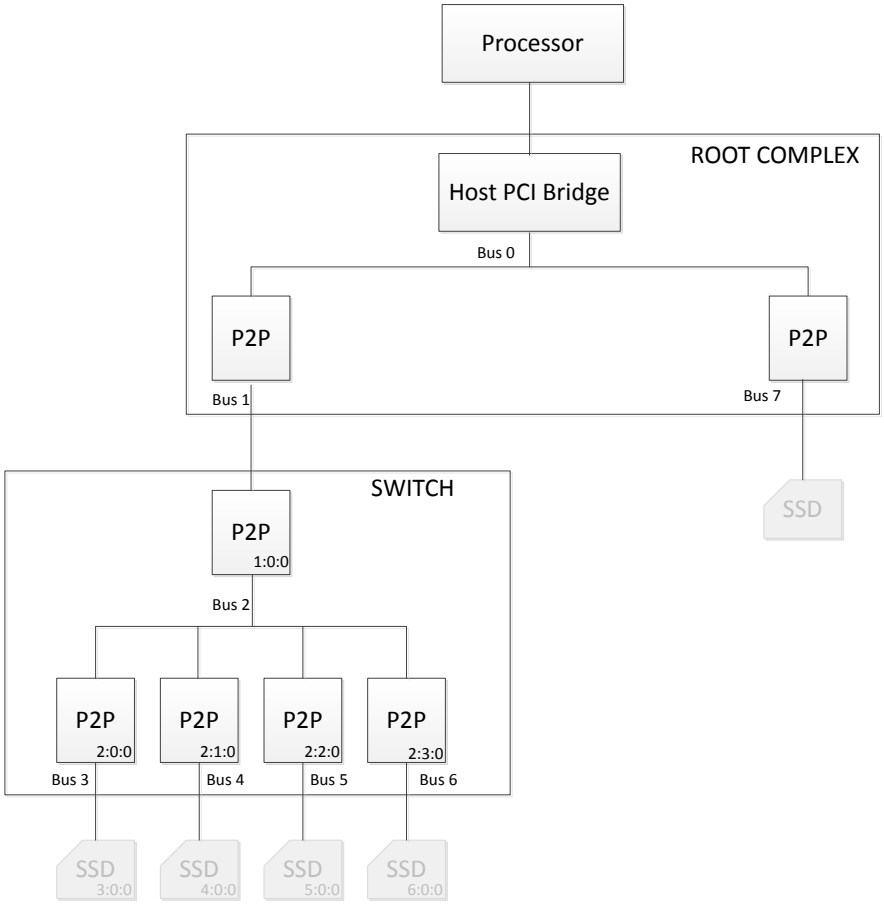
BIOS Considerations

This section discusses the impact of switches on BIOS design and planning. Consider Figure 2, in which a root complex is connected to a switch over one bus (0) and a potential target on the other bus (7).

Resource Automatic Allocation for Supporting Hot Add of Endpoints

When using the PCIe Switch in a transparent mode of operation where endpoints can be hot added or removed, the switch simply needs to be configured with P2P bridges for each hot pluggable endpoint in the topology. When the host enumerates the switch, each P2P is initialized and resources for endpoints are automatically allocated by the BIOS (this is true for most current BIOS designs). At this stage, endpoints can be hot added and removed and the switch will facilitate the in-band messaging required to signal to the host and OS that new endpoints have been added or removed.

Figure 2 • Root Complex Connected to Switch and Target



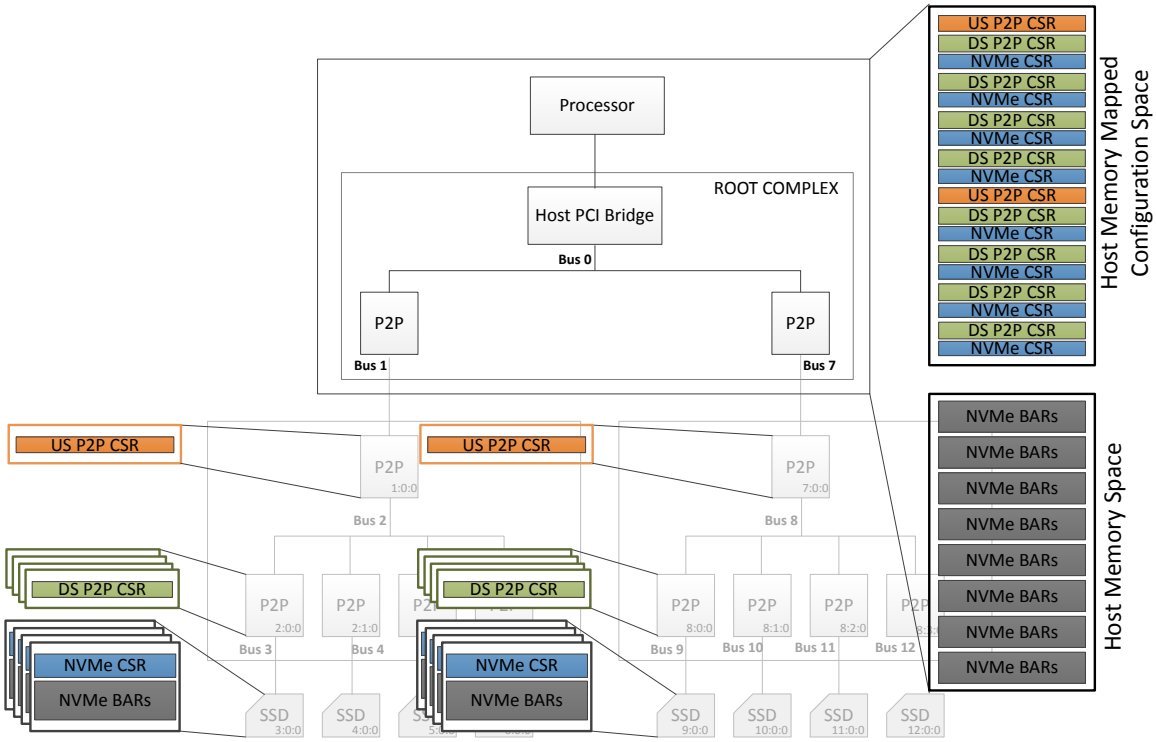
For a system that boots with this topology, the BIOS will perform a depth-wise search of the PCIe topology during enumeration. This results in bus/device/function assignment as illustrated in Figure 2. In this example, each DS P2P has been assigned a bus/device/function but none of the DS P2Ps have physical endpoints. During enumeration, the host automatically allocates resources for each P2P to account for an endpoint that could be added at a later time. The amount of resource allocation will vary from BIOS vendor to BIOS vendor, but most vendors will allocate sufficient resources to support an NVMe drive.

Resource Preallocation for Supporting Hot Add of a Switch

When planning for a system topology where an entire switch can be hot plugged to a running host, the BIOS needs to reserve (preallocate) sufficient resources to support the connectivity to and management of all possible endpoints attached to the system. This includes the PCI identifiers (bus/device/function), configuration, and memory space for the NVMe targets and the configuration space for the downstream and upstream P2Ps.

Figure 3 illustrates a transparent topology where two switches can be hot added to the system with four drives each below them.

Figure 3 • Transparent Topology



For the host DS P2Ps to properly support these switches, the BIOS needs to preallocate resources for the maximum topology (8 SSDs and 2 switch upstream P2Ps). The BIOS also needs to reserve bus/device//function identifiers for all of the available ports in the topology (in this case, 2 upstream P2Ps, 8 downstream P2Ps, and 8 SSDs). If this host were to boot with nothing attached, a poorly preallocated BIOS would assign bus numbers 1 and 2. This would prevent a switch to be added to Bus 1. In the example in Figure 3, Bus 1 and Bus 7 are preallocated in preparation for supporting up to 5 more buses to be connected to Bus 1. This preallocation must be planned by the BIOS and systems engineers in advance.

Once the BIOS is prepared in this way, hot add and hot removal of devices can be supported by the higher level software/OS.

> **Note:**
> When communicating to NVMe targets over PCIe, host-initiated PCIe memory I/O reads occur only during initialization and error handling. This is because data reads from an NVMe drive are actually performed by the drive writing to host memory space and an associated doorbell mechanism is used to synchronize the data flow. This architecture makes it simpler, from the BIOS perspective, to support NVMe targets of varying capacity. The core memory requirements are fixed and unrelated to the capacity of the endpoint.

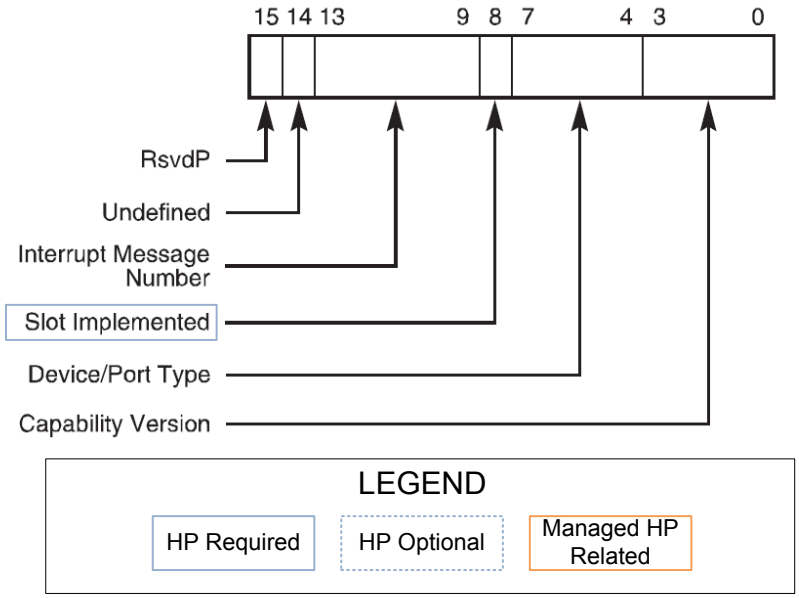
Hot Plug-Related PCI Express Capability Registers

The following downstream P2P configuration registers are used in hot plug management with NVMe drives. Each section outlines a PCI Express Capability Register used in hot plug management.

PCI Express Capabilities Register (Offset 0x02h)

In Figure 4, the SLOT IMPLEMENTED field must be set in order to support an endpoint on that port. SLOT IMPLEMENTED is required even for statically attached endpoints.

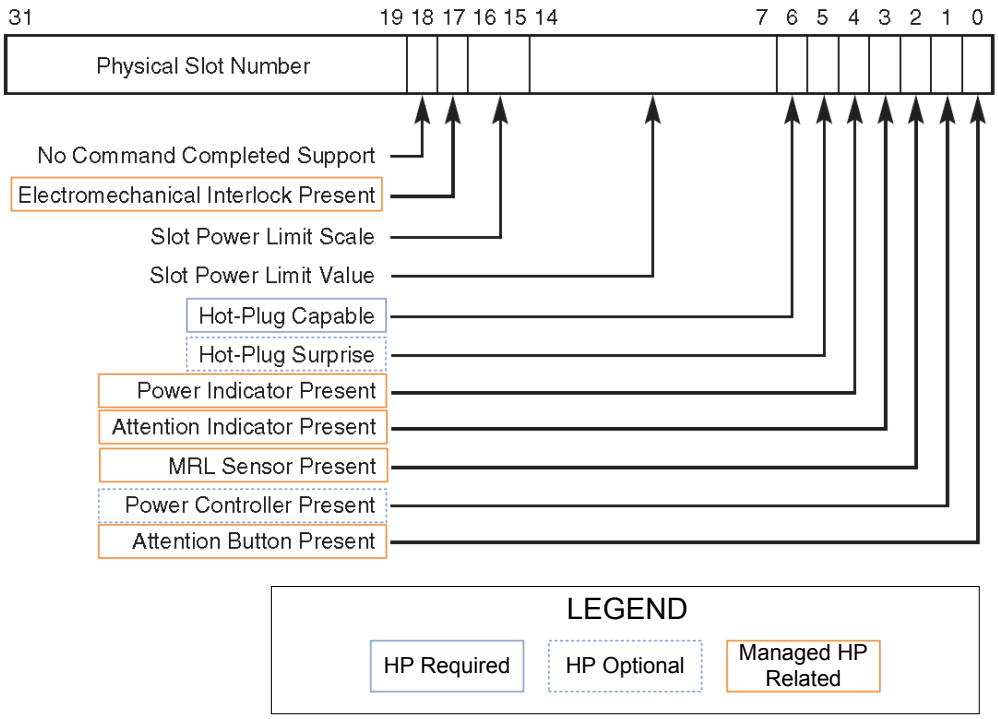
Figure 4 • PCI Express Capability Register Fields Used in Hot Plug



Slot Capabilities Register (Offset 0x14h)

In Figure 5, a mixture of registers are used to support hot plug. Some fields are specific to managed hot plug while others are general hot plug signals. HOT PLUG SURPRISE is a special case, which is noted below.

Figure 5 • Slot Capability Register Fields Used in Hot Plug

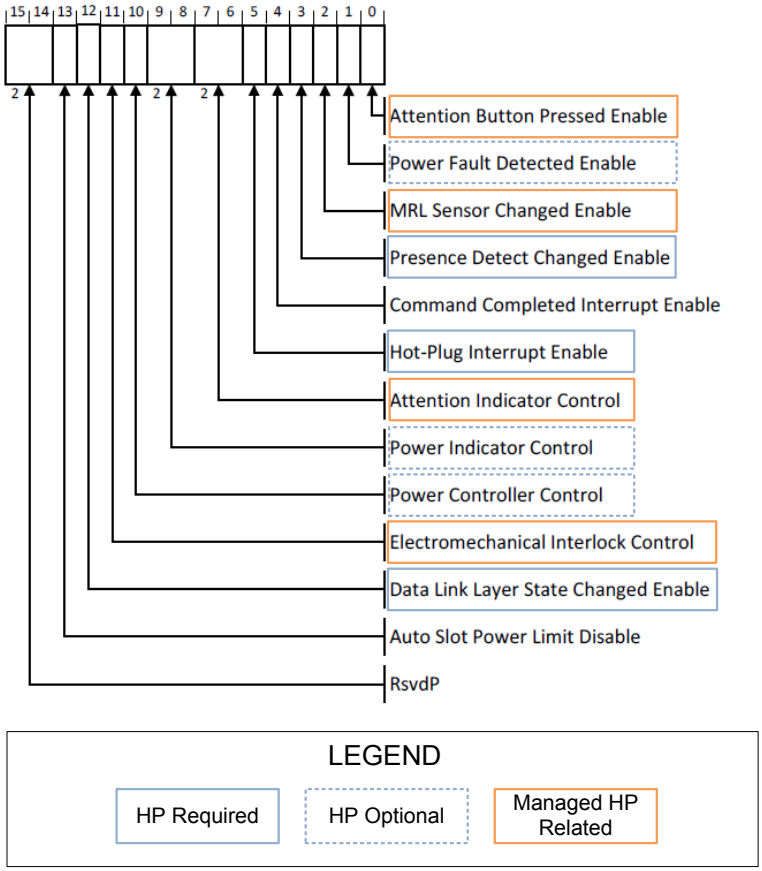


Note:
HOT-PLUG SURPRISE should be set to 0 when DPC is enabled, which will allow PCIe standard DPC mechanisms to manage surprise removal events. For more information see Downstream Port Containment on page 5.

Slot Control Register (Offset 0x18h)

The slot control registers shown in Figure 6 are used by the host to control and coordinate hot plug operations. Many fields are optional, based on the implementation of the system. The host may not have power control of the drives in some systems, so power indicator and power controller are likely not used.

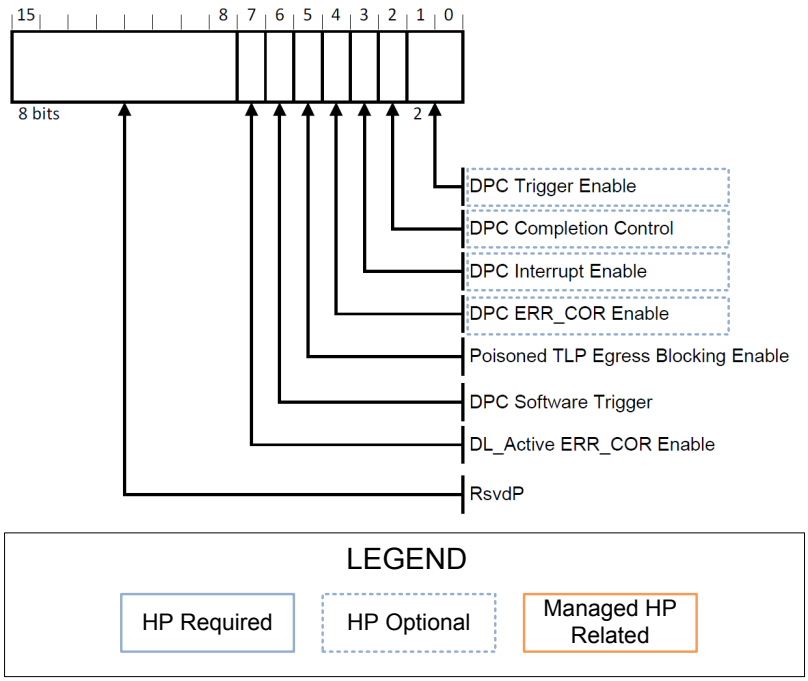
Figure 6 • Slot Control Register Fields Used in Hot Plug



DPC Control Register (Offset 0x06h in DPC Extended Capability)

The following control register enables or disables the use of DPC in the switch. This capability is written by the host and the default state of DPC as defined in the *PCIe Base Specification 3.1* is DISABLED. In Figure 7, the DPC Trigger Enable bit is used to enable or disable DPC. There are several other optional DPC features that can be controlled in this register.

Figure 7 • DPC Control Register Fields Used in Hot Plug



Note: DPC must be enabled if setting Hot Plug Surprise to '0', otherwise there will be no mechanism to stop ERR_FATAL from being propagated to the host.

Software-managed Hot Add/Remove of a Device

This section addresses the common use cases of software-managed hot add/remove. In all cases, the host is given a prior knowledge of the event that is about to happen and given time and control of when the event occurs. This provides the most reliable method of adding and removing PCIe endpoints through a switch.

Content available under NDA

The following topic is covered in the Long Form version of this white paper which is available under NDA via MSCC. Please contact your MSCC sales team for access.

Surprise Hot Add/Remove of a Device

This section addresses the use cases of a surprise (or asynchronous) hot add/remove. In all cases, the host is not given prior knowledge of the event that is about to happen and not given time and control of when the event occurs. The use case will cover a situation where a device is hot added while I/O is running through the switch to an adjacent drive and a situation where a device is hot removed while I/O is running to the removed device.

Content available under NDA

The following topic is covered in the Long Form version of this white paper which is available under NDA via MSCC. Please contact your MSCC sales team for access.

Conclusion

This white paper illustrates that it is indeed possible to support PCIe hot and surprise plug with existing software and NVMe targets when using them alongside an Enterprise PCIe storage switch. Special considerations are required when planning for surprise or managed hot plug, and it is important to keep in mind the specific sequences involved in each type of hot plug. BIOS planning is required when handling cases where an entire switch topology can be hot added to a system after the management host has already booted. Error containment mechanisms such as DPC, AER, and CTS should be considered when planning a hot plug system, as it will impact the types of capabilities the PCIe storage switch should advertise on its downstream P2Ps.

Microsemi Corporate Headquarters
One Enterprise, Aliso Viejo,
CA 92656 USA

Within the USA: +1 (800) 713-4113
Outside the USA: +1 (949) 380-6100
Sales: +1 (949) 380-6136
Fax: +1 (949) 215-4996
E-mail: sales.support@microsemi.com
www.microsemi.com

Microsemi Corporation (Nasdaq: MSCC) offers a comprehensive portfolio of semiconductor and system solutions for aerospace & defense, communications, data center and industrial markets. Products include high-performance and radiation-hardened analog mixed-signal integrated circuits, FPGAs, SoCs and ASICs; power management products; timing and synchronization devices and precise time solutions, setting the world's standard for time; voice processing devices; RF solutions; discrete components; enterprise storage and communication solutions; security technologies and scalable anti-tamper products; Ethernet solutions; Power-over-Ethernet ICs and midspans; as well as custom design capabilities and services. Microsemi is headquartered in Aliso Viejo, Calif., and has approximately 4,800 employees globally. Learn more at **www.microsemi.com**.

The technology discussed in this document may be protected by one or more patent grants.