PCI-SIG ENGINEERING CHANGE NOTICE

TITLE:	Root Complex Integrated Endpoint & IOV Updates
DATE:	Updated 19 Nov 2015 PWG Approved for release 19 Nov 2015
AFFECTED DOCUMENTS:	PCIe Base Specification Rev 3.1, Single Root I/O Virtualization and Sharing Specification Rev 1.1
SPONSOR:	Intel Corporation

Part I

1. Summary of the Functional Changes

This ECR implements a variety of spec modifications intended to correct inconsistencies related to, and to support more consistent implementation of, Root Complex integrated Endpoints, with a particular focus on issues relating to Single Root IO Virtualization (SR-IOV).

2. Benefits as a Result of the Changes

By addressing inconsistencies and gaps in our current specifications, we will enable more consistent hardware and software implementations.

3. Assessment of the Impact

The spec changes have been made in such a way as to minimize the expected disruption, but there is the potential that hardware and software changes could be required.

4. Analysis of the Hardware Implications

Hardware, particularly Root Complex implementations implementing support for SR-IOV, may require modifications to conform to and take advantage of this set of changes.

5. Analysis of the Software Implications

Software, particularly OS/VMM implementing support for SR-IOV, may require modifications to conform to and take advantage of this set of changes.

6. Analysis of the C&I Test Implications

Some PCIe register tests may be affected.

Part II

Detailed Description of the change

In the SR-IOV Specification, Sect 2.1.2. ("VF Discovery"), edit as shown:

…

VFs may reside on different Bus Number(s) than the associated PF. This can occur if, for example, First VF Offset has the value 0100h. A VF shall not be located on a Bus Number that is numerically smaller than its associated PF. A VF that is located on the same Bus Number as its associated PF shall not be located on a Device Number that is numerically smaller than the PF.<*ADD FOOTNOTE*: SR-IOV Devices immediately below a Downstream Port always have a Device Number of 0 and thus always satisfy this condition.>

VFs of a SR-IOV Root Complex Integrated Endpoint Device are associated with the same Root Complex Event Collector (if any) as their PF.

As in …

In the SR-IOV Specification, Section 3.3.2, edit as shown:

Table 3-2: SR-IOV Capabilities

Bit Location	Register Description	Attributes
…		
1	**ARI Capable Hierarchy Preserved** *PCI Express Endpoint:* If Set, the ARI Capable Hierarchy bit is preserved across certain power state transitions. *Root Complex Integrated Endpoint:* Not applicable – it is strongly recommended that this bit be hardwired to 0b.	RO

In the SR-IOV Specification, Section 3.3.2.2, add as shown:

3.3.2.2. ARI Capable Hierarchy Preserved

ARI Capable Hierarchy Preserved is Set to indicate that the PF preserves the ARI Capable Hierarchy bit across certain power state transitions (see Section 3.3.3.5). Components

must either Set this bit or Set the No_Soft_Reset bit (see Section 6.2). It is recommended that components set this bit even if they also set No_Soft_Reset.

ARI Capable Hierarchy Preserved is only present in the lowest numbered PF of a Device (for example PF_0). ARI Capable Hierarchy Preserved is Read Only Zero in other PFs of a Device.

ARI Capable Hierarchy Preserved does not apply to Root Complex Integrated Endpoints, and its value is undefined (see Section 3.3.3).

In the SR-IOV Specification, Section 3.3.3, edit as shown:

Table 3-3: SR-IOV Control

Bit Location	Register Description	Attributes
…		
4	**ARI Capable Hierarchy** *PCI Express Endpoint:* ~~The Device is permitted to locate VFs in Function numbers 8 to 255 of the captured Bus Number. Default value is 0b. This field is RW in the lowest numbered PF of the Device and is Read Only Zero in all other PFs.~~ This bit must be RW in the lowest numbered PF of the Device and hardwired to 0b in all other PFs. If the value of this bit is 1b, the Device is permitted to locate VFs in Function Numbers 8 to 255 of the captured Bus Number. Otherwise, the Device must locate VFs as if it were a non-ARI Device. Default value is 0b. *Root Complex Integrated Endpoint:* Not applicable – ~~must hardwire the bit~~ this bit must be hardwired to 0b. Within the Root Complex, VFs are always permitted to be assigned to any Function Number allowed by First VF Offset and VF Stride rules (see Sections 3.3.9 and 3.3.10).	RW or RO (see description)

In the SR-IOV Specification, Section 3.3.3.5, edit as shown:

3.3.3.5. ARI Capable Hierarchy

For Devices associated with an Upstream Port, ARI Capable Hierarchy is a hint to the Device that ARI has been enabled in the Root Port or Switch Downstream Port immediately above the Device. …

…

IMPLEMENTATION NOTE

ARI Capable Hierarchy

For a Device associated with an Upstream Port, that ~~The~~ Device has no way of knowing whether ARI has been enabled in the Downstream Port immediately above it. If ARI is enabled, the Device can conserve Bus Numbers by assigning VFs to Function Numbers greater than 7 on the captured Bus Number. ARI is defined in the *PCI Express Base Specification*.

Since Root Complex Integrated Endpoints are not associated with an Upstream Port, ARI does not apply, and VFs may be assigned to any Function Number within the Root Complex permitted by First VF Offset and VF Stride (see Sections 3.3.9 and 3.3.10).

In the SR-IOV Specification, Section 3.7.2, edit as shown:

3.7.2. Access Control Services (ACS) Extended Capability

ACS is an optional extended capability. If an SR-IOV Capable Device other than one in a Root Complex implements internal peer-to-peer transactions, ACS is required with additional requirements described below.

PF and VF functionality is defined in the PCI Express Base Specification except where noted in Table 3-23.

All Functions in SR-IOV Capable Devices (Devices that implement at least one PF) other than Root Complex Integrated Endpoints that support peer-to-peer transactions within the Device shall implement ACS Egress Control.

Implementation of ACS in Root Complex Integrated Endpoints is permitted but not required. It is explicitly permitted that, within a single Root Complex, some Root Complex Integrated Endpoints implement ACS and some do not. It is strongly recommended that Root Complex implementations ensure that all accesses originating from Root Complex Integrated Endpoints (PFs and VFs) without ACS capability are first subjected to processing by the Translation Agent (TA) in the Root Complex before further decoding and processing. The details of such Root Complex handling are outside the scope of this specification.

Table 3-23: ACS Capability Register

Bit Location	PF and VF Register Differences from Base	PF Attributes	VF Attributes
5	**ACS P2P Egress Control (E)** – Required to be 1b for Functions that are not Root Complex Integrated Endpoints if peer-to-peer transactions within the Device are supported.	**Base**	**Base**

In the PCIe Base Specification, at the end of Section 6.20.2.1. "PASID field", add Impl. Note:

Implementation Note: PASID Width Homogeneity

The PASID value is unique per Function and thus the original intent was that the width of the PASID value supported by that Function could be based on the needs of that Function. However, current system software typically does not follow that model and instead uses the same PASID value in all Functions that access a specific address space. To enable this, system software will typically ensure a common system PASID width for Root Complex and persistent translation agents. Such system software will typically disable ATS on any hot plugged Endpoint Functions or translation agents reporting PASID width support which is less than that of the common system PASID width.

The Root Complex, Endpoints, and translation agents, are often implemented independently of system software, therefore it is highly recommended that hardware implement the maximum width of 20 bits to ensure interoperability with system software.

Endpoints may, in an implementation-specific way, be able to map the 20 bit system PASID to an internal representation carrying a smaller width. If this is done, it is critical that the Endpoint do so without impacting system software, which has no mechanism to differentiate such implementation from those that implement the full 20 bit width natively.

In the PCIe Base Specification, Section 6.12, edit as shown (changes from related errata (released as B12) highlighted in green, new changes highlighted in yellow):

6.12. Access Control Services (ACS)

ACS defines a set of control points within a PCI Express topology to determine whether a TLP should be routed normally, blocked, or redirected. ACS is applicable to RCs, Switches, and multi-Function devices.[89] For ACS requirements, single-Function devices that are SR-IOV capable must be handled as if they were multi-Function devices, since they essentially behave as multi-Function devices after their Virtual Functions (VFs) are enabled.

Implementation of ACS in Root Complex Integrated Endpoints is permitted but not required. It is explicitly permitted that, within a single Root Complex, some Root Complex Integrated Endpoints implement ACS and some do not. It is strongly recommended that Root Complex implementations ensure that all accesses originating from Root Complex Integrated Endpoints (PFs and VFs) without ACS capability are first subjected to processing by the Translation Agent (TA) in the Root Complex before further decoding and processing. The details of such Root Complex handling are outside the scope of this specification.

…

6.12.1.2. ACS Functions in SR-IOV Capable and Multi-Function Devices

This section applies to multi-Function device ACS Functions, with the exception of Downstream Port Functions, which are covered in the preceding section. For ACS requirements, single-Function devices that are SR-IOV capable must be handled as if they were multi-Function devices.

…

- ACS P2P Request Redirect: must be implemented by Functions that support peer-to-peer traffic with other Functions. This includes SR-IOV Virtual Functions (VFs).

  …

  When ACS P2P Request Redirect is enabled in a multi-Function device that is not a Root Complex Integrated Endpoint, peer-to-peer Requests between Functions of the device must be redirected Upstream towards the RC.

  It is permitted but not required to implement ACS P2P Request Redirect in a Root Complex Integrated Endpoint. When ACS P2P Request Redirect is enabled in a Root Complex Integrated Endpoint, peer-to-peer Requests, defined as all Requests that do not target system memory, must be sent to implementation-specific logic within the Root Complex that determines whether the Request is directed towards its original target, or blocked as an ACS Violation error. The algorithms and specific controls for making this determination are not architected by this specification.

  Completions are never affected by ACS P2P Request Redirect.

- ACS P2P Completion Redirect: must be implemented by Functions that implement ACS P2P Request Redirect.

  …

When ACS P2P Completion Redirect is enabled in a ~~multi-Function~~ device that is not a Root Complex Integrated Endpoint, peer-to-peer ~~Read~~ Completions that do not have the Relaxed Ordering bit set must be redirected Upstream towards the RC. Otherwise, peer-to-peer Completions must be routed normally.

Requests are never affected by ACS P2P Completion Redirect.

- ACS Upstream Forwarding: must not be implemented.
- ACS P2P Egress Control: implementation is optional; is based on Function Numbers or Function Group Numbers; controls peer-to-peer Requests between the different Functions within the multi-Function or SR-IOV capable device.

…

Each Function within a multi-Function or SR-IOV capable device that supports …

With ACS P2P Egress Control in multi-Function and SR-IOV capable devices, …

- ACS Direct Translated P2P: must be implemented if the multi-Function or SR-IOV capable device Function …

When ACS Direct Translated P2P is enabled in a multi-Function device or SR-IOV capable Function, …

6.12.1.3. Functions in Single-Function Devices

This section applies to single-Function device Functions, with the exception of Downstream Port Functions and SR-IOV capable Functions, which are covered in ~~a~~ preceding sections. For ACS requirements, single-Function devices that are SR-IOV capable must be handled as if they were multi-Function devices.

…

In the PCIe Base Specification, Section 7.12, edit as shown (changes from related errata (B10) highlighted in green, new changes highlighted in yellow):

7.12. Device Serial Number Capability

The PCI Express Device Serial Number Capability is an optional Extended Capability that may be implemented by any PCI Express device Function. The Device Serial Number is a read-only 64-bit value that is unique for a given PCI Express device. Figure 7-56 details allocation of register fields in the PCI Express Capability structure.

It is permitted but not recommended for Root Complex Integrated Endpoints to implement this Capability.

Root Complex Integrated Endpoints that implement this Capability are permitted but not required to return the same Device Serial Number value as that reported by other Root Complex Integrated Endpoints of the same Root Complex.

All multi-Function Dd~~d~~evices ~~that~~other than Root Complex Integrated Endpoints that implement this Capability must implement it for Function 0; other Functions that implement this Capability must return the same Device Serial Number value as that reported by Function 0.

Root Complex Integrated Endpoints are permitted to implement or not implement this Capability on an individual basis, independent of whether they are part of a multi-Function device.

A PCI Express ~~multi-device~~ component other than a Root Complex containing multiple Devices such as a PCI Express Switch that implements this Capability must return the same Device Serial Number for each device.

In the PCIe Base Specification, Section 7.25, edit/add:

7.25. Latency Tolerance Reporting (LTR) Capability

…

For a multi-Function device associated with the Upstream Port of a component that implements the LTR mechanism, this Capability structure must be implemented only in Function 0, and must control the component's Link behavior on behalf of all the Functions of the ~~d~~Device.

Root Complex Integrated Endpoints implemented as Multi-Function Devices are permitted to implement this Capability structure in more than one Function of the Multi-Function Device.