Single Root I/O Virtualization Configuration

Eric DeHaemer (Intel)

David Kahn (Sun Microsystems)

Disclaimer

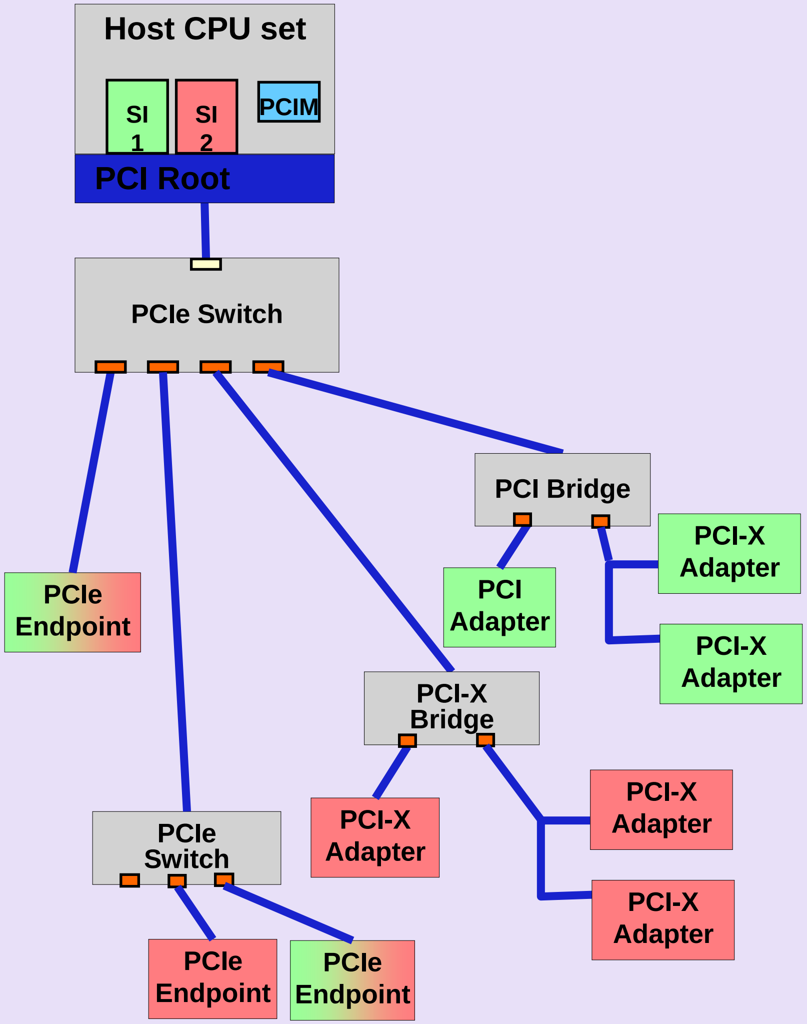
- NOTE: The information in this presentation refers to a specification still in the development process. This presentation reflects the current thinking of the workgroup, but all material is subject to change before the specification is released.

Outline

- Single Root Configuration Space Overview
- SR IOV Extended Capability
- PF/VF Configuration Space – Type 0 Header
- PCI Express® Capability
- PCI Standard Capabilities
- PCI Extended Capabilities
- Single Root IOV Error Handling

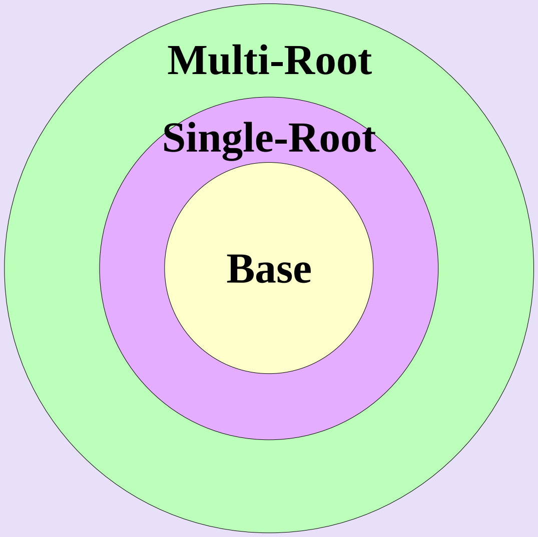
SR Configuration Space Overview

Single Root Overview



- A single Root Complex with multiple System Images sharing SR-IOV aware devices.
- A single root fabric consists of a single set of PCI address spaces (just like PCI Express base)
- A VI is required to manage access to the fabric (permissions, etc.)

Single Root Overview (Cont.)



- SR is built on the PCI Express base protocol.
- SR requires no changes to the root complex or the PCI Express fabric.
- Some implementations may decide to include some optional changes to switches and possibly the root complex to implement SR. (examples: ARI, ATPT). Note: ATPT is not specified or required by any IOV specification.
- Changes to CPU complex to support virtualization. (protection, etc.) Note: CPU changes to support virtualization are not specified by the IOV specifications.

SR Overview – PF/VF

- Physical Function (PF)
  - ✓ A PCI Express function that includes the SR-IOV Capability.
  - ✓ A PF contains the SR-IOV capability for configuration and management of the PF and its associated VFs.
  - ✓ Used by SR PCIM to manage a set of virtual functions.
- Virtual Function (VF)
  - ✓ Simply, a name for a virtual view of the device.
  - ✓ Used by SIs to access resources on the endpoint.
  - ✓ VFs are created/managed by SR-PCIM
  - ✓ Each VF is associated with a single PF
  - ✓ Once created, it can be probed and accessed through the root complex using normal access methods.

SR Overview – SR PCIM

- SR PCI Manager (SR-PCIM)
  - ✓ The entity responsible for configuration and management of an IOV-enabled fabric and devices.
  - ✓ Creates and manages VFs
  - ✓ Handles events that cannot be associated with a single VF/SI.

SR Overview – Role of the VI

- Provides protection between SIs
  - ✓ AKA hypervisor, etc.
  - ✓ Physical resources (memory, devices, privileged registers)
  - ✓ PCI resources (memory, io, config space)
  - ✓ DMA addresses
  - ✓ Routing of messages (Interrupts, etc)
  - ✓ Can be (and usually will be) a combination of software and hardware

SR – Key Config Space Requirements

- SR PCIM must be able to discover PFs and configure them.
  - ✓ SR-IOV Extended Capability
- Each VF must have a unique Routing ID.
  - ✓ Unique configuration space address to discover the VF instance.
  - ✓ Unique Routing ID used in interrupts, messages, R/W requests, etc.
- Compatibility with the PCI Express Base
  - ✓ Retain header layout for type 0 and 1 headers.
  - ✓ No need to implement all bits.
  - ✓ Maintain configuration space read/write semantics. (ordering …)
  - ✓ Maintain routing rules defined by the base spec.
- Minimize bits that must be implemented per VF.
  - ✓ Alias bits where possible.
  - ✓ Implement bits where required.
  - ✓ VI emulation where “alias” or “implement” is not practical.

SR-IOV Extended Capability

SR-IOV Extended Capability

31 24	23 20	19 16	15 0	Byte Offset
Next Capability Offset		Capability Version	PCI Express Extended Capability ID	00h
SR IOV Capabilities				04h
SR IOV Status			SR IOV Control	08h
TotalVFs (RO)			InitialVFs (RO)	0Ch
RsvdP	Function Dependency Link (RO)		NumVFs (RW)	10h
VF Stride (RO)			First VF Offset (RO)	14h
VF Device ID			RsvdP	18h
Supported Page Sizes (RO)				1Ch
System Page Size (RW)				20h
VF BAR0 (RW)				24h
VF BAR1 (RW)				28h
VF BAR2 (RW)				2Ch
VF BAR3 (RW)				30h
VF BAR4 (RW)				34h
VF BAR5 (RW)				38h
VF Migration State Array Offset (RO)				3Ch

SR-IOV Capability: SR-IOV Capabilities

Bit Location	Register Description	Attributes
0	**VF Migration Capable – Migration Capable Device running under Migration Capable MR-PCIM**	RO
20 .. 1	**Reserved – These fields are currently reserved**	RsvdP
31 .. 21	**VF Migration Interrupt Message Number – Indicates the MSI/MSI-X vector used for migration interrupts**	RO

SR-IOV Capability:
SR-IOV Capabilities fields

- VF Migration Capable (RO)
  - ✓ VF Migration is supported in systems that implement MR-IOV.
  - ✓ VF Migration Capable (RO) must be read-only zero if the device is “single root” only.
- VF Migration Interrupt Message Number (RO)
  - ✓ MSI or MSI-X interrupt “number” used for migration events.
  - ✓ Not used if VF Migration Capable is zero.

SR-IOV Capability: SR IOV Control

Bit Location	Register Description	Attributes
0	**VF Enable** – Enables / Disables VFs. Default value is 0b	RW
1	**VF Migration Enable** – Enables / Disables VF Migration Support. Default value is 0b	RW
2	**VF Migration Interrupt Enable** – Enables / Disables VF Migration State Change Interrupt. Default value is 0b.	RW
3	**VF MSE** – Memory Space Enable for Virtual Functions. Default value is 0b.	RW
4	**VF ARI Enable** – Device may locate VFs in Function numbers 8 to 255 of the captured Bus number. Default value is 0b.	RW
15..5	**Reserved** – These fields are currently reserved	RsvdP

SR-IOV Capability:
SR IOV Control fields

- VF Enable (RW)
  - ✓ NumVFs VFs exist (are created) when VF Enable is Set
  - ✓ If VF Enable is Clear, VFs do not exist.
- VF Migration Enable (RW)
  - ✓ Migration not permitted if this field is zero.
  - ✓ May be hardwired zero on Devices that are SR-only or don’t support MR migration features.
  - ✓ Allows software to override migration capability.
- VF Migration Interrupt Enable (RW)
  - ✓ Enables use of the VF Migration Interrupt for migration events.
- VF MSE (RW)
  - ✓ Memory space enable bit for all VFs
- VF ARI Enable (RW)
  - ✓ Set by software if ARI forwarding is enabled in the switch/root port above this Device. The PF may use this value to determine optimal settings for First VF Offset and VF Stride.

SR-IOV Capability: SR-IOV Status

Bit Location	Register Description	Attributes
0	**VF Migration Interrupt Pending** – Indicates a VF Migration In or Migration Out Request has been issued by MR-PCIM. Details are available through scanning the VF State Array.	RW1C
15..1	**Reserved** – These fields are currently reserved	RsvdZ

SR-IOV Capability: Number of VFs fields

- InitialVFs (RO)
  - ✓ Maximum number of “allocated” VFs associated with this PF.
- TotalVFs (RO)
  - ✓ Total number of VFs that could be associated with this PF
  - ✓ Describes additional “VF slots” that may or may not be backed by resources.
  - ✓ Used with migration only. If Migration Capable and Enable are set:
    - TotalVFs must be >= MaxVFs
- NumVFs (RW)
  - ✓ Describes the number of VFs actually in use.
  - ✓ Written by SR-PCIM prior to setting VF Enable to 1.

SR-IOV Capability: First VF Offset and VF Stride

- First VF Offset (RO)
  - ✓ RID offset (from the PF's RID) of the first VF.
  - ✓ May change if NumVFs and/or VF ARI Enable are changed (but before VF Enable is Set).
- VF Stride (RO)
  - ✓ RID offset to subsequent VFs
  - ✓ Algorithm to determine the RID of VF_n
    - RID_{PF} + First VF Offset + ((n-1) * (VF Stride))
    - VF's are numbered 1 .. n
    - All arithmetic is unsigned 16-bit ignoring any carry (modulo 2^{16})
- Use these fields to determine the number of buses that the Device "needs" when VF Enable is Set. (When programming the downstream switch or root ports bus number ranges fields.)
  - ✓ Configure All PFs, Setting NumVFs and VF ARI Enable if applicable.
  - ✓ Calculate max bus number from all PFs in any given Device.

SR-IOV Capability: Function Dependency Link

- 8-bit Function number of dependent PF (linked list).
- Contains the function number of this PF, if no dependencies or if the last dependent function in a dependency list.
- Describes a linked list of PFs that should have their VFs allocated together.
- Function dependencies are vendor specific.
- Example: A multi-function Device with a network PF plus a crypto PF, implemented as separate functions, but the crypto function can be used to accelerate the network function in a vendor specific manner.

SR-IOV Capability: VF Device ID

- In VF config space, the Device ID and Vendor ID fields are RO and return the value FFFFh when read.
  - ✓ Legacy, non IOV-aware probing software may “ignore” configured VFs when they “see” FFFFh in these fields.
  - ✓ The VI can return the proper values for these fields when read, if applicable to a system vendors implementation.
- VF Device ID field contains the actual Device ID of all VFs associated with this PF.
  - ✓ All VFs associated with a PF use the same Device ID value.
- VF Vendor ID is the same as the PFs Vendor ID value.

SR-IOV Capability:
Page Size Related Fields

- System Page Sizes (RO) and Supported Page Size (RW)
  - ✓ Allows software to specify a system page size alignment for each VF BARx
- Supported Page Sizes (RO)
  - ✓ Bitmask of supported “page sizes”
  - ✓ If bit *n* is set, $2^{(n+12)}$ page size is supported
  - ✓ Devices must support 4k, 8k, 64k, 256k, 1M and 4M page sizes.
  - ✓ Support for other page sizes is optional.
- System Page Size (RW)
  - ✓ Same encoding as Supported Page Sizes
  - ✓ Affects VF BARx “size” and “alignment”
    - – Each VF BARx will be aligned on a “system page size” boundary
  - ✓ Set this field before setting VF Enable and before sizing VF BARs
  - ✓ Results are undefined if more than 1 bit is set in System Page Size.
  - ✓ Results are undefined if a bit is Set that is not Set in Supported Page Sizes

SR-IOV Capability: VF BARx

- Base Address registers for all VFs
  - ✓ One set of decoders per PF for all its VFs.
  - ✓ Size and alignment are for a single VF instance
    - Use standard BAR sizing algorithm described in *PCI Local Bus Spec 3.0*
  - ✓ Set System Page Size prior to using the BAR sizing algorithm
    - System Page Size requirements affect VF BARx alignment
  - ✓ After NumVFs, VF Enable and VF MSE are Set
    - Each VF BARx decodes *NumVFs* address spaces.
    - Actual address space decoded per VF BARx:
      - NumVFs * (probed BARx size)
  - ✓ Each VF's BARx is aligned on a System Page Size boundary
    - Permits software to use separate MMU mappings for each VF for each BARx

SR-IOV Capability: VF Migration State Array Offset

Bit Location	Register Description	Attributes
31..3	**VF Migration State Offset** – Used as an offset from the address contained by one of the function's Base Address registers to point to the base of the VF Migration State Array. The lower 3 MVF Migration State BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.	RO
2..0	**VF Migration State BIR** – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the function's VF Migration State Array into Memory Space. **BIR Value Base Address register** 0 BAR0 10h 1 BAR1 14h 2 BAR2 18h 3 BAR3 1Ch 4 BAR4 20h 5 BAR5 24h 6 Reserved 7 Reserved For a 64-bit Base Address register, the VF Migration State BIR indicates the lower DWORD.	RO

SR-IOV Capability: VF Migration State Array

Bit Location	Register Description	Attributes
1..0	**VF Migration State** – State of the associated VF	RW
7..2	**Reserved** – These fields are currently reserved	RsvdP

VF State	VF Exists	Description
00b	No	**Inactive.Unavailable** – VF does not exist to SR nor is it being migrated in or out.
01b	No	**Dormant.MigrateIn** – VF is available for use by SR. VF exists but can not initiate transactions.
10b	Yes	**Active.MigrateOut** – SR has been requested to relinquish use of the VF.
11b	Yes	**Active.Available** – Fully functional. Could be assigned to an SI.

SR-IOV Capability:
VF Migration State Array: Transitions

Current State	New State	Change Initiated By	SR Visible Effects of Change
Active.Available	Active.MigrateOut	MR-PCIM	**VF Migrate Out Request** VF continues to exist. Sets VF Migration Status.
Inactive.Unavailable	Dormant.MigrateIn	MR-PCIM	**VF Migrate In Request** VF remains non-existent. Sets VF Migration Status.
Dormant.MigrateIn	Inactive.Unavailable	MR-PCIM	**VF Migrate In Retract** VF remains non-existent. Sets VF Migration Status.
Active.MigrateOut	Active.Available	MR-PCIM	**VF Migrate Out Retract** VF continues to exist. Sets VF Migration Status.

PF/VF Configuration Space: Type 0 Header

Configuration Space: Key

Register Attribute	Description
LB 3.0	Attribute is same as specified in PCI Local Bus Specification 3.0.
Base	Attribute is same as specified in PCI Express Base Specification, Revision 1.1
HwInit	Hardware Initialized: Register bits are initialized by firmware or hardware mechanisms …
RO	Read-only register: Register bits are read-only and cannot be altered by software. …
RW	Read-Write register: Register bits are read-write and may be either set or cleared by software to the desired state.
RW1C	Read-only status, Write-1-to-clear status register: Register bits indicate status when read …
ROS	Sticky - Read-only register: Registers are read-only and cannot be altered by software. …
RWS	Sticky - Read-Write register: Registers are read-write and may be either set or cleared …
RW1CS	Sticky - Read-only status, Write-1-to-clear status register: Registers indicate status …
RsvdP	Reserved and Preserved: Reserved for future RW implementations …
RsvdZ	Reserved and Zero: Reserved for future RW1C implementations; …

NB: Any field/register not shown has the same definition as the Base specification.

Type 0 Header fields (1)

Field Name	PF	VF
Vendor ID	Base	RO FFFFh
Device ID	Base	RO FFFFh
Command Register	Base	***
Status Register	Base	***
Class Code	Base	Base Same value in each VF:PF
Revision ID	Base	Base Same value in each VF:PF
Cacheline Size	Base	RO 00h
Latency Timer	Base	RO 00h
Header Type	Base	RO 00h
BIST	Base	RO 00h

Type 0 Header fields (2)

Field Name	PF	VF
Base Address Registers	Base	***
Cardbus CIS Pointer	Base	RO 00h
Subsystem Vendor ID	Base	Base Same value in each VF:PF
Subsystem Device ID	Base	Base Same value in each VF:PF
Expansion ROM BAR	Base	***
Capabilities Pointer	Base	Base
Interrupt Line	Base	RO 00h
Interrupt Pin	Base	RO 00h
Min_Gnt	Base	RO 00h
Max_Lat	Base	RO 00h

Command Register

Bit Location	PF and VF Register Differences from Base Specification	PF Attributes	VF Attributes
0	**I/O Space Enable –** VF: Hardwire 0.	Base	RO 0b
1	**Memory Space Enable –** VF MSE controls VFs	Base	RO 0b
2	**Bus Master Enable**	Base	Base
6	**Parity Error Enable** – See Error section	Base	RsvdP
8	**SERR Enable –** See Error section	Base	RsvdP
10	**Interrupt Disable –** VF: Hardwire zero	Base	RO 0b

Status Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
3	**Interrupt Status –** Does not apply to VFs. Must be hardwired to 0 for VFs.	Base	RO 0b

VF Base Address Registers

- VF Base Address registers are implemented in the SR-IOV Capability in the PF.
- The VI may provide emulation for VF BAR reads, if required by system software.

Expansion ROM BAR

- Expansion ROM BAR
  - ✓ Not applicable to VFs
  - ✓ Emulate using PFs expansion ROM BAR
  - ✓ Shared ROM BAR decoding is not permitted

PCI Express Capability

Device Capabilities Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
4:3	**Phantom Functions Supported** – Unsupported with VFs	Base	00b
25:18	**Captured Slot Power Limit Value**	Base	00b
27:26	**Captured Slot Power Limit Scale**	Base	00b
28	**Function Level Reset Capability –** Required for SR-IOV devices (PFs and VFs). Must be hardwired to 1.	1b	1b

Device Control Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
0	**Correctable Error Reporting Enable**	Base	RsvdP
1	**Non-Fatal Error Reporting Enable**	Base	RsvdP
2	**Fatal Error Reporting Enable –** PF bit setting applies to all associated VFs as well.	Base	RsvdP
3	**Unsupported Request Reporting Enable –** PF bit setting applies to all associated VFs as well.	Base	RsvdP
4	**Enable Relaxed Ordering –** PF bit setting applies to all associated VFs as well.	Base	RsvdP
7:5	**Max_Payload_Size –** PF bit setting applies to all associated VFs as well.	Base	RsvdP
8	**Extended Tag Field Enable –** PF bit setting applies to all associated VFs as well.	Base	RsvdP
9	**Phantom Functions Enable** – If SR-IOV is enabled, this bit is hardwired to 0.	Base	RsvdP
10	**Auxiliary (AUX) Power PM Enable**	Base	RsvdP
11	**Enable No Snoop –** PF bit setting applies to all associated VFs as well.	Base	RsvdP
14:12	**Max_Read_Request_Size** – PF bit setting applies to all associated VFs as well.	Base	RsvdP
15	**Initiate Function Level Reset –** Required for PFs and VFs	Base	Base

Device Status Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
4	**AUX Power Detected**	Base	RO 0b

Link Control Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
1:0	**Active State Power Management (ASPM) Control**	Base	RsvdP
3	**Read Completion Boundary (RCB)**	Base	RsvdP
6	**Common Clock Configuration**	Base	RsvdP
7	**Extended Synch**	Base	RsvdP
8	**Enable Clock Power Management**	Base	RsvdP
9	**Hardware Autonomous Width Disable**	Base	RsvdP

Link Status Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
3:0	**Current Link Speed**	Base	Rsvd
9:4	**Negotiated Link Width**	Base	Rsvd
10	**Undefined** – The value read from this bit is undefined in Base 1.1 (was previously Training Error).	Base	Rsvd
11	**Link Training** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base	Rsvd
12	**Slot Clock Configuration**	Base	Rsvd
13	**Data Link Layer Link Active**	Base	Rsvd
14	**Link Bandwidth Management Status** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base	Rsvd
15	**Link Autonomous Bandwidth Status** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base	Rsvd

Device Control 2 Register

Bit Location	PF and VF Register Differences from Base 2.0	PF Attributes	VF Attributes
3:0	**Completion Timeout Value**	Base	RsvdP
4	**Completion Timeout Disable**	Base	RsvdP

Link Status 2 Register

Bit Location	PF and VF Register Differences from Base 2.0	PF Attributes	VF Attributes
0	**Current De-emphasis Level**	Base	RsvdP

PCI Standard Capabilities

PCI Standard Capabilities

Capability Name	PF	VF
PCI Power Management	Base (required)	Base (optional)
PCI Hot Plug	Base	N/A
VPD	Base	Base **
Slot ID	Base	N/A
MSI	Base	Base
MSI-X	Base	Base

PCI Express Extended Capabilities

PCI Express Extended Capabilities

Capability Name	PF	VF
AER	Base	** See Error Section
VC (02h and 09h)	Base	N/A
Device Serial No.	Base	N/A
Power Budgeting	Base	N/A
MFVC	Base	N/A
ACS	Base **	Base **
ARI	Base (Required)	Base ** (Required)
ATS	Base	Base **
SR-IOV	Base	N/A
MR-IOV	N/A	N/A

SR IOV Error Handling

Key Error Reporting Requirements

- VI owns the first response to error messages.
  - ✓ Error messages sent to Root Port
  - ✓ VI can triage errors before sending to SI
- Error Control Bits are only located in the PF
  - ✓ Includes control, mask, and severity bits
  - ✓ VF uses the controls in associated PF when making decisions
- Function Specific Error Status Bits are present in VFs
  - ✓ Independent error status for logging Function Specific Errors
  - ✓ Poison TLP, Completer Timeout, CA, UR, …
- Non-Function Specific Errors are logged in the PF
  - ✓ Physical Layer, Link Layer, Malformed, ECRC, …

Uncorrectable Error Status

Bit Location	PF and VF Register Differences from Base	PF Attributes	VF Attributes
4	**Data Link Protocol Error Status**	Base	RsvdP
5	**Surprise Down Error Status**	Base	RsvdP
12	**Poisoned TLP Status**	Base	Base
13	**Flow Control Protocol Error Status**	Base	RsvdP
14	**Completion Timeout Status**	Base	Base
15	**Completer Abort Status**	Base	Base
16	**Unexpected Completion Status**	Base	Base
17	**Receiver Overflow Status**	Base	RsvdP
18	**Malformed TLP Status**	Base	RsvdP
19	**ECRC Error Status**	Base	RsvdP
20	**Unsupported Request Status**	Base	Base
21	**ACS Violation**	Base	Base

Correctable Error Status

Bit Location	PF and VF Register Differences from Base	PF Attributes	VF Attributes
0	**Receiver Error Status**	Base	RsvdP
6	**Bad TLP Status**	Base	RsvdP
7	**Bad DLLP Status**	Base	RsvdP
8	**REPLAY_NUM Rollover Status**	Base	RsvdP
12	**Replay Timer Timeout Status**	Base	RsvdP
13	**Advisory Non-Fatal Error Status**	Base	Base

Error Mask and Severity

Uncorrectable Mask Register

Uncorrectable Severity Register

Correctable Mask Register

- ✓ Only meaningful in the PF
- ✓ VF versions are all RsvdP
- ✓ VF uses values in associated PF when making error logging/signaling decisions.

Advanced Error Capabilities and Control Register

Bit Location	PF and VF Register Differences from Base	PF Attributes	VF Attributes
4:0	**First Error Pointer**	Base	Base
5	**ECRC Generation Capable**	Base	Base
6	**ECRC Generation Enable**	Base	RsvdP
7	**ECRC Check Capable**	Base	Base
8	**ECRC Check Enable**	Base	RsvdP

ADVERR Header Log

- Mechanism defined to allow sharing of header logs across VFs
  - ✓ All associated VFs must implement at least 2 logs
  - ✓ PF operates under Base rules and has independent header log.
- Header Log is locked error is serviced
  - ✓ Bit set in Uncorrectable Error Status
  - ✓ First Error Pointer updated in AdvErr Capabilites and Control
  - ✓ Header is logged and the entry in shared logs is locked
  - ✓ Header entry is freed when corresponding bit in Uncorrectable Error Status is cleared
- A function may not have room to log a header
  - ✓ Function shall update Error Status registers
  - ✓ Function will return all 1's when the Header Log is read to indicate 'overflow' condition

Questions

Thank you for attending the PCI-SIG Developers Conference 2007

For more information please go to www.pcisig.com