
PCI
SIG®

Single Root IOV Configuration

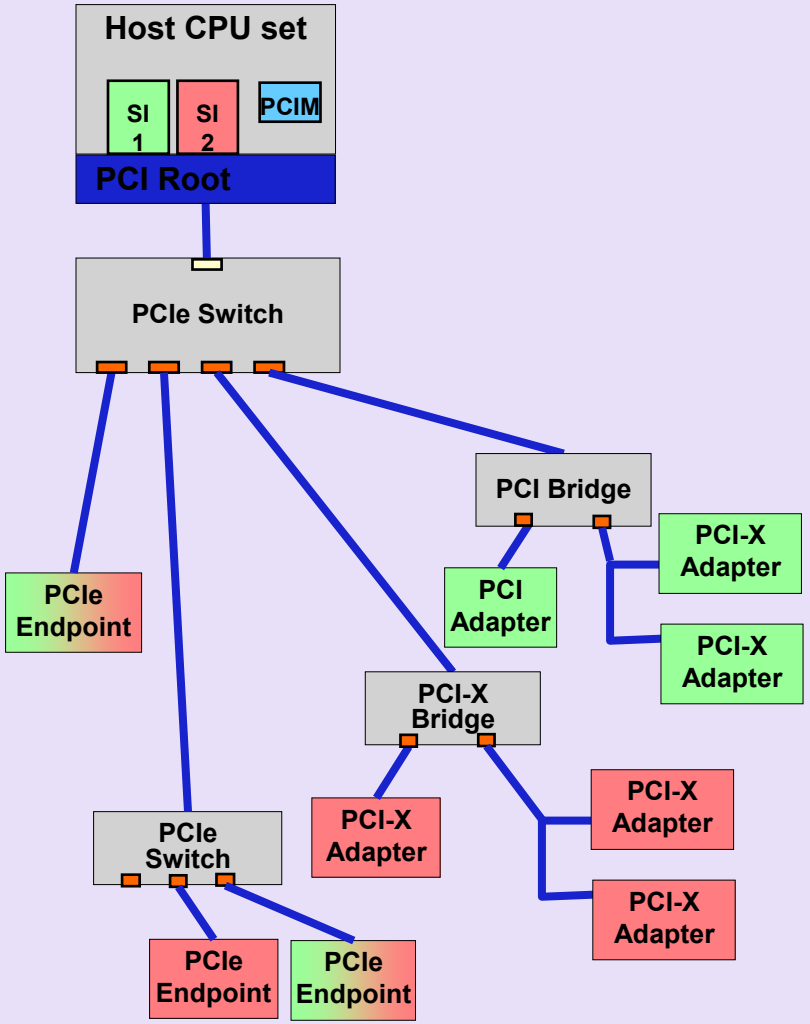
David Kahn (Sun Microsystems)

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

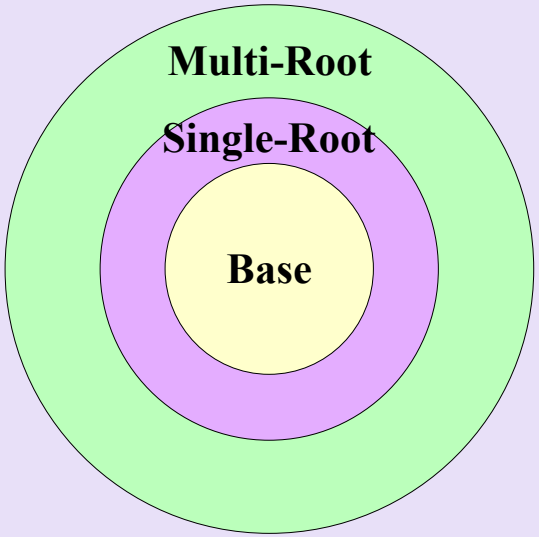
SR Configuration Space Overview

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.
  - ✓ An SR iov-aware PF contains IOV capabilities for configuration and management of the PF.
  - ✓ 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 Capabilty

SR-IOV Extended Capability

31 20	19 16	15 0	Byte Offset
Next Capability Pointer	Cap Version	Capability ID	00h
SR IOV Capabilities			04h
SR IOV Status		SR IOV Control	08h
TotalVFs (RO)		MaxVFs (RO)	0Ch
ReservedZ		NumVFs (RWS)	10h
VF Stride (RO)		First VF Offset (RO)	14h
Supported Page Sizes (RO)			18h
System Page Size (RWS)			1Ch
VF BAR0 (RW)			20h
VF BAR1 (RW)			24h
VF BAR2 (RW)			28h
VF BAR3 (RW)			2Ch
VF BAR4 (RW)			30h
VF BAR5 (RW)			34h
VF Migration State Array Offset (RO)			38h

SR-IOV Capabilities Register

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**	ReservedZ
31 .. 21	**VF Migration Interrupt Message Number – Indicates the MSI/MSI-X vector used for the interrupts**	RO

SR IOV Capabilities Register 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 Control Register

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
15..4	**Reserved** – These fields are currently reserved	ReservedZ

SR-IOV Control Register fields

- VF Enable (RW)
  - ✓ NumVFs VFs exist when VF Enable is Set
  - ✓ If VF Enable is reset to zero, 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

SR-IOV Status Register

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	ReservedZ

SR-IOV Capability: Number of VFs fields

- MaxVFs (RO)
  - ✓ Maximum number of 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 (RWS)
  - ✓ 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.
  - ✓ Valid only when VF Enable is Set, otherwise undefined.
- 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})

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 “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

- BAR-relative offset (in memory space) of the VF Migration State Array
- Field is undefined if either VF Migration Capable is zero or VF Migration Enable is zero.
- Low order 3 bits define the BAR (as with MSI-X, etc)
  - ✓ 0 = BAR0, 1=BAR1, … 5=BAR5.
- Upper 29 bits define the upper 29 bits of the BAR-relative offset. The low order 3 bits of the actual offset is zero.

SR-IOV Capability: VF Migration State Array

VF State	VF Active	Description
1	No	**Inactive VF** – MR-PCIM may map or unmap resources to the VF while in this state. Resource mappings may not be changed in any other state. Any mapping / unmapping of resources is invisible to SR software.
2	No	**VF being Migrated In** – after mapping resources to the VF (if needed), MR-PCIM transitions the VF to state 2 to initiate the Migrate In operation. This transition causes a VF Migration Event, allowing SR software to detect the state change and either accept or refuse the Migrate In.
4	Yes	**Active VF / PF** – Fully functional.
8	Yes	**VF being Migrated Out** – MR-PCIM transitions the VF to state 8 to request a Migrate Out operation. SR software either accepts the Migrate Out operation (by transitioning to state 1), or refuses the Migrate Out (by transitioning back to state 4). Once accepted, MR-PCIM may then unmap resources from the VF.

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 1.1	Attribute is same as specified in PCI Express Base Specification, Revision 1.1
Base 2.0	Attribute is same as specified in PCI Express Base Specification, Revision 2.0.
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. …
ROP	**Read-only register that must match the value of the PF field of the same name. Used in Virtual Functions.**
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; …

Type 0 Header fields (1)

Field Name	PF	VF
Vendor ID	Base 1.1	ROP ***
Device ID	Base 1.1	ROP ***
Command Register	Base 1.1	***
Status Register	Base 1.1	***
Class Code	Base 1.1	ROP
Revision ID	Base 1.1	ROP ***
Cacheline Size	Base 1.1	ROP
Latency Timer	Base 1.1	ROP
Header Type	Base 1.1	ROP
BIST	Base 1.1	n/a

Type 0 Header fields (2)

Field Name	PF	VF
Base Address Registers	Base 1.1	***
Cardbus CIS Pointer	Base 1.1	n/a
Subsystem Vendor ID	Base 1.1	ROP ***
Subsystem Device ID	Base 1.1	ROP ***
Expansion ROM BAR	Base 1.1	***
Capabilities Pointer	Base 1.1	Base 1.1
Interrupt Line	Base 1.1	ROP
Interrupt Pin	Base 1.1	ROP
Min_Gnt	Base 1.1	ROP
Max_Lat	Base 1.1	ROP

Command Register

Bit Location	PF and VF Register Differences from Base Specification	PF Attributes	VF Attributes
0	**I/O Space Enable –** VF: Hardwire 0.	LB3.0	RO
1	**Memory Space Enable –** VF MSE controls VFs	Base 1.1	ROP
2	**Bus Master Enable**	Base 1.1	Base 1.1
3	**Special Cycle Enable –** n/a 0	Base 1.1	Base 1.1
4	**Memory Write and Invalidate –** n/a 0	Base 1.1	Base 1.1
5	**VGA Palette Snoop –** n/a 0	Base 1.1	Base 1.1
6	**Parity Error Enable** – See Error section	Base 1.1	ROP
7	**IDSEL Stepping/Wait Cycle Control –** n/a 0	Base 1.1	Base 1.1
8	**SERR Enable –** See Error section	Base 1.1	ROP
9	**Fast Back-to-Back Transactions Enable –** n/a	Base 1.1	Base 1.1
10	**Interrupt Disable –** VF: Hardwire zero	Base 1.1	RO

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 1.1	RO
4	**Capabilities List –** Must be hardwired to 1	Base 1.1	Base 1.1
5	**66 MHz Capable –** n/a 0	Base 1.1	Base 1.1
7	**Fast Back-to-Back Transactions Capable –** n/a	Base 1.1	Base 1.1
8	**Master Data Parity Error –** See Error section	Base 1.1	ROP
10:9	**DEVSEL Timing –** n/a 0	Base 1.1	Base 1.1
11	**Signaled Target Abort –** See Error Section	Base 1.1	ROP
12	**Received Target Abort –** See Error Section	Base 1.1	ROP
13	**Received Master Abort –** See Error Section	Base 1.1	ROP
14	**Signaled System Error –** See Error Section	Base 1.1	ROP
15	**Detected Parity Error –** See Error Section	Base 1.1	ROP

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

PCI Express Capabilities Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
3:0	**Capability Version**	Base 1.1	Base 1.1
7:4	**Device/Port Type** – SR-IOV capable devices must indicate a Device Type of either 0000b (PCI Express Endpoint Device) or 1001b (Root Complex Integrated Endpoint Device).	Base 1.1	ROP
8	**Slot Implemented** – Does not apply to PFs or VFs. Must be hardwired to 0.	Base 1.1	Base 1.1
13:9	**Interrupt Message Number**	Base 1.1	Base 1.1
14	**TCS Routing Supported** Not applicable to end point devices. Must be hardwired to 0. Note: This field is added in the Base 2.0 specification.	Base 2.0	Base 2.0

Device Capabilities Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
2:0	**Max_Payload_Size Supported**	Base 1.1	ROP
4:3	**Phantom Functions Supported** – Unsupported with IOV	Base 1.1	RO
5	**Extended Tag Field Supported**	Base 1.1	ROP
8:6	**Endpoint L0s Acceptable Latency**	Base 1.1	ROP
11:9	**Endpoint L1 Acceptable Latency**	Base 1.1	ROP
12	**Undefined** – (was previously Attention Button Present).	Base 1.1	Base 1.1
13	**Undefined** – (was previously Attention Indicator Present).	Base 1.1	Base 1.1
14	**Undefined** – (was previously Power Indicator Present).	Base 1.1	Base 1.1
15	**Role-Based Error Reporting – Required to be Set**.	Base 1.1	Base 1.1
25:18	**Captured Slot Power Limit Value**	Base 1.1	ROP
27:26	**Captured Slot Power Limit Scale**	Base 1.1	ROP
28	**Function Level Reset Capability –** Required for SR-IOV devices (PFs and VFs). Must be hardwired to 1.	Base 2.0	Base 2.0

Device Control Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
0	**Correctable Error Reporting Enable**	Base 1.1	ROP
1	**Non-Fatal Error Reporting Enable**	Base 1.1	ROP
2	**Fatal Error Reporting Enable –** PF bit setting applies to all associated VFs as well.	Base 1.1	ROP
3	**Unsupported Request Reporting Enable –** PF bit setting applies to all associated VFs as well.	Base 1.1	ROP
4	**Enable Relaxed Ordering –** PF bit setting applies to all associated VFs as well.	Base 1.1	ROP
7:5	**Max_Payload_Size –** PF bit setting applies to all associated VFs as well.	Base 1.1	ROP
8	**Extended Tag Field Enable –** PF bit setting applies to all associated VFs as well.	Base 1.1	ROP
9	**Phantom Functions Enable** – If SR-IOV is enabled, this bit is hardwired to 0.	Base 1.1	RO
10	**Auxiliary (AUX) Power PM Enable**	Base 1.1	ROP
11	**Enable No Snoop –** PF bit setting applies to all associated VFs as well.	Base 1.1	ROP
14:12	**Max_Read_Request_Size** – PF bit setting applies to all associated VFs as well.	Base 1.1	ROP
15	**Initiate Function Level Reset –** Required for PFs and VFs	Base 2.0	Base 2.0

Device Status Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
0	**Correctable Error Detected**	Base 1.1	Base 1.1
1	**Non-Fatal Error Detected**	Base 1.1	Base 1.1
2	**Fatal Error Detected**	Base 1.1	Base 1.1
3	**Unsupported Request Detected**	Base 1.1	Base 1.1
4	**AUX Power Detected**	Base 1.1	ROP
5	**Transactions Pending** – When set indicates that a particular function (PF or VF) has issued Non-Posted Requests which have not been completed. A function reports this bit cleared only when all Completions for any outstanding Non-Posted Requests have been received.	Base 1.1	Base 1.1

Link Capabilities Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
3:0	**Supported Link Speeds**	Base 1.1	ROP
9:4	**Maximum Link Width**	Base 1.1	ROP
11:10	**Active State Power Management (ASPM) Support**	Base 1.1	ROP
14:12	**L0s Exit Latency**	Base 1.1	ROP
17:15	**L1 Exit Latency**	Base 1.1	ROP
18	**Clock Power Management**	Base 1.1	ROP
19	**Surprise Down Error Reporting Capable –** For Upstream Ports, this bit must be hardwired to 0b.	Base 1.1	Base 1.1
20	**Data Link Layer Link Active Reporting Capable –** For Upstream Ports, this bit must be hardwired to 0b.	Base 1.1	Base 1.1
21	**Link Bandwidth Notification Capability** – Not applicable to Endpoint devices. Must be hardwired to 0b.	Base 2.0	Base 2.0
31:24	**Port Number**	Base 1.1	ROP

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 1.1	ROP
3	**Read Completion Boundary (RCB)**	Base 1.1	ROP
4	**Link Disable** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base 1.1	Base 1.1
5	**Retrain Link** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base 1.1	Base 1.1
6	**Common Clock Configuration**	Base 1.1	ROP
7	**Extended Synch**	Base 1.1	ROP
8	**Enable Clock Power Management**	Base 1.1	ROP
9	**Hardware Autonomous Width Disable**	Base 2.0	ROP
10	**Link Bandwidth Management Interrupt Enable** – Not applicable to Endpoint devices. Must be hardwired to 0b.	Base 2.0	Base 2.0
11	**Link Autonomous Bandwidth Interrupt Enable** – Not applicable to Endpoint devices. Must be hardwired to 0b.	Base 2.0	Base 2.0

Link Status Register

Bit Location	PF and VF Register Differences from Base 1.1	PF Attributes	VF Attributes
3:0	**Current Link Speed**	Base 1.1	ROP
9:4	**Negotiated Link Width**	Base 1.1	ROP
10	**Undefined** – The value read from this bit is undefined in Base 1.1 (was previously Training Error).	Base 1.1	Base 1.1
11	**Link Training** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base 1.1	Base 1.1
12	**Slot Clock Configuration**	Base 1.1	ROP
13	**Data Link Layer Link Active**	Base 1.1	ROP
14	**Link Bandwidth Management Status** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base 2.0	Base 2.0
15	**Link Autonomous Bandwidth Status** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base 2.0	Base 2.0

Device Capabilities 2 Register

Bit Location	PF and VF Register Differences from Base 2.0	PF Attributes	VF Attributes
3:0	**Completion Timeout Ranges Supported**	Base 2.0	ROP
4	**Completion Timeout Disable Supported**	Base 2.0	ROP

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 2.0	ROP
4	**Completion Timeout Disable**	Base 2.0	ROP

Link Control 2 Register

Bit Location	PF and VF Register Differences from Base 2.0	PF Attributes	VF Attributes
3:0	**Target Link Speed**	Base 2.0	ROP
4	**Enter Compliance**	Base 2.0	ROP
5	**Hardware Autonomous Speed Disable**	Base 2.0	ROP
6	**Selectable De-emphasis** – Reserved for Endpoint devices. Must be hardwired to 0b.	Base 2.0	Base 2.0
9:7	**Transmit Margin**	Base 2.0	ROP
10	**Enter Modified Compliance**	Base 2.0	ROP

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 2.0	ROP

PCI Standard Capabilities

PCI Standard Capabilities

Capability Name	PF	VF
PCI Power Management	Base 1.1 (required)	Base 1.1 (optional)
PCI Hot Plug	Base 1.1	N/A
VPD	Base 1.1	Base 1.1
Slot ID	Base 1.1	N/A
MSI	Base 1.1	Base 1.1
MSI-X	Base 1.1	Base 1.1
Vendor Specific	Base 1.1	Base 1.1

PCI Express Extended Capabilities

PCI Express Extended Capabilities

Capability Name	PF	VF
AER	Base 1.1	See Error Section
VC (02h and 09h)	Base 1.1	N/A
Device Serial No.	Base 1.1	N/A
Power Budgeting	Base 1.1	N/A
MFVC	Base 1.1	N/A
Vendor Specific	Base 1.1	Optional per vendor
CAC	Base 1.1	N/A
ACS	Base 1.1	Base 1.1
ARI (Required if > 8 total functions)	Base 1.1	Base 1.1
ATS	Base 1.1	Base 1.1

Questions


PCI
SIG®