Overview of Changes to PCI Express^{TM} Specification 1.1

We write the Books!

Ravi Budruk
MindShare, Inc.
ravi@mindshare.com

June 2005

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designators appear in this document, and MindShare was aware of the trademark claims, the designations have been printed in initial capital letters or all capital letters.

The author has taken care in preparation of this document, but makes no expressed or implied warranty of any kind and assumes no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

First Printing, June 2005

Find MindShare on the World-Wide Web at:
http://www.mindshare.com

For technical training on the subject of this document, contact MindShare at 1-800-633-1440 or send email to info@mindshare.com

Introduction

PCI Express Specification 1.1 is the latest release as of June 2005. This spec release is available through the PCI Special Interest Group (PCISIG) or from the website: www.pcisig.com. The 1.1 spec was released in March 2005 and has been the major spec release since the spec revision 1.0a was released in April 2003.

This document details the differences between the PCI Express spec 1.1 and 1.0a. Both *Major* changes as well as spec *Clarifications* have been documented. The author has documented these changes in sections that align to Chapters of MindShare's *PCI Express System Architecture* textbook. This textbook can be obtained from www.mindshare.com/store/

Summary of Changes

This next section outlines major changes to the 1.1 spec. For details on each bullet, proceed to the next section. Spec changes are:

- MSI-X interrupt generation capability added.
- A Root complex that supports peer-to-peer transactions is allowed to split packets.
- Addition of Root Complex Integrated Endpoint and Event Collector definition as well as related extended capability registers.
- Additional Byte Enable usage rule.
- Tag[7:0] field restricted use in Vendor-Defined Messages.
- Hot-Plug related Attention Indicator On/Off/Blink, Power Indicator On/Off/Blink and Attention Button Pressed messages are no longer supported.
- Endpoints must not delay the acceptance of a Posted Request for more than the Posted Request Acceptance Limit of 10 µs.
- New optional feature called CRS Software Visibility which is enabled via a new configuration bit in the Root Control register of Root Complex's PCI Express Capability Block.
- Additional transaction ordering rules apply for traffic within the same traffic class.
- A new Multi-Function Virtual Channel (MFVC) Capability structure is added and optionally implemented in upstream ports.
- The timeout flow control mechanism must be disabled if an Infinite Credit advertisement has been made during initialization for P, NP and Cpl flow control classes.
- InitFC1 and InitFC2 DLLPs must be transmitted in the order InitFC for P, NP, then Cpl.

- The three types of InitFC1 or InitFC2 DLLPs must be transmitted at least once every 34 µs.
- While in InitFC2 stage, transaction layer must block transmission of any other new TLPs.
- If a received TLP ends with EDB, but the TLP does not have an LCRC that is the inverse of what it should be, then discard the TLP and free up any TLP resources as well as if the NAK_SCHEDULED flag is clear, set it and immediately schedule a Nak DLLP for transmission.
- Detection of Link Training errors which was optional is removed from the specification. Link Training register also removed.
- It is required to detect 8b/10b decode errors and report them as Physical Layer receiver errors.
- It is possible to design a port in which the Lanes support different data rates as indicated via the Link Data Rate Identifier within the TS1/TS2 packets.
- Detection of crosslink during training.
- During Configuration state while in the step required for initializing the Lane number, the upstream port may wait up to 1ms before accepting the Lane number indicated in the received TS1s.
- Loopback Slave behavior.
- During Detect state, it is not required that the detect sequence be performed on both conductors of a differential pair.
- The definition of the “Extended Sync” bit (bit 7 of the Link Control register) has been changed.
- A new “Eye measurement Clock Recovery Function” is described.
- The new capability defined allows a device to transition to L2 or L3 via L2/L3 Ready without software first placing a device into the D3 state first.
- A new Link power state called Link Down (LDn) is added.
- Reference Clock (REFCLK) can be gated while a device’s Link is in the L1 or L2/L3_Ready state.
- Configuration registers relating to turning REFCLK On/Off are added.
- Configuration and Message requests are the only TLPs accepted by a function in the D1 or D2 or $D3_{Hot}$ device power state.
- Link L1 power management state entry and exit procedure is further clarified.
- A deadlock avoidance mechanism for cases were one or more devices does not respond to a PME_Turn_Off request with a PME_TO_Ack Message is explained.
- Multi-function device behavior is explained given that the functions are programmed for entry into different Link power management states.

- If an upstream device rejects entry into L1 power state from a downstream device, downstream device does not have to transition Link to L0s, but instead can keep the link in the L0 state.
- In order to avoid errors, a wait time has been specified between two consecutive requests for Link entry into L1 power state.
- Default value of "Active State PM Control" bits (bit[1:0] of the Link Control Register) has been changed to 00b which implies L0s and L1 Active states entry is disabled by default.
- A new capability reported via the Role-Based Error Reporting bit in the Device Capability register is added.
- New feature called Advisory Non-Fatal Error Handling and related configuration registers are defined.
- New optionally supported Data Link Layer related error state called Surprise Down is added. Related configuration registers are also added.
- New configuration bit indicated whether the Data Link Layer is in the active or inactive state.
- Timing parameter from end of reset to start of Link Training is changed to 20ms.
- Optional power controller element added to list of Hot-Plug capability related elements.
- Related to Hot-Plug capability, "Electromechanical Interlock" related registers are added and a new read-only "No Command Completed Support" register bit is added.
- Associated with the PCI Express Enhanced Configuration Mechanism support (memory mapped configuration space), the size and base address for the range of memory addresses mapped to the Configuration Space is host bridge design specific.
- CPU Memory Writes to the address range that converts to configuration transactions in the host bridge is implemented as non-posted transactions. Thus, transaction ordering is preserved.
- SERR# Enable Bit in the Bridge Command register, when set, enables transmission by the primary interface of ERR_NONFATAL and ERR_FATAL error messages forwarded from the secondary interface.
- The usage of the Interrupt Message Number register is clarified when either MSI or MSI-X interrupt generation mechanism is enabled.
- Seven additional PCI Express Extended Capability register IDs/blocks are defined.

Overview of Changes to PCI Express Spec 1.1

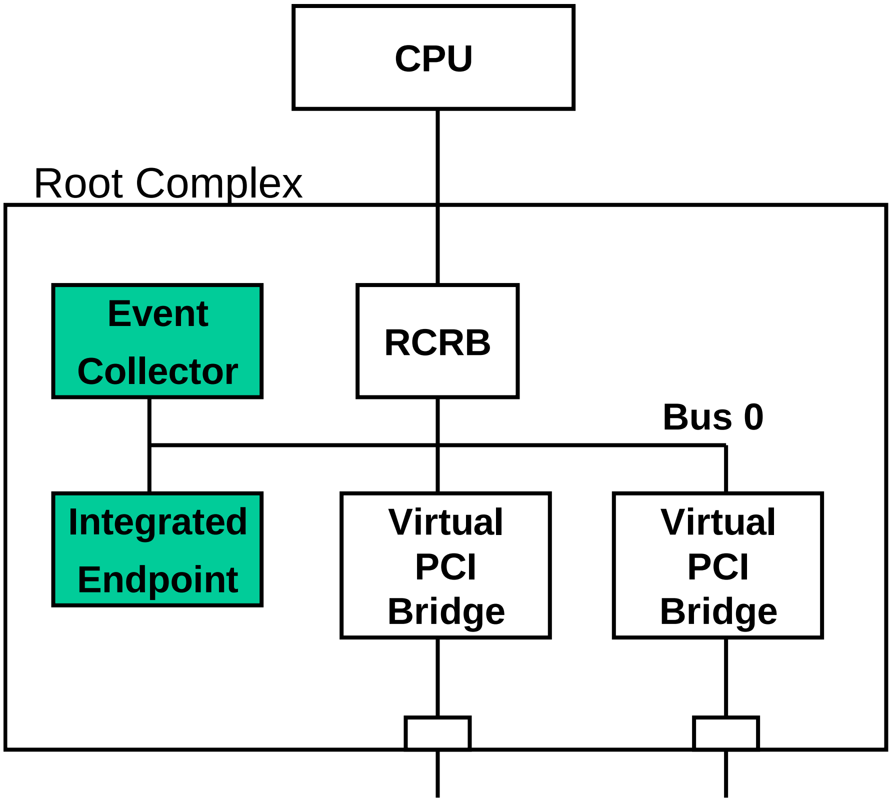
The author has documented these changes in sections that align to Chapters of MindShare's *PCI Express System Architecture* textbook.

Chapter 1: Architectural Perspective

Major Changes:

- MSI-X capability is documented in addition to the existing MSI interrupt mechanism. For interrupt support, legacy and native endpoints are required to support either MSI or MSI-X or both interrupt generation capability. The author recommends that designs which support interrupt generation capability support all three interrupt generation mechanisms: 1) Legacy mechanism, 2) MSI mechanism, and 3) MSI-X mechanism. This way, the device will function with all existing and future operating systems (OSs).
- Unlike switch designs which are not allowed to split packets, root complex devices that support peer-to-peer transactions between hierarchical domains are allowed to split packets into smaller packets that meet packet format rules. An exception to this rule is that Root Complexes are not allowed to split vendor-defined message packets into smaller packets except at 128-byte boundaries so as to allow PCIe-to-PCI/PCI-X bridges to forward messages across to secondary buses.
- Definition of root complex integrated endpoint and root complex event collector is added. Root complex integrated endpoints are embedded endpoints and are implemented on root complex internal logic that contains the root port. Root complex event collectors provide support for terminating PME and error messages generated by root complex integrated endpoints.

Figure 1-1: Event Collector and Integrated Endpoint



Clarifications:

- Definition of Root Complex is device that contains a host bridge with 1 or more root ports plus 0 or more of the following: integrated endpoints, event collectors.
- Definition of endpoints to include legacy endpoints, native PCI Express endpoints and root complex integrated endpoints.
- Switch integrated endpoints must not appear to configuration software on the switch's internal bus at the same level as downstream virtual bridges. In other words, these integrated endpoints must be implemented downstream of a virtual bridge.

Chapter 2: Architecture Overview

Major Changes:

NONE

Clarifications:

- When a switch generates its own request, e.g. for error reporting, it must use the Requester ID associated with the primary side of the bridge logically associated with the Port causing the request generation.
- Unless a device has a valid Requester ID (it will have one after the first configuration write it receives), it is not allowed to initiate non-posted requests.

Chapter 3: Address Spaces and Transaction Routing

Major Changes:

- Unlike switch designs which are not allowed to split packets, root complex devices that support peer-to-peer transactions between hierarchical domains are allowed to split packets into smaller packets that meet packet format rules. An exception to this rule is that Root Complexes are not allowed to split vendor-defined message packets into smaller packets except at 128-byte boundaries so as to allow PCIe-to-PCI/PCI-X bridges to forward messages across to secondary buses.

Clarifications:

- Bytes 8 through 15 of a message header are reserved except for messages that employ ID based routing. Message that employ ID based routing use these bytes to specify destination bus, device and function number as well as message specific information.

Chapter 4: Packet-Based Transactions

Major Changes:

- In addition to the existing First and Last BE usage rules, all non-QW aligned Memory Requests with length of 2 DW (1 QW) must enable only bytes that are contiguous with the data between the first and last DW of the Request.
- Whereas the Tag[7:0] field in posted requests is undefined and can contain any value, Vendor Defined Messages that are designed to be interoperable with PCI-X Device ID Messages have restrictions on the contents of the Tag[7:0] field.
- New optional feature called CRS Software Visibility which is enabled via a new configuration bit in the Root Control register of Root Complex's PCI Express Capability Block. This feature allows the Root Complex to inform

software of the reception of CRS completion status so that software can perform other tasks while the device under self-initialization completes its initialization process. When the CRS Software Visibility bit is set and initialization software access a device's Vendor ID register for which the device returns a Completion with CRS completion status, the Root Complex returns the data of 0001h instead of the Vendor ID and all '1's for any additional bytes included in the request. For all configuration writes or configuration reads other registers the Root Complex automatically re-issues the transaction as a new transaction.

Clarifications:

- Definition of Completer Abort (CA) and Unsupported Request (UR) updated to not only include completion packets with CA or UR. completion status but to include the condition in the completer that results in CA or UR states for both posted and non-posted transactions.
- Multi-function devices that implement the Max_Payload_Size register can be configured to the same or different values across the various functions. A function that generates packets must have a data payload size that does not exceed the size specified in that function's Max_Payload_Size. Multi-function devices are encouraged at the very least to ensure that the packet size does not exceed the smallest Max_Payload_Size setting across all functions in the device. Software should not set the Max_Payload_Size in different functions to different values unless software is aware of the specific implementation.
- To flush posted write buffers, the address for the zero-length Read must target the same device as the Posted Writes that are being flushed. One recommended approach is to use the same address as one of the Posted Writes being flushed.
- When a switch generates its own request, e.g. for error reporting, it must use the Requester ID associated with the primary side of the bridge logically associated with the Port causing the request generation.
- Unless a device has a valid Requester ID (it will have one after the first configuration write it receives), it is not allowed to initiate non-posted requests.
- Receivers may optionally check for violations of header format rules in a received packet. If a Receiver implementing these checks determines that a TLP violates these rules, the TLP is a Malformed TLP.
- Bytes 8 through 15 of a message header are reserved except for messages that employ ID based routing. Message that employ ID based routing use these bytes to specify destination bus, device and function number as well as message specific information.

- If a received Request Type is not supported either by design or a configuration setting, the request is an Unsupported Request and may be logged/reported as such an error.
- If a received Message has a Message Code that is undefined or not supported by the receiver (other than Vendor Defined Message Type 1 which is not treated as an error), the Message is an Unsupported Request. If the Message Code is an ignored Message, ignore the Message without reporting any error.
- When a device receives a Configuration Request and the device is busy, the condition under which that device can return a Completion with Configuration Request Retry Status (CRS) include the time after a Cold, Warm and Hot Resets as well as reset initiated in response to a D3hot to D0uninitialized device state transition. Devices are not permitted to return CRS following a device software-initiated reset due to device's software driver writing to a device-specific reset bit. Additionally, a device is not permitted to return CRS after having previously returned a Successful Completion without an intervening valid reset condition mentioned above.

Chapter 5: ACK/NAK Protocol

Major Changes:

- If a received TLP ends with EDB, but the TLP does not have an LCRC that is the inverse of what it should be, then discard the TLP and free up any TLP resources. In addition, if the NAK_SCHEDULED flag is clear, set it and immediately schedule a Nak DLLP for transmission.

Clarifications:

- If the Data Link Layer is reporting that a port is in the DL_Down state (i.e. port state that indicates no connection with another component in the Link), the Transaction Layer is not required to accept received TLPs from the Data Link Layer provided these TLPs have not been acknowledged. Also, flow control credits are not updated. In this DL_Down state, the Data Link Layer is allowed to discard TLPs as long as they have not been acknowledged.
- When the NAK_SCHEDULED flag is clear and a Nak DLLP is to be transmitted, set the NAK_SCHEDULED flag and "immediately" schedule a Nak DLLP to be transmitted. Set the NAK_SCHEDULED flag.
- The AckNak_LATENCY_TIMER must be restarted from 0 each time an Ack or Nak DLLP is scheduled for transmission.

- When the NAK_SCHEDULED flag is set, the device cannot send Nak DLLPs.
- The Ack DLLP may be scheduled more often for transmission than required.

Chapter 6: QoS/TCs/VCs and Arbitration

Major Changes:

- A new multi-function arbitration model that defines an optional arbitration mechanism within a multi-function device has been created. This functionality provides an arbitration scheme for the device's Upstream Egress Port from its multiple functions via a Multi-Function Virtual Channel (MFVC) capability register structure. The optionally implemented MFVC Capability structure manages TC/VC mapping, optional Function Arbitration, and optional VC Arbitration for the device's Upstream Egress Port. It does not provide arbitration support for peer-to-peer requests between functions nor does it provide arbitration support for downstream requests arriving at the Endpoint port targeting the functions.
- For Endpoints that do not support the MFVC Capability structure, Function 0's VC Capability structure provides the TC/VC mapping for the Link. The arbitration between the functions for the Link is beyond the scope of the specification and is implementation specific.
- Regarding isochronous transaction support in multi-function devices. The MFVC capability structure should not apply backpressure to isochronous requests coming from it functions. The function simply drops packets that cannot be transmitted at the desired injection rate. The MFVC capability structure must support Time-Based arbitration for those VCs capable of supporting isochronous traffic.

Clarifications:

- Definition of Port Arbitration is clarified. For switches, Port Arbitration refers to the arbitration at an Egress Port between traffic coming from other Ingress Ports that is mapped to the same VC. For Root Ports, Port Arbitration refers to the arbitration at a Root Egress Port between peer-to-peer traffic coming from other Root Ingress Ports that is mapped to the same VC. For RCRBs, Port Arbitration refers to the arbitration at the RCRB (e.g., for host memory) between traffic coming from Root Ports that is mapped to the same VC.

Chapter 7: Flow Control

Major Changes:

- The timeout flow control mechanism must be disabled if an Infinite Credit advertisement has been made during initialization for P, NP and Cpl flow control classes.
- During flow control initialization, InitFC1 and InitFC2 DLLPs must be transmitted in the order InitFC for P, NP, then Cpl.
- The three types of InitFC1 or InitFC2 DLLPs must be transmitted at least once every 34 µs, but it is strongly recommended that these DLLPs be transmitted as often as possible particularly when there are no other TLPs or DLLPs to transmit.
- While in InitFC2 stage, transaction layer must block transmission of any other new TLPs.

Clarifications:

- All received malformed TLPs (such as those with undefined Type field in the TLP header) must be discarded without updating receiver flow control information unless it is unambiguous which buffer to release in which case it is optional to update receiver flow control information.
- Flow Control mechanisms used internally within a multi-function device is outside the scope the specification.
- For the Posted Data VC buffer, the minimum advertised buffer credits are the largest possible setting of the Max_Payload_Size for the component divided by FC Unit Size. For a multi-function device, this credit advertisement takes into account the largest Max_Payload_Size setting of all functions in the device.
- If the Data Link Layer is reporting that a port is in the DL_Down state (i.e. port state that indicates no connection with another component in the Link), the Transaction Layer is not required to accept received TLPs from the Data Link Layer provided these TLPs have not been acknowledged. Also, flow control credits are not updated. In this DL_Down state, the Data Link Layer is allowed to discard TLPs as long as they have not been acknowledged.

Chapter 8: Transaction Ordering

Major Changes:

- Under normal operating conditions, Endpoints must not delay the acceptance of a Posted Request for more than the Posted Request Acceptance Limit of 10 µs. The device must either (a) be designed to process received Posted Requests and return associated Flow Control credits within 10 µs, or (b) depend on a restricted programming model to ensure that a Posted Request is never sent to the device either by software or by other devices while the device is unable to accept a new Posted Request within 10 µs. Under the following conditions, the 10 µs limit does not apply:
  -The period immediately following a Fundamental Reset.
  -TLP retransmissions or Link retraining.
  -One or more dropped Flow Control Packets.
  -The device is in a diagnostic mode.
  -Device is in an abnormal use mode.
- Some additional transaction ordering rules apply for traffic within the same traffic class. These rules are:
  -For Endpoints and Bridges, acceptance of a posted or non-posted request must not depend upon the transmission of a posted or non-posted request in the same Traffic Class.
  -Acceptance of a posted request must not depend upon the transmission of a completion in the same Traffic Class.
  -Completions issued for non-posted requests must be returned in the same Traffic Class as the corresponding non-posted request.
  -Acceptance of a completion must not depend upon the transmission of a posted or non-posted request or a completion in the same Traffic Class.
  -Root Complexes that support peer-to-peer operation and Switches must enforce these transaction ordering rules for all forwarded traffic.
  -Devices should not forward traffic from one Virtual Channel to another, though if this is done, one must guarantee deadlock free operation.

Clarifications:

- To flush posted write buffers, the address for the zero-length Read must target the same device as the Posted Writes that are being flushed. One recommended approach is to use the same address as one of the Posted Writes being flushed

Chapter 9: Interrupts

Major Changes:

- MSI-X capability is documented in addition to the existing MSI interrupt mechanism. For interrupt support, legacy and native endpoints are required to support either MSI or MSI-X or both interrupt generation capability. The author recommends that designs which support interrupt generation capability support all three interrupt generation mechanisms: 1) Legacy mechanism, 2) MSI mechanism, and 3) MSI-X mechanism. This way, the device will function with all existing and future operating systems (OSs).

Clarifications:

- PME and Hot-Plug Event interrupts always share the same MSI or MSI-X vector, as indicated by the Interrupt Message Number field in the PCI Express Capabilities register.

Chapter 10: Error Detection and Handling

Major Changes:

- Multi-function devices that receive a packet for which the TC does not map to any enabled VCs in the MFVC Capability structure is treated as a malformed TLP.
- Detection of Link Training errors which was optional is removed from the specification. The corresponding error reporting Status, Mask, and Severity bits in the Advanced Error Capability block are changed to Undefined. Software is supposed to write a one to the bit position (bit 0) for the formerly Link Training Error Mask bit.
- It is required to detect 8b/10b decode errors and report them as Physical Layer receiver errors. It remains optional to trigger a Receiver Error on Framing Error, Loss of Symbol Lock, Lane De-skew Error, and Elasticity Buffer Overflow/Underflow.
- A new capability reported via the Role-Based Error Reporting bit in the Device Capability register is added. This feature allows non-fatal errors to be sometimes signaled with ERR_NONFATAL message, sometimes signaled with ERR_COR message, and sometimes not signaled at all, depending upon the role of the agent that detects the error and whether the agent implements Advanced Error Reporting Capability register block.

On some platforms, sending ERR_NONFATAL will prevent software from attempting recovery or determining the ultimate disposition of the error. For cases where the detecting agent is not the appropriate agent to determine the ultimate disposition of the error, a detecting agent with Advanced Error Reporting Capability register block can signal the non-fatal error with ERR_COR, which serves as an advisory notification to software. For cases where the detecting agent is the appropriate one, the agent signals the non-fatal error with ERR_NONFATAL message.

E.g. A poisoned TLP might be signaled by intermediate Receivers (Switches) with ERR_COR message, while the ultimate destination Receiver might signal it with ERR_NONFATAL message.

Here is another example. In 1.0a devices with no Role-Based Error Reporting capability, if the UR Reporting Enable bit is clear, then the completer is prevented from reporting any error messages when an UR error is detected (on both posted and non-posted requests received). A device with Role-Based Error Reporting capability, when the SERR# Enable bit is set, even with Unsupported Request Reporting Enable bit clear, the device will send an ERR_NONFATAL or ERR_FATAL messages to signal UR errors on bad posted requests received. The completer will not send an uncorrectable error message for non-posted requests received, thus keeping PC-compatible configuration space probing using configuration read requests. By reporting errors on bad posted requests received, silent data corruption is prevented. It is recommended that software/firmware keeps the UR Error Reporting Enable bit clear for devices not capable of Role-Based Error Reporting, but set the UR Reporting Enable bit for devices that are capable of Role-Based Error Reporting. That way, UR errors are reported on bad posted requests received, but no errors are reported on non-posted requests (such as configuration space probing transactions).

- To avoid "Error Pollution" for errors detected in the Transaction layer, it is permitted and recommended that no more than one error be reported for a single received TLP. The following order (from highest to lowest) should be used when multiple errors are detected at the same time for the same received TLP:

  -Receiver Overflow.
  -Flow Control Protocol Error.
  -ECRC Check Failed.
  -Malformed TLP.
  -Unsupported Request (UR), Completer Abort (CA), or Unexpected Completion.

-Poisoned TLP Received.

- Advisory Non-Fatal Error Handling associated with devices that support Role-Based Error Reporting. On an error-by-error case basis (documented below in this bullet), a device that detects an uncorrectable error (of non-fatal class) reports it using an ERR_COR (not ERR_NONFATAL) message if it implements the Advanced Error Reporting (AER) Capability registers. If the device does not implement AER, it does not report an error message. This allows software to more intelligently handle errors classified as non-fatal. If the detected error severity is increased to fatal however, the device will report ERR_FATAL instead of ERR_NONFATAL or ERR_COR.

  Errors that are handled as Advisory Non-Fatal Errors are:
  -Completer sending a completion with UR/CA status.
  -Intermediate receiver that is not the final destination of a TLP that detects a non-fatal error.
  -Ultimate receiver detects a poisoned TLP.
  -Requester detects a completion timeout.
  -Receiver receives an unexpected completion.
  -Requestor receives a completion with status UR/CA.
  -Error Forwarding (Data Poisoning).
- Registers associated with Advisory Non-Fatal Error Handling are 1) Advisory Non-Fatal Error Status in the AER Correctable Status register and 2) Advisory Non-Fatal Error Mask in the AER Correctable Mask register. When a non-fatal error is detected (of Advisory Non-Fatal Error classification), the Advisory Non-Fatal Status bit is set. The Advisory Non-Fatal Mask bit is checked to see if it is clear, in which case error logging and reporting proceeds as follows. A corresponding error status bit in the Uncorrectable Error Status register is set. If a corresponding Mask bit in the Uncorrectable Error Mask register is clear, then the First Error Pointer and Header Log register are updated. Finally, an ERR_COR message is sent if the Correctable Error Reporting Enable bit is set in the Device Control register.
- A new Data Link Layer related error state called Surprise Down is added. This error detection is optional. Its default severity is Fatal. Surprise Down is a Data Link Layer state caused by the Physical Layer reporting to the Data Link Layer that the Link is Down (Physical LinkUp = 0) this causing the Data Link Layer state to transition from DL_Active to DL_Inactive. Upstream components are optionally permitted to treat this transition from DL_Active to DL_Inactive as a Surprise Down error, except in the following cases where this error detection is blocked:
  -If the Secondary Bus Reset in Bridge Control register has been set to 1b by software, then the subsequent transition to DL_Inactive is not an error.

-If the Link Disable bit has been set to 1b by software, then the subsequent transition to DL_Inactive is not an error.
-If a PME_Turn_Off Message has been sent through this port, then the subsequent transition to DL_Inactive is not an error. DL_Inactive transition for this condition will not occur until a power off, a reset, or a request to restore the link is sent to the Physical layer. In the case where the PME_Turn_Off/PME_TO_Ack handshake fails to complete successfully, a Surprise Down error may be detected.
-If the port is associated with a hot-pluggable slot, and the Hot Plug Surprise bit in the Slot Capabilities register is set to 1b, then any transition to DL_Inactive is not an error.
-If the port is associated with a hot-pluggable slot (Hot-Plug Capable bit in the Slot Capabilities register set to 1b), and Power Controller Control bit in Slot Control register is 1b (Off), then any transition to DL Inactive is not an error.

- Configuration register bits related to Surprise Down include: Surprise Down Error Status, Surprise Down Error Mask, Surprise Down Error Severity bits. These bit are added to the Uncorrectable Error Status, Mask and Severity registers respectively of the Advance Error Extended Capability register block. Software can determine if a device supports detection of this error via the Link Capability register bit called Surprise Down Error Reporting Capable bit.
- When the Data Link Layer status is DL_Active, a new Link Status register bit (bit 13) called "DLL Link Active" is set. When the Data Link Layer status goes to DL_Inactive, this bit clears. Support of this optional feature is indicated via the Link Capability bit (bit 20 called DLL Active Reporting Capable) if it is set. During a hot-plug event, the DLL Link Active status bit allows software to determine when a device has been powered. Software must then wait 100 us after this status bit has been detected set before configuring the newly added device. If DLL Active Reporting Capable bit is clear, then the DLL Link Active bit must be hardwired to 0.

Clarifications:

- When a switch generates its own request, e.g. for error reporting, it must use the Requester ID associated with the primary side of the bridge logically associated with the Port causing the request generation.
- Receivers may optionally check for violations of header format rules in a received packet. If a Receiver implementing these checks determines that a TLP violates these rules, the TLP is a Malformed TLP.
- All received malformed TLPs (such as those with undefined Type field in the TLP header) must be discarded without updating receiver flow control

information unless it is unambiguous which buffer to release in which case it is optional to update receiver flow control information.

- If a received Request Type is not supported either by design or a configuration setting, the request is an Unsupported Request and may be logged/reported as such an error.
- If a received Message has a Message Code that is undefined or not supported by the receiver (other than Vendor Defined Message Type 1 which is not treated as an error), the Message is an Unsupported Request. If the Message Code is an ignored Message, ignore the Message without reporting any error.
- A requester that receives a completion with completion status of Unsupported Request (UR) or Completer Abort (CA) reports these errors in a similar manner to the mechanism of PCI devices reporting Master Aborts or Target Aborts. The completer who reports a UR or CA completion reports these detected errors via the PCI Express error handling mechanism.
- Link Errors are defined as 8b/10b decode errors, loss of Symbol lock, Elasticity Buffer Overflow/Underflow, or loss of Lane-to-Lane de-skew.

Chapter 11: Physical Layer Logic

Major Changes:

It is now required to detect 8b/10b decode errors and report them as Physical Layer receiver errors. It remains optional to trigger a Receiver Error on Framing Error, Loss of Symbol Lock, Lane De-skew Error, and Elasticity Buffer Overflow/Underflow.

Clarifications:

NONE

Chapter 12: Electrical Physical Layer

Major Changes:

- A new “Eye measurement Clock Recovery Function” is described.

Clarifications:

- For Beacon pulses with a width greater than 500 ns, the minimum and maximum beacon amplitude is –6 dB down from the minimum and maximum differential peak to peak output voltage ($V_{TXDIFFp-p}$) respectively.

- Electrical Idle Exit does not occur if a signal smaller than $V_{RX-IDLE-DET-DIFFp-p}$ = 65 mV minimum is detected at a Receiver. Electrical Idle Exit occurs if a signal larger than $V_{RX-IDLE-DET-DIFFp-p}$ = 175 mV maximum is detected at a Receiver.

Chapter 13: System Reset

Major Changes:

- A device must enter Detect within 20ms of end of Fundamental Reset. This number has been changed from the previous 80ms.
- Software waits 100ms after Fundamental Reset to start generating configuration traffic. Software also waits 1s after Fundamental Reset before it may determine that a device which fails to return a Successful Completion status for a valid Configuration Request is a broken device. If software has no mechanism of determining the end of Fundamental Reset, then it uses some known event after Fundamental Reset de-assertion to base the above 2 timing parameters.

Clarifications:

- When a device receives a Configuration Request and the device is busy, the condition under which that device can return a Completion with Configuration Request Retry Status (CRS) include the time after a Cold, Warm and Hot Resets as well as reset initiated in response to a D3hot to D0uninitialized device state transition. Devices are not permitted to return CRS following a device software-initiated reset due to device's software driver writing to a device-specific reset bit. Additionally, a device is not permitted to return CRS after having previously returned a Successful Completion without an intervening valid reset condition mentioned above.
- Any type of reset (cold, warm, hot, or DL_Down) has the same effect at the Transaction Layer and above as would RST# assertion and de-assertion in conventional PCI system.
- When software writes to the Secondary Bus Reset bit of the Bridge Control register in Header 1, the switch or root downstream port must guarantee a minimum reset time of Trst = 1ms as specified in the PCI specification.

Chapter 14: Link Initialization and Training

Major Changes:

- Detection of Link Training errors which was optional is removed from the specification. The corresponding error reporting Status, Mask, and Severity bits in the Advanced Error Capability block are changed to Undefined. Software is supposed to write a one to the bit position (bit 0) for the formerly Link Training Error Mask bit.
- It is possible to design a port in which the Lanes support different data rates as indicated via the Link Data Rate Identifier within the TS1/TS2 packets. i.e. some Lanes support 2.5Gbit/s protocol and some Lanes support 5Gbit/s data rates. The Link may continue to train with a single LTSSM or the LTSSM may split into multiple LTSSMs based on a common speed for all Lanes per LTSSM.
- Upon entry into Configuration.Linkwidth.Start, if a port supports crosslink capability, then the behavior of the downstream port during training is clarified. All downstream Lanes of a downstream port that detected a receiver during Detect must first transmit 16-32 TS1s with a non PAD Link number and PAD Lane number. After this occurs if the downstream port receives two consecutive TS1 ordered sets with a Link number different than PAD (K23.7) and a Lane Number set to PAD, the Downstream port is now designated as Upstream port and a new random cross Link timeout is chosen. Re-enter Configuration.Linkwidth.Start as an upstream port.
- It is possible for a port to detect a crosslink on some Lanes of a Link, in which case training can continue with a single LTSSM or optionally split into multiple LTSSMs.
- Upon entry into Configuration.Linkwidth.Start, if a port supports crosslink capability, then the behavior of the upstream port during training is clarified. All upstream Lanes of an upstream port that detected a receiver during Detect must first transmit 16-32 TS1s with PAD Link number and PAD Lane number. After this occurs the port can proceed with the training process as described in the specification.
- During Configuration state while in the step required for initializing the Lane number, the upstream port may wait up to 1ms before accepting the Lane number indicated in the received TS1s. The reason for waiting before accepting the Lane numbers is to prevent received errors or any Lane-to-Lane skew from affecting the final Link width.
- For Loopback support, a Loopback Slave is required to retransmit the received 10-bit information as received, with the polarity inversion determined during Polling applied. The Loopback Slave must also continue to perform clock tolerance compensation.

- During Detect state, it is not required that the detect sequence be performed on both conductors of a differential pair.
- The definition of the "Extended Sync" bit (bit 7 of the Link Control register) has been changed. If this bit is set, a device whose Link is in L0s is forced to transmit additional ordered sets in L0s prior to entering L0. This extended sync provides external devices monitoring the link time to achieve bit and symbol lock before the Link exits L0s and enters the L0 state where the Link resumes normal communication.

Clarifications:

NONE

Chapter 15: Power Budgeting

Major Changes:

NONE

Clarifications:

NONE

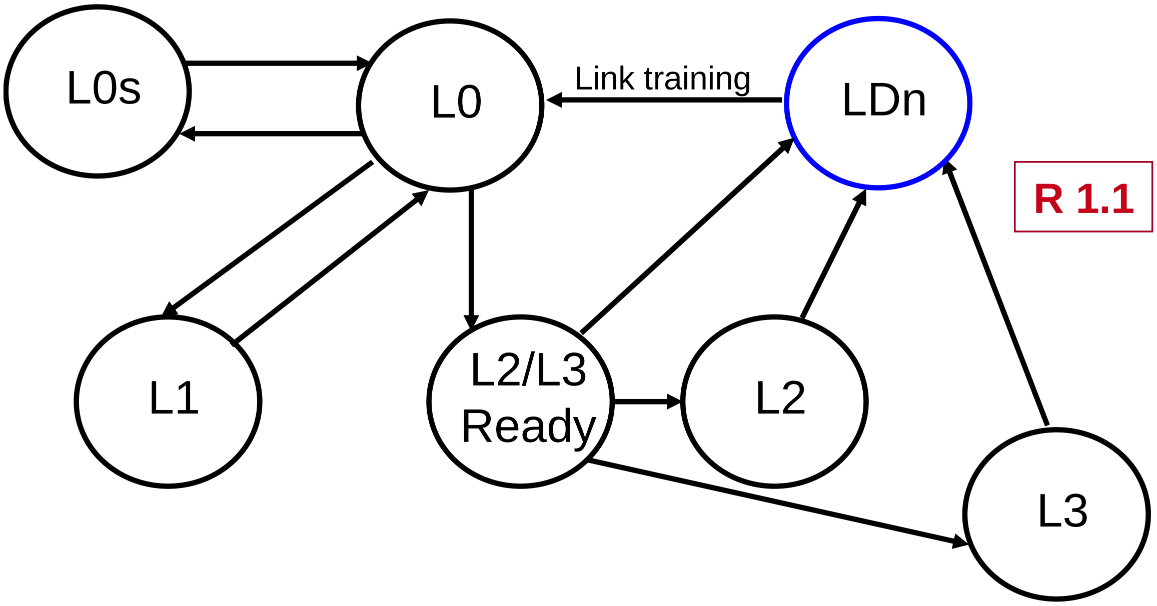
Chapter 16: Power Management

Major Changes:

- The L2/L3_Ready state is now defined as pseudo transitional state (corresponding to the LTSSM L2 state) that a given Link enters into in preparation for the removal of power and clocks from the downstream component or both attached components. Formerly, this state was only entered as a result of PCI PM Software placing a device into D3hot power state and then causing a PME_Turn_Off/PME_TO_Ack handshake. Upon completing the handshake, the device would transition to L2/L3_Ready. The Link state transitions involved are: L0->L1->L0->L2/L3_Ready. When Vcc power and clocks are removed by the system, then the device would transition either to L2 if Vaux is available or L3 if Vaux is not available. The new capability defined allows a device to transition to L2 or L3 via L2/L3 Ready without software first placing a device into the D3 state first. While the devices are in any of the D states (D0, D1, D2, D3hot), software can cause the Root to initiate a PME_Turn_Off/PME_TO_Ack handshake which will result in the devices transitioning to the L2/L3_ready state. The Link transitions from L0 directly to L2/L3_ready. Then software can cause the system to remove Vcc power and clocks and the downstream component must be ready for this event.

- On receiving a PME_Turn_Off Message from the upstream Root, a device must block the transmission of PM_PME Messages and return a PME_TO_Ack Message upstream. This device is allowed to send a PM_PME Message only after the Link is returned to L0 state through LDn.
- A new transitional power state called Link Down (LDn) is added. This pseudo transitional state is entered prior to L0. LDn is associated with the LTSSM states Detect, Polling, and Configuration, and, when applicable, Disabled, Loopback, and Hot Reset. Entry into LDn can be from L2/L3_Ready, L2 or L3. From LDn the transition to L0 is via Link re-initialization through LTSSM Detect state. During the LDn state, Vcc power has been turned back on as well as Reference Clock is also turned on. The devices internal PLL may either be On or Off. Vaux may or may not be available. LDn state is entered upon a Fundamental Reset, Hot Reset, transmission from an upstream component, or Vcc/Clock reapplication when device was in L2 or L3.

Figure 16-1: Link Down Power State



- While the Link is in L1 or L2/L3_Ready power state, the Reference Clock was required to be on. The new feature allows the system to gate Reference Clock while a device's Link is in the L1 or L2/L3_Ready state. This is an especially useful feature for low power mobile or handheld applications. If the Reference Clock is turned off, the system must be tolerant of the additional latency associated with turning back on the Reference Clock during low power state exit.

- A configuration register bit in the Link Capability register (bit 18) called "Clock Power Management Capable" indicates the device is capable of signaling REFCLK to be turned off (via the usage of the CLKREQ# signal) in L1 or L2/L3_Ready. Another bit in the Link Control register (bit 8) called "Enable Clock Power Management" causes CLKREQ# output from the device to be de-asserted which results in gating of REFCLK to the device.
- Configuration and Message requests are the only TLPs accepted by a function in the D1 or D2 device power state. All other received Requests must be handled as Unsupported Requests, and all received Completions must be handled as Unexpected Completions. If an error caused by a received TLP is detected while in D1 or D2, and error reporting is enabled, the link must be returned to L0 if it is not already in L0 and an error message must be sent. If an error caused by an event other than a received TLP (e.g. a Completion Timeout) while in D1 or D2, an error message must be sent when the device is programmed back to the D0 state.
- Configuration and Message requests are the only TLPs accepted by a function in the $D3_{hot}$ state. All other received Requests may optionally be handled as Unsupported Requests, and all received Completions may optionally be handled as Unexpected Completions. If an error caused by a received TLP is detected while in $D3_{hot}$, and error reporting is enabled, the link must be returned to L0 if it is not already in L0 and an error message must be sent. If an error caused by an event other than a received TLP (e.g. a Completion Timeout) while in D3hot, an error message may optionally be sent when the device is programmed back to the D0 state.
- During Link L1 entry from L0, the downstream component and upstream component perform a PM_Enter_L1 DLLP/PM_Request_Ack DLLP handshake. These DLLPs must be transmitted continuously until the handshake is complete with no more than 4 symbol times of Logical Idle between the DLLPs. Similarly, during Link L1 Active entry from L0, the downstream component and upstream component perform a PM_Active_Request_Enter_L1 DLLP/PM_Request_Ack(or _Nak) DLLP handshake. These DLLPs must be transmitted continuously until the handshake is complete with no more than 4 symbol times of Logical Idle between the DLLPs The transmission of SKIP Ordered-Sets is permitted during this handshake and does not contribute to the 4 symbol times of Logical Idle limit.
- Upon exit from L1 to L0, it is recommended that the Downstream Component send flow control update DLLPs for all enabled VCs and FC types within 1 µs of L1 exit.
- In order to avoid deadlock in the case where one or more devices do not respond with a PME_TO_Ack Message, the power manager must implement a timeout after waiting for a certain amount of time, after

which it proceeds as if the PME_TO_Ack Message had been received and all links put into the L2/L3 Ready state. The recommended limit for this timer is 1 ms to 10 ms.

- In a multi-function device that is in the D0 state, if at least one function has been enabled for only L0s active state entry and at least one other function has bee enabled for only L1 active state entry, then ASPM is disabled for that device. Similarly, if at least one of the functions is enabled for L1 active state entry only, then ASPM is enabled for L1 only. Similarly if at least one of the functions is enabled for L0s state entry only, then ASPM is enabled for L0s only.
- The following requirement has been removed: A downstream component must as soon as possible transition its upstream Link to L0s if its request to transition the Link to L1 is rejected. The Link can instead continue to remain in the L0 state.
- If any TLPs become available from the Transaction Layer for transmission during the L1 negotiation process, the transition to L1 must first be completed and then the downstream component must initiate a return to L0.
- Between two handshake events to negotiate entry into L1 Active state, a downstream component must either enter and exit L0s or wait 10 us. If this is not done, there is the risk that the two components will get out of sync with each other, and the results may be undefined.

  Similarly, if the upstream component rejects request for entry into L1, then it must either detect the Link entering L0s or detect a break in reception of PM_Request_L1 DLLP for 9.5us before it accepts a subsequent request for entry into L1.
- The default state of the of the "Active State PM Control" bits (bit[1:0] of the Link Control Register) has been changed to 00b which implies L0s and L1 Active states entry is disabled. A specific form factor specification could indicate a different default.

Clarifications:

- In addition to endpoints which can initiate negotiation for entry into ASPM L1, Switch upstream ports can also initiate negotiation for entry into ASPM L1.
- The ASPM L1 power state is an optionally supported power state unless otherwise specified by a particular form factor specification.
- PME and Hot-Plug Event interrupts always share the same MSI or MSI-X vector, as indicated by the Interrupt Message Number field in the PCI Express Capabilities register.

Chapter 17: Hot-Plug

Major Changes:

- Attention Indicator On/Off/Blink, Power Indicator On/Off/Blink and Attention Button Pressed messages are no longer supported. Transmitters are strongly encouraged not to transmit these messages, but if message transmission is implemented, it must conform to the requirements of the 1.0a version of this specification. Receivers are strongly encouraged to ignore receipt of these messages, but are allowed to process these messages in conformance with the requirements of 1.0a version of this specification.

  The related Attention and Power Indicator configuration registers are converted to Undefined registers.
- When the Data Link Layer status is DL_Active, a new Link Status register bit (bit 13) called "DLL Link Active" is set. When the Data Link Layer status goes to DL_Inactive, this bit clears. Support of this optional feature is indicated via the Link Capability bit (bit 20 called DLL Active Reporting Capable) if it is set. During a hot-plug event, the DLL Link Active status bit allows software to determine when a device has been powered. Software must then wait 100 us after this status bit has been detected set before configuring the newly added device.
- One more optional element is added to the list of elements related to hot-plug support. This is the power controller(s) that are software controlled and controls power to a slot or adapter as well as monitor power for fault conditions.
- If a power controller is implemented for a slot, the slot main power must be automatically removed from the slot when the MRL Sensor indicates that the MRL is open. If signals such as Vaux and SMBus are switched by the MRL, then these signals must be automatically removed from the slot when the MRL Sensor indicates that the MRL is open and must be restored to the slot when the MRL Sensor indicates that MRL is closed.
- In the absence of an MRL sensor, for some form factors, staggered presence detect pins may be used to handle the switched signals such as slot main power, Vaux and SMBus signals. In this case, when the presence pins break contact, the switched signals are automatically removed from the slot.
- New "Electromechanical Interlock" registers are added. These register bits are: Electromechanical Interlock Present bit in the Slot Capability register, Electromechanical Interlock Control bit in the Slot Control register and the Electromechanical Interlock Status bit in the Slot Status register. The optional Electromechanical Interlock feature prevents adapter removal

when the interlock is turned on. The state of the Electromechanical Interlock must be maintained even when power to the slot is removed.

- A new read-only "No Command Completed Support" register bit in the Slot Capability register indicates when set that this slot does not generate software notification when an issued command is completed by the Hot-Plug Controller. This bit is only permitted to be set to 1b if the hot-plug capable port is able to accept writes to all fields of the Slot Control register without delay between successive writes. When this bit is hardwired to 0, then the "Command Completed" bit in the Slot Status register will be set whenever a hot-plug command has completed and the Hot-Plug Controller is ready to accept a subsequent command. It is set as an indication to host software that the Hot-Plug Controller has processed the previous command and is ready to accept the next command.

Clarifications:

- PME and Hot-Plug Event interrupts always share the same MSI or MSI-X vector, as indicated by the Interrupt Message Number field in the PCI Express Capabilities register.

Chapter 18: Add-in Cards and Connectors

Major Changes:

NONE

Clarifications:

NONE

Chapter 19: Configuration Overview

Major Changes:

- It is strongly recommended that PCI Express devices place no registers in Configuration Space other than those in Headers or Capability structures. Instead they should be placed in Memory Space that is allocated by one or more Base Address registers. Device-specific registers that need to be accessible before Memory Space is allocated should instead be placed in a Vendor-Specific Capability Structure or a Vendor-Specific Extended Capability Structure.

Clarifications:

- Reserved registers must be read-only and return a 0 when read.

Chapter 20: Configuration Mechanism

Major Changes:

- For PCI Express Enhanced Configuration Mechanism support (memory mapped configuration space), the size and base address for the range of memory addresses mapped to the Configuration Space is host bridge design specific. They are reported by the firmware to the operating system in an implementation-specific manner. The size of the memory range is determined by the number of bits that the host bridge maps to the Bus Number field in the configuration address. Systems that support more than 4 GB of memory addresses are encouraged to map eight bits of Memory Address to the Bus Number field there by utilizing 256 MB of memory space to map to configuration space.
- Given that Memory Write transactions from the CPU tend to be posted type transactions in the root complex, the system hardware must provide a method for the system software to guarantee that a write transaction using the enhanced configuration access mechanism is completed by the completer before system software execution continues with the next command. If this is not done, write ordering problems may be created for software. One solution to this problem is to have the host bridge/root complex treat Memory Writes to the address range that maps to configuration space as non-posted. I.e. to have the host bridge not return a READY to the processor until the configuration write to the completer has completed. That way the processor does not proceed to the next command until the configuration write has completed.

Clarifications:

NONE

Chapter 21: PCI Express Enumeration

Major Changes:

- New optional feature called CRS Software Visibility which is enabled via a new configuration bit in the Root Control register of Root Complex's PCI Express Capability Block. This feature allows the Root Complex to inform software of the reception of CRS completion status so that software can perform other tasks while the device under self-initialization completes its initialization process. When the CRS Software Visibility bit is set and initialization software access a device's Vendor ID register for which the device returns a Completion with CRS completion status, the Root Complex returns the data of 0001h instead of the Vendor ID and all '1's for any additional bytes included in the request. For all configuration writes or configuration reads other registers the Root Complex automatically re-issues the transaction as a new transaction.

Clarifications:

- The assignment of Device Numbers to the downstream ports within a switch, may be done in an implementation specific way.
- When a device receives a Configuration Request and the device is busy, the condition under which that device can return a Completion with Configuration Request Retry Status (CRS) include the time after a Cold, Warm and Hot Resets as well as reset initiated in response to a D3hot to D0uninitialized device state transition. Devices are not permitted to return CRS following a device software-initiated reset due to device's software driver writing to a device-specific reset bit. Additionally, a device is not permitted to return CRS after having previously returned a Successful Completion without an intervening valid reset condition mentioned above.

Chapter 22: PCI Compatible Configuration Registers

Major Changes:

- For Type 1 configuration space header implemented in Bridges, the SERR# Enable Bit in the Command register, when set, enables transmission by the primary interface of ERR_NONFATAL and ERR_FATAL error messages forwarded from the secondary interface. This bit does not affect the transmission of forwarded ERR_COR messages.

Clarifications:

- When software writes to the Secondary Bus Reset bit of the Bridge Control register in Header 1, the switch or root downstream port must guarantee a minimum reset time of Trst = 1ms as specified in the PCI specification.

Chapter 23: Expansion ROMs

Major Changes:

NONE

Clarifications:

NONE

Chapter 24: Express-Specific Configuration Registers

Major Changes:

- New optional feature called CRS Software Visibility which is enabled via a new configuration bit in the Root Control register of Root Complex's PCI Express Capability Block. This feature allows the Root Complex to inform software of the reception of CRS completion status so that software can perform other tasks while the device under self-initialization completes its initialization process. When the CRS Software Visibility bit is set and initialization software access a device's Vendor ID register for which the device returns a Completion with CRS completion status, the Root Complex returns the data of 0001h instead of the Vendor ID and all '1's for any additional bytes included in the request. For all configuration writes or configuration reads other registers the Root Complex automatically re-issues the transaction as a new transaction.
- A new Multi-Function Virtual Channel (MFVC) Capability structure is added and optionally implemented in upstream ports such as endpoint and upstream switch ports. This MFVC Capability Structure serves as QoS logic between functions of a multi-function device's upstream egress port.
- Detection of Link Training errors which was optional is removed from the specification. The corresponding error reporting Status, Mask, and Severity bits in the Advanced Error Capability block are changed to Undefined. Software is supposed to write a one to the bit position (bit 0) for the formerly Link Training Error Mask bit.
- The Device/Port type field of the PCI Express Capability Structure includes two additional encodings of 1001b and 1010b which represent Root Complex Integrated Endpoint and Root Complex Event Collector respectively.
- The Advanced Error Interrupt Message Number field of the AER Extended Capability Structure is used by functions that request and receive multiple MSI interrupt Numbers. This register contains the offset from the base Message Data register of the MSI Capability Register set.

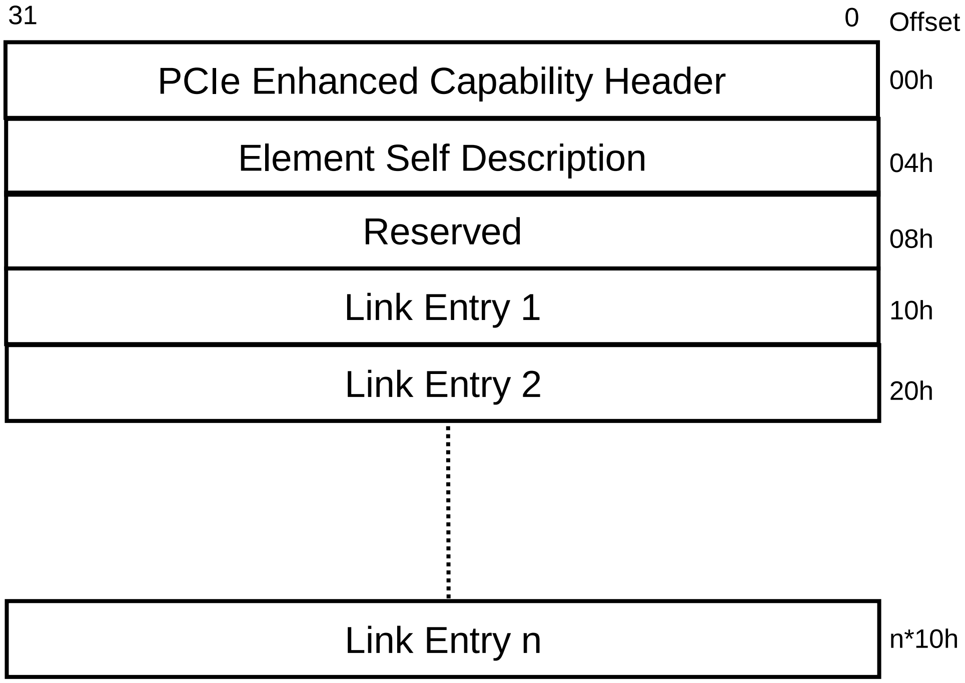
The interrupt vector pointed to by this Interrupt Message Number is shared by both PME and Hot-Plug event interrupts.

For MSI-X support, this field contains the MSI-X table entry used to generate the interrupt. The entry must be in the first 32 entries of the MSI-X table. If both MSI and MSI-X are supported, this register must point to the offset vector of which ever interrupt mechanism (MSI or MSI-X) is enabled.

- Additional PCI Express Extended Capabilities have been added. Below is a complete list of all Extended Capabilities and their associated IDs. Those IDs in bold are spec 1.1 additions
    - 0001h = Advanced Error Reporting Extended Capability
    - 0002h = Virtual Channel Extended Capability implemented in a device without an MFVC structure
    - 0003h = Serial Number Extended Capability
    - 0004h = Power Budgeting Extended Capability
    - **0005h** = Root Complex Link Declaration Extended Capability
    - **0006h** = Root Complex Internal Link Control Extended Capability
    - **0007h** = Root Complex Event Collector Endpoint Association Extended Capability
    - **0008h** = Multi-Function Virtual Channel (MFVC) Extended Capability
    - **0009h** = Virtual Channel Extended Capability implemented in a Multi-Function device with an MFVC structure
    - **000Ah** = RCRB Header Extended Capability
    - **000Bh** = Vendor-Specific Extended Capability

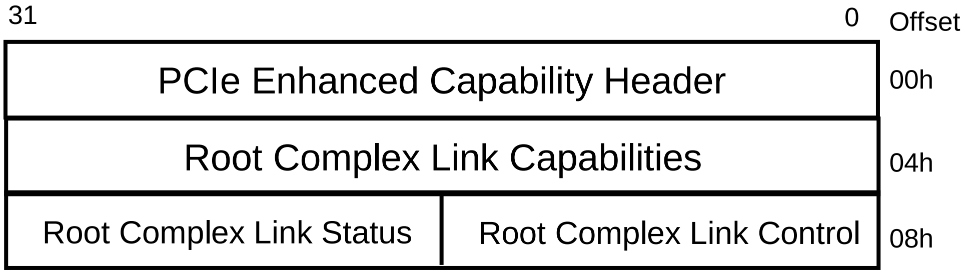
- The optional Root Complex Link Declaration Extended Capability register block is used to declare the Root Complex internal topology.

Figure 24-1: Root Complex Link Declaration Extended Capability Register



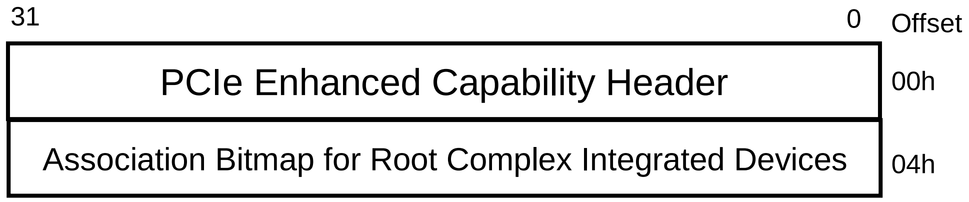
- The optional Root Complex Internal Link Control Extended Capability register block is used to control the internal Link connecting two Root Complex Components.

Figure 24-2: Root Complex Internal Link Control Extended Capability Register


31
0
Offset
PCIe Enhanced Capability Header
00h
Root Complex Link Capabilities
04h
Root Complex Link Status
Root Complex Link Control
08h

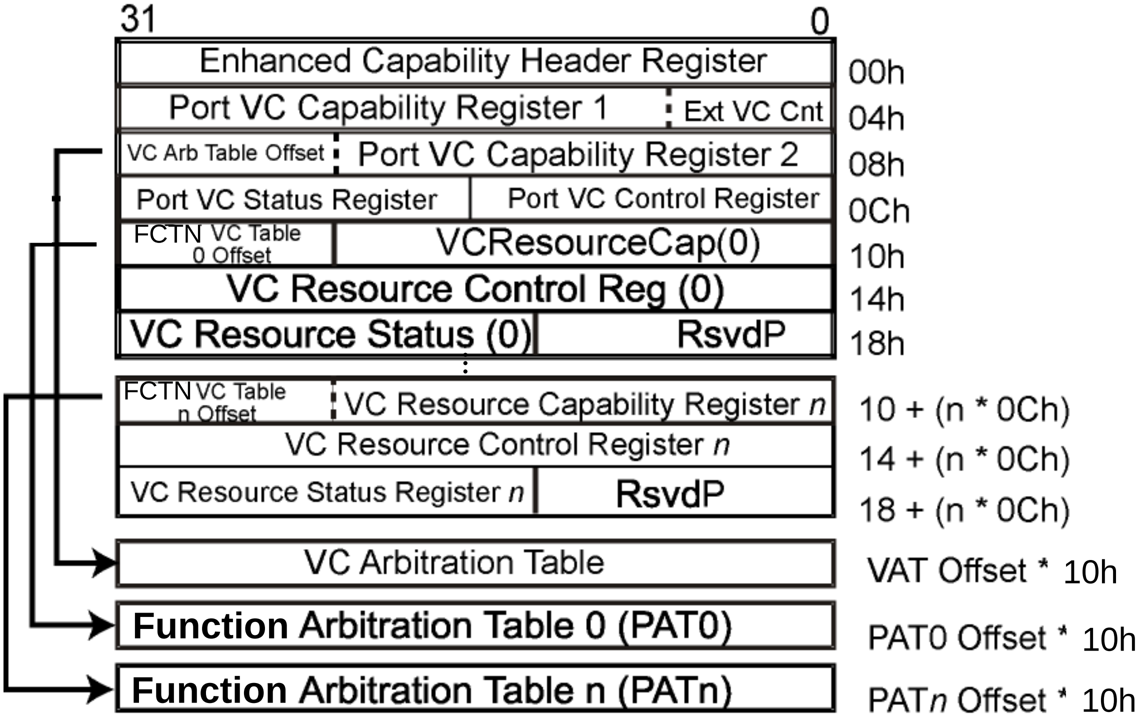
- The optional Root Complex Event Collector Endpoint Association Extended Capability register block declares the Root Complex Integrated Endpoints supported by the Root Complex Event Collector on the same logical bus on which the Root Complex Event Collector is located. A Root Complex Event Collector must implement the Root Complex Event Collector Endpoint Association Capability register block.

Figure 24-3: Root Complex Event Collector Endpoint Association Extended Capability Register


31
0
Offset
PCIe Enhanced Capability Header
00h
Association Bitmap for Root Complex Integrated Devices
04h

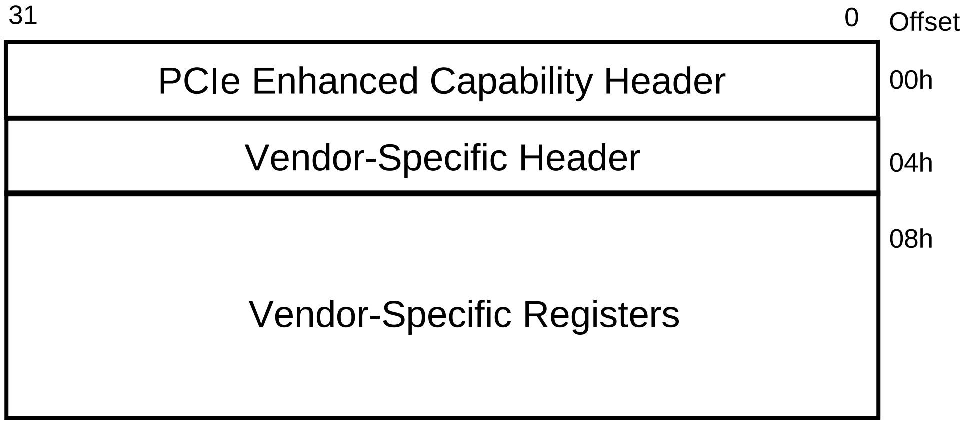
- The optional Multi-Function Virtual Channel (MFVC) capability register block is an extended capability required for PCI Express multi-function devices that support functionality beyond the default Traffic Class (TC0) over the default Virtual Channel (VC0). The MFVC capability structure must be present in the Extended Configuration Space of Function 0 of the multi-function device's Upstream Port. This MFVC capability structure controls Virtual Channel assignment at the PCI Express Upstream Port of the multi-function device, while a VC capability structure if present in a function controls the Virtual Channel assignment for that individual function. A multi-function device is permitted to have an MFVC Capability structure even if none of its functions have a VC Capability structure. However, an MFVC Capability structure is permitted only in Function 0 in the Upstream Port of a multi-function device.

Figure 24-4: Multi-Function Virtual Channel (MFVC) Extended Capability Register



- The Vendor-Specific Capability Structure is an optional extended capability that may be implemented by any PCI Express Function or RCRB. This allows PCI Express component vendors to use the extended capability mechanism to expose vendor-specific registers. A single PCI Express Function or RCRB is permitted to contain multiple Vendor-Specific Capability structures.

Figure 24-5: Vendor-Specific Extended Capability Register



- This optional RCRB Header Capability register block may be implemented in an RCRB to provide a Vendor ID and Device ID for the RCRB and to provide management of parameters that affect the behavior of Root Complex functionality associated with the RCRB.

Figure 24-6: RCRB Header Extended Capability Register

31	0	Offset
PCIe Enhanced Capability Header		00h
Device ID	Vendor ID	04h
RCRB Capabilities		08h
RCRB Control		0Ch
RsvdZ		10h

Clarifications:

- Multi-function devices that implement the Max_Payload_Size register can be configured to the same or different values across the various functions. A function that generates packets must have a data payload size that does not exceed the size specified in that function's Max_Payload_Size. Multi-function devices are encouraged at the very least to ensure that the packet size does not exceed the smallest Max_Payload_Size setting across all functions in the device. Software should not set the Max_Payload_Size in different functions to different values unless software is aware of the specific implementation.
- The Maximum Link Width register in the Link Capability register is allowed to indicate a Link width that is greater than the number of Lanes connected to a slot, adapter or component.
- Software must enable power to slots prior to reading the presence detect state for slots that do not implement a power controller.
- The Reference Clock Field of the VC Extended Capability Structure which specifies an encoding of 100ns applies to all ports that support time-based WRR for port arbitration.
- Port Arbitration related registers apply not only to Switch Ports, but also to Root Ports that support peer-to-peer traffic.