Compute Express Link™ (CXL™)

Engineering Change Notice to the Specification 2.0

July 2021

Error Isolation on CXL.mem and CXL.cache

C6124.0003 BN/FCURCI 39895222v1

CXL ENGINEERING CHANGE NOTICE

TITLE:	Error Isolation on CXL.mem and CXL.cache
DATE:	Introduced date (09/23/2020)
	Updated date (12/07/2020)
AFFECTED DOCUMENT:	CXL 2.0
SPONSOR:	Ishwar Agarwal, Microsoft

Part I

1. Summary of Functional Changes

This ECN defines a formal transaction timeout mechanism for Host initiated transactions on CXL.mem and CXL.cache. It defines capability, control, and status for a mechanism to isolate uncorrectable errors, including transaction timeout, on CXL.mem and CXL.cache at the CXL Root Port.

2. Benefits as a Result of the Changes

PCIe and CXL.io support Downstream Port Containment (DPC), which is intended to halt all traffic below a Downstream Port after an unmasked uncorrectable error is detected at or below the Port, avoiding the potential spread of any data corruption, and supporting Containment Error Recovery (CER), if implemented by software. This means that an uncorrectable error on a given PCIe Port may be recoverable, given a suitable software stack. Examples of such errors are surprise link down, completion time-outs and unexpected transaction layer errors.

CXL does not have any similar mechanisms for CXL.mem and CXL.cache, nor is it feasible to support the same type of error containment model, due to their interactions with system memory caching. The "isolation" mechanism defined for CXL.mem & CXL.cache protocols is designed to enable continued platform operation following uncorrectable errors with CXL.mem and/or CXL.cache transactions, which necessarily indicate corrupted system memory and/or corrupted system caching.

Isolation is usually triggered at a root port when a tracked CXL.mem or CXL.cache request times out. Isolation can also be triggered if the CXL link goes down due to any reason; for example, DPC. For each outstanding and subsequent Read transaction, the CXL Root Port synthesizes a synchronous exception response; e.g., returning poisoned data. For each outstanding and subsequent Write or Invalidate transaction, the CXL Root Port synthesizes a completion response and drops the data. Isolation mechanisms for CXL.mem and CXL.cache transactions are orthogonal.

The primary purpose of isolation is to complete pending and subsequent transactions associated with the isolated root port quickly with architected semantics once isolation is triggered. Since system memory and system caches must generally be assumed to be corrupted, software recovery generally relies on software identifying all software threads, VMs, containers, etc. whose system state might be corrupted, and then terminating them. Other software recovery mechanisms are also possible, and they are outside the scope of this specification.

3. **Assessment of the Impact**

This is an optional normative feature. It will have no impact on existing implementations. For implementations that do support it, the hardware changes are restricted to the Host only. No change is required for CXL switches or endpoint devices.

4. **Analysis of the Hardware Implications**

This is an optional normative feature. New hardware will be required in the Host to implement Isolation. For Hosts that choose to implement this, the primary hardware impact will be to track outstanding transactions on CXL.mem and CXL.cache.

5. **Analysis of the Software Implications**

This is an optional normative feature. To take advantage of this feature, software will be required to discover and enable it via new capability bits. This feature relies on software for terminating any threads, containers, VMs etc. that are impacted by the errors that triggered isolation. Further, this feature relies on software for recovery actions.

6. **Analysis of the Compliance and Test Implications**

This ECN does not repurpose any reserved bits and as such, does not impact existing C&I tests. New vendor-specific tests will be necessary to ensure components that implement this feature are compliant to the requirements listed in the spec.

Part II

Detailed Description of the change

Replace Section 12.3 to:

12.3 Isolation on CXL.mem and CXL.cache

Isolation on CXL.mem and CXL.cache is an optional normative capability of a CXL Root Port. Isolation on CXL.mem and CXL.cache halts traffic on the respective protocol. Further, once triggered, the Root Port synthesizes response for all pending and subsequent transactions on that protocol. This is further described in Section 12.3.1 and Section .12.3.2.

Two trigger mechanisms are defined by the specification

1. Link down – If a Root Port supports CXL.cache isolation and software enables CXL.cache isolation, a link down condition shall unconditionally trigger CXL.cache isolation. If a Root Port

supports CXL.mem isolation and software enables CXL.mem isolation, a link down condition shall unconditionally trigger CXL.mem isolation.

2. Transaction timeout – A Root Port that supports CXL.cache isolation, may be capable of being configured in such a way that a CXL.cache timeout triggers CXL.cache isolation. A Root Port that supports CXL.mem isolation, may be capable of being configured in such a way that a CXL.mem timeout triggers CXL.mem isolation.

Implementation Note: - Transaction Timeout Value settings for CXL.cache and CXL.mem: The system needs to make sure that timeouts are setup appropriately. For example, a timeout should not be so short that isolation is triggered due to a non-erroneous, long latency access to a CXL device. Software may need to temporarily disable the triggering of isolation upon timeout if one or more device is being transitioned to a state (e.g., firmware update) where it may violate the timeout.

The primary purpose of the isolation action is to complete pending and subsequent transactions associated with the isolated root port quickly with architected semantics once isolation is triggered. Since system memory and system caches must generally be assumed to be corrupted, software recovery generally relies on software identifying all software threads, VMs, containers, etc. whose system state might be corrupted, and then terminating them. Other software recovery mechanisms are also possible, and they are outside the scope of this specification.

A Root Port indicates support for Isolation by implementing CXL Timeout and Isolation Capability structure (Section 8.2.5.17). It contains the capability, control, and status bits for both Transaction Timeout and Isolation on both CXL.mem and CXL.cache. Both Timeout and Isolation are disabled by default and must be explicitly enabled by software for each protocol individually before they can be triggered. When Isolation is enabled for either CXL.mem or CXL.cache, software can optionally configure the Root Port to force a link down condition if the respective protocol enters into Isolation.

When Isolation is entered, the Root Port, if capable, signals an MSI/MSI-X or send an ERR_COR Message if enabled. Software may also choose to rely purely on mandatory synchronous exception handling (see Section 12.3.1 and Section 12.3.2). Software may read the CXL Timeout and Isolation Status register to determine if a Timeout or Isolation has occurred on CXL.mem and/or CXL.cache and if the Isolation was triggered due to a Timeout or due to a link down condition.

12.3.1 CXL.mem Transaction Layer Behavior During Isolation

This section specifies the CXL.mem Transaction Layer's behavior while the CXL Root Port is in Isolation.

The Root Port shall handle host requests that it would ordinarily map to (M2S) CXL.mem messages in the following manner.

- For each host request that would ordinarily be mapped to CXL.mem Req and RwD,
  - For Read transactions, the CXL Root Port synthesizes synchronous exception response. The specific mechanism of synchronous exception response is CXL Root Port implementation specific. An example of a synchronous exception response would be returning data with Poison.
  - For non-read transactions, the CXL Root Port synthesizes a response as appropriate. The specific mechanism of the synthesized response is implementation specific. An example would be returning a completion (NDR) for a write (RwD) transaction.

12.3.2 CXL.cache Transaction Layer Behavior During Isolation

This section specifies the CXL.cache Transaction Layer's behavior while the Root Port is in Isolation.

The Root Port shall handle host requests that would ordinarily be mapped to (H2D) CXL.cache messages in the following manner.

For each host snoop that would ordinarily be mapped to (H2D) CXL.cache request messages:

- If the host is tracking the device as a possible exclusive owner of the line, then data is treated as poison.

- Else if the host knows the device can only have a Shared or Invalid state for the line, then the device cache is considered Invalid (no data poisoning is needed).

  Implementation Note: Exclusive vs Shared/Invalid may be known based on internal state in the host.

  Timeout detection logic shall account for partial responses. For example, if the Root Port observes that the data is returned on D2H Data channel in a timely manner, but if no D2H Rsp was observed for a long enough time, the Root Port shall treat it as a CXL.cache timeout.

For each pending Pull that is mapped to H2D CXL.cache Response of type *WritePull* which expects a data return, the Root Port must treat the return data as poison.

In Section 8.2.5 Table 142, make the following change

Capability	ID	Highest version	Mandatory1	Not Permitted	Optional
..	…	…	…	…	…
CXL Timeout and Isolation Capability	9	1		P, D1, D2, LD, FMLD, UP1, USP, DSP	R

- 1. P- PCI Express device, D1 - CXL 1.1 Device, D2 - CXL 2.0 Device, LD - Logical Device, FMLD - Fabric Manager owned LD 0xFFFF, UP1 - CXL 1.1 Upstream Port RCRB, DP1 - CXL 1.1 Downstream Port RCRB, R - CXL 2.0 Root Port, USP - CXL Switch Upstream Port, DSP - CXL Switch Downstream Port

Insert new sub-section to Section 8.2.5.9 right below the current section 8.2.5.8 as follows:

8.2.5.9 CXL Timeout and Isolation Capability Header (Offset: Varies)

Bit Location	Attributes	Description
15:0	RO	**CXL_Capability_ID**: This defines the nature and format of the CXL Capability register. For the CXL Timeout and Isolation Capability Header register, this field shall be 0x9.
19:16	RO	**CXL_Capability_Version**: This defines the version number of CXL Capability structure present. For this version of the specification, this field must be 0x1.
31:20	RO	**CXL_Timeout_and_Isolation_Capability_Pointer**: This defines the offset of the CXL Timeout and Isolation Capability Structure relative to the beginning of CXL_Capability_Header register.

Insert new sub-section 8.2.5.17 to Section 8.2.5 right below the current section 8.2.5.15 as follows:

8.2.5.17   CXL Timeout and Isolation Capability Structure

Offset	Register Name
0x0	CXL Timeout and Isolation Capability Register
0x4	Reserved
0x8	CXL Timeout and Isolation Control Register
0xC	CXL Timeout and Isolation Status Register

8.2.5.17.1 CXL Timeout and Isolation Capability Register (Offset 0x0)

Bit Location	Register Description	Attributes
3:0	***CXL.mem Transaction Timeout Ranges Supported*** – This field indicates support for transaction timeout ranges on CXL.mem. Four time value ranges are defined: **Range A** Default range: 50us to 10ms. **Range B** 10ms to 250ms **Range C** 250ms to 4s **Range D** 4s to 64s Bits are set according to the table below to show timeout value ranges supported. **0000b** Transaction Timeout programming not supported – the function must implement a timeout value in the range 50us to 10ms. **0001b** Range A **0010b** Range B **0011b** Range A and B **0110b** Range B and C **0111b** Range A, B and C **1110b** Range B, C and D **1111b** Range A, B, C and D All other values are Reserved.	RO
4	***CXL.mem Transaction Timeout Supported*** – The value of 1b indicates support for CXL.mem Transaction Timeout mechanism.	RO
7:5	***Reserved***	RsvdP
11:8	***CXL.cache Transaction Timeout Ranges Supported*** – This field indicates support for transaction timeout ranges on CXL.cache. Four time value ranges are defined: **Range A** Default range: 50us to 10ms. **Range B** 10ms to 250ms **Range C** 250ms to 4s **Range D** 4s to 64s	RO

CXL 2.0 ECN

	Bits are set according to the table below to show timeout value ranges supported. **0000b** Transaction Timeout programming not supported – the function must implement a timeout value in the range 50us to 10ms. **0001b** Range A **0010b** Range B **0011b** Range A and B **0110b** Range B and C **0111b** Range A, B and C **1110b** Range B, C and D **1111b** Range A, B, C and D All other values are Reserved.	
12	*CXL.cache Transaction Timeout Supported* – The value of 1b indicates support for CXL.cache Transaction Timeout mechanism.	RO
15:13	*Reserved*	RsvdP
16	*CXL.mem Isolation Supported* – This field indicates support for Isolation on CXL.mem.	RO
17	*CXL.mem Isolation Link-Down Supported* - This field indicates support for triggering Link-Down on the CXL port if CXL.mem enters Isolation mode. This field can only be set to 1b if *CXL.mem Isolation Supported* is set to 1b.	RO
18	*CXL.cache Isolation Supported* – This field indicates support for Isolation on CXL.cache.	RO
19	*CXL.cache Isolation Link-Down Supported* - This field indicates support for triggering Link-Down on the CXL Root Port if CXL.cache enters Isolation mode. This field can only be set to 1b if *CXL.cache Isolation Supported* is set to 1b.	RO
24:20	*Reserved*	RsvdP
25	*Isolation ERR_COR Signaling Supported* – If Set, this bit indicates that the Root Port supports the ability to signal with ERR_COR when Isolation is triggered.	RO
26	*Isolation Interrupt Supported* – This field indicates support for signaling interrupt when Isolation is triggered.	RO
31:27	*Isolation Interrupt Message Number* - This field indicates which MSI/MSI-X vector is used for the interrupt message generated in association with the CXL Timeout and Isolation Capability structure. This field is valid only if Isolation Interrupt Supported is 1b. For MSI, the value in this field indicates the offset between the base Message Data and the interrupt message that is generated. Hardware is required to update this field so that it is correct if the number of MSI Messages assigned to the Function changes when software writes to the Multiple Message Enable field in the Message Control	RO

Register for MSI.

For MSI-X, the value in this field indicates which MSI-X Table entry is used to generate the interrupt message. The entry must be one of the first 32 entries even if the Function implements more than 32 entries. For a given MSI-X implementation, the entry must remain constant.

If both MSI and MSI-X are implemented, they are permitted to use different vectors, though software is permitted to enable only one mechanism at a time. If MSI-X is enabled, the value in this field must indicate the vector for MSI-X. If MSI is enabled or neither is enabled, the value in this field must indicate the vector for MSI. If software enables both MSI and MSI-X at the same time, the value in this field is undefined.

8.2.5.17.2 CXL Timeout and Isolation Control Register (Offset 0x8)

Bit Location	Register Description	Attributes
3:0	**CXL.mem Transaction Timeout Value** – In CXL Root Port Functions that support Transaction Timeout programmability, this field allows system software to modify the Transaction Timeout Value for CXL.mem. Functions that support Transaction Timeout programmability must support the values given below corresponding to the programmability ranges indicated in the CXL.mem Transaction Timeout Ranges Supported field. Defined encodings: **0000b** Default range: 50us to 10ms Values available if Range A (50us to 10ms) is supported. **0001b** 50us to 100us **0010b** 1ms to 10ms Values available if Range B (10ms to 250ms) is supported. **0101b** 16ms to 55ms **0110b** 65ms to 210ms Values available if Range C (250ms to 4s) is supported. **1001b** 260ms to 900ms **1010b** 1s to 3.5s Values available if Range D (4s to 64s) is supported. **1101b** 4s to 13s **1110b** 17s to 64s Values not defined above are Reserved. Software is permitted to change the value in this field at any time. For Requests already pending when the Transaction Timeout Value is changed, hardware is permitted to use either the new or the old	RW/RO

	value for the outstanding Requests and is permitted to base the start time for each Request either on when this value was changed or on when each request was issued. The default value for this field is 0000b.	
4	**_CXL.mem Transaction Timeout Enable_ –** When Set, this bit enables CXL.mem Transaction Timeout mechanism. Software is permitted to Set or Clear this bit at any time. When Set, the CXL.mem Transaction Timeout detection mechanism is enabled. If there are outstanding Transaction when the bit is set, it is permitted but not required for hardware to apply the completion timeout mechanism to the outstanding Transactions. If this is done, it is permitted to base the start time for each Transaction on either the time this bit was set or the time each Request was issued. This bit must be RW if the CXL.mem Transaction Timeout Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the CXL.mem Transaction Timeout Supported bit is Set. The default value for this bit is 0b.	RW/RO
7:5	**_Reserved_**	RsvdP
11:8	**_CXL.cache Transaction Timeout Value_** – In CXL Root Port Functions that support Transaction Timeout programmability, this field allows system software to modify the Transaction Timeout Value for CXL.cache. Functions that support Transaction Timeout programmability must support the values given below corresponding to the programmability ranges indicated in the CXL.cache Transaction Timeout Ranges Supported field. Defined encodings: **0000b** Default range: 50us to 10ms Values available if Range A (50us to 10ms) is supported. **0001b** 50us to 100us **0010b** 1ms to 10ms Values available if Range B (10ms to 250ms) is supported. **0101b** 16ms to 55ms **0110b** 65ms to 210ms Values available if Range C (250ms to 4s) is supported. **1001b** 260ms to 900ms **1010b** 1s to 3.5s Values available if Range D (4s to 64s) is supported. **1101b** 4s to 13s **1110b** 17s to 64s Values not defined above are Reserved.	RW/RO

CXL 2.0 ECN

	Software is permitted to change the value in this field at any time. For Requests already pending when the Transaction Timeout Value is changed, hardware is permitted to use either the new or the old value for the outstanding Requests and is permitted to base the start time for each Request either on when this value was changed or on when each request was issued. The default value for this field is 0000b.	
12	**CXL.cache Transaction Timeout Enable –** When Set, this bit enables CXL.cache Transaction Timeout mechanism. Software is permitted to Set or Clear this bit at any time. When Set, the CXL.cache Transaction Timeout detection mechanism is enabled. If there are outstanding Transaction when the bit is set, it is permitted but not required for hardware to apply the completion timeout mechanism to the outstanding Transactions. If this is done, it is permitted to base the start time for each Transaction on either the time this bit was set or the time each Request was issued. This bit must be RW if the CXL.cache Transaction Timeout Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the CXL.cache Transaction Timeout Supported bit is Set. The default value for this bit is 0b.	RW/RO
15:13	*Reserved*	RsvdP
16	**CXL.mem Isolation Enable** – This field allows System Software to enable CXL.mem Isolation actions. If this field is set, Isolation actions will be triggered if either a CXL.mem Transaction Timeout is detected or if the CXL link went down. This bit must be RW if the CXL.mem Isolation Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the CXL.mem Isolation Supported bit is Set. The software is required to quiesce the CXL.mem traffic passing through the Root Port when changing the state of this bit. If Software modifies this bit in presence of CXL.mem traffic, the results are undefined.	RW/RO
17	**CXL.mem Isolation Link-Down Enable** - This field allows System Software to trigger link down on the CXL Root Port if CXL.mem enters Isolation mode. This bit must be RW if the CXL.mem Isolation Link-Down Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the CXL.mem Isolation Link-Down Supported bit is Set.	RW/RO
18	**CXL.cache Isolation Enable** – This field allows System Software to enable CXL.cache Isolation actions. If this field is set, Isolation actions	RW/RO

	will be triggered if either a CXL.cache Transaction Timeout is detected or if the CXL link went down. This bit must be RW if the CXL.cache Isolation Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the CXL.cache Isolation Supported bit is Set. The software is required to quiesce the CXL.cache traffic passing through the Root Port when changing the state of this bit. If Software modifies this bit in presence of CXL.cache traffic, the results are undefined.	
19	**CXL.cache Isolation Link-Down Enable** - This field allows System Software to trigger link down on the CXL Root Port if CXL.cache enters Isolation mode. This bit must be RW if the CXL.cache Isolation Link-Down Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the CXL.cache Isolation Link-Down Supported bit is Set.	RW/RO
24:20	**Reserved**	RsvdP
25	**Isolation ERR_COR Signaling Enable** – When Set, this bit enables the sending of an ERR_COR Message to indicate Isolation has been triggered. Default value of this bit is 0b. This bit must be RW if the Isolation ERR_COR Signaling Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the Isolation ERR_COR Signaling Supported bit is Set.	RW/RO
26	**Isolation Interrupt Enable** – When Set, this bit enables the generation of an interrupt to indicate that Isolation has been triggered. Default value of this bit is 0b. This bit must be RW if the Isolation Interrupt Supported bit is Set; otherwise, it is permitted to be hardwired to 0b. Software must not Set this bit unless the Isolation Interrupt Supported bit is Set.	RW/RO
31:27	**Reserved**	RsvdP

8.2.5.17.3 CXL Timeout and Isolation Status Register (Offset 0xC)

Bit Location	Register Description	Attributes
0	**CXL.mem Transaction Timeout** – When Set, this indicates that a CXL.mem transaction timed out.	RW1CS/RsvdZ
3:0	**Reserved**	RsvdZ
4	**CXL.cache Transaction Timeout** – When Set, this indicates that a CXL.cache transaction timed out.	RW1CS/RsvdZ
7:5	**Reserved**	RsvdZ
8	**CXL.mem Isolation Status** – This field indicates that Isolation mode for CXL.mem was triggered. When this bit is set, CXL.mem is in	RW1CS/RsvdZ

	isolation mode and the link is forced to be down if CXL.mem Isolation Link-Down Enable is set. Software is permitted clear this bit as part of recovery actions regardless of the state of other status bits, after which, the CXL Root Port is no longer in Isolation mode for CXL.mem transactions. The link must transition through the link-down state before software can attempt re-enumeration and device recovery.	
9	*CXL.mem Isolation Link-Down Status* - This field indicates that Isolation Mode for CXL.mem was triggered due to link down.	RW1CS/RsvdZ
11:10	*Reserved*	RsvdZ
12	*CXL.cache Isolation Status* – This field indicates that Isolation mode for CXL.cache was triggered. When this bit is set, CXL.cache is in isolation mode and the link is forced to be down if CXL.cache Isolation Link-Down Enable is set. Software is permitted clear this bit as part of recovery actions, after which, the CXL Root Port is no longer in Isolation mode for CXL.cache transactions. The link must transition through the link-down state before software can attempt re-enumeration and device recovery.	RW1CS/RsvdZ
13	*CXL.cache Isolation Link-Down Status* - This field indicates that Isolation Mode for CXL.cache was triggered due to link down.	RW1CS/RsvdZ
14	*CXL RP Busy -* When either the CXL.mem Isolation Status bit or the CXL.cache Isolation Status bit is Set and this bit is Set, the Root Port is busy with internal activity that must complete before software is permitted to Clear the CXL.mem Isolation Status bit and the CXL.cache Isolation Status bit. If software violates this requirement, the behavior is undefined. This field is valid only when either the CXL.mem Isolation Status bit or the CXL.cache Isolation Status bit is Set; otherwise the value of this field is undefined. Default value of this bit is undefined.	RO/RsvdZ
31:15	*Reserved*	RsvdZ

CXL 2.0 ECN