

Implementing PCIe® Link Subdivision

White Paper

March 2025

Revision 1.0
822831

Revision History

Revision	Date	Comments
1.0	March 6, 2025	Initial release.

Contents

PCIe Express Link Subdivision Overview 5
Normative References 6
BIOS Manageability 7
Implementation Examples 8
Baseboard Management Controller (BMC) Manageability 10
Operating System Manageability 10
Conclusion 10

NOTE: ***This page intentionally left blank.***

PCIe Express Link Subdivision Overview

Link Subdivision, also known as bifurcation, is typically implemented by PCIe® CEM Slot and OCP 3.0 NIC adapters that have two or more PCIe devices. These devices typically consume individual PCIe lanes. For example, an Intel® E810-2CQDA2 100GbE dual port QSFP28 Adapter has two PCIe devices, each consuming 8 PCIe lanes. When this adapter is used, the BIOS configures a 16 PCIe lane slot into a 2 x8 PCIe lane bus. Other Intel adapters such as the Intel® Ethernet Network Adapter E610-XT4 for OCP3.0 and Intel® Ethernet Network Adapter E610-IT4 require the BIOS to configure two (2) PCIe Links, each having four (4) PCIe lanes.

The Peripheral Component Interconnect Special Interest Group (PCI-SIG®) and Open Compute Project specifications prescribe methods to describe any number of link configurations, with any variation of port counts, and with fully flexible variable widths, as described in the following documents:

- *PCI CEM Specification*, Chapter 13, "Link Subdivision"
- *PCI Base Specification*, Chapter 12, "Architectural Out-of-Band Management"

Link subdivided adapters advertise their PCIe lane divisions as described in:

- *PCI Base Specification*, Chapter 12, Section 12.6.1, "FRU Information Device Requirements"

  The PCI-SIG defined MultiRecord format supplied by the adapter is described in:

  - Section 12.6.2, "FRU Information Format"
- *OCP NIC 3.0 Specification*, Section 3.5, "PCIe Bifurcation Mechanism"

To retrieve the PCI-SIG MultiRecord, or the OCP3.0 NIC MultiRecord, the Baseboard Management Controller (BMC), or some other platform device, accesses the FRU Information device, which complies with the *IPMI FRU Specification*, and fetches the Link Subdivision option for the adapter. The Link Subdivision information is communicated to the platform BIOS in a method defined outside of this White Paper.

The BIOS uses the Link Subdivision record information to configure the PCIe root port lanes, physically connected to the subdivided adapter slot. This information is also used to create the DMTF SMBIOS Type 9 System Slot Record, which describes the slot adapter link subdivision to the operating system and applications.

The BIOS also communicates the slot adapter link subdivision to the BMC in a method defined outside of this White Paper. Typically, the SMBIOS Type 9 System Slot Record is the format communicated to the BMC.

The operating system parses the SMBIOS Type 9 System Slot Record to align the PCIe bus enumeration to a device instance in a slot.

The intent of this document is to inform the reader of how the industry standard specifications, when combined as a unified view, can inform the BIOS, BMC and Operating System the PCIe Link topology to a PCIe connector (or slot).

Normative References

The following reference documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

- *IPMI Platform Management FRU Information Storage Definition, Revision 1.3, March 24, 2015* (***IPMI FRU Specification***)

  https://www.intel.com/content/dam/www/public/us/en/documents/specification-updates/ipmi-platform-mgt-fru-info-storage-def-v1-0-rev-1-3-spec-update.pdf

- *PCI-SIG®, PCI Express® Base Specification Revision 6.2, January 25, 2024* (***PCI Base Specification***)

  http://www.pcisig.com/specifications/

- *PCI-SIG®, PCI Express® Card Electromechanical Specification, Revision 6.0, October 23, 2024* (***PCI CEM Specification***)

  http://www.pcisig.com/specifications/

- *OCP NIC 3.0 Design Specification, Version 1.5.0* (***OCP NIC 3.0 Specification***)

  https://www.opencompute.org/wiki/Server/NIC

- *DMTF DSP0134 System Management BIOS (SMBIOS) Reference Specification 3.7* (***SMBIOS Specification***)

  https://www.dmtf.org/standards/smbios

- *DMTF DSP0246, Platform Level Data Model (PLDM) for SMBIOS Data Transfer Specification 1.0.1* (***DSP0246 SMBIOS Data Transfer Specification***)

  http://www.dmtf.org/standards/published_documents/DSP0246_1.0.1.pdf

For the Intel® Xeon® Processor Scalable Family, the following specification provides the methods to query the PCIe root port for the status and PCI Link/Lane ordering. All these specifications can be searched for and downloaded at the link below and may require an Intel Confidential Non-Disclosure Agreement (CNDA) account.

- *4th Gen Intel® Xeon® Processor Scalable Family Registers Specification*, Rev. 2.4, Doc. No. 612246 (**Intel® Xeon® Processor Scalable Family**)

  https://www.intel.com/content/www/us/en/resources-documentation/developer.html

BIOS Manageability

The BIOS uses Link Subdivision record information to configure the PCIe root port lanes physically connected to the subdivided adapter slot. This information is also used to create the SMBIOS Type 9 System Slot Record, describing the slot adapter Link Subdivision to the operating system and applications. The BIOS also communicates the slot adapter Link Subdivision to the BMC in a method defined outside of this White Paper. Typically, the format communicated to the BMC is the SMBIOS Type 9 System Slot Record.

In some architectures, the BIOS may not have control over the PCIe Bus enumeration (assignment) which, in the event of a PCIe lane reversal, may lead to inconsistent PCIe device enumeration for a Link Subdivided adapter. In the case of PCIe Lane Reversal, typically the low PCIe Bus number of a PCIe Link Subdivided adapter will be assigned to the Higher physical PCIe Lanes and the higher PCIe Bus Number will be assigned to the lower physical PCIe Lanes of the PCIe Slot Link Subdivision.

The BIOS, BMC or Operating System shall not make any assumptions about the relationship between the PCIe Bus Number and the PCIe Device Instance when the PCIe Link to an add-in adapter is subdivided.

The SMBIOS Type 9 System Slot Record has provisions to describe a PCIe Slot, such as a PCIe CEM slot or an OCP 3.0 NIC slot, including the Link Subdivision to the slot.

For a PCIe Adapter that requires Link Subdivision, the BIOS uses the information from *PCI Base Specification*, Section 12.6.2, “FRU Information Format” to complete the SMBIOS Type 9 Slot Record. Each slot connected to the PCIe bus shall have a single SMBIOS Type 9 Slot Record, no matter how many PCIe buses are assigned to the slot.

Implementation Examples

"Normative References" shows an example of a SMBIOS Type 9 Slot Record for an adapter in a PCIe CEM slot. This record shows a full data lane width of 16 lanes, subdivided into 2 PCIe buses, each with a data width of 8 lanes. The lowest bus number is always assigned to the base device and is directly correlated to the lowest physical (lane) pins on the adapter. The adapter (slot) physical placement rule applies to both normal lane enumeration and PCIe lane reversal enumeration, to define consistent enumeration for the BIOS, BMC, and operating system. The SMBIOS Type 9 Slot Record is also used for OCP3.0 NIC slots.

Offset	Name	Value	Description
04h	Slot Designation	PCIe Slot 2	Slot name as a STRING
06h	Slot Data Bus Width	0Dh	Designates a x16 PCIe Lane
08h	Slot Length	04h	Long Length (x16 PCIe CEM Slot)
0Ch	Slot Characteristics 2	0Ch	SMBus Supported, Link Subdivision Supported
0Dh	Segment Group Number (Base)	00h	Single PCIe Hierarchy (Segment) System
0Fh	Bus Number (Base)	12h	Assigned to lowest adapter physical lanes
10h	Device/Function Number (Base)	00h	Typically, only list Device and PF0
11h	Bus Data Width (base)	08h	8 lanes, connected to lowest physical lanes
12h	Peer Grouping Count	01h	Only 1 Peer Group
13h	Segment Group Number (Peer)	00h	Single PCIe Hierarchy (Segment) System
15h	Bus Number (Peer)	14h	Assigned to Highest adapter physical lanes
16h	Device/Function Number (Peer)	00h	Typically, only list Device and PF0
17h	Bus Data Width (Peer)	08h	8 lanes, connected to Highest physical lanes
18h	Slot Information	Varies	Platform Dependent
19h	Slot Physical Width	0Dh	Designates Physical Slot Width (x16)

Table 1. Example SMBIOS Type 9 Slot Record x16 PCIe Lane Link Subdivision

Table 2 shows the SMBIOS Type 9 Slot Record from the *SMBIOS Specification*. This table is from the *DMTF DSP0134 System Management BIOS (SMBIOS) Reference Specification 3.7*. See "Normative References".

A PCIe Link Subdivided adapter shall be a BASE Device Entry, comprised of:

- Segment Group (Base)
- Bus Number (Base)
- Device/Function Number (Base)
- Data Bus Width (Base)

There shall also be a single Peer Device, comprised of:

- Segment Group (Peer)
- Bus Number (Peer)
- Device/Function Number (Peer)
- Data Bus Width (Peer)

Offset	Spec. Version	Name	Length	Value	Description
02h	2.0+	Handle	WORD	Varies	
04h	2.0+	Slot Designation	BYTE	STRING	String number for reference designation EXAMPLE: 'PCI-1', 0
05h	2.0+	Slot Type	BYTE	ENUM	See 7.10.1.
06h	2.0+	Slot Data Bus Width	BYTE	ENUM	See 7.10.2.
07h	2.0+	Current Usage	BYTE	ENUM	See 7.10.3.
08h	2.0+	Slot Length	BYTE	ENUM	See 7.10.4.
09h	2.0+	Slot ID	WORD	Varies	See 7.10.5.
0Bh	2.0+	Slot Characteristics 1	BYTE	Bit Field	See 7.10.6.
0Ch	2.1+	Slot Characteristics 2	BYTE	Bit Field	See 7.10.7.
0Dh	2.6+	Segment Group Number (Base)	WORD	Varies	See 7.10.8.
0Fh	2.6+	Bus Number (Base)	BYTE	Varies	See 7.10.9.
10h	2.6+	Device/Function Number (Base)	BYTE	Bit Field	Bits 7:3 - device number Bits 2:0 - function number See 7.10.8.
11h	3.2	Data Bus Width (Base)	BYTE	Varies	Indicate electrical bus width of base Segment/Bus/Device/Function/Width
12h	3.2	Peer (S/B/D/F/Width) grouping count (n)	BYTE	Varies	Number of peer Segment/Bus/Device/Function/Width groups that follow. Zero if no peer groups
13h	3.2	Peer (S/B/D/F/Width) groups	5*n BYTES	Varies	Peer Segment/Bus/Device/Function/Width present in the slot; see 7.10.9. This field is absent (not empty) if there are no peer groups (n=0).
13h + 5*n	3.4	Slot Information	BYTE	Varies	See 7.10.10.
14h + 5*n	3.4	Slot Physical Width	BYTE	Varies	See 7.10.11.
15h + 5*n	3.4	Slot Pitch	WORD	Varies	See 7.10.12.
17h + 5*n	3.5	Slot Height	BYTE	Varies	See 7.10.13.

Table 2. SMBIOS Type 9 Slot Record Reference

Baseboard Management Controller (BMC) Manageability

The Baseboard Management Controller (BMC) shall be aware of any adapter that requires PCIe Link Subdivision. This specification recommends that the BMC receives the SMBIOS Type 9 Slot Records from the BIOS to provide a platform view of the platform slots, consistent with both the BIOS and the operating system view.

The BIOS to BMC communication protocol is typically proprietary, but there is a defined industry standard method. Refer to the DSP0246 SMBIOS Data Transfer Specification.

The SMBIOS Type 9 Slot Record is used by the BMC to enumerate device instances on a PCIe Link Subdivided adapter, starting with the base device and following with the peer devices in the order listed in the record.

Operating System Manageability

The operating system uses the *SMBIOS Specification* to enumerate the device instances and provide the adapter (slot) information. The operating system should enumerate the physical PCIe devices with their associated PCIe functions on the same adapter, in the order listed in the SMBIOS Type 9 Slot Record.

Conclusion

The PCI-SIG® and DMTF have standardized the discovery and management of adapters that require PCIe Link Subdivision. The referenced standards provide a mechanism to align the adapter requirements across the BIOS, BMC, and operating systems, which provides a consistent user facing view of the platform topology.

Future Intel adapters, such as the Intel® E610 Ethernet Network Adapters, will support the *PCI Base Specification*, Section 12.6.1, "FRU Information Device Requirements" with a self-describing record detailing the PCIe Link Subdivision requirements. Currently shipping Intel® Ethernet Network Adapters for OCP NIC 3.0 adapters have compliance to OCP 3.0 Specifications, version 1.2.x for PCIe Link Subdivision and have specific IPMI FRU Specification MultiRecords. Future Intel® Ethernet Network Adapters for OCP NIC 3.0 Adapters will update compliance to the OCP 3.0 NIC Specifications, version 1.5.x.

For currently shipping adapters, such as the Intel® E810-2CQDA2 100GbE 2P QSFP28 Adapter, the BIOS needs a manual PCIe Link Subdivision method to configure the slot and then generate the SMBIOS Type 9 Slot Record.



LEGAL

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

This document (and any related software) is Intel copyrighted material, and your use is governed by the express license under which it is provided to you. Unless the license provides otherwise, you may not use, modify, copy, publish, distribute, disclose or transmit this document (and related materials) without Intel's prior written permission. This document (and related materials) is provided as is, with no express or implied warranties, other than those that are expressly stated in the license.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors which may cause deviations from published specifications.

Copies of documents that are referenced in this document can be obtained by visiting the Intel Resource and Documentation Center.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

Other names and brands may be claimed as the property of others.

© 2025 Intel Corporation. All rights reserved.