OPEN
Compute Project

k

PCIe Extended Connectivity Requirements Document

Revision 0.07c
July 31th, 2023

Contributing Company – Contact Person
Seagate Technologies – Mohamad El-Batal
Ayar-Labs – LK Bhupathi
Astera-Labs – Richard Ward

Revision History

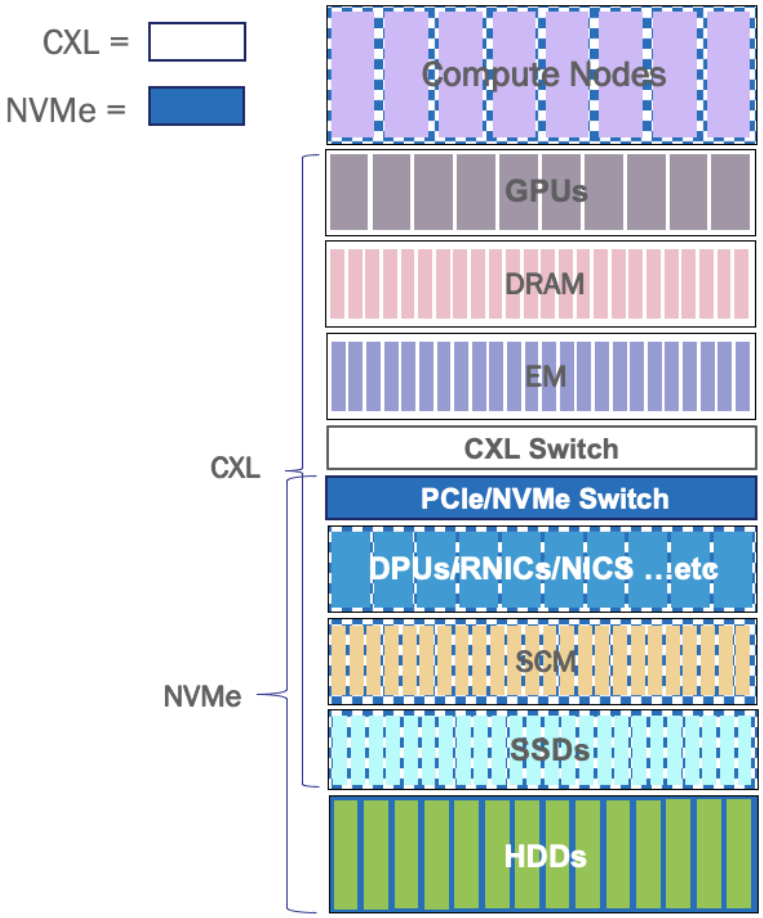
Version	Date	Details
0.1	1/3/2023	First draft – Mohamad El-Batal
0.11	1/9/2023	Added Top level rack diagram and Out-Of-Band Security/Management
0.12	2/12/2023	Added channel model diagrams for DAC and AEC
0.13	2/13/2023	Added Rack level Diagrams for CXL Scenarios
0.14	2/23/2023	Added cabling suggestions for Sideband signals & changed title from External Connectivity to Extended Connectivity
0.15	5/10/2023	Updated diagrams
0.20	6/16/2023	Added some more content to the AEC section
0.5	6/29/2023	Various edits and corrections
0.7a	7/17/2023	More edits and corrections
0.7b	7/31/2023	Added Sustainability section and participant's names
0.7c	8/3/2023	Editorial comments and corrections

Table of Contents

1 OCP

1.1 Scope

To document the various industry usage-model scenarios and requirements for disaggregated NVMe and CXL inter-connected devices including Switch, Compute, Accelerator, Networking, Memory, and Storage modules using PCIe-5.0 & PCIe-6.0 external Direct Attached Connection (DAC), Active Electrical Connection (AEC), Direct Optical Connection (DOC), and Active Optical Connection (AOC) Cables. We shall not specify any specific detailed connector or cabling implementations specification at this time; however, a future industry standards committee will be formed by a team of willing members aiming to materialize the requirement output of this Workstream into a complete standardized cabling and connector specification that leverages existing connector and cabling designs and provide the most optimal solution level Total Cost of Ownership (TCO) results and the best industry wide economy of scale benefits. This workstream shall explore and document only the most relevant external PCIe physical layer connectivity scenarios that are currently in plan or are foreseen for future datacenter and enterprise applications in conjunction with NVMe and CXL protocol based external connectivity applications that are intra-rack and/or inter-rack in scale.



1.2 Acknowledgement

The OCP External PCIe Connectivity Requirements Document was created with collaboration from many OCP member companies and facilitated by the OCP External PCIe Connectivity Workstream under the OCP Server Project Workgroup. The OCP Extended PCIe Connectivity Requirements Workstream would like to acknowledge the following member companies for their contributions to the Extended PCIe Connectivity Requirements Document Rev1.0:

OCP Member Company
Amphenol Corporation - Sam Kocsis
IIVI/Coherent Corporation - Deyu Zhou
TE Connectivity Corporation - Brian Costello
Molex Connectivity Corporation - Muria
Credosemi - Sheng Huang
Samtec Connectivity corporation - jignesh.shah
Broadcom Limited - Bruce Trunk
Intel Corp. - Kay Xiao
Marvell Semiconductor, Inc. - Henry Chen
Google - Pierre-Luc Michael , Sunny Saini
Microsoft - Shane kavanagh
Ghiasi Quantum - Ali Ghiasi
Seagate Technology Corporation - Alvin Cox
Super Micro Corporation - Jessica T. , Paul M.
Western Digital Corporation - Barrett Edwards
Astera-Labs Corporation - Richard Ward
Serial Cables Corporation - Justin Mutschler
Ayar Labs Corporation - LK Bhupathi
Enfabrica - shimon@enfabrica.net
Siamak Tavallaei (OCP Server Project IC Rep.)

2 Compliance with OCP Tenets

2.1 Openness

The OCP External PCIe Connectivity Requirements Document was developed in dialogue with the OCP Server, Datacenter, Memory and Storage Project membership and the resulting requirements document represents the initial requirement set for this new category of external PCIe physical layer connectivity solutions. This ecosystem is nascent, so we are confident that this document will evolve over time, and that this requirements document represents the wisdom of the collective ecosystem on the scenarios and requirements for creating NVMe and/or CXL based solutions over external PCIe physical layer connectivity in the near future datacenter and enterprise applications.

2.2 Sustainability

This PCIe Extended connectivity initiative will enable the physical disaggregation of hardware components and allow for superior scaleout architectures that are composable in nature, thus the minimization of stranded resources and overprovisioning (eliminating unreachable and underutilized resources) in the modern datacenter, by starting small and expanding on a temporal basis depending on the datacenter requirements. This architecture also enables reuse and re-deployment of hardware infrastructure across multiple platforms and multiple hardware generations. While not prohibited, this architecture is to eliminate the need for using other distance enabling physical interfaces like Ethernet or SAS in order to achieve scalable disaggregation.

2.3 Efficiency

The efficiency goals of this workstream are to identify the key scenarios and requirements that connector and cabling manufacturers across the industry can build external PCIe physical layer connectivity solutions that follow a consistent and unified set of requirements, vetted by vendors and customers. Architecturally, such external PCIe connectivity infrastructure enables datacenter and enterprise solution designers to take advantage of using the same connectivity infrastructure and drive the best economy of scale TCO benefits to the NVMe and CXL connectivity industry. Additionally, it is not a preset requirement to define unified connectivity requirements for both NVMe and CXL, but to explore both protocol solution scenarios separately and then explore any potential intersecting solution requirements along with the pros and cons of such possible unified solutions if any. Solution level Total Cost of Ownership (TCO) optimization is king, and any decision to unify requirements must pass the TCO analysis checkpoint relative to the Value gained (ToV.)

2.4 Impact

The standardization of External PCIe physical layer connectivity for NVMe and CXL will have a huge impact on the overall datacenter and enterprise ecosystems. It shall extend the once considered internal server-centric PCIe converged bus to enabling the PCIe physical layer to become the ideal

high-bandwidth and ultra-low-Latency optimized disaggregation interface of choice for future NVMe and CXL protocol interconnected compute, acceleration, memory, and storage solutions.

2.5 Scale

The scale of this new interface connectivity ecosystem will be governed by the PCIe physical layer attributes at PCIe-5.0 and PCIe-6.0 differently; however, the scale limits will be different relative to each of the two protocols.

3 References

From https://www.pcisig.com:
- PCI Express® Base Specification, Revision 6.0
- PCI Express® Base Specification, Revision 5.0
- ECN - Component Measurement and Authentication (CMA) (June 25, 2020)

From https://nvmexpress.org/:
- NVM Express® Base Specification revision 2.0b
- NVM Express Management Interface Specification, Revision 1.2b
- NVM Express NVM Command Set Specification, Revision 1.0b
- NVM Express Zoned Namespace Command Set Specification, Revision 1.1b
- NVM Express Key Value Command Set Specification, Revision 1.0b
- NVM Express NVMe over PCIe Transport Specification, Revision 1.0b

From https://nvmexpress.org/:
- NVM Express® Base Specification revision 2.0b
- NVM Express Management Interface Specification, Revision 1.2b
- NVM Express NVM Command Set Specification, Revision 1.0b
- NVM Express Zoned Namespace Command Set Specification, Revision 1.1b
- NVM Express Key Value Command Set Specification, Revision 1.0b
- NVM Express NVMe over PCIe Transport Specification, Revision 1.0b

From https://www.opencompute.org/:
- Datacenter NVMe® SSD Specification, Revision 2.0
- Storage NVMe® HDD Specification, Revision 0.7

From https://trustedcomputinggroup.org/:
- TCG Storage Security Subsystem Class: Opal, Version 2.02, Revision 1.0
- TCG Storage Interface Interactions Specification (SIIS), Version 1.11
- TCG DICE Attestation Architecture, Version 1.00

From https://dmtf.org
- DSP0238 Management Component Transport Protocol (MCTP) PCIe VDM Transport Binding Specification
- DSP0274 Security Protocol and Data Model (SPDM) Specification, Version 1.2.0

From https://osfpmsa.org/
- Specification for OSFP, Octal Small Form Factor Pluggable Module, Rev 5.0
- Specification for OSFP-XD,Octal Small Form Factor eXtra Dense Pluggable Module, Rev 1.0

From https://www.qsfp-dd.com/
- QSFP-DD/QSFP-DD800/QSFP112 Hardware Specification for QSFP Double Density 8X and QSFP 4X Pluggable Transceivers, Rev 6.3

4 Extended PCIe Connectivity Scenarios for CXL

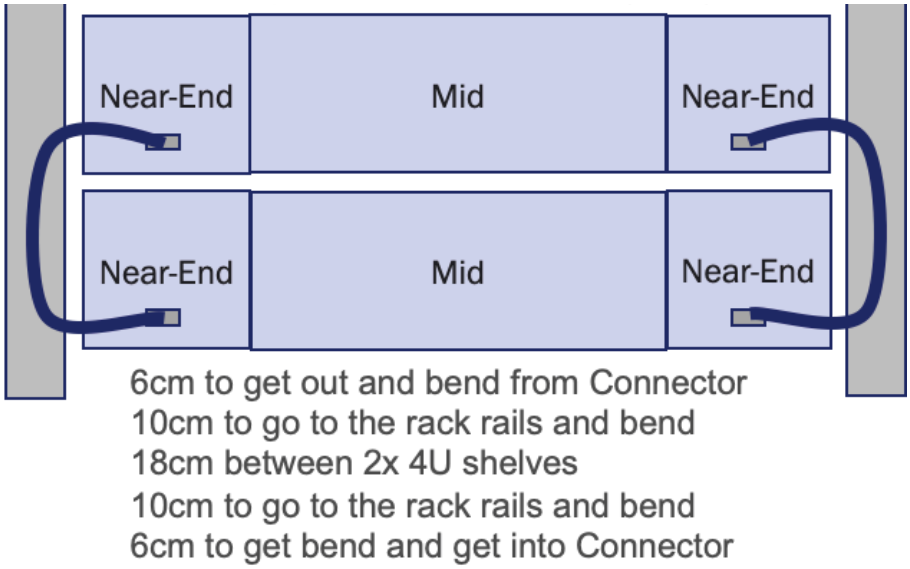
4.1 U-Heights of Various Related Chassis

- 1U Servers and/or Switches
- 2U Servers, GPUs blades, EDSFF Memory or Storage
- 3U PCIe CEM cards, NICs, RNICs, DPUs, IPUs …etc.
- 4U Enterprise Servers or High-Density Storage
- 4U High-Density Storage Enclosure
- 5U Potential 19” Half-width chassis vertically mounted solutions
- 6U Potential 21” Half-width chassis vertically mounted solutions

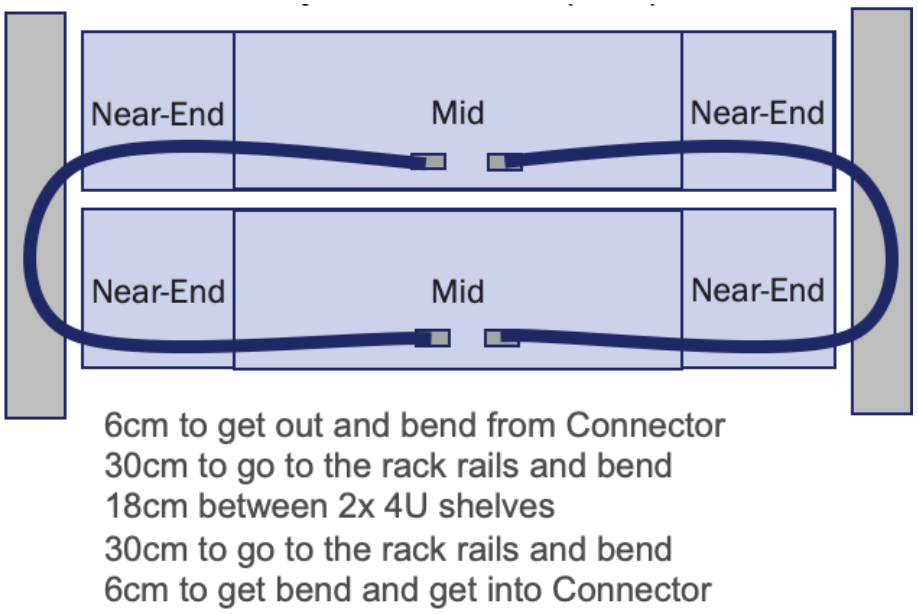
An enclosure with Cable Management Arm (CMA) would require a minimum of 1m longer cable than a standard chassis. All above cable length scenarios in section-4 are analyzed without CMAs due to CXL latency budget concerns.

4.2 Rack Level Configuration Considered

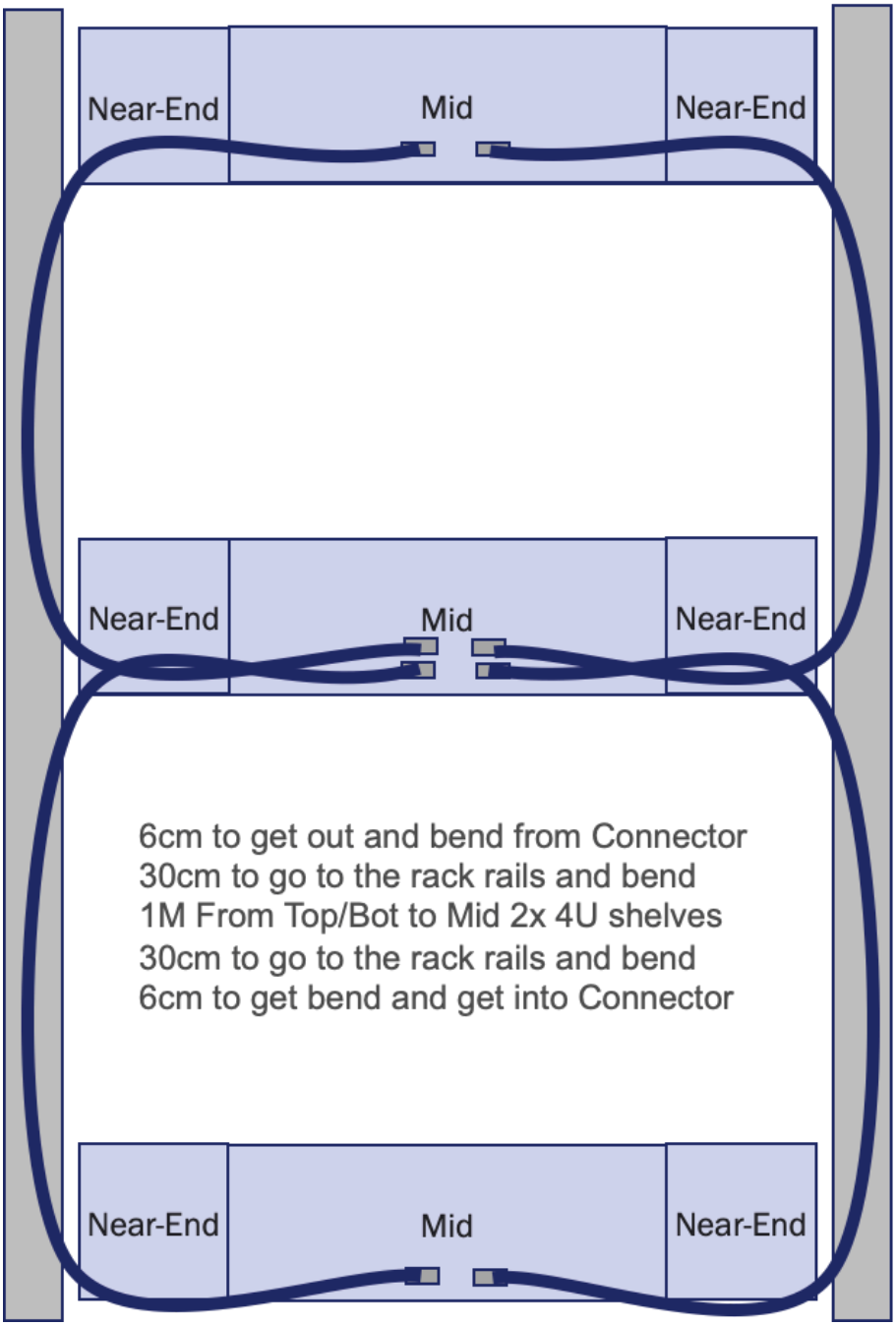
The cable shall come in different widths and lengths to allow for various connectivity scenarios.



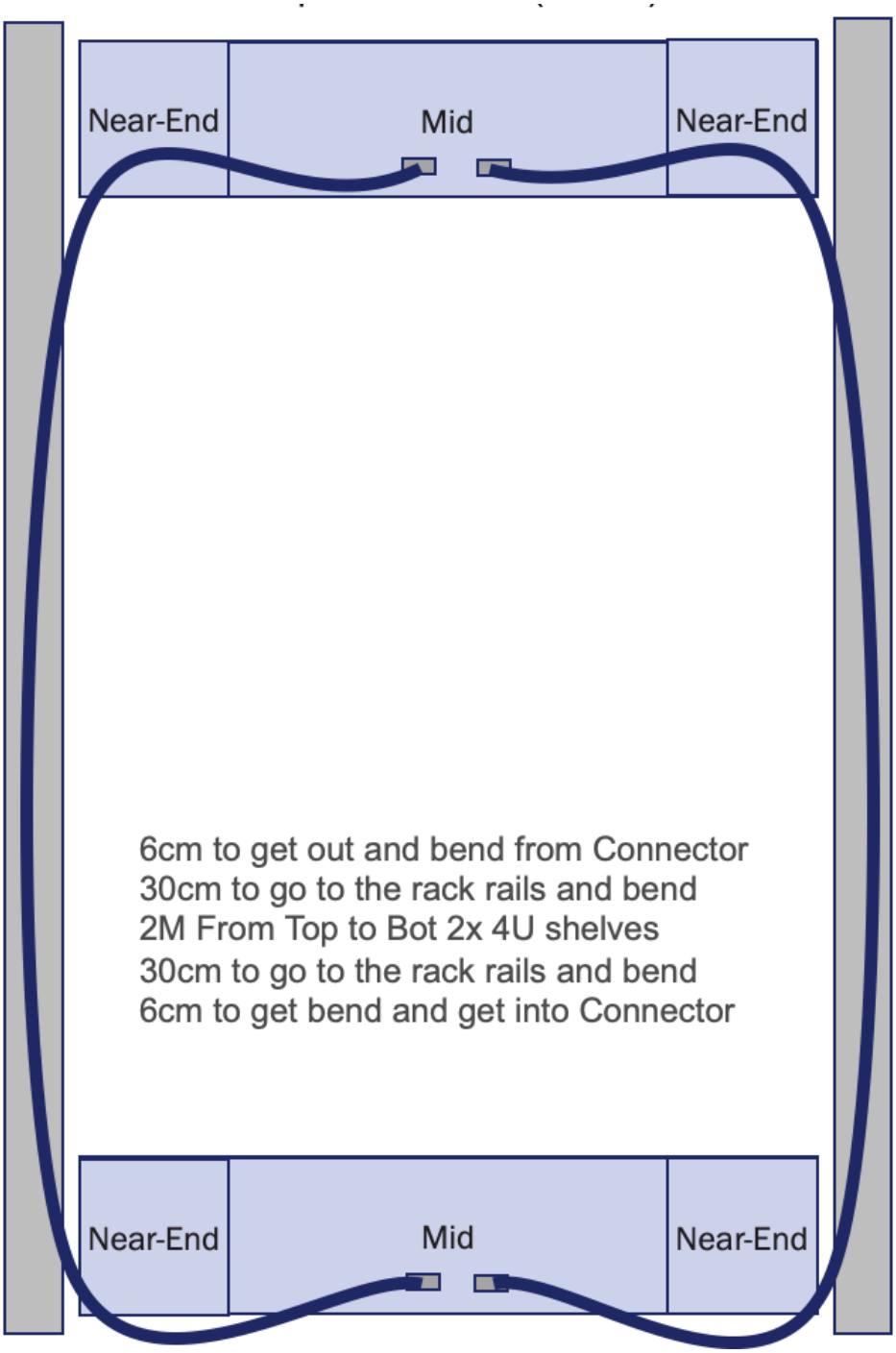
0.5m from Near-End-Shelf (NES) to adjacent Near-End-Shelf (NES) of 1U to 4U chassis heights. For >4U chassis a 1M cable may be required.



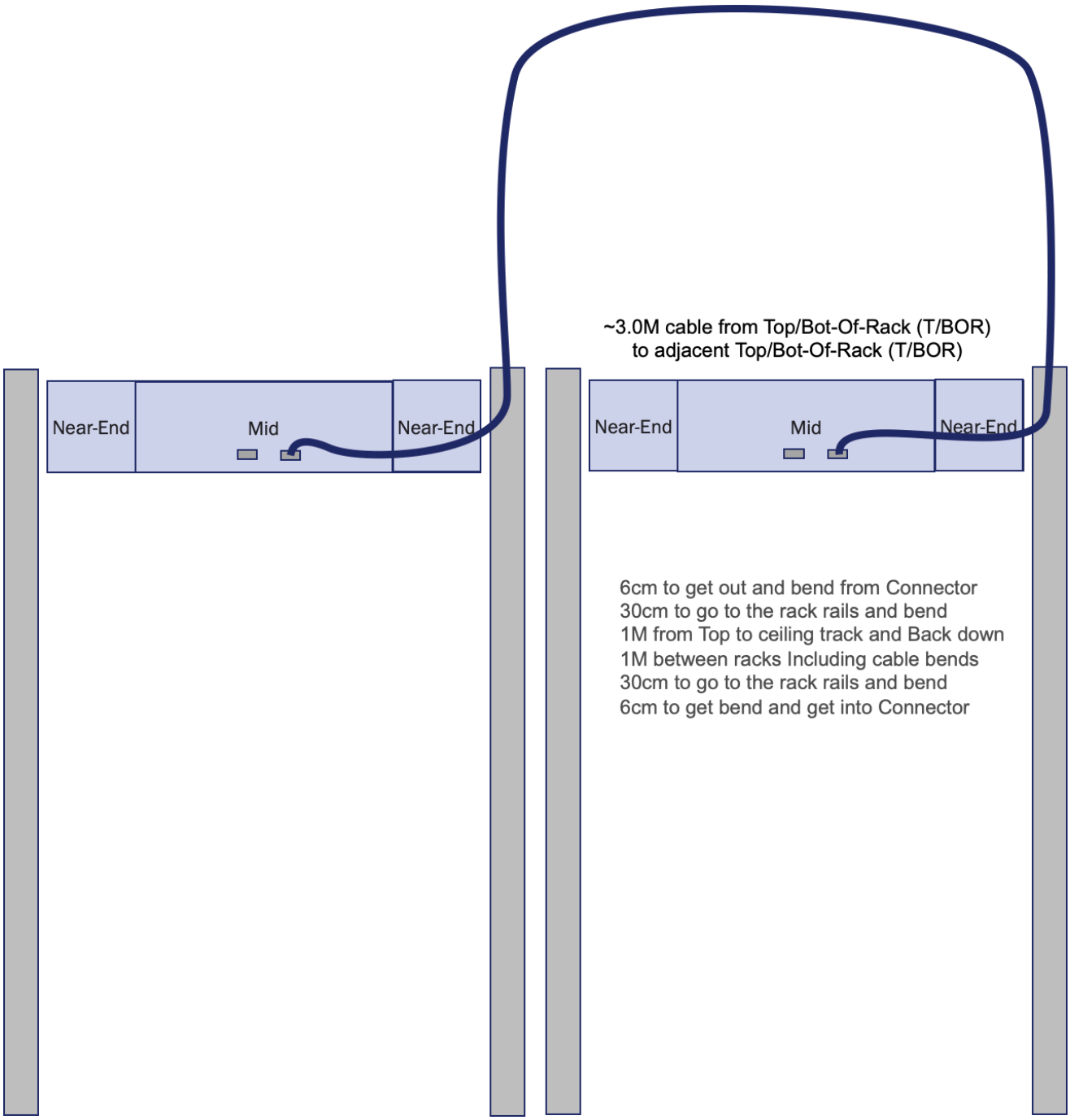
1.0m from Mid-Shelf (MS) to adjacent Mid-Shelf (MS) should be sufficient for up to 6U Shelves



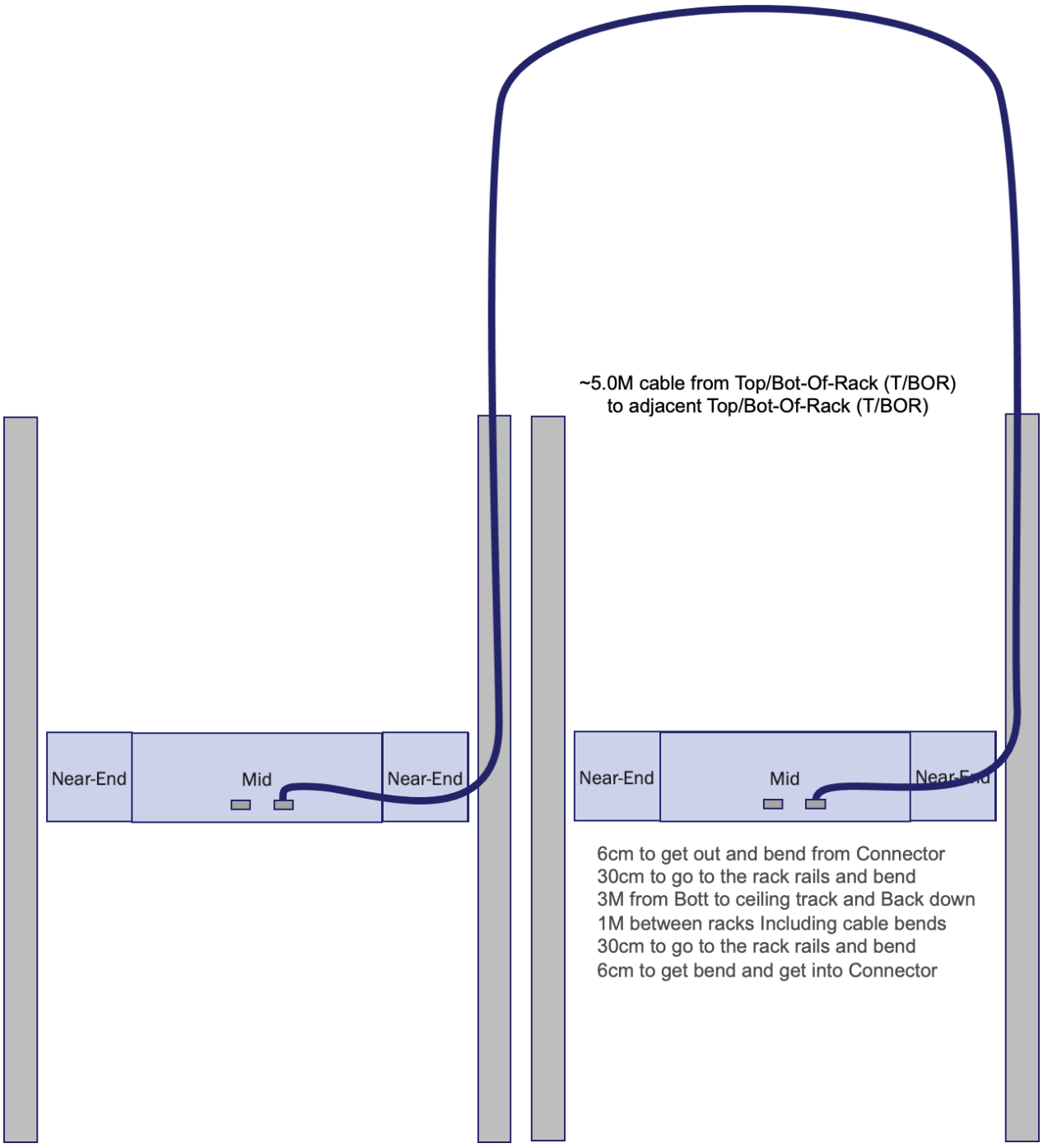
2.0m from Mid-Of-Rack (MOR) to Top/Bott-Of-Rack (T/BOR)



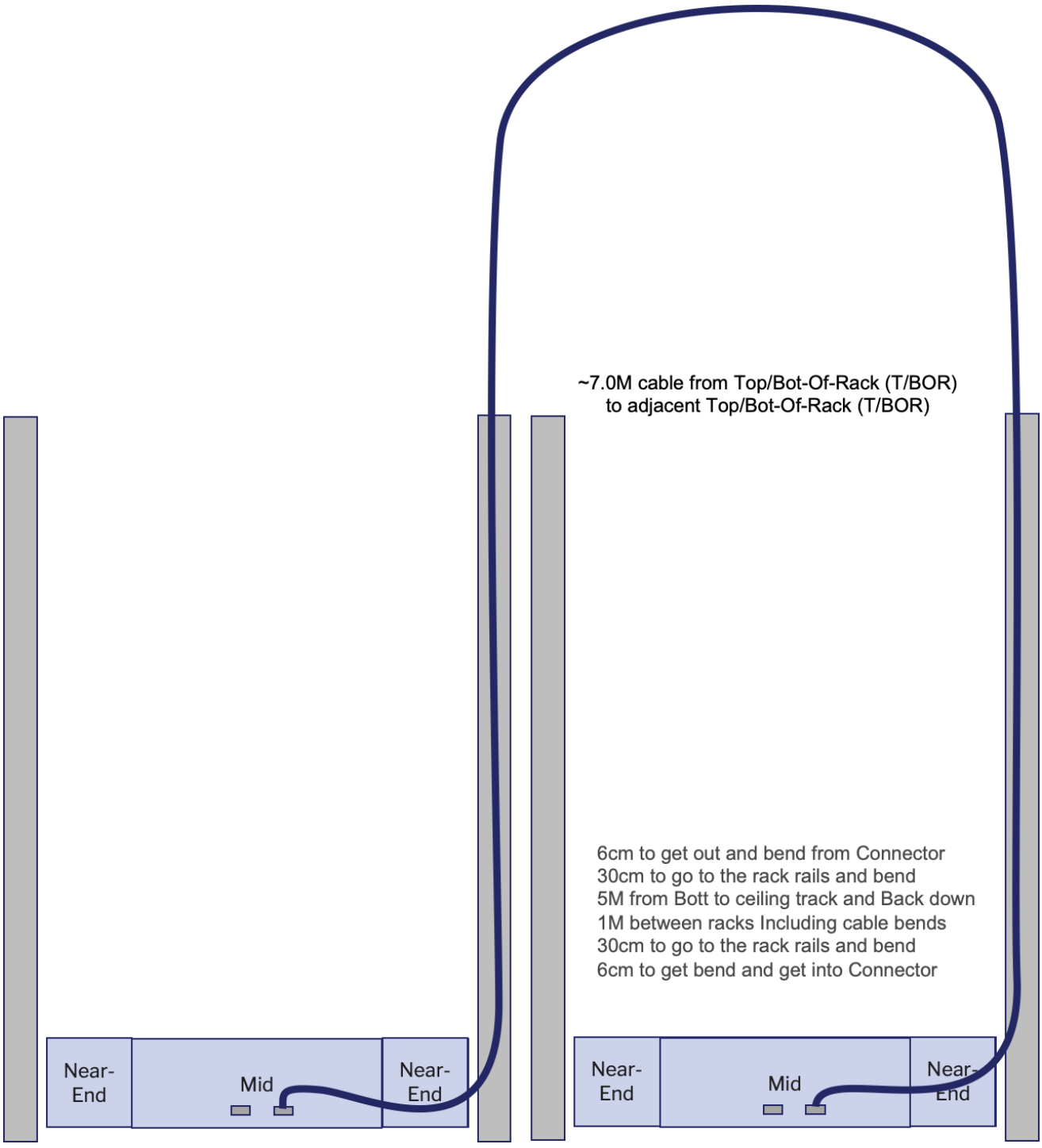
3.0m from Bott/Top-Of-Rack (B/TOR) to Top/Bott-Of-Rack (T/BOR)



3.0m from Top-Of-Rack (TOR) to adjacent Top-Of-Rack (TOR) w/ Ceiling Tracks
3.0m from Bott-Of-Rack (BOR) to adjacent Bott-Of-Rack (BOR) w/ Subfloor Tracks



5.0m from Mid-Of-Rack (MOR) to adjacent Mid-Of-Rack (MOR) w/ Ceiling of Subfloor Tracks
5.0m from Top-Of-Rack (TOR) to adjacent Bott-Of-Rack (BOR) w/ Ceiling of Subfloor Tracks
5.0m from Bott-Of-Rack (BOR) to adjacent Top-Of-Rack (TOR) w/ Ceiling of Subfloor Tracks



7.0m from Top-Of-Rack (TOR) to adjacent Top-Of-Rack (TOR) w/ Subfloor Tracks
7.0m from Bott-Of-Rack (BOR) to adjacent Bot-Of-Rack (TOR) w/ Ceiling Tracks

4.3 Compute Cluster

Req ID	Speed	Config/Width	Lengths
Scenario-1	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-2	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m

4.4 Compute to Accelerator

Req ID	Speed	Config/Width	Lengths
Scenario-3	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-4	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m

4.5 Accelerator Cluster

Req ID	Speed	Config/Width	Lengths
Scenario-4	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-5	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m

4.6 Compute to Memory

Req ID	Speed	Config/Width	Lengths
Scenario-6	PCIe-5.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-7	PCIe-5.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-8	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-9	PCIe-6.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-10	PCIe-6.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-11	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m

4.7 Accelerator to Memory

Req ID	Speed	Config/Width	Lengths
Scenario-12	PCIe-5.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-13	PCIe-5.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-14	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-15	PCIe-6.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-16	PCIe-6.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-17	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m

4.8 Memory Cluster

Req ID	Speed	Config/Width	Lengths
Scenario-18	PCIe-5.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-19	PCIe-5.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-20	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-21	PCIe-6.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-22	PCIe-6.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m
Scenario-23	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m & 7.0m

5 Extended PCIe Connectivity Scenarios for NVMe

The cable shall come in different lengths to allow for various connectivity scenarios:

- 0.5m from Near-End-Shelf (NES) (w/o CMA) to adjacent Near-End-Shelf (NES) (w/o CMA)
- 1.0m from Mid-End-Shelf (MES) (w/o CMA) to adjacent Mid-End-Shelf (MES) (w/o CMA)
- 2.0m from Near-End-Rack (TOR) (w/o CMA) to adjacent Near-End-Shelf (NES) (w/ CMA)
- 2.0m from Mid-Of-Rack (MOR) (w/o CMA) to Top/Bott-Of-Rack (T/BOR) (w/o CMA)
- 3.0m from Bott-Of-Rack (BOR) (w/o CMA) to Top-Of-Rack (TOR) (w/o CMA)
- 3.0m from Top-Of-Rack (TOR) (w/o CMA) to adjacent Top-Of-Rack (TOR) (w/o CMA)
- 5.0m from Top-Of-Rack (TOR) (w/o CMA) to adjacent Top-Of-Rack (TOR) (w CMA)
- 5.0m from Mid-Of-Rack (MOR) (w/o CMA) to adjacent Mid-Of-Rack (MOR) (w/o CMA)
- 7.0m from Bott-Of-Rack (BOR) (w/o CMA) to adjacent Bott-Of-Rack (BOR) (w/o CMA)
- 7.0m from Mid-Of-Rack (MOR) (w/o CMA) to adjacent Mid-Of-Rack (MOR) (w CMA)
- 10.0m from Bott-Of-Rack (BOR) (w/o CMA) to adjacent Bott-Of-Rack (BOR) (w/ CMA)
- 10.0m from Bott-Of-Rack (BOR) (w/ CMA) to adjacent Bott-Of-Rack (BOR) (w/ CMA)

An enclosure with CMA would require a minimum of 1m longer cable than a standard chassis. Some of the above NVMe cable length scenarios in section-5 are analyzed with CMAs since the extra cable latency is less of a concern.

5.1 Compute to Storage

Req ID	Speed	Config/Width	Lengths
Scenario-24	PCIe-5.0	1x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-25	PCIe-5.0	2x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-26	PCIe-5.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-27	PCIe-5.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-28	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-29	PCIe-6.0	1x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-30	PCIe-6.0	2x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-31	PCIe-6.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-32	PCIe-6.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-33	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m

5.2 Accelerator to Storage

Req ID	Speed	Config/Width	Lengths
Scenario-34	PCIe-5.0	1x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-35	PCIe-5.0	2x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-36	PCIe-5.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-37	PCIe-5.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-38	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-39	PCIe-6.0	1x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-40	PCIe-6.0	2x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-41	PCIe-6.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-42	PCIe-6.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-43	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m

5.3 Storage Cluster

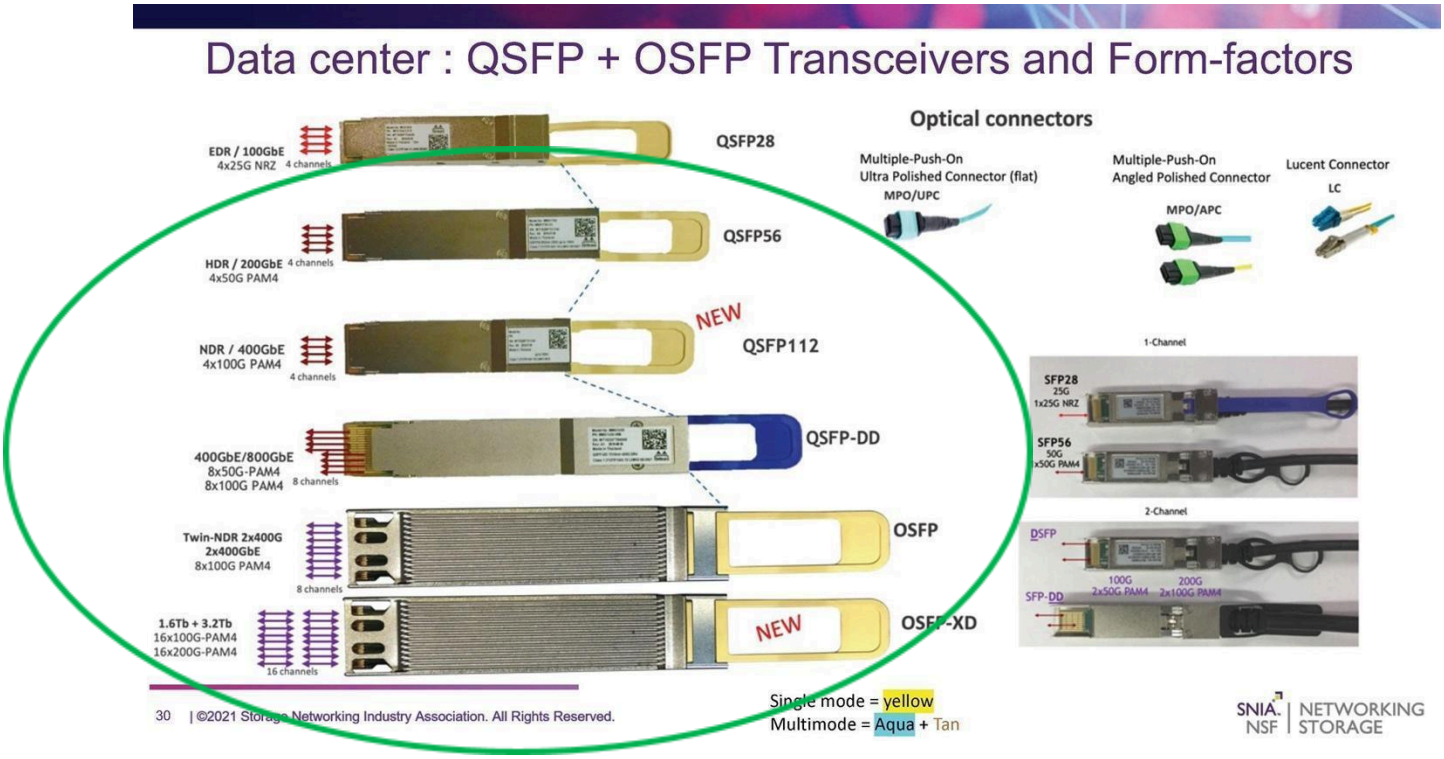
Req ID	Speed	Config/Width	Lengths
Scenario-44	PCIe-5.0	1x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-45	PCIe-5.0	2x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-46	PCIe-5.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-47	PCIe-5.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-48	PCIe-5.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-49	PCIe-6.0	1x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-50	PCIe-6.0	2x4	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-51	PCIe-6.0	1x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-52	PCIe-6.0	2x8	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m
Scenario-53	PCIe-6.0	1x16	0.5m, 1.0m, 2.0m, 3.0m, 5.0m, 7.0m & 10.0m

6 Extended Connectivity Types

6.1 Common Extended Connectivity Options

For PCIe-5.0 x4 we would need smaller devices to fit within a QSFP module PCBA,
For PCIe-5.0 x8 these package sizes can fit inside a QSFP-DD connector module PCBA.
For PCIe-5.0 x16 these packages would need to fit inside the OSFP-XD PCB. The x16 standard retimer package size requires a 90-degree rotation to fit inside the OSFP-XD form factor, or the use of non-standard packaging.

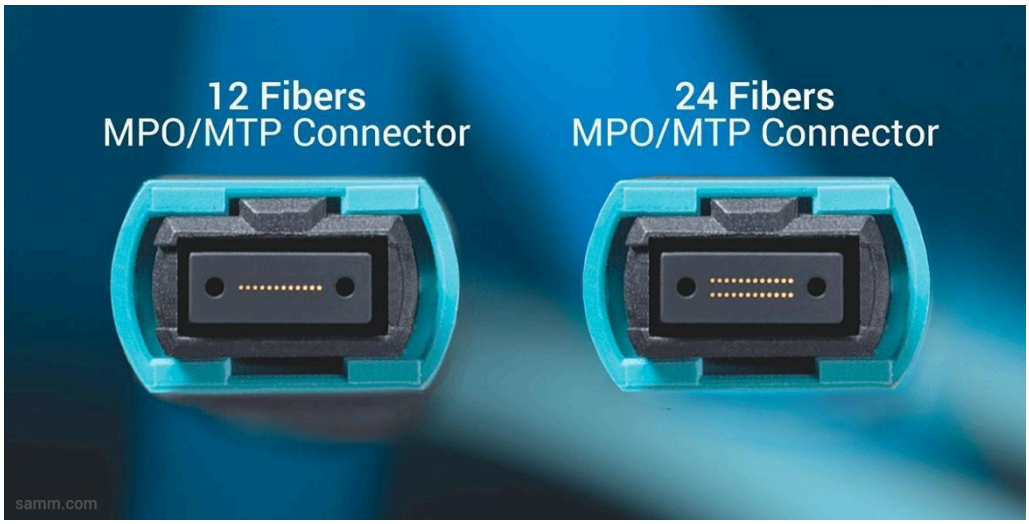
Module form factors under consideration for the PCIe NVMe & CXL in common between DAC, AEC and AOC are QSFP, QSFP-DD, OSFP and OSFP-XD:



Source:
https://www.snia.org/sites/default/files/ESF/Next-generation-Interconnects-the-Critical-Importance-of-Cables-and-Connectors.pdf

Additional proposed connector form-factor by Molex for DAC and AEC is the CDFP-stye2: https://members.snia.org/document/dl/27465

Although the standard x16 retimer package fits without rotation within the CDFP-style2 x16 connector PCBA; however, the dual PCBA aspect of the CDFP connector architecture does not lend itself to the current retimer specification: https://members.pcisig.com/wg/PCI-SIG/document/17251 that requires a single PCBA application. Furthermore, any optical AOC solution would also require a single PCBA connector architecture.



Source: https://www.samm.com/mpo-mtp-frequently-asked-questions

The above MPO connector is optimal for DOC solutions. Other connector options are available in the market, and we will leave it up to the specification team to pick the final optimal connector implementation solution.

6.2 Direct Attached Connection (DAC)

DAC cables must be completely passive on data-path signals but must contain an SMBus target controller and signals to provide Vital Product Data (VPD) information for the cable on both ends, but not used to pass communication across the cable to identify the endpoint or root-port. Adequate power must be provided to power the VPD controller on each end of the cable separately and independently. Power is not intended to be propagated from one connector end to another. It is proposed that the PCIe module control follows Cable Modem to Customer Premise Equipment Interface (CMIS) functionality for the module, otherwise we'll need to define our own set of register pages and functions (for example for FW upgrade) into the OCP specifications.

Out-of-Band management between devices on extended connectivity solutions shall not require the use of sideband signals like I2C or I3C, which adds more wires across the cable thus more overall cost. The physical PERST# line shall be optional and could be replaced by either Hot-Reset transport commands to reset the PCIe controller, or through a separate side-band management communication via I2C, I3C, USB, Ethernet …etc commands that would internally reset the various controllers that are required to be initialized. A DAC cable must either provide full Common Clock, PERST# and sideband connectivity or none of the above.

Such Out-of-Band (control-plane) management communication could make use of standard PCIe out-of-band transports that are available to us from various Baseboard Management Controllers (BMCs) devices with Management Component Transport Protocol (MCTP) support. Firmware downloads and Reset functions can also be done using MCTP over DOE or over VDM transports commands. Security attestation could also make use of Security Protocol and Data model (SPDM) over PCIe Data Object Exchange (DOE) or over MCTP over Vendor Defined Messages (VDM) transport layer.

Alternatively, users can make use of other sideband cabled interfaces between connected devices over Ethernet or USB to enable Security, Out-of-band Management, Firmware download or Reset functions.

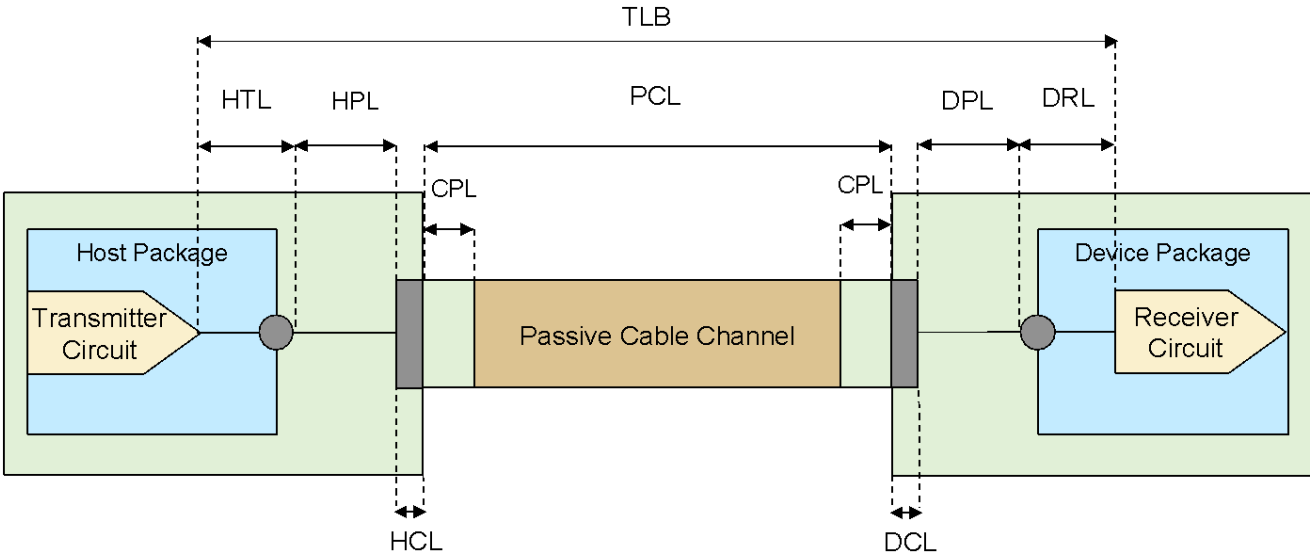
DAC cables shall require host and endpoint support for SRIS with Spread Spectrum Clocking (SSC) and SRNS with standard single frequency clocking both as mandatory functions in order to save on the added cabling cost and EMI reduction.

Alternatively, a shared synchronization clock can be connected between Root and Endpoint devices over the same sideband control-plane management interface or a separate clock synchronization cable if desired by the architecture.

DAC cables must allow for standard PCIe physical layer signal requirements including connector discontinuities and both PCB routes on both the Source and Device ends.

DAC cables shall support PCIe port bifurcation in SRIS/SRNS mod. All additional optional SMbus PCIe differential REFCLK(+/-) and PERST# signals are not required.

PCIe-Gen5 Total Loss Budget (TLB) = 36dB => 10^{-12} BER
PCIe-Gen6 Total Loss Budget (TLB) = 32dB => 10^{-6} BER



TLB	Total Loss Budget
PCL	Passive Cable Loss
HTL	Host Transmitter Loss
HPL	Host PCB Loss
CPL	Connector PCB Loss
DPL	Device PCB Loss
DRL	Device Receiver Loss
HCL	Host Connector Loss
DCL	Device Connector Loss

Cable AWG	IL*(db/m) @16GHz 100-Ohm	IL*(db/m) @16GHz 85-Ohm
26	3.6	4.1
28	4.3	5.2
30	5.3	6.4
32	7.2	8.3

Cable Loss db/m per gauge for 100-Ohm and 85-ohm Twinax provided as a reference by Amphenol via Sam Kocsis based on American Wire Gauge standards for wire conductor size. Bulk cable performance is ultimately a function of cable construction and process technology. Impedances can be adjusted to optimize loss performance based on the constraints of the form factor requirements.

Example viability test for 1-meter DAC cable in PCIe-5.0:

Cable AWG	IL* (db/m) @16GHz (PCIe-5.0 NRZ 36 db)	HTL(db)	HCL(db)	CPL(db)	DRL(db)	HPL(db)	DPL(db)
30	5.3	9	1	1	4	7.6	7.6
32	7.2	9	1	1	4	6.6	6.6

Example viability test for 1-meter DAC cable in PCIe-6.0:

Cable AWG	IL* (db/m) @16GHz (PCIe-6.0 PAM4 32db)	HTL(db)	HCL(db)	CPL(db)	DRL(db)	HPL(db)	DPL(db)
30	5.3	8	1	1	4	6.6	6.6
32	7.2	8	1	1	4	5.6	5.6

The Printed Circuit Board (PCB) dB budget loss depends greatly on the dielectric material used, copper trace width and weight, and the system designer must tune their solution specification to achieve the required balance in the above equation not to exceed the 36dB budget of the specific PCIe generation targeted.

There is more to signal integrity than the loss budget, since many factors play a big role in how long of a cable can be used including the trace routing and impedance matching as well as any reflections or crosstalk. The system designer must qualify the complete end to end solution.

The Passive Cable Channel dB loss budget also depends greatly on the cable gauge and type used. The DAC cable provider must specify their loss per distance as well as their connector's loss.

DAC cables should use specific Micro-Twinax cables with the above gauges that are optimal/acceptable across the full link. Depending on the cable characteristics used the system designer must calculate the total loss and keep it under the total loss budget.

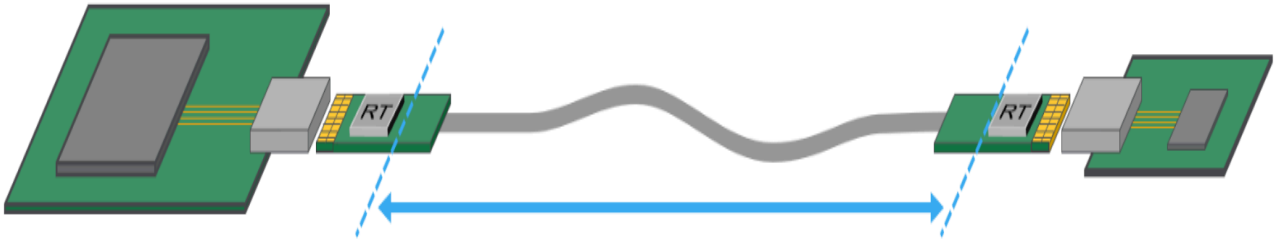
DAC cables can optionally support different connector width on each end to allow for Octopus cable for 1 x8 to 2 x4, and 1 x16 to 2 x8 or 4 x4 configurations with PCIe bifurcated connections.

Req ID	DAC	# Ports	Lengths Meters	Bend Radius
CNNCT-1	1x4	1	Up to 1	5x CD
CNNCT-2	1x8	1	Up to 1	5x CD
CNNCT-3	1x16	1	Up to 1	5x CD
CNNCT-4	x8(2x4)	2	Up to 1	5x CD
CNNCT-5	x16(2x8)	2	Up to 1	5x CD
CNNCT-6	x16(4x4)	4	Up to 1	5x CD
CNNCT-7	x8 to 2x4	2	Up to 1	5x CD
CNNCT-8	x16 to 2x8	2	Up to 1	5x CD
CNNCT-9	x16 to 4x4	4	Up to 1	5x CD

DAC Copper Cable Bend Radius - 5x the outer diameter of the cable jacket

6.3 Active Electrical Connection (AEC)

AEC cables support Re-Timer devices on both end modules, forming an Active-Active connection type.



Each Re-Timer that is added into the path resets the PCIe link budget, allowing for an extra 36dB channel loss (PCIe-5.0 for 1x10-12 BER) or 32dB channel loss (PCIe-6.0 for 1x10-6 BER) in the system, with equivalent reach extension.

AEC cables shall contain an SMBus target controller and signals to provide Vital Product Data (VPD) information for the cable on both connector ends. The SMbus controllers shall not be used to pass communication across the cable to identify the endpoint or root-port. Adequate power must be provided to power the VPD controller on each end of the cable separately and independently. Power shall not be propagated from one connector end to another. It is proposed that the PCIe module control follows CMIS functionality for the module, otherwise a new definition of register pages and functions (for example for FW upgrade) must be created.

Since we expect several Data Links to interconnect external Chassis, to eliminate duplication and reduce individual interconnect cost, we expect any sideband signal for reset, clock, and OoB (out-of-band) management be on a separate cable. Out-of-Band management between devices on extended connectivity solutions shall not require the use of sideband SMBus signals, which adds more wires across the cable thus more overall cost. The physical PERST# line shall not be required and can be replaced by either Hot-Reset transport commands to reset the PCIe controller, or through side-band management communication via I2C, I3C, USB, Ethernet …etc commands that would internally reset the various controllers that are required to be initialized. An AEC cable must either provide full Common Clock, PERST# and sideband connectivity or none of the above as currently defined by the PCI-SIG on SRIS/SRNS capable implementations.

Such Out-of-Band (control-plane) management communication could make use of standard PCIe out-of-band transports that are available to us from various Baseboard Management Controllers (BMCs) devices with Management Component Transport Protocol (MCTP) support. Firmware downloads and Reset functions can also be done using MCTP over DOE or over VDM transports commands. Security attestation could also make use of Security Protocol and Data model (SPDM) over PCIe Data Object Exchange (DOE) or over MCTP over Vendor Defined Messages (VDM) transport layer.

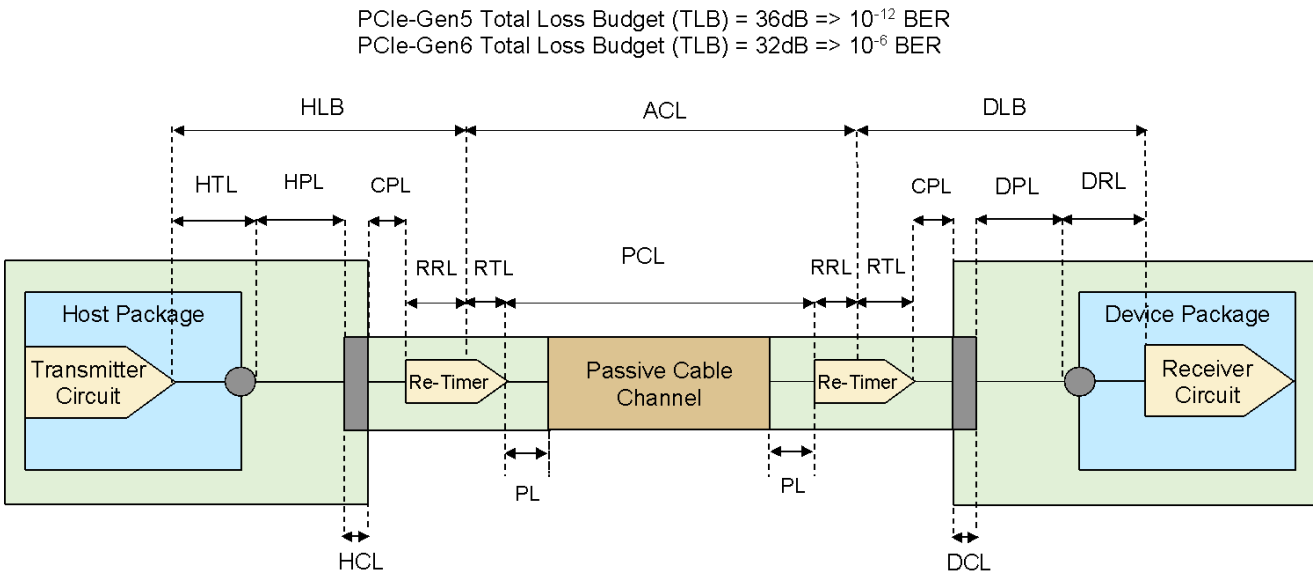
Alternatively, users can make use of other sideband cabled interfaces between connected devices over Ethernet or USB to enable Security, Out-of-band Management, Firmware download or Reset functions.

AEC cables shall require host and endpoint support for SRIS with Spread Spectrum Clocking (SSC) and SRNS with standard single frequency clocking both as mandatory functions in order to save on cabling cost and EMI reduction.

Alternatively, a shared synchronization clock can be connected between Root and Endpoint devices over the same sideband control-plane management interface or on a separate clock synchronization cable if desired by the architecture.

AEC cables must allow for standard PCIe physical layer signal requirement including connector discontinuities and both PCB routes on both the Source and Device ends.

AEC cables shall support PCIe port bifurcation in SRIS/SRNS mod. All additional optional SMbus, PCIe differential REFCLK(+/-) and PERST# signals are not required.



TLB	Total Loss Budget
HLB	Host Loss Budget
DLB	Device Loss Budget
ACL	Active Cable Loss
PCL	Passive Cable Loss
HTL	Host Transmitter Loss

HPL	Host PCB Loss
CPL	Connector PCB Loss
PL	PCB Loss
RRL	Retimer Receiver Loss
RTL	Retimer Transmitter Loss
DPL	Device PCB Loss
DRL	Device Receiver Loss
HCL	Host Connector Loss
DCL	Device Connector Loss

Example viability test for 3-meter AEC cable in PCIe-5.0:

Cable AWG	IL* (db/m) @16GHz (PCIe-5.0 NRZ 36db)	RTL(db)	2x PL(db)	RRL(db)	Margin (db)
30	16	4	2	4	10
32	22	4	2	4	4

Example viability test for 3-meter AEC cable in PCIe-6.0:

Cable AWG	IL* (db/m) @16GHz (PCIe-6.0 PAM4 32db)	RTL(db)	2x PL(db)	RRL	Margin (db)
30	16	4	2	4	6
32	22	4	2	4	0

With the additional link budget from including retimers, the maximum length of the cable will depend on the line-side Package and PCB loss and the gauge of the cable (AWG).

The Active Cable Channel dB-loss budget also depends on the cable gauge and type used. The AEC cable vendor shall provide the HCL/DCL and CPL losses to calculate Root and Device link margins to the Retimer IC.

AEC cables can optionally support different connector widths on each end to allow for Octopus cable for 1 x8 to 2 x4, and 1 x16 to 2 x8 or 4 x4 configurations with PCIe bifurcated connections.

Req ID	AEC	# Ports	Lengths Meters	Bend Radius
CNNCT-10	1x4	1	Up to 3	5x CD
CNNCT-11	1x8	1	Up to 3	5x CD
CNNCT-12	1x16	1	Up to 3	5x CD
CNNCT-13	x8(2x4)	2	Up to 3	5x CD
CNNCT-14	x16(2x8)	2	Up to 3	5x CD
CNNCT-15	x16(4x4)	4	Up to 3	5x CD
CNNCT-16	x8 to 2x4	2	Up to 3	5x CD
CNNCT-17	x16 to 2x8	2	Up to 3	5x CD
CNNCT-18	x16 to 4x4	4	Up to 3	5x CD

AEC Copper Cable Bend Radius - 5x the outer diameter of the cable jacket.

The PCI-SIG defines the PCIe-4.0 Retimer Specification, here:
https://members.pcisig.com/wg/PCI-SIG/document/17251

The PCIe-5.0 & PCIe-6.0 Retimer Specification are under development.

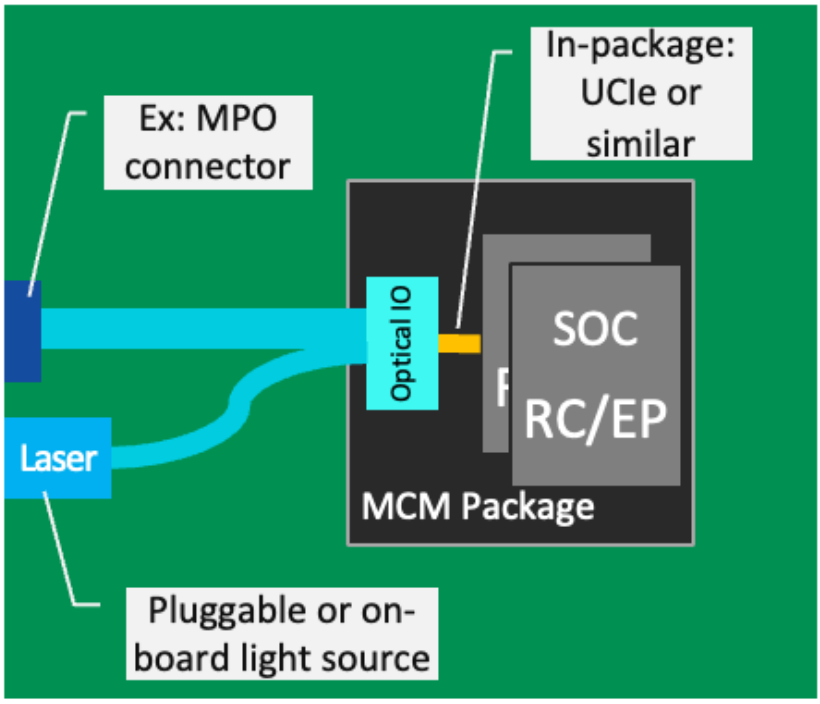
Rate	# of Lanes	Max Power/Lane/End Input + Output (Watts)	Total Package Power in Watts	Package Size (mm)
PCIe-5.0	1 x 4	1.2	Up to 5	tbd
PCIe-5.0	1 x 8	1.2	Up to 10	8.5 x 13.4
PCIe-5.0	1 x 16	1.2	Up to 20	8.9 x 22.8
PCIe-6.0	1 x 4	1.2	Up to 5	tbd
PCIe-6.0	1 x 8	1.2	Up to 10	tbd
PCIe-6.0	1 x 16	1.2	Up to 20	tbd

6.4 Direct Optical Connection (DOC)

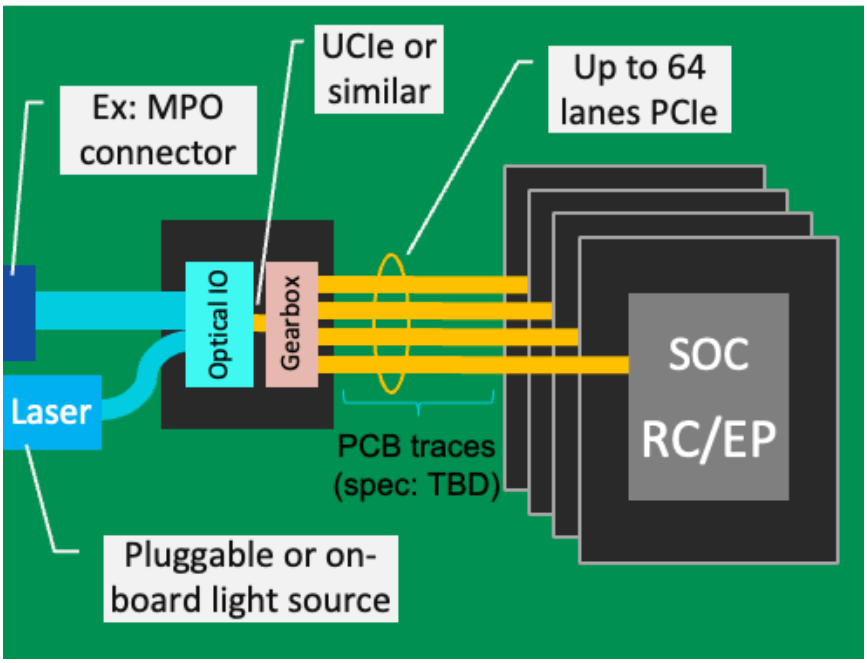
In this scenario, the primary physical medium for data transfer is optical fibers. In some cases, especially as data-rates increase, optical fibers may offer longer reach, better bend radius, lower

weight, higher power efficiency, and lower latency relative to electrical cables. As compute, memory, storage, and switch devices work in the electrical domain, there is a need to convert electrical signals into the optical domain. With Direct Optical Connection (DOC) the transition from electrical to optical is performed in one of two ways:

1. In-package optical IO uses an interface like UCIe between the Root Complex or End Point die and the E/O IO chiplet or engine



2. Near-package electrical to optical conversion uses a PCIe retimer / gearbox between the RC/EP and the E/O IO solution.



The two DOC scenarios are represented above. In both scenarios, it is expected that the PCIe root-complex or endpoint will follow the PCI Express specifications. Channel budgets described in section 8.4 of the PCI Express® Base Specification Revision 6.0 should be applicable. We suggest that PCI-SIG consider adding specifications tailored to “short-reach modes” in order to reduce power and complexity when the channel is considerably short. In-package optical IO might also warrant optimizations around traces and coupling approaches.

We recommend that an option for an external light source be included. Standard external laser form-factors like ELSFP specified by OIF can be used. Three different options of the light source populated as a mid-board module on the PCP, pluggable on the front-panel or as a pluggable in a separate shelf may be considered. These three scenarios are shown in figure 3 of the ELSFP specification (https://www.oiforum.com/wp-content/uploads/OIF-ELSFP-01.0.pdf).

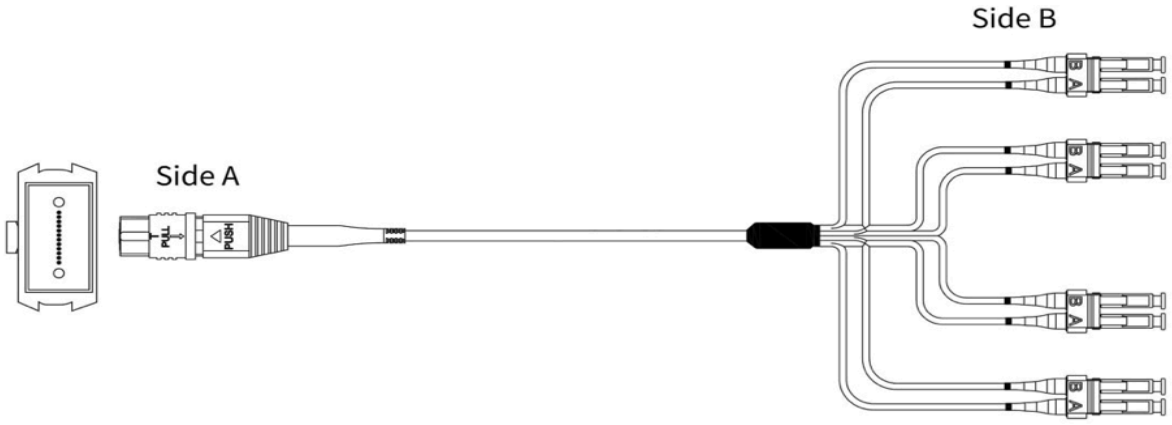
One of the characteristics of DOC is that copper traces do not travel very far from the RC/EP device. In both cases, optical connectors are presented at the front-panel as shown. Passive optical fibers can be connected using standardized connectors. The DOC fiber cables must be completely passive on data path signals, and for cost optimization reasons we don’t require SMBus target controller and signals to provide Vendor Product Data (VPD) information on both ends.

As there is a desire to achieve a different data-rates that match with some natural use-cases like a DDR5 interface or a x8 CXL 2.0 interface, each fiber pair may represent data rates that match up with x4/x8/x16 bifurcations of PCIe speeds. One or more fiber-pairs may map to a desired PCIe interface.

Max data-rate matrix	Spec 5.0	Spec 6.0	Spec 7.0
x4 PCIe	128 Gbps	256 Gbps	512 Gbps
x8 PCIe	256 Gbps	512 Gbps	1024 Gbps
x16 PCIe	512 Gbps	1024 Gbps	2048 Gbps

The diagrams shown in the previous page for example may offer the ability to bifurcate each x16 down to x4 which would require the optical engine to support a minimum of 16 ports.

DOC cables could support Multi-Mode (MM) or Single Mode (SM) fibers.



Source: https://edgeoptic.com/wp-content/uploads/2021/02/12A-MPO-12-4LC.jpg

DOC cables may be split in order to fanout via bifurcation of the CXL interface. A typical scenario may allow for a x8 CXL 2.0 or x4 CXL 3.0 interface to be connected to a different node in the fabric. In order to avoid extra overhead, DOC cables may only operate in SRIS with Spread Spectrum Clocking (SSC) and SRNS modes with standard single frequency clocking.

Source: https://www.oiforum.com/technical-work/current-work/#ELSFP

In typical DOC implementations, the light source that contains lasers that power the optical engine are expected to be external to the data-path devices. They could be incorporated into the system either as a module mounted on the PCB or as a pluggable like the ELSF (External Laser Small Form Factor Pluggable) module specified by OIF. Solutions maybe realized with integrated lasers if power and cost are within requirements.

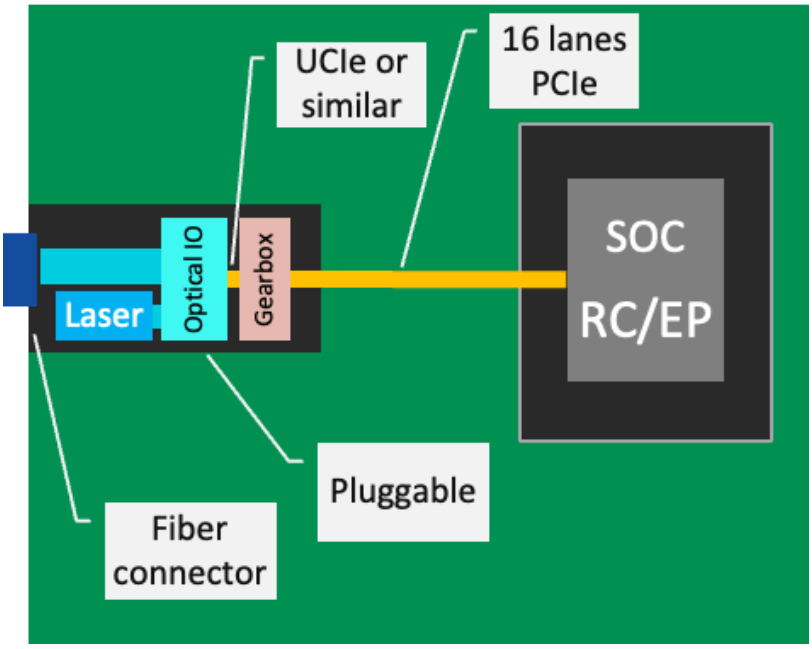
Req ID	DOC	# Ports	Lengths Meters	Bend Radius
CNNCT-19	1x4	1	Up to 10	10x COD
CNNCT-20	1x8	1	Up to 10	10x COD
CNNCT-21	1x16	1	Up to 10	10x COD
CNNCT-22	x8(2x4)	2	Up to 10	10x COD
CNNCT-23	x16(2x8)	2	Up to 10	10x COD
CNNCT-24	x16(4x4)	4	Up to 10	10x COD
CNNCT-25	x8 to 2x4	2	Up to 10	10x COD
CNNCT-26	x16 to 2x8	2	Up to 10	10x COD
CNNCT-27	x16 to 4x4	4	Up to 10	10x COD

DOC Fiber - 10x the outer diameter of the fiber bundle

6.5 Active Optical Connection (AOC)

AOC cables must support built-in Copper to Optical transceiver devices on both end connectors, in order to provide support for the complete standard PCIe-5.0 at 1x10-12 BER requirements, as well as PCIe-6.0 at 1x10-6 BER requirements as far as the Physical transport layer is concerned, while keeping in mind the need for the overall connectivity solution to enhance the BER at the Link layer for the specific protocols that we are addressing being NVMe and CXL.

Rate	# of Lanes	Max Retimer + Optical I/O Circuit/Lane/End (Watts)	Package Size (mm)
PCIe-5.0	1 x 4	1.0	tbd
PCIe-5.0	1 x 8	1.0	tbd
PCIe-5.0	1 x 16	1.0	tbd
PCIe-6.0	1 x 4	1.0	tbd
PCIe-6.0	1 x 8	1.0	tbd
PCIe-6.0	1 x 16	1.0	tbd



AOC cable source and destination PCB route and db-loss budget must also be specified to meet the standard PCIe driver compliance requirements.
AOC cables must contain an SMBus target controller and signals to provide Vendor Product Data (VPD) information on both ends.
AOC cables must use Multi-Mode (MM) or Single-Mode (SM) Micro-Fiber cabling that comply with the optical transceiver's requirements used on both connector ends.
AOC cables must support PCIe port bifurcation in conjunction with support for SRIS with Spread Spectrum Clocking (SSC) and SRNS with standard single frequency clocking are both mandatory.
AOC cables must have the same connector type and widths on both ends of the cable.

Req ID	AOC	# Ports	Lengths Meters	Bend Radius
CNNCT-28	1x4	1	Up to 10	10x COD
CNNCT-29	1x8	1	Up to 10	10x COD
CNNCT-30	1x16	1	Up to 10	10x COD
CNNCT-31	x8(2x4)	2	Up to 10	10x COD
CNNCT-32	x16(2x8)	2	Up to 10	10x COD
CNNCT-33	x16(4x4)	4	Up to 10	10x COD
CNNCT-34	x8 to 2x4	2	Up to 10	10x COD
CNNCT-35	x16 to 2x8	2	Up to 10	10x COD
CNNCT-36	x16 to 4x4	4	Up to 10	10x COD

AOC Fiber - 10x the outer diameter of the fiber bundle