Consortium For On-Board Optics

Design Considerations of Optical Connectivity in a Co-Packaged or On-Board Optics Switch

White Paper

2022

Table of Contents

Definitions and Terms 5
1. Introduction 7
2. System Overview of CPO Switch 8
    2.1 CPO Switch System 8
    2.2 Connectivity Options for Electrical Interface 8
    2.3 Connectivity Options Between Compute and Front Panel 11
3. Optical Engine Interface 15
    3.1 Electrical Separation: Socketed Optical Engine 15
    3.2 Optical Separation: Soldered Optical Engine with Fiber Pigtail 16
    3.3 Optical Separation: Soldered Optical Engine with Fiber Receptacle 16
4. Mid-board Connection 17
    4.1 Fiber Routing Complexity 17
    4.2 Mid-board Connector as a Solution to Fiber Routing Concerns 18
5. Laser Source 20
    5.1 Internal versus External Laser Source 20
    5.2 Options for External Laser Sources 20
    5.3 Polarization Maintaining Fiber for High Power Mating 22
6. Faceplate Design for CPO Switch Interface 23
    6.1 Design Options 23
    6.1.1 Optical Connectors for Data Ports 23
    6.1.2 External Light Source 24
    6.2 Faceplate Comparison 25
7. Optical Test and Measurement 27
    7.1 Location of TP2 and TP3 27
    7.2 Faceplate Connector Loss 28
    7.3 Power Budget Adjustment for CPO 29
8. Cooling Systems 30
    8.1 Thermal Challenge 30
    8.2 Implementation Factors for Cooling Designs 30
    8.3 Cooling Methods for CPO Switch Assembly 35
    8.4 Comparison of Cooling Methods 38
9. Summary 39
Bibliography 40
Co-packaged Optics Working Group Members 43
Points of Contact 44
Copyright Information 44

Copyright © 2022 COBO. All Rights Reserved.

FIGURES

Figure 2-1: CPO electrical interconnect using package or interposer designs 9
Figure 2-2: OBO electrical interconnect using interposer or cable designs 9
Figure 2-3: CPO/OBO non-retimed electrical interface block diagram 10
Figure 2-4: CPO/OBO 'half linear' electrical interface block diagram 10
Figure 2-5: CPO/OBO XSR electrical interface block diagram 11
Figure 2-6: CPO/OBO XSR+ electrical interface block diagram 11
Figure 2-7: CPO/OBO VSR/C2M electrical interface block diagram 11
Figure 2-8: 51.2 Tbps switch CPO layout example 13
Figure 3-1: Electrically separable socketed optical engine 15
Figure 3-2: Soldered optical engine with optically separable pigtail 16
Figure 3-3: Soldered optical engine with optically separable receptacle 16
Figure 4-1: Example fiber layout for 16 co-packaged module design 17
Figure 4-2: Fiber pigtail lengths required for interior connections 18
Figure 4-3: Example of mid-board connector solutions 18
Figure 4-4: Connection points between transmit and receive PMDs 19
Figure 5-1: Potential External Laser Source (ELS) configurations 21
Figure 6-1: Faceplate port requirements 23
Figure 6-2: ELS optical connectivity design options 25
Figure 6-3: Example faceplate configurations 25
Figure 7-1: Test point definitions 27
Figure 7-2: Potential additional test points for signal measurements 28
Figure 8-1: Classification of cooling methods 31
Figure 8-2: Air cooled heat sink cooling mechanism 31
Figure 8-3: Two phase enhanced air cooling mechanisms 33
Figure 8-4: Single phase liquid cooling designs 33
Figure 8-5: Two phase liquid cooling designs 35
Figure 8-6: Liquid-to-air CDU (left) and switch rack with 2 CDUs (right) 36
Figure 8-7: Simulated temperature distribution in a 25.6 Tbps CPO assembly 37
Figure 8-8: Simulated temperature distribution in a 51.2 Tbps CPO assembly 37
Figure 8-9: PUE of different cooling methods in a datacenter 38

Copyright © 2022 COBO. All Rights Reserved.

Tables

Table 2-1: Electrical capabilities of Pluggable, OBO, and CPO designs 8

Table 2-2: Optical capabilities of Pluggable, OBO, and CPO designs 12

Table 2-3: Packaging and design capabilities of Pluggable, OBO, and CPO 12

Table 2-4: Optical fiber connectors and 1RU connector density 14

Table 5-1: Comparison of External Laser Source (ELS) configurations 21

Table 6-1: Faceplate port granularity 24

Table 8-1: 25.6-51.2 Tbps CPO power consumption estimation 30

Table 8-2: Thermal interface material properties 32

Table 8-3: 25.6 Tbps and 51.2 Tbps CPO assembly simulation condition 36

Table 8-4: Comparison of different cooling methods 39

Copyright © 2022 COBO. All Rights Reserved.

Defintions and Terms

Defintions

Term	Definition
Connectorized Module	A PMD with a separable fiber optic connection on the module, with a separate optical fiber patch cord (also called Receptacled Module)
Data Center Cabling	The IEEE 802.3 copper or fiber optic infrastructure between the two MDI points
MDI	Media Dependent Interface - the IEEE 802.3 link location at which the copper or fiber optic connector mates to the PMD
OE	Optical Engine – the chiplet or sub-assembly for a co-packaged optical module the ASIC
OSFP	Octal Small Form Factor Pluggable
Pigtailed Module	A PMD with an inseparable length of fiber, typically terminated with a fiber optic connector, exiting the module and routing to the card edge
PMD	Physical Medium Dependent – IEEE 802.3 compliant transceiver module minus any copper or fiber optic cables necessary to take the signal to the card edge
QSFP-DD	Quad Small Form-factor Pluggable Double Density
TP2	Optical Test Point 2, as described in IEEE 802.3 Ethernet standards, at which optical power is measured from the source PMD
TP3	Optical Test Point 3, as described in IEEE 802.3 Ethernet standards, at which optical receive signal is measured going to the destination PMD

Terms

Term	Definition
ASIC	Application Specific Integrated Circuit
BER	Bit Error Rate
CDR	Clock and Data Recovery
CLTE	Continuous-time Linear Equalizer
COBO	Consortium for On-Board Optics
CPO	Co-Packaged Optics
DAC	Direct Attached Cable
dB	10 times log10 of a power ratio
DFE	Decision-feedback Equalizer
DR4	Module output with 4 optical channels using 8 single-mode fibers, 500 m reach

Copyright © 2022 COBO. All Rights Reserved.

Defintions and Terms (con't)

Terms

Term	Definition
ELS	External Laser Source
FFE	Feed-forward Equalizer
FR4	Module output with 4 optical channels using 2 single-mode fibers, 2 km reach
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IC	Integrated Circuit
MDI	Medium Dependent Interface
MSA	Multi-Supplier Agreement
NPO	Near Packaged Optics
OBO	On Board Optics
PIC	Photonic Integrated Circuit
PUE	Power Usage Effectiveness
SiPh	Silicon Photonics
SN(R)	Signal-to-Noise (Ratio) - the ratio of signal power to the noise power, often expressed in decibel (dB) units
SR4	Module output with 4 optical channels using 8 multi-mode fibers, 100 m reach
SR8	Module output with 8 optical channels using 16 multi-mode fibers, 100 m reach
TIA	Telecommunications Industry Association
VSR	Very Short Reach
XSR, XSR+	Extra Short Reach

Copyright © 2022 COBO. All Rights Reserved.

1. Introduction

Next generation applications within datacenters continue their trend toward data intensive usage models and resource disaggregation. To meet the next generation application requirements, cost, power, latency, and overall bandwidth of each component needs to be considered for both existing applications and new use cases. The shift to cloud-based networks focuses this data load on hyperscale-level datacenters were networking equipment upgrades focus on denser, cheaper, and faster implementations which can expand data network bandwidth. These upgrades are occurring as the network design has changed, shifting to mesh architectures where leaf and spine switches are used to provide greater reliability and greater increases to the interconnectivity of the network. The required interconnectivity demands high speed switches and optical interconnects in order to maintain and sustain application developments.

As these market drivers create larger bandwidth requirements, technical forces make using pluggable technologies for interconnects difficult as sufficient bandwidth begin to enact space and size penalties versus existing designs. Embedded optics at the board or module have been discussed and deployed for generations, and the concept is not new in the networking industry. Optics mounted directly on the host PCB or directly near the compute module or sockets were target for specific applications such as in supercomputers. On-board optical (OBO) or co-packaged optical (CPO) modules both provide interconnects within the server PCB. The Consortium for On-Board Optics (COBO) members developed an embedded optical module form factor: the On-Board Optical Module Specification to support 400G, 800G, and beyond. For CPO modules, module requirements on shape, size, and power are being aligned through various consortia, multi-supplier agreements (MSA), and standard agencies such as IEEE or OIF. [1]

The removal of the optical pluggables from the faceplate provides an opportunity to re-evaluate the faceplate design configuration for optical connections, heat management, and power into a network switch or datacenter server. This white paper will review design options available to accommodate the optical connectivity required and evaluate potential impacts on optical signal, thermal, and safety criteria. Topics will include issues, potential solutions, and design implementation considerations to inform the implementer community. It should be noted that the descriptions in this application note are intended as examples of possible applications and implementations. In no case are they intended to be prescriptive as certain CPO specifications may change and the overall options may require adjustment for implementation to meet other new or changed specifications.

Copyright © 2022 COBO. All Rights Reserved.

2. System overview of CPO switch

2.1 - Design changes with a CPO switch system

The CPO switch system still performs the same structure as a pluggable-based switch, routing data between individual serves and toward the overall data center. A CPO based switch system would integrate, or co-package, the ASIC and the optical engine onto a single substrate. Solutions would need to co-package the ASIC with the correct number of optical engines to reach the target bandwidth. The shift of optical transceivers away from the front plate creates new opportunities for revised layouts and modifications across the entire switch. Capabilities for electrical, optical, and thermal cooling, and packaging designs will depend on the design choices for various solutions.

2.2 - Connectivity options for electrical interface

The electrical interconnect between the host ASIC and the (optical) engine/module should be selected to minimize power and cost at the same time achieving the required performance metrics and mechanical requirements. The electrical interface performance is not only a function of the insertion loss of the host electrical interconnect, it is also dependent on the reflections and crosstalk of the interconnect. A variety of options have been proposed as bandwidth requirements continue to increase. A high level summary is provided in Table 2-1.

Characteristics	Pluggable	On-Board Optics	Co-Packaged Optics
Electrical Interface Reach	LR, VSR	VSR, XSR, XSR+	XSR, XSR+, Linear
Electrical Interface Width	Serial	Serial	Serial or Wide
Bandwidth Density (Electrical)	Serial PCB phy & bump pitch dependent	OBO module size dependent	High. Package level bump pitch dependent
Target Link Bandwidths	All link bandwidths	Link BW: 1, 2, 4, 8+ Lanes: 25 to 100 Gbps	Link BW: 1, 2, 4, 8+ Lanes: 50, 100+ Gbps

TABLE 2.1 - Electrical Capabilities of Pluggable, OBO, and CPO Designs

The signal integrity of the host channel is controlled by PCB design or cabled interconnect performance and connector/socket selection. The following sections provide guidance for selecting the optimal host ASIC to optical engine/module electrical interconnect and interface.

Electrical interconnect options for CPO

The electrical interconnect between the Host ASIC and a CPO optical engine has no connector (but may have a high performance socket). The channel uses package or interposer traces to achieve good signal integrity. Implementations of CPO designs may use package traces (Figure 2-1a) or interposer traces (Figure 2-1b). The signal integrity of the electrical channel is determined by the channel routing, the socket performance and package parasitics. Note that all insertion loss values in this section are calculated at the Nyquist frequency for 112G PAM4 signaling.

Copyright © 2022 COBO. All Rights Reserved.

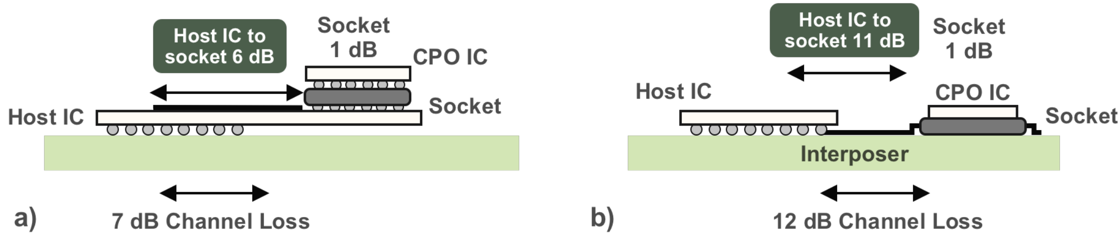

FIGURE 2.1 - CPO electrical interconnect using package or interposer designs

The electrical interconnect between the Host ASIC and an On-Board Optics (OBO) socket and IC has no connector (but may have high performance sockets on both ends of the channel). These connection types have also been called near packaged optics (NPO), as the ASIC and OBO IC are ideally near each other in order to facilitate shorter electrical interconnects. The interconnect may use host PCB traces connected to the OBO socket, or interposer traces connected to an intra box cable which is connected to a OBO socket. The connections to the package/interposer and the OBO module are done via sockets. Figure 2-2 shows the Host IC to OBO interconnect using a interposer traces (2-2a) and an intra box cable (2-2b). The signal integrity of the electrical channel is determined by the package/interposer channel, the intra box cable, the socket performance and package parasitics.

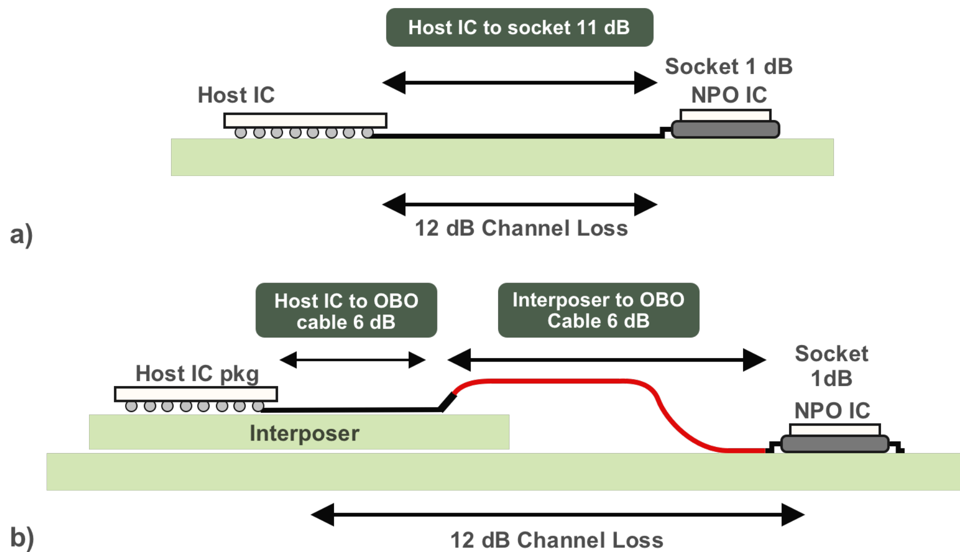

FIGURE 2.2 - OBO electrical interconnect using interposer or cable designs

Electrical interface options for the CPO and OBO interconnects

There are multiple electrical interface options that are available for the CPO and OBO interconnects. The selection of the optimum electrical interface depends on the insertion loss, the use of a connector and the signal integrity of the interconnect.

Copyright © 2022 COBO. All Rights Reserved.

Designs can use a linear interface, long reach (LR), a very short reach (VSR), or a retimed extra short reach (XSR and XSR+). A wide variety of electrical interoperability designs are available and the following summarizes some relevant designs for CPO or OBO interconnects. A continuous-time linear equalizer (CTLE), feed-forward equalizer (FFE), and decision-feedback equalizer (DFE) are needed in all cases. The electrical interface options addressed here are listed in increasing complexity of module Clock and Data Recovery (CDR) design.

The non-retimed interface uses the ability of the Host IC to equalize both the electrical and optical interconnect. The module design requires only a CTLE in the transmit path and pre-emphasis (labeled as Pre-Emph) on the receive path. The linear interface has a CDR of 22 tap DFE and 40 tap FFE, and is shown in Figure 2-3.

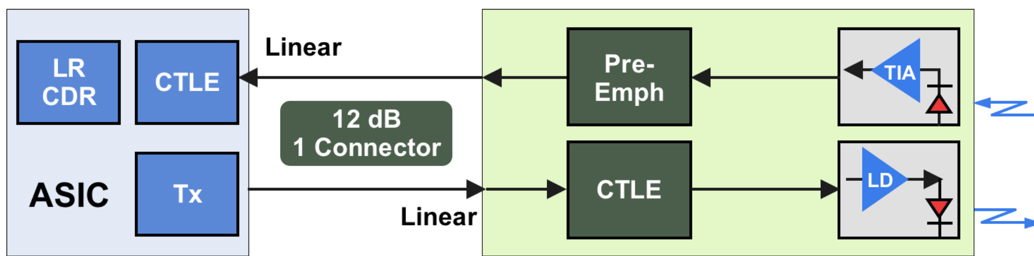

Figure 2-3: CPO/OBO non-retimed electrical interface block diagram

The Half Linear interface uses the XSR+ retimed interface in the Tx direction and a linear interface for the Rx direction. This option can be used for optical module designs that require retiming in the Tx direction because of the design of the laser driver/modulator. The half linear interface would have the same LR CDR and a XSR+ CDR of 1 tap DFE and 3 tap FFE, and is shown in Figure 2-4.

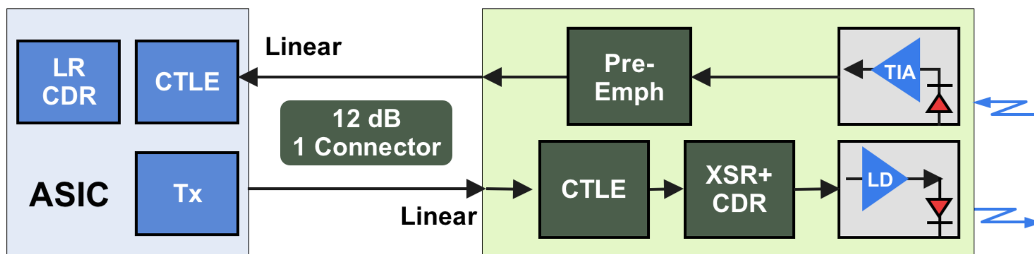

Figure 2-4: CPO/OBO 'half linear' electrical interface block diagram

The XSR electrical interface uses the lowest power retimer circuits in the module. The XSR CDR has no digital equalizer, resulting in the optical CDR being between 5 – 15 tap FFE only. However, it does not support the higher insertion loss interconnects and does not support a connector. The XSR electrical interface is shown in Figure 2-5.

Copyright © 2022 COBO. All Rights Reserved.

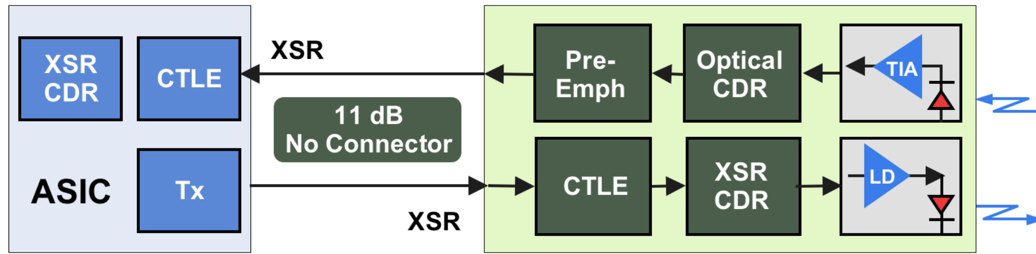

Figure 2-5: CPO/OBO XSR electrical interface block diagram

The XSR+ electrical interface uses low power retiming in both Tx and Rx directions. The XSR+ CDR is 1 tap DFE and 3 tap FFE, and the optical CDR would be 5 – 15 tap FFE. It requires retiming of the optical interface to be done inside the optical module. It supports one connector/socket. The XSR+ electrical interface is shown in Figure 2-6.

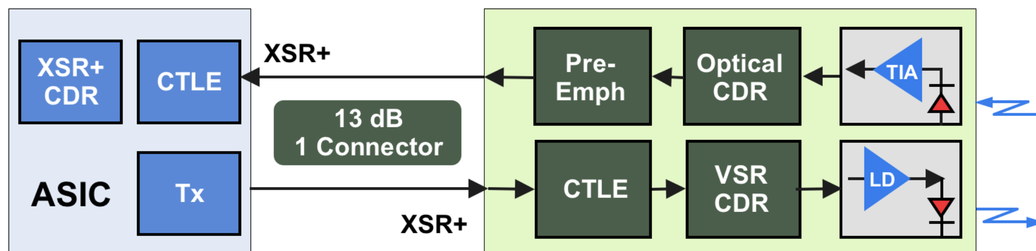

Figure 2-6: CPO/NPO XSR+ electrical interface block diagram

The VSR/Chip to Module (VSR/C2M) electrical interface is traditionally used for Host IC to face-plate pluggable modules. It supports the highest loss interconnect but also requires the highest optical module power dissipation. The VSR electrical interface has a CDR of 4 tap DFE and 8 tap FFE, and is shown in Figure 2-7.

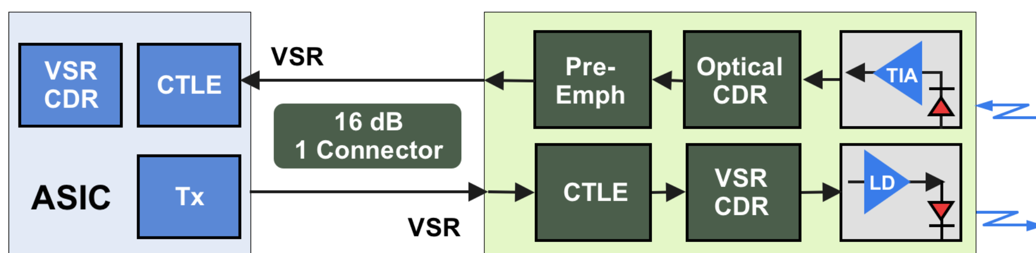

Figure 2-7: CPO/NPO VSR/C2M electrical interface block diagram

2.3 Connectivity options between compute and front panel

For a 51.2 Tbps switch, most of the ports on the front panel will be optical connections. Optical connectivity options have been reviewed extensively in other reports [2], but a brief discussion of connections and definitions is needed to define the optical path from the compute engine to the front panel.

Copyright © 2022 COBO. All Rights Reserved.

The term PMD (Physical Medium Dependent) from the IEEE is used in the same context as prior COBO whitepapers [3] to incorporate the electro-optical package mounted to the system circuit board without the fiber necessary to take the signal to and from the card edge. In a co-packaged optical design, the PMD is the integrated optical engine (OE). Some form of optical cabling is required to reach the Medium Dependent Interface (MDI), which is defined in the same manner as IEEE ethernet standards. The end-user of the system simply sees an MDI interface at the faceplate regardless of connectivity type. A summary of potential performance changes in the optical and packaging capabilities for on-board optics and co-packages optics are provided in Table 2-2 and Table 2-3, respectively.

Characteristics	Pluggable	On-Board Optics	Co-Packaged Optics
Internal or External Laser	Internal	Internal	Internal or External
Signaling Technology	56G and 112G PAM4	56G and 112G PAM4 (112G focus)	56G and 112G PAM4 (112G focus)
Signal Integrity	Reflected in power and BER rows.	Reflected in power and BER rows.	Reflected in power and BER rows.
Retimers	Potentially (Dependent on rate, distance)	No	No
Bandwidth Density (Optical)	Determined by transceiver module width	OBO module size dependent (likely similar to pluggable)	High. PIC size dependent
Optical Interface Reach	Wide range: cost dependent options	Intra-data center	Intra-data center
Optical Lane Formats	SR4/SR8, DR4, FR4	SR4/SR8, DR4, FR4	SR4/SR8, DR4, FR4

Table 2-2: Optical Capabilities of Pluggable, OBO, and CPO Designs

Characteristics	Pluggable	On-Board Optics	Co-Packaged Optics
Connector Locations	1: Attached to Module	2: OBO module and faceplate	2: ASIC and faceplate
Packaging Complexity	+No impact on ASIC package (other than traces)	+No impact on ASIC package (other than traces)	Intrusive to package design by definition
Manufacture Capability	Evolutionary approach for transceiver modules	Not dependent on high SiPh integration	Not mature: SiPh integration expected.
Serviceability	Known model with access at faceplate	Evolving	Challenging. New approach required
Module Standardization	Exists	Exists	Required for package level interoperability
Ecosystem Support	Exists	Exists	Expected if standardized
Electrical-Optical Fan-In/ Fan-Out	No opportunity. ASIC lanes are individually routed to faceplate	Electrical lanes can fan-in from ASIC. PCB routing complexity limits efficiency	Density advantage limited by USR/XSR electrical interface
Blast Radius (with OE module failure)	Does not require replacement of ASIC	Does not impact ASIC	ASIC package replacement may be required

Table 2-3: Pluggable, OBO, and CPO Packaging and Design Capabilities

Copyright © 2022 COBO. All Rights Reserved.

Both the connection method and connection type are design considerations that impact overall system requirements for signal, heat, and cabling density. Pigtailed cabling is directly attached to the PMD, while connectorized cabling uses a patch cord with connectors at both ends. In either case, connections at the panel are needed for data transmit/receive along with optional external laser sources. A design showing the basic elements of a switch ASIC with co-packaged optics is shown in Figure 2-8.

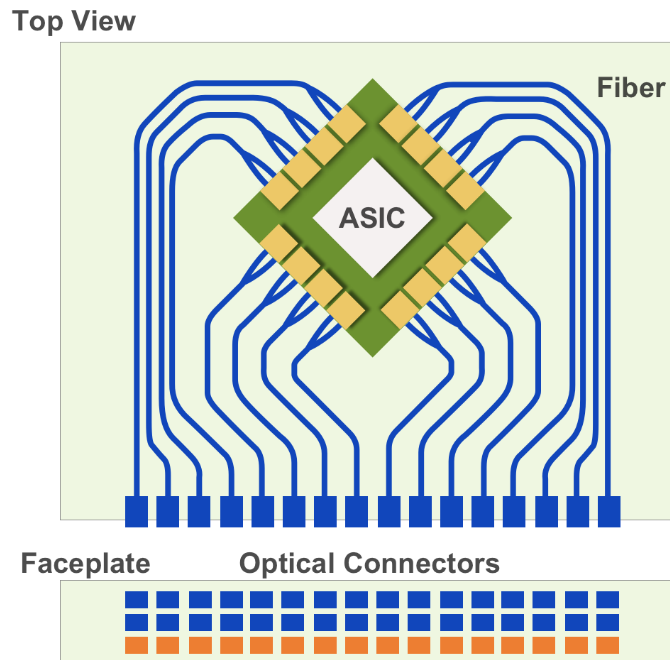

Figure 2-8: 51.2 Tbps switch CPO layout example

Detail descriptions of fiber optic connectors, cleaning of connector end faces, and connector reliability and performance standards can be found in [3]. Connector types vary dependent on implementation and the selection of PMD. For a 51.2 Tbps switch, multiple multi-fiber connectors are expected to be required due to single channel bandwidth constraints. For external laser sources which would plug into the face plate, current form factors are not standardized. This whitepaper follows the proposal of the Co-Packaged Optics Collaboration [4], where either OSFP form factor dimensions [5] or QSFP-DD form factor dimensions [6] would be used. A list of commonly known connector types is provided in Table 2-4. These designs have different fiber densities and area requirements, which define the overall concentration and layout of connectors on the faceplate layout.

Copyright © 2022 COBO. All Rights Reserved.

Fiber Connector	Fibers per Connector	Connectors per 1RU	Fibers per 1RU	Reference Standard
Duplex LC	2	144	144	IEC 61754-20:2012 [7]
CS	2	320	320	TIA-604-19 [8]
MDC	2	432	432	IEC 61754-37 [9]
SN	2	432	432	IEC 61754-36 [10]
MPO-12	12	80	960	TIA-604-5 [11] IEC 61754-7-1 [12]
MPO-16	16	80	1280	TIA-604-18 [13] IEC 61754-7-4 [14]
AirMT-12	12	128	1536	
MPO-24	24	80	1920	TIA-604-5 [11] IEC 61754-7-2 [16]
MPO-32	32	80	2560	TIA-604-18 [13] IEC 61754-7-3 [17]
AirMT-24	24	128	3072	
MXC-32	32	104	3328	
MMC-16	16	216	3456	
SN-MT16	16	216	3456	

Table 2-4: Optical fiber connectors and 1RU connector density

The size of the overall connection area on the rack faceplate will vary depending on the overall design and spacing requirements of each receptacle. Blind mate connectors, used to protect the end-user from eyesight damage, may also require designs with receptacles significantly larger than the minimum receptacle size requirements. These connector designs are referenced later in the whitepaper when faceplate connections are required.

Copyright © 2022 COBO. All Rights Reserved.

3. Optical engine interface

One of the primary challenges of co-packaged optics is the large amount of fiber that connects directly to the CPO substrate. At the time of this white paper, CPO fiber counts for 51.2 Tbps bandwidth would require hundreds to over 1000 fibers. Fibers may be different lengths and have optical connectors attached to the ends, including MT-type multi-fiber connectors. End users will not have easy access to this fiber since these connectors are on-board and behind the front panel. Although there are many optical factories trained in handling on-board fiber, connecting and routing the required number of fibers to an ASIC is a new challenge.

Design safeguards to the optical engine (OE) interface are recommended to allow for replacement of components, thus impacting the yield and reliability of the entire CPO switch. It is critical that the components are designed such that they can be physically separated from the CPO. As shown in the following sub-sections, this separation can be achieved electrically, optically, or both depending on design needs.

3.1 Electrical separation: socketed optical engine

Electrical separation is typically realized by means of a socket between the OE and the CPO substrate (see Figure 3-1). It allows CPO factories to plug in optical engines after the solder reflow process. This avoids exposing the OE, optical fiber, and connectors to high temperatures, thus improving switch yield and reliability. An electrical socket also allows for the use of standard optical connectors and transceiver components. Using standard optics reduces cost and enables a mature, diverse supply chain.

The tradeoff to electrically socketed optical engines is reduced electrical performance and higher component costs. At some speed threshold, it may be difficult to maintain signal integrity. Therefore, some future optical engines may not be able to use an electrical socket. Current development plans use electrical sockets, but alternative options are available and discussed in later sections.

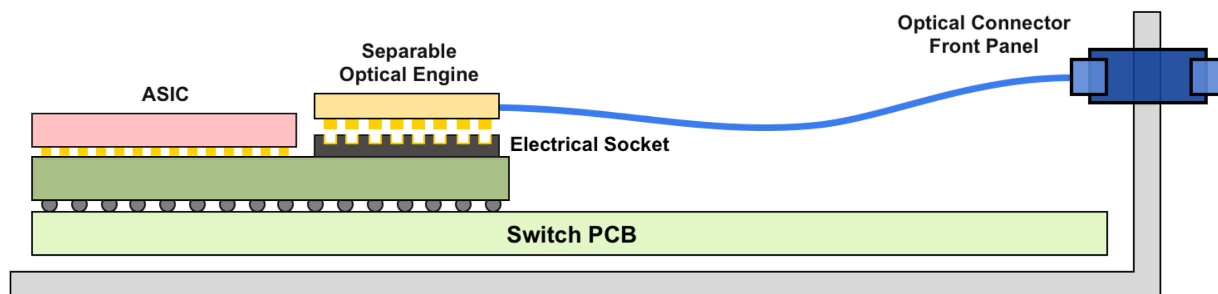

Figure 3-1: Electrically separable socketed optical engine

Copyright © 2022 COBO. All Rights Reserved.

3.2 Optical separation: soldered optical engine with fiber pigtail

If instead of an electric socket the OE is soldered to the substrate, optical separation is highly desired. Optical separation is typically realized by means of an optical connector between the OE and the on-board fiber. It allows for most of the optical fiber and connectors to be installed after the solder reflow process. However, some type of optical interface must remain integrated into the OE.

One method of achieving this optical separation is for the OE to integrate a short length of fiber ending with an optical connector (i.e. a fiber pigtail, See Figure 3-2). Again, because the fiber pigtail is permanently fused to the OE, it too must be compatible with solder reflow temperatures. At the time of this whitepaper, commercialized MT-type optical connectors/pigtails compatible with reflow are not commercially available.

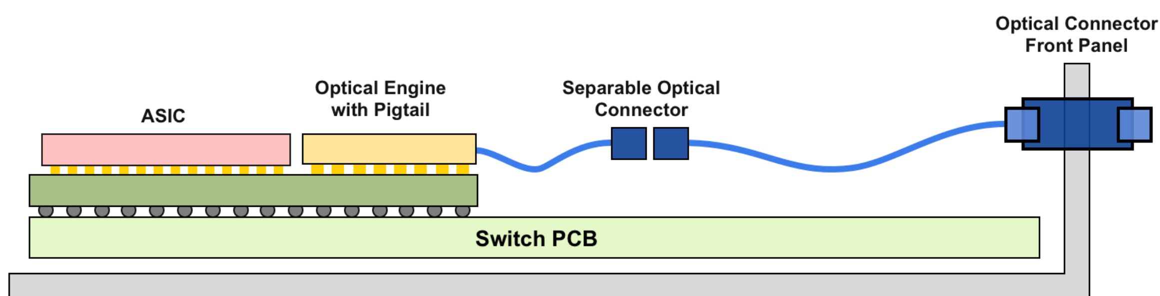

Figure 3-2: Soldered optical engine with optically separable pigtail

3.3 Optical separation: soldered optical engine with fiber receptacle

Another method to achieve optical separation is to integrate a fiber optic receptacle into the OE, illustrated in Figure 3-3. While similar to the fiber pigtail in Section 3.2, this subtle change eliminates exposure of the fiber and connector to solder reflow temperatures. Much like electrically socketed optical engines in Section 3.1, it enables the use of standard optical connectors to mate to the receptacle.

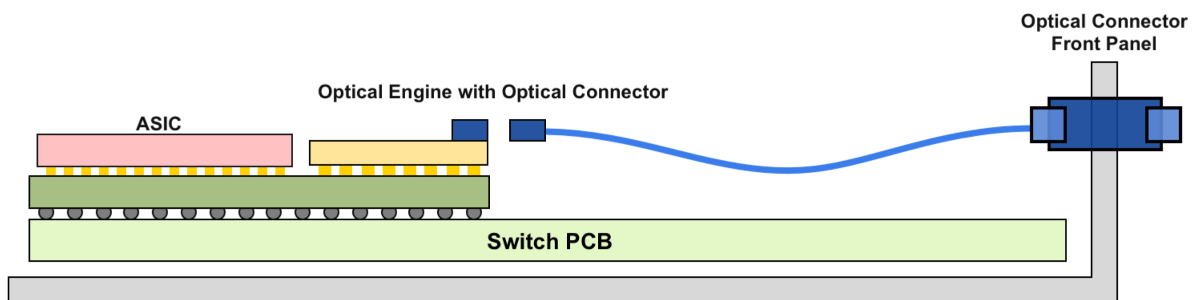

Figure 3-3: Soldered Optical Engine with Optically Separable Receptacle

Any failure modes of the optical engine in this design are non-recoverable without removal of the co-packaged module. The optical connector attached directly to the engine could potentially take up valuable space for component cooling. The connector on board may also create risks during optical fiber replacement. since space may be constrained within the switch enclosure. At the time of this whitepaper, commercialized MT-type optical receptacles compatible with reflow have been demonstrated, but are not yet commercially available.

Copyright © 2022 COBO. All Rights Reserved.

4. Mid-board connection

4.1 Fiber routing complexity

For a CPO switch ASIC, dealing with fiber routing is inevitable from fibers from optical engines to the front panel inside the switch chassis. The optical engines are positioned closely around the ASIC to minimize the distance of the electrical path, thereby maximizing the electrical performance and lowering the overall ASIC power consumption. However, design selection complicates the fiber routing inside the system. Even though electrical paths can be minimized between OE and ASIC, the distance from each OE to the faceplate would vary. In addition, the fibers would exit the optical engine in four different directions in the case where the optical engines are positioned as illustrated in Figure 4-1.

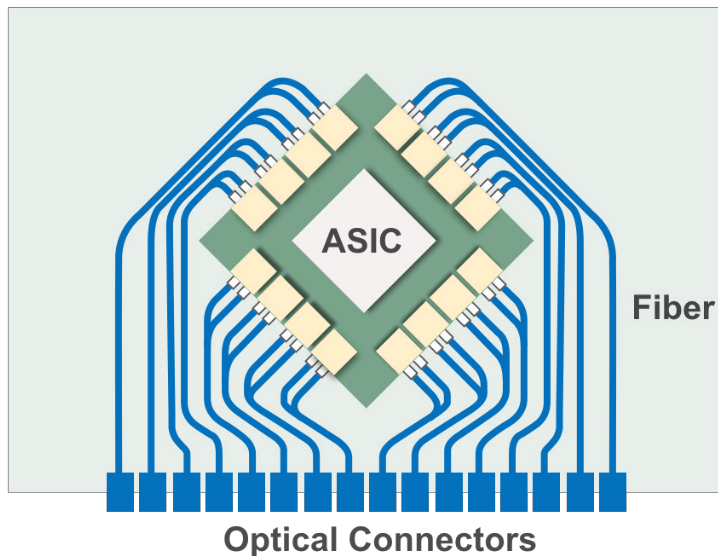

Figure 4-1: Example fiber layout for 16 co-packaged module design

For an electrically separable socketed optical engine, fibers in the Figure 4-1 design would need to be long enough to extend to the faceplate, requiring various lengths of fiber to route each optical engine. For a pigtailed optical engine design as shown in Figure 4-1, there may be 4 to 8 variants of pigtail length required. Multiple lengths increase the total components in the BOM of the switch system, and the OE manufacturers would be required to prepare the pigtails in various lengths as illustrated in Figure 4-2.

Copyright © 2022 COBO. All Rights Reserved.

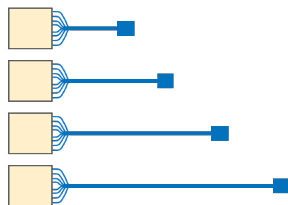

Figure 4-2: Fiber pigtail lengths required for interior connections

Pigtail length variation brings concerns not only about design but also the risk of fiber damage during installation. The fibers pigtails will need to be routed in tight spaces surrounded by other components. This can impact the overall production yield, especially in the case that the OEs are soldered in the switch package and are not replaceable, where damage to a single fiber in the pigtail will compromise the entire CPO switch package. This yield risk can contribute negatively to the overall costs of the switch production. In the case where pigtails would be connectorized, connector positions and size could constrain options for heating or fiber density. Pigtail lengths, if varied, may documentation detailing fiber layouts to prevent unwanted breakage.

4.2 Mid-board connector as a solution to fiber routing concerns

Implementing a mid-board/on-board optical interconnect solution can solve these concerns. By adding a mid-board connector between the OE and the faceplate, and providing jumper cords of various lengths, the OE pigtail length can be reduced to just one design. This simplifies the manufacturing for the OE vendors, while also reducing the risk of damage to the OE and the attached pigtail. Furthermore, the shorter pigtails simplify the installation of the ASIC OE to the faceplate. Although the jumper cords require routing around the system, fiber breakage do not compromise the CPO switch ASIC subassembly. Jumper fibers can be easily replaced with another jumper. Yield-affecting risks are shifted to jumper cables from the more expensive OE and ASIC components. The change helps to improve the cost impact of the CPO switch ASIC. There are multiple solutions for mid-board/on-board connectivity. Options for mid-board connectors are shown in Figure 4-3.

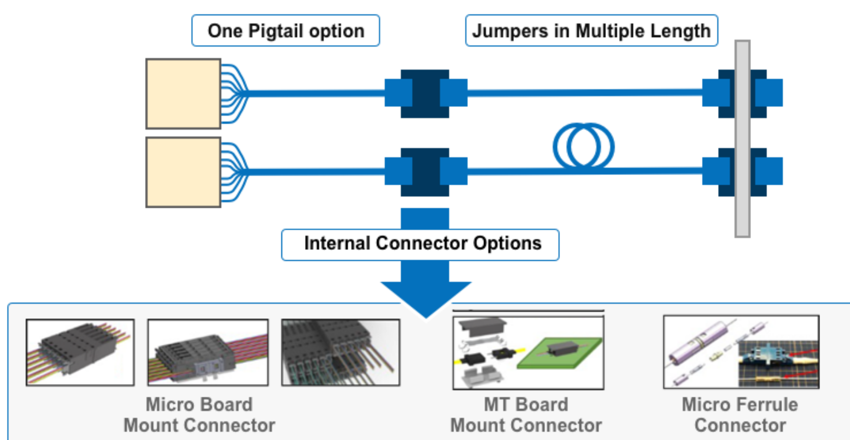

Figure 4-3: Example of mid-board connector solutions

Copyright © 2022 COBO. All Rights Reserved.

This approach does add an additional connection and the associated connector loss. The system suppliers need to consider a way to keep this connector loss as low as possible, and desirably the end operators can maintain the other connection points in the network at lower loss as well. The number of potential connector points increase, as shown in Figure 4-4. CPO assemblies would add additional fiber optical cables and connectors to the design, with jumper cables adding one additional set beyond the ones included for a CPO solution. The use of airgap or expanded-beam connectors, which are more dust-insensitive and easy to clean, could help reduce the operational costs associated with the assembly and maintenance of these connections.

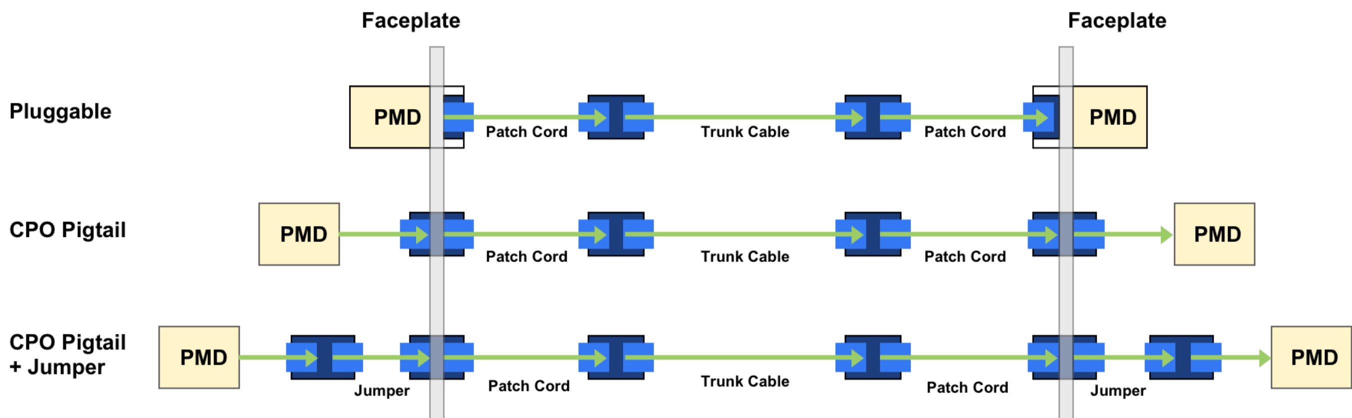

Figure 4-4: Connection points between transmit and receive PMDs

The use of IEC 61753-1 Grade B rated connectors can be considered to achieve a lower loss consistently. It is one of the best performance grades for the fiber optic connector insertion loss. This performance grade is providing a manufacturer independent interoperability, and the connector loss is guaranteed by a test with randomly mated pairs among different manufacturers and manufacturing batches. Based on this random mating specification for MPO connector (IEC 61755-3-3), Grade B MPO 12F single mode fiber connectors are to meet 0.25dB or better at 97.3% of each fiber channel. A summary of fiber interconnections was previously provided in Table 2-4.

Copyright © 2022 COBO. All Rights Reserved.

5. Laser source

5.1 Internal versus external laser source

For optical engines placed near the ASIC, multiple factors should be considered prior to the selection of an internal (or integrated) laser source (ILS) versus an external laser source (ELS). Failure in the pluggable design currently used in most datacenter applications requires the replacement of the pluggable on the faceplate of the switch. Using the design previously shown in Figure 4-1, the reliability of the ASIC and 16 ILS-based OE components would be combined. Failure of one ILS-based OE would reduce the switch network efficiency. The replacement and service cost of the full ASIC/OE subassembly would be needed to restore full functionality.

Designs with external laser sources would locate the laser in separate packages or modules. This design should remove one component failure risk off the subassembly. These ELS modules would use optical fiber to supply the laser source to each OE, resulting in an increase in the quantity and density of fibers required to reach the optical engine. Ideally, these ELS connectors use standard pluggable form factors and fiber connectors which could be incorporated on the switch faceplate. Multiple ELS modules may be required to supply the laser intensity necessary to power every lane. To connect these to the OEs inside the switch, similar design considerations on pigtails and fiber path cords would be needed for ELS.

The ELS modules can also be optimized for heat management and temperature control. ASIC temperature consistency depends on the utilization rate of the network switch, while separate ELS modules would have improved thermal management or be isolated from the ASIC heat generation. Other design options for ELS modules are dependent on the module design and implementation requirements.

5.2 Options for external laser sources

There are multiple options to create ELS modules, and designs can be categorized into 3 types: an On-Board Optics (OBO) design, a front plate pluggable design with a fiber pigtail (Pluggable Pigtail), and a front plate pluggable design with a blind mate optical connector (Pluggable Blind Mate). Polarized maintaining fibers (PMF) are used in most ELS module designs at this time. Design options are shown schematically in Figure 5-1.

Copyright © 2022 COBO. All Rights Reserved.

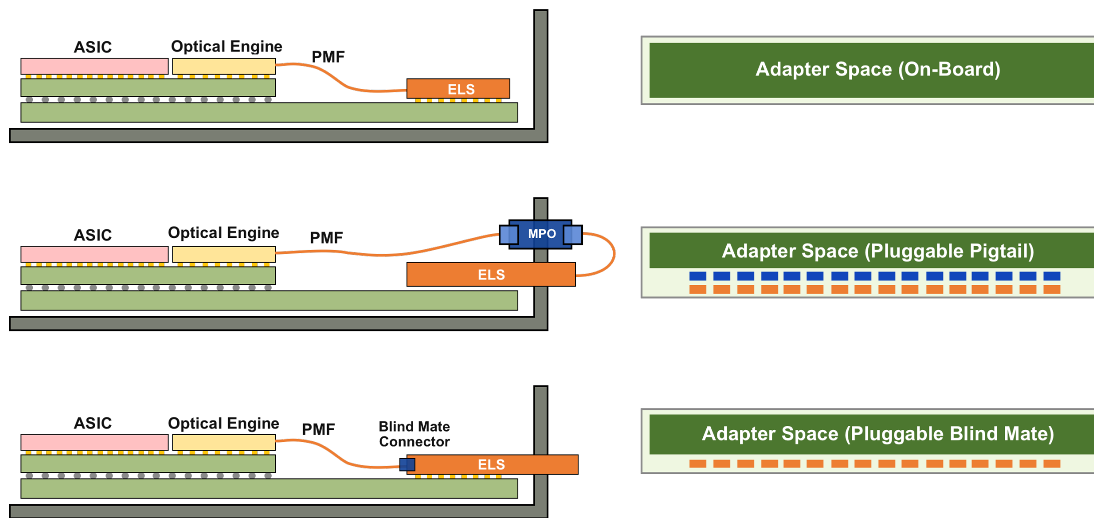

Figure 5-1: Potential External Laser Source (ELS) configurations

For these three options, there are trade-offs in the design constraints on the module and the switch by selecting one option over another option. For these three options, a summary comparison is provided in Table 5-1.

Item	On-Board Optic	Pluggable Pigtail	Pluggable Blind Mate
Serviceability	**Difficult** Entire switch powered down	**Easy** Individually serviceable	**Easy** Individually serviceable
Eye Safety	**Easy** No field concerns	**Difficult** Specialized shutter and/or dust cap required	**Easy** Laser not directed towards user and no power until mated
Optical Connector Servicing	**Easy**	**Difficult**	**Challenging**
Cleaning	Performed at factory only	Tools existing	Requires specialized tools
Inspection in field	Performed at factory only	Field technicians must ensure pristine end face for high power fiber mating	Field technicians must ensure pristine end face for high power fiber mating
Thermal / Cooling	**Easy**	**Challenging**	**Difficult**
Front Panel Space	Requires no faceplate space	Require both ELS ports and optical adapters	Requires ELS ports
Module Cooling	Heatsink on each ELS	Similar to legacy QSFP style module	Similar to legacy QSFP style module
Switch Cooling	Lowest impact on faceplate space for air vents	Least space at faceplate for air vents	Less space at faceplate for air vents

Table 5-1: Comparison of External Laser Source (ELS) Configurations

Copyright © 2022 COBO. All Rights Reserved.

5.3 Polarization maintaining fiber for high power mating

Each of the three ELS types use an array of polarization maintaining (PM) fibers to connect to the optical engine. Use of PM fiber is not common in the data center or the switch, especially in array connector types (e.g. MPO). Array connector manufacturers are studying methods to optimize the use of PM fiber, which need to be aligned towithin a certain rotational precision to properly isolate the two polarization states - expressed as polarization extinction ratio (PER).

Connector manufacturers are also studying the end face cleanliness requirements for mating of PM fiber in high power applications. Data centers are familiar with the challenges of debris on optical connectors. Links brought down by debris are common in normal operations. Cleaning and inspection can resolve these issues, but it is often mitigated reactively as links go down in the field. The penalty for debris on ELS PM mated fibers is much higher. The ELS PM fibers may carry hundreds of mW of power, which is two orders of magnitude greater than a typical data fiber. With these power levels, even the smallest bit of debris may result in permanent catastroph-ic damage to the mating connector. Ensuring clean ELS PM fiber mating connectors is critical. Regardless of the type of ELS module, PM fiber for high power mated connections will require a more diligent and proactive approach to cleaning & inspection in the field.

Copyright © 2022 COBO. All Rights Reserved.

6. Faceplate design for CPO switch interface

The faceplate can be configured in multiple ways, with different combinations of the optical port granularity and connector types. It is important that these are carefully chosen to serve the operational needs of the data center as well as to assist in the thermal management of the switch box, particularly if this involves the air-cooling of >1kW systems. For example, as shown in Figure 6-1, the faceplate of a 51.2 Tbps switch box may be configured as 32 ports with 1.6 Tbps throughput per port, or 128 ports x 400 Gbps per port, with impact on the required panel height (rack real estate) and faceplate surface available for airflow. Basic guidance is provided below on the possible configurations and their influence on thermal management.

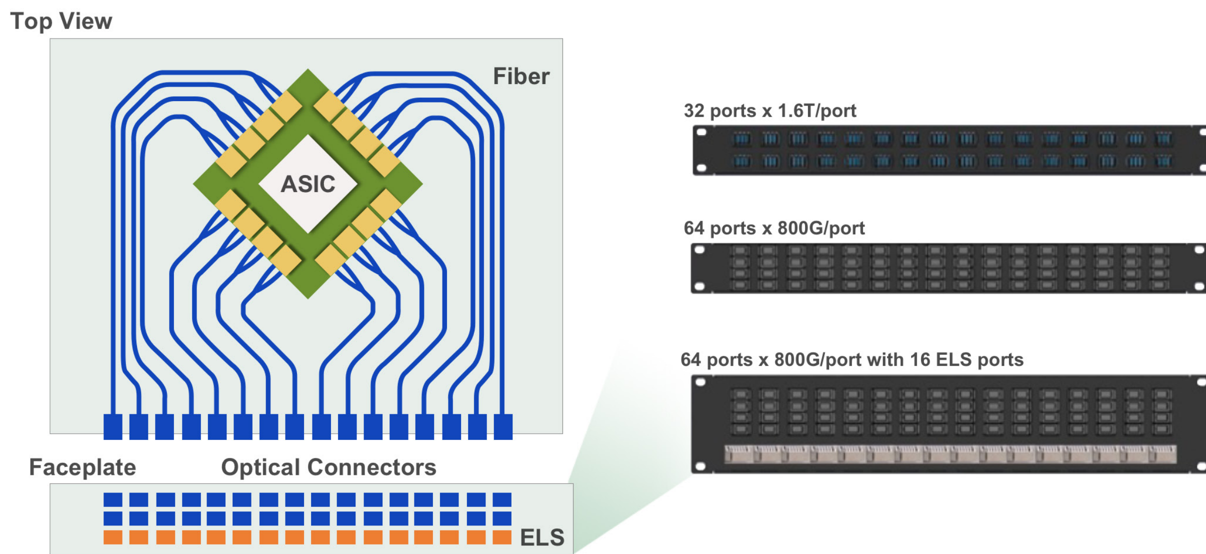

Figure 6-1: Faceplate port requirements

6.1 Design options

6.1.1 Optical connectors for data ports

Table 6-1 lists the fiber channel density for various optical connectors, and the corresponding panel height required DR-based and FR-based switch systems, assuming 100Gb/s/lane communication. For ease of discussion, DR-based 51.2 Tbps bandwidth switches were assumed to require 1024 fibers, while FR-based 51.2 Tbps bandwidth switches were assumed to require 256 fibers. Due the reduced fiber requirements for FR-based front panels, additional connector options could be available for a 1U design since a smaller number of ports would be necessary to fit 256 fibers on the faceplate. For DR-based front panels, only connectors with high number of fibers per port would fit within 1U switch design.

Copyright © 2022 COBO. All Rights Reserved.

Fiber Connector	Fibers per Connector	Connectors per 1RU	Fibers per 1RU	DR (1024f)	FR (256f)
LC	2	72	144	2U+	2U
CS	2	160	320	2U+	1U
MDC	2	216	432	2U+	1U
SN	2	216	432	2U+	1U
MPO-12	12	80	960	2U	1U
MPO-16	16	80	1280	1U	1U
AirMT-12	12	128	1536	1U	1U
MPO-24	24	80	1920	1U	1U
MPO-32	32	80	2560	1U	1U
AirMT-24	24	128	3072	1U	1U
MXC-32	32	104	3328	1U	1U
MMC-16	16	216	3456	1U	1U
SN-MT16	16	216	3456	1U	1U

Figure 6-1: Faceplate Port Granularity

Connector choices also differ in size and functionality: the air-gap (AirMT) and expanded beam connectors are more dust resistant. Connectors such as AirMT, MXC, MMC, or SN-MT16, have a smaller connector size and high fiber/port density to minimize the number of ports and the overall area required on the faceplate. The optical and operational characteristics of these connectors are described and referenced in Section 2, and connector selection should be chosen according to user needs.

6.1.2 External light source

The implementation of the optical source, as described in Section 5, is also a matter of consideration for the faceplate design. In the case that the optical source is implemented on the front panel in the form of a pluggable external light source (ELS) modules, its form factor and method of optical connection will influence the faceplate design. A blind-mate connection is preferred, whereby the optical connection between the ELS and the CPO is made through the host-connector inside the module cage, as opposed to having a dedicated connector on the faceplate, as shown in Figure 6-2. This not only improves the eye-safety of the system, but also reduces the footprint, thereby enabling a smaller panel height and/or increased space for airflow.

Copyright © 2022 COBO. All Rights Reserved.

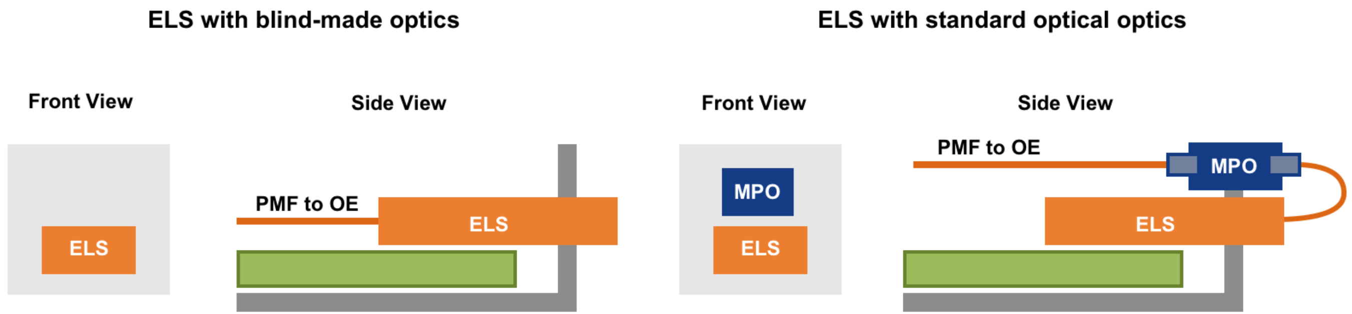

Figure 6-2: ELS optical connectivity design options

The number of ELS modules required for a CPO system depends on the optical power required by the optical engine, and the output power of the laser channels. For example, if the output of each laser channel is sufficient to power the transmission of 4 x 100G lanes, this would require 128 laser channels in a 51.2 Tbps system. For ELS module carrying 8 laser channels, 16 of such modules would be required.

6.2 Faceplate comparison

Once connectors have been selected, any available remaining space could be used for cooling or other components at the switch faceplate. Figure 6-3 provides a visual representation of potential connector footprint area. Higher density connector configurations all allow significant reduction in overall space, providing room for other components or reducing the switch from a 2U to a 1U design.

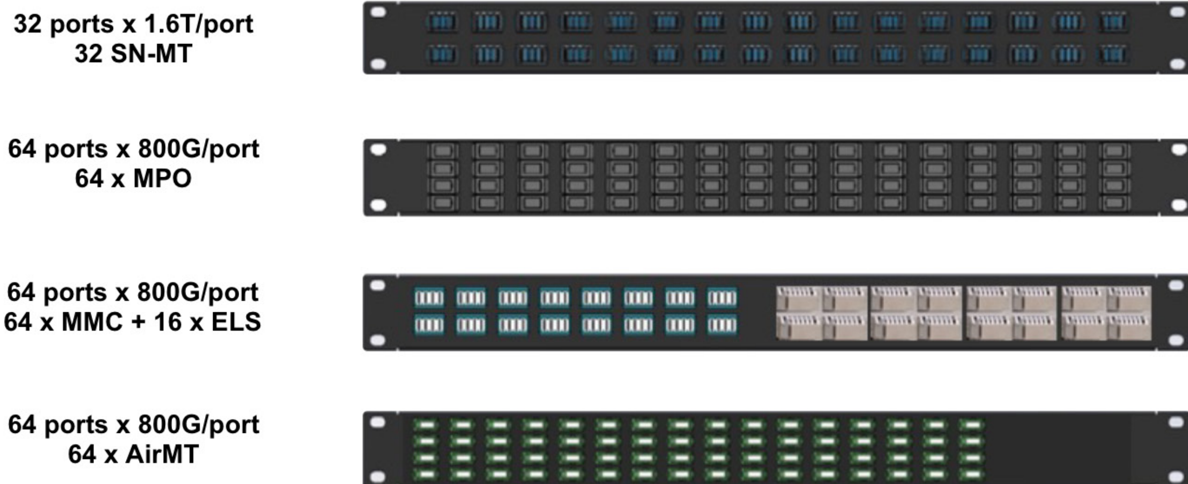

Figure 6-3: Faceplate connector footprint (left) and air-hole footprint (right)

Copyright © 2022 COBO. All Rights Reserved.

While 64 MPO ports in a 1U panel is possible, there is only a small amount of faceplate area left for airflow or ELS modules. Smaller, denser, or ganged connectors would be needed to free front panel area for other components. Ganged variants or multiport adapters are shown for SN-MT and MMC connectors in Figure 6-3, which also help reduce overall faceplate area used by optical connectors. Higher density connectors such as AirMT, SN-MT and MMC shown in Table 6-2 carry 16-fiber per connector and would have lower panel area use for an equivalent number of fibers. These benefits are apparent when comparing configurations in Figure 6-3.

Configurations could be envisioned where ELS modules and a connectors could be incorporated within a 1U based design. The addition of both the ELS module and optionally the ELS optical connector would require additional faceplate area to incorporate. The configuration and layout of such a design is likely to be specific to that module and require layout of both ELS and fiber connectors in a manner that matches the internal fiber layout.

With a CPO or OBO switch having electrical to optical connections within the switch itself, connector choice availability and placement is no longer constrained to only the pluggable module design. Smaller footprint connectors can be implemented and still maintain the number of fibers or ports necessary for a 51.2 Tbps bandwidth switch. The connector choice should be made in combination with other design considerations. Operational factors such as connector cleanability and breakout capability will need to be analyzed to ensure acceptable ease of use while in operation. The layout of the connectors and ELS modules is also important, as it affects the fiber routing and management inside the switch box. However, specific port layout recommendations are beyond the scope of this whitepaper. The faceplate configuration requires careful consideration of thermal management, particularly since the CPO switch box could potentially need to dissipate a kW or more.

Copyright © 2022 COBO. All Rights Reserved.

7. Optical test and measurement

Compared to fiber links that use pluggable transceivers, those that use co-packaged transceivers contain more fiber connectors. These additional connectors include those on the faceplate of the co-packaged switch as well as any mid-board connectors. The characteristics of these connectors must be carefully considered to ensure that co-packaged optics are 1) interoperable with pluggable modules and 2) backwards compatible with structured cabling already installed in a data center. Satisfying these two criteria are essential for the wide adoption of CPO.

7.1 Location of TP2 and TP3

Application standards include test points throughout a link often labeled “TP”. In IEEE 802.3 Ethernet standards two important test points are TP2 and TP3. The optical transmit signal properties are defined and measured at TP2 and the optical receive signal is defined at TP3. For pluggable transceivers TP2 is at the output of a short patch cord that connects to the pluggable module transmit port. TP3 is at the output of the fiber cable that connects to the pluggable module. The optical loss between TP2 and TP3 due to fiber attenuation and connector and splice loss is included in the power budget. Note that any loss between the transceiver and the fiber connectors at the transceiver port is not included in the power budget. These connection points are shown in Figure 7-1.

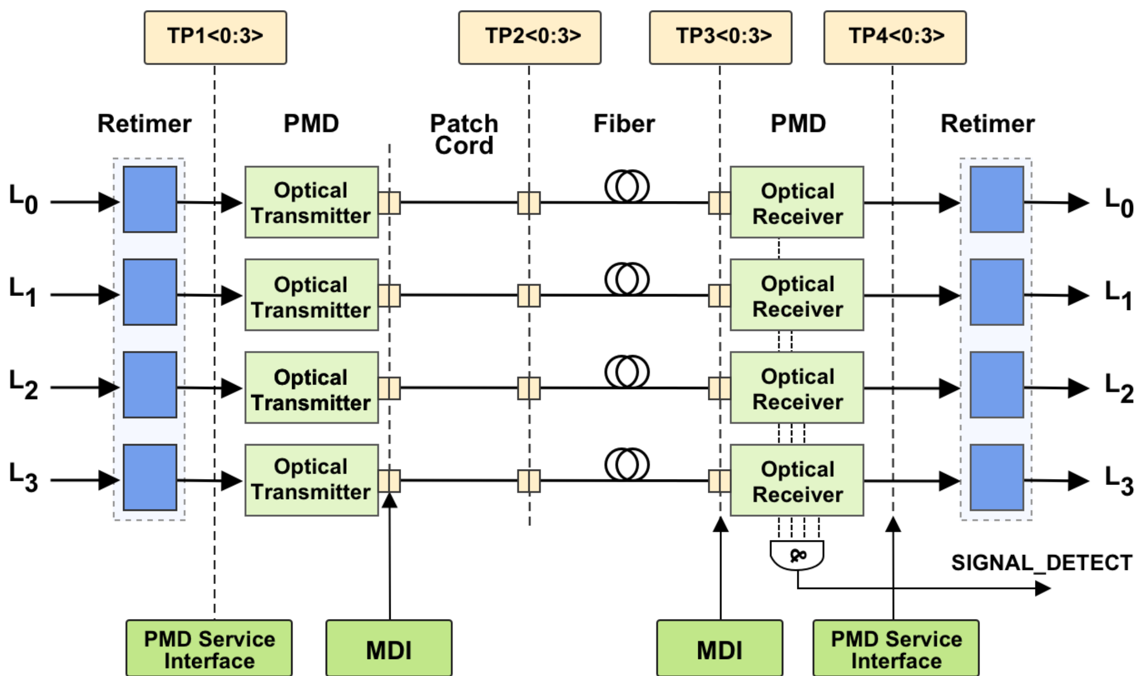

Figure 7-1: Test Point Definitions

Copyright © 2022 COBO. All Rights Reserved.

For co-packaged switches there is not a transceiver interface accessible at the faceplate. Instead there are fiber connectors on the faceplate. To maintain backwards compatibility with installed cable plants in data centers the optical loss of the faceplate connectors must be considered as a separate power budget line item.

Installed cable plants are built with pluggable standards in mind and it cannot be assumed that there is extra loss available in the link budget. TP2 and TP3 must be located at the same place as with pluggable modules. It may be useful to define additional test points (named TP2' and TP3') at the mid-board connectors to characterize the optical engines during switch assembly. A proposed point for these two test points are shown in Figure 7-2.

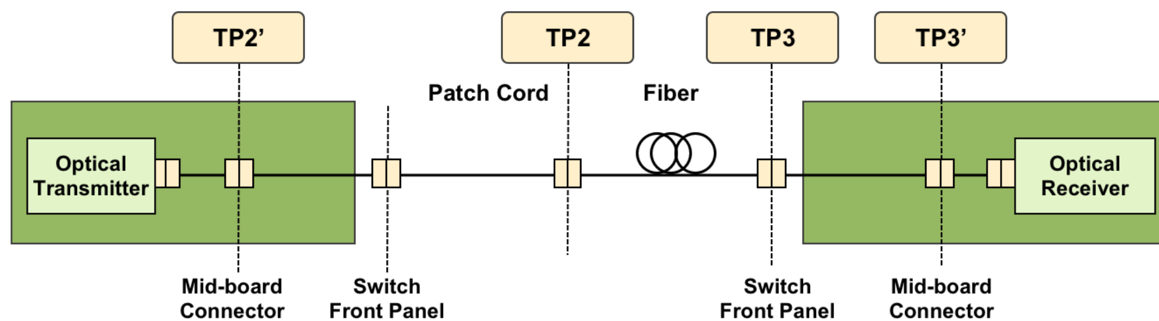

Figure 7-2: Potential additional test points for signal measurements

7.2 Faceplate connector loss

The reliability of CPO switches is important to end users looking to adopt this new technology. The optical loss of the faceplate connectors must be known and well controlled. Co-packaged switches may have up to 1024 fibers on the faceplate, so the loss statistics must be considered. If standard loss values such as those found in IEC 61753-3-31 are used, the long distribution "tail" will lead to connector loss budgets up to 1 dB. This is much too high for cost-effective, low power co-packaged optics systems. Instead, co-packaged switches should only use faceplate connectors with loss values better than standard. These connectors may have names like Ultra Low Loss and will have maximum loss values of 0.35 dB for single mode and 0.2 dB for multimode connectors. Using these connectors will ensure that power budgets will be practical and loss targets will be met with high reliability.

Copyright © 2022 COBO. All Rights Reserved.

7.3 Power budget adjustment for CPO

Compared to power budgets written with pluggable modules in mind, the additional loss due to extra CPO connectors must be added to both the transmit power and receiver sensitivity. This is necessary to ensure interoperability with pluggable modules. Compared to transmission between two pluggable modules, the case where a CPO switch transmits, and a pluggable receiver includes more loss due to the faceplate and mid-board connectors. The standard for the pluggable module is already written and the receiver sensitivity of the pluggable cannot be adjusted. The only option is to increase transmit power. Likewise, if a pluggable module transmits and the CPO switch receives, the only available option is to improve the sensitivity of the CPO receiver to compensate for the additional loss. The CPO connector loss will show up twice in the power budget: in the increased transmit power and in the improved receiver sensitivity. It is imperative that the connector loss be maintained at reasonable levels to ensure that CPO switches operate at low power and are cost effective.

Any loss incurred coupling from the optical engine to the first fiber need not be included in the application power budget. Like pluggable modules, the transmit signal will be characterized after the optical engine couples into the first fiber. This loss will be included in the required transmit power.

Copyright © 2022 COBO. All Rights Reserved.

8. Cooling systems

There are five levels of data center cooling system: chip level, server (device) level, rack level, plenum level and room level [18]. The device level cooling system of a data center is to transfer the heat generated by the operating electronic devices to the rack level room air or cooling distribution unit (CDU) in a timely manner. Combining all the five levels, the focus of the data center cooling system is to ensure that the temperature of all devices is stable within a safety range. The device level cooling system should work properly using the thermal budget between the hot devices and the ambient cold source, applying efficient cooling methods. Solutions described here are limited to the cooling process for the switch assembly and not the overall datacenter, as cooling methods can vary dependent on location, size, and layout.

8.1 Thermal challenge

With the rapid growth of switching ASIC bandwidth and computational power consumption, the temperature of electronic devices in a switch will increase if cooling systems cannot remove the increased heat generation. With a new design, CPO type switches will challenge thermal management due to the compact layout of hot optics and the ASIC switch. An ideal device level cooling system should work robustly and efficiently between the hot CPO parts and the ambient cold source, with easy assembly and maintenance. Table 8-1 provide the estimated power consumption values of 25.6 and 51.2 Tbps CPO assemblies.

Device	Power (W)	Quantity	Total Power (W)
Switch ASIC	600 est.	1	600
CPO	32 est.	16	512
Total Power	1112 W		
Switch ASIC	830 est.	1	830
CPO	64 est.	16	1024
Tower Power	1854 W		

Table 8-1: 25.6-51.2 Tbps CPO Power Consumption Estimation

Under the existing technical conditions, the power of 25.6 Tbps CPO assembly and the next-generation 51.2 Tbps CPO assembly are estimated to reach 1112 W and 1854 W, respectively, excluding the power consumed by electrical frequency conversion.

8.2 Implementation factors for cooling designs

Figure 8-1 shows a hierarchy of cooling methods, starting from the two main cooling methods: air cooling and liquid cooling. Air cooling can be further divided as solid metal air cooling and two phase enhanced air cooling, based on the implementation choices for thermal management and heat distribution. For liquid cooling methods, classifications depend on the liquid coolant remaining a single phase or designed with two phase cooling.

Copyright © 2022 COBO. All Rights Reserved.

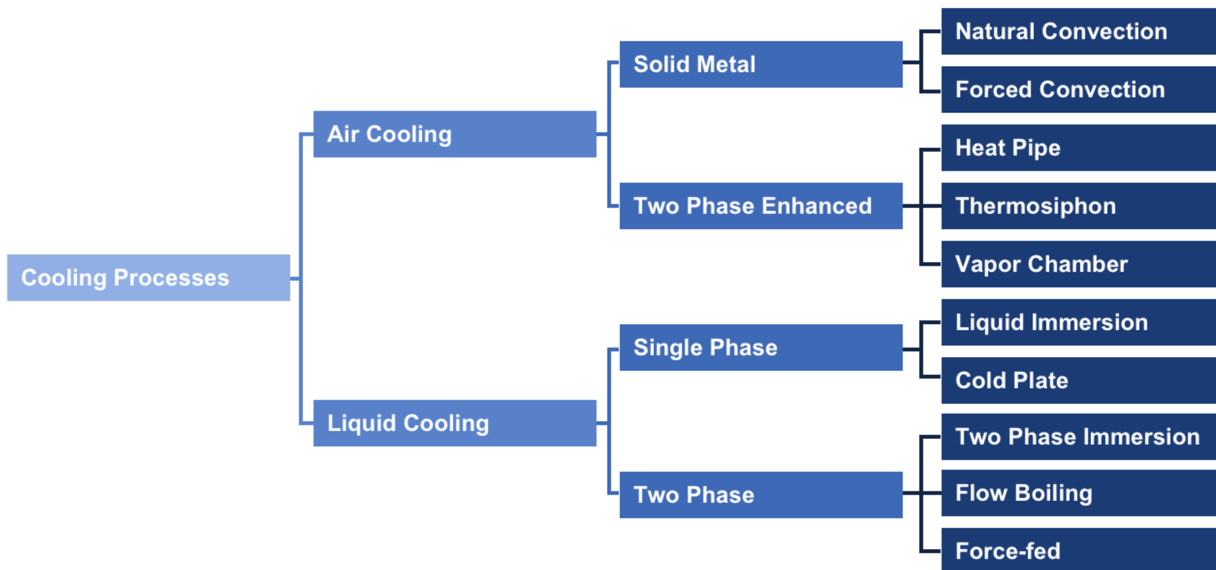

Figure 8-1: Cooling Methods Classification

Solid metal cooling methods

The air-cooling method is a well-established cooling method used within the current data center infrastructure. Solid metal air cooling adopts solid metal as the base structure and thermal spreader to achieve the heat dissipation of devices. The heat generated by the working electronic devices is conducted to the heat sink through the thermal interface material (TIM) and the solid metal spreader. Heat is then removed by the air flowing over the heat sink surfaces. The cooling mechanism for a heat sink is shown in Figure 8-2 showing air blowing across the ASIC heatsink.

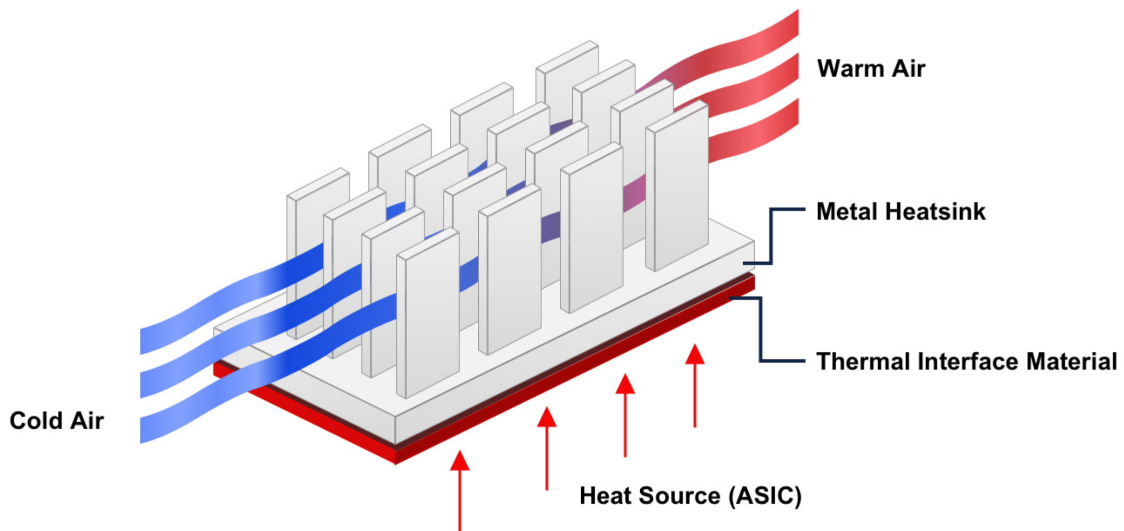

Figure 8-2: Air cooled heat sink cooling mechanism

Copyright © 2022 COBO. All Rights Reserved.

Thermal interface material (TIM) is used to cover the incomplete contact between the heat sink and the chip components. The TIMs family includes but is not limited to the following materials: phase changing material / composition (PCM / PCC), metal solders, thermally anisotropic composite, carbon-based materials, polymer-based materials, or liquid metals. Materials are provided as a solid thermal pad, thermal grease, or thermal gel. Property ranges for commercially available materials are shown in Table 8-2.

TIM Type	Thermal Conductivity (W/ (m•K))	TIM Thickness (mm)	Temperature Range (°C)	Dimensional Clearance (mm)
Thermal Pad	1-20	0.5 - 14	- 40 to 150	Different fit clearances possible
Thermal Grease	0.6~7.0	<0.1	- 50 to 260	less than 0.1mm
Thermal Gel	0.6~9.0	<0.15	- 50 to 260	less than 0.1mm

Table 8-2: Thermal Interface Material Properties

Fans are key components for the air cooling system. The key fan issues are flow direction, pressure drop, flow rate, redundancy, and ease of replacement. These factors influence the flow control of air through the switch box. To feed the cold air to the hot regions in a switch, the main heat sources' physical layout, front panel openings, guiding structure, side baffles, fin type and direction should be optimized together with heatsink and fans power characteristics. Simulation and testing are needed to validate temperature profiles. For narrow switch chassis design configurations, the heat sink plus fan assembly may be too tall and more complicated design types are needed. Fans and heat sinks can be separated and the commonly adopted fin types includes aluminum profile, interrupted-fin, folded-fin, or radiation enhanced fin designs.

Two phase enhanced air cooling methods

Two phase enhanced air-cooling systems refer to the air-cooling systems which are enhanced by locally installed two phase fluid containers. These designs leverage the latent heat of vaporization to absorb and transfer heat, using heat pipes, thermosiphons, or vapor chambers. The suitability of each design varies dependent on the geometry and available cooling systems. Vapor chambers are preferred for local heat sinks, thermosiphons for remote heat sinks, while the heat pipe allows for flexibility in shape and relative position. Figure 8-3 shows heat transfer schematics for heat pipes and vapor chamber designs. Heat pipes use a structured wall surface to transport liquid back to the evaporator zone. When air cooled, heat pipes need heat sinks with an open area in the chassis to contain cooling elements and equipment. Thermosiphons are modifications of heat pipes, where gravity is used to move condensed liquid back to the evaporator.

Copyright © 2022 COBO. All Rights Reserved.

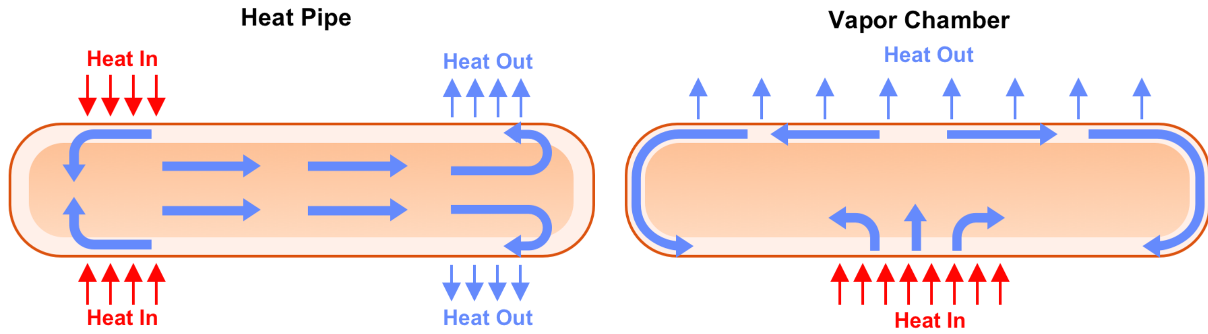

Figure 8-3: Two phase enhanced air cooling mechanisms

For a given diameter, thermosiphons have a higher maximum heat transport capacity than heat pipes. Thermosiphons can also carry heat farther distances than heat pipes because the working fluid flows back to the evaporator along the smooth or grooved inner walls. Vapor chambers allow for heat to spread evenly using the same mechanism as heat pipes. Vapor chambers can be used with air-cooled heat sinks or incorporated into more complex designs to spread heat across a large surface area. Since two phase methods provide additional geometric flexibility, height constraints are less critical and can allow 1U space designs.

Single phase liquid cooling methods

As power generation within switches and other datacenter components continue to increase, the maximum capacity of air heat removal is around 37 W/cm2 begins to constrain designs and options for only air cooled technologies. Liquid cooling allows for higher heat transfer and is an efficient method that can be directly or indirectly implemented in the datacenter. Single phase liquid cooling has three forms which are fundamentally the same process with different engineering designs. In Figure 8-4, two cooling designs for single phase liquid cooling are shown.

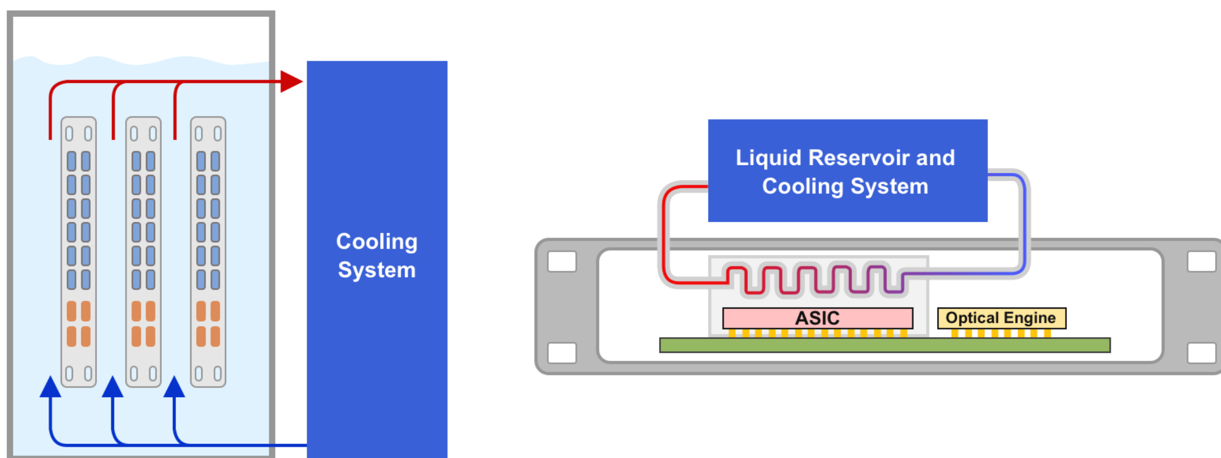

Figure 8-4: Single phase liquid cooling designs

Copyright © 2022 COBO. All Rights Reserved.

The left schematic is liquid immersion cooling, which applies non-evaporating coolant to transfer the heat to CDU. If all components are fully immersed, this would be difficult to implement where optical connectors may be contaminated with coolant. While ruggedized connectors exist, parts are bulky and may limit the overall serviceability and design density requirements.

Chip immersion cooling limits immersion to only the chips or surrounding area and is shown in the center image of Figure 8-4. Circulation and coolant volume are limited due to design requirements, and cooling systems are currently custom to the co-packaged module. An alternative design to chip immersion is cold plate cooling, which uses forced convection in channels to cool one or more components. This method is an application of a demonstrated similar technology in automotive engines and radiators. Single-phase cold plate cooling is a heating process of circulating coolant where no phase change occurs. Water is the most practical coolant due to superior thermophysical properties and high boiling point, but a leak risks component damage. Selecting non-conductive coolants may lower damage risks and help with leakage mitigation.

Two phase liquid cooling methods

Two phase liquid cooling takes advantage of the large latent heat of evaporation to absorb large amount of energy. Immersion and isolated systems are shown in Figure 8-5. A two phase fluid near its boiling point is an ideal coolant to absorb, transfer and discharge heat. Two phase immersion cooling, or pool boiling, immerses the heat generating components in a large pool of liquid. Vapor from the pool is cooled at the on-site condensing unit. Immersion cooling can support more than 200 kW power consumption and can reach a PUE (Power Usage Effectiveness) around 1.1 over 20 years lifespan. This method is limited to components which are compatible with liquid immersion because of sealing issues, but remains an option due to significant benefits in heat dissipation and vibration free cooling.

Flow boiling methods are in development to utilize two phase liquid cooling in a closed loop circulating fluid system. Two possible options are loop heat pipes or loop thermosiphons. Liquid evaporates while flowing through the hot spot, similar to the heat pipe designs shown in Figure 8-3. The condenser to remove heat is remote and independent of the server rack. The key disadvantage to this method is dry out risks at the hot spot, which would significantly reduce the cooling performance. Research and reliability testing is ongoing on new liquids and cooler designs to improve performance and reduce the risk of dry out.

Copyright © 2022 COBO. All Rights Reserved.

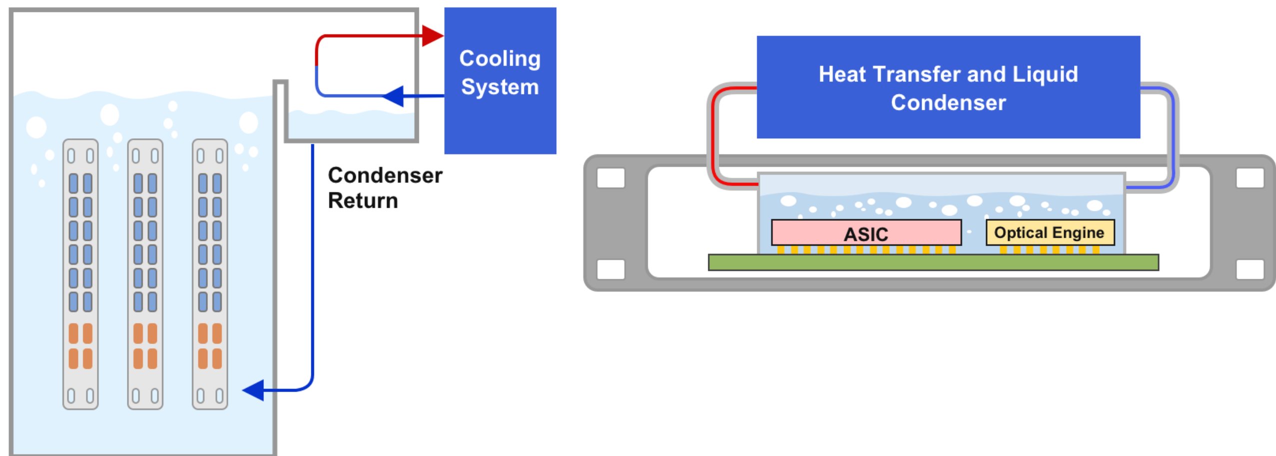

Figure 8-5: Two phase liquid cooling designs

Force-fed cooling is a more advanced design of flow boiling which attempts to enhance liquid convection across a hot surface. While this cooling method one of the most efficient cooling methods in theory, limited implementations are in use due to potential sealing issues, complex installation, and process control.

8.3 Cooling methods for CPO switch assembly

For the CPO assembly, the concentration of high-power devices creates challenges for air cooling. Since all components are now centrally located, heat must be transported away or sufficient air space provided to effectively cool the ASIC and OE components temperature, Tj, to an acceptable thermal level. As an example, the liquid cooling options of cold plate heat exchange design integration is provided.

A proposed template is shown in Table 8-2. This design is a type of cold plate cooling rack with 8 liquid cooling modules (each module's cooling capacity is 700-1200 W and maximum heat flux density is 150 W/cm^2). Meanwhile, 2 liquid-to-liquid cooling distribution units (CDU) are assembled, while each of them has 2 pumps (1+1 redundancy). A liquid-to-air CDU is assembled to cool the coolant from the two liquid-to-liquid CDUs by the room air. For the entire liquid cooling system, 3M™ Fluorinert™ Electronic Liquid FC-40 is chosen as the cooling liquid at rack side, avoiding short circuit caused by working fluid leakage. For a sense of scale, Figure 8-6 shows a cooling unit next to a switch rack.

Copyright © 2022 COBO. All Rights Reserved.

Figure 8-6: Liquid-to-air CDU (left) and Switch Rack with 2 CDUs (right)

The results of thermal simulation analysis show that the use of cold plate heat exchanger can effectively control the temperature of the CPO Assembly within a sufficient margin from the device specifications. Figure 8-7 and Figure 8-8 show the CPO Assembly temperature distribution as defined in Table 8-3.

25.6Tbps	Power (W)	Quantity	Total Power (W)	Tj, °C
Switch ASIC	600 est.	1	600	92.2
CPO	32 est.	16	512	52.4
Switch cold plate flow	4 L/min, inlet temp 40 °C, outlet temp 44.3 °C			
CPO cold plate flow	4 L/min, inlet temp 40 °C, outlet temp 45.7 °C			
Total Power	1112 W			
51.2Tbps	**Power (W)**	**Quantity**	**Total Power (W)**	**Tj, °C**
Switch ASIC	830 est.	1	830	97.2
CPO	64 est.	16	1024	59.6
Switch cold plate flow	7 L/min, inlet temp 40 °C, outlet temp 43.3 °C			
CPO cold plate flow	7 L/min, inlet temp 40 °C, outlet temp 46 °C			
Total Power	1854 W			

Table 8-3: 25.6 Tbps and 51.2 Tbps CPO Assembly Simulation Condition

In Figure 8-7 and Figure 8-8, heat is effectively managed in both cases. The increase in heat for the 51.2 Tbps is reflected in the higher overall temperature of all components through the simulated design.

Copyright © 2022 COBO. All Rights Reserved.

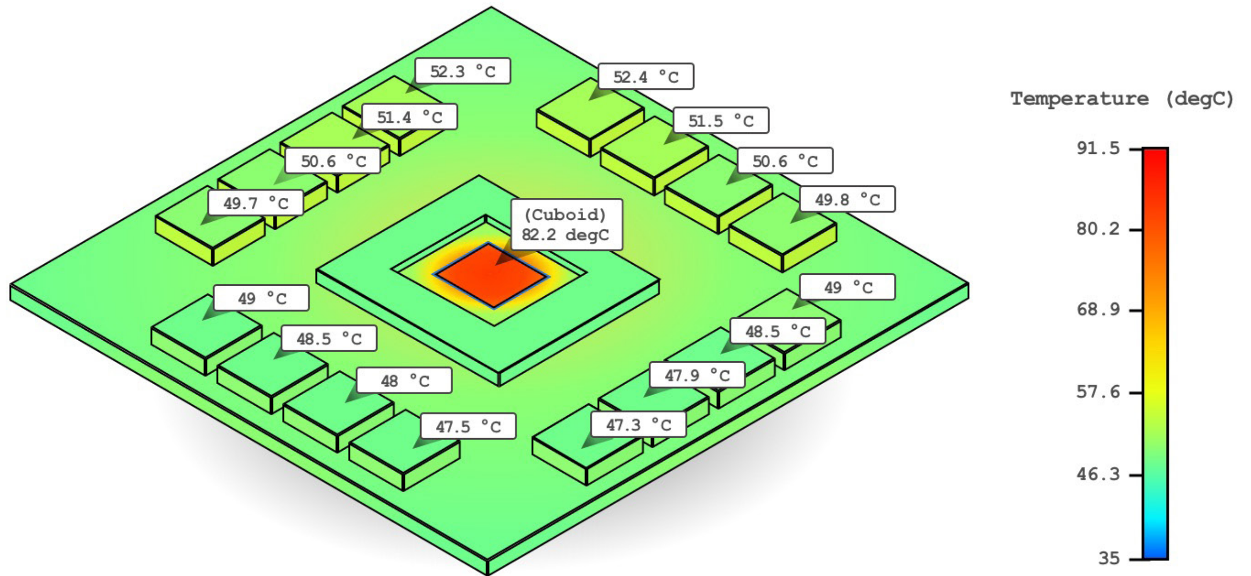

Figure 8-7: 25.6 Tbps CPO Assembly Temperature Distribution Simulation

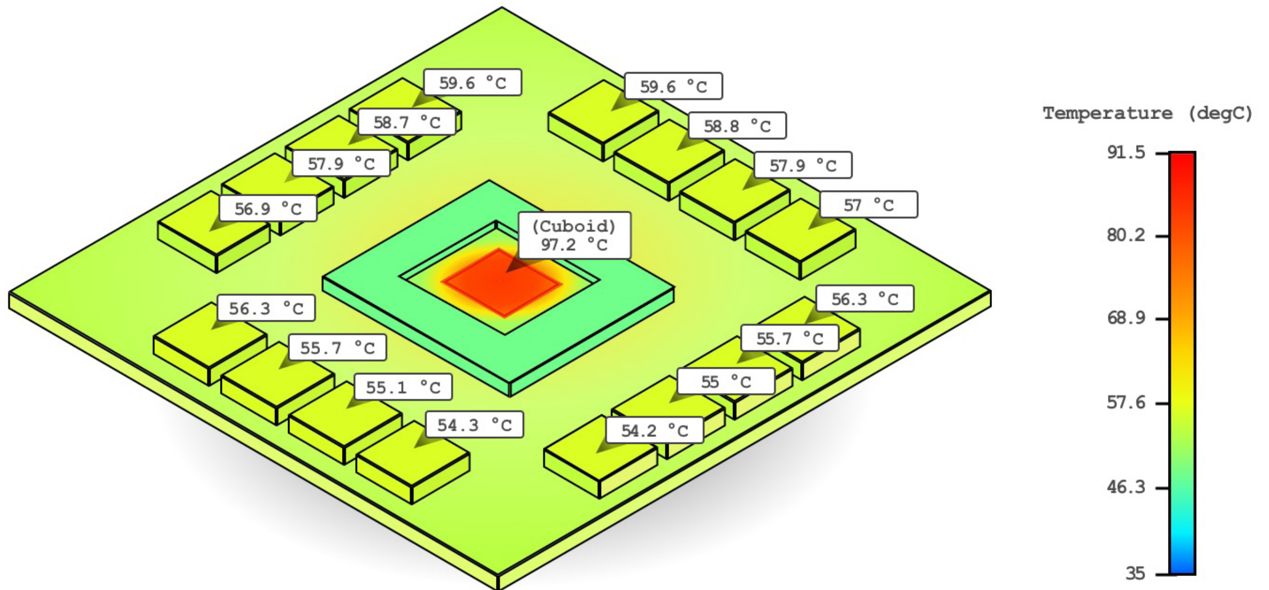

Figure 8-8: 51.2 Tbps CPO Assembly Temperature Distribution Simulation

Copyright © 2022 COBO. All Rights Reserved.

8.4 Comparison of cooling methods

With the understanding that the system could be cooled using cold plate technology as shown in the example design, another key consideration is the energy efficiency of the overall design. PUE is the index to evaluate the energy efficiency of data centers. PUE is calculated using the following parameters:

$$PUE = \frac{Datacenter\ total\ power\ consumption}{IT\ device\ power\ consumption}$$

For the application of different cooling methods in data center, the commonly seen PUE range is shown in Figure 8-9, however, it may vary based on the equipment and system design. It can be seen that the PUE of immersion cooling can reach the lowest value, approaching the theoretical limitation value 1.0.

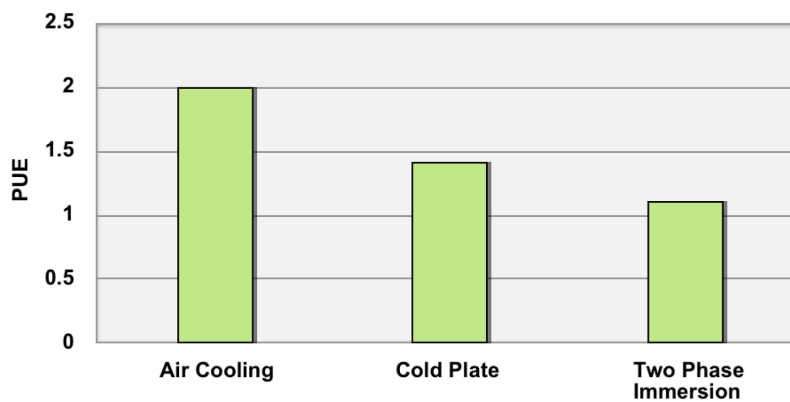

Figure 8-9: PUE of Different Cooling Methods in a Data Center

The PUE itself does not take into account any reuse of energy (heat recovery), therefore the Green Grid defined the Energy Reuse Effectiveness (ERE) as:

$$ERE = \frac{P_{cooling} + P_{power} + P_{lighting} + P_{IT} - P_{reuse}}{P_{IT}}$$

Where $P_{cooling}$ is the input power of the cooling equipment, P_{power} is the power energy lost in the power distribution system through line-loss and other infrastructure (UPS or PDU) inefficiencies, $P_{lighting}$ is the power used to light the data center and support spaces, P_{IT} is the input power of the IT equipment, P_{reuse} is the heat reuse factor. The cold plate and immersion liquid cooling technology have the advantages not only on PUE but also provide easy accessible energy recover and reuse option through the CDU.

Cooling method selection will impact the overall space within the switch design, and methods will need to be implemented in order for the optical engines and ASIC to remain within operational conditions. Air cooling, cold plate cooling. immersive cooling, and other cooling methods have benefits and drawbacks to the overall design of the switch. Key considerations are provided in Table 8-4.

Copyright © 2022 COBO. All Rights Reserved.

Property	Air Cooling	Cold Plate Cooling	Immersive Cooling
Heat Carrying Capacity	Low	High	High
Reliability	Good	Good	High
Maintenance	Easy	Complicated	Complicated
Failure Risks	Blocked air flow	Leakage	Leakage
Cost	Low	High	High
Noise / Vibration	High	None	None
Setup	Easy	Complicated	Complicated
Space Requirements	Low	Medium	High

Table 8-4: Comparison of Different Cooling Methods

Thermal efficiency is only a portion of the design requirements for a comprehensive cooling solution. Cooling capacity requirements, available space, environmental requirements, budget, and other factors will need to be evaluated to ensure switch components remain within operating temperature ranges.

9. Summary

Optical engine design, co-packaging method, connector selection, laser source requirements, and heat management all impact the design and space requirements necessary for co-packaged designs to exist within an optical switch. Industrial solutions to handle new requirements have started to become available to integrate all components into a smaller footprint than previously possible. Smaller connector options have recently been announced which allow for denser connections without significant impact to server size or airflow design. External laser sources and co-packaged designs with better heat management could enable a full design within a 1RU space footprint.

Copyright © 2022 COBO. All Rights Reserved.

Bibliography

[1] "Trends in Optical Networking," COBO, 2020.

[2] International Electrotechnical Commission, "IEC 61753-1:2018 Fibre optic interconnecting devices and passive components - Performance standard - Part 1: General and guidance," International Electrotechnical Commission, 2018.

[3] Consortium For On-Board Optics, "Optical Connectivity Options for 400 Gbps and Higher On-Board Optics," COBO, 2019.

[4] Co-Packaged Optics Collaboration, "Co-packaged Optics External Laser Source Guidance Document, v 1.0," Co-Packaged Optics Collaboration, 2020.

[5] OSFP MSA, "OSFP Octal Small Form Factor Pluggable Module Specification Rev 2.0," OSFP MSA, 2021.

[6] QSFP-DD MSA, "QSFP-DD/QSFP-DD800/QSFP112 Hardware Specification Rev 6.01," QSFP-DD MSA, 2021.

[7] International Electrotechnical Commission, "IEC 61754-20:2012 Fibre optic interconnecting devices and passive components - Fibre optic connector interfaces - Part 20: Type LC connector family, Ed 2," International Electrotechnical Commission, 2012.

[8] Telecommunications Industry Association, "FOCIS 19 Fiber Optic Connector Intermateability Standard- Type Sen Connector," Telecommunications Industry Association, 2021.

[9] International Electrotechnical Commission, "Fibre optic interconnecting devices and passive components – Fibre optic connector interfaces- Part 37: Type MDC connector family," International Electrotechnical Commission, 2021.

[10] International Electrotechnical Commission, "IEC 61754-36 Fibre optic interconnecting devices and passive components - Fibre optic connector interfaces. - Part 36: Type SAC connector family (Draft)," International Electrotechnical Commission, 2021.

Copyright © 2022 COBO. All Rights Reserved.

11] Telecommunications Industry Association, "TIA 604-5: FOCIS 5 Fiber Optic Connector Intermateability Standard- Type MPO, Rev E," Telecommunications Industry Association, 2015.

[12] International Electrotechnical Commission, "IEC 61754-7-1:2014 Fibre optic interconnecting devices and passive components - Fibre optic connector interfaces - Part 7-1: Type MPO connector family - One fibre row," International Electrotechnical Commission, 2014

[13] Telecommunications Industry Association, "TIA 604-18: 2015 FOCIS 18 Fiber Optic Connector Intermateability Standard- Type MPO- 16," Telecommunications Industry Association, 2015.

[14] International Electrotechnical Commission, "IEC 61754-7-4 Fibre optic interconnecting devices and passive components – Fibre optic connector interfaces – Part 7-4: Type MPO connector family – One fibre row 16 fibres wide," International Electrotechnical Commission, Draft 2019.

[15] Telecommunications Industry Association (TIA), "FOCIS 5 Fiber Optic Connector Intermateability Standard- Type MPO," Telecommunications Industry Association (TIA), 2019.

[16] International Electrotechnical Commission, "IEC 61754-7-2:2017 Fibre optic interconnecting devices and passive components - Fibre optic connector interfaces - Part 7-2: Type MPO connector family - Two fibre rows," International Electrotechnical Commission, 2017.

[17] International Electrotechnical Commision, "IEC 61754-7-3:2019 Fibre optic interconnecting devices and passive components - Fibre optic connector interfaces - Part 7-3: Type MPO connector family - Two fibre rows 16 fibre wide," International Electrotechnical Commision, 2019.

[18] A. H. Khalaj and S. K. Halgamuge, "A Review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system," Applied Energy, vol. 205, pp. 1165-1188, 2017.

[19] M. Saini and R. L. Webb, "Heat rejection limits of air cooled plane fin heat sinks for computer cooling," IEEE Transactions on Components and Packaging Technologies, vol. 26, no. 1, pp. 71-79, 2003.

Copyright © 2022 COBO. All Rights Reserved.

11] S. V. Garimella, L.-T. Yeh and T. Persoons, “Thermal Management Challenges in Telecommunication Systems and Data Centers,” IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 2, no. 8, pp. 1307-1316, 2012.

[12] S. V. Garimella, T. Persoons, J. A. Weibel and V. Gektin, “Electronics Thermal Management in Information and Communications Technologies: Challenges and Future Directions,” IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 7, no. 8, pp. 1191-1205, 2017.

[13] International Electrotechnical Commission, “Fibre optic interconnecting devices and passive components. Fibre optic connector interfaces Part 37. Type MDC connector family,” International Electrotechnical Commission, 2021.

Copyright © 2022 COBO. All Rights Reserved.

Co-packaged optics working group members

3M

Accelink Technologies Co., Ltd.

ADVA Optical Networking SE

AIO Core Co., Ltd.

Amphenol Corporation

Applied Optoelectronics Inc

Arista

Broadcom

Celestica Inc.

Ciena Corporation

CommScope Of North Carolina

Cudoform Inc.

Dell Inc.

DuPont Specialty Products USA, LLC

Foxconn Interconnect Technology, Ltd. (FIT)

Fujikura Ltd.

Fujitsu Optical Components

Furukawa Electric

Globalfoundries

Hirose Electric (U.S.A.) Inc.

Hisense

I-PEX Inc.

II-VI Incorporated

Intel Corporation

LESSENGERS Inc.

LIPAC Co., Ltd.

Marvell Asia Pte Ltd.

Microsoft

NEC Corporation

NVIDIA

Optomind Inc.

Panasonic

PETRA

Quanta Computer Inc.

Ragile Networks

RANOVUS

Reichle & De-Massari AG

Rosenberger

Sabic

Semtech

SENKO Advanced Components

Sicoya GmbH

Sumitomo Electric Industries

TE Connectivity Corporation

TTM Technologies Trading (Asia) Co., Ltd.

US Conec Ltd.

Vario-Optics AG

Copyright © 2022 COBO. All Rights Reserved.

Points of contact

Tiger Ninomiya, SENKO, CPO Working Group Chairman, Tiger.Ninomiya@senko.com

Peter Johnson, SABIC, CPO Working Group Technical Editor, Peter.Johnson@sabic-hpp.com

Melissa Kallos, Consortium for On-Board Optics Editor, Melissa@onboardoptics.org

Copyright

Copyright © 2022 COBO. All Rights Reserved.

Limitations

THIS DOCUMENT IS PROVIDED "AS IS." The publisher and contributors expressly disclaim any warranties (express, implied, or otherwise), including but not limited to implied warranties of merchantability, non-infringement, fitness for a particular purpose, or title. The entire risk of using or implementing this document is assumed by the user. IN NO EVENT IS THE PUBLISHER OR ANY CONTRIBUTOR LIABLE TO ANY PARTY FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING LOST PROFITS) BASED ON ANY THEORY OF ACTION (INCLUDING NEGLIGENCE), EVEN IF THE OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE. This document may be subject to third party rights.