System Vendor Requirements Document for Energy Efficient Interfaces

OIF-EEI-Requirements-RD-01.0

May 9, 2024

Requirements Document created and approved

OIF

www.oiforum.com

The OIF is an international non profit organization with over 100 member companies, including the world's leading carriers and vendors. Being an industry group uniting representatives of the data and optical worlds, OIF's purpose is to accelerate the deployment of interoperable, cost-effective and robust optical internetworks and their associated technologies. Optical internetworks are data networks composed of routers and data switches interconnected by optical networking elements.

With the goal of promoting worldwide compatibility of optical internetworking products, the OIF actively supports and extends the work of national and international standards bodies. Working relationships or formal liaisons have been established with EA, IEEE 802.3, IETF, PCI-SIG, INCITS T11, Infiniband, IPEC, ITU-T SG15, SNIA-SFF.

For additional information contact:
OIF
5177 Brandin Ct, Fremont, CA 94538
510-492-4040 info@oiforum.com

www.oiforum.com

Working Group: **Physical Layer User Group**

TITLE: **System Vendor Requirements Document for Energy Efficient Interfaces**

SOURCE:	**TECHNICAL EDITOR**	**WORKING GROUP CHAIR**
	Eric Bernier	Jeffery Maki
	Huawei Technologies	Juniper Networks
	303 Terry Fox Dr. - Suite 100	1133 Innovation Way
	Kanata, ON, K2K 3J1	Sunnyvale, CA, 94089
	Canada	United States
	Phone: + 1 613 876 0196	Phone: + 1 408 936 8575
	Email: eric.bernier@huawei.com	Email: jmaki@juniper.net

ABSTRACT: The purpose of this document is to establish a common set of system vendor requirements for energy efficient I/O which addresses optical interfaces for high density applications, and which are limited by the total energy consumed on each interface. The document provides an overview of the applications requiring the interface as well as the key specification attributes needed for those applications to achieve maximum performance.

Table of Contents

List of Figures

List of Tables

Document Revision History

Working Group: Physical Layer User Group

SOURCE:	**TECHNICAL EDITOR**	**WORKING GROUP CHAIR**
	Eric Bernier	Jeffery Maki
	Huawei Technologies	Juniper Networks
	303 Terry Fox Dr. - Suite 100	1133 Innovation Way
	Kanata, ON, K2K 3J1	Sunnyvale, CA, 94089
	Canada	United States
	Phone: + 1 613 876 0196	Phone: + 1 408 936 8575
	Email: eric.bernier@huawei.com	Email: jmaki@juniper.net

DATE: **May 9, 2004**

---

1 Introduction

In today's fast-paced technological landscape, where data processing and communication demands continue to soar, the pursuit of energy-efficient solutions has become paramount. This document serves as a vital component aimed at providing a requirement document for Energy Efficient Interfaces (EEI). Rooted in the practice of engaging with end users to understand their unique needs and expectations, this project seeks to prioritize applications and establish fundamental criteria for the next generation of energy-efficient electrical and optical links.

In addressing the primary objective of this document, the pressing need is for Energy Efficient Interfaces for use in high-density application scenarios. The document outlines critical requirements to achieve optimal performance for various applications that necessitate the utilization of such interfaces. These requirements encompass a range of factors, including energy targets, desired form factors, degrees of interoperability, traffic types, and more.

2 Applications for Energy Efficient Interconnect EEI

This section delves into several distinct applications that play a pivotal role in shaping the future of data centers, each demanding enhancement in efficiency, scalability, and performance. These applications encompass optimizing data transmission, reducing latency, harnessing computational power, and streamlining resource management. AI applications particularly drive the need for scaling computer node performance, necessitating alignment with network structures and applications in the context of AI infrastructure.

Three primary networks currently command significant attention within the global industry, as illustrated in Figure 1: Front End Interconnect networks that also serve as the intra-datacenter interconnect; Back End Interconnect that serves to connect GPU-to-GPU in large scale clusters; and the Compute Interconnect used as the Back End Network within compute nodes.

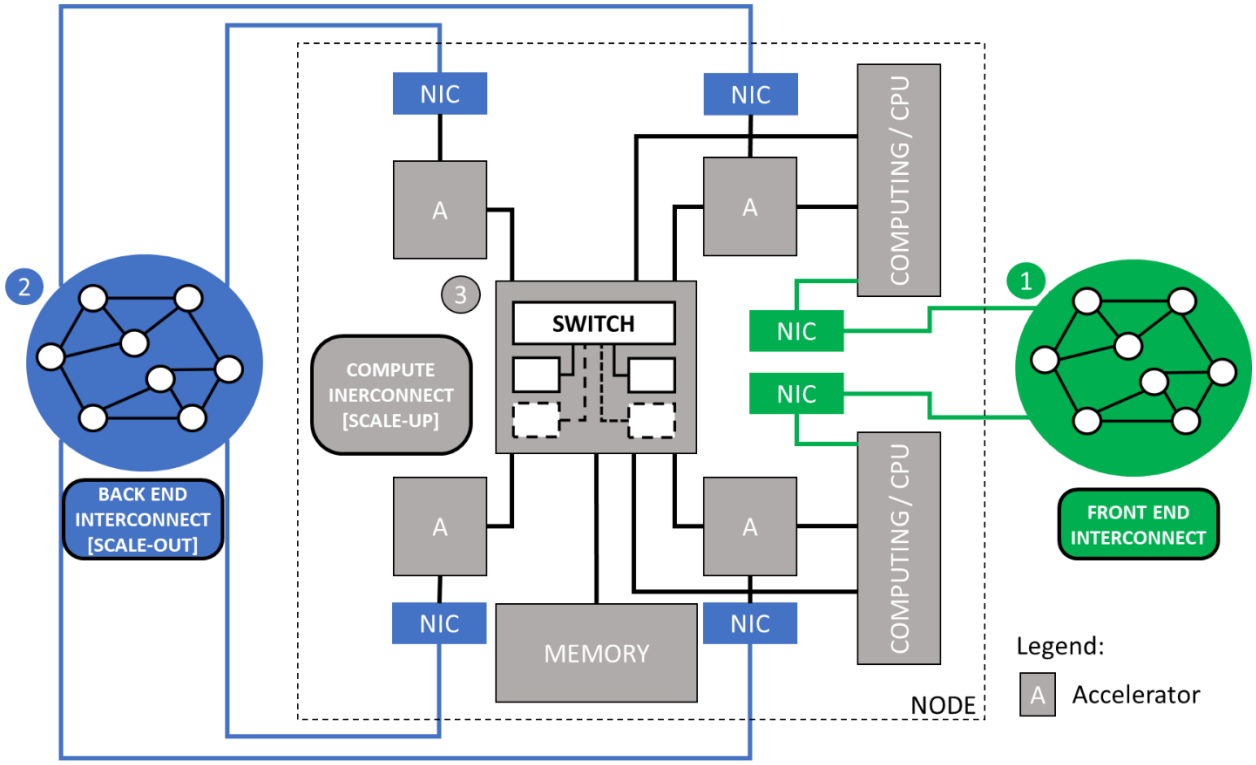


Figure 1 Conceptual representation of modern high-performance networks. (1) Front End Interconnect (FEI), (2) Back End Interconnect (BEI), and (3) Compute Interconnect

Front end networks encompass optical interconnections designed to facilitate seamless communication between servers. Whether it's server-to-server, server-to-storage, or server-to-edge communication, these networks define how servers interact with the external world. They cater to both general-purpose data center servers and specialized servers dedicated to AI, analytics, HPC systems, and more.

Back end networks are the specialized interconnects tailored for AI intra-cluster accelerator communication, HPC intra-cluster communication, and other high-performance computing tasks. Their purpose is to scale out the computing infrastructure, often extending to 1,000-10,000 nodes. This use case revolves around optical interconnections that facilitate accelerator-to-accelerator communication for executing or training AI models. The term 'accelerator' encompasses GPUs, TPUs, analog computers, and similar devices.

Compute Interconnect is the specialized interconnect that serves as the vital link connecting computing infrastructure, memory, and I/O with the compute nodes. Historically, these applications relied on PCI buses and have evolved to be connected through a switched configuration forming a simplified network. They are integral to scaling up the computing infrastructure. For instance, the implementation of pooled memory allows multiple nodes to access the same memory, reducing overall system memory consumption. This extends to pooled I/O, further optimizing resource utilization.

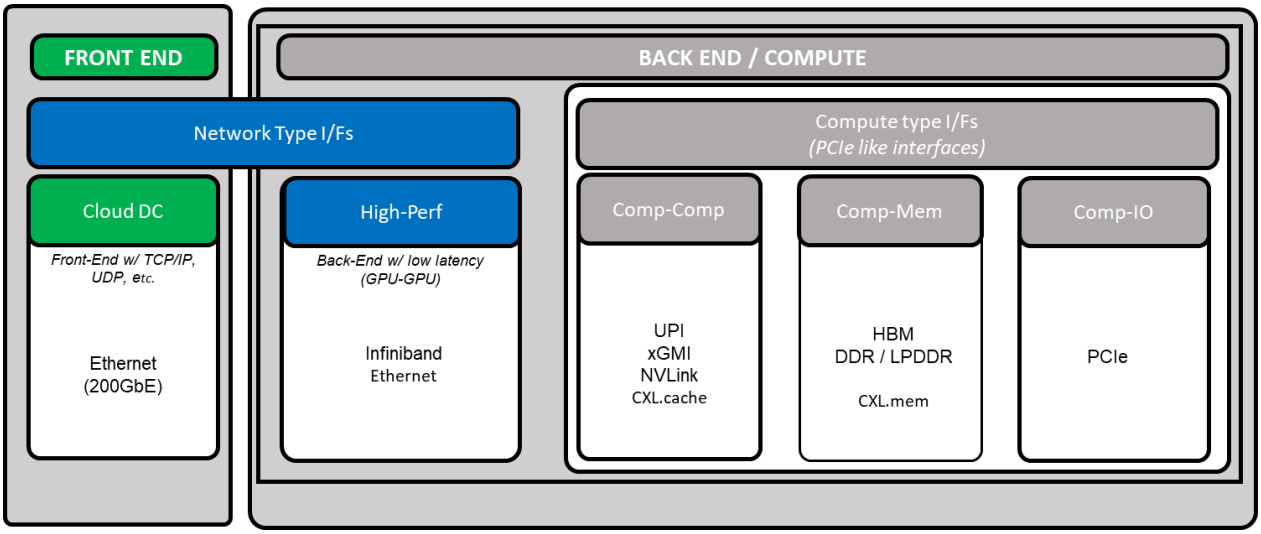


Figure 2 Representation of the modern high-performance networks with current busses [9]

2.1 Front End Interconnect (FEI)

Front End Interconnect (FEI), often referred to as "intra-datacenter interconnect," constitutes the foundational networking and communication infrastructure within a single data center facility. This comprehensive system encompasses switches, routers, cables, and protocols, all meticulously orchestrated to manage the seamless flow of data among servers, storage systems, and various hardware components within the data center environment. At its core, the primary objective of the FEI is to facilitate efficient data transmission, ensure low-latency communication, and maintain high resource availability, collectively sustaining the uninterrupted operation of critical applications and services.

The FEI Use case: Compatibility and Efficiency

Within the FEI use case, a multitude of systems coexist, demanding robust infrastructure compatibility over extended timeframes. This network places a paramount emphasis on system interoperability, both backward and forward compatibility. The pursuit of energy efficiency enhancements is also a driving force, allowing for increased density and bandwidth. However, it is noteworthy that these interfaces are already meticulously designed to achieve optimal efficiency, considering the specific implementation objectives, suggesting limited potential for further improvements in energy efficiency. In contrast, the current interconnect standards were not optimized to the BEI requirements. As new requirements and new specifications are explored for the BEI there is an opportunity to define new interfaces for new applications that will result in lower power consumption.

Connecting Accelerator Pods Control Interface

In the context of an AI cluster, the FEI is key in ensuring the seamless operation of these high-performance computing ecosystems. Its primary function lies in providing uninterrupted connectivity to the control interface of GPU pods. In the intricate orchestration of an AI cluster, these control interfaces play an indispensable role, serving as the command center for managing distributed computing resources accelerators like GPUs, TPUs, and others.

They facilitate instantaneous execution of control commands, swift data transfers, and real-time monitoring signals between the central management system and individual accelerator pods, all executed with minimal delay and maximum throughput. The network required robust connectivity imbues the cluster with responsiveness, empowering real-time adjustments, workload distribution, and resource allocation.

Figure 3 Representation of a large AI/ML cluster. In blue, the Front End Interconnect (FEI). In green, the Back End Interconnect (BEI). [6]

Table 1 Main characteristics: Front End Interconnect

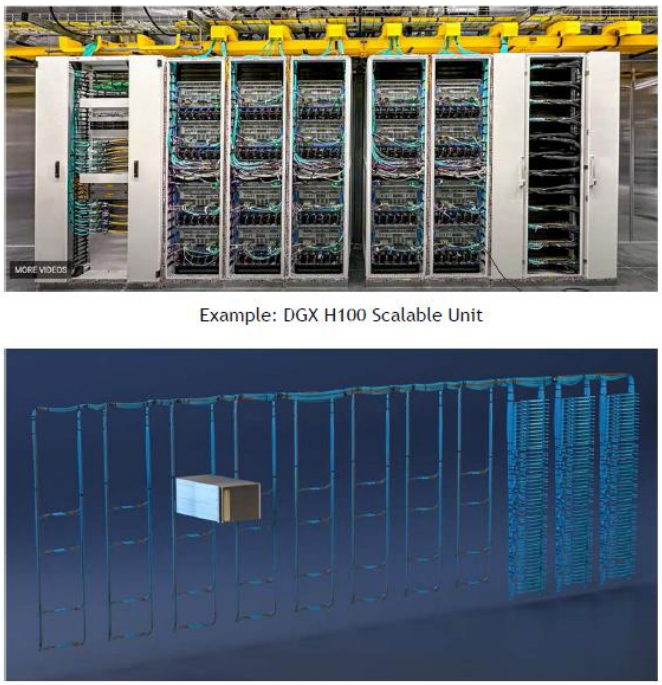
1st priority	Standard based, interoperability, backward and forward compatibility
2nd priority	Energy Efficiency
3rd priority	High-density implementations targeting throughput increase
Timeframe	Already in use
Market Size	Annual Sales Market Volume in 2023: 2141 million dollars (USD) Annual Sales Market Volume in 2027: 4390 million dollars (USD) (2x Increase)

2.2 Back End Interconnect

(Sometimes called Accelerator-to-Accelerator Remote Interconnect)

The Back End Interconnect (BEI), represents a critical component within the infrastructure of large GPU clusters and HPC systems, engineered to facilitate dedicated communication and seamless data exchange among individual GPUs housed within GPU cluster or communication between processors in a HPC system case. This interconnect serves as the enabler of parallel processing and data sharing between processors, a fundamental requirement for accelerating compute-intensive tasks such as deep learning, scientific simulations, and data analytics. Harnessing cutting-edge fabrics, the processor-to-processor interconnect establishes an ecosystem of ultra-low latency and high-bandwidth connectivity. There is currently limited implementation of the interconnect to implement the BEI. Current examples of BEI interconnect would be HPE Slingshot™ in the case of HPC Systems, or InfiniBand™.

Within specialized interconnects, the BEI use case emerges as the linchpin for AI intra-cluster accelerator communication. It's tailored to accommodate the scaling-out of the computing infrastructure, often spanning a scale of 1,000 to 10,000 nodes. This unique application spotlights the optical interconnections that facilitate accelerator-to-accelerator communication, essential for executing and training AI models. The term 'accelerator' encompasses a diverse range of devices, from GPUs and TPUs to Analog Computers and beyond.



Example: DGX H100 Scalable Unit

Figure 4 Large GPU cluster. Top: multiple GPU nodes of the DGX™ H100 are disposed into racks. Bottom: a single GPU node is illustrated with all else grayed out. It is the chassis within a scaled-out server. [6]

Looking ahead, the industry anticipates a transition toward modified interfaces based on Ethernet technology, with customized protocol stacks designed to ensure peak throughput and minimal latency.

Table 2 Main characteristics: Back End Interconnect

1st priority	Latency / Throughput
2nd priority	Low error rate for training efficiency
3rd priority	Energy Efficiency
4th priority	High-density implementations
Timeframe	Proprietary implementation in use. Major evolution within 2024-2025
Market Size	Annual Sales Market Volume in 2023: 1321 million dollars (USD) Annual Sales Market Volume in 2027: 3883 million dollars (USD) (3x increase)

2.3 Compute Interconnect

(sometimes called Accelerator-to-Accelerator Local Interconnect)

The Compute Interconnect (CI) represents a critical component within the node enabling the use of a large number of Accelerators (GPU). It is engineered to facilitate seamless data exchange among individual GPUs housed within computing node. The integration of optical interconnect technology within computing nodes marks a groundbreaking shift in the field. This advancement addresses the need to integrate an increasing number of accelerators, such as GPUs, into single nodes. As a result, these nodes are not just scaling up to accommodate hundreds or even thousands of GPUs, but also expanding in size, sometimes across multiple racks. In the CI application switch such as a CXL switch may be used to interconnect the various components. It means the CI is a single CXL domain network, FEI and BEI operate across multiple CXL network domains.

Figure 5 illustrates the current system architecture with PCIe as its central component. NVLink™ is also a current bus in use for Compute interconnect. These nodes, typically the size of a server, have historically used electrical interconnects for all interfaces. These interfaces include Accelerator-to-Accelerator (both remote and local), Accelerator-to-CPU, and Accelerator-to-Memory connections. Figure 2 presents a comprehensive list of the current protocols employed for these various functions.

Compute disaggregation, a revolutionary architectural concept, involves the separation of key computing components (processors, memory, storage) into distinct, standalone units. This modular approach allows for dynamic resource allocation, precisely matching the requirements of specific workloads. This strategy dramatically enhances resource utilization, scalability, and adaptability in data centers and cloud environments. It facilitates the efficient distribution of computing resources, tailored to the distinct demands of various applications. This leads to enhanced performance, cost-efficiency, and optimal resource management. Disaggregation might involve externally connecting memory blocks when they exceed the capacity of a single High Bandwidth Memory (HBM) or physical enclosure. While resource pooling is an effective strategy, it represents just one of many avenues in this field.

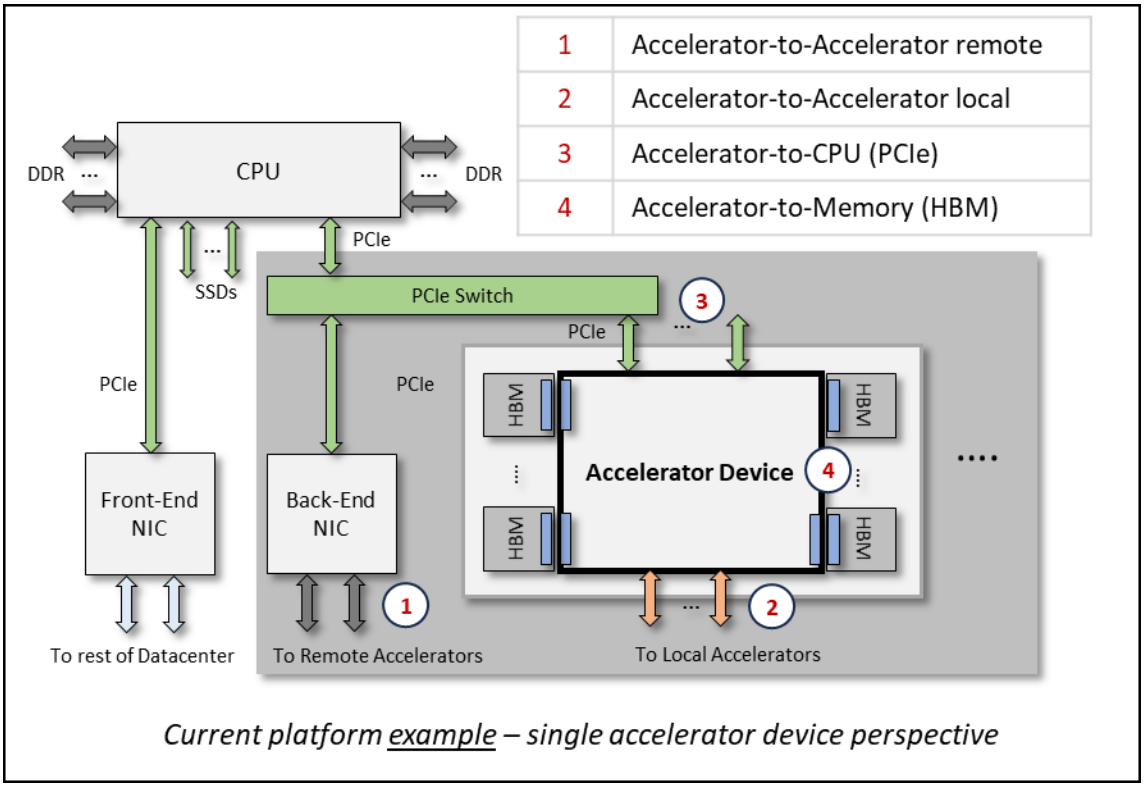
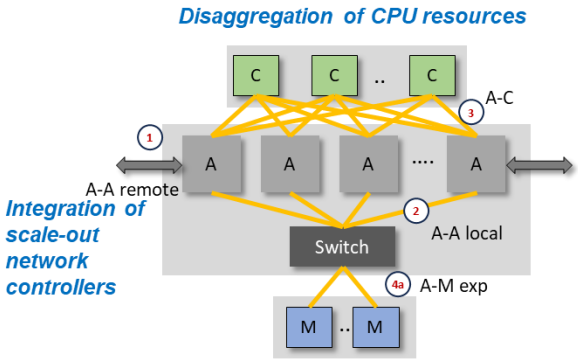


Figure 5 Current platform implementations within single server/node. [1]

It's crucial to acknowledge the challenges in achieving full system disaggregation in the near term, particularly considering memory access speed as a pivotal factor in system throughput. The introduction of extensively disaggregated memory could impede access speeds.

The most critical need for optical interconnects lies in Accelerator-to-Accelerator connectivity, especially for AI/ML applications, where each accelerator requires a substantial amount of interconnection to achieve optimal performance. The industry is leaning towards developing interfaces that are largely homogeneous, facilitating reuse, expansion, sharing of devices, and cost reduction. Given the massive scale required for in-system interfaces, two key parameters for adopting optical technology are cost and energy efficiency.



	Connectivity	Function
1	A-A remote	High BW Communication
2	A-A local	High BW Communication
3	A-C (PCIe)	CPU / control plane, DRAM, Datacenter network
4	A-M (HBM)	Model data

Figure 6 Network map for Scaling-up nodes using optical interconnections. List of the interface connectivity type. [1]

Legend: A: Accelerator, C: Computing/CPU, M: Memory

Table 3 Main characteristics: Compute Interconnect

1st priority	Cost N x 10 Tbps of aggregate bandwidth per accelerator
2nd priority	Latency / Throughput
3rd priority	High-density implementations
4th priority	Energy Efficiency
Timeframe	Optical interfaces to appear around 2025-2027

2.4 Optical Computing Accelerator

Optical computing accelerators represent specialized instruments that harness the power of light to perform complex computations. While these systems are currently under extensive research, they hold the promise of becoming a future solution to specific computation challenges. While not a primary focus of the industry at present, and therefore not part of the focus for this requirement document, optical computing accelerators deserve recognition for their potential to achieve unparalleled energy efficiency, leveraging the inherent properties of light to naturally derive solutions to computational problems.

Diverse Forms of Optical Computing Accelerators

Optical computing accelerators encompass a diverse range of technologies, including Quantum Computers, Photonic Neural Networks, Programmable Photonic Processors, Optical FFT, and more. These accelerators elevate traditional electronic computing systems by capitalizing on the speed and bandwidth of optical signals, facilitating rapid data transmission and parallel processing of information.

Their true value shines in tasks involving vast datasets, such as machine learning, artificial intelligence, and complex simulations, where they have the capacity to significantly enhance overall system performance and energy efficiency.

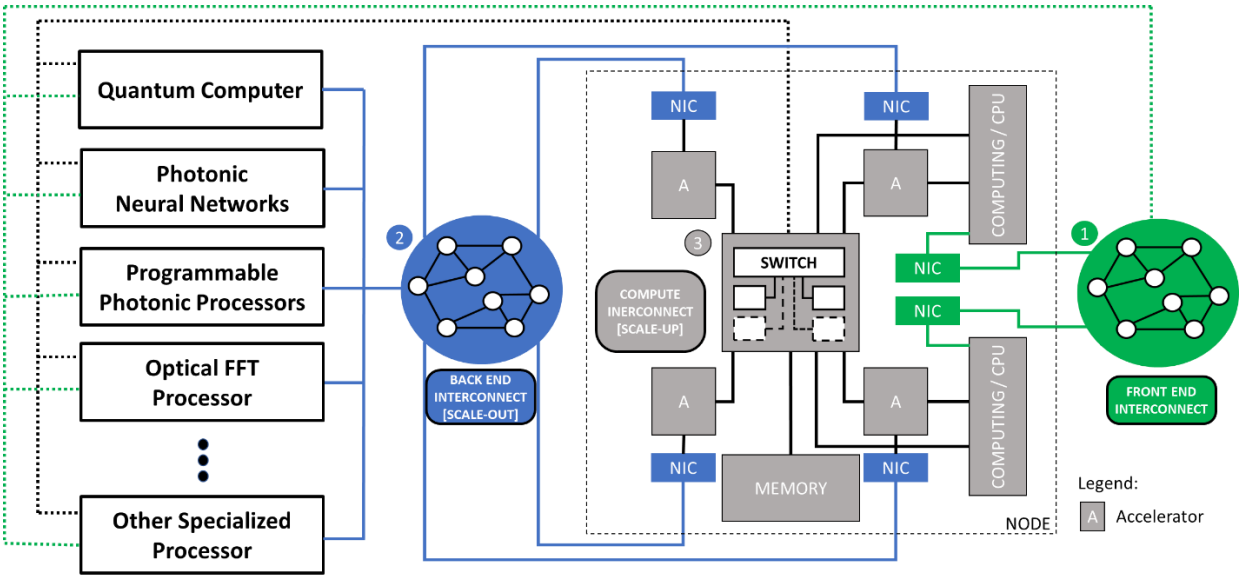


Figure 7 System integration of various optical accelerator functions.

Futureproofing for Accelerator Compatibility

Crucially, any new interface designed must be future-ready, accommodating all forthcoming accelerator technologies. This forward-thinking approach ensures that evolving computational paradigms can seamlessly integrate with existing infrastructure, fostering innovation and adaptability in the ever-evolving landscape of high-performance computing.

Table 4 Main characteristics: Optical Computing Accelerator

1st priority	Interfaces tolerant to High-Loss
2nd priority	Latency / Throughput
Timeframe	Beyond 2030

3 Key requirements for EEI applications

This section delves into the criteria for developing Energy Efficient Interfaces (EEI). Over an eight-month period, the Optical Internetworking Forum (OIF) conducted surveys with multiple users, gathering data to delineate interface specifications for various applications. A complete table of these findings can be found in Appendix D of this document and the detailed spreadsheet in contribution OIF2023.231.12.

When considering the advancement of next-generation optical interconnections, five primary characteristics emerge:

1. **Power Consumption:** This aspect evaluates the energy usage of the interfaces, emphasizing efficiency.
2. **Latency:** A measure of the time delay in data transmission, vital for performance in high-end applications.
3. **Requirement for Interoperability:** Ensuring compatibility between devices from various components vendors.
4. **Form Factor and Density:** Concerning the physical size and connection capacity within a given space.
5. **Transmission Distance:** The range over which data can be transmitted effectively, at which data-rate/traffic type.

For each application, it is vital to conduct a relative comparison, not just to benchmark current capabilities but to identify potential trade-offs and enhancements. By understanding the interplay between these characteristics, one can ascertain the improvements achievable by possibly relaxing certain criteria. This approach aids in striking an optimal balance between performance, efficiency, and practicality within each type of optical interconnect technology.

3.1 Power Consumption

Efficiency in power consumption is a critical aspect that underpins the effectiveness and sustainability of the applications and technologies discussed in this document. This section emphasizes the key considerations related to power efficiency.

In serving the industry, it is imperative that implementations minimize power consumption. The less power an implementation uses, the better it serves the broader technological landscape. Lower energy needs translate directly to higher density of applications, and this is what is needed to allow a system to utilize optics deeply in the system and exploit the system advantages of optical interconnect.

To achieve alignment with industry standards and expectations, discussions surrounding power targets were facilitated. Establishing a consensus on targeted power consumption is crucial to guide the development of energy-efficient interfaces effectively.

Within the context of FEI, specific considerations come into play. There is a growing need for longer-distance capabilities and a broader range of interoperability requirements. It is essential to maintain compatibility with already deployed modules, ensuring seamless integration and continuity.

For FEI, the widely accepted energy consumption target stands at 10 picojoules per bit (10pJ/b). This serves as a foundational benchmark for power efficiency within this application.

Acknowledging the varying transmission reach requirements within FEI, one recognizes the existence of various solutions, each tailored to specific distances. However, the median requirement for power consumption remains at 10 pJ/bit.

Beyond FEI interconnect, other critical applications need to be explored in priority, including BEI in Large GPU Clusters, Compute Interconnect, and Optical Computing Accelerators.

These applications come with their unique requirements. They often involve shorter distances, higher densities, and a heightened need for power efficiency to maximize performance. Operational distances for these applications typically fall within the range of 10 to 100 meters.

To meet the distinctive demands of these applications, the targeted power consumption is set at 5pJ/b with some operators calling for a 3pJ/b as a target. This underscores the paramount importance of power efficiency in these scenarios.

As the specifications and requirements for each application is studied, it is essential to emphasize the commitment to prioritizing power efficiency. This commitment ensures that data center environments can meet evolving needs effectively.

$$Front\ End\ Interconnect:\ 10pJ/b$$

$$Other\ applications:\ 5pJ/b$$

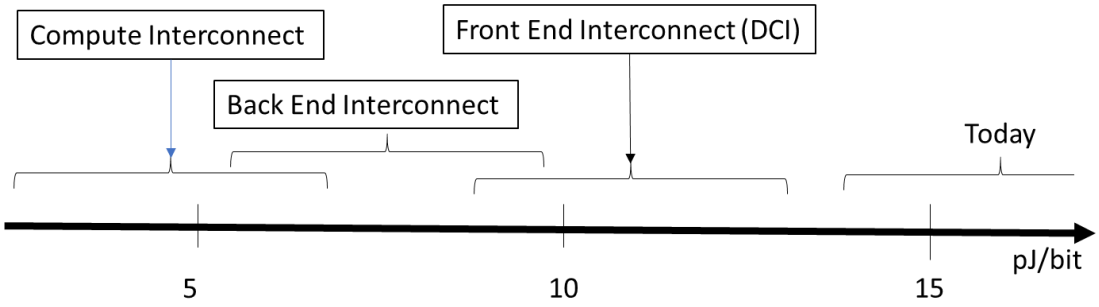


Figure 8 Relative requirement for energy efficiency depending on the applications.

3.2 Latency in Optical Interconnects.

Latency, a critical aspect of optical interconnect technology, is defined as the time delay from when a signal is sent from a transmitter to when a receiver receives it . This delay has a significant negative impact on system performance; for instance, Reference [6] demonstrated that a 1-microsecond increase in latency within an AI cluster can result in a 25% decrease in performance throughput.

Addressing latency requires innovative approaches at the system level. The goal is to minimize both peak and average latency, which involves optimizing the time taken to move data from a usable state in one subsystem to another. Strategies such as reducing the number of re-transmissions can significantly lower peak latency.

This document focuses specifically on Physical Medium Dependent (PMD) latency, excluding other factors that contribute to overall latency, e.g., MAC latency and flow control. The importance of PMD latency in the larger picture needs to be understood to determine the value of any potential latency improvements in the PMD.

Regarding PMD latency targets, the survey has identified specific requirements for three applications:

Front End Interconnect: The current latency levels are generally adequate and act as a benchmark for future improvements. In these applications, latency is consistent with End-to-End Reed Solomon Forward Error Correction (FEC) (RS(544,514)) as defined by the IEEE, as detailed in Equation [1]. Here, 'd' represents the length of fiber traversed in meters.

$$PMD\ Latency\ <\ 20ns\ +\ d * 5ns/m\ +\ (FEC\ delay)$$

Equation 1

Back End Interconnect: It would be desirable to reduce latency by approximately 15 nanoseconds from the current baseline through architectural innovations. These interfaces are expected to remain similar to standard Ethernet devices to leverage existing supply chains and expertise. Notably, the peak latency in this application is naturally lower on average than in Datacenter Interconnect applications, considering the shorter expected interconnection lengths of no more than 300 meters.

$$PMD\ Latency\ <\ 5ns\ +\ d * 5ns/m\ +\ (FEC\ delay)$$

Equation 2

The 15ns baseline difference between Front End and Back End Interconnects reflects the adaptive nature of latency in back-end interconnects. In scenarios involving very short links with minimal impairments, the latency is designed to be close to that of Computer Interconnects. Conversely, for longer links, the latency adjusts to be more akin to Front End Interconnects. This adaptability allows for optimized performance across varying link lengths and conditions, ensuring efficiency in different network configurations. Gains in latency reduction may be extracted from shorter equalization chains in the link or in FEC processing gains although, as the equation suggests, the industry is currently of the opinion that FEC will likely remain Reed Solomon for BEI.

Compute Interconnect: in Compute Interconnects, there is an expectation to uphold FEC that is reliant on the PCIe FLIT (Flow Control Unit) architecture. The use of FLIT-based encoding in PCIe is particularly important in maintaining data fidelity over extended distances or in high-bandwidth scenarios.

The primary focus in this area is on leveraging optical technology to minimize the need for Link Layer Retry. By reducing the frequency of these retries, which are necessary when data packets are lost or corrupted during transmission, there is a consequent lowering of overall system latency. This improvement in latency is vital for enhancing the performance and efficiency of Compute Interconnects.

Through these targeted efforts, the goal is to advance optical interconnect technology. The emphasis is to be optimizing FEC through PCIe FLIT architecture and reducing Link Layer Retry is a strategic approach to making systems faster and more efficient. This is especially important in catering to the demanding requirements, where high performance, low latency, and reliable data transfer.

3.3 Level of Interoperability

In optical interconnect technology, the requirement for "plug and play" interoperability varies across applications. This aspect forms a key area of trade-offs, where designers can make decisions to achieve gains in energy efficiency and latency reduction. Interoperability can be segmented into three distinct types:

Type-1 (Vendor-Specific Interworking): This involves devices working together within systems and equipment from a single vendor, known as bookended links.

Type-2 (Cross-Vendor Similar Devices): This refers to the interworking between similar devices from various vendors within the same generation.

Type-3 (Cross-Vendor Compatibility, Cross-Generational): This encompasses interworking between similar devices from various vendors and includes backward compatibility with devices of various generations.

Application-Specific Interoperability Requirements

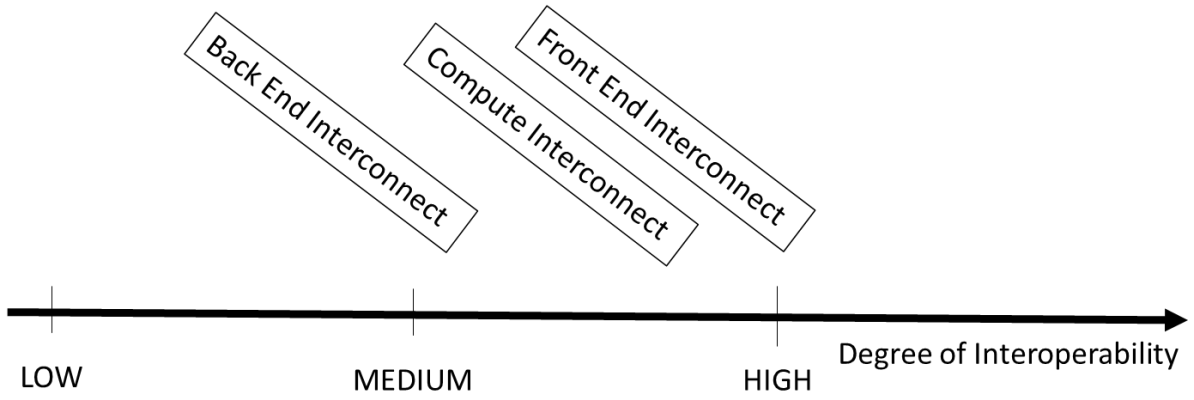
Front End Interconnects are expected to maintain the highest level of interoperability; these systems allow for the use of devices from various vendors and technologies. PMDs are specified in a way that permits multiple implementation options, facilitating the required level of interoperability. As energy efficiency improves, maintaining this high interoperability degree is crucial, especially as Front End Interconnect systems evolve over time.

Back End Interconnect, interoperability is desired but not mandatory. Typically originating from a single system house or hyperscale operator, these interconnects are expected to be deployed uniformly over a short period. The nature of these deployments, often controlled by a single entity, allows for reduced interoperability requirements even within the same generation. The uniformity in the type and number of accelerators used further supports this.

Similar to the Front End Interconnect, the Compute Interconnect application has additional degrees of freedom but a slightly higher interoperability requirement. Given the expected involvement of multiple suppliers providing sub-systems for scaled-up nodes, a diverse array of components like accelerators, memory, and CPUs may be involved. Drawing from the PCIe deployment model's success in enabling plug-and-play operations without tight inter-vendor collaborations, optical interconnects are envisioned to operate similarly, necessitating a degree of interoperability.

Cross-Application Interoperability

While interoperability across different applications is not a primary focus, there is an expectation that network intermediation will occur through system appliances. The drive for interoperability in this context is motivated by the desire to share technological advancements across applications, ultimately aiming to reduce costs and streamline the technology ecosystem.



3.4 Form Factor and Density requirement

In high-performance computing (HPC), the density and form factor of optical interconnections play a pivotal role in system performance and throughput. The efficiency of these interconnections is a critical factor influencing the final throughput of HPC systems, as noted in reference [11]. However, quantifying this impact in terms of a simple, comprehensible figure of merit remains challenging.

Density, in this context, refers to the data-rate – the amount of data per second that can be transmitted in and out of a system. A practical approach to measuring density is by assessing the data throughput across the system, node, or server enclosure, often concentrated at the faceplate. This can be expressed as terabits per second per square millimeter (Tb/s/mm²). For systems built on 2D printed circuit boards, density can also be evaluated by the data transmission and reception capacity per linear millimeter of the transmitter-receiver width (Tb/s/mm)[1].

System designers often prefer Faceplate Pluggable (FPP) form factors like OSFP and QSFP-DD for their numerous advantages:

1. **Modularity:** FPP form factors allow for easy interchangeability of components from any vendor, as long as they meet the specifications.
2. **Reliability:** These form factors enhance system reliability due to their ease of replacement, facilitating in-service interface swaps when failures occur.
3. **Supply Chain Efficiency:** Being incumbent interfaces, the supply chain is well-equipped for low-cost fabrication and testing.

However, new form factors like Near Package Optics (NPO) and Co-Packaged Optics (CPO) are emerging, offering breakthroughs in signal quality and energy efficiency.

The aim is to scale up systems and increase compute capacity per system. Maintaining a stable energy density is essential. This necessitates improving the energy efficiency of interfaces. CPO, in particular, shows promise in reducing power requirements by minimizing power usage in the chip-to-module interface.

The drive to enhance fan-in/out from ASICs, CPUs, TPUs, etc., to over 200G/lane demands improvements in signal quality and power efficiency. Implementing the chip-to-module interface over shorter distances directly translates into these enhancements. CPO can facilitate low-power, high-quality chip-to-module interfaces by reducing the length of this interface and the number of connectors a signal traverse. Implementing “linear” interconnect technologies may be more feasible in a CPO context at high data rates.

Tradeoffs between pluggable interfaces and the energy improvement need further evaluation.

[1] The linear density is defined as transmitter bandwidth plus the receiver bandwidth together over the edge length.

Evolutionary Path of Server and Switch Interconnects

The progression of server to switch interconnects is expected to evolve from electrical to pluggable optical transceivers, then to a combination of pluggable transceivers on the server and CPO to the switch, and finally, CPO on both servers and switches. Similarly, switch to switch interconnects, currently pluggable optical on both sides, are anticipated to transition to CPO over time. This roadmap may include an intermediate step transitioning to Linear technologies.

In conclusion, as industry forges ahead in the development of high-performance optical interconnections, a balanced consideration of form factor, density, and energy efficiency will be crucial. These elements will collectively dictate the future trajectory of interface standardization and innovation in the field of optical interconnect technology.

3.5 Transmission Distance

Transmission distance plays a critical role in the design and application of optical interconnects. The required transmission distance varies based on the specific use case and application environment. This requirement is particularly relevant in the context of high-performance computing, AI clusters, and data center networks, where distance can significantly impact the overall system performance and connectivity.

Front End Interconnect: Front End Interconnects have a more varied range, spanning from 0 to 500 meters and even extending up to 2,000 meters in some cases. These distances correspond to the expected size of data center networks. The 0-500 meter range caters to most typical data center layouts, while the extended 0-2,000 meter range accommodates larger data center environments. These distances are defined in line with IEEE standards, ensuring compatibility and efficiency in broader data center network architectures. The flexibility in distance requirements for Front End Interconnects reflects the diverse scale and design of modern data centers, addressing the need for both localized and extended network connectivity.

Back End Interconnect: The BEI requirement extends up to 300 meters. This distance aligns with the viable size of an AI cluster machine, where the compactness of the cluster is key to delivering expected performance. A 300-meter range allows for effective interconnection within a large AI cluster, ensuring that data can be transmitted efficiently across various nodes and components of the system. This intermediate distance range is fundamental in maintaining the balance between physical size and performance in AI clusters. While the IEEE requirements baseline for FEI would make them suitable BEI implementations, it is important to note that the users are expecting the industry to leverage the reduced distance requirement to produce gains in power and latency performance in the BEI applications.

Compute Interconnect: For Compute Interconnects, the transmission distance typically ranges between 7 to 10 meters. This range is ideal for interconnecting systems within a single rack or between adjacent racks in a data center. This short-range connectivity is crucial in environments where high-speed data transfer between closely situated systems is essential. The 7-10 meters range ensures efficient communication within these confined spaces, facilitating rapid data exchange necessary for compute-intensive tasks.

Implementation of Photonics switch within the Interconnect Network

The implementation of photonic switches in a photonic network is an emerging innovation not originally outlined in the requirements gathered through surveys, but it holds potential for mid to long-term advancements. Commercial deployment of photonic switches with millisecond switching speeds has been documented in references [12, 13]. These switches, varying in type and architectures, have been discussed in past publications primarily based on laboratory experiments.

The primary impact of integrating photonic switches into optical networks is on the loss budget of the optical link. This increased loss is often equated to an additional length of optical fiber. For energy-efficient interfaces (EEI), this implies a higher demand for optical power at the interface, leading to increased power consumption. Additionally, the added loss might necessitate receivers to handle lower signal-to-noise ratios, thereby requiring more forward error correction (FEC) gain to maintain signal integrity.

Future contributions and studies are essential to clearly establish the trade-offs and impacts of incorporating photonic switches within systems. As technology evolves, understanding these nuances will be crucial for optimizing optical networks in terms of efficiency, performance, and power consumption.

4 Prioritization for Future Work in Standardization of Interfaces

As optical interconnect technology continues to evolve, especially in high-performance computing areas like AI/ML clusters, it's imperative for the Optical Internetworking Forum (OIF) to strategically guide the standardization of interfaces. This section outlines prioritized recommendations for future work, emphasizing two critical areas: Back End Interconnect (Accelerator-to-Accelerator Interconnection Remote) and Compute Interconnect (Accelerator-to-Accelerator Interconnection Local). The objective is to align the OIF's standardization efforts with the latest industry trends and technological developments.

4.1 Recommendations for Back End Interconnect

Short-Term Focus: Immediate attention should be given to the Back End Interconnect, a domain where optical interconnections are already and rapidly being adopted in AI/ML clusters. This area already benefits from the Ethernet component supply chain and utilizes bookended solutions. The OIF is advised to focus on developing Implementation Agreements (IAs) that encourage common solutions, enhancing interoperability and reducing costs.

Key Areas for Technology Standardization:

1. **Energy-Reducing Technologies:** Development of technologies aimed at decreasing energy consumption in optical interconnections should be a priority.
2. **Latency-Reducing Technologies:** Innovations focused on reducing latency are critical for maintaining system performance and minimizing the impact of interconnect delays.
3. **Maximizing Link Transmission Quality:** Efforts should be made to lower the raw bit error rate on links, which would enable weaker FEC and lower final BER, thus increasing link throughput. Transitioning from a margin-based link budget to actively optimized link transmission will ensure optimal BER and reduce overall energy consumption.
4. **Interface Densification:** To meet the growing demand for optical interconnections in AI/ML clusters, it's essential to increase the density of transmission per module, thereby enhancing throughput and economic efficiency. The initial focus should be on increasing module throughput (measured in $Tb/s/mm^2$) before moving to on-package optical interconnects.

4.2 Recommendations for Compute Interconnect

Medium-Term Focus: The Compute Interconnect, particularly for local accelerator-to-accelerator applications, will need considerable advancements. This technology is also relevant to Accelerator-to-CPU and Accelerator-to-Memory applications, with many technologies from the BEI expected to be adapted for very short reach, cost-sensitive interconnections.

Key Areas for Technology Standardization:

1. **Cost Reduction Technologies:** IAs should aim for implementations that minimize costs, focusing on module architectures that reduce packaging and testing expenses. Interface densification can further reduce costs by consolidating multiple links into a single module, provided yields are maintained.

2. **Energy Efficiency:** There should be a continued emphasis on energy-reducing technologies within optical interconnection.

3. **Latency Reduction:** As the trend moves towards scaling up single nodes — both in the number of cores per accelerator and the number of accelerators per node — minimizing latency becomes increasingly crucial. This is particularly important in rack-scale nodes, where optical interconnects serve as the bus linking components like accelerators, memory, and CPUs.

In conclusion, these recommendations for Back End and Compute Interconnects are formulated to direct the OIF's standardization efforts in the Physical and Link Layer (PLL) working group in a trajectory that aligns with both current and future industry requirements. By focusing on energy efficiency, latency reduction, cost-effectiveness, and interface densification, the OIF can ensure its standards remain relevant and beneficial in the swiftly evolving world of optical interconnect technology.

5 Implementation of a common language for the industry

The increasing application of optical interconnections, traditionally used in data communication within data centers and telecommunication equipment, has fostered the development of a common industry language. This standard nomenclature has enhanced clarity and efficiency in communication across the sector.

Recent trends, as highlighted in this document, show optical interconnects expanding into different system areas and adapting to new architectural designs. However, this evolution, largely driven by a handful of suppliers, has introduced a variety of terminologies for similar concepts, leading to confusion regarding applications and their requirements.

Furthermore, as optical interconnections penetrate deeper into systems, aiding in scaling up capabilities, a discrepancy in terminology becomes apparent. Standards and computer buses, which have long relied on electrical interconnects, have developed their own set of terms. These sometimes overlap with, but often differ from, the terminology used in the optical interconnect community, either representing different concepts or the same concepts with varied meanings.

This scenario underscores the necessity of establishing a unified industry nomenclature. A common language should be developed to facilitate clearer expression of concepts and enable the industry to reach consensus more efficiently.

6 Summary

In conclusion, this requirement document presents a comprehensive overview and strategic recommendations for advancing optical interconnect technology in high-performance computing environments. It underscores the importance of energy efficiency, latency reduction, and interface densification in meeting the growing demands of AI/ML clusters. The document also highlights the need for standardization in technology to achieve interoperability and cost-effectiveness. These insights and guidelines set a clear path for the development and standardization of energy-efficient optical interconnections, ensuring they remain relevant and beneficial in the face of rapid technological advancements.

References

Normative references

This section is intentionally left blank

Informative references

[1] R. Huggahalli, “Microsoft Use Cases, Optical Connectivity for AI Clusters”, Presentation - OCP October 19-17, 2023

[2] R. Huggahalli, “Energy Efficient Interface Requirements – Microsoft,” oif2023.420.01

[3] D. Alduino, et al., “Energy Efficient Interfaces - Meta Perspective,” oif2023.273.01

[4] R. Huggahalli, “Electrical Interface Considerations – Microsoft, “ oif2023.269.02

[5] S. P. Sundararajan, “Use case for energy efficient interfaces at HPE,” oif2023.268.02

[6] C. Thompson, “NVIDIA Motivation for Energy Efficient Interfaces,” oif2023.270.00

[7] D. Piehler, “Linear pluggable optical modules – more on key decision points,” oif2023.151.01

[8] E. Bernier, “Linear Optical Interface: A system Vendor’s Perspective,” oif2023.126.00

[9] J. Hutchins, “Energy Efficient Optical Links for PCIe AI/ML Scale-Out”, Presentation - OCP October 19-17, 2023

[10] Contributed table. More info within MEGA DATACENTER OPTICS REPORT JULY 2023, Lightcounting

[11] Q. Cheng, et al., “Recent advances in optical technologies for data centers: a review”, Optica, November 2018, Vol. 5, No. 11

[12] L. Poutievski, et al., “Jupiter Evolving: Transforming Google’s Datacenter Network via Optical Circuit Switches and Software-Defined Networking”, SIGCOMM’22, August 22-26, 2022

[13] N. P. Jouppi, et al., “TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings”, ISCA ‘23, June 17–21, 2023

Appendix A: Glossary

Accelerator: Accelerator is a hardware device designed to perform specific types of computations more efficiently than a general-purpose CPU. Commonly used in high-performance computing and AI, accelerators such as GPUs and TPUs optimize tasks like machine learning, enhancing overall system performance.

Computing/CPU: Computing refers to the process of utilizing computer systems for managing, processing, and manipulating data, especially in high-performance environments like data centers and AI/ML clusters. The CPU (Central Processing Unit) is the primary hardware component in these computer systems, responsible for executing instructions and processing data.

GPU: A GPU, or Graphics Processing Unit, is a specialized processor designed to handle parallel computation efficiently, making it crucial in AI, particularly in deep learning and neural networks. Its capacity for simultaneous multiple calculations optimizes the data-intensive tasks of training and running AI models, resulting in faster and more effective processing.

Node: Any device within a network, FEI, BEI, or CI that can send, receive, or forward information.

PCI: The PCI bus, or Peripheral Component Interconnect bus, is a standardized bus for connecting peripheral devices to a computer's motherboard, overseen by the PCI-SIG (PCI Special Interest Group).

Scale-out: Involves expanding capacity by adding more units to a system, such as adding more servers to a network to distribute workloads across multiple machines, enhancing performance without altering the capacity of individual units.

Scale-up: Refers to the process of increasing the capacity of a single unit, such as adding more resources like CPUs, Cores, or memory to a single server/computer unit, to enhance its performance.

Server-to-edge: Involves network interactions between central servers and terminal devices, such as user clients or applications requiring enhanced processing capabilities.

Server-to-server: Refers to data exchanges and communications between servers within a network, facilitating tasks like synchronization and computing.

Server-to-storage: Describes the connections between servers and storage systems, essential for data access, management, and backup operations. In AI cluster this may be used as a common repository of the AI model.

TPU: A TPU, or Tensor Processing Unit, is a type of processor developed by Google specifically designed for neural network machine learning. It is optimized for the high-volume, low-precision computation required in deep learning applications, offering significant performance improvements over traditional CPUs and GPUs in certain tasks.

Appendix B: Open Issues / current work items

Appendix C: Summary Table per Applications

	Data Center Network / Front end network to AI clusters	High-Performance Computing and AI GPU interconnect	Large Compute Node/ Compute Disagregation
Requirement	**Front End Interconnect**	**Back End Interconnect**	**Compute Interconnect**
Description of use case	This use case describes the optical interconnections that are used to ease servers communications. (Server to servers, servers to storage, servers to edge, etc.) This is about how servers communicate with the outside world. They may be general purpose datacenters servers or specialized servers for AI, analytics, HPC systems etc.	Specialized interconnect for AI intra-cluster Accelerator communication, HPC intra-cluster communication, and others. This is to scale-out the computing infrastructure. Going to 1,000-10,000 nodes. This use case describes the optical interconnections that are used to ease Accelerator to Accelerator communication to execute the AI models or train the models. (Accelerator refers to GPU, TPU, Analog Computers, etc.)	Pooled memory allows multiple nodes to access the same memory reducing the system memory consumption. (Extending to pooled memory and pooled IO)
End user deployment	Next Generation Ethernet Infrastructure. (Front end network)	AI/ML and HPC (Back end network)	Within nodes. Comp-IO, Comp-Mem, Comp-Comp. Note : (The Disaggregated node may be much larger than nodes today.)
Timeframe	NOW (Already in pervasive use)	Emerging Now (2023)	2025-2027
Form factor(s)	Traditional Ethernet traffic (Server to server, server to storage and server to edge communication)	Pluggable, NPO/CPO	Co-Packaging / AOC
Traffic Type	Traditional Ethernet traffic (Server to server, server to	ND	PCIE Express Gen 5/6 and CXL 3.0 or later
Latency	Consistent with End-to-End KP4 FEC or <20ns + d*5ns + (FEC delay)	<5ns + d*5ns + (FEC delay)	Target < 1 us or less write, target < 2 us or less read
End-to-end FEC	KP4, optionally, KP4-IEEE defined concatenated FEC	ND	PCIe + FEC
Data rate	200G / 100G per lane 50Gper lane for backward compatibility 800GbE (Main use case), 1.6TbE for future 400GbE for backward compatibility	400GbE, 800GbE, and future 1.6TbE 200G / 100G per lane 50Gper lane for backward compatibility 800GbE (Main use case), 1.6TbE for future 400GbE for backward compatibility	PCIE Express Gen 5/6
Expectation of capability for the host	Unclear	200G / Lane linear channel for CPO (TBD)	ND
Expectation for compliance (@TP1x, @TP4)	IEEE Standards: SRx, DRx, FRx, etc.	Compliance to TPx to be defined in electrical link	ND
Expectation for compliance (@TP2, @TP3)	IEEE standard 800GE-FR4, 400G-FR4 and 200G-FR4	Compliance with TPx to be defined in opitcal link	ND
Target IL for link (for optical links)	Unclear	2.5dB, 500m	ND
Degree of interoperability	Full interoperability with all module types and technologies	Fully interoperable with pluggable modules desirable, but depends upon power/spec trade-off.	ND
Universal host requirement	ND	ND	PCIe, UCIe
Energy efficiency for link	<10pj/bit for OE+Laser	<<4pj/bit for OE+Laser	ND
Energy efficiency for link + host	ND	ND	ND
Power saving target (modules only)	ND	ND	ND
Power saving target (end-to-end)	ND	ND	Target < 75% current implementation
Density (Beachfront, Areal)	>1Tbps/mm (End goal in a complete CPO evolution)	>1Tbps/mm	
Link accountability (out-of-service)	System and module parameter measurements will result in a determination of the location of the fault.	System and module parameter measurements will result in a determination of the location of the fault.	ND
Link accountability (in-service)	Easy to identify the link element with issue	Easy to identify the link element with issue	SW assumes reliable link to memory for CPU/GPU Loads/Stores
Link margin requirements and allocation	Sufficient link margin to match pluggable interoperability desirable	Sufficient link margin to match pluggable interoperability desirable, but	ND
Compliance verification	PAM-4, NRZ	Ability to determine compliance of each link element	ND
Modulation format	ND	PAM-4, NRZ	ND
Designated Management Interface	Definition to be developed by CMIS	Definition to be developed by CMIS	ND