OPEN
Compute Project

Short Reach, Photonic Interconnect for Compute: A Market and Technology Landscape

Editor:

Dr. Michael Bortz

Release 01 April 2024

Executive Summary

As is well- known, photonic interconnects play a critical role in the scale-out of data center networks. Typical photonic interconnect distances span 10 meters up to 2,000 meters and bandwidths of 400 Gbps are in full deployment and 800 Gbps is approaching commercial applicability. But what is less well- known and less well understood is the role photonic interconnects will play in the next few years at much shorter length scales, from 10 cm to 10 meters. Photonics will be crucial to facilitate a new range of applications inside computers and to enable increased xPU-xPU scale-up capability as well as vastly improved xPU-memory access. This Open Compute Project (OCP) White Paper is aimed at elucidating the physical layer limitations, use-cases, and technologies that can motivate and enable short distance, “short reach'' photonic interconnects and is based on an in-person Future Technologies Initiative (FTI) workshop conducted at the OCP 2023 Global Summit.

Editor’s Note: This document is assembled as a white paper contribution to the Open Compute Project (OCP) in the style of a conference proceedings document with participants from the OCP 2023 Global Summit FTI summit on photonics contributing individual papers towards the topic. See the references at the end of this document for slides and video presentations from the conference.

Table of Contents

1. Introduction

There are 3 well-known and interrelated physical layer limitations that are driving the use of photonics for directly interfacing to ASICs for xPU-XPU interconnects as well as xPU-Memory interconnects.

1) SERDES and copper transmission limitations

As bit rates increase, the distance tradeoff for photonics to replace copper decreases to shorter distance scales. The tradeoff is based on capital cost, power, and other practical factors. Electrical retiming, while always an option, is a stopgap measure at best. By late 2023, high speed 200G/lane SERDES can enable 10’s of cm on a PCB or twin-ax flyover cable and ~ 1-2 meters on a DAC cable with powers of 4-6 pJ/bit. As per-pin bandwidths move above 200Gbps, there is no way to avoid photonics for lengths scales exceeding 10’s of centimeters. The copper-optical tradeoff is unrelenting and pushes photonics to shorter and shorter lengths scales over time, as has been the case over the past 40 years.

2) xPU-Memory Bandwidth

Many, important compute workloads are memory intensive, and workload performance is nearly always gated by memory access. The tiered memory hierarchy hides much of this limitation, but the insatiable demand for memory access is not going away, and is now being driven by new AI workloads, huge LLM models, and synchronous training strategies. As is well-known in computer science circles, DRAM access is limited by the xPU-memory interconnect. The compute FLOPs to memory access (GB/sec) ratio for traditional DRAM access has been trending in the wrong direction for decades, from the ideal of ~ 1 to ~100, and this dramatically affects compute performance. In some applications, this “Memory Wall” is partially overcome by the use of HBM, with its in-package, very high bandwidth interface, but it remains to be seen if the costly HBM-enabled SOC’s are practical for broad, widespread compute applications. Photonics can enable much higher memory access by using serialized memory and low latency protocols like CXL/PCIe to enable high speed access to separately packaged commodity memory. Photonics offers the ability to also disaggregate memory and pool memory for vastly improved utilization, with the memory not being tied to the specific processor.

3) Packaged ASIC I/O Escape Bandwidth

BGA packages have fundamental size and ball (pin) pitch limitations and the largest packages have ~ 8000 or so pins. In a large CPU, about 40% of these pins are used for DRAM memory access (and about 40% are used for power and ground). HBM solves the escape IO problem for memory by bringing memory in package with an interposer connection, but this is a very costly solution. I/O bandwidth obviously increases as the per pin bandwidth increases, and this is the main driver for serialized memory access at 64 Gbps/pin and beyond (PCIe 7.0 will be on par with Ethernet at 128 gbps/lane). However, photonic interfaces directly to xPU’s promise to increase the bandwidth by > 10X over foreseeable electrical solutions, due to the vastly higher shoreline and areal density made possible with various photonic architectures. Ethernet switch chips, with little or no external memory access, have 50 Tbps I/O capability at 100 Gbps/lane, which corresponds to 2000 balls for I/O (differential transmission in each direction means 4 balls per lane). Expanding this to 200Gbps/lane will be challenging, but feasible, as discussed above. Compute xPU's with vastly more pins allocated to memory and other functions will be hard pressed to allocate more than a few 10's of Tbps to I/O, and many firms feel that once xPUs require ~ 10Tbps of I/O bandwidth optical interfaces will be required. It is important to note that these optical interfaces can be allocated to both memory and scale-up /scale-out capability.

Underpinning the three fundamental issues discussed above is power - the power required for moving data in and out of the chip, either to memory or to another chip. Interconnect power is becoming a larger and larger fraction of the total power consumption of the system. Typical long-reach electrical SERDES consume 4-6 pJ/bit (at each end), so a 1 Tbps Ethernet or serialized memory interface might consume ~40-60W of the ~ 400-600W chip TDP. While water cooling promises to make this more palatable from a thermal management perspective, the raw electrical power consumption still remains a problem for DC operators. Photonics has historically required higher power than LR SERDES, but new photonic technologies and approaches have opened the door to sub 1 pJ/bit transmission capabilities, which can overcome most of the power problems associated with data transmission over these relatively short distances.

The FTI Workshop we conducted at the OCP Global Summit consisted of six talks on use cases and system related applications and then seven talks on photonic technologies. Nearly all participants that presented have also provided a section in this White Paper, and this White Paper follows the agenda of the workshop session. Both the Workshop and the White Paper are “use case and technology agnostic” meaning that this effort is not meant nor intended to promote a specific use case nor a specific technology for implementation, other than the general use of photonics. There are other bodies that are promoting implementation standards, but our intent here is to canvas the use cases and technologies to see what is possible in the 3-5 year time frame.

In the first session of the workshop, Dell discussed what they feel will be the first use case for broad adoption of photonics inside servers, which would be the photonic enabled NIC cards. Meta discussed the use of photonic-enabled GPUs to enhance AI-LLM workloads, focusing in part on photonics alleviating the I/O escape bandwidth problem discussed above. LBL-Stanford then discussed the use of photonics for compute resource disaggregation, where the photonic interconnect enables resources such as CPUs or GPUs and different types of memory, to be disaggregated, then combined selectively and temporarily based on workload requirements. Rockport discussed the novel use of photonics for interconnecting processors in a perfect shuffle scale up topology. Lightelligence then discussed CXL/PCIe/photonics to enable memory resource pooling, which was effectively a first implementation of the LBL-Stanford vision. Finally, to end the first session, Light Counting gave an overview of the market potential for photonic interconnects inside computers.

In the second session the technology providers offered their vision and described how their technologies could address the use cases outlined in the first section. Ranovus, Ayar Labs, and Nubis described their silicon photonics platforms and discussed the first use cases that they felt would be instrumental towards developing the market for the short reach photonic interconnects. Quintessent gave a presentation on their comb laser transceiver platform. Coherent (Finisar) discussed the use of VCSEL's for these applications, leveraging a long tradition in the HPC sector as well as discussing the manufacturability of their approach. Avicena discussed their novel, highly parallel LED based solution, which is tailored to the very short distances and per lane bitrates one sees in proposed memory interfaces. Finally Marvell gave an overview of their linear drive platform that promises to reduce power by eliminating retimers in high speed interfaces.

This Workshop was the first step OCP has taken to develop a community around the use of photonics for short reach in-server interconnects. The next step in 2024 is to launch a workstream within the Future Technologies Initiative of OCP. This will enable the Community to evolve and grow, mature the technologies, and evolve this workstream to where it can be coordinated with other OCP Projects and industrial activity. By late 2024 there should be sufficient clarity to enable a formal OCP Project workstream to be kicked off that will enable tangible contributions to be made to OCP. This white paper memorializes the workshop, which was the first step in this journey, to provide future participants with a clear understanding of what came before them.

Michael Bortz

December 2023

2. Compliance with Open Compute Project Tenets

This whitepaper describes the landscape and examines some implementation details for Short Reach Photonics Interconnects. As it is explanatory and not necessarily implementation specification, compliance with The OCP Tenets is aspirational and/or related to the inherent properties of the underlying technologies described.

2.1. Openness

This document is published to be a cornerstone reference for future development of open specifications. It is published with a permissive license that allows for all to reuse as needed.

2.2. Efficiency

Continuous improvement has been a fundamental value of the industry. Many of the technologies and approaches within this document target and/or contemplate improving efficiency and are often related in the sections in terms of reducing power consumption/bit.

2.3. Impact

The authors have proposed ideas and technologies whose impact can be measured by the creation of future workstreams and/or resulting new technologies, specifications, and potential time-to-market advantages based upon the ideas within this white paper.

2.4. Scale

If this paper is successful as a catalyst in the creation of new technologies and workstreams, it's likely that Short Reach Photonic Interconnects will have the scale of inclusion as essential components/subsystems within IT Equipment that is typically found in the datacenter.

2.5. Sustainability

The entirety of Short Reach Photonic Interconnects look to enable massive amounts of bandwidth available at a scale that is, in its advanced forms, likely to be less energy intensive than copper interconnects.

3. Opportunities for Computer Interconnects: a Meta Platforms Perspective

OPEN Compute Project

Opportunities for Computer Interconnects: a Meta Platforms Perspective

Andrew Alduino, Meta Platforms

Executive Summary

We identify the three major IO regimes for GPUs in future AI/ML architectures and chart the growth in their performance versus the growth in Large Language Model (LLM) parameter sizes as a metric for AI/ML cluster performance. The growing gap between LLM requirements and component performance provides an opportunity for the integration of optical computer IO into future GPU based systems. We point to the GPU to memory IO pathway as a prime candidate for high bandwidth, low power, reliable integrated optics to support the growth in performance of future Generative AI workload.

Introduction

As a social media company with a mission of “Giving people the power to build community and bring the world closer together” [Reference 1], Meta Platforms operates over 20 different worldwide Data Center locations, representing some of the largest Data Centers in the World. Meta Platforms deploys optical communication technologies across a wide array of applications to provide the world-wide connected services required to support the current family of apps experience and required to scale in support of future user experiences such as the Metaverse. However, in alignment with the goal of this white paper, in this discussion we will limit ourselves to considering optical computer IO within AI/ML Cluster architectures. This application space is also undoubtedly the product area most ripe for innovation, since we currently project that the evolution of currently deployed ethernet networking technologies will be capable of meeting Meta’s next generation Data Center requirements in the front-end Networking and long-haul applications.

Some of the most challenging design decisions for AI/ML clusters are created by the requirements for training the emerging Large Language Model (LLM) applications. These design challenges center upon enabling the Graphics Processing Unit (GPU), which performs the massive amount of calculations required to train these models, where the sheer scale of these clusters creates unprecedented communication challenges to support the calculations and data sets required for these applications. Even limiting the discussion to consideration of GPU IO requirements still highlights 3 different IO regimes whose performance must be considered, they are GPU communication to CPUs, to other GPUs and to Memory which is currently served with HBM attached memory stacks. These three different IO regimes are illustrated in Figure 1 shown below,

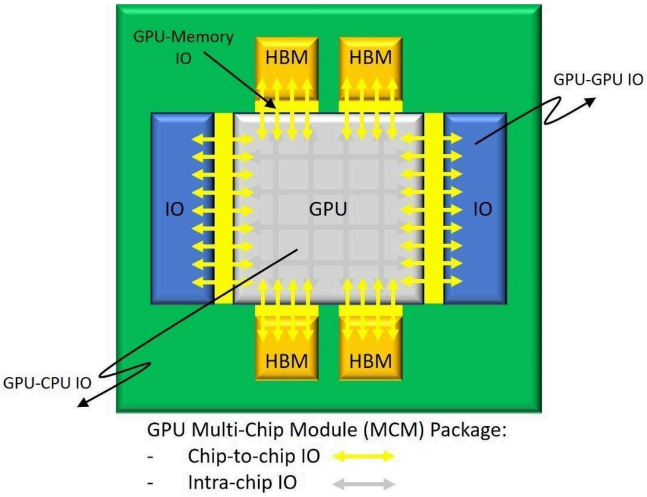


Figure 1: A drawing of one potential implementation of a GPU ASIC package identifying the 3 major IO communication regimes. The GPU MCM package contains High Bandwidth Memory (HBM), IO chiplets and the GPU ASIC. [Reference 2]

In this figure we see a drawing of a potential future advanced GPU package illustrating the three regimes of IO which were mentioned above. In this drawing the GPU is represented by the central gray square, with the lightly colored grid of arrows representing the intra-ASIC data flow between processing elements and to and from the die edges. The GPU IO traffic communication path to the CPU, which is not represented in this drawing, is illustrated in the arrow pointed down and to the left, this IO communication traffic connects the GPU to the front-end communication network and also to the storage network. The inter-GPU IO traffic, which is supported by the back-end communication network, is illustrated in the arrow pointing right in the image, and for the purposes of this drawing are illustrated as originating in the IO chiplets which are connected to the GPU ASIC itself through a die-to-die on-package interconnect. Finally, the memory traffic from the GPU is illustrated with the arrow pointing down in the upper left hand quadrant of the image, illuminating the IO data flow between the labeled HBM memory stacks and the GPU ASIC.

Considering these three IO interfaces in more detail can provide a deeper understanding of their use case, their bandwidth requirements and current expectations towards increased performance.

1 GPU Package IO Use Cases

The growth of LLMs and the broader field of generative AI have led to a vast increase in the size of the data sets used in model training applications. Additionally, the scale of the AI/ML clusters is leading to a vast scaling in the number of parameters used in the neural nets which make up these generative AI architectures. These two trends are not being met by the rate of improvement of the various contributing components which is partially illustrated in Figure 2 shown below.

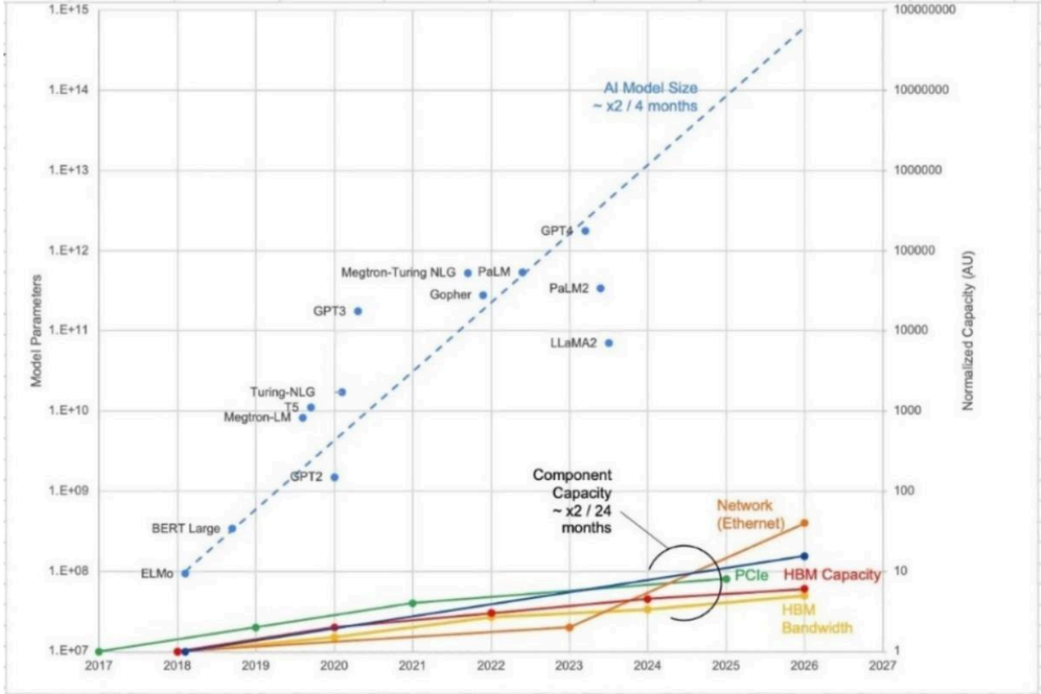


Figure 2: A plot of the number of model parameters used for a collection of LLMs as a function of time over the last 5 years. Also included is the normalized capacity scaling of 4 different IO and system parameters for AI/ML clusters; Network IO, PCIe and both HBM Bandwidth and HBM capacity. These curves highlight the increasing scaling gap between the LLM requirements and component level improvements.

In this chart we see a comparison plot of the number of parameters used in various LLMs versus the growth in capacity of Network (Ethernet) performance, PCI performance, HBM Bandwidth and HBM Capacity; all of which are increasing at a rate of ~2x every 24 months. The ~2x growth every 4 months in LLM model parameters is being met both through an increase in the size of the AI/ML clusters, as well as through architectural innovations such as the aforementioned NVidia Grace Hopper architecture. The potential for optical interconnects to address the rate of growth of HBM-like memory capacity and bandwidth is an exciting opportunity to address the AI/ML cluster architectures and use cases of the future.

1.1. GPU-CPU Communication

GPU-CPU communication links have traditionally been supported, as almost all CPU communication links are, with the PCIe physical layer. Currently deployed PCIe g4 links have a maximum aggregate bandwidth of ~64GBps/512Gbps over a physical layer of 16 channels each operating at 32Gbps. It is expected that future GPU generations will support more advanced PCIe physical layers and therefore that the bandwidth of these PCIe supported GPU-CPU connections could grow to 128GBps/1Tbps in PCIeg5 and eventually to 256GBps/2Tbps in PCIeg6. The relatively lower line rate and constrained channel counts of the PCIe physical layer lends itself well to electrical signaling solutions, with on-board routing solutions to plug-in cards in combination with external cabling very common implementations of PCIe communication links.

Nvidia has chosen to pursue a different direction to support the GPU-CPU communication link performance scaling in their latest Grace-Hopper architecture [Reference 3]. In this architecture the GPU and CPU silicon die are integrated onto the same organic substrate and connected through a highly parallel, short-distance, low power NVlink C2C interconnect capable of supporting ~900GBps total aggregate bandwidth. The impact of the co-package integration of the GPU and CPU ASICs limits the connectivity of the link to only these two ASICs, however it also results in a higher bandwidth and lower power link. The more limited connectivity model of this co-packaged integration of the GPU and CPU thus presents the AI/ML cluster architect some trade-offs, but generally the larger bandwidth available can be valuable for providing a higher bandwidth link to an additional memory tier therefore alleviating some of the pressure on the GPU to memory IO communication path discussed below.

1.2. GPU-GPU Communication

GPU to GPU communication links are the basis of the so called “back-end” fabric of an AI/ML cluster and form a large portion of the design trade-off space with which designers must contemplate. A representative back-end fabric, consisting of two switching stages and both a scale-up and a scale-out domain is illustrated in Figure 3 below.

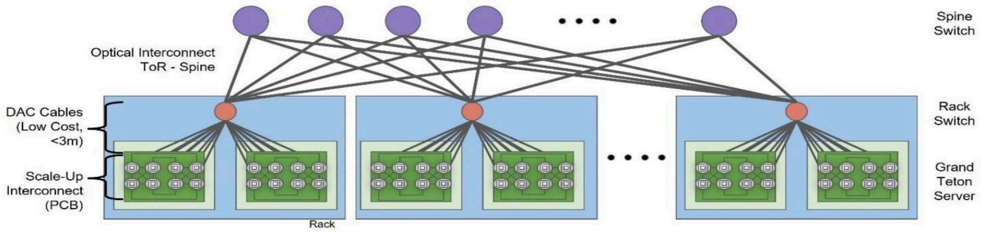

Spine Switch
Optical Interconnect ToR - Spine
DAC Cables (Low Cost, <3m)
Scale-Up Interconnect (PCB)
Rack
Rack Switch
Grand Teton Server

Figure 3: A schematic drawing of a representative AI/ML cluster back-end fabric. The fabric consists of two tiers of switching, Rack switches and Spine switches, as well as a scale-up on-board connected domain and a scale-out cabled communication domain. The PHY connection technologies, specifically on-board routing, electrical cables and optical connections, are also identified in this drawing. [Reference 4]

The back-end network of a representative Meta Grand Teton AI/ML cluster architecture is shown in this drawing. Here the GPU ASICs are represented as the circular objects which are connected through the on-board electrical traces of the scale-up interconnect which are denoted as the gray lines within the green PCB board.

The Grand Teton Server contains two of these GPU scale-up domains which are both connected to a rack switch through Direct Attach Copper (DAC) Cables which are limited to <3m for within rack connectivity. Collections of these Grand Teton Server racks are then connected through a Spine Switching tier to enable the large AI/ML cluster sizes required to support the LLM training task requirements. As illustrated in this image, the back-end network consists of two domains: the direct attached "scale-up" domain and the switched "scale-out" domain. The locality and the direct attached connectivity model of the scale-up domain lends itself to a lower cost and higher bandwidth connection, whereas the switched network fabric and the inter-rack distances of the scale-out domain generally lead to scaled down bandwidth (BW) deployments. The back-end network BW requirements between the scale-up and scale-out domains are therefore usually not equivalent. This BW performance difference is primarily driven by the cost and performance of the available interconnect technologies. As shown in this image, the scale-up connections have often been realized through on board or short distance cabled links, while the longer distances required of the rack-to-rack connections associated with the scale-out domain drive a requirement for optical connectivity. The Nvidia Grace Hopper architecture referenced above [Reference 3] supports an aggregate back-end network connectivity of 900GBps which will typically be delivered through some combination of 400Gbps and 800Gbps optical "ports". It is anticipated that both the radix, or number of ports, and the BW per port must increase to support the growth in cluster size and model size for future AI/ML workloads.

1.3. GPU-Memory Communication

A full discussion of the memory system of AI/ML clusters is far beyond this work and is also simultaneously an area of deep research and innovation as AI/ML cluster requirements and the application space continues to grow. The balance of GPU ASIC design, memory sub-system design, memory IO BW and cluster architecture is multi-variant even when not considering the software and application design. For this discussion on possible optical interconnect intercepts for the GPU to memory connections we broadly consider three different communication connectivity models. First, connectivity of the storage network, which captures the initial training model data set for instance, is provided through the front-end network. As this storage connectivity is not currently a performance limiter to the way in which a job is run on the AI/ML cluster, the BW and performance of these links are not further considered here. The second memory connectivity pathway is the connection of the GPU through to DDR memory, currently accessed through the attached CPU, which is currently supported by the lower bandwidth PCIe links and soon to be Grace Hopper architecture. This connectivity pathway was described above and will not be discussed further here. The third connection for the GPU to memory is through on-package traces connecting the GPU directly to High-Bandwidth Memory (HBM) memory stacks. These on- package die-to-die interfaces are highly efficient, short distance, highly

parallel links which currently support memory bandwidths of ~1TBps and which historically scale at a rate of approximately 50% per generation.
Therefore, the expected GPU to memory connection bandwidth for HBM memory interfaces is expected to be over 2TBps by the end of the decade. [Reference 5] These interfaces are not obvious candidates for optical interconnects due to their highly efficient and low cost nature, however the growing demands upon memory capacity and bandwidth are a significant challenge.

We believe that GPU and cluster architectures which incorporate GPU to Memory optical links with the appropriate performance, power, and cost points, could result in AI/ML cluster architectures with performance beyond what can be achieved with purely electrical links.

2 GPU Package IO Opportunity & Direction

Taken in sum, the aggregate IO requirements for future GPUs are on track to exceed 10TBps in the next 5 years, with most of this bandwidth contributed by the GPU to the memory subsystem. This high-performance requirement, coupled with the challenges of scaling the component performance as illustrated above in Figure 2 provide an opportunity for optical computer interconnects to enable GPU performance scaling for future AI/ML architectures.

2.1. Opportunity

Of the three primary communication pathways described in the chapter above, we believe that the GPU to memory connection is most ready for innovative disruption by optical computer interconnects.

Although the CPU to GPU communication pathway is an area which is undergoing significant innovation, best illustrated by the Grace Hopper architecture, the goal of this innovative architecture is to create a pathway for additional memory bandwidth and capacity per GPU. Both because the Grace Hopper architecture co-packaged integration is a non-optical way to address this IO bandwidth and also because of the affinity of this IO bandwidth on the high-bandwidth memory subsystem, we will put off consideration of this specific IO pathway and focus on the likely more impactful GPU to HBM memory communication path.

The back-end network fabric IO is the only one of the three identified GPU IO regimes currently being addressed by optical IO, at least for the longer distance, inter-rack portion of the network fabric. The optical BW requirements of this IO regime are currently being addressed with existing OSFP ethernet pluggable modules, although we anticipate that there will be significant strain on these fabrics as the IO bandwidth, and hence the relative cluster cost contribution, rise. In this IO regime, the opportunity for optical computer interconnects is to displace the on-board traces and intra-rack cabling with optical IO links, which will have significant cost and power challenges for the optical IO technology, but does provide a clear roadmap of required IO bandwidth targets. As efforts to create optical standards to address the requirements for this intra-rack switched fabric IO are under active development [Reference 6], we also avoid additional discussion here.

The GPU to memory subsystem is however a currently performance limiting element in GPU AI/ML cluster architectures and is therefore an area ripe for innovation. Although there are a wide variety of AI/ML production use cases which are deployed at Meta, there are certain applications which are limited by the available HBM capacity and which could benefit from access to additional memory capacity beyond that which can be supported by the current HBM roadmap. Advanced remote memory sub-system architectures based upon optical interconnects have the potential to address these system performance constraints and would be the target of this optical computer IO deployment. Assuming that the optical IO requirements must match or exceed the existing HBM bandwidth roadmap, leads to a BW requirement of ~2TBps HBM-equivalent bandwidths with HBM-equivalent power performance. Clearly these high bandwidth and low power targets lead to a careful consideration of co-packaged integration as a pathway to achieve these technical metrics.

Perhaps an ideal future optical computer IO solution to address these future GPU IO requirements for all three of these GPU IO pathways is one which will support a low-power host-side interface scalable to 2TBps and beyond, with HBM equivalent host side interface power of ~0.5pJ/b. Further, these IO solutions should support an optical interface for connection distances of up to cluster-scale ≤500m, capable of cost effectively supporting the 2TBps BW over a “reasonable” fiber plant for package-to-package communication. The architecture of these links should be capable of supporting multiple (>4) of these links from a single package and must support end to end memory traffic with sufficient fidelity to deliver this BW target “effectively”. It is insufficient to specify ONLY the BER required from the optical interconnect link, as the end-to-end performance of the memory transfer activities will need to be delivered at >1e-25 BER fidelity to ensure appropriate operation of the GPU to memory data transfer. An understanding of the error rates and burst error profile of the optical PHY will need to be integrated into the appropriate Forward Error Correcting (FEC) architecture and likely Link Level Retry (LLR) mechanism to ensure this data transfer fidelity.

2.2. Direction

As mentioned above, co-packaged integration of these 2TBps optical links with the GPUs seems the only feasible way to escape this much (>8TBps) bandwidth from a single GPU package. Any optical technology able to support this bandwidth, while integrated onto the GPU package and able to support the highly parallel, low power host-side electrical links, and a link distance of up to ~500m is viable. This description of a very high- bandwidth, co-packaged optical link drives interest in the density of the optical solution, bringing credence to a desirable metric of the solution as “shoreline bandwidth density”, where the amount of data which can be escaped per edge of the GPU die is measured. HBM3e is projected to achieve ~1TBps aggregate bandwidth with a shoreline bandwidth density of

~0.75Tbps/mm; we anticipate that the aforementioned 2TBps optical link will need to support this bandwidth over the same shoreline therefore setting a target of >1.5Tbps/mm shoreline bandwidth density.

The co-packaged environment presents many challenges to the optical technology and the manufacturing flow which must be addressed in order to make the technology viable. These include, but are not limited to, the issue of reliability of the optical technology; repairability and replaceability are not as desirable attributes as simple reliability, since it is most likely that the Field Replaceable Unit (FRU) will encompass the GPU package, support card and all integrated IO. This GPU FRU element will likely be very expensive and so rare failure events will be critical both to ensure that the AI/ML cluster efficiently performs the required work, but also to ensure that the optical IO enabled memory solution is economically viable. Manufacturability will be another important consideration, to support the same required cost infrastructure, the high value GPU FRU assemblies must be manufacturable with very high yield for the solution to be viable. Additionally, the GPU FRU will be very dense thermally, and so it is likely that a large liquid cold plate will be required to be integrated on the package, making vertical fiber escape solutions more challenging to implement. Highly parallel optical solutions, VCSELs for instance, are certainly acceptable solutions on many levels, however high bandwidth per fiber, to support “reasonable” fiber cost, fiber escape area and fiber connector performance are also important considerations which will impact the adoption of this technology for future AI/ML architectures.

3 Conclusion

Optical computer interconnects are a potentially enabling technology for new AI/ML cluster memory systems, if they can achieve the up to ~2TBps bandwidth required with the right combination of power, cost and reliability. These AI/ML memory systems have the potential to address the memory constraints challenging current cluster designs.

An effort by the Open Compute Project (OCP) to standardize on the physical boundary conditions for future GPU packages which would support all of the projected optical IO, commensurate with the thermal density, the cooling requirements, and the power delivery requirements of these future AI/ML systems with potentially disaggregated memory would be helpful. Further contributions to OCP from memory system vendors who could support future remote memory appliances connected with HBM equivalent IO bandwidth with large capacity and low latency would further be welcomed.

4 References

1. Meta info: https://about.meta.com/company-info/
2. Drew Alduino and Rob Stone; “AI/ML - Opportunities for Optical Interconnects” presented in XX @ OCP

   2022
3. Nvidia’s Grace Hopper architecture is described here:

   https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

4. Rob Stone; “Future Data Center Network Architectures” presented in Data Center Summit Panel @ OFC
   2023
5. One vendor’s HBM roadmap is captured in this article:
   https://www.anandtech.com/show/18982/micron-publishes-updated-dram-roadmap-32-gb-ddr5- drams-gddr7-hbmnext
6. OIF Co-Packaging standardization efforts are captured here:
   https://www.oiforum.com/wp-content/uploads/OIF-Co-Packaging-3.2T-Module-01.0.pdf

4. Optical Resource Disaggregation

Optical Resource Disaggregation

Author(s): George Michelogiannakis, Lawrence Berkeley National Laboratory

John Shalf, Lawrence Berkeley National Laboratory

Release 01 April 2024

Executive Summary

The diversity of workload requirements and increasing hardware heterogeneity in emerging high performance computing (HPC) systems motivate a system that is capable of allocating compute and memory resources in a fine-grain manner to applications; this capability is known as resource disaggregation. However, it is unclear how to efficiently realize this capability and cost-effectively meet the stringent bandwidth and latency requirements of HPC applications in order to maintain correctness guarantees and minimize performance degradation. To that end, in this white paper we summarize recent resource utilization measurements we made at NERSC's Perlmutter to motivate resource disaggregation in the diverse, scientific workload that NERSC serves. Then, we describe how we can efficiently implement resource disaggregation within racks using modern photonics. Our implementation meets the escape bandwidth of all compute and memory resources in NERSC's Perlmutter and fully meets the bit error rate (BER) requirements in modern systems. We also briefly describe recent advancements in modern photonics that make resource disaggregation more practical than in the past. Since intra-rack resource disaggregation leads to a more modular design, synergistic technologies and protocols that are being developed by open compute project (OCP) partners such as CXL are crucial to resource disaggregation's success. Our photonic-based disaggregated rack provides an average application speedup of 11% for 25 CPU and 61% for 24 GPU benchmarks compared to a similar system that instead uses modern electronic switches for disaggregation. We estimate that an iso-performance HPC system with our disaggregated racks would require 4x fewer memory modules.

Introduction

In this white paper, we start with motivating data from NERSC's Perlmutter focusing on how memory capacity is used by the diverse workload that NERSC serves. This workload represents a wide range of scientific problems that are of interest to the research community and highlights the challenge that systems face today because while they are provisioned for important workloads, those workloads are infrequent whereas the majority of workloads use only a fraction of available resources; this inevitably leads to underutilization of resources that are expensive to procure and maintain.

Then, we discuss recent advancements in modern photonics and how components that have been demonstrated or available for purchase today can be used to implement resource disaggregation within racks of compute and memory resources in NERSC's Perlmutter. As part of this discussion, we show how we can meet bit error rate (BER) and escape bandwidth requirements. Finally, we focus on the application performance impact that the added latency to and from memory due to the additional hardware to implement resource disaggregation (compared to today's non-disaggregated systems); in particular, we quantify how much faster applications execute on a photonically-disaggregated rack compared to an equivalent disaggregated rack with electronic components, and estimate how many fewer memory modules our disaggregated system has compared to a non-disaggregated system, while preserving overall system compute throughput. Due to the variety of vendors that implement the necessary components, realizing resource-disaggregated systems especially if disaggregation operates down to the chip level presents a major opportunity for defining open-source standards and protocols by open compute project (OCP) partners.

1 Perlmutter System Profiling

NERSC's Perlmutter was fully accepted in July 2023, is number 8 in the June 2023 top 500 HPC list, and consists of 1536 GPU and 3072 CPU nodes. For our analysis, we collected information from Perlmutter in the period of November 1 to December 1 of 2022 from Slurm (the job scheduler), operating system statistics, and hardware counters. We found this period of time to be representative based on sampling of other periods of usage. Since NERSC traditionally serves a diverse, open-science workload from numerous scientific domains, the variety of resource requirements will provide us with useful information. Below is a summary of key findings. You can find the full study with more details in [1].

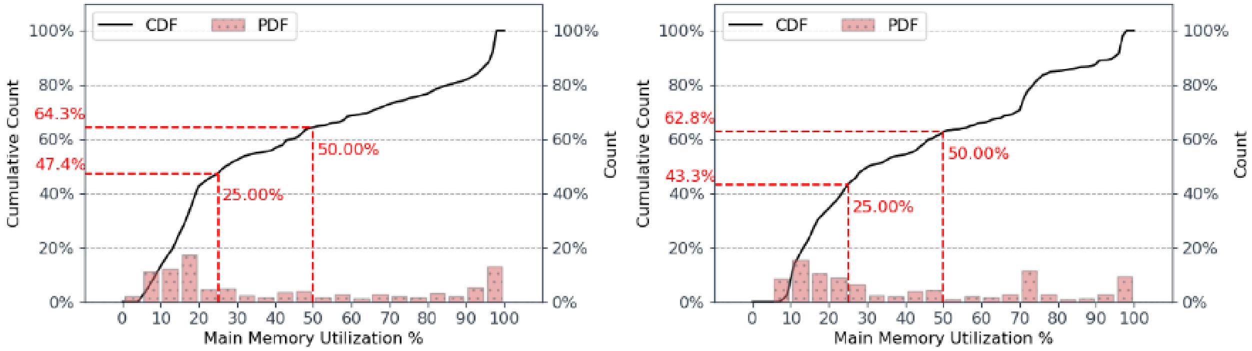


Figure 1. Per-job maximum memory capacity utilization.

Figure 1 shows per job its maximum memory capacity utilization at any point throughout the job's lifetime (left for CPU nodes and right for GPU nodes). Memory capacity here refers to the host memory found in nodes (DDR5). As shown, a considerable amount of jobs use at most between 5% and 25% of host memory. A small fraction of jobs use all or almost all of available memory.

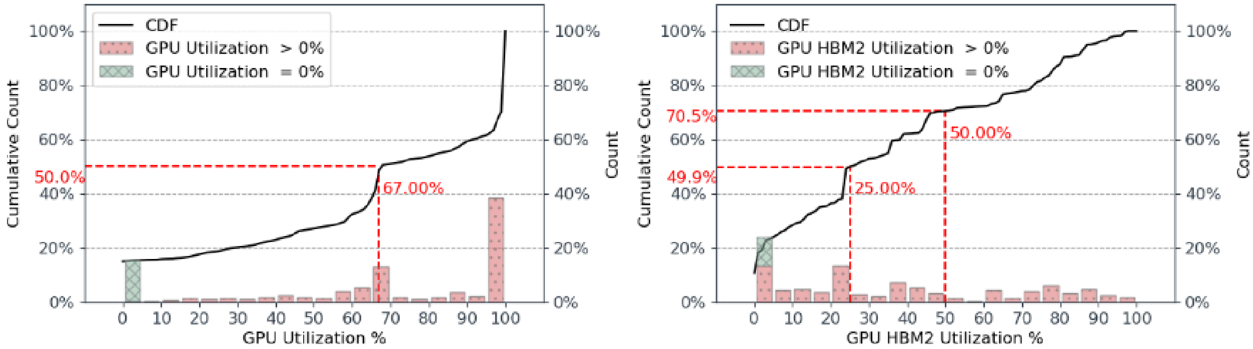


Figure 2. GPU utilization: Compute throughput on the left and HMB2 memory capacity on the right.

Figure 2 shows the utilization of GPUs in terms of their compute throughput on the left, and on the right HBM2 memory capacity. The overall trends remain the same compared to figure 1. A small fraction of jobs request GPU nodes and do not use any GPU compute resources or HBM2 capacity possibly due to software bugs, user error, or allocation policies. However, we do not notice that a larger fraction of jobs fully utilize GPU compute throughput as compared to memory capacity; this is intuitive based on the large amounts of efforts dedicated to accelerating GPU workloads.

In other analysis on the same data that can be found in [1], we notice that jobs in Perlmutter show a spectrum of spatial and temporal imbalance in memory capacity usage, defined as the variability across a job's lifetime (temporal) or among the nodes a job occupies (spatial). However, the distribution is skewed towards less variation for temporal and spatial. Finally, we notice correlation between resource usage. For instance, jobs with higher memory usage tend to use CPUs more heavily, and longer-running jobs tend to have more imbalance.

These two figures illustrate the challenge we alluded to that systems which serve diverse workloads such as the open-science workload of NERSC face: workloads that use vast amounts of memory are important but infrequent, leaving memory capacity to remain underutilized under the majority of other workloads. This presents an opportunity for resource disaggregation in HPC.

2 Disaggregated Photonic Rack

As our next step, we designed and modeled a Perlmutter rack with CPU nodes but with additional photonic components to implement intra-rack resource disaggregation. That is, the rack provides optical paths between any two CPU, NIC, and memory chips in the rack such that jobs get allocated exactly the resources they require. You can find the full study in [2].

Photonic Components

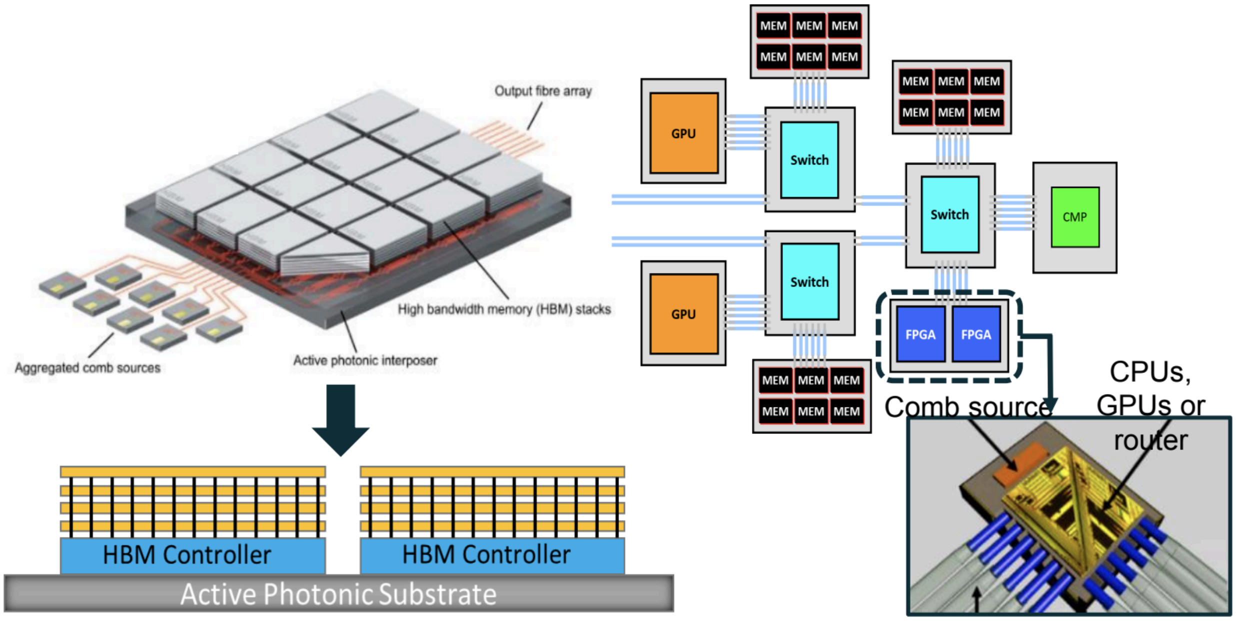
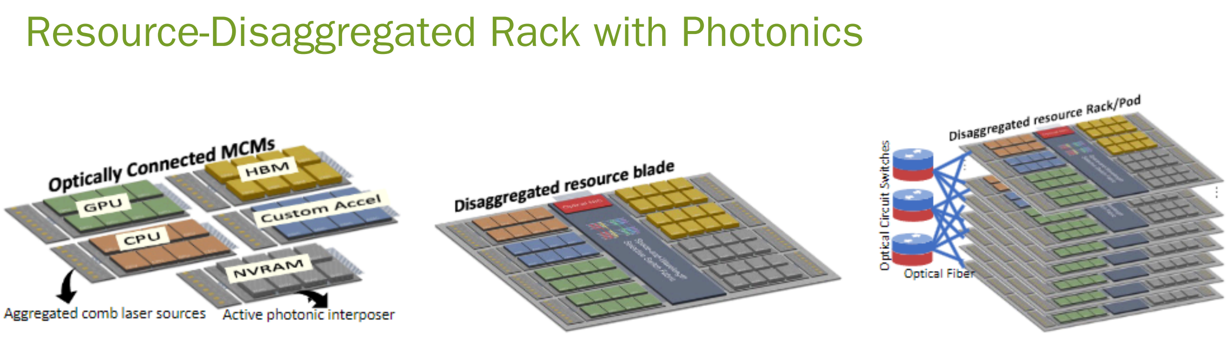


Figure 3. Left: Multi-chip modules (MCMs) with an active photonic substrate take in light from comb laser sources and output optically-encoded data to an array of fibers. Right: With this building block and small-radix optical switches, we can arrange multiple such MCMs and construct a node.

The photonic components in this study have been demonstrated or are available to purchase. An important building block are photonically-enabled multi-chip modules (MCMs) that are able to place multiple chips on an active photonic substrate. An array of comb laser sources on the side of the MCM generates multiple lasers of light, each of which has a certain number of wavelengths, each of which supports data at specific bandwidths. Information to and from chips in the MCM is driven to an array of optical fibers shown on the right. Figure 3 illustrates such an MCM example on the left. On the right, we show an illustration of how with multiple MCMs and low-radix optical switches we can form direct optical paths between resources and thus form a node.


Resource-Disaggregated Rack with Photonics
Optically Connected MCMs
HBM
GPU
Custom Accel
CPU
NVRAM
Aggregated comb laser sources
Active photonic interposer
Disaggregated resource blade
Disaggregated resource Rack/Pod
Optical Circuit Switches
Optical Fiber

(a) Active optical MCM (b) Blade (c) Rack/Pod

Figure 4. We can optically connect MCMs and place photonic switches in the middle to construct a blade (node) shown in the middle. We can then stack multiple such nodes with multiple parallel optical switches to implement a disaggregated rack.

Building on the photonic components developed in other thrusts of this project, figure 4 shows how we hierarchically compose a disaggregated rack. We arrange chips in MCMs as shown on the left. For simplicity, each MCM holds chips of the same type, such as CPUs, GPUs, etc. Each MCM is based on an active photonic interposer and aggregated comb laser sources as their source. Then, we arrange a number of those MCMs into a blade that replaces a node in today's systems. Finally, a collection of three blades make up the entire rack. MCMs connect to each other through photonic switches.

Table 1. Chips per MCM and MCMs per rack to match NERSC's Perlmutter and fully satisfy each chip's escape bandwidth.

Chip Type	Chips per MCM	MCMs per Rack
CPU	14	10
GPU	3	171
NIC	203	3
HBM	4	128
DDR4	27	38
Total	-	350

A primary goal of our disaggregated rack is to satisfy each chip's maximum escape bandwidth to avoid performance loss from bandwidth starvation. As such, table 1 shows the number of MCMs in a rack and the number of chips per type in each MCM, in order for our rack to have the same resources as a Perlmutter rack. Through our analysis, we confirm that this arrangement provides full escape bandwidth to each chip to avoid bandwidth starvation. In addition, we achieve comparable BER to and from memory than today's modern HPC system with electronics (10^{-18}) by using forward error correction and account for its latency. In particular, we used lightweight forward error correction that was proposed for CXL and can have latency as low as 2ns [3]. With these goals, we carefully modeled the additional latency from our photonic hardware, its impact to application performance, and how it compares to equivalent electronic components, explained later.

An important consideration is whether to use spatial or wavelength-selective switches. Architecturally, this choice makes a substantial difference since spatial switches provide high bandwidth between a small number of source-destination pairs, but require reconfiguration to do so. In turn, this requires an electronic control plane combined with accurate monitoring or prediction of application traffic. In contrast, wavelength-selective switches such as arrayed waveguide grating routers (AWGRs) distribute one wavelength per input port to every other output port. AWGRs do not require reconfiguration thus providing a simpler solution, but provide low bandwidth between any particular source-destination pair.

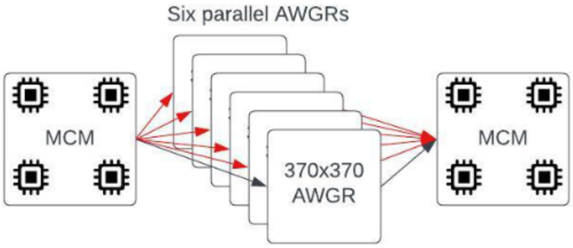


Figure 5. With multiple parallel AWGRs we can provide multiple paths between a particular source-destination pair, thus increasing point-to-point bandwidth.

Ultimately, our disaggregated rack used a series of parallel AWGRs, shown in figure 5. For this, we assume state of the art demonstrated AWGRs each of which has 370 inputs, 370 outputs, 370 wavelengths per port, and 25 Gbps per wavelength. In order to satisfy the full escape bandwidth of each chip, we require 6 parallel AWGRs for the number of MCMs and chips in MCMs shown in table 1.

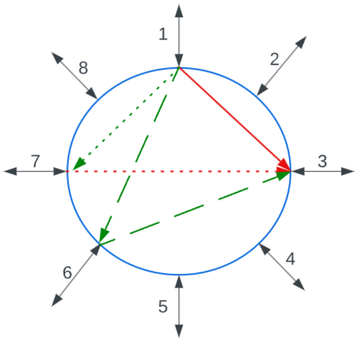


Figure 6. Indirect routing allows a particular source-destination to use available wavelengths between other pairs.

To overcome the limited point-to-point bandwidth of AWGRs, we use a distributed indirect routing scheme shown in figure 6. In that, each endpoint (i.e., each chip) has a small amount of extra logic that (i) advertises to all other endpoints through piggybacking or through extra control messages, which of its wavelengths to other endpoints are in use. Also, this endpoint distributed logic (ii), by knowing what other endpoints are busy, can pick an indirect path if the direct path to the desired destination does not offer sufficient bandwidth. In the figure above and to the right, endpoint 1 wants to send to 3 but its direct bandwidth is not enough (indicated with a solid red arrow). It could choose 7 as an intermediate destination thus making the path 1 to 7 to 3, but the path from 7 to 3 is occupied. Therefore, since it knows that the path from 6 to 3 is available, it chooses 6 as its intermediate destination thus forming the path as 1 to 6 to 3.

Performance Implications

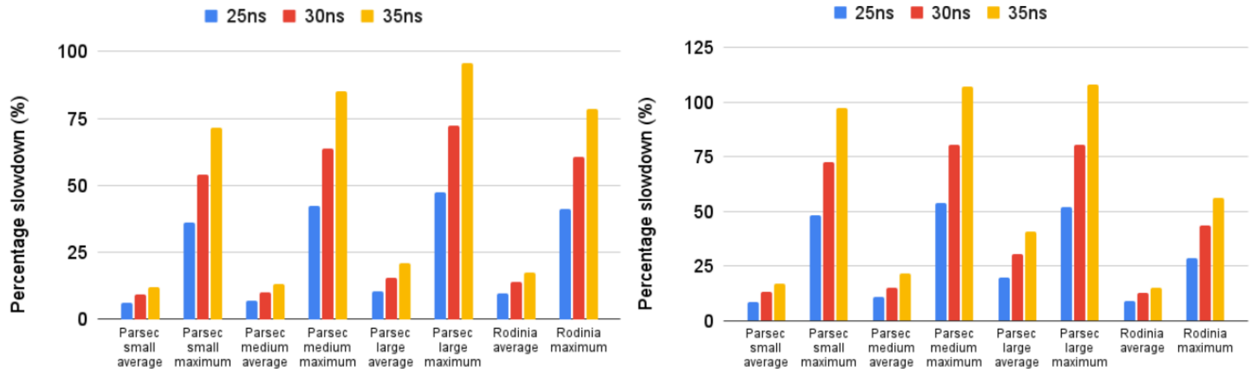
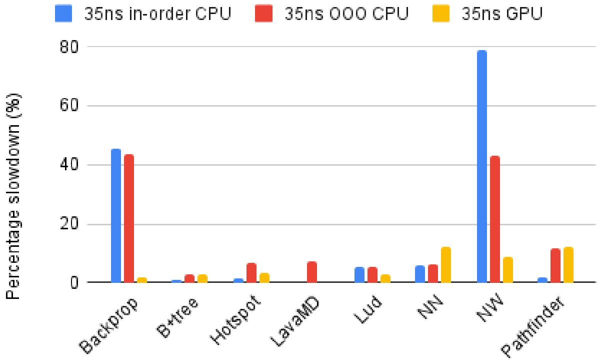


Figure 7. Application performance penalty (slowdown) as a function of additional latency to access main memory.

Having satisfied the BER and bandwidth requirements, our focus now is to evaluate the impact of the added 35ns to application performance. For this, we rely on architectural simulation for CPU and GPU benchmarks executed with a single compute core in order to better isolate the impact of the added latency to and from memory. Figure 7 shows the average and maximum slowdown of applications, by average per benchmark suite and input data size, for an extra 25ns, 30ns, and 35. The left figure is for in-order CPU compute cores and the right figure for out-of-order CPU cores. As expected, applications show a variety of behavior that we found strongly correlates with the resulting miss rate of the last level cache (LLC). That is, the higher the LLC miss rate, the more intense slowdown an application experiences. We notice that the average for 35ns ranges about 10% to 20%. While our photonic hardware results in an added 35ns latency, we show 25ns and 30ns to illustrate the importance of photonic interconnect components and forward error correction with a lower latency since the slowdown for 25ns is considerably slower.



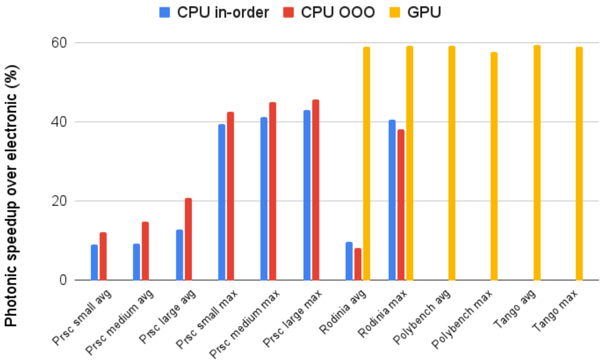


Figure 8. Left: Application slowdown for GPU benchmarks. Right: How much faster applications execute on our photonically-disaggregated rack versus an equivalent one with electronic links and switches to implement resource disaggregation.

Comparing similar results with GPU executions, figure 8 (left) shows the same benchmarks in their GPU versions and their slowdown. As shown, GPUs hide memory latency better and thus are good candidates for disaggregation. Figure 8 (right) compares our photonic disaggregated rack against an equivalent rack that implements disaggregation with electronic links and switches. As shown, our photonic switches provide considerable speedup compared to their equivalent electronic counterparts, which makes a strong case for photonics to implement disaggregation. Here we see that the benefit is strong for GPUs, because we notice that GPUs can hide a certain amount of latency, but beyond a certain amount the compute resources in GPUs start becoming starved of data, which results in a large performance penalty. Finally, through other calculations on system-wide performance, we estimate that our disaggregation approach can result in an iso-performance system with 4x fewer memory and 2x fewer NIC modules.

3 Conclusion

We show how photonic components that have been demonstrated today can be used to implement a resource-disaggregated rack that allows applications to execute faster than alternative implementations with electronic components. Such systems require interoperability of multiple components thus open-source standards and implementations that the OCP can contribute will provide a major boost in making resource-disaggregated systems a reality.

4 Glossary

AWGR: Arrayed waveguide grating routers.

BER: Bit error rate.

MCM: Multi-chip module.

OCP: Open Compute Project.

5 References

[1] Li, J., Michelogiannakis, G., Cook, B., Cooray, D., Chen, Y. (2023). Analyzing Resource Utilization in an HPC System: A Case Study of NERSC’s Perlmutter. In: Bhatele, A., Hammond, J., Baboulin, M., Kruse, C. (eds) High

Performance Computing. ISC High Performance 2023. Lecture Notes in Computer Science, vol 13948. Springer, Cham. https://doi.org/10.1007/978-3-031-32041-5_16

[2] George Michelogiannakis, Yehia Arafa, Brandon Cook, Liang Yuan Dai, Abdel Hameed Badawy, Madeleine Glick, Yuyang Wang, Keren Bergman, and John Shalf. 2023. Efficient Intra-Rack Resource Disaggregation for HPC Using Co-Packaged DWDM Photonics. arXiv preprint arXiv:2301.03592 (2023).

[3] S. Van Doren, “Abstract - hoti 2019: Compute express link,” in 2019
IEEE Symposium on High-Performance Interconnects (HOTI), 2019, pp. 18–18.

5. OPTICAL INTERCONNECT – PATHWAYS TO AN OPEN AI INFRASTRUCTURE

WHITE PAPER: OPTICAL INTERCONNECT – PATHWAYS TO AN OPEN AI INFRASTRUCTURE

Author: Matthew Williams, CTO, Rockport Networks

Release 01 April 2024

Executive Summary

Artificial intelligence is transforming every industry, generating massive amounts of data and driving demand for more capacity and specialized hardware. Closed server systems designed for capacity computing are hitting a wall, increasing overhead, limiting choice and creating an unsustainable environmental footprint.

Rockport Networks designs open AI infrastructure systems to optimize capacity and heterogeneity without inflating costs or power consumption. Our approach aligns with the Open Compute Project (OCP) goals of open design and best practices for data center efficiency, at-scale operations and sustainability.

Proven in HPC, Rockport's distributed switching and optical interconnect technology are now being applied to achieve the efficiency and cost advantages of Composable Disaggregated Infrastructure (CDI) at data center scale. Our open AI model makes it possible to use the same pluggable transceivers and fiber optic cabling to build advanced networks that aren't limited by the constraints of spine-and-leaf architectures. Using off-the-shelf optics, operators can achieve the best performance per watt per dollar in the industry. Already, we've modeled a 35% reduction in per-rack power consumption for AI infrastructure.

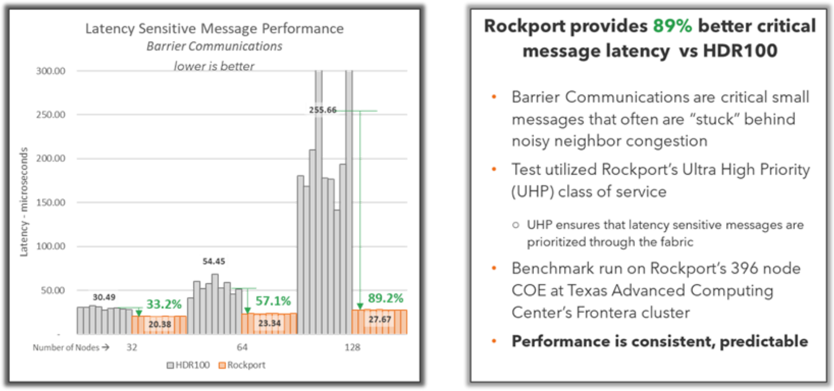
Extending this open AI approach to protocol-agnostic transport, we can deliver the advantages of CXL memory pooling for greater efficiency and scale years before the 3.0 standard is available.

Introduction

Artificial intelligence (AI) and Machine Learning (ML) applications have a wide range of point-to-point latency and bandwidth requirements with multi-directional traffic that is highly sensitive to congestion. Today, data centers have centralized spine-and-leaf switching architectures for Ethernet, InfiniBand, and other protocols – all built for homogenous computing. Emerging use cases require greater capacity and highly optimized topologies to handle complex traffic patterns and heterogeneous workloads.

Rockport has deployed a fully distributed fabric architecture that embeds the control forwarding and data plane into every node. Distributed fabric nodes are connected by the Rockport SHFL, a completely passive optical interconnect that requires zero power, cooling, or configuration. Topology-independent, the SHFL can be used with off-the-shelf optics and scales to hundreds of nodes in servers or PCIe expansion chassis for a high-radix infrastructure with 15-150 Tbps of fabric bandwidth. The advantages of this distributed architecture have been proven in performance-intensive AI/ML and HPC workloads. SHMEM benchmark results, shown in Figure 1, demonstrate an 89% improvement in latency performance for critical barrier messages.

The next generation of direct interconnect technology is being applied to Composable Disaggregated Infrastructure (CDI) to provide an open, sustainable model for AI infrastructure. Topologies can be optimized by use case, and transport decoupling ensures that the right transport environment is applied for each protocol. This topology-and-transport independence allows for robust PCIe scaling and scalable CXL memory pooling years in advance of CXL 3.



Rockport provides 89% better critical message latency vs HDR100

- Barrier Communications are critical small messages that often are "stuck" behind noisy neighbor congestion
- Test utilized Rockport's Ultra High Priority (UHP) class of service
  - UHP ensures that latency sensitive messages are prioritized through the fabric
- Benchmark run on Rockport's 396 node COE at Texas Advanced Computing Center's Frontera cluster
- **Performance is consistent, predictable**

osu_shmem_barrier test using Open MPI 4.1.2 UCX 1.13

Figure 1: SHMEM benchmarks show lower latency for critical messages in a Rockport distributed fabric.

Sustainable AI Infrastructure

Direct interconnect topologies enable highly efficient bandwidth utilization and low-latency performance. Typically, network fabrics are composed of point-to-point links between endpoints and first-level switches, and a set of point-to-point links between each switch layer. When high-performance and/or distance are required in these links, multi-lane optics are the principal method of achieving sufficient per-link bandwidth. Commonly, links are composed of four, eight or 16 parallel lanes, with each lane supporting usable bandwidth of 25, 50, 100 or 200 Gbps. These lanes are combined to provide the desired per-link bandwidth. For example, a 200 Gbps link may be composed of four 50 Gbps lanes or eight 25 Gbps lanes, depending on the optical solution chosen. Higher per-lane bandwidth drives significantly higher transceiver cost and power consumption along with corresponding increases in SERDES power consumption and complexity. The additional cost and power of utilizing these high per-lane bandwidth transceivers can exceed the increase in performance, creating an overall increase in both dollar/performance and power/performance.

MAXIMIZING EFFICIENCY

In a distributed direct interconnect, there are no centralized switches. Instead, each node in the network contains both a host interface (Ethernet, PCIe, CXL, etc.) and a distributed hardware-based switch. Instead of using a set of parallel lanes to create a single point-to-point link, these distributed switches treat each lane as an individual link. Each of these independent single-lane links is used to create direct-connections to a large community of direct-neighbor nodes forming a mesh. Nodes that aren't directly connected to each other utilize other nodes to provide transit switching capability. The high degree of connectivity in a direct interconnect fabric ensures very high path diversity, ensuring consistently high performance for advanced applications. Each network flow leverages all available network capacity, addressing a key performance challenge of networks that are limited to single paths per flow or destination.

Commodity pluggable transceivers such as eight-lane QSFP-DD or OSP modules or 16-lane OSFP-XD modules are currently used to provide low-cost, power-efficient access to a high density of optical lanes. As direct-interconnect solutions can aggregate the available capacity of multiple paths, the most power- and cost-efficient per-lane bandwidth can be utilized, typically based on 28 Gbps NRZ or 56 Gbps PAM4 technologies, depending on the target use case.

OVERCOMING SCALE LIMITS

In spine-and-leaf networks, there are scaling breaking points where the radix of the centralized switch drives the requirement for an additional layer of switches to support additional endpoints. In a distributed direct interconnect model, there are no scaling breakpoints. As the endpoint node is the only active component, these solutions provide a perfectly linear per-endpoint cost and power profile.

Distributed direct interconnect solutions will also have an inherent advantage when co-packaged optics (CPO) become widely available. By co-packaging the optics with the distributed hardware switch, the power efficiency gains will allow a higher lane density (e.g. 24 or 32 lanes) and a higher per-lane bandwidth (e.g. 100 Gbps lanes utilizing 4x 25 Gbps NRZ/DWDM). One of the challenges of CPO is the impact of component failures in the overall system model. Since the direct interconnect hardware switches are distributed, the impact of a CPO failure is restricted to the endpoint node. The high path diversity of the overall fabric ensures that the blast radius is restricted to that failed node. In a centralized switching application, a CPO failure may require a costly and disruptive full-switch replacement with a large blast radius that would affect all connected nodes.

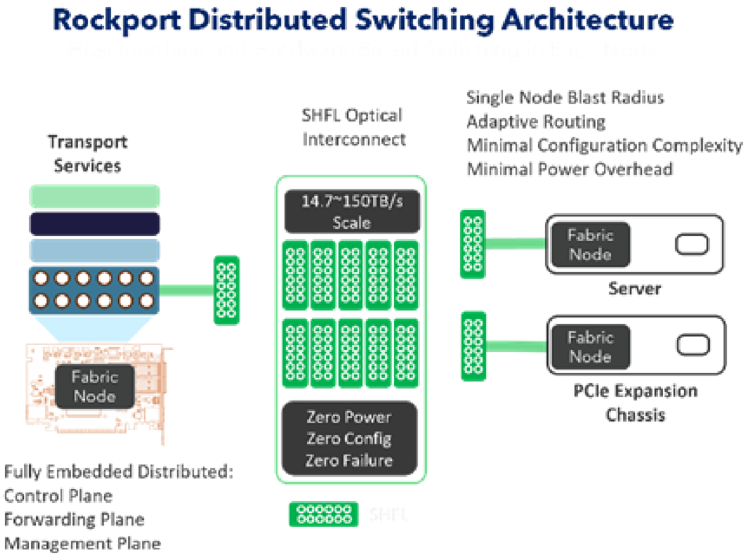


Figure 2: Distributed design enables greater efficiency, low-latency performance and scalability.

Pre-wired Topologies

One of the challenges with traditional direct interconnects has been the complexity of manually wiring the target topology. With a high number of links per endpoint that each need to be wired to the correct neighbor node, direct interconnect solutions have typically been limited to very large scale HPC systems.

To address this wiring challenge in the distributed model, Rockport uses a commodity high-density passive optical cable (e.g. MTP/MPO 24/32) to connect 12 or 16 links from each node to a port on a passive SHFL. This cable contains multiple fiber pairs that carry the independent single-lane links. Inside the SHFL, each of these links is broken out of the high density-cable and connected to the port associated with the correct target node. In this way, the complexity of the wiring pattern is addressed during manufacturing. A single passive optical cable connects each node to a SHFL, and additional passive optical cables connect SHFLs together to create large-scale inter-rack connectivity. The power- and cost-savings, simplicity, and flexibility of passive cabling, along with the reduction in overall cable count, deliver significant advantages over typical multi-layer centralized switching environments.

The pattern of connectivity of the target topology is pre-wired into the rack-mountable SHFL. The SHFL breaks out each of the links and physically routes them to the target neighbor node, forming a direct optical path. Traffic is spread across multiple active paths, to avoid creating network hot-spots. Bandwidth-sensitive traffic is spread across physically independent paths, providing simultaneous access to the full fabric bandwidth. Critical latency-sensitive messages (e.g., barrier operations) are designated ultra-high priority for routing across the shortest available paths.

While multi-lane pluggables (QSFP, OSFP, etc.) are generally used for “A” to “B” connections, the SHFL connects “A” to multiple neighbors to create much more advanced topologies with any number of racks and any number of nodes in every rack. The model supports efficient inter-chassis and intra-chassis communication for CPU-x, GPU-GPU or any resource type. Any number of nodes can be added (or removed) at any time without having to pre-purchase and pre-wire the switching for simplified in-place scaling.

Open System Architecture

Composable Disaggregated Infrastructure (CDI) is an open system architecture that allows provisioning on-demand by connecting any remote device such as GPUs, other accelerators, SSDs and DRAM into nodes at job-execution time. The Rockport SHFL implements supercomputer topologies in CDI systems by making simple, predefined connections between fabric nodes in the host and chassis. Network topology and route information automatically update when compositions are added, removed, unavailable, or if an individual link path is down. The topology can be changed at any time by adding a new SHFL configuration.

For AI/ML workloads, Rockport achieves GPU density by placing PCIe devices in chassis, where communications between large numbers of GPUs are now local. This is a substantial advantage, since up to 70% of the execution time for AI/ML tasks can be consumed in GPU communication. Routing traffic across multiple paths in the underlay fabric makes it possible to connect multiple chassis and servers without any interfering cross-traffic on the network. Dedicated paths between the chassis and compute nodes increase performance predictability by avoiding network congestion from other workloads or chassis.

Rockport overcomes the limitations of PCIe in scaling CDI by providing transport adaptation services that decouple the native PCIe transport from the underlay fabric. From a host point of view, it looks like the native protocol, but the virtual PCIe switch can now leverage Rockport's traffic transport services at scale while enabling reliable, 24x7 operation with built-in failover and hot-swapping capability. A remote PCIe switch complex enables local communication between devices located in a physical chassis. Placing an end-to-end reliable network connection between the host PCIe root complex and a remote PCIe device addresses the reliability issue that causes system failures inherent to native PCIe expansion. The PCIe service translates between global and local PCIe addresses, alleviating the memory space exhaustion issue by isolating remote devices to their own local memory spaces inside their virtual switch.

CDI USE CASES

GPU Density: Large Language Models (LLMs), neural networks and other large-scale AI/ML applications require the capacity to solve problems faster and at a much larger scale. Composable Disaggregated Infrastructure offers a cost-effective alternative for GPU density to optimize utilization and ensure that processing capacity is matched to the workload. Rockport can be deployed at multi-row scale, so that GPUs can be added in any quantity or location based on their workloads. A single host CPU attached to 32 GPUs can serve multiple resource pools for different workload types or schedules. Resources can be reallocated after the workload or application is finished running. Implementing CDI architecture, Rockport has modeled a 35% reduction in per-rack power consumption for GPU infrastructure.

Storage Agility: Managing "always-on" storage resources such as NVMe create challenges for scheduling upgrades or replace failed hardware. The cost and complexity of ripping and replacing hardware drives 3-year upgrade cycles. The distributed architecture of the Rockport CDI system model enables built-in failover and hot-swapping without manual intervention or powering down for systems that run continuously. Rockport connects resources in a fully redundant underlay fabric that isolates lost system links and reroutes traffic across a diverse set of paths to ensure high reliability. When devices fail, they can be replaced or upgraded in software from shared resource pools and deployed at any time without service disruptions.

CXL Memory Pooling

The advantages of Composable Disaggregated Infrastructure (CDI) and transport adaptation services are also applied to CXL. By decoupling the host protocol from the underlay fabric, Rockport can bring Type 3 memory pooling capability early and apply advanced topologies and transport capabilities at scale. Our CXL design will enable large-scale memory composability based on CXL 1.x and 2.x endpoints. Since the Rockport underlay fabric is designed to be agnostic to the protocol, it's possible to take CXL 1.x and 2.x systems and scale them to hundreds or even thousands of endpoints. Years before CXL 3.x is available, Rockport's memory composability will enable full memory scaling. Once CXL 3.1 endpoints are available, Rockport will provide full CXL 3.1 compatibility in addition to our large-scale CXL 1.x and 2.x support.

With CXL deployments, SHFLs are designed to ensure that servers accessing remote memory pools have one or more direct connections to the memory pool. This flexibility in design allows the solution to address the requirements of a wide range of memory pooling use cases.

Rockport's distributed design also simplifies upgrades to future generations of CXL. Only the fabric node within each endpoint needs to be replaced. The existing node-to-SHFL cabling, SHFLs, and SHFL-to-SHFL cabling can remain as-is, making upgrades time and cost efficient. Additionally, an existing deployment could change to a completely different topology simply by replacing the passive SHFL architecture without replacing the fabric nodes or cabling to the SHFLs. The fabric nodes will simply discover the new topology and build sets of source routes to the destinations.

Conclusion

Rockport delivers an open AI infrastructure model that enables a more sustainable data center by maximizing resource utilization, minimizing costs and energy consumption, and supporting device choice. Very high path diversity ensures that all connected devices can access the entire fabric bandwidth, ensuring consistently high performance for advanced AI applications.

By optimizing Composable Disaggregated Infrastructure (CDI) at operational scale, we offer data centers greater control over hardware life-cycles and help to eliminate vendor lock-in. Using the same off-the-shelf optics, cables and other components, data centers can build different networks that are calibrated to deliver the best efficiency and performance at the lowest possible cost. Bringing advances such as scalable CXL memory pooling early allows data centers to future-proof their AI infrastructure.

This open approach to AI infrastructure is of interest to data centers in Enterprise, Cloud and HPC segments that need cost-effective scale and predictable performance for complex, heterogeneous workloads and latency-sensitive traffic.

We invite members of the Open Compute Project (OCP) working stream to apply for our Early Access Program to validate our platform and optical approach in different environments. Input from the OCP Community will help us to gather additional requirements for using optical direct interconnect technology in open AI systems.

6. Optical CXL Interconnect for Large Scale Memory Pooling

OPTICAL CXL INTERCONNECT
FOR LARGE SCALE MEMORY POOLING

Authors:

Huaiyu Meng, Lightelligence

Weifeng Zhang, Lightelligence

Ron Swartzentruber, Lightelligence

Optical CXL Interconnect to Enable Large Scale Memory Pooling

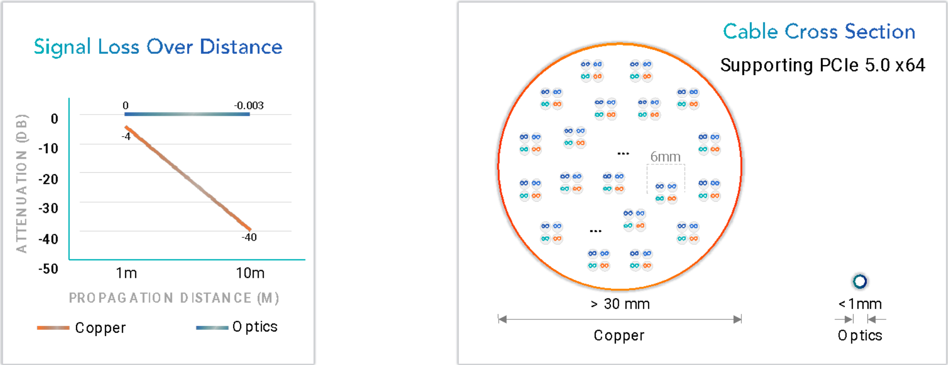
1 Problem Definition

Artificial intelligence (AI) applications have been widely used to intelligently analyze the petabyte-scale data generated by human society every day. Data analysis at this scale not only requires huge computation power, but also consumes vast storage capacity, posing unprecedented challenges to the computing and memory capabilities at data centers and edge devices. Meanwhile, AI models are still rapidly evolving with average parameter size increasing tenfold every year. The trend of AI evolution leads to several challenges to the existing computing infrastructure:

1) Limitation of computing resource capacity: The high computational and memory requirements of AI models, such as large language models (LLM), make it infeasible to completely store the model in the accelerator's on-device memory. Hundreds, even thousands, of high-end accelerators with fast interconnect are required to conduct AI training or inference.
2) Limitation of scalability: Large-scale distributed computing devices between datacenter racks usually rely on Ethernet-based data communication due to relatively large physical distance. So, it is very difficult to achieve the near-linear expansion of large-scale computing power due to bandwidth limitation and high communication delay of Ethernet-based data transportation.
3) Low resource utilization: Servers with fixed configurations often suffer resource underutilization due to unbalanced allocation among CPUs, accelerators, and memories. For example, memory stranding [1] has been identified as a dominant source of memory waste (up to 25%) from Azure, thus a potential source of massive cost savings.

In response to the challenges above, resource disaggregation has emerged as a new computing paradigm to achieve scalability and flexibility of computing power. By pooling and dynamically sharing the resources, flexible allocation of computing resources and elastic expansion of computing power can be implemented. The stranded resources, unused otherwise in the traditional data center, can be recomposed to create more virtual computing machines under the same provision of resources.

However, efficiency of resource disaggregation largely depends on bandwidth and latency of interconnection technologies. At the same time, the signal loss due to the extended reach of the interconnect needs to be solved for reliable data transportation. As shown in Figure 1, the optical interconnect provides much smaller signal loss over distance and better signal density. Thus, we propose the optical CXL-based [2] interconnect as a viable approach to address these problems and enable large-scale resource pooling.



Assuming AWG26 wire, PCIe 5.0 signal

32 cables with diameter > 6mm (CAT8)

16 fibers with diameter of 0.125mm

Figure 1. Comparison of copper cable and optical fiber to support CXL scaling

2 Optical CXL Interconnect

Compute Express Link (CXL) is an open industry standard interconnect running on the PCIe 5.0 and 6.0 physical layer (PHY) infrastructure, offering high-bandwidth and low-latency connectivity between host and CXL devices. CXL supports dynamic multiplexing of three protocols: CXL.io, CXL.cache, and CXL.mem. CXL.mem uses coherent load/store semantics to access data, avoiding memory barriers or even software-defined critical conditions. Compared with Ethernet RDMA based communication, CXL can result in much lower performance overhead and much shorter latencies, from the order of 10us with Ethernet to the order of 100ns with CXL.

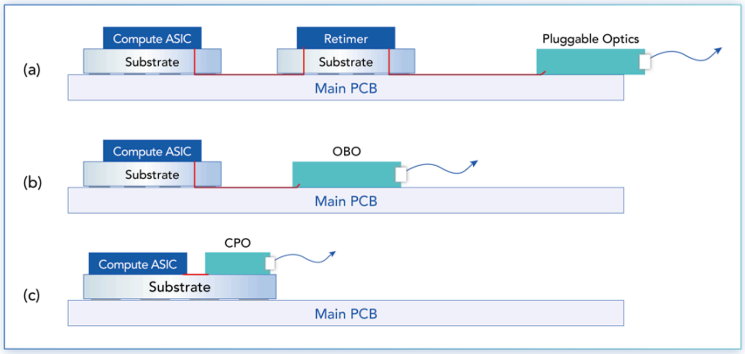


Figure 2. Interconnect between optical transceiver and compute/memory module

To improve optical CXL interconnect, the distance between optical module and computing / memory module can be further shortened to help reduce system power consumption and latency. Figure 2 shows a potential evolution of optical interconnect, from a pluggable optical transceiver module, an on-board optics (OBO) module, to co-packaged optics (CPO).

Figure 3 shows one type of system architecture via optical interconnect. Optical interconnect is realized by a set of 3D stacked electronic and photonic chips. The combined structure can be packaged around the computing chip and CXL memory controller, realizing a co-packaged optics module.

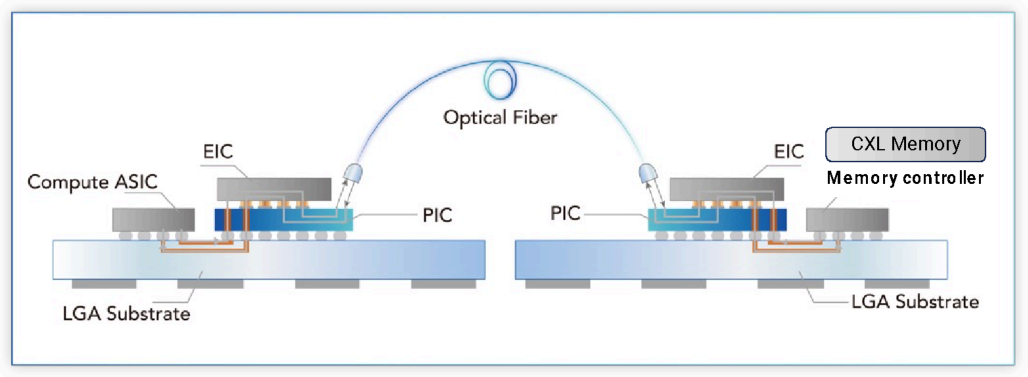


Figure 3. System architecture of silicon photonics-based CPO solution

3 Simulation and Results

To demonstrate the potential and basic value of CXL memory pool with optical interconnect, we implemented a much-simplified simulation platform based on the existing AMD Genoa CPU and Samsung CXL memory. Here is the hardware system configuration and the application setup for simulation.

Hardware Configurations			Workload and Inference framework	
Server	o AMD EPYC 9124 16-Core CPU o Samsung DDR5 4800 MT/s o System MEM0 size: 256GB o System MEM1 size: 256GB o Bandwidth: 307GB/s		LLM model	o OPT-66B o Batch size = 24 o Context length: 512 o Output length = 8
GPU	o Nvidia A10 GPU, Gen4x16 o Device memory: 24GB o Bandwidth: 32GB/s			
NVMe	o Samsung, Gen4 x4 o MEM size: 1.92TB o Bandwidth: 8GB/s		Inference Framework	FlexGen
CXL 1.1 memory	o Samsung, Gen5 x8 o Memory size: 128GB o Bandwidth: 32GB/s			

AMD Genoa CPU can support CXL 1.1 protocol. Samsung CXL memory has a size of 128GB with PCIe Gen5 x8 configuration. The optical interconnect is CXL 2.0 compliant with a bandwidth of 32GB/s and a latency of 60ns.

As AI models grow bigger and bigger, we envision that an AI model and associated data won't be able to fit into GPU no-device memory and not even the system memory during training or inference. CXL memory pool can be leveraged to offload the training data and model parameters rather than the NVMe storage. This scenario resembles the legacy server which uses the disaggregated memory box to expand its memory capacity and solve the power provision problem at the same time.

The large language model (LLM) OPT-66B[3] is selected to do inference on the system above. We also choose the open-source AI inference framework FlexGen[4] which is a high-throughput generation engine for running LLMs with limited GPU memory. The OPT decode throughput performance is compared between FlexGen NVMe-based offloading (baseline) and CXL-memory based offloading.

Table 1 shows the measured performance using the simulation configurations above. CXL memory achieves ~2.4x higher throughput than the NVMe baseline. As expected, pure CXL memory offloading still performs a little worse than the native system memory due to its bandwidth and latency. However, by implementing a proper memory tiering policy (e.g., mixed allocation of 60% CXL memory and 40% system memory), the overall performance is completely comparable with the native system memory performance.

OPT-66B model	NVMe (baseline)	CXL Memory	System Memory	Allocation Policy (60% : 40%)
Decode throughput (tokens/s)	1.984	4.859	6.216	6.237
Decode latency (s)	338.7	138.2	108.1	107.7

Table 1. FlexGen decode performance comparison between NVMe-based offloading and CXL memory offloading

Finally, Figure 4 shows the computing resource utilization during FlexGen inference. Both CPU and GPU utilization are improved due to the reduced offloading overhead via CXL memory.

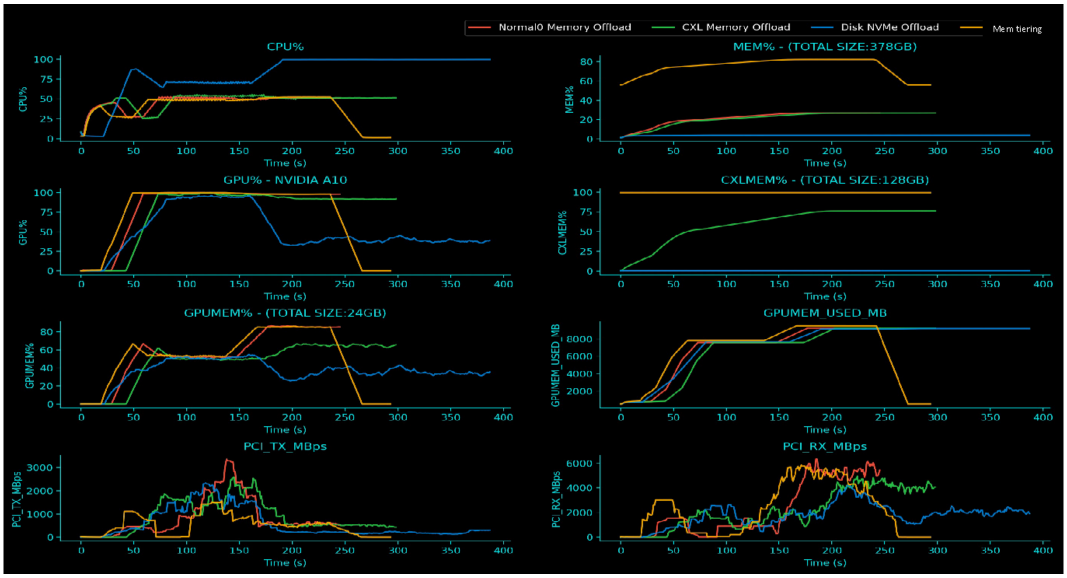


Figure 4. Resource utilization during FlexGen inference

4 References

1) Huaicheng Li, Daniel S. Berger, et al, “Pond: CXL-Based Memory Pooling Systems for Cloud Platforms”, ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS ’23)
2) Compute Express Link (CXL), https://www.computeexpresslink.org/
3) Susan Zhang, Stephen Roller, et al, “OPT: Open Pre-trained Transformer Language Models”, https://arxiv.org/abs/2205.01068
4) Ying Sheng, Lianmin Zheng, et al, “FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU”, the 40th International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023

7. The Rack Scale Market Opportunity

The Rack Scale Market Opportunity

Authors: Bob Wheeler

LightCounting

The Rack Scale Market Opportunity

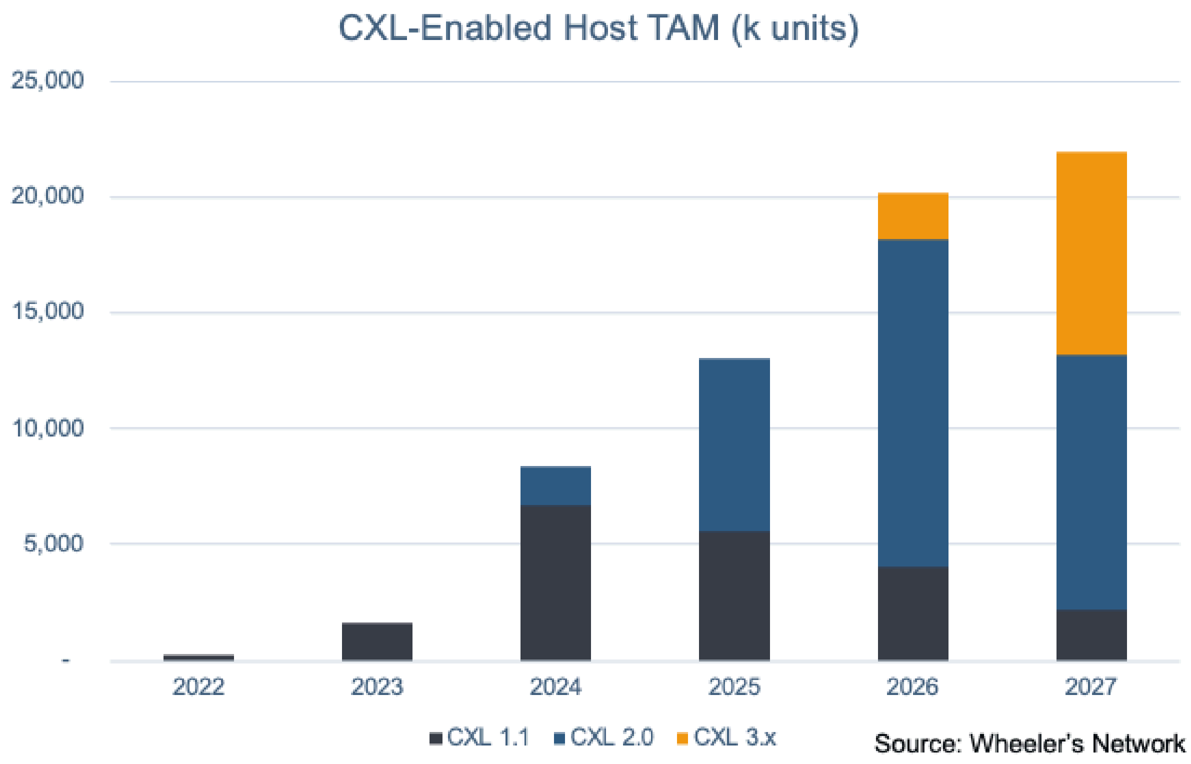
We view main-memory disaggregation as the primary opportunity for photonic interconnects within the rack. Storage and networking disaggregation are well established, with various architectures deployed in hyperscale data centers. By contrast, progress toward main-memory disaggregation has been slow. Intel promoted pooled-memory concepts a decade ago with its Rack-Scale Architecture (RSA) using silicon photonics, but its roadmap changed before these concepts could be productized. The Gen-Z Consortium formed in 2016 backed by OEMs and memory vendors, but products reached only the proof-of-concept phase before being terminated. Compared with prior efforts like RSA and Gen-Z, CXL is moving quickly to market.

CXL Server Penetration Sets the Stage

To size the potential market for CXL over photonics, we begin by forecasting the total available market (TAM) of CXL-enabled servers. We forecast server-unit shipments by CXL generation, that is, CXL 1.1, 2.0, and 3.x. The CXL 1.1 server ramp officially began in 2022 with AMD’s introduction of Genoa. Although Intel shipped some Sapphire Rapids processors beginning in 3Q22, those early versions may not support CXL. The Xeon versions (SKUs) announced in 1Q23 support CXL 1.1, matching AMD’s Epyc. Both vendors’ processors offer CXL support on a subset of their PCIe lanes, specifically four x16 CXL ports per socket. This means a two-socket (2P) server can offer up to eight x16 slots that handle CXL 1.1 in addition to PCIe Gen5.

We expect next-generation server processors will support CXL 2.0, which reuses the PCIe Gen5 physical layer. These should include Intel’s Sierra Forest in 1H24 followed by Granite Rapids in 2H24. The Birch Stream platform, which will handle both Sierra Forest and Granite Rapids, should continue to offer four x16 ports that handle CXL. AMD’s Turin should ship in mid-2024, although the company hasn’t narrowed the launch timing and platform details remain under wraps.

Expected in 2026, the next server cycle should add CXL 3.x support based on PCIe Gen6. Given the Gen6 upgrade will double per-lane bandwidth, we don’t expect the new platforms to increase the number of CXL ports per socket. In other words, CXL bandwidth will double if lane counts remain the same as the prior CPU generation. With three generations shipping in 2026, we expect virtually all servers will support CXL, as Figure 1 shows.



Source: Wheeler's Network

Figure 1. CXL-enabled server forecast through 2027

CXL-Attached Memory Use Cases

One reason CXL will succeed where prior efforts failed is that it starts by enabling simple single-host use cases. As shown on the left in Figure 2, the CXL memory expander disaggregates the memory controller from the host CPU. This use case enables memory-capacity and memory-bandwidth expansion while decoupling the memory technology (DDR/LPDDR/etc.) from the CPU. It also represents the best case for CXL-memory performance, assuming a direct connection without intermediate devices or layers such as retimers and switches. This use case does not represent a high-volume opportunity for photonics, however, as it is generally implemented within a server chassis.

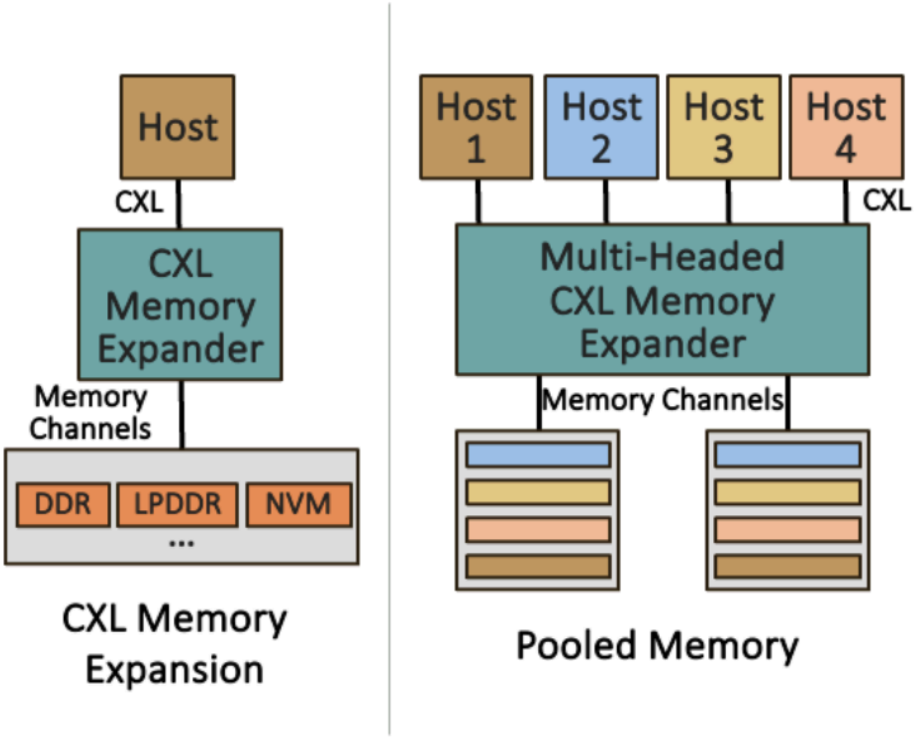


Figure 2. Initial CXL use cases

The next use case is pooled memory, which enables flexible allocation of memory regions to specific hosts. In pooling, memory is assigned and accessible to only a single host—that is, a memory region is not shared by multiple hosts simultaneously. When connecting multiple processors or servers to a memory pool, CXL enables two approaches. The original approach added a CXL switch component between the hosts and one or more expanders. The downside of this method is that the switch adds latency, which we estimate at around 80ns for CXL 2.0.

The alternative approach, shown on the right in Figure 2, instead uses a multi-headed (MH) expander to directly connect a small number of hosts to a memory pool. Memory pooling can alleviate the problem of stranded memory, which impacts the capital expenditures of hyperscale data-centers operators. System designers can build several different physical topologies for multi-host memory pooling. One option is a large chassis like Meta's Grand Teton platform, which integrates an ExaMAX cabled backplane for PCIe Gen5 signals. Alternatively, designers can physically disaggregate the hosts and pooled memory into compute nodes (e.g., 1U chassis) and a separate memory appliance, respectively. This case requires cables to connect the compute nodes to the appliance's CXL ports.

Server Memory Pooling Challenges

Latency is the key metric for memory access, and even the simplest CXL topology can double latency compared with CPU-integrated memory controllers. To minimize the performance impact of added latency, software must become memory-tier aware. This work has already begun, with Meta's Transparent Page Placement (TPP) as a prime example. TPP was upstreamed to the Linux kernel, and Meta presented early performance simulations across several workloads at the 2022 OCP Global Summit. Public-cloud providers, however, don't control their customers' workloads, so new memory tiers must be transparent to applications.

The added cost of CXL components presents another challenge. The primary goal of memory pooling is to reduce total cost of ownership (TCO) by recovering stranded memory. The cost of CXL components, however, offset the reduction in DRAM cost. For the appliance topology, these components include a chassis, power supply, and cabling, as well as the cost of CXL expander chips. Even a quick TCO analysis will show that memory pooling cannot support the pricing of existing optical solutions. For example, an active optical cable (AOC) with 8x50Gbps lanes currently sells for more than $400.

CXL in GPUs and Accelerators

The CXL end game is in enabling new disaggregated rack-level architectures, but achieving this vision requires many new features that add complexity. Moreover, whereas server penetration is a certainty, CXL penetration of GPUs and accelerators is far more speculative. Today, data-center GPUs employ proprietary GPU-to-GPU interconnects such as NVlink, xGMI, and Xe Link. Support for CXL 3.x in a GPU is unlikely before 2025.

The CXL Consortium recently published the CXL 3.1 specification, which adds features required for true fabrics. These include peer-to-peer communications, port-based routing, multi-level switch hierarchies, and shared fabric-attached memory. Whereas memory pooling enables flexible allocation of DRAM to servers, CXL 3.1 enables true shared memory. The shared-memory expander is called a global fabric-attached memory (G-FAM) device, and it allows multiple hosts or accelerators to coherently share memory regions. G-FAM enables a large number of GPUs to share a large pool of fabric-attached memory, removing dependence on CPU-attached DRAM.

In the long term, CXL 3.x promises more-composable rack-level architectures, which disaggregate compute, memory, storage, and networking resources. As Figure 3 shows, CXL-attached DPUs can serve as shared NICs, eliminating the traditional Ethernet top-of-rack (ToR) switch. Adoption of shared NICs could be an important inflection point, as it would consolidate rack-level cabling around CXL. By contrast, the initial memory-pooling use case requires CXL cabling in addition to existing Ethernet server-to-ToR cabling.

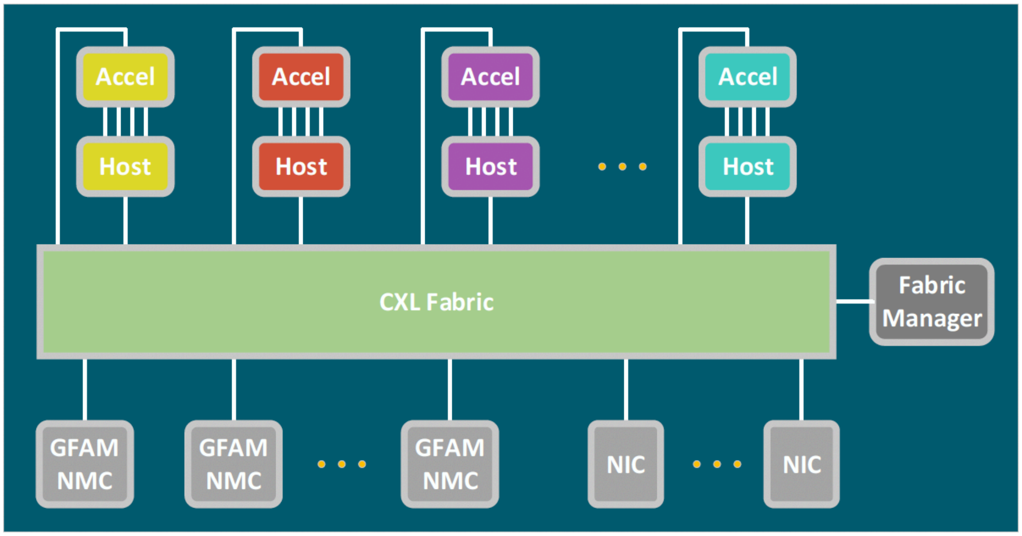


Figure 3. CXL 3.x fabric example topology (Source: CXL Consortium)

The Role of Photonics

The incumbent solution for CXL cabling is passive copper cables (DACs) driven by electrical retimers. Although bulky, these DACs provide sufficient reach for memory pooling within one rack. Customers would prefer optical fibers, as they reduce weight and remove reach limitations. Once pooling scales beyond one rack, optical interconnects become a requirement.

The biggest barrier to optics adoption is cost—existing solutions are simply too expensive. Radically new optical interconnect solutions are needed to compete with copper connectivity. Existing optical solutions may serve a few niche CXL applications, but they won't penetrate memory pooling in hyperscale data centers. To be successful, this Future Technology Initiative must start with cost as a central requirement, perhaps building off of the work already being done in the Extended PCIe Connectivity workstream.

Improving the utilization of costly data-center GPUs has high value, providing more room in the TCO equation for optical interconnects. An open question is how the FTI will address proprietary GPU interconnects. Does the OCP tenet of openness require this work to consider only open standards such as CXL? Or can its scope include proprietary protocols (e.g., NVLink) carried over short-reach photonics?

8. Energy Efficient optical Links for PCIe

Energy Efficient optical Links for PCIe

Author (s): Ranovus

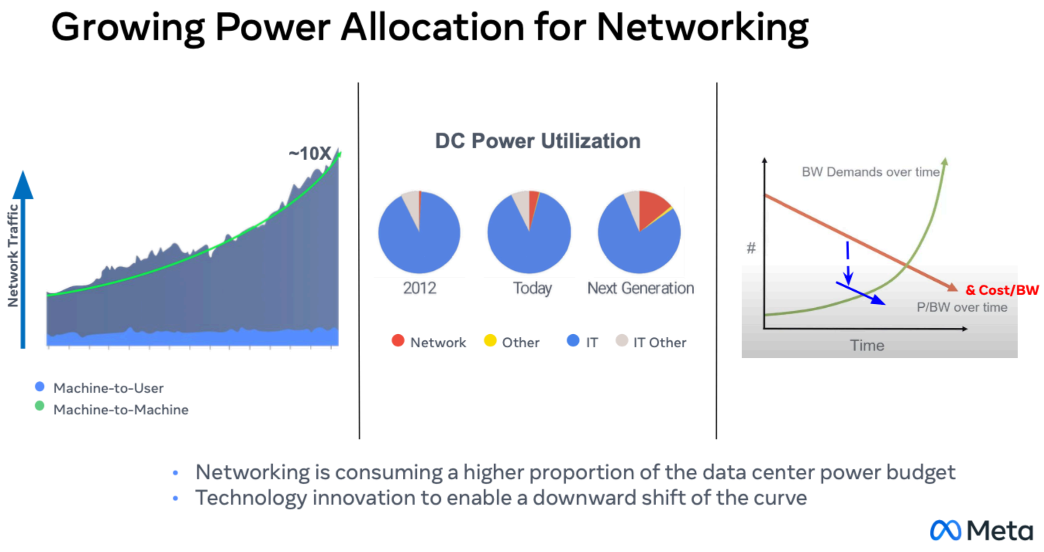
Release 01 April 2024

Market Drivers

There has been significant interest implementing links with Energy Efficient interfaces for a variety of reasons:

- Reduced power consumption
- Improved density
- Reduced latency

Figure 1 tells a well-known story that networking power has been increasing its share of data center power with each new generation and has led to calls for improving energy efficiency of the next generation links.


Growing Power Allocation for Networking

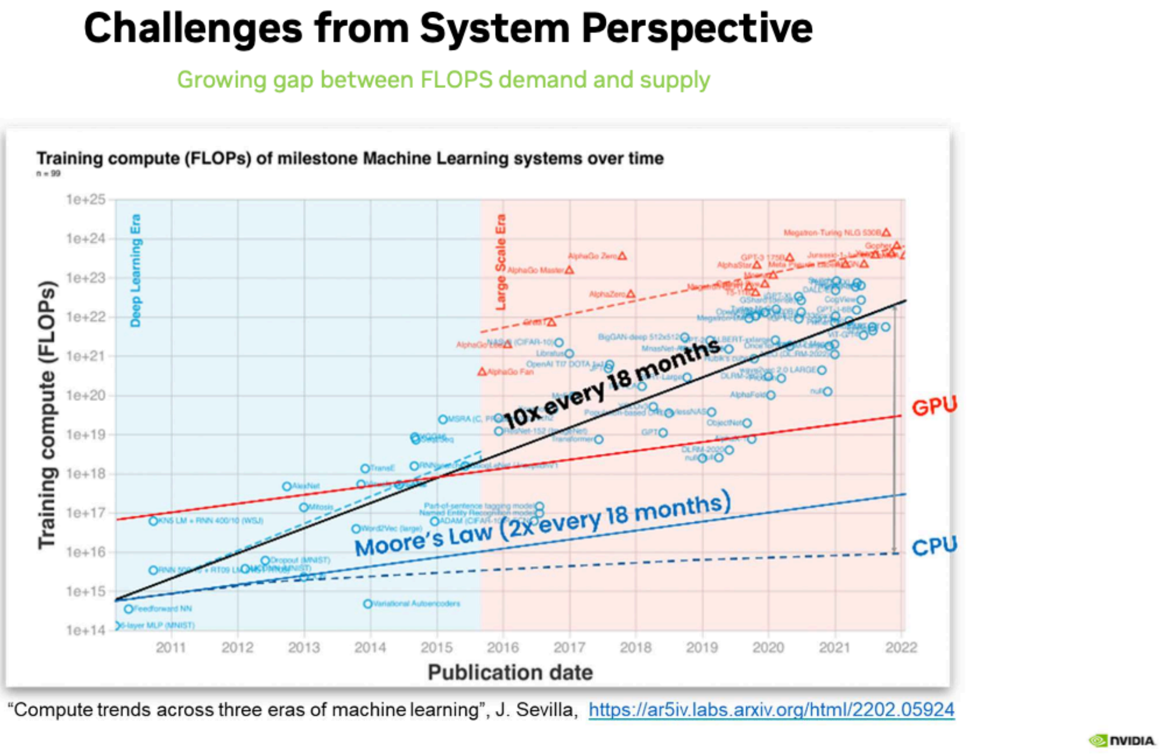
- Networking is consuming a higher proportion of the data center power budget
- Technology innovation to enable a downward shift of the curve

Meta

Figure 1 - Network Power Trend[1]

Figure 2 tells us that the growth in the size of the AI models is outstripping what can be implemented with current industry solutions leading to the call for technologies for increased density and reduced latency.

[1] Presented at Lightcounting's "Linear Drive Enables Green All-Optical Connectivity for Datacenters – Webinar", March 2023



Challenges from System Perspective

Growing gap between FLOPS demand and supply

"Compute trends across three eras of machine learning", J. Sevilla, https://ar5iv.labs.arxiv.org/html/2202.05924

Figure 2 - Growth of Models[1]

This whitepaper will focus on solutions for low latency, energy efficient PCIe links supporting AI in the data center. Many of the backend data center links utilize low latency PCIe connections with higher level protocols like CXL.

Challenges of using Optical Links with PCIe Protocol

The PCIe protocol was designed around copper-based links. Many of the features in PCIe protocol depend on physical properties of copper connections. For example, optical links are inherently unidirectional, whereas the copper link transmitters can sense whether an electrical link is terminated. These features "break" when optics are inserted into the link.

The list of features which are problematic is quite extensive, but a few key ones are summarized below. Note that a given PCIe link likely won't support all PCIe features. However, to be PCIe compliant, an optical engine (OE) would need to support them to guarantee interoperability in all cases.

First a note on terminology to avoid confusion. In the optics world, transmitter refers to an optical transmitter, but in the PCIe world, transmitter refers to an electrical transmitter. To avoid confusion, the Tx and Rx functions are prepended with either an “e” for electrical or an “o” for optical to indicate the type of transmitter or receiver as shown in Figure 3.

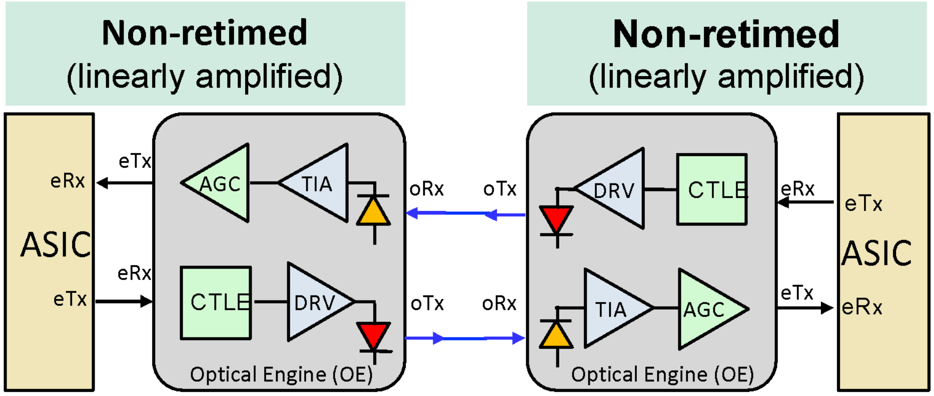


Figure 3 - A Non-retimed Link

Rx Detect

PCIe electrical receivers can signal to a remote electrical transmitter that they are ready for communication by lowering their impedance from high-Z to a typical 100Ω differential level. Unfortunately, optical links typically don’t support a variable electrical receiver (eRx) impedance feature based on what the other end of the optical sees, so a workaround or a PCIe protocol change is needed.

Another related issue is that optical links often take longer than copper links to be ready for traffic. With some optical technologies, this delay may cause the PCIe state machine timeout. One approach would be to use Rx Detect on the eRx to signal when the optical link is ready for carrying traffic.

Inconsistent Signaling

Optical links typically function best when there is consistent data rate. However, PCIe has a variety of states where a consistent data rate is interrupted. Low power states and data rate changes are such two examples where the electrical transmitter (eTx) is in electrical idle.

During electrical idle, the electrical transmitter (eTx) holds the differential pairs to approximately zero-differential voltage. The electrical receiver (eRx) detects whether its inputs are at approximately zero-differential voltage. In addition, when an optical link is placed between an ASIC’s electrical transmitter (eTx) and an ASIC’s electrical receiver (eRx), the optical link may have difficulty reproducing the zero-differential voltage within the required PCIe specification limits.

When traffic resumes, optical engines usually require some time to return to their optimum bit error rate (BER) level. The PCIe link can fail if the optical link's delay to reach sufficient BER performance interferes with the timeout associated with the electrical receiver's adaptation.

There are also some signaling states where the transmitted signal can approach the lower cut-off frequency for the PCIe link and can be challenging for some optical engine technologies to support.

Other PCIe Considerations

There are a few other items to consider.

- Hot plug events may need to be supported even when an optical link is involved.
- Spread spectrum clocks may not work well with optical engines and so it may be necessary to confine the link to SRNS (Separate Reference clocks with no Spread spectrum clock) mode.
- Support for side band signals may also need to be considered for the optical links.
- Finally, the additional latency of longer links possible with optical fiber needs to be considered and adds roughly 5 ns/m.

Optical PCIe Link Architectures

There are a variety of ways optics can be inserted into PCIe links. These links can be characterized by several figures of merit:

1. **Link Accountability**: The ability of determining the reason a link doesn't perform well. An excellent link accountability score indicates that the link will perform with a variety of electrical and optical channel configurations ensuring the link will work once connected up.
2. **Energy Efficiency**: The amount of energy required per bit transferred. Energy efficiency is typically quantified in terms of pico-joules per bit and includes the power consumed by optical light sources.
3. **PCIe Compatibility**: Indicates whether the root complex and end point need to adjust their protocol to support the inclusion of an optical link for any optical technology. An excellent PCIe compatibility indicates that any version of PCIe can be supported by the root complex and end points. As described in § 3.0, pure optical links typically have a low PCIe compatibility score.
4. **Configuration Frozen**: Indicates when the optical link configuration is frozen. This can be at the time of deployment or perhaps earlier in the manufacturing process.
5. **Link Latency**: A measure of the latency incurred by including an optical link. The latency of an optical link is a combination of the optical engine's (OE) circuitry as well as the time-of-flight (TOF) of an optical signal which is roughly 5 ns/m in fiber.

6. **Serviceability**: Indicates how easy it is to repair a failed optical link element. For example, a Fault Tolerant Overlay Network (FTON) panel pluggable optical module would have a high serviceability score.
7. **Interoperability Margin**: Indicates a relative amount of additional margin available to deal with a variety of link configurations.

Examples of PCIe Links with Optics

Figure 4 shows a variety of PCIe optical link configurations to highlight various interoperability cases.

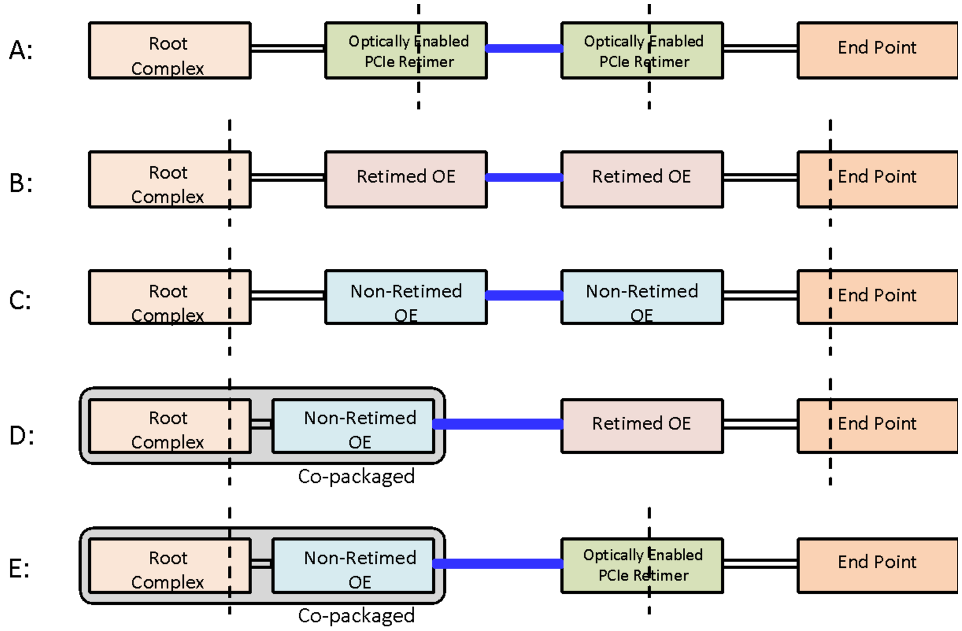


Figure 4 - PCIe Optical Link Interoperability

The vertical dashed lines indicate the boundaries encapsulating the optical link from the PCIe protocol.

Optically Enabled PCIe Retimer (A)

This link utilizes an *optically enabled PCIe retimer* where the PCIe retimer is co-packaged with the optical engine. This configuration allows the optical link to be encapsulated by the two retimers. Most, if not all of the PCIe protocol optical link issues, can be hidden from the rest of the PCIe link which enables the link to be highly PCIe compatible. Although the addition of PCIe retimers add latency and power, there are significant advantages in link accountability and backwards compatibility.

PCIe Link with Retimed Optics (B)

This link utilizes a retimed optical engine (OE) which is not protocol aware, therefore, the root complex and endpoint may need adaptations to support optics. An OE with a CDR retimer provides retiming boundary which helps to isolate the performance of the electrical channels from the optical channel providing increased link accountability compared to a non-retimed OE.

PCIe Link with Non-retimed Optics (C)

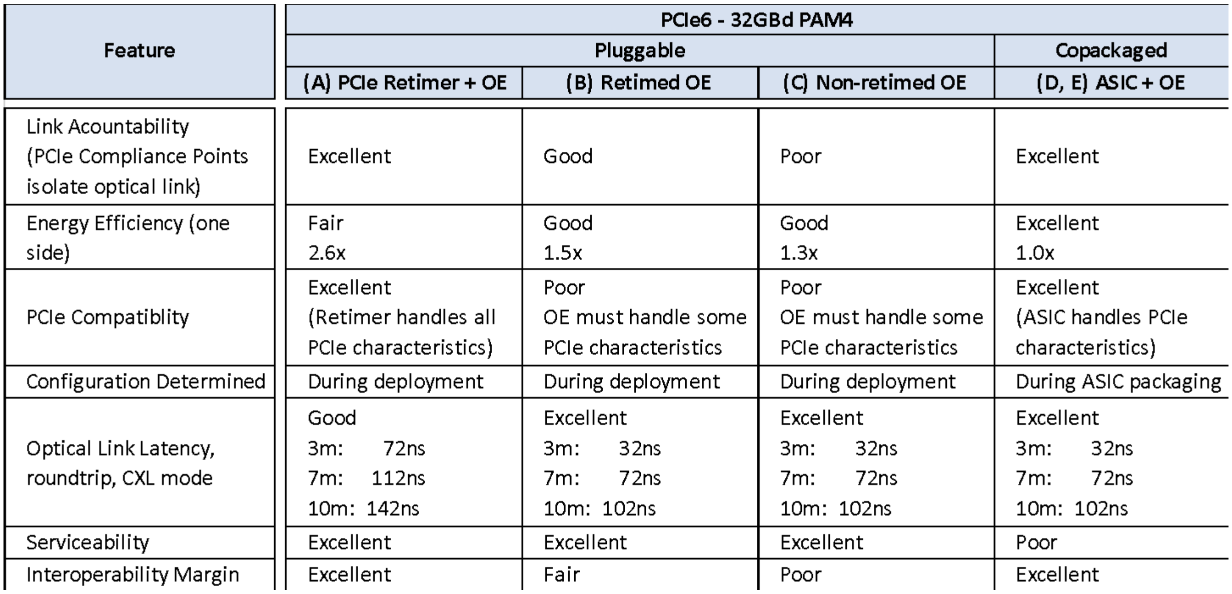
This link utilizes a non-retimed optical engine (OE) which is not protocol aware, therefore, the root complex and endpoint may need adaptations to support optics. An OE without a CDR retimer doesn’t isolate the performance of the electrical channels from the optical channel and so the link accountability is poor.

PCIe Link with Co-packaged Root Complex or End Point (D&E)

Co-packaging a non-retimed optical engine (OE) with a root complex (RC) or end point (EP) greatly improves link accountability. This is due to the short and well controlled electrical channel between the root complex and OE. The co-packaging of the OE and RC adds complexity and minimizes the power consumption.

Comparison of Optical PCIe Architectures

Table 1 shows a summary comparison of these links.



Feature	PCIe6 - 32GBd PAM4			
	Pluggable			Copackaged
	(A) PCIe Retimer + OE	(B) Retimed OE	(C) Non-retimed OE	(D, E) ASIC + OE
Link Acountability (PCIe Compliance Points isolate optical link)	Excellent	Good	Poor	Excellent
Energy Efficiency (one side)	Fair 2.6x	Good 1.5x	Good 1.3x	Excellent 1.0x
PCIe Compatiblity	Excellent (Retimer handles all PCIe characteristics)	Poor OE must handle some PCIe characteristics	Poor OE must handle some PCIe characteristics	Excellent (ASIC handles PCIe characteristics)
Configuration Determined	During deployment	During deployment	During deployment	During ASIC packaging
Optical Link Latency, roundtrip, CXL mode	Good 3m: 72ns 7m: 112ns 10m: 142ns	Excellent 3m: 32ns 7m: 72ns 10m: 102ns	Excellent 3m: 32ns 7m: 72ns 10m: 102ns	Excellent 3m: 32ns 7m: 72ns 10m: 102ns
Serviceability	Excellent	Excellent	Excellent	Poor
Interoperability Margin	Excellent	Fair	Poor	Excellent

Table 1 - Comparison of Optical PCIe Architectures

Recommended Architectures

There are two configurations which are excellent on most figures of merit and have differing trade-offs:

1. (A) Optical enabled PCIe retimer which is excellent in all categories at the expense of energy efficiency and latency.

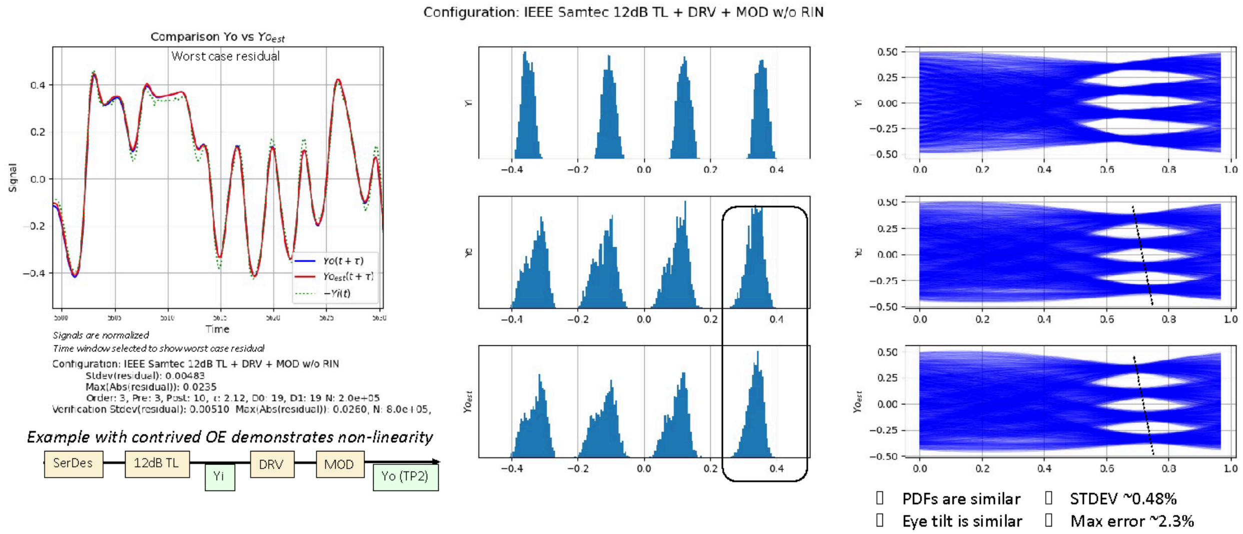
2. (E) Co-packaged root complex or end point which is excellent in all categories at the expense of serviceability and requires early commitment to the optical configuration.

Ranovus believes that these two options represent the best paths forward for the industry.

Modeling Non-retimed Links

Non-retimed links can provide significant energy savings but need to be carefully designed to provide sufficient link accountability. Link modeling is typically performed using linear time invariant (LTI) models. However, optical links are unlikely to be LTI systems, therefore need to be simulated as non-LTI systems and contain static and dynamic non-linearities. Finally, non-LTI models from differing suppliers utilize different simulation platforms making it difficult to develop interoperable standards.

To address this, Ranovus has developed a Generalized Non-linear Optical Engine Model (GNOEM) which is expected to support a wide variety of optical engine technologies for use by standards entities to aid in the development of standards.



Signals are normalized
Time window selected to show worst case residual
Configuration: IEEE Samtec 12dB TL + DRV + MOD w/o RIN
Stdev(residual): 0.00483
Max(Abs(residual)): 0.0235
Order: 3, Pre: 3, Post: 10, τ: 2.12, D0: 19, D1: 19 N: 2.0e+05
Verification Stdev(residual): 0.00510 Max(Abs(residual)): 0.0260, N: 8.0e+05,

Example with contrived OE demonstrates non-linearity

- PDFs are similar
- Eye tilt is similar
- STDEV ~0.48%
- Max error ~2.3%

Figure 5 – GNOEM Modeling Results

Figure 5 shows the input and output of the generalized model. The upper right row shows a PDF and eyes for a 100GbE PAM4 input signal, *Yi.* The right middle row shows likewise for an OE output, *Yo*. Finally, the right bottom row shows the predicted PDF and eye using a generalized, technology independent model, *Yo,est*, which matches the OE output quite well.

Having such a model enables the industry to confidently move forward with implementations of energy efficient, low latency optical links.

Summary

Embedding optical links in a PCIe link to realize low latency, energy efficient link is of great interest to the industry. However, there are several challenges as PCIe is not optics friendly and while the PCI-SIG may adapt the PCIe protocol to make it more optics friendly, backwards compatibility issues will likely still exist.

There are a two PCIe optical link architectures which provide the best trade-offs for the next generation of optical links:

1. Optically Enabled PCIe retimer (PCIe retimer co-packaged with an optical engine)
2. Optical Engine Co-packaged with a Root Complex or End Point.

9. Ayar Labs Optical I/O For The Next Wave of Computing

AYAR LABS OPTICAL I/O FOR THE NEXT WAVE OF COMPUTING

Author (s): Ayar Labs

Release 01 April 2024

Executive Summary

Ayar Labs has demonstrated working monolithic in-package optical I/O (OIO) solutions delivering 4Tbps bi-directional throughput per chiplet. This marks the arrival of a new universal I/O solution based on micro-rings that enables chips to communicate with each other within a chassis, across chassis, between racks and even across a row, at the power, latency, and bandwidth density of in-package interconnect. Several technology trends point to the arrival of optical I/O chiplets as a critical industry inflection point.

Increasing demand for artificial intelligence, high-performance computing, hyper-scale data centers, disaggregated memory and aerospace data collection and communication has led to the development of a variety of powerful, high-throughput, system-on-chip (SoC) solutions that redefine traditional computing architectures. These new architectures accelerate the data rate of communications between die, sockets, boards, systems, and racks. Electrical I/O has been a bottleneck to build scale-out and disaggregated systems that are needed for the next wave of computing. Optical I/O technology delivers the metrics needed to build the computing infrastructure needed for tomorrow's AI, HPC and datacenter workloads.

Introduction

Experts agree that electrical SerDes, the most common form of electrical I/O, is hitting a wall. Going beyond 112 gigabits per second (Gbps) is extremely challenging because the large signal losses in copper interconnects at the board level make it hard to transmit data further than a few centimeters at such a high data rate. The next wave of high-performance computing architectures requires a new form of universal I/O that eliminates the bottlenecks created by electrical I/O.

Optical communications began replacing electrical cabling in high-performance computing (HPC) and datacenter applications several years ago, changing copper cables between electrical faceplate ports to pluggable optical transceivers with connectorized fiber cables as shown in Figure 1. As faceplates become constrained by mechanical and thermal limits, optical communications are moving from the faceplate to inside the package. In-package optical I/O is a revolutionary approach that integrates silicon-photonic chiplets built with CMOS processes inside a multi-chip package (MCP). OIO chiplets eliminate electrical I/O bottlenecks and transcend process limitations to unleash the next wave of innovation in semiconductor and datacenter design [7]-[9].

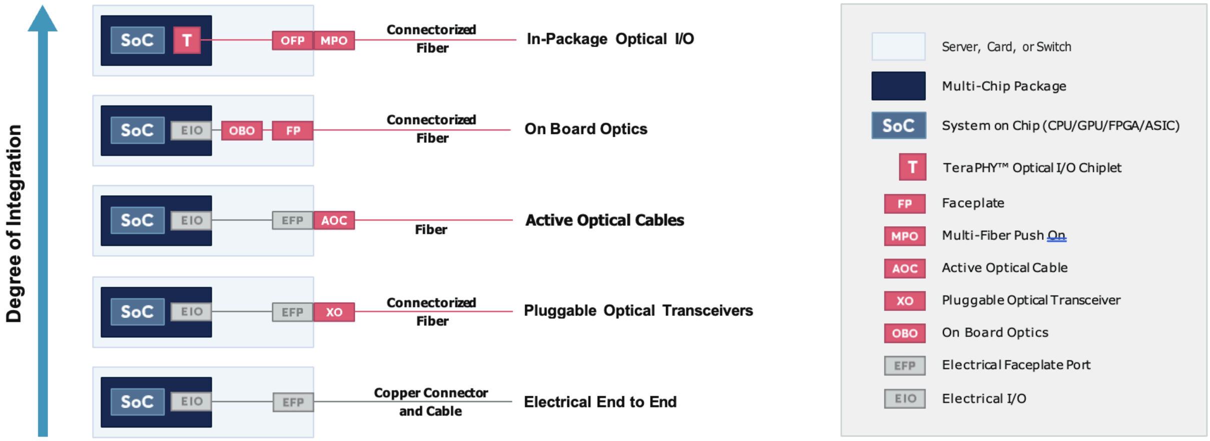


Figure 1 - Industry Approaches to Optical Integration

Barriers to High-Performance Architectures

When designers need to deliver terabit data rates between chips at distances beyond the faceplate, electrical I/O can't deliver the power efficiency, latency and bandwidth density required for the next wave of computing architectures. Even getting to the faceplate electrically and using traditional pluggable solutions involve overcoming significant hurdles with power, latency, and bandwidth density. These old approaches are definitely not scalable as high-performance workloads like Generative AI demand ever increasing throughput density.

Power Efficiency

Power efficiency is a critical limitation when designing electrical systems and data centers since it directly impacts heat and reliability [10]. Today, the power efficiency of long-reach (LR) electrical I/O at 112 Gbps is 6-to-10 pico Joules per bit (pJ/b). Reaching from the package to the edge of a printed circuit board (PCB) is possible at this data rate but draws a lot of power. Reaching from the package to go across systems, racks and datacenters draws significantly more power, requiring a combination of electrical I/O and pluggable optics.

Latency

Latency is another critical design factor, limiting the size and number of components interconnected into a system. On-board and off-board electrical I/O at data rates above 50 Gbps require forward error correction (FEC) coding, which introduces added latency of ~100 ns. Such latency, while tolerated in networking applications, is not tolerated in distributed computing systems (memory semantic fabrics) such as those used for machine learning training, inference, memory-pooling, and other high-performance computing applications.

Bandwidth Density

Today electrical I/O provides bandwidth density around 200 Gbps/mm, supporting 25.6-Tbps (terabits per second) Ethernet switch chips. Next-generation 51.2-Tbps Ethernet switch chips will require bandwidth density around 500 Gbps/mm. Future 102.4-Tbps Ethernet switch chips will require bandwidth density around one Tbps/mm. While the latest generation of 112-Gbps long-reach electrical SerDes solution can deliver bandwidth density at 200-500 Gbps/mm, there is no long-term roadmap for SerDes technology to achieve increasing bandwidth densities. Long-reach electrical SerDes at high data rates also suffer from increased package complexity, cooling requirements, and costs.

Reach

At Tbps throughputs, electrical I/O is suitable for applications within a package but does not scale for connections outside the package due to signal loss, the need for repeaters, and the amount of error correction required. Correcting for these inherent electrical limitations quickly drives power curves beyond sustainable levels.

Ayar Labs In-Package Optical I/O Chiplet

Ayar Labs' solution combines TeraPHY™, an in-package optical I/O chiplet, with SuperNova™, a multi-wavelength optical source, to eliminate I/O bottlenecks, transcend process limitations, and unleash innovative architectures. TeraPHY chiplets disrupt the traditional performance, cost, and efficiency curves of the semiconductor and computing industries by combining silicon photonics with standard CMOS manufacturing processes to deliver up to 1000x bandwidth density improvement at 1/10th the power compared to electrical I/O [12].

Developed in a high-volume GlobalFoundries 45 nanometer process, TeraPHY chiplets integrate millions of transistors with hundreds of photonic devices to drive tens of Tbps of bandwidth up to 2 km out of the package with unmatched power efficiency of less than 5pJ/b. Latency is only 10ns + 5ns/m, point-to-point, with no need for repeaters or FEC, allowing designers to create logically connected, physically distributed compute architectures that scale across racks. TeraPHY delivers bandwidth density more than 200 Gbps/ mm today with a roadmap to more than one Tbps/mm for future generations of high-performance architectures.

Figure 3 shows an opened-lid MCP containing four TeraPHY chiplets integrated into the package with an SoC. MCP technologies (Embedded Interconnect Bridge (EMIB), Silicon-interposer, High-density fanout) provide many advantages over large monolithic dies, including improved yield, support for multiple process nodes in one package, and increased power efficiency.

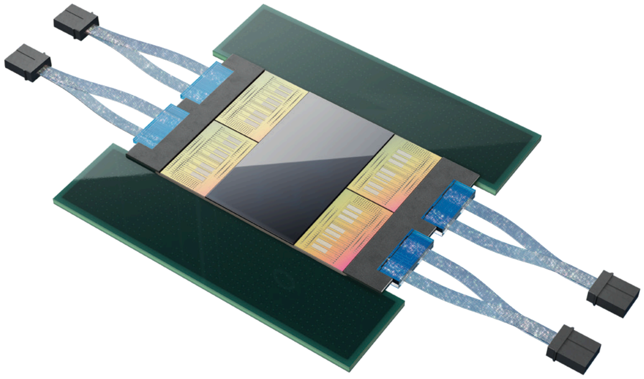


Figure 3 - TeraPHY™ Optical I/O Multi-Chip Package Integration Example

Figure 6 zooms in to show the floorplan for the currently available TeraPHY chiplet. This TeraPHY chiplet supports up to 8 ports, each operating at up to 256 Gbps, enabling four Tbps of bi-directional optical I/O bandwidth per chiplet. Each port connects to three fibers, one for transmit (Tx), one for receive (Rx), and one for the laser light source. The light originates at the Ayar Labs SuperNova laser (multi-wavelength optical source), supplying eight O-band wavelengths to each TeraPHY optical port.

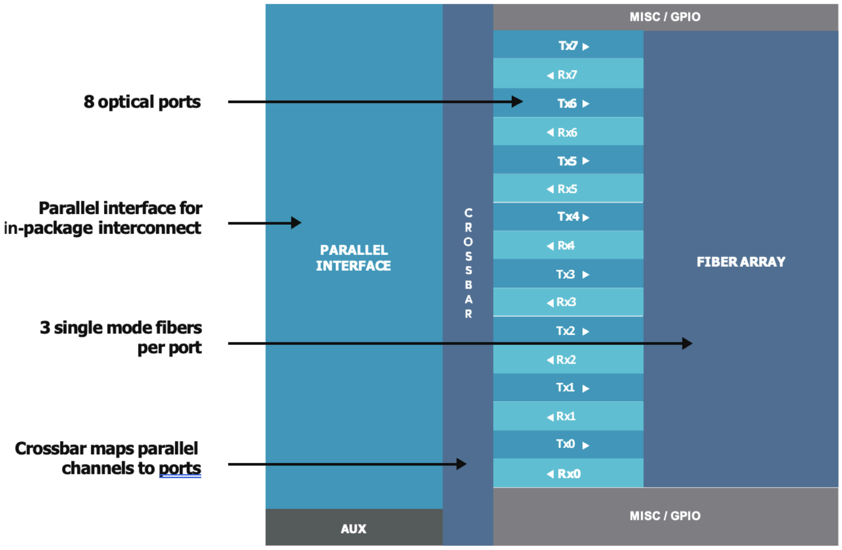


Figure 6 - Ayar Labs TeraPHY™ OIO Chiplet Floor Plan

Port capacity scales with the number of wavelengths per port and data rate per wavelength. Data may be encoded at up to 32 Gbps/λ and scaled up to 64Gbps/λ and beyond. The number of wavelengths may be also scaled up from eight to 16 and more achieving increasing bandwidth densities as shown in Table 3 to satisfy ever increasing needs of scale-out architectures that need resources interconnect via memory-semantic fabrics. Ayar Labs has demonstrated technology components that can pave the way for 10s of Tbps in connectivity per chiplet.

Chiplet bi-directional BW	# of ports / chiplet	# of wavelengths/port	Data rate/wavelength
4.096 Tbps	8	8	32 Gbps
8.192 Tbps	8	16	32 Gbps
16.384 Tbps	8	16	64 Gbps
32.768 Tbps	16	16	64 Gbps

Table 3 - TeraPHY™ Data Rates and Bandwidth Density

Ayar Labs has developed SuperNova, the industry's first multi-wavelength optical source designed to be compliant with the CW-WDM MSA specification released in 2021. The Ayar Labs SuperNova remote light source (Figure 8) can be deployed across a wide range of applications including high-speed I/O, artificial intelligence, optical computing, and high density, co- packaged optics.

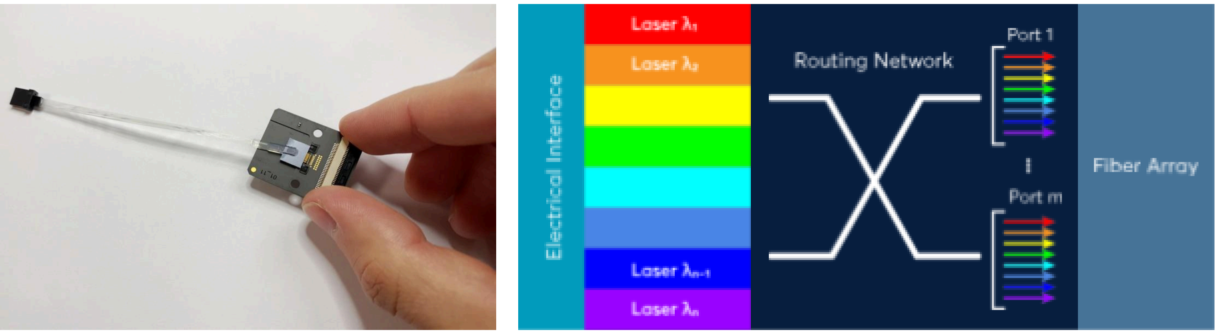


Figure 8 – Ayar Labs SuperNova™ Optical Source

Applications of Optical I/O

By eliminating the constraints imposed by electrical I/O, TeraPHY OIO enables new system architectures for applications across several markets. These markets include HPC and artificial intelligence (AI), cloud computing, telecommunications, and an increasing array of intelligent edge applications.

HPC and AI

HPC and AI are rapidly evolving algorithms that drive the need for greater processing power, high-bandwidth and low-latency access to higher memory capacity, and low-latency between computational elements. By bringing optics to the CPU and memory, OIO is able to dramatically improve performance, while enabling truly disaggregated and distributed architectures that support more flexible systems – all while requiring only a 10th of the power required for copper-based data movement. With the ability to provide high- bandwidth density, new architectures are being explored that enable every element in the system to communicate with low latency, and effectively eliminating many of the traditional impediments to scaling.

Cloud Computing

Cloud applications include rack-scale architectures and resource pooling, scaling the power of systems that were previously constrained within a server to extend across rows of racks in a datacenter. The emergence of OIO presents new opportunities for much more powerful data centers, providing technological avenues that were previously impossible, or cost-prohibitive. With the ability to deliver up to a 1000x improvement in interconnect bandwidth density at one-tenth of the power, Ayar Labs is disrupting the traditional performance, cost, and efficiency curves and lowering the bar to higher performing architectures. OIO is helping to democratize access to more powerful computing, creating opportunities and avenues for innovation in virtually every industry that accesses cloud computing. With OIO, the computing industry is being transformed, providing the performance and scalability needed for advanced applications and creating more opportunity for innovation.

Connectivity

Bandwidth demands are scaling to astronomical levels as mmWave technologies such as 5G (and eventually 6G) put new pressures on telecommunications systems. As these new networking technologies drive faster data rates over existing infrastructures, bottlenecks at interconnection points create network congestion and unacceptable points for failure. Replacing copper cables, OIO technology provides interconnection that enables high-bandwidth and low-power connectivity between antenna/sensing elements and the digital signal processing infrastructure – all within a compact package that can be used in fields where space and weight constraints are critical factors for success. With the ability to deploy new architectures in places where low-latency, high-bandwidth, and low-power capabilities are paramount, OIO technology has the potential to completely revamp previous-generation components with more nimble and capable equipment and architectures. As the edge-computing ecosystem being built around 5G networking evolves and matures, OIO technology will increasingly deliver fast, high-bandwidth connectivity required to process and distribute data at the edge.

Intelligent Edge

The capabilities of OIO, particularly the unique characteristics of the TeraPHY chiplet, which efficiently couples the electrical and optical domains, opens a vast realm of opportunities for intelligent edge use cases, such as those being employed in autonomous vehicles and in aerospace and defense organizations. The capabilities of OIO enable the development of new sensing and computer architectures that disaggregate units within the system – providing a range of benefits for improving communications, strengthening defensive measures, and enabling improvements in real-time decision- making. With capabilities for high-bandwidth, low-latency, and low-power interconnection resistant to electromagnetic interference, OIO is being used to create improvements in systems of all types. Applications include improvements in air traffic control systems, ground-to-air communications with orbital vehicles, phased array radar and communications systems, unmanned aircraft systems, communications with interconnected satellites and more. OIO effectively solves the key challenges of accelerating compute performance while reducing size, weight, and power requirements, making it an ideal choice for the growing number of intelligent edge applications.

Conclusion

Semiconductor designers are creating new, chiplet based architectures that deliver the second phase of Moore's Law to lower costs, improve yields, enhance reliability, and deliver faster time-to-market for next- generation designs. At the same time, the limitations of electrical I/O are driving the search for a new form of universal I/O that quickly delivers terabit data rates from die-to-die and across data centers with less power. TeraPHY OIO chiplets eliminate electrical I/O bottlenecks and transcend process limitations to unleash the next wave of innovation in semiconductor and datacenter design.

Ayar Labs is the first to deliver monolithic in-package optical I/O chiplets, a new universal I/O solution that replaces traditional electrical I/O and enables chips to communicate with each other from millimeters to kilometers, to deliver orders of magnitude improvements in latency, bandwidth density, and power consumption.

Contact us at www.ayarlabs.com to learn more about how TeraPHY OIO can accelerate your design performance today.

10. High-density Optical Interconnect for ML Clusters

High-density Optical Interconnect for ML Clusters

Author :

Peter Winzer, Karen Liu, Nubis Communications

Revised Dec 23, 2023

INTRODUCTION

ML/AI clusters span many racks today but with large differences in interconnection performance between chips, between pods and between racks. This section of the whitepaper describes an application for optical interconnect where extremely dense linear electro-optic components are needed to extend the high bandwidth communications available only within pods today. Erasing the distinction between intra-pod and intra-cluster, essentially flattening the interconnect hierarchy, simplifies the task of the software. This new need can be met by a class of optics called High Density I/O or HDI/O HDI/O which today can achieve the density required to match in-package bandwidth at the perimeter of the pod. The same hardware building blocks can also be used for Linear Pluggable Optics (LPO) in order to streamline development and increase common manufacturing volume. Over time the roadmap extends direct full-bandwidth I/O to more processors including on the inside of the pod array as well. From an industry ecosystem perspective, the same hardware building blocks can also be used for Linear Pluggable Optics (LPO) in order to streamline development and increase common manufacturing volume. Desired roadmap features for Cluster-cluster for inclusion in a future OCP project are discussed.

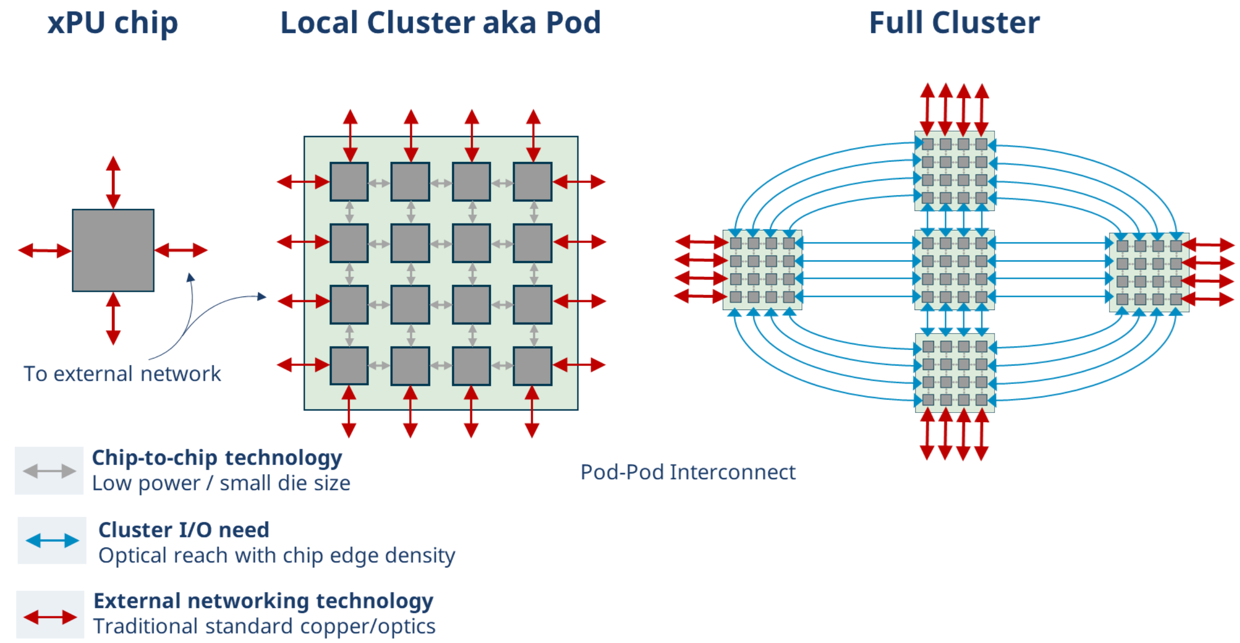

xPU chip
Local Cluster aka Pod
Full Cluster
To external network
Pod-Pod Interconnect
Chip-to-chip technology
Low power / small die size
Cluster I/O need
Optical reach with chip edge density
External networking technology
Traditional standard copper/optics

Figure 1 Classes of I/O needed to support cluster sizes expanding from chip to multi-chip to multi-pod

OPTICS DENSIFIED TO MATCH IN-PACKAGE INTERCONNECT

Pods are delimited today by advanced chip packaging. Figure 1 shows the growth of the "computer" from a single stand-alone chip to a small cluster of chips in-package to the full compute cluster which now spans multiple racks. At all scales, the "computer" entity also talks to the outside world via traditional networking interfaces. But communications within a pod—indeed this is what defines a pod functionally--relies on high bandwidth chip-chip connections using technologies such as parallel BoW/UCIe or serialized USR/XSR. The full cluster by contrast is made of multiple local pods interconnected at much lower bandwidths, whether across a board or across many racks. The available technologies used for this internal connection today were developed for the external networking interface. What is desired instead is something that has as similar as possible the same high bandwidth density and low power consumption as in-package I/O but extended to meters of reach. This goal is described as HDI/O.

Bandwidth density

Table 1 shows that the bandwidths available to the computer architect can fall off dramatically by an order of magnitude off-package.

I/O Type	Full-Duplex I/O Density	Technology
Intra-pod	500 Gbps/mm – 5 Tbps/mm	Die-to-die parallel or serial
Pod-to-pod	40 Gbps/mm – 250 Gbps/mm	Copper cable, some optics
Pod-to-external network	40 Gbps/mm – 250 Gbps/mm	Copper cable, optical module
Pod-to-pod or external network	250 Gbps/mm – multi-Tbps/mm	HDI/O optics

Table 1: Comparison of I/O Densities

Array packaging has overcome the single chip size constraint and stitching together multi-packaging technology is now the size limitation to fully connect clusters at full bandwidth. Chipletized arrays that form local clusters are limited in size by ceramic or high-density build-up substrate size to ~100mm x 100mm; on-wafer chiplet packaging is limited to silicon interposer made from 300-mm silicon wafers. Systems such as Cerebras' Wafer Scale Engine have been limited to the 300mm wafer size for two generations already. Conventional technologies that go off-package today are optimized for networking to the external world rather than efficient communications in-system. These include technologies such as DAC (direct-attach cables) limited by both connector density and reach. Traditional pluggable optics have similar density limits to copper cables and additionally incur too much cost and power.

Benefit of optical reach

The goal of HDI/O is to relieve the complexity of design trade-offs for ML/AI hardware. The driver for optical reach in this application is subtly different from traditional networking where a network hierarchy dictates different reaches by tier. Here the point is reach-independence such that hardware can be flexibly configured. A given HW I/O port could be used to talk to a peer within the same rack, or to a peer in a different rack e.g. for a large torus, or even to a switch for external networking. While the real reach constraint will likely be latency, the point of providing arbitrary reach within a datacenter is simplification, removing the physical media constraint from consideration.

- Reach: < 1 meter – 10 meters is necessary to remove package layout and power/thermal constraints across growing ML/AI clusters.
- Networking compatibility: Extending reach to 500 m allows flexible assignment of HW to external either networking interfaces or inter-pod links depending on specific installation architecture

Optimized Power

Ideally the optics should offer reach independence without creating new complications. This is possible now in an era where the electrical and optical channels both require comparable levels of equalization. At the same time, electrical interface power and real estate are an undesirable tax on the processor chips, so keeping only the minimum required features is essential.

- Latency: avoid additional latency by avoiding additional retimers.
- Power: Minimize power especially but not only locally to the array where delivery is already strained
- Thermal: minimize thermal load locally to the array and make cooling compatible with existing pod cooling.

HDI/O vs LPO Comparison

Linear Pluggable Optics (LPO) has recently gained interest as a way to decrease power consumption cost and latency , relying on the SerDes at the chip edge. While LPO does not require the density of HDI/O it equally benefits from purpose-built equalization in the analog front-end electronics and linearization of the optical components as well as the efficiencies to be gained from a co-design of the two. HDI/O, by moving the optics closer to the xPU, can further reduce the overall digital burden by reducing the length of electrical trace the SerDes must support.

HDI/O operating point optimized for power and LPO optimized for interoperability

The same building blocks can be shared between LPO and HDI/O, the difference being different optimizations for operating points. The ML/AI Array Edge use case differs fundamentally from LPO in its goals that drive differing partitioning of equalization between the electrical and optical components. The priority for ML/AI array edge is total system efficiency: maximizing data bandwidth, minimizing power and real estate on both the host chips and the optics . This is best done by optimizing the end-to-end electro-optical link as a whole. In contrast, the priority for LPO design is to support pluggable optics' separation of the electrical and optical segments with standardized hand-offs.

To illustrate the nature of the difference, Figure 2 shows two deliberately over-simplified partitioning schemes. In the "Clean Up Your Own Mess" scheme, the equalization of the host is primarily responsible for its own electrical channel impairments which are primarily loss and some reflections. Taking this even further, the host Tx needs to take care of most of its own channel, and the host Rx takes care of the far-end channel loss. The LPO module's equalization is responsible for its impairments which include optics-specific nonlinearities as well as bandwidth limitations.   This assignment of responsibility is grossly inefficient in view of the different capabilities of the digital host Tx, analog module and digital host Rx.

Higher performance and lower power can be achieved when the combined electro-optic link and all its equalization elements are allowed to be co-designed. As a general rule, equalization is generally done better on the Rx side than the Tx side as equalization at Tx results in reduced usable signal swing. A second principle is that the analog equalization implementable with low power in the optical module can deal well with host electrical loss, but is not equipped to handle nonlinearity or reflections. Under the "Expertise-based" scheme, the optical module can take care of much of the linear electrical loss. In particular, theTx-side host trace loss is handled better by post-equalization in the analog optical module than pre-distortion at the Serdes transmitter, while in exchange the digital Tx and Rx handle the optics nonlinearities as well as electrical channel reflections.

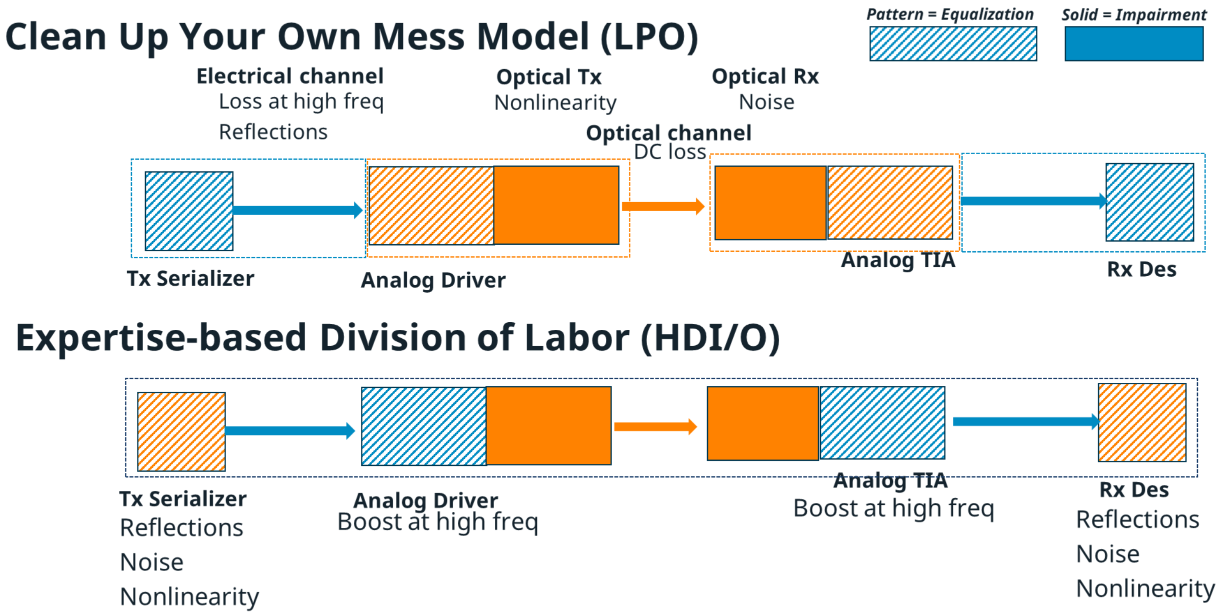


Figure 2 Contrasting models for equalization of multiple types of transmission impairments

Nubis optical engine as an example of HDI/O

The Nubis optical engine shown in Figure 3 is a chiplet with 16x112 Gbps full-duplex that encompasses the full electro-optical conversion function. This multi-purpose engine addresses LPO and retimed modules as well but is distinguished by its low power and ultra-high density that allow it uniquely to address the HDI/O need. Its 5.5-mm x 7-mm footprint includes modulators and photodetectors on the SiPh die with dual octal driver and TIA BiCMOS chips flip-chip mounted on top.

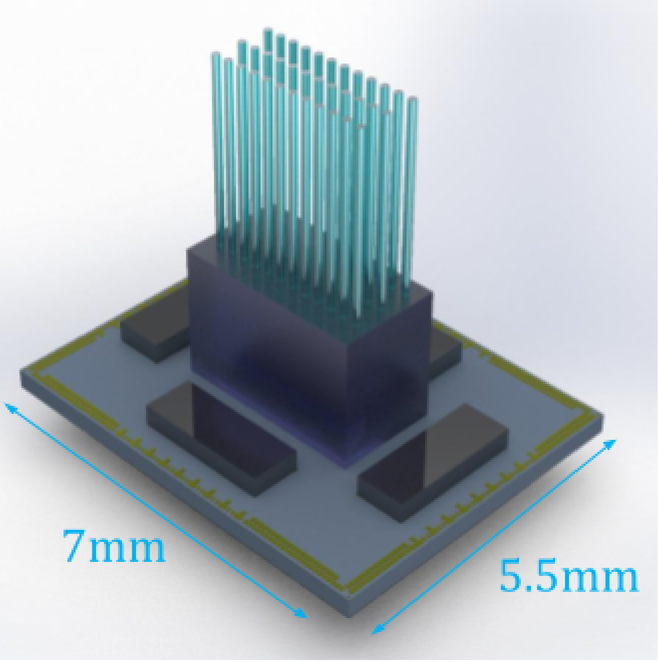


Figure 3 Nubis HDI/O chip supports 16 full-duplex channels up to 112Gbps

The record-setting density of the optical I/O as well as electrical I/O is thanks to a 3D design approach that starts with the fiber escape. An array of grating couplers in the middle of the PIC couple directly into a 3x12 array of standard single-mode fiber. The 2D fiber further extends to functional density by enabling tiling in multiple rows.

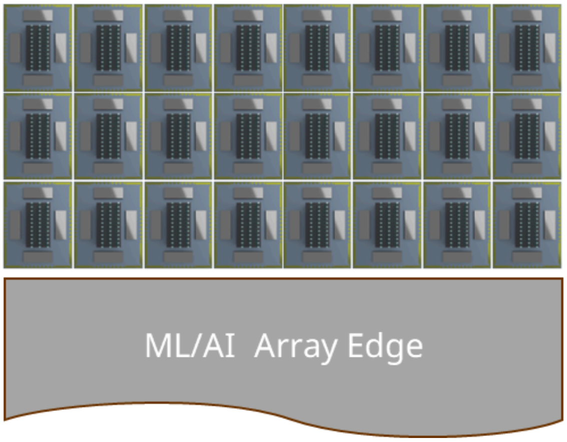


Figure 3 2D tiling allows scaling of Array Edge bandwidth

Roadmap and Potential OCP Project

Roadmap direction

The HDI/O example presented as sampling today achieves the edge density required to interface to the outside chips in a multi-chip package. The largest packages contain more than 2x2 chips, e.g. Tesla's Dojo uses a 5x5 array. The internal chips are still unable to access the rest of the cluster directly. At the same time, chip I/O continues to advance. An HDI/O roadmap study can include both keeping up with the chip advances such as to 200G/lane serialize rates, as well as evaluating radical breakthrough approaches that aim to access the chips on the inside of the in-package array.

Bandwidth

Today's high speed SerDes of 100 Gbps (PAM-4 50 Gbaud) are widely used. 200 Gbps SerDes are expected to sample by the end of 2024 but there are diminishing returns in power per bit expected from the speed increase. Some back-end ML/AI networks, i.e. for long pod-to-pod links, are migrating to 200G faster than the front-end network under pressures to increase total chip I/O bandwidth per chip under pin-limited I/O constraints. The ML/AI pod interconnect is currently limited to well below the per-chip bandwidth. The introduction of HDI/O at 100Gb/s represents a 2x – 9x improvement in throughput; the extension of the same HDI/O architecture 200 Gbps/lane is another 2x improvement.

SerDes can be optimized for hybrid electro-optical links

Today's SerDes are designed either for copper cables or PCB connections to optical retimer chips. With optical links becoming more common, and furthermore with the availability of HDI/O, future SerDes can be optimized for the feature set required to exist in DSP while relying on analog equalization in the linear optics to keep the Serdes as lightweight as possible.

Future standardization of ancillary high-density optical components

The density of HDI/O optical engines has been demonstrated. Associated technologies including reduced size optical connectors and simplified fiber management is needed. HDI/O uses external lasers to displace the footprint, power and heat of these low-speed components from a highly constrained array edge. By definition, this application requires a large number of I/O lanes and hence a large number of lasers. Despite moving the lasers out of the most critical real estate, higher density packaging of the external lasers is still highly desirable.

References

Lauterbach, G. (2021). The Path to Successful Wafer-Scale Integration: The Cerebras Story. *IEEE Micro, 41*(6), 52-57.

Talpes, E. (2023). The Microarchitecture of DOJO, Tesla's Exa-Scale Computer. *IEEE Micro*, 31-39.

Winzer, P. (2023). Ultra-Dense and Ultra-Low Power 16x 112-Gbps Linear-Drive Silicon Photonics for ML/AI. *2023 OCP Global Summit Symposium.* San Jose.

11. TECHNOLOGY FOR SCALABLE AND RELIABLE OPTICAL COMPUTE INTERCONNECTS

TECHNOLOGY FOR SCALABLE AND RELIABLE OPTICAL COMPUTE INTERCONNECTS

Authors:

Alan Liu, Quintessent

Chris Cole, Quintessent

Robert Herrick, Quintessent

Justin Norman, Quintessent

Brian Koch, Quintessent

John Bowers, Quintessent

Release 01 April 2024

Introduction

A new class of power optimized, dense interconnect solutions is needed to scale system level performance of AI clusters and accelerated datacenters. Quintessent is developing highly scalable and highly reliable optical interconnect technology to deliver massively accelerated and power efficient bandwidth across the datacenter. We combine advances in multi-wavelength quantum dot lasers and amplifiers with novel silicon photonic integrated circuits to simultaneously reduce power consumption and required component count over the state of the art while multiplying the achievable bandwidth (density) per fiber. By leveraging multiple innovations spanning materials, device/circuit design, and link architecture, we can enable connectivity solutions that are untethered by the scaling limitations of current technology.

1 Requirements for Optical Compute IO

Optical compute IO will require new technology and architectures capable of matching the progression of bandwidth (density) scaling from compute IO interfaces while minimizing power, latency, fiber count, (photonic) chip size, and cost. In addition, reliability levels comparable to electrical solutions are desirable, which demands significantly more reliable optical components – and in particular lasers – as compared to mainstream optical technology used for Ethernet transceivers today.

DWDM

Scaling bandwidth in the wavelength domain is the most attractive architecture for optical compute interconnects. In this approach, a single wavelength can be encoded with a modest data rate to optimize for power efficiency, while many additional wavelength colors can be deployed in parallel to simultaneously achieve a high aggregate bandwidth while minimizing the number of fibers required via dense wavelength-division multiplexing (DWDM) in a single fiber. Having the flexibility of trading off wavelength count and signaling rate can allow for optical rate matching to the host electrical interface signaling if desirable and, in particular, low power wide-and-parallel interfaces to optimize overall system IO power and latency.

Wavelength Division Multiplexing (WDM) has been used in telecommunication links to increase the capacity of fiber optic cable since the 1980s. Once installed, fiber infrastructure is fixed and expensive, so there is strong motivation to increase the amount of data transmitted over each fiber to reduce the per wavelength cost. For applications where fiber capacity is not an issue, ITU-T Rec. G.694.2 specifies Coarse WDM (CWDM) 20nm spaced grid, with 18 wavelengths in O, C and L band. CWDM enables use of uncooled laser sources over a 70˚C temperature range, which minimizes packaging cost. Data communications is almost entirely in O-band. Mainstream 40 Gb/s and 100 Gb/s links, using 10 Gb/s and 25 Gb/s lane rates, respectively, use four NRZ lanes, either four parallel fiber pairs or the four shortest CWDM wavelengths. For 200 Gb/s links, using 50 Gb/s lane rate, modulation is PAM4 to keep the four-lane paradigm and not increase Baud rate. For 400 and 800 Gb/s links, using 100 and 200 Gb/s lane rates, Baud rates double with each rate increase, as moving to higher order beyond PAM4 is impractical because of steep SNR degradation. Early variants of new rates have used eight lanes, either eight parallel fiber pairs, or eight LAN WDM (LWDM) 800GHz spaced wavelengths. The latter is more difficult to manufacture using the currently dominant free space optics paradigm. PAM4 compared to NRZ for the same data rate reduces the SNR by 3.5dB, which doubles the required optical power. It also requires more sophisticated and power-hungry equalization and associated data conversion. Coherent DSP is now being proposed for the data center but that increases power further. Increasing Baud rate also disproportionately increases power consumption because the copper channel links have a logarithmic attenuation vs. frequency and require significant signal boost to overcome high frequency loss.

Computer I/O applications have wider links than datacenter networks. For example, PCIe is specified up to 16 lanes. Dedicated chip to chip buses, like between CPU and memory, are factors of two wider. As rate increases reduce the practical reach of copper interconnects, optics become the only alternative for computer I/O. To support 8, 16 and 32 lanes, CW-WDM MSA has specified DWDM grid wavelength sets, with 100, 200, 400 and 800 GHz spacing in O-band, suitable for datacom applications. Traditionally DWDM links have required large photonic circuits, tight control of individual laser wavelengths, and higher link loss due to multiplexing and demultiplexing requirements. More recent architectures and technologies such as discussed later in this document eliminate these problems while maintaining the benefits of high lane count.

RELIABILITY

Computer interconnect reliability requirements are typically two orders of magnitude greater than in networking applications. Optical systems using traditional approaches like discrete components cannot meet the stringent requirements of computer I/O. Although all parts of the optical system must be reliable, all optical elements other than the laser can be designed and engineered to avoid reliability problems, while traditional lasers have fundamental limitations that ultimately limit the overall reliability of datacom links today [1].

The current generation of networking equipment utilizes front-panel pluggable fiber optic transceivers to provide connectivity. One reason is to allow easy access to the transceivers so that they can be replaced in minutes. Large cloud network operators report transceiver failure rates of 300-500 FITs (i.e., ~0.3-0.5% annual failure rate). Thus, over a 5-7 year deployment life, nearly half of all 32-transceiver boxes require maintenance. Looking forward, more integrated system designs with co-packaged optics are envisioned for future switches, XPUs, and other applications, with the optical I/O located next to the host compute or switch IC. Some of these integrated designs have optical I/O in a socketed form, where (with much greater expense than front-panel transceiver replacement), the box can be opened, and socketed optical I/O can be replaced. But most have soldered optical I/O that is not considered cost effective to maintain if the optical I/O fails – any channel failing will result in failure of the entire switch system. One reason such systems have not yet been adopted is the mismatch between the demanding reliability requirements compared to what is available with the dominant technology today. While half, or more, of systems today require transceiver replacement in their service life, the target for the next generation of systems is 0.1%-1% failure rate with no maintenance over the expected service lifetime. In many cases, this is for systems with even higher bandwidth, and a higher number of optical links.

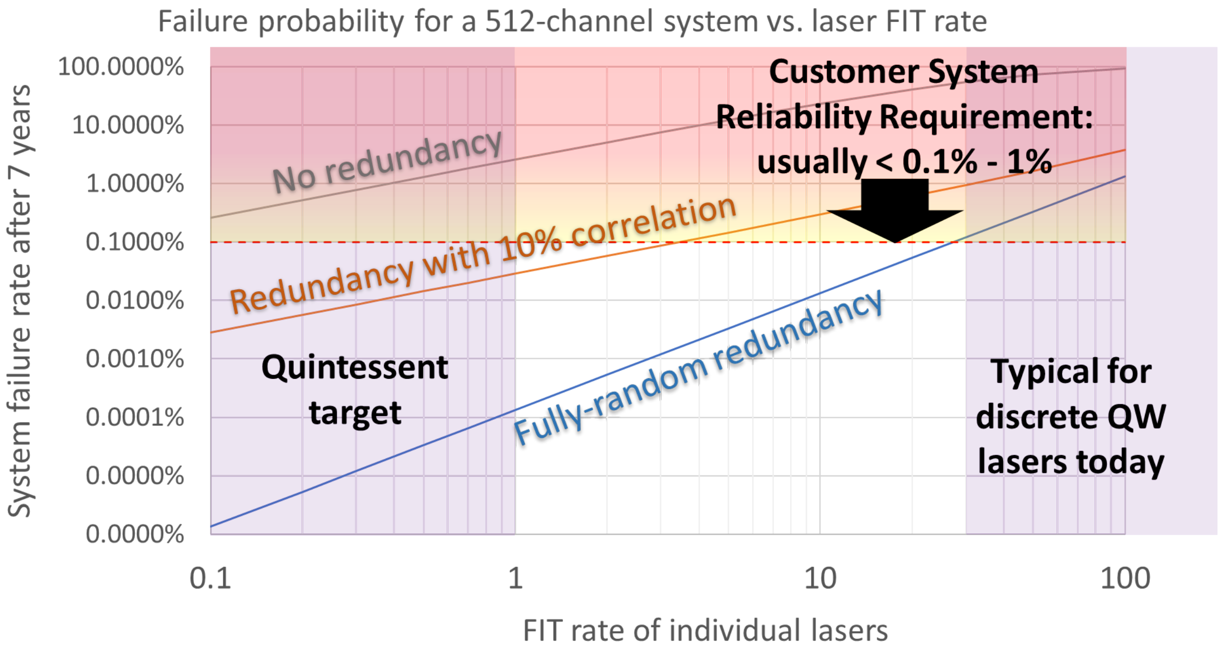
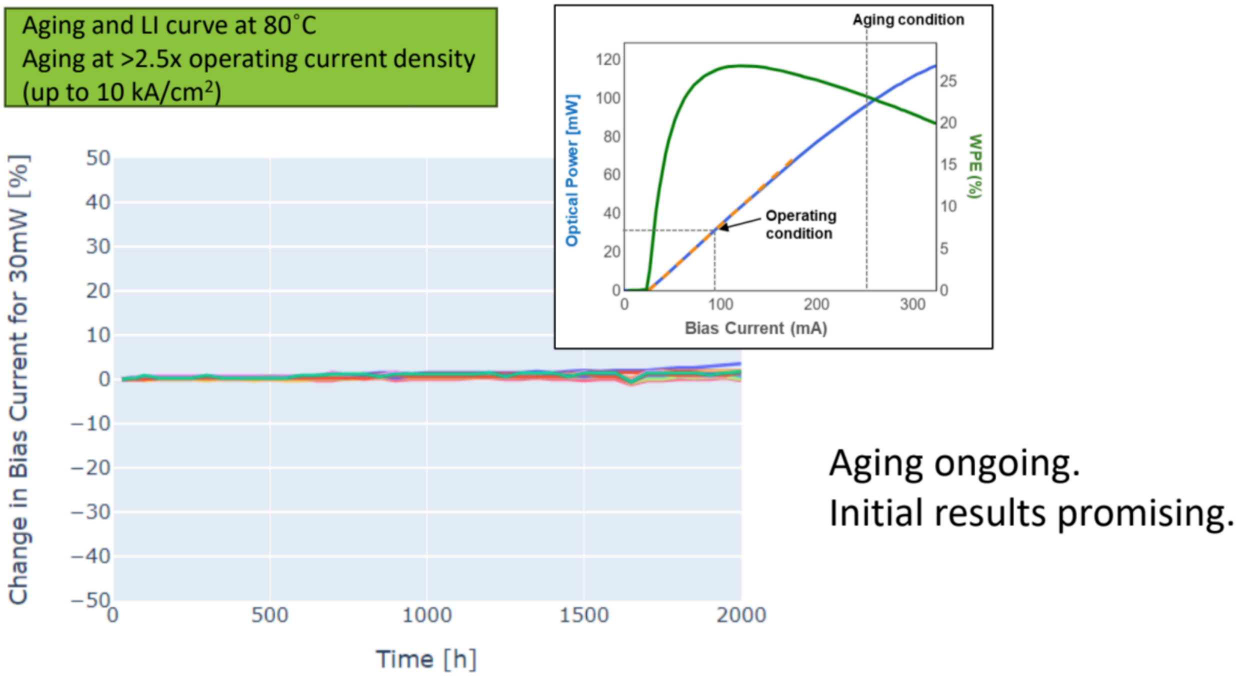


Figure 1: Predicted failure rate after 7 years for a hypothetical switch ASIC system with 512 optical channels of integrated optical I/O (non-serviceable) with different redundancy assumptions.

Closing this gap requires two changes: an improvement in laser reliability and the introduction of redundant backup channels. As shown in Figure 1, either change alone is inadequate to meet the target on a 512-channel system operated over a seven-year deployment lifetime. The implementation of redundancy is relatively straightforward: it requires two or more lasers and an output switch, as previously reported [2]. The difficulty is in how to detect a failure and when and how to switch between lasers. This is driven by individual datacenter operational requirements and is outside the scope of this paper. The calculation of the benefit of redundancy, however, is less straightforward, given the concern as to whether failure is truly random, or whether the issue that caused one laser to fail would also cause its neighboring backup to fail. In most cases, the defects are random, and relatively few types of defects will impact neighboring devices. Here, we design the system with the conservative assumption of 10% correlation of failures.

The improvement in reliability can come from several sources, but it is helpful to start with an explanation of what limits the reliability of lasers used in the majority of low-cost transceivers today: either directly modulated lasers (DMLs) or externally modulated lasers (EMLs, silicon photonics) in the latest high lane rate transceivers. Such emitters typically use AlGaInAs-based multi-quantum-well cleaved ridge lasers. On the positive side, such lasers have excellent high-temperature performance, low power consumption, are relatively inexpensive to produce, and are available from many suppliers. On the negative side, integration usually requires complex optomechanical packaging with many piece parts, and the lasers are vulnerable to dark-line defects, usually originating from cleave flaws, particularly near the ridge corners. In the DML case, directly modulating the lasers at high baud rates usually requires a very high current density >20 kA/cm^2 to get acceptable fall-times on the trailing edge. This high-current density accelerates the degradation rate in lasers with inherent flaws.

One way to reduce failure rates is to use heterogeneously integrated lasers. Such lasers use a taper to gently “push” the light into a silicon waveguide, and thus have no cleaved or powered III-V facets, eliminating a key failure mode that plagues discrete III-V lasers. Since the laser is only providing CW light, with modulation being done elsewhere, the laser can also be larger, and operate at much lower current densities (usually <5kA/cm^2) than a DML. This has proven to give laser failure rates approximately two orders of magnitude lower than DMLs, in deployments of millions of transceivers in large cloud data centers [1, 3]. In addition, in Quintessent’s case, the implementation of strained InAs quantum dots as a gain medium provides an additional layer of protection. The quantum dot material appears to be immune to the dark-line defect growth which dominates random failure in conventional quantum well lasers [4]. It is believed that the irregular strain pattern created by the quantum dots tangles any dislocations that try to propagate and creates a barrier to their movement. Capture of the carriers in isolated dots also reduces the recombination between the dots that is necessary for the dislocations to move.



Aging and LI curve at 80˚C
Aging at >2.5x operating current density (up to 10 kA/cm^2)

Aging ongoing.
Initial results promising.

Figure 2 a) Preliminary aging results are shown for 13 cleaved-facet lasers being aged at an ambient temperature of 80˚C and 250 mA bias current with the relative change in bias needed for 30 mW of output power being plotted vs. time. b) A representative LI curve at 80˚C is shown for one of the devices illustrating the aging condition and the intended operating condition in a system.

Editor's Note: Missing diagram 2a substitution.

Preliminary aging studies on Quintessent cleaved-facet quantum dot lasers are yielding promising results. Figure 2(a) shows results for 13 lasers being aged at an ambient temperature of 80˚C and a constant bias current of 250 mA. On average, we observe <1.3% increase in the required bias current for 30 mW of output power after 2,000 hours of continuous aging. Figure 2(b) shows a representative LI curve (taken at 80˚C) for one of these devices before aging. The aging condition and target operating condition are marked on the plot with the aging bias set at ~2.5× the intended operating point. By combining the reliability advantages of heterogeneous laser integration and those conferred by quantum dot material, Quintessent expects to reach FIT rates below 1, which would otherwise be unachievable for quantum well based lasers of a similar design. These ultra-low FIT rates will enable highly reliable optical compute I/O solutions which can power systems with demanding reliability requirements that are otherwise not possible with today's photonics.

2 Quintessent Technology Enabling Compute IO

PIC BUILDING BLOCKS AND LINK ARCHITECTURES

Wavelengths per fiber	8 λ -32 λ
Signaling rate per λ	25 Gbps – 128 Gbps
Bandwidth per fiber	200G – 4.1 Tbps
Die edge bandwidth density (including laser)	1 – 4 Tbps / mm
Link distance	Up to 2 km
Laser power	0.15-0.5 pJ/bit
Link power	< 5 pJ/bit

Table I: Quintessent's initial DWDM technology roadmap targets.

Quintessent is developing photonic integrated circuit technology to support the range of configurations shown in Table I. The transmit building block proposed by Quintessent consists of a multi-wavelength (quantum dot) comb laser coupled directly to a serial array of ring resonator modulators, while the receive side consists of a serial array of ring filters each coupled to a photodetector. This approach is attractive for its compact size and simplicity, but there are limits to the number of rings that can practically be cascaded serially and thus there are limits to the number of wavelengths that can be combined to a single optical waveguide or fiber. One reason is the small incremental loss incurred when bypassing each ring accumulates to become significant when there are many rings. Depending on link budget, the output power a single comb source can realistically achieve for each wavelength can also become a challenge. Perhaps most importantly, the free spectral range of each ring resonator combined with the wavelength bandwidth requirement for each signal limits the usable bandwidth for a serial ring array. To scale beyond such limitations, we propose using separate groups of DWDM wavelength arrays (with a modest wavelength count) populating multiple CWDM or CWDM-like wavelength bands [5]. Wavelength band widths and guard bands between each wavelength band can be chosen appropriately for a given application. This architecture is uniquely enabled and simplified by multi-wavelength comb lasers, as each DWDM wavelength array within a single band can be generated from a single comb laser with constant and fixed intra-band wavelength spacing, while the envelope of individual comb lasers can drift independently of one another over a given temperature range. DWDM ring Tx/Rx elements can track the wavelengths while CWDM filter elements can pass/accept all wavelengths regardless of their drift. With this approach we anticipate the ability to scale to 4 Tb/s per fiber with 4 wavelength bands at modest signaling rates per wavelength. More bands and/or wavelengths can be added to scale bandwidth without requiring significant technological advances beyond the base photonic building blocks.

QUANTUM DOT GAIN MATERIAL

Quantum dot (QD) gain material is one fundamental technological advancement that Quintessent is leveraging to significantly enhance the capabilities of our photonic platform for optical compute interconnects. Quantum dots possess several benefits over quantum wells, enabling performance and cost improvements for integrated multiwavelength laser sources [6-9]. These technological benefits are derived from the discrete, atom-like density of states and spatial localization inherent to QDs which contrasts with the stair-step density of states and large in-plane diffusion lengths in quantum wells. These fundamental differences in material physics translate into improved device properties including:

- Lower relative intensity noise (RIN) in multiwavelength lasers
- Optical amplifiers with near ideal noise figure
- Improved optical feedback tolerance for isolator-free, on chip lasers
- The ability to operate efficiently at high ambient temperatures
- Improved reliability, as discussed above

Quantum dot lasers enjoy low RIN due to their highly damped relaxation oscillation and, in multiwavelength lasers and amplifiers, low mode partition noise. High damping derives from the long capture/escape times of charge carriers due to the large separation of dot energy levels. Low mode partition noise results from the spatially and energetically isolated nature of an individual QD relative to its neighbors. In-plane diffusion lengths are < 1 µm in QD materials leading to limited gain competition via the carrier reservoir, and differences in size and strain between dots in the same device lead to different emission wavelengths that help reduce coupling via the intracavity photon density.

Quantum dot SOAs work well for multiwavelength systems due to the reduced gain competition between wavelengths described above, and due to their low spontaneous emission levels. Low spontaneous emission is the result of QDs having population inversion factors closer to unity and has led to SOAs with noise figures approaching the ideal value of 3 dB. For example, we have demonstrated a quantum dot SOA with 4 dB noise figure, which is on par with the best SOA results ever demonstrated to our knowledge, while simultaneously achieving 9 dB gain, 15 dBm saturation power, and 20% wall plug efficiency.

The highly damped gain and near zero linewidth enhancement factor of QDs also reduces their sensitivity to external reflections. This effect is significant and may enable isolator-free systems and highly stable on-chip light sources where eliminating reflections can require extensive design and manufacturing efforts. The critical feedback level for coherence collapse scales as, $f_{crit} \propto \gamma^{2}(1 + \alpha_{H}^{2})/\alpha_{H}^{4}$, where γ is the damping coefficient and α_{H} is the linewidth enhancement factor. In QWs, α is generally >1 and often 3-5; whereas α in QDs can be <<1.

Uncooled operation at high temperatures is another attractive feature of QD lasers. When properly designed, the wide energy level spacing relative to kT gives QD lasers unique temperature insensitivity and has enabled the highest maximum operating temperatures of any semiconductor laser. The near zero change in output power from QD lasers over relatively wide temperature ranges can eliminate the need for active cooling, and the ability for lasers to operate up to a 220˚C ambient [9] can enable new integration concepts previously inaccessible to co-packaged optics with integrated (quantum well) light sources.

LASER INTEGRATION

Laser integration with silicon photonics is another technology that Quintessent views as critical for long term scalability of optical compute interconnects. Integrating lasers and silicon photonics in a volume manufacturing foundry presents significant technological and manufacturing challenges, but the benefits of this approach are significant. These benefits include:

- New product architectures and functionalities otherwise not achievable using external lasers, for example complete self-test at the chip level, or on-chip amplification
- Reduced cost
- Reduced power consumption and improved link margin
- Improved reliability, as discussed above

The cost of most high volume, low-cost optical components is driven largely by the packaging costs, i.e. electrically and optically interfacing to the device. This is one of the primary benefits claimed for photonic integration including silicon photonics, which has begun to prove true in recent years. Integrating photonic elements together allows for fewer packages, including less material and fiber coupling costs. This integration also uses fewer optical fibers, reducing total cost and complexity of the fiber infrastructure and reducing the chip cost by reducing area on the chip used for fiber interfacing. The same reasoning applies to integrating lasers onto photonic circuits. Heterogeneously integrated lasers also avoid laser to fiber coupling loss and fiber to Tx chip coupling loss, typically at least 1 dB at each interface in practice, amounting to 2-3 dB of additional link loss. Incurring a 50% power loss has major implications for power dissipation as well as placing more stress and reliability concern on the laser by requiring it to deliver more optical power. Instead, the coupling between III-V material and silicon waveguides in heterogeneously integrated lasers can be well below 0.5 dB.

The same exact quantum dot material, coupling mechanisms, and processing steps used for integrated lasers can also be used to make integrated SOAs. Even modest amounts of gain placed at strategic locations along a link can offer significant advantages. Integrated SOAs can be placed at the output of a Tx chip or input of an Rx chip to compensate for some link losses, and by using a smaller amount of electrical power than would be used to close the link with a higher power laser, this approach can reduce the total power dissipation of the link. This approach may also enable network architectures that require optical links with higher loss, such as optical circuit switching networks.

Figure 3 shows heterogeneously integrated quantum dot lasers on a 200mm SOI wafer. The lasers are made with O-band GaAs quantum dot material and heterogeneously integrated with standard 220 nm thick silicon photonic waveguides. Thus far we have demonstrated integrated quantum dot lasers with threshold currents below 10 mA, output powers near 10 dBm, and operating temperatures up to 100 ° C.

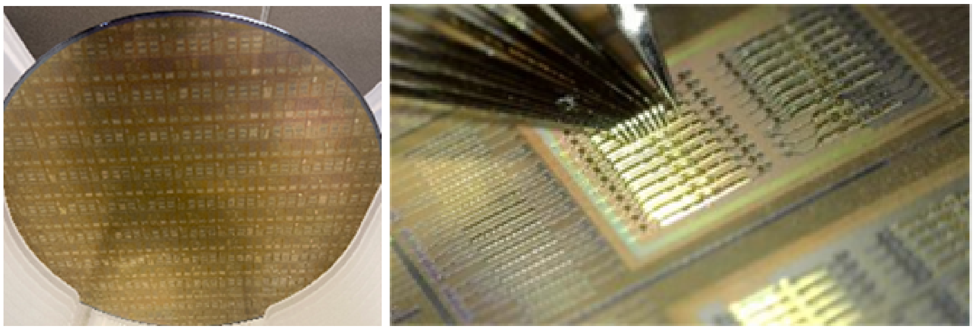


Figure 3. 200 mm silicon photonic wafer containing thousands of lasers and a close-up of several lasers being electrically tested at wafer level. The lasers are hermetically encased within the wafer surface.

MULTI-WAVELENGTH COMB LASERS

In traditional WDM architectures, the incremental cost of scaling bandwidth (by adding more wavelengths of light) can quickly grow to be prohibitive at modest wavelength counts as an additional laser is needed for each new wavelength. Multi-wavelength comb lasers – where a single laser generates multiple wavelengths from the same cavity with all wavelengths emitted into a single optical output port – enables a new scaling curve where the incremental cost of scaling wavelengths (and bandwidth) is negligible and unleashes the full potential of DWDM by providing accessibility to high wavelength count regimes. In contrast to arrays of single wavelength lasers, comb lasers can offer several advantages such as saving chip area (and thus reducing cost and increasing bandwidth density), reducing yield concerns by only requiring one laser instead of many, avoiding the need to combine wavelengths in an additional photonic element, ensuring that all lines always have the desired frequency spacing by design, and avoiding the need to monitor and control the output power and wavelength of each line. Quintessent's approach is to use a (quantum dot) comb laser with only one or two electrically contacted sections in a single compact laser cavity, with a goal of requiring no dynamic control for varying ambient conditions, which enables very simple operation and control. Figure 4 summarizes Quintessent comb laser demonstrations to date. These results are obtained from native GaAs substrate quantum dot lasers, but we anticipate achieving similar performance from lasers integrated on silicon once optimized. Also included in the figure are our goals for 200 GHz spaced comb lasers, which are currently in development.

Frequency spacing	**25 GHz (demonstrated)**	**50 GHz (demonstrated)**	**100 GHz, (demonstrated)**	**200 GHz target (in progress)**
#of comb lines	>100 (within 6 dB)	>32 (within 3 dB)	8 (within 3 dB)	8 (within 1 dB)
Power of lowest line	-5 dBm	~2 dBm	~8 dBm	6 dBm
RIN of worst line	<-129 dBc/Hz	<-133 dBc/Hz	< -147 dBc/Hz	<-145 dBc/Hz
Wall plug efficiency (usable comb power)	>10%	>10%	>25%	20%

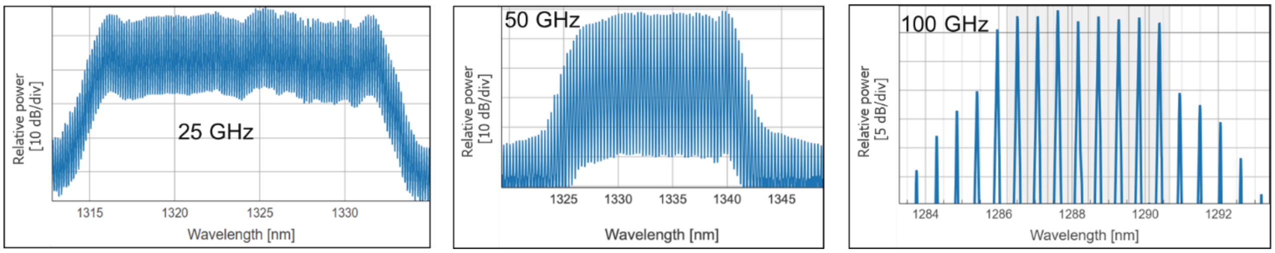


Figure 4. Summary of demonstrated and targeted comb laser performance.

Common concerns for multi-wavelength comb lasers include challenges in achieving high enough optical output power for each comb line and achieving low enough RIN for each comb line. Figure 5 a) shows the spectrum and comb line power levels for an eight wavelength x 100 GHz line spacing comb laser, which emits ~8 dBm or more optical power for each of the eight lines within 3 dB of the peak power. Combs with more wavelengths tend to have lower power per line as shown in Figure 4, but nonetheless even these lower power levels are reasonable and can be sufficient for many links/applications. Figure 5 b) shows the RIN (30 kHz to 25 GHz) measured separately for each comb line emitted by the same 8 wavelength by 100 GHz spacing laser after

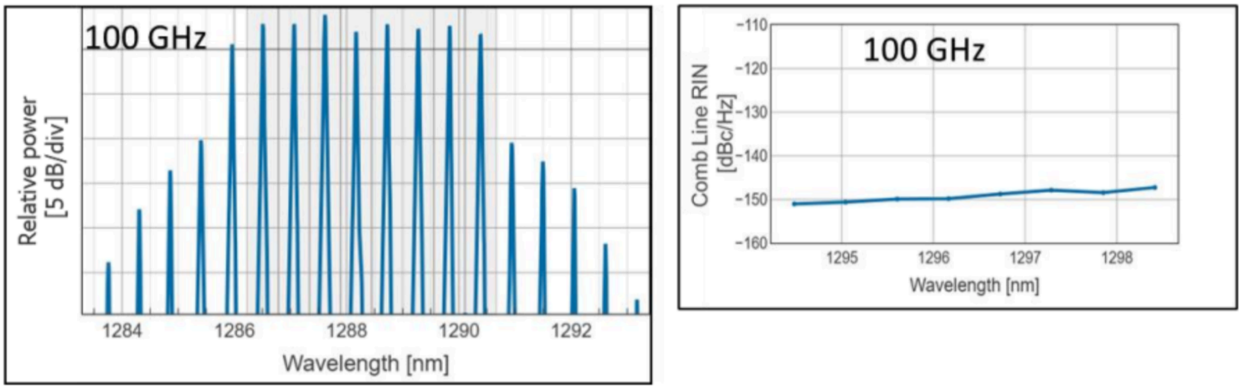
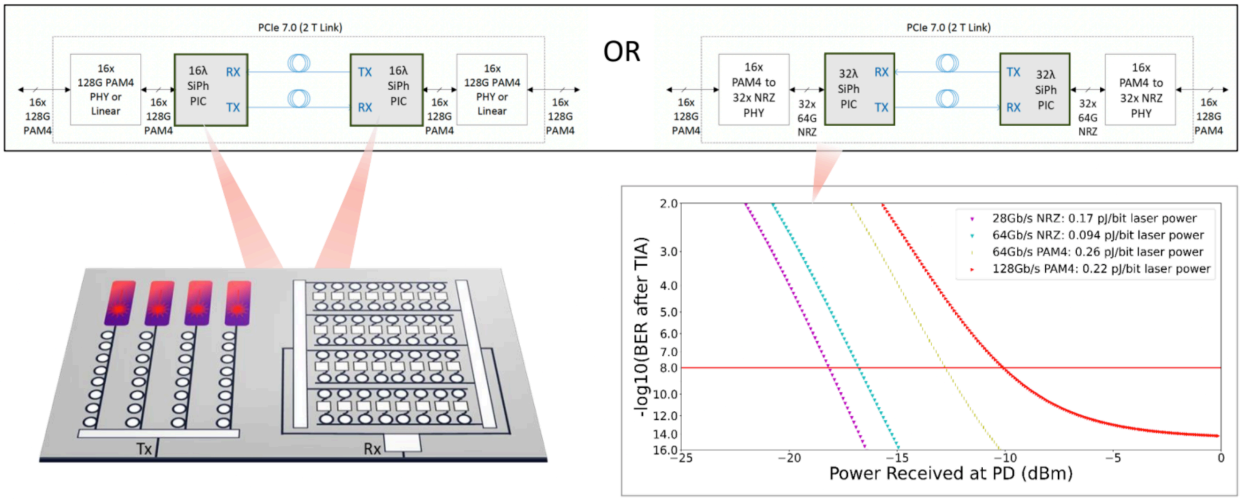


Figure 5. (a) Optical spectrum emitted from an 8 wavelength 100GHz spacing comb laser. (b) RIN of each comb wavelength.

Editor's Note: Initially Missing diagram 5, a substitution from the presentation is included.

module packaging, demonstrating < -147 dBc/Hz RIN for each comb wavelength.



Editor's Note: Missing diagram 6 substitution.

Figure 6: Two configurations for optical PCIe Gen 7 enabled by DWDM technology. Top: 16 lane PCIe interface with each lane mapped to a single DWDM wavelength at 128 Gbps PAM4. Bottom: Scaling the PIC to 32 wavelengths allows each of the 16 PCIe lanes to be split into two NRZ streams and mapped to two separate DWDM wavelengths, each at a more modest 64G NRZ optical rate.

Figure 6 shows how Quintessent technology can enable optical compute IO interfaces, in this case an optical PCIe link targeting speeds expected in PCIe Gen 7. The simple and reliable wavelength scaling of our technology enables flexibility to co-optimize with the host interface depending on system needs and constraints. For example, one can choose to match the optical rate to the host electrical interface to minimize gearboxing or split the incoming electrical data stream into dual NRZ streams to relax bandwidth and power requirements for the optical components. With Quintessent technology, the 16 wavelengths can be supplied by one or two quantum dot comb lasers, as opposed to requiring 16 individual single wavelength lasers. The link itself in our proposed configuration uses a single pair of fibers which benefits escape density, fiber attachment, cable size, and ease of routing. If the transmitter requires an external laser source, then additional (PM) fibers are required to connect the laser module to the transmit modulators. This alternative configuration suffers increased coupling loss on both ends of the connection, requiring increased laser power, reducing reliability, and increasing packaging and cost. A multi-wavelength comb laser integrated with the transmit modulators requires no CW laser fiber connections, simplifies component count and minimizes required laser power, which improves system reliability.

Conclusion

Optical interconnects for emerging advanced compute applications require tremendous scalability in bandwidth per fiber and bandwidth density while improving well beyond today's state of the art in energy efficiency and reliability. Quintessent combines multiple innovations in high performance quantum dot material, compact and efficient comb lasers, dense integration of all DWDM photonic devices into a single silicon chip, and new photonic circuit architectures to enable connectivity solutions that are untethered by the scaling limitations of current technology for optical compute interconnects.

References

1. *Practical Reliability Guidance for PIC Design* (see at 49-51 minutes for discussion of cause of transceiver failures). Accessed: Jun. 16, 2020. [Online Video]. Available: https://www.youtube.com/watch?v=w18Lyiju1v8&feature=youtu.be&t=2739
2. D. Kuchta, "MOTION: Multiwavelength Optical Transceivers Integrated on Node," presented at the ENLITENED Annual Meeting, Long Beach, California: ARPA-E, U.S. Department of Energy, Jul. 2022. [Online]. Available: https://arpa-e.energy.gov/sites/default/files/2022-09/1130_Kuchta.pdf
3. R. Blum, *Integrated silicon photonics for high-volume data center applications*, in Optical Interconnects XX, H. Schröder and R. T. Chen, Eds., San Francisco, United States: SPIE, Feb. 2020, p. 19.
4. J. Selvidge et al, *Reduced dislocation growth leads to long lifetime InAs quantum dot lasers on silicon at high temperatures*, Applied Physics Letters, vol. 118, no. 19, p. 192101, May 2021.

5. Liu, A, *Scaling optical connectivity with DWDM silicon photonics*, OCP Global Summit 2022. https://drive.google.com/file/d/1LOpmKMp9WdBT9M5dZip6ch86dQdZISYR/view (slide 10)
6. Norman, J.C., R.P. Mirin, and J.E. Bowers, *Quantum dot lasers—History and future prospects.* Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films, 2021. **39**(2): p. 020802.
7. Norman, J.C., et al., *A Review of High-Performance Quantum Dot Lasers on Silicon.* IEEE Journal of Quantum Electronics, 2019. **55**(2): p. 1-11.
8. Norman, J.C., et al., *Reliability of lasers on silicon substrates for silicon photonics*, in *Reliability of Semiconductor Lasers and Optoelectronic Devices*. 2021, Woodhead Publishing. p. 239-271.
9. T. Kageyama et al., *Extremely high temperature (220°C) continuous-wave operation of 1300-nm-range quantum-dot lasers,* 2011 Conference on Lasers and Electro-Optics Europe

12. Unleashing the Potential of VCSELs in Co-Packaged Optics for Short-reach Applications

UNLEASHING THE POTENTIAL OF VCSELS IN CO-PACKAGED OPTICS FOR SHORT-REACH APPLICATIONS

Vipul Bhatt, Coherent Corp.

Frank Flens, Coherent Corp.

Computer Interconnect Photonics Workshop

Future Technologies Initiative

OCP Global Summit San Jose, October 18, 2023

Release 01 April 2024

Executive Summary

There is an excellent match between the properties of multimode optical links and the requirements of short-reach applications in data centers, such as HPC networks, AI fabric, and server connectivity for mainstream networks.

Multimode optics enters the CPO arena with a strong track record of success in the front panel pluggable market, in terms of both the unit volumes and robustness of specifications. This track record continues as the IEEE develops specifications for 800G-SR8 with sufficient manufacturing margins. Multimode CPO is on track to offer a low-cost and low-power solution when the market is ready.

IBM and Finisar (now Coherent Corp.) have implemented an ARPA-E sponsored project of building an 800G multimode CPO solution with a power dissipation of less than 4 pJ/bit. Phase 2 will aim for a 3.2T co-packaged transceiver.

Reliability of such CPO products can be enhanced by three orders of magnitude if cold sparing of VCSELs is deployed.

Looking ahead, we see strong potential for further progress in the arrival of 200G VCSELs, multiple wavelengths, and new packaging advancements. The main challenge with higher speed and higher channel density will be to maintain a very low bit error rate (BER) floor for those applications that must use lightweight FEC.

We think the OCP community can complement the work of OIF by going beyond the framework document and fostering a vibrant CPO ecosystem, including multimode CPO.

Introduction

There is an excellent match between the properties of multimode optical links and the requirements of short-reach applications in data centers. In this paper, we describe that match, present some data from our implementation experience with multimode co-packaged optics (CPO), and conclude with some thoughts on future directions.

Short-Reach Application Areas

We focus here on three application areas of interest:

1. High-performance computing (HPC) – a network of supercomputing systems that process data and perform complex operations at high speeds.

2. Connectivity for Artificial Intelligence (AI) systems, such as machine learning networks that perform training and inference. Connectivity for such systems can be seen as both the back-end fabric as well as the front-end connections via Ethernet or Infiniband.

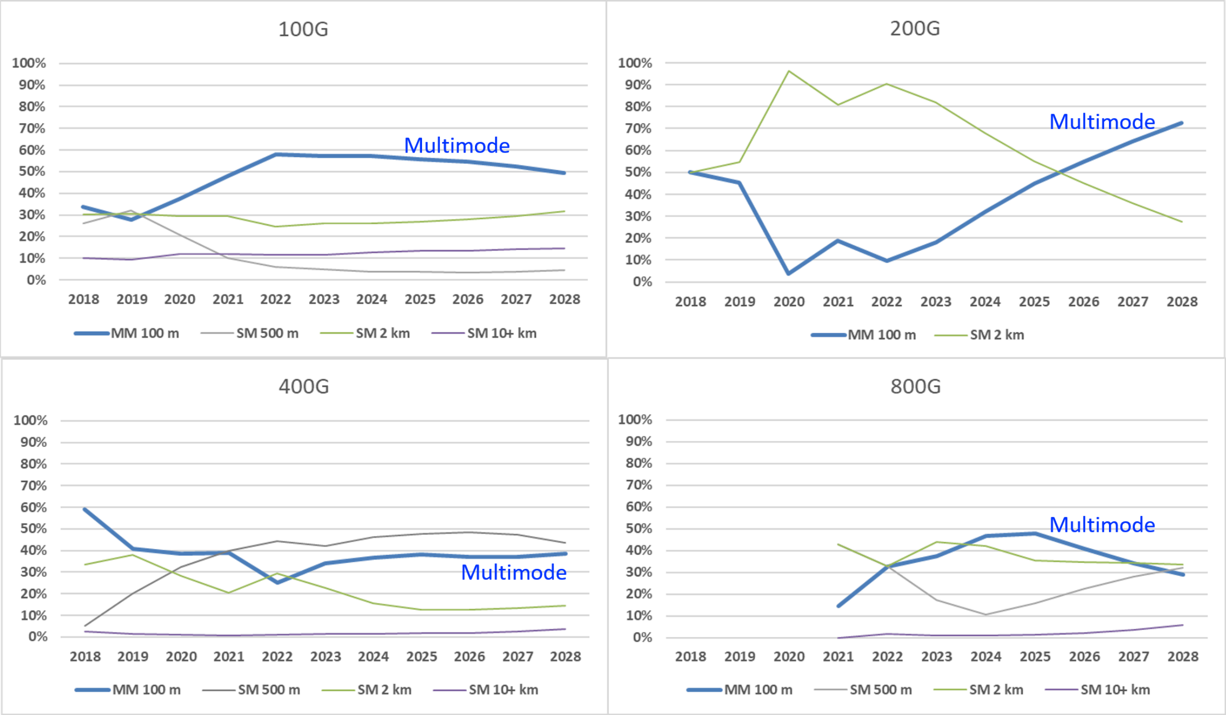
3. The first two tiers of traditional networking – server to Top of Rack (ToR) or Leaf switches, and in an increasing number of short-span data centers, Leaf to Spine switches as well.

The common theme among these three areas is that the link length is less than 100 meters. By specifically targeting such short-reach applications, we can leverage a technology that is best suited for them – and in our opinion, that technology is the multimode link technology.

Multimode Pluggable Transceivers: A Track Record of Success

Multimode transceivers have an impressive track record of success in data centers. The figure below shows the percentages of the number of units of various types of transceivers, ranging from 100G to 800G. It shows not just the historic track record, but also projected volume over the next five years. Clearly, multimode transceivers have held a significant share, and are projected to continue to do so.

Figure 1 Multimode transceivers have been very successful in data centers and are expected to continue their dominance. Source: LightCounting



Specifications and Performance of 800G Multimode Optical Link

For the three short-reach applications mentioned above, an example of a highly suitable optical transceiver – whether in a front panel pluggable form or co-packaged – is the 800G-SR8. Each transceiver unit has an information carrying capacity of 800 Gbps, although a co-packaged tile or chiplet may contain multiple 800G transceivers. It uses parallel fibers, which is an excellent design decision for short reaches, because it reduces the transceiver cost, the dominant cost element for short reaches. It supports a reach of 100 meters on OM4 fiber.

The IEEE P802.3df task force is currently developing specifications for 800G-SR8.[1] We list below the draft specifications for transmitter and receiver, along with an illustrative link power budget. Notice the conservative nature of these specs – this is essential for commercial success.

Table 1 Transmitter Specifications, 800G-SR8

Description	Value
Signaling rate, each lane	53.125 GBaud +/- 50 ppm
Modulation format	PAM4
Center wavelength (range)	844 to 863 nm
RMS spectral width (max)	0.6 nm
Average launch power, each lane (max)	4 dBm
Average launch power, each lane (min)	-4.6 dBm
Outer OMA, each lane (max)	3.5 dBm
Outer OMA, each lane (min) For max (TECQ, TDECQ) ≤ 1.8 dB For 1.8 < max (TECQ, TDECQ) ≤ 4.4 dB	 -2.6 dBm -4.4 + max (TECQ, TDECQ) dBm
TDECQ, each lane (max)	4.4 dB
TECQ, each lane (max)	4.4 dB
Overshoot / undershoot (max)	29%
Transmitter power excursion, each lane (max)	2.3 dBm
Extinction ratio, each lane (min)	2.5 dB
Transmitter transition time, each lane (max)	17 ps
Average launch power of OFF transmitter, each lane (max)	-30 dBm
$\text{RIN}_{14}\text{OMA}$ (max)	-132 dB/Hz
Optical return loss tolerance (max)	14 dB
Encircled flux	≥ 86% at 19 µm ≤ 30% at 4.5 µm

Table 2 Receiver Specifications, 800G-SR8

Description	Value
Signaling rate, each lane	53.125 GBaud +/- 50 ppm
Modulation format	PAM4
Center wavelength (range)	842 to 948 nm
Damage threshold (min)	5 dBm
Average receive power, each lane (max)	4 dBm
Average receive power, each lane (min)	-6.4 dBm
Receive power, each lane (OMAouter) (max)	3.5 dBm
Receiver reflectance (max)	-15 dB
Receiver sensitivity ($\text{OMA}_{\text{outer}}$) (max) For TECQ ≤ 1.8 dB For 1.8 < TECQ ≤ 4.4 dB	-4.6 dBm -6.4 + TECQ
Stressed receiver sensitivity (OMAouter) (max)	-2.0 dBm

Table 3 Illustrative Link Power Budget, 800G-SR8, OM4 Fiber

Description	Value
Signaling rate, each lane	53.125 GBaud +/- 50 ppm
Effective modal bandwidth at 850 nm	4700 MHz.km
Power budget (for max TDECQ)	6.4 dB
Operating distance	0.5 to 100 meters
Channel insertion loss (fiber, splices, connectors)	1.8 dB
Allocation for penalties (for max TDECQ)	4.6 dB

In what way has the conservative nature of multimode link specs by IEEE contributed to commercial success of products?

Notice how the transmitter power output is specified for a flexible range of eye closure. This allows designers of lower-noise or higher-bandwidth components to dial down the optical modulation amplitude to -2.6 dBm instead of some value closer to 0 dBm. That reduces power consumption and improves manufacturing yields.

Similarly, at the receiver end, if a receiver's photodiode and TIA combination offers good sensitivity, the manufacturer can use the available margin to improve manufacturing yields. Receiver sensitivity range of -4.6 dBm to -2.0 dBm is comfortably achievable over a practical range of eye closure conditions.[2]

All of these "robust by design" features of a multimode optical link carry over to co-packaged implementations. We illustrate this with two examples in the next section.

Benefits of Using VCSELs and Multimode Fibers in Co-Packaged Optics Applications

Table 4 Mapping the features of multimode optics to the benefits for co-packaged optics

Feature	Benefit for co-packaged optics
On-wafer testing, high yields	Reduced cost
Small die size, compact arrays	Smaller transceiver footprint around ASIC
Small and circular output beam	Higher coupling efficiency, lower link loss
Larger alignment tolerances	High assembly yields for transceivers
Higher temperature range of operation	Greater flexibility in cooling system
Ease of adding spare VCSELs	Significant improvement in reliability
Low-loss fiber connections	Greater link margin, higher yields

Let's take connector loss as the first example to illustrate how comfortably a multimode link can adapt to a co-packaged optics implementation and still maintain its key strengths of low cost and low power. Let's consider a hypothetical scenario where we implement 800G-SR8 in co-packaged optics form.

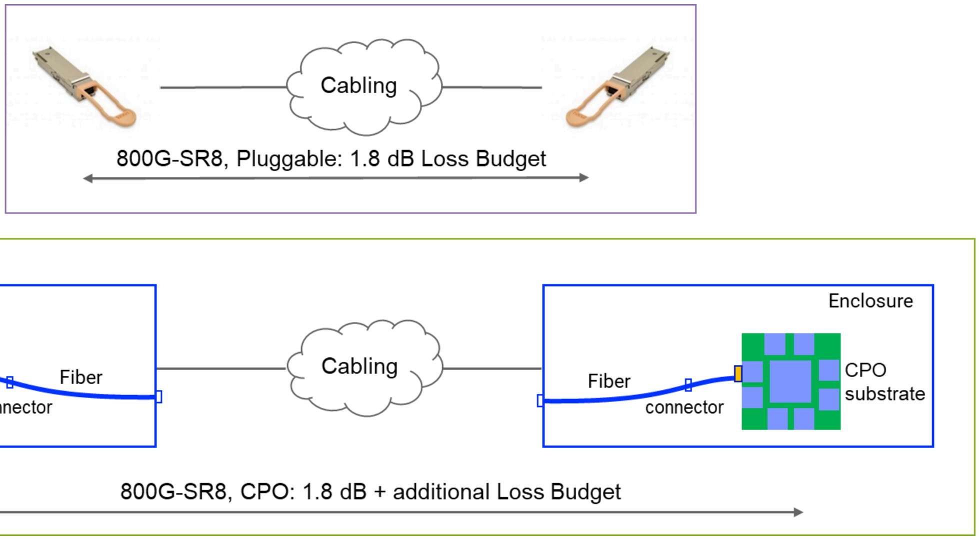
The first step is to decide the location of compliance test points. In data centers, CPO implementations will coexist and interoperate with front panel pluggable implementations, operating over the same cable plant. For optical transceivers, the transmitter test point is TP2, defined by the IEEE as the end of a 2-meter cable attached to a transmitter. For typical front panel pluggable transceivers, the cable simply clicks into its LC receptacle. Practical constraints suggest that an equivalent scenario for CPO is a 2-meter cable attached to the enclosure (the system box), thus adding at least one extra connection, as shown in Figure 2. Inside the box, there will likely be at least one additional loss mechanism to account for – typically to join the pigtails epoxy-bonded to a transceiver chiplet (tile) with the rest of the fiber harness, or to create a strain-relief mechanism, or to allow tight fiber bends, or some other such arrangement. A similar configuration will repeat at the receiver end.

Table 5 IEC 61753-1 Connector Loss Specifications

Grade	Mean Loss	97% Loss
D (single-mode)	≤ 0.5 dB	≤ 1.0 dB
B (multimode)	≤ 0.3 dB	≤ 0.6 dB

Therefore, a CPO implementation must bear an additional burden equivalent of one or two fiber-to-fiber connections at the transmitter end and receiver end. Referring to Table 5, if we model it as four extra connections per link, this can make a difference of up to 1.6 dB per link for the 97% case when compared to a single-mode CPO implementation (4*(1 – 0.6)). A multimode link can be up to 1.6 dB better off than a single-mode link for accommodating a CPO implementation. Other reasons for considering this additional loss are described in the Co-Packaging Framework Document published by the Optical Internetworking Forum.[3]

Figure 2 Additional loss budget of a CPO implementation is easier to manage with multimode fiber connections



Given that the economic success of these products relies on reuse of the component ecosystem of Ethernet-compliant front panel pluggable transceiver modules, this seemingly small loss budget difference can make a big impact on manufacturing yields and, therefore, on cost.

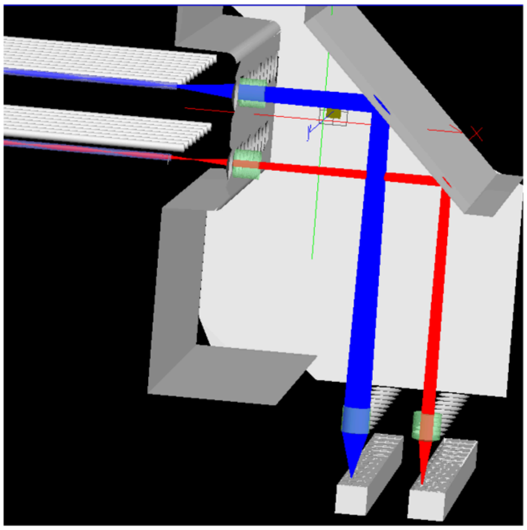


Figure SEQ Figure * ARABIC 3 A cross-section of a VCSEL-based Transmitter Optical Subassembly

In our second example, this theme of larger tolerances leading to high efficiency and high yields repeats for the transmitter assembly as well. The low threshold current and the narrow circular output beam of a VCSEL, combined with the large core of the fiber, enable a very efficient launch of an optical signal into the fiber. Figure 3 illustrates a cross-section of a VCSEL-based 12-fiber Transmitter Optical Subassembly (TOSA). Notice the utter simplicity of it – light comes out of the VCSEL, makes a 90-degree turn, and focuses the beam on the fiber. The injection-molded block is made of low-cost polyetherimide, a high-strength plastic material. The alignment tolerance on this TOSA is greater than +/- 10 microns, and the loss from VCSEL to fiber is less than 2 dB. Contrast this with a single-mode silicon photonics arrangement, where the tolerance is an order of magnitude smaller, and the loss from CW laser output to fiber launch is an order of magnitude higher.

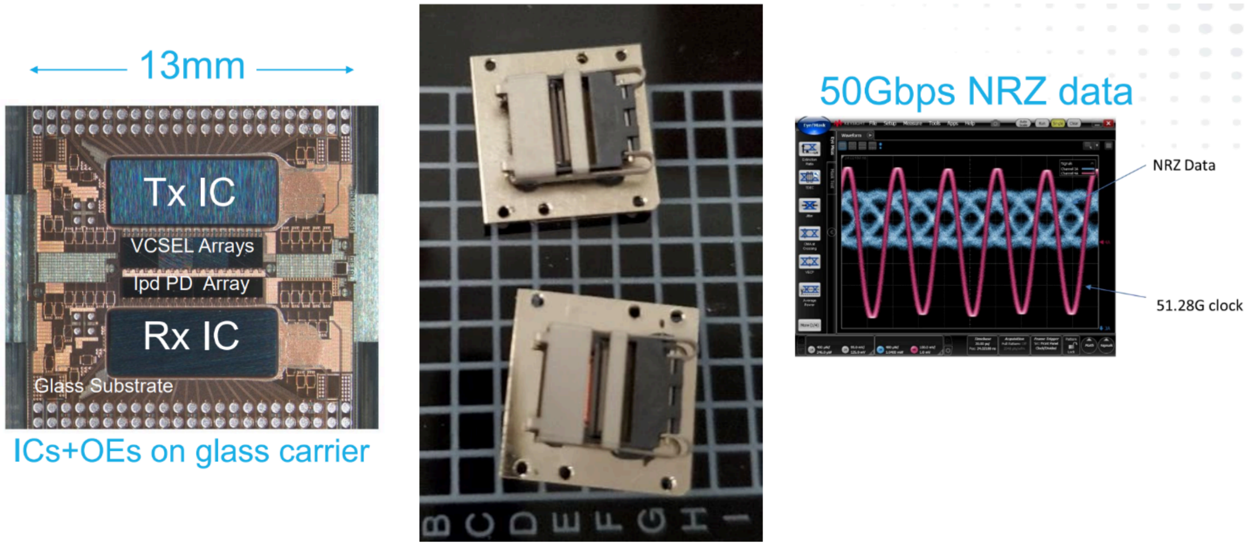
Multimode CPO Implementation Experience

In this section, we briefly report on our experience of implementing a multimode CPO solution. It is a co-packaging project named MOTION (Multi-wavelength Optical Transceivers Integrated on Node), sponsored by the U.S. Advanced Research Projects Agency-Energy (ARPA-E) and collaboratively executed by IBM and Finisar (now Coherent).[4]

Phase 1 of this project is now complete. Each module carries greater than 800G (16 x 50G NRZ) with a pre-FEC bit error ratio (BER) of less than 1e-12. Its electrical interface has 6 dB budget, and optical link supports 2 dB channel insertion loss, more than sufficient for the targeted 30 meters of reach with connectors. With a power dissipation of less than 4 pJ/bit (3.2 watts for 800G), and a footprint of just 13 mm x 13 mm, it is a demonstrated proof of the ability of multimode optics to enable compact, low-power, co-packaged optical transceivers. Figure 4 shows pictures of the completed MOTION modules. It supports both soldered and socketed operation. This solution was demonstrated at OFC'23.

The goal of Phase 2 of this project, now in progress, is to increase the capacity of a module to 3.2T (16 fibers, 2 wavelengths per fiber, 112G PAM4 per wavelength).

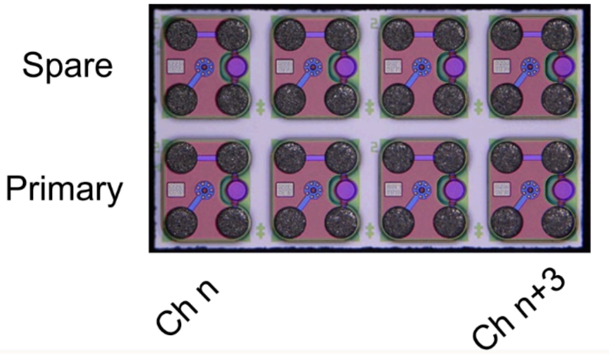
Figure SEQ Figure * ARABIC 4 MOTION co-packaged transceiver modules, 800G each



Enhancing Reliability with 2-to-1 Sparing

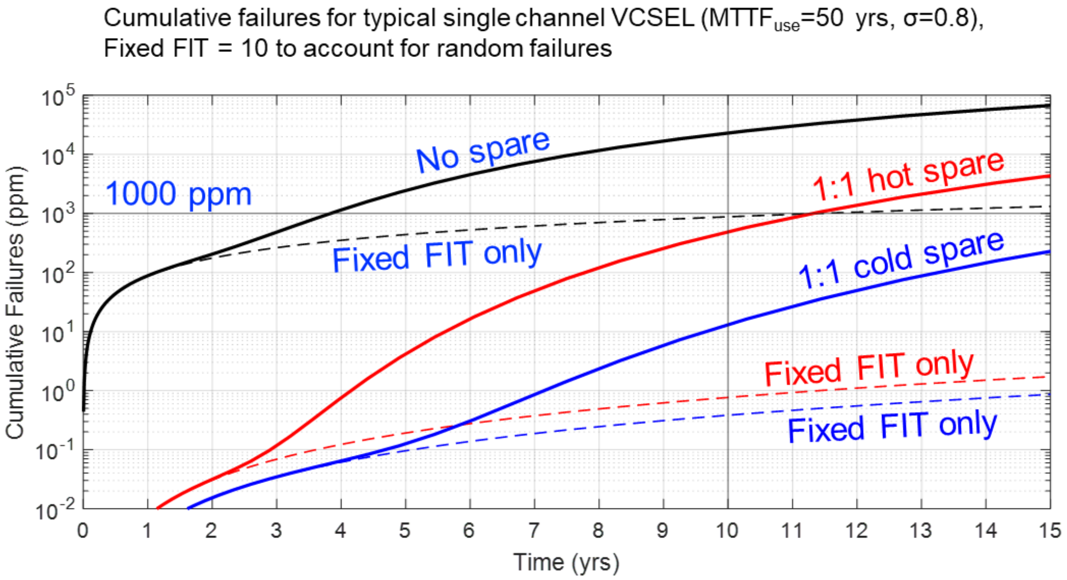
The small size and low cost of VCSELs suggest a simple and powerful way to enhance the reliability of multimode CPO systems – the use of 2-to-1 sparing.

Figure SEQ Figure * ARABIC 5 VCSEL array with spare units



In this approach, we speak of hot sparing and cold sparing. In hot sparing, lasers degrade simultaneously, and the transceiver unit fails only if both independent components fail. In cold sparing, the spare component is called into service only when the primary component fails. Both sparing methods are helpful, but cold sparing improves reliability more, as shown in Figure 6.

Figure SEQ Figure * ARABIC 6 With cold sparing, failure rate can be reduced by a factor of 1,000



Looking Ahead: Opportunities and Challenges

Opportunities

Increasing Transceiver Capacity

To keep up with future growth, we will need to ensure that a multimode CPO transceiver chiplet (tile) can scale to 6.4T and 12.8T. If 100G PAM4 signaling is worth preserving as the modulation option with optimum complexity, one easy path forward is to increase the number of wavelengths (see Figure 7). Four to eight wavelengths can be accommodated in the range of 840 nm to 980 nm. The SWDM Alliance may be able to play a helpful role here since they have already defined the wavelength plan and other specifications that can be easily adopted for co-packaged optics.[5]

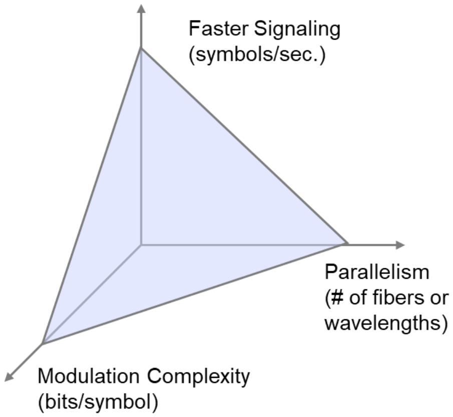
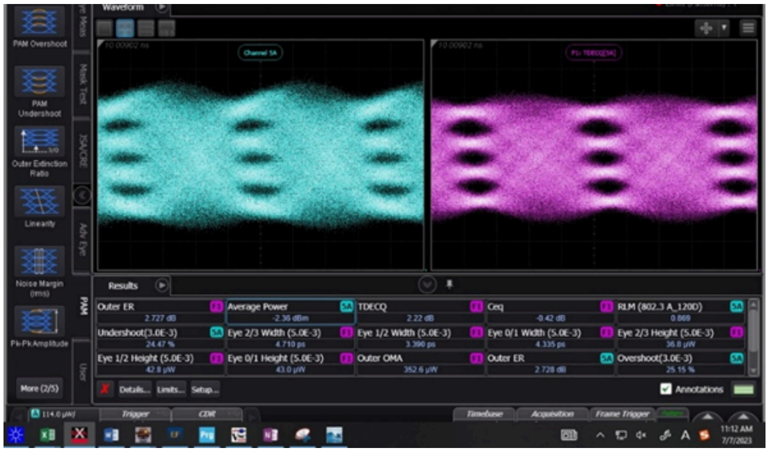


Figure SEQ Figure * ARABIC 7 The three axes of increasing information carrying capacity of an optical transceiver

Linear Electrical Interface

With improvements in digital signal processing (DSP), new generations of system chips are equipped with stronger equalization, signal shaping, and nonlinearity compensation capabilities. At the same time, some system vendors appear to be more open to specifying and managing the end-to-end link performance that need not meet the IEEE 802.3ck specifications at both ends of a link. This creates an opportunity to deploy un-retimed optical transceivers over a linear electrical interface. Recent results for 800G-SR8 have shown promising results. An example is shown in Figure 8. In a limited set of scenarios where linear pluggable optics (LPO) is technically and operationally feasible, there is an opportunity to further save power and cost.

Figure SEQ Figure * ARABIC 8 Linear interface 800G-SR8 transmitter output eye diagram, TDECQ 2.69 dB



Outer ER F1 2.727 dB	Average Power 5A -2.36 dBm	TDECQ F1 2.22 dB	Ceq F1 -0.42 dB	RLM (802.3 A_120D) 5A 0.869
Undershoot(3.0E-3) 5A 24.47 %	Eye 2/3 Width (5.0E-3) F1 4.710 ps	Eye 1/2 Width (5.0E-3) F1 3.390 ps	Eye 0/1 Width (5.0E-3) F1 4.335 ps	Eye 2/3 Height (5.0E-3) F1 36.8 µW
Eye 1/2 Height (5.0E-3) F1 42.8 µW	Eye 0/1 Height (5.0E-3) F1 43.0 µW	Outer OMA F1 352.6 µW	Outer ER 5A 2.728 dB	Overshoot(3.0E-3) 5A 25.15 %

Packaging Advancements

Performance of future high-speed CPO links will be determined by advances in packaging. Fortunately, some innovative approaches are being investigated. Here are two examples.

One approach envisages a glass substrate solution where both optical and electrical connectivity can be fabricated using ion-exchange waveguides embedded into the glass and low-loss electrical routing. Through-glass vias provide both power delivery for the mounted components and data management connectivity. If successful, this approach has the potential to simplify or even eliminate the fiber harness assembly.[6]

Another approach looks beyond current packaging and focuses on using interposers and 2.5D assembly. By making dense wiring available, this approach can help scale the CPO systems to a higher capacity, provided that the slow-and-wide approach is accepted in the market. Temperature within this aggressive thermal environment will need to be suitably regulated.[7]

200G VCSEL

There is significant commercial product development activity under way to enable VCSELs to support 200 Gbps. As 200G electrical interfaces and Serializer-Deserializer (SerDes) speeds arrive in the market, 200G VCSELs will offer transceiver designers a way to avoid using “reverse gearboxes” to convert electrical signals to optical.

A helpful development here may be to recognize that multimode fibers likely have a higher bandwidth than is currently assumed. The current assumption about the fiber's effective modal bandwidth (EMB) is based on outdated profiles of VCSEL output modes. Modern VCSELs can support a set of mode weighting functions that lead to a higher EMB spec.

Further improvements for 200G VCSEL links may involve more specialized (yet inexpensive) optics that can tailor the modal structure of the input to the fiber to best match the performance of the installed fiber.

Challenges

Achieving a Low Floor for Pre-FEC BER.

Some short-reach applications may benefit from lower overall latency. Once we are in the region of less than 30 meters link reach, the link latency is limited by the encoding and decoding times of forward error correction (FEC). A lower floor of pre-FEC BER permits the use of a lightweight FEC or no FEC at all, thus reducing FEC-processing latency. Effectively, this boils down to improving the signal-to-noise ratio (SNR) of an optical link. This is not an easy task as speeds go up.

Potential Role of the OCP Community

We believe the OCP community can play a very helpful role by complementing the work of the Optical Internetworking Forum (OIF), which has developed a co-packaging framework document. Possible avenues available to the OCP community include raising awareness of the growing ecosystem of users and suppliers and holding technology symposiums to exchange ideas about use cases and innovative new solutions.

Concluding Remarks

Considering the relative ease with which its form factor can transition from front panel pluggable to CPO, a multimode transceiver is an ideal candidate for use in short-reach applications. The industry's vast experience with it, having shipped hundreds of millions of units over two decades, is an asset we cannot afford to ignore. Since three major applications – HPC, AI, and server connectivity – are decidedly short-reach applications, we can expect multimode links to continue to thrive, including in co-packaged form where feasible and appropriate.

References

“Adopted IEEE P802.3df Objectives,” IEEE P802.3df 400 Gb/s and 800 Gb/s Ethernet Task Force, https://www.ieee802.org/3/df/proj_doc/objectives_P802d3df_221117.pdf

"800G (2×400 SR4) OSFP Optical Transceiver," product summary. https://ii-vi.com/product/800g-2x400-sr4-osfp-optical-transceiver/

Jackson, Kenneth, et al. “Co-Packaging Framework Document,” Optical Internetworking Forum, 23 Feb. 2022, www.oiforum.com/wp-content/uploads/OIF-Co-Packaging-FD-01.0.pdf

Kuchta, Daniel. “Co-Packaging on Organic Laminates: MOTION Phase 2.” ARPA-E ENLITENED Kickoff Meeting, 13 Jan. 2021, http://arpa-e.energy.gov/sites/default/files/2021-01/DAY1_Kuchta_ENLITENED%20Kickoff.pdf

Lewis, David. “100G SWDM4 MSA Technical Specifications.” The SWDM Alliance, 6 Nov. 2017, https://swdm.org/wp-content/uploads/2017/11/100G-SWDM4-MSA-Technical-Spec-1-0-1.pdf

L. Yeary, L. Brusberg, C. Kim, S.-H. Seok, J. Noh, and A. Rozenvax, "Co-packaged Optics on Glass Substrates for 102.4 Tb/s Data Center Switches," *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)*, Orlando, FL, USA, 2023, pp. 224-227, doi: 10.1109/ECTC51909.2023.00046.

B. G. Lee, N. Nedovic, T. H. Greer, and C. T. Gray, "Beyond CPO: A Motivation and Approach for Bringing Optics Onto the Silicon Interposer," in *Journal of Lightwave Technology*, vol. 41, no. 4, pp. 1152-1162, 15 Feb.15, 2023, doi: 10.1109/JLT.2022.3219379.

13. High Density, Ultra-Low Power microLED based Optical Interconnects for Chip-to-Chip Communications

High Density, Ultra-Low Power microLED based Optical Interconnects for Chip-to-Chip Communications

Authors: Chris Pfistner, Bardia Pezeshki, Rob Kalman

Company: AvicenaTech Corp.

Release 01 April 2024

Executive Summary

Avicena is pioneering a revolution in data transmission with high-density, ultra-low power, short-reach optical interconnects. Unleashing the untapped potential of Gallium Nitride (GaN) microLEDs, Avicena's LightBundle™ provides a solution better suited for short-reach interconnects than traditional, energy intensive SerDes based technology. Connecting microLED arrays with silicon photodetectors through cost-effective fiber bundles, Avicena's interconnects offer unparalleled energy efficiency at <1pJ/bit, and exceptional bandwidth density starting at 1Tbps/mm with a roadmap to 10Tbps/mm and beyond. Furthermore, GaN microLED transmitters can sustain high-temperature operation beyond 125°C. Drawing on the existing IC and display manufacturing ecosystems, Avicena has the capability for a swift volume ramp, with affordable production, transforming the landscape of chip-to-chip communications with a brighter future of faster, more energy-efficient computing.

Introduction

In the relentless pursuit of higher performance and efficiency in modern High-Performance Computing (HPC) systems, the demand for interconnect solutions that can effectively bridge the ever-widening gap in processor-to-processor and processor-to-memory communications has reached unprecedented levels. As electrical interconnects face fundamental limits in reach and power efficiency at higher data rates the HPC and IC industry has been exploring different optical link technologies. The diagram in Figure 1 below shows a figure of merit for bandwidth density and energy efficiency (Y Axis) for different transmission technologies against their respective link reach (X Axis).

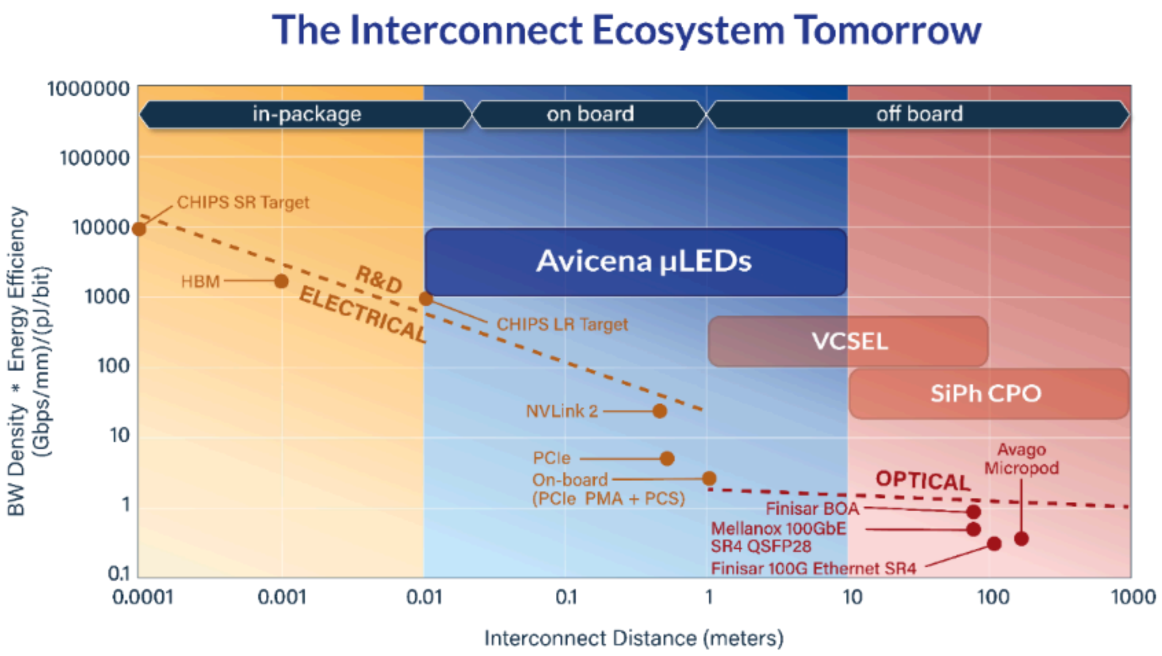


Figure 1. This graph shows a figure of merit of bandwidth density and energy efficiency vs reach for different transmission technologies. For short reach up to a few centimeters electrical links offer the best performance. For long distance communication conventional optical technologies offer the best performance with multi-mode VCSEL links reaching up to 100m, and single-mode SiPh, DML and EML interconnects covering distances of many kilometers depending on configuration. For intermediate reach up to 10m Avicena's microLED based interconnects offer the best performance (shown as blue area in the center).

Multi-mode Vertical-Cavity Surface-Emitting Lasers (VCSEL) links are used in a wide range of datacom applications today, but they are not well suited for linking hot running processor ICs due to their limited tolerance for high temperature operation [1]. Silicon Photonics (SiPh) links represent another optical link technology under evaluation. However, originally developed for medium to long-haul applications SiPh interconnects show limited energy efficiency on the order of 5pJ/bit for short reach applications [2,3]. Moreover, both VCSEL and SiPh based links typically use Serializers/De-serializers (SerDes) to achieve high aggregate link bandwidth of > 100Gbps per lane which introduces latency, increases power consumption, and often requires Forward Error Correction (FEC) to achieve the required Bit Error Rates (BER). Because of all the limitations of traditional SerDes based optical interconnects the HPC and IC industry keeps evaluating innovative solutions to high bandwidth density, high energy efficiency and low latency for short to intermediate reach of up to a few meters.

This white paper delves into the revolutionary realm of GaN microLED-based optical links, offering a transformative approach to chip-to-chip interconnects that promises unparalleled levels of ultra-low power consumption, exceptional bandwidth density, and minimal latency, in stark contrast to existing SerDes-based solutions.

MicroLED technology, renowned for its application in high-resolution displays and lighting systems, has demonstrated the potential to redefine the landscape of data communication at the chip level. By combining the intrinsic advantages of light as a medium for data transmission with the 2D layout of LED arrays, microLED-based optical links present a groundbreaking avenue for achieving previously unattainable levels of performance in interconnect architectures. This white paper covers the fundamental principles, design considerations, benefits, and potential challenges of employing microLED-based optical links for parallel chip-to-chip communication, heralding a new era of interconnected computing systems.

GaN microLEDs have been used in free-space Visible Light Communications (VLC) at limited data rates [4]. Using our patented technology, we have demonstrated GaN microLEDs running at data rates up to 14Gb/s [5]; Fig. 2 shows a 10Gb/s eye diagram from a microLED. Despite being under development for just the last few years, multiple parallel optical links using GaN microLEDs have already demonstrated bandwidth density of > 1Tbps/mm2 and power < 1pJ/bit [6,7].

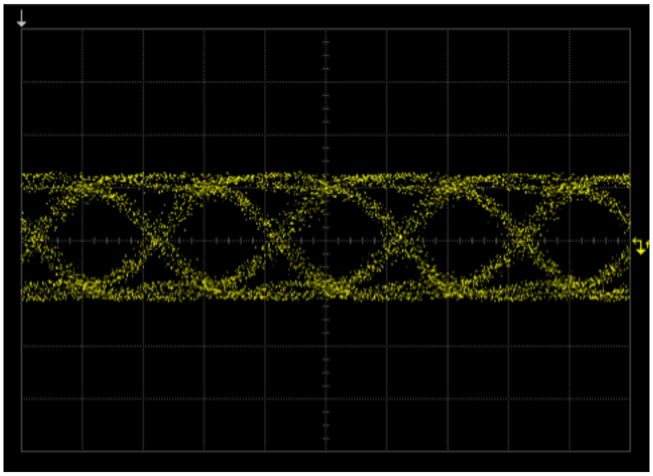


Figure 2. 10Gbps eye from GaN mircoLED transmitter

The architecture of LightBundle^{TM} micro LED interconnects

The basic LightBundle^{TM} interconnect architecture is shown in Fig. 3. An array of GaN microLEDs is bonded to a transceiver ASIC and connected via a fiber bundle to a matching array of Si PDs on another transceiver ASIC. A typical LightBundle link has a few hundred channels each operating at a few Gb/s, providing aggregate throughputs of > 1Tbps per link. The modest per-channel speeds are well-matched to typical IC clock speeds and obviate the need for SerDes operating at tens of Gb/s. Arrays of these microLEDs can be transferred to silicon ICs with very high yields, for instance using a laser lift-off (LLO) process [8]. This process is used for displays with > 100,000 pixels so optical interconnects with a few hundred channels can be fabricated with very high yield.

LightBundle uses an array of GaN microLEDs emitting blue light centered near 425nm with a spectral width of ~10nm. These microLEDs have our patented epitaxial structure optimized for speed at the cost of somewhat reduced quantum efficiency (QE). Despite this reduced QE, links based on microLEDs still achieve energy efficiencies of < 1pJ/bit.

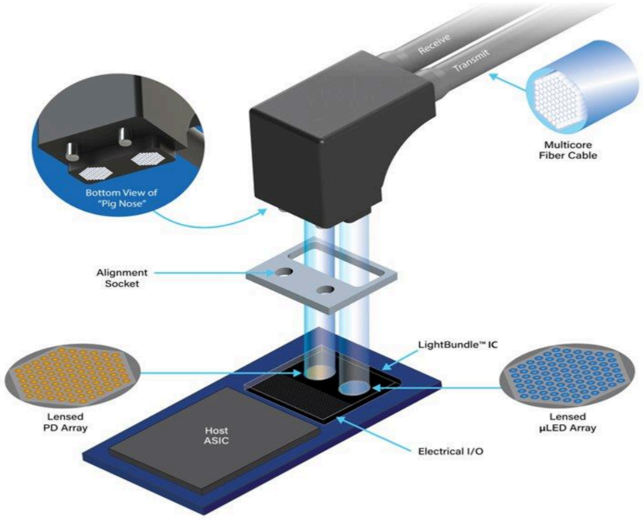


Figure 3. The architecture of the Avicena LightBundle^{TM} microLED based interconnects

The microLED array is optically connected to an array of silicon photodetectors (PDs) using a fiber bundle. Typical microLED and PD arrays have a few hundred elements on a hexagonal close-packed (HCP) or square grid with a grid pitch in the 50µm range. The fiber bundle consists of multimode fibers, each with a diameter matching the microLED/PD array grid pitch (typically ~ 50µm) such that each microLED is coupled to a PD through a single fiber. The large core diameter enables greatly relaxed packaging tolerances and low packaging cost relative to single-mode fiber packaging. Reach is typically limited by modal and chromatic dispersion to ~ 10m.

A great benefit of utilizing blue light emitters is the very high absorption of Si at this wavelength. The ~0.2µm absorption depth of Si at this wavelength allows the use of simple lateral PD structures that are compatible with standard CMOS processes. These PDs can have very low capacitance per unit area, enabling simple receivers with excellent sensitivity and very low power consumption. By contrast, longer wavelengths require the use of vertical PD structures that are not compatible with CMOS and with much higher capacitance per unit area.

Table 1. below shows a high level comparison of the key link parameters of a microLED vs a SiPh interconnect.

Table 1. Key Performance Parameters: LightBundle microLED vs SiPh Links

Parameters	LightBundle^{TM} microLED Link	Typical Silicon Photonic Link
Bandwidth Density	1Tbps/mm - 10Tbps/mm	~ 0.25Tbps/mm - 7.2Tbps/mm
Energy Efficiency	< 1pJ/bit	~ 5pJ/bit
Latency	2ns + Time of Flight	~ 5ns + Time of Flight
Operating Temperature	-55 to 125°C	0 to 70°C
Reach	10m	500m - 2000m

LightBundle^{TM} Demo ASIC

A great benefit of the LightBundle technology is that it is not tightly tied to a specific IC process. We are currently demonstrating a bidirectional 304-channel LightBundle^{TM} interconnect using a custom ASIC fabricated in TSMC's 16nm finFET process (Fig. 4). Each channel operates at 4Gb/s, giving a total throughput of ~ 1.2Tbps in each direction. The ASIC contains optical transmitter (Tx) and receiver (Rx) arrays. It also includes a pulse pattern generator (PPG) and error checker, an Rx open eye monitor (OEM), and various loopback paths. This allows the IC to be operated without external high-speed data connections for a stand-alone testbed, and to execute various built-in self-test functions.

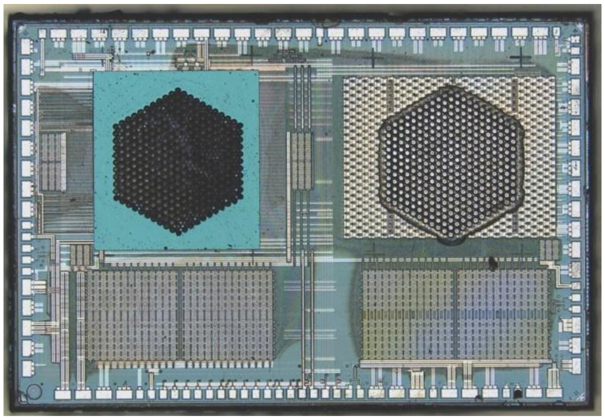


Figure 4. Integrated 1Tbps transceiver chip in 16nm CMOS. The hexagon on the right shows the microLED transmitter array and the one on the left the receiver PD array.

The optical transmitter array has 331 channels. Each channel has an LED driver with programmable bias + drive current. Each LED driver is connected to a pad to which the anode of a microLED is bonded. Another metal connection links each microLED's cathode to a corresponding pad on the ASIC. Each microLED is ~8µm in diameter and located on a 50um pitch hexagonal close-packed (HCP) grid (Fig. 5). At a 4Gb/s per-channel data rate, resulting in a 2D bandwidth density of 1.75Tb/s/mm^2. A polymer micro-lens array was fabricated on top of the LED array to optimize coupling of each microLED into the corresponding fiber.

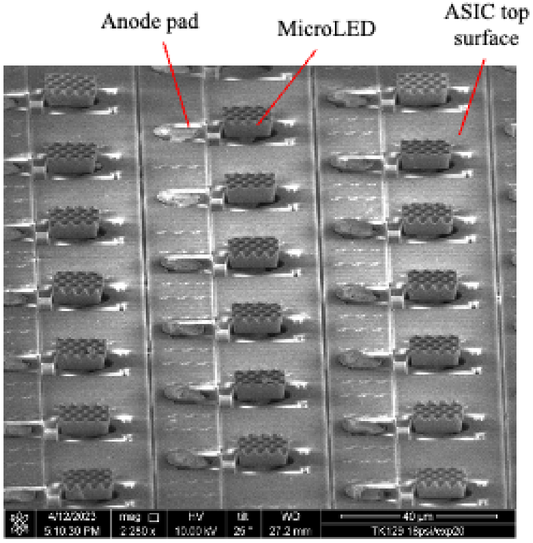


Figure 5. Integrated 1Tbps transceiver ship in 16nm CMOS.

On the receiver side there is a matching array of 331 PDs fabricated in Silicon. The PDs are located on the same 50µm-spaced HCP grid as the microLEDs. Like the LED array, a polymer micro-lens array was fabricated and attached to the PD array to optimize coupling from each fiber to its associated PD. Each of the 331 PDs is reverse biased and connected to one input of a differential TIA.

The Tx array was butt-coupled to a fiber bundle containing 331 fibers on a 50µm HCP grid, matching the Tx array grid geometry. The other end of the fiber bundle was butt-coupled to an Rx array that has the same grid geometry as the Tx array. As noted previously, the relatively large diameter of each fiber in the bundle gives alignment tolerances on the order of ±5µm.

Target power consumption for the LightBundle link is ~ 1pJ/bit at 4Gb/s, including both the optical transmitter and receiver (with associated Tx and Rx electronics). Iterative optimizations of the microLEDs, PDs, and micro-lenses are ongoing, and we expect to achieve link power consumption significantly below 1pJ/bit.

Applications for LightBundle^{TM} Interconnects

The bandwidth density requirements in HPC and Cloud Computing have increased steadily over the past years. Now the Large Language Models (LLM) in Generative AI are driving the need for energy efficient high-bandwidth interconnects to unprecedented levels. Figure 6 below illustrates how the release of ChatGPT-4 in late 2022 started a new era when the number of parameters in the underlying model expanded beyond the one trillion mark.

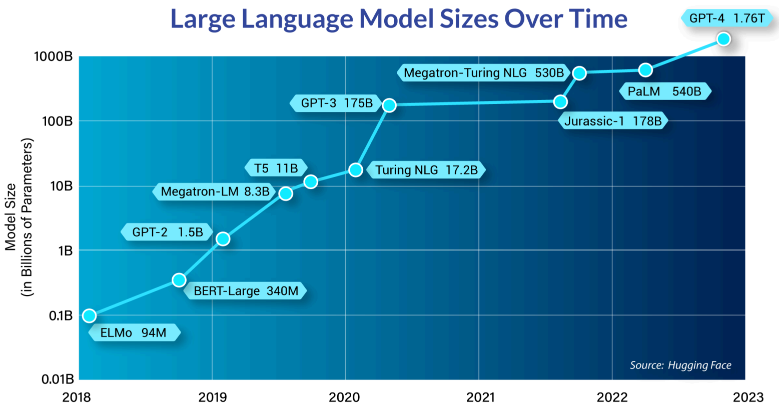


Figure 6. Number of parameters in different Large Language Models (LLM) (Source: Hugging Face)

As illustrated in Figure 1. electrical links will keep dominating short reach interconnects and conventional optical technologies like VCSELs, Silicon Photonics and pluggable transceivers using DML and EML will cover the long reaches. However, for intermediate reach of up to 10m microLED interconnects offer a clear advantage in terms of ultra-low energy of < 1pJ/bit and high bandwidth density of multiple Tbps/mm. It is precisely for these intermediate reaches where the AI/ML computing clusters show the highest demand for efficient high-bandwidth interconnects. Moreover, the parallel nature of the LightBundle architecture with hundreds of lanes running at relatively low data rates of a few Gigabits makes these interconnects an excellent match for chip-to-chip communication since internally processors run wide and slow buses. The key applications that stand out for immediate applications are:

- High Bandwidth Memory (HBM) interconnects
- Parallel Chiplet Interfaces: e.g., UCIe, BoW, OpenHBI, AIB

High Bandwidth Memory - HBM

The current generation of HBM is referred to as HBM3 and features a parallel interface of 1024 lanes running at 6.4Gbps each for an aggregate data rate of ~ 6.5Tbps or ~ 0.8TBps. Today these HBM3 modules are co-located with the logic die like a GPU or CPU on a silicon interposer because the electrical interface between the logic die and the HBM3 module can only support a reach of ~ 2mm which limits architectural design options. Figure 7 shows an NVIDIA Hopper H100 GPU with six co-located HBM modules in one package.

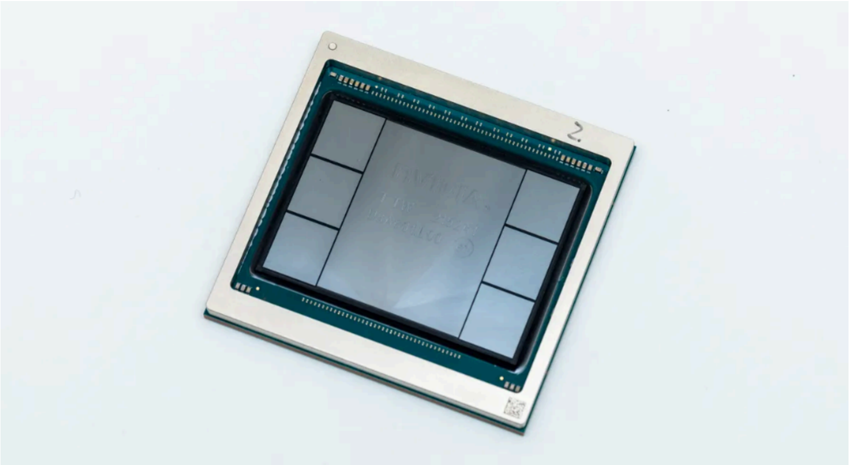


Figure 7. NVIDIA Hopper H100 GPU with six co-located HBM modules in one package

Inserting a LightBundle interconnect between the HBM module and the logic die offers several benefits and can be done in a phased approach as illustrated in Figure 8 below:

- **Phase 1:**
  Remoting the HBM module away from the hot logic die. HBM modules need to be refreshed periodically and the cycle frequency increases with temperature. Re-locating the HBM module to a cooler location of the board reduces the number of refresh cycles which increases overall throughput and reduces energy consumption.
- **Phase 2:**
  With the HBM module removed from the logic die the I/O density can be roughly doubled which means the processor can now access double the memory at the same latency. Moreover, the HBM modules can be placed wherever most convenient on the PCB to achieve the overall design targets.
- **Future:**
  As chiplet interfaces keep gaining acceptance in the processor and memory domain, LightBundle

interconnects will enable ever more flexible architectures with increased bandwidth density yet virtually constant energy per bit and overall link latency.

Phase 1:

Remote HBM from hot logic die

- Bandwidth Density: 1Tbps/mm
- Energy Efficiency: 1pJ/bit
- Latency: 2ns + ToF

Phase 2:

Double HBM Bandwidth per die

- Bandwidth Density: 2Tbps/mm
- Energy Efficiency: 1pJ/bit
- Latency: 2ns + ToF

Future:

Advanced UCIe increases Bandwidth 10x

- Bandwidth Density: 10Tbps/mm
- Energy Efficiency: 1pJ/bit
- Latency: 2ns + ToF

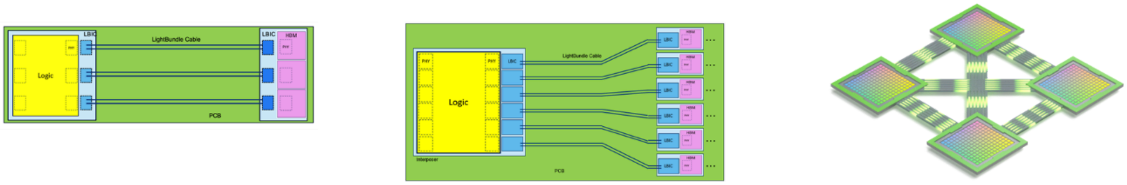
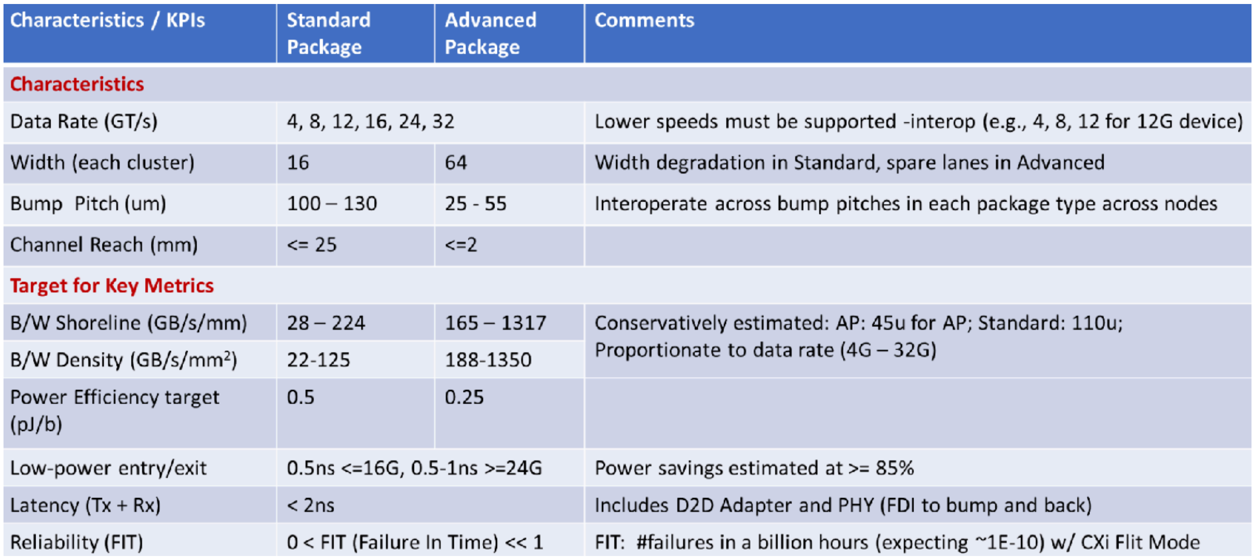


Figure 8: Roadmap for optical HBM interfaces using LightBundle interconnect.

Parallel Chiplet Interfaces, e.g., UCIe

Fundamentally, any electrical interface can be supported by the LightBundle interconnect. The LightBundle transceiver ASIC will convert the electrical data format to match the optical microLED transmission format. However, any parallel electrical interface will be best suited to work in combination with the parallel microLED array interface of the LightBundle chiplet since no power hungry SerDes will be needed. The IC will either match the electrical lane rate to the data rate of each individual microLED link or provide a simple muxing/demuxing function to keep latency and energy consumption to a minimum. A prime candidate is the advanced version of Universal Chiplet Interface Express (UCIe) which is gaining broad industry support (Figure 9).



Characteristics / KPIs	Standard Package	Advanced Package	Comments
Characteristics			
Data Rate (GT/s)	4, 8, 12, 16, 24, 32		Lower speeds must be supported -interop (e.g., 4, 8, 12 for 12G device)
Width (each cluster)	16	64	Width degradation in Standard, spare lanes in Advanced
Bump Pitch (um)	100 – 130	25 - 55	Interoperate across bump pitches in each package type across nodes
Channel Reach (mm)	<= 25	<=2	
Target for Key Metrics			
B/W Shoreline (GB/s/mm)	28 – 224	165 – 1317	Conservatively estimated: AP: 45u for AP; Standard: 110u; Proportionate to data rate (4G – 32G)
B/W Density (GB/s/mm^2)	22-125	188-1350	
Power Efficiency target (pJ/b)	0.5	0.25	
Low-power entry/exit	0.5ns <=16G, 0.5-1ns >=24G		Power savings estimated at >= 85%
Latency (Tx + Rx)	< 2ns		Includes D2D Adapter and PHY (FDI to bump and back)
Reliability (FIT)	0 < FIT (Failure In Time) << 1		FIT: #failures in a billion hours (expecting ~1E-10) w/ CXi Flit Mode

Figure 9: Preliminary specifications of the standard and advanced versions of the Universal Chiplet Interface Express (UCIe) Interface according to the UCIe Consortium 2022.

As the microLED interconnect technology matures there will be many more possible applications. Because GaN based microLEDs are inherently tolerant of operating in harsh environments and under extreme temperatures, these interconnects will be well suited for automotive and aerospace applications as well as camera sensor to processor links in smart phones or antenna interconnects in cellular towers or radar stations.

Conclusion

As AI/ML, HPC and Cloud computing networks are growing ever more complex the need for energy efficient high-bandwidth interconnects is increasing relentlessly. Electrical interfaces only offer a viable solution for short reach applications, and conventional optical solutions using VCSEL or Silicon Photonics are well positioned to serve the longer distance needs. However, a lot of the emerging AI/ML computing clusters require efficient interconnects for intermediate reach of up to a few meters and here microLED links offer a compelling alternative because of their high-bandwidth density and excellent energy efficiency. The inherent parallel nature of microLED arrays makes them well suited for chip-to-chip and chip-to-memory communications with no need for power hungry SerDes and FEC to support high data rates at low latency. As microLED based interconnects mature we expect to see many more applications beyond computing in automotive, aerospace and sensor markets.

Glossary

Term: definition.

References

1. J. E. Proesel, B. G. , C. W. Baks, and C. L. Schow "32-Gb/s VCSEL-based optical link using 32nm SOI CMOS circuits," 2013 Optical Fiber Communications conference and Exhibition, paper OM2H.2
2. M. Wade, E. Anderson, S. Ardalan, W. Bae, B. Beheshtian, S. Buchbinder, K. Chang, P. Chao, H. Eachempatti, J. Frey, E. Jan, A. Katzin, A. Khilo, D. Kita, U. Krishnamoorthy, C. Li, H. Lu, F. Luna, C. Madden, L. Okada, M. Patel, C. Ramamurthy, M. Raval, R. Roucka, K. Robberson, M. Rust, D. Van Orden, R. Zeng, M. Zhang, V. Stojanovic, F. Sedgwick, R. Meade, N. Chan, J. Fini, B. Kim, S. Liu, C. Zhang, D. Jeong, P. Bhargava, M. Sysak, C. Sun, "An error-free 1Tbps WDM optical I/O chiplet and multi-wavelength multi-port laser," 2021 Optical Fiber Communications conference and Exhibition (OFC).
3. D. F. Logan, S. Gebrewold, K. Murray, A. Dewanjee, E. Huante-Ceron, D. Kim, A. Baker, M. Kukiela, F. Znidarsic, M. Koehler, J. Whiteaway, R. Chen, C. Dorschky, G. Roell, "800 Gb/s Silicon Photonic Transmitter for CoPackaged Optics," 2020 IEEE Photonics conference (IPC).
4. R. X. G. Ferreira, E. Xie, J. J. D. McKendry, S. Rajbhandari, H. Chun, G. Faulkner, S. Watson, A.. E. Kelly, E. Gu, R. V. Penty, I. H. White, D. C. O'Brien,and M. D. Dawson, "High Bandwidth GaN-Based Micro-LEDs for Multi-Gb/s Visible Light," IEEE Photonics Technology Letters, vol. 28, vol. 19, p. 2023, (2016).
5. B. Pezeshki, A. Tselikov, R. Kalman, and C. Danesh "Wide and parallel LED-based optical links using multi-core fiber for chip-to-chip communications," 2021 Optical Fiber Communications Conference and Exhibition (OFC)
6. B. Pezeshki, A. Tselikov, R. Kalman, E. Afifi, "Sub 1pJ/bit dense optical interconnects using microLEDs on CMOS transceiver ICs," Proceedings Volume 12441, Light-Emitting Devices, Materials, and Applications XXVII; 1244106 (2023) https://doi.org/10.1117/12.2647762, SPIE OPTO, 2023, San Francisco, California, United States
7. B. Pezeshki, A. Tselikov, C. Danesh, R.Kalman, "8x 2Gb/s LED-Based Optical Link at 420nm for Chip-to-Chip Applications," 2021 European Conference on Optical Communication (ECOC) 2020
8. R. Delmdahl, R. Pätzel, J. Brune, "Large-area laser-lift-off processing in microelectronics," Physics Procedia 41, pp. 241 – 248, 2013.

14. Chiplet Technology for Advancing Optical Interconnects

Chiplet Technology for Advancing Optical Interconnects

Author: Radha Nagarajan, Senior Vice President and CTO of Optical and Cloud Connectivity at Marvell and a Visiting Professor at the National University of Singapore.

Executive Summary

Chiplet technologies based on 2.5D and 3D packaging and through silicon vias (TSVs) hold tremendous promise for manufacturing higher performance optical components and raising the performance, capabilities, and energy efficiency of data centers and AI clusters. These techniques effectively allow manufacturers to combine all of the critical elements of an optical module into a single semiconductor package to reduce board space, simplify manufacturing, and lower power consumption while also providing flexibility in overall design. Commercialization of these techniques potentially could lower the power consumption of optical modules (or components that perform the function of modules today) to 5 picojoules per bit (pJ/bit) and lower while smoothing the glide path for new optical technologies and designs.

Introduction: The Evolution of Optical Transceivers

Optical digital signal processors (DSPs) and transceivers entered data centers approximately 20 years ago and have played a critical role in the evolution of the cloud. The vast majority of connections above 5 meters inside data centers are made through PAM4 DSPs; in the near future, optical DSPs will also likely be employed to connect devices with racks closer than 5 meters to enable further disaggregation of processors, storage, and memory.

Over the last twenty years, data rates in optical pluggable modules powered by PAM4 DSPs have increased by 1000x, while the energy consumption per bit has dropped by 100x. Power consumption is now approaching 10 picojoules per bit. See Figure 1[1]. (Coherent DSPs and coherent modules for connecting data centers over long geographic distances have enjoyed a similar trajectory. Coherent DSP modules will also benefit from packaging. This paper, however, focuses on the PAM4 DSP technology largely deployed inside data centers and telco networks.)

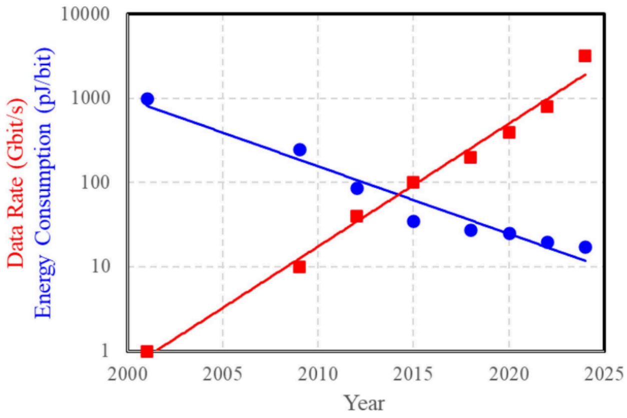
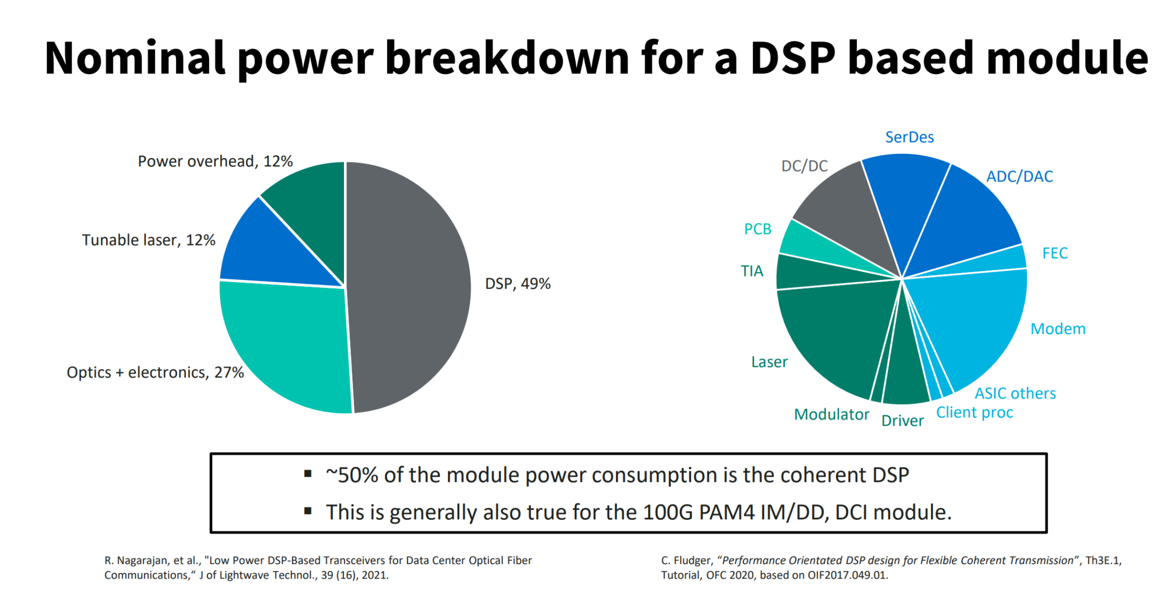


Figure 1

The aggregate power consumption of modules, however, has increased because of the enormous increases in bandwidth. At the same time, the number of optical modules is on the rise: A high-performance AI cluster today might contain 32,000 GPUs and 74,000 optical modules. By 2025, some clusters could contain hundreds of thousands of modules that consume several megawatts of power[2].

Optical DSPs consume roughly half of the power of a module. (Figure 2 is a set of pie charts from a slide from a talk at the Open Compute Project (OCP) Global Summit in San Jose in October 2023.) That large percentage has prompted many to develop design concepts that eliminate optical DSPs within modules and rely instead on the DSP embedded in a switch. Relying solely on embedded DSPs, however, can reduce performance and signal integrity. It also requires increasing SerDes power. Additionally, various analysts and customers have noted the potential for incompatibilities between components from different vendors. The challenges in overcoming these trade-offs in part is behind the fact that neither co-packaged optics (CPO) nor linear packaged optics (LPO) have been deployed in commercial volumes.



Nominal power breakdown for a DSP based module

- ~50% of the module power consumption is the coherent DSP
- This is generally also true for the 100G PAM4 IM/DD, DCI module.

R. Nagarajan, et al., "Low Power DSP-Based Transceivers for Data Center Optical Fiber Communications," J of Lightwave Technol., 39 (16), 2021.

C. Fludger, "*Performance Orientated DSP design for Flexible Coherent Transmission*", Th3E.1, Tutorial, OFC 2020, based on OIF2017.049.01.

Figure 2

Enter Chiplet Technologies

In contemporary modules, optical DSPs, transimpedance amplifiers (TIAs), drivers and other components are mounted on a common substrate and connected by wire bonds 500µm or less in length (top row in Figure 3[1]). A large contributor to the total module power consumption is the host electrical interface. Modules based around this device can operate at 14 pJ/bit.

The terms 2.5D and 3D packaging refer to how the discrete chips are integrated on a common substrate. With 2.5D integration (middle row), a wire bond connects the optical DSP to a vertical stack consisting of a silicon interposer and TIA and/or other active components linked via TSVs: in such a design, power can be lowered to 10 pJ/bit. By eliminating the second wire bond and shifting to 3D packaging where all of the active connections are vertically connected on top of the silicon interposer (bottom row), power can further be reduced.

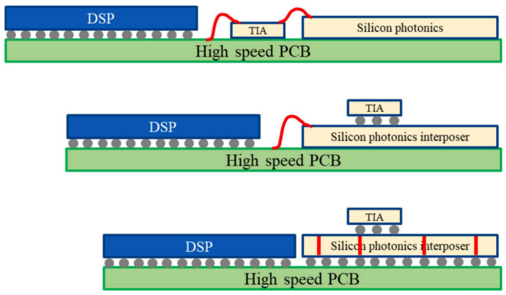


Figure 3

OEMs benefit as well: a unified integrated system allows them to combine components from different process generations or foundries while also reducing board space. 2.5D and 3D packaging techniques can be employed for traditional modules, producing linear packaged modules (which contain TIAs and drivers only) or co-packaged modules for insertion into switches. Module manufacturers can also integrate traditionally discrete components such as TIAs and drivers and combine them through packaging to a DSP. Experience with 2.5D and 3D packaging has been percolating across the industry rapidly[3].

Another aspect worth noting are the thermal properties of advanced packaging. Wire bonded assemblies are harder to cool, and the heat needs to be pulled out from the bottom, thru the PCB. In 2.5D integrated structures, the wire bonds are not “in the way”, and the heat can be channeled to the top through the package lid.

Results

In tests conducted at Marvell (Figure 4), the single wire bond module outperforms the first, but both wire bond-based approaches have limited analog bandwidth, and signal discontinuities, as shown in the modeling results of Fig. 5. The cleanest, from the high-speed packaging point of view, is the last row, where the silicon photonics device is also an interposer with TSV’s, and the assembly is completely free of wire bonds. Higher levels of integration also allow for closer placement of devices which minimizes the parasitic power consumed to compensate for the frequency dependent losses in the interconnect traces.

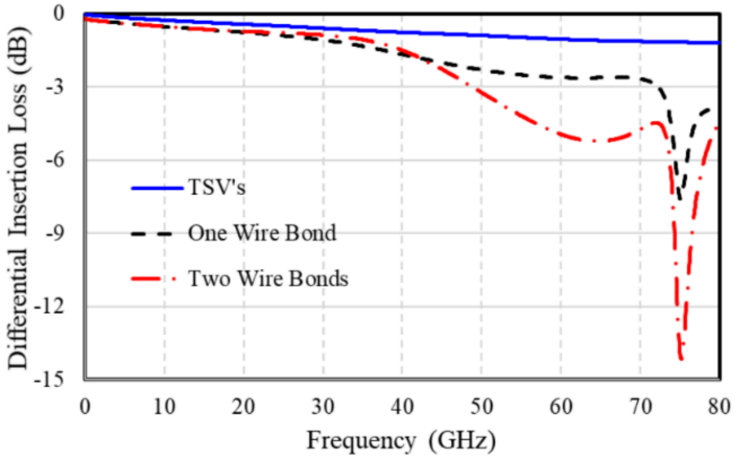


Figure 5

As the optical modules get smaller and are co-packaged with an electrical host ASIC, the power at this interface can be lowered further still. With even tighter integration, we may not need a DSP inside the optical module, and it can be directly driven by the host ASIC. At this point, the power consumption may be reduced to about 5 pJ/bit.

We also demonstrated 2.5D light engines transmitting data within a 1.28T Teralynx 7 switch at OFC in 2022: the module carried live traffic successfully with 5km link[5]. Similarly, Cisco has demonstrated how power consumption and thermal density can be lowered through a combination of co-packaging optics and advanced packaging with silicon photonics[5]. An SEM image of the device is below:

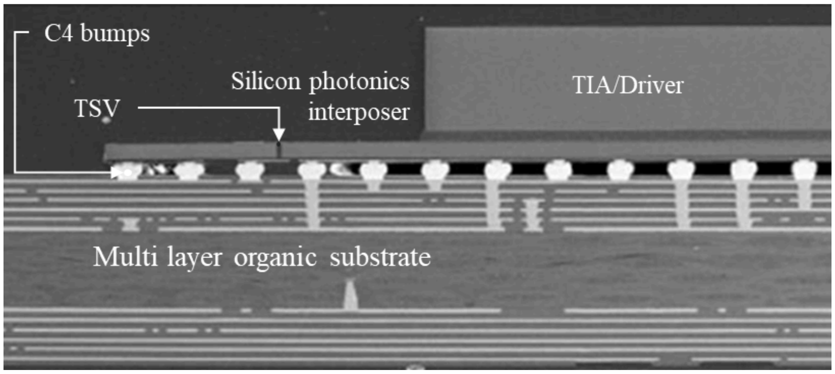


Figure 6

Conclusion

The use of optical in networking will continue to grow as data infrastructures expand and the performance of applications escalate. Optical components, however, will have to undergo fundamental changes to meet these challenges. Leveraging existing technologies fine-tuned in other semiconductor markets like 2.5D and 3D packaging can help the module makers, semiconductor designs and, ultimately, their customers accelerate the path to adoption while delivering gains in performance, economics and sustainability.

References

1. IEEE Journal of Selected Topics in Quantum Electronics, Vol. 29 Number 3 May/June 2023
2. Marvell, industry estimates
3. McKinsey & Company, May 2023.
4. Marvell, March 7, 2022.
5. Cisco, March 7, 2023

15. Conclusion

This whitepaper, assembled in a “Conference Proceedings” style (see reference) from Wikipedia, is composed of some 11 independently contributed perspectives, that in total, describes the overall photonics landscape of drivers and examines some implementation details for Short Reach Photonics Interconnects. As the reader will be able to gather, while there is agreement on the nature of workloads, specifically around Artificial Intelligence (AI) and High Performance Computing (HPC) and that these workloads are demanding step change increases in available compute and networking performance, there is not a consensus on the path forward. The goal of this type of proceedings/white paper is to provide the reader with knowledge in an open way such that they may form their own opinions outside of the market or purchasing context and act as a marker on the state of the art. Since the time this document was assembled, a new OCP Future Technologies Initiative called Short Reach Optical Interconnects (SROI) has been formed to examine the differences in the roadmaps of copper and optical interconnects and form analysis that will elucidate the tradeoffs in areas like performance, reliability, and manufacturability resulting in potential ways forward for the industry. From this paper, it is expected that the proceeding SROI work will result in future industry based roadmaps and ultimately products within the open community that will impact the market.

The OCP would like to acknowledge the great efforts of the contributors. **Thank you** for all of your time contributing to open photonics knowledge.

Lastly, the OCP encourages the entire photonics community to join and participate to have their voice heard, collaborate and influence the future of the industry.

Editorial comments: Bijan Nowroozi, CTO, OCP

16. Glossary

AI/ML - Artificial Intelligence / Machine Learning. Types of computer workloads and applications that are driving increased demands for computing performance.

BER - Bit Error Rate. A measure of the number of bit errors that occur in data transmission.

CDI - Composable Disaggregated Infrastructure. An architecture that allows computing resources like GPUs, memory, and storage to be provisioned flexibly on-demand.

CPO - Co-Packaged Optics. Optical components like modulators and detectors integrated in the same package as the main computing chip or switch ASIC.

CXL - Compute Express Link. An open interconnect standard designed for high-speed CPU-to-device and CPU-to-memory links.

DML - Directly Modulated Laser. A type of optical transmitter where the laser current is modulated to transmit data.

DWDM - Dense Wavelength Division Multiplexing. Transmitting multiple optical signals at different wavelengths on the same fiber.

EML - Externally Modulated Laser. A type of optical transmitter where a laser provides continuous light that is modulated by a separate device.

FEC - Forward Error Correction. Adding redundancy to transmitted data to allow errors to be corrected.

HDI/O - High Density I/O. An approach to achieve very high bandwidth density interconnects between chips, enabled by photonic technologies.

LLM - Large Language Model. An AI model with a very large number of parameters, used for natural language processing tasks.

LPO - Linear Pluggable Optics. Optical transceivers that connect directly to the host ASIC electrical I/O without requiring a retimer.

OCP - Open Compute Project. A collaborative community developing open innovations for data centers and computing infrastructure.

PAM4 - Pulse Amplitude Modulation 4-level. A signaling method that transmits two bits per symbol for higher data rates.

PCIe - Peripheral Component Interconnect Express. A high speed serial computer expansion bus standard.

QD - Quantum Dot. A type of semiconductor nanostructure used as the gain medium in some lasers.

SerDes - Serializer/Deserializer. Circuitry that converts between serial and parallel data formats, often used in high-speed communications.

SiPh - Silicon Photonics. Photonic integrated circuits based on silicon, allowing integration of optical and electronic components.

VCSEL - Vertical Cavity Surface Emitting Laser. A type of semiconductor laser diode with the laser cavity perpendicular to the chip surface.

17. References

17.1. See 2023 OCP Global Summit, FTI, Presentations on Photonics https://www.opencompute.org/events/past-events/2023-ocp-global-summit#symposium

Overview of the White paper, the Agenda and the Workstream	Michael Bortz (OCP)	Video	Slides
NIC PhotoNICs	David Piehler (Dell)	Video	Slides
Opportunities for Optical Computer Interconnects: A Meta Platforms Perspective	Drew Alduino (Meta)	Video	Slides
Photonics-Based Resource Disaggregation for HPC	Georgios Michelogiannakis (LBL/Stanford)	Video	Slides
Optical CXL Interconnect for Large Scale Memory Pooling	Ron Swartzentruber (Lightelligence)	Video	Slides
Optical Interconnect: Pathways to an Open Infrastructure for AI	Matthew Williams (Rockport Networks)	Video	Slides
The Long Road to Rack Scale Disaggregation	Bob Wheeler (LightCounting)	Video	Slides

Energy Efficient Optical Links for PCIE	Jeff Hutchins (Ranovus)	Video	Slides
Scalable Optical I/O for Disaggregated Infrastructure	LK Bhupathi (Ayar Labs)	Video	Slides
High Density Optical Interconnect for the ML Array Edge	Karen Liu (Nubis)	Video	Slides
Multi-wavelength Technology for Scalable and Reliable Optical Compute Interconnects	Brian Koch (Quintessent)	Video	Slides
VCSELs in co-packaged optics for SR applications	Vipul Bhatt (Coherent)	Video	Slides
High Density Low Power Micro-LED based Optical Interconnects for Chip-to-Chip Communications	Chris Pfistner (Avicena)	Video	Slides

Heterogeneous Integration and Linear Optical Engines	Radha Nagarajan (Marvell)	Video	Slides
Panel/Open session/Rump Session	Michael Bortz (OCP)	Video	

18. License

18.1. Creative Commons

OCP encourages participants to share their proposals, specifications and designs with the community. This is to promote openness and encourage continuous and open feedback. It is important to remember that by providing feedback for any such documents, whether in written or verbal form, that the contributor or the contributor's organization grants OCP and its members irrevocable right to use this feedback for any purpose without any further obligation.

It is acknowledged that any such documentation and any ancillary materials that are provided to OCP in connection with this document, including without limitation any white papers, articles, photographs, studies, diagrams, contact information (together, “Materials”) are made available under the Creative Commons Attribution-ShareAlike 4.0 International License found here: https://creativecommons.org/licenses/by-sa/4.0/, or any later version, and without limiting the foregoing, OCP may make the Materials available under such terms.

As a contributor to this document, all members represent that they have the authority to grant the rights and licenses herein. They further represent and warrant that the Materials do not and will not violate the copyrights or misappropriate the trade secret rights of any third party, including without limitation rights in intellectual property. The contributor(s) also represent that, to the extent the Materials include materials protected by copyright or trade secret rights that are owned or created by any third-party, they have obtained permission for its use consistent with the foregoing. They will provide OCP evidence of such permission upon OCP’s request. This document and any "Materials" are published on the respective project's wiki page and are open to the public in accordance with OCP's Bylaws and IP Policy. This can be found at http://www.opencompute.org/participate/legal-documents/. If you have any questions please contact OCP.

Footer:

19. About Open Compute Foundation

The Open Compute Project (OCP) is a collaborative Community of hyperscale data center operators, telecom, colocation providers and enterprise IT users, working with the product and solution vendor ecosystem to develop open innovations deployable from the cloud to the edge. The OCP Foundation is responsible for fostering and serving the OCP Community to meet the market and shape the future, taking hyperscale-led innovations to everyone. Meeting the market is accomplished through addressing challenging market obstacles with open specifications, designs and emerging market programs that showcase OCP-recognized IT equipment and data center facility best practices. Shaping the future includes investing in strategic initiatives and programs that prepare the IT ecosystem for major technology changes, such as AI & ML, optics, advanced cooling techniques, composable memory and silicon. OCP Community-developed open innovations strive to benefit all, optimized through the lens of impact, efficiency, scale and sustainability. Learn more at www.opencompute.org.