MOSAIC: Breaking the Optics versus Copper Trade-off with a Wide-and-Slow Architecture and MicroLEDs

Kaoutar Benyahya[*] Ariel Gomez Diaz[*] Junyi Liu[*] Vassily Lyutsarev[*]
Marianna Pantouvaki[*] Kai Shi[*] Shawn Yohanes Siew[*] Hitesh Ballani[*] Thomas Burridge[*]
Daniel Cletheroe[*] Thomas Karagiannis[*] Brian Robertson[*] Ant Rowstron[*] Mengyang Yang[*]
Arash Behziz[†] Jamie Gaudette[†] Paolo Costa[*]
[*]Microsoft Research [†]Microsoft Azure

Abstract

Link technologies in today's data center networks impose a fundamental trade-off between reach, power, and reliability. Copper links are power-efficient and reliable but have very limited reach (< 2 m). Optical links offer longer reach but at the expense of high power consumption and lower reliability. As network speeds increase, this trade-off becomes more pronounced, constraining future scalability.

We introduce MOSAIC, a novel optical link technology that breaks this trade-off. Unlike existing copper and optical links, which rely on a *narrow-and-fast* architecture with a few high-speed channels, MOSAIC adopts a *wide-and-slow* design, employing hundreds of parallel low-speed channels. To make this approach practical, MOSAIC uses directly modulated microLEDs instead of lasers, combined with multicore imaging fibers, and replaces complex, power-hungry electronics with a low-power analog backend. MOSAIC achieves $10\times$ the reach of copper, reduces power consumption by up to 68%, and offers $100\times$ higher reliability than today's optical links. We demonstrate an end-to-end MOSAIC prototype with 100 optical channels, each transmitting at 2 Gbps, and show how it scales to 800 Gbps and beyond with a reach of up to 50 m. MOSAIC is protocol-agnostic and seamlessly integrates with existing network infrastructure, providing a practical and scalable solution for future networks.

CCS Concepts

• **Networks** → **Physical links**; **Data center networks**; • **Hardware** → **Networking hardware**.

Keywords

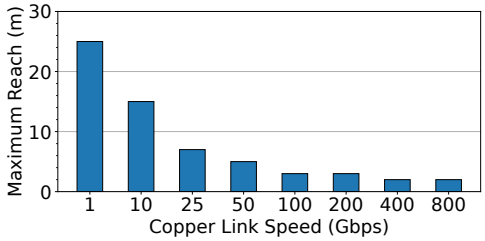
Optical interconnects, Wide-and-slow architecture, MicroLEDs, Data center networks, Cloud computing, AI infrastructure

ACM Reference Format:

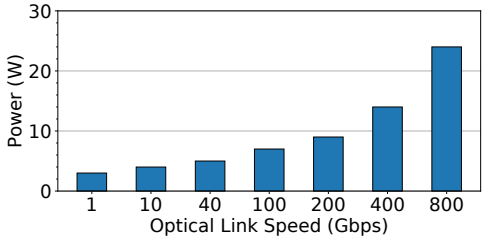
Kaoutar Benyahya, Ariel Gomez Diaz, Junyi Liu, Vassily Lyutsarev, Marianna Pantouvaki, Kai Shi, Shawn Yohanes Siew, Hitesh Ballani, Thomas Burridge, Daniel Cletheroe, Thomas Karagiannis, Brian Robertson, Ant Rowstron, Mengyang Yang, Arash Behziz, Jamie Gaudette, and Paolo Costa. 2025. MOSAIC: Breaking the Optics versus Copper Trade-off with a Wide-and-Slow Architecture and MicroLEDs. In *ACM SIGCOMM 2025 Conference (SIGCOMM '25), September 8–11, 2025, Coimbra, Portugal.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3718958.3750510

SIGCOMM '25, Coimbra, Portugal

ACM ISBN 979-8-4007-1524-2/25/09
https://doi.org/10.1145/3718958.3750510



(a) Copper link reach.



(b) Optical link power (two transceivers per link).

Figure 1: As network speeds increase, the reach of copper links shortens and the power of optical links grows.

1 Introduction

To meet the increasing demands of cloud and AI workloads, data center operators are forced to make a hard compromise by choosing between reach, power, and reliability for the network links. Electrical (copper) links are power-efficient and highly reliable but can only transmit up to a very short distance (< 2 m), restricting their use to within a single rack. On the other hand, predominant optical links used in data centers (active optical cables or AOCs [2]) span tens of meters but at the cost of high power consumption and high failure rates (up to $100\times$ worse than copper).

Looking ahead, as network speeds increase, these limitations are worsening as optical links' power consumption continues to rise while copper's reach keeps shrinking (Fig. 1). Therefore, in the near future, we expect an increased reliance on optical links even within a rack. This perfect storm suggests that data centers might face a *networking wall*, similar to the memory wall [14], leading to higher deployment costs, greater power consumption, and reduced reliability.

To solve these challenges, we need a *fundamental* rethinking of the link technology. In this paper, we introduce MOSAIC, a novel optical link technology that breaks the optics versus copper trade-off, enabling long reach, low power, and high reliability *simultaneously*. MOSAIC is backward-compatible with existing standard link form factors (e.g., pluggable QSFP/OSFP [54]) and electrical host interfaces (e.g., PCIe or VSR/MR [57, 63]), offering a drop-in replacement for today's optical and copper links without requiring any server or switch changes.

The key insight behind this work is that the optics versus copper trade-off stems from their reliance on a *narrow-and-fast* (NaF) model, which utilizes only a few high-speed serial channels (e.g., 8×100 Gbps channels for an 800 Gbps link). In copper links, higher channel speeds lead to greater signal integrity challenges, which limits their reach. In optical links, high-speed transmission is inherently power-inefficient, requiring power-hungry laser drivers and analog/digital converters (ADC/DAC) as well as complex digital signal processing (DSP) and forward error correction (FEC) to compensate for transmission impairments. Sustaining high speeds also pushes the limits of optical components (e.g., lasers and modulators), which increases failure rates and reduces overall reliability (§2). These challenges worsen as channel speeds increase.

In contrast, MOSAIC employs a *wide-and-slow* (WaS) architecture, shifting from a small number of high-speed *serial* channels to hundreds of *parallel* low-speed optical channels. This is reminiscent of memory and chip buses, where parallel low-speed channels are preferred due to their lower power consumption, higher reliability, and simpler design [19, 25]. Unfortunately, today's copper and optical technologies make such a design impractical due to *i)* electromagnetic interference challenges in high-density copper cables and *ii)* the high cost and power consumption of lasers as well as the increase in packaging complexity. MOSAIC overcomes these issues by leveraging directly modulated microLEDs (§3.2), a technology originally developed for displays [44]. MicroLEDs are significantly smaller than traditional LEDs (ranging from a few to tens of μm) and, due to the small size, can be modulated at several Gbps using a simple ON-OFF scheme. MicroLEDs are manufactured in large arrays with over half a million microLEDs in a small physical footprint for high-resolution displays, e.g., head-mounted devices or smartwatches [3, 8]. For our purposes, small arrays of these devices are sufficient to meet high (aggregate) speeds. For example, assuming 2 Gbps per microLED channel, an 800 Gbps MOSAIC link can be realized by using a 20×20 microLED array, which can fit in less than 1 mm $\times$ 1 mm die.

MOSAIC's WaS design provides four core benefits. First, operating at low speed improves power efficiency, achieving up to 68% power reduction compared to today's optical links (§5). Second, by leveraging optical transmission (via microLEDs), it sidesteps copper's reach issues, supporting distances up to 50 m, i.e., $> 10\times$ longer than copper and comparable to current AOCs. Third, microLEDs are more reliable than lasers due to their simpler structure and temperature insensitivity [65]. The parallel nature of WaS also makes it easy to add redundant channels, further increasing reliability — two orders of magnitude better than AOCs (§7). Finally, the WaS approach is also scalable as higher aggregate speeds (e.g., 1.6 Tbps or 3.2 Tbps) can be achieved by increasing the number of channels and/or raising per-channel speed (e.g., to 4-8 Gbps).

While conceptually simple, realizing this architecture posed a few key challenges. First, using individual fibers per channel would be prohibitively complex and costly due to the large number of channels. We addressed this by employing imaging fibers, which can support thousands of cores in a single fiber, enabling multiplexing many channels within a single fiber (§3.3). Second, compared to lasers, microLEDs are a less pure light source with *i)* a larger beam shape (which complicates fiber coupling) and *ii)* a broader spectrum (which negatively affects fiber transmission due to chromatic dispersion). We tackled these issues through innovative optical lens design (§3.4) and a power-efficient analog-only electronic backend, which does not require any expensive digital signal processing (§3.5). We also leverage MOSAIC's high channel count to explore new system design opportunities to improve reliability (§4.1), reduce electronics complexity (§4.2), and achieve power proportionality (§4.3).

We demonstrate the feasibility of MOSAIC through an end-to-end prototype comprising 100 channels, each transmitting at 2 Gbps over 20 m (1.6 Gbps over 30 m) and show how this can naturally scale to higher aggregate speeds (800 Gbps and beyond) and longer distances (up to 50 m) in §6. MOSAIC is protocol-agnostic as it simply relays bits from one endpoint to another without terminating or inspecting the connection. We have validated our prototype using Ethernet and InfiniBand stacks (§6) and have confirmed its compatibility with newer protocols such as NVLink [38] and CXL [17].

In this work, we focus specifically on link technology with the goal of providing a practical path to scaling existing network designs (e.g., mainstream Clos topologies) to future generations. Historically, however, step changes in network technologies have triggered transformative advances in computing and applications [64]. By overcoming the reach, power and/or reliability limitations of existing link technologies, we hope that MOSAIC will also act as an enabler for many recently proposed (and hopefully new) topologies and architectures for next generation of data center and AI clusters. We briefly outline some of these opportunities in §8.

[This work does not raise any ethical issues.]

2 Motivation

In this section, we discuss the limitations of existing copper and optical interconnects, and explain how MOSAIC's WaS architecture overcomes these challenges.

The curse of narrow-and-fast (NaF) architectures. Fig. 1 illustrates the fundamental scaling bottleneck faced by current link technologies: as data rates increase, copper reach shrinks while optical power consumption rises. In both copper and optical domains, the root cause of this poor scalability is the continued reliance on a NaF architecture with a small number of high-speed channels.

For copper links, higher data rates imply higher modulation frequencies, which suffer from greater signal attenuation over the wire due to the skin effect and dielectric loss [30]. As signal loss increases approximately linearly with frequency, this results in roughly halving the reach at every generation.

For optical links, scaling to higher data rates is challenging both on the electronics and photonics front. On the electronics side, the power of analog components such as drivers, clock-and-data recovery (CDR) circuits, and ADCs/DACs grows proportionally to the

modulation frequency and, unlike digital logic, it has not scaled efficiently to smaller CMOS process nodes [24]. On the photonics side, higher speeds stretch the performance of optical components such as lasers and modulators, reducing system margins and tolerance to noise, thus requiring advanced DSP and forward error correction (FEC) logic to compensate for impairments. As a concrete example of these scaling challenges, in the latest 800 Gbps generation, the optics industry has resorted to doubling the channel count, i.e., moving from 4×100 Gbps lanes for 400 Gbps to 8×100 Gbps lanes for 800 Gbps, rather than doubling the channel speed as in previous generations, due to the difficulty of scaling to 200 Gbps per channel.

Beyond power consumption, higher speeds are also directly linked to higher failure rates. As optical margins shrink, the impact of aging effects, temperature fluctuations (causing laser wavelength drift), mechanical stress, and environmental contamination (e.g., dust and humidity) becomes more pronounced. Further, high-speed DSP chips consume more power, increasing heat dissipation and accelerating optical component wear, which further shortens component lifetime. Finally, the high cost per channel makes it impractical to adopt any form of redundancy to protect against individual channel failure.

The future looks even bleaker. For example, the next-generation 1.6 Tbps copper links will support <1m reach [41], i.e., less than half of the size of a rack. In principle, signal loss could be compensated by adding retimers on-path (active electrical cables or AECs [32]), which would avoid halving the reach, but this would incur higher power consumption, bringing it closer to optical solutions. In optics, the difficulty of scaling beyond 100/200 Gbps per lane leaves the industry with two costly paths: doubling channel count or moving to even more complex modulation schemes (e.g., PAM8 or QAM), both of which will lead to significantly higher power and lower reliability, not to mention cost implications.

AI clusters: a case in point. The recently announced NVIDIA NVL72 pod, which comprises 72 Blackwell B200 GPUs in a single NVLink scale-up domain [40], provides a prime example of the impact of these limitations on modern AI clusters. A single B200 GPU consumes approximately 1 kW and supports 7.2 Tbps (per direction) of network connectivity through NVLink [39]. NVIDIA has estimated that connecting them using optics would have increased rack power consumption by 20 kW per rack (i.e., the equivalent of 20 GPUs), a prohibitive penalty given fixed data center power budget [22]. Further, assuming a 100,000-GPU cluster [46] and typical failure rates for 800 Gbps links (§7), using optics would also result in a link failure every 6-12 hours, which would be particularly disruptive for AI workloads due to their synchronous nature. As a result, NVIDIA has opted to use copper to support NVLink connectivity. Due to copper's short reach, however, all 72 GPUs had to be hosted within a *single* rack with a total power consumption of 120 kW per rack [45]. This resulted in extremely high power density, requiring complex liquid cooling solutions, which have already caused deployment delays [61]. Looking ahead, the continuous increase in GPU power, combined with the shrinking distances of copper interconnects, will make this approach even more challenging. For example, next-generation NVIDIA NVL576 is projected to consume as much as 600 kW per rack [35], further raising the complexity of efficiently scaling future GPU clusters. These trends reinforce the need for link solutions that are low-power, long-reach, and highly reliable.

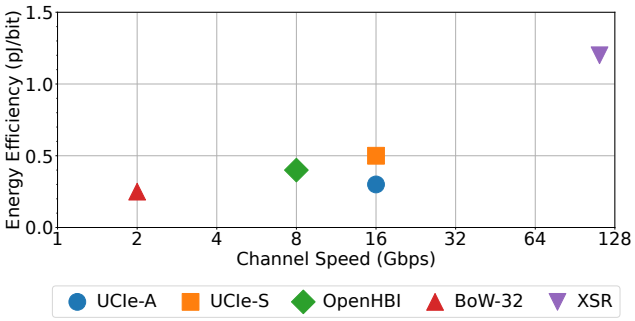


Figure 2: Energy efficiency in pJ/bit (equivalent to W/Tbps) for different memory and die-to-die interconnects operating at different data rates per channel.

Can a wide-and-slow (WaS) architecture help? The challenges associated with NaF architectures suggest that a disruptive shift is necessary. Interestingly, memory and chip interconnects typically adopt WaS architectures with many low-speed parallel channels to reduce power consumption. Fig. 2 shows the relationship between channel speed and energy efficiency in pJ/bit (which is equivalent to W/Tbps) across various chip interconnect technologies, comparing traditional high-speed serial links (XSR [59]) with wide-and-slow approaches (UCIe [36], BoW [42, 56], OpenHBI [43]). The results demonstrate that lower per-channel data rates consistently lead to higher overall energy efficiency. Beyond power savings, as previously discussed, slower speeds also provide lower channel loss (for copper links) and higher reliability (for optical links). Given these advantages, one might wonder *why networking link technologies deviated from memory and die-to-die interconnects, adopting a narrow-and-fast design instead?*

The reason lies in fundamental physical constraints. Unlike board traces, which can be densely routed with fine-pitch wiring, meter-long copper cables would suffer from electromagnetic interference (EMI) and crosstalk when multiple lanes are closely packed. As a result, increasing the number of lanes in a copper cable is impractical beyond a certain point, forcing high-speed serial transmission to maximize bandwidth over fewer lanes. While optics eliminate EMI, laser power consumption does not scale well with increasing channel count. A single laser used for communication typically consumes 10s to 100s of mW, and scaling up to hundreds of lasers would result in excessive power consumption. Additionally, the complexity of packaging multiple lasers and fibers at scale would incur severe reliability and manufacturing constraints. Finally, due to their reliability issues, increasing laser count would proportionally increase failure rates.

By adopting microLEDs as light source, MOSAIC overcomes these limitations, providing a practical way to implement an optical WaS solution. First of all, unlike copper, microLEDs use optical transmission, eliminating EMI and allowing channels to be densely packed without interference. Second, a microLED operates at just a few 100s of μW, orders of magnitude lower than traditional lasers, making it possible to scale to hundreds of channels without excessive power consumption. Third, a monolithically-integrated microLED

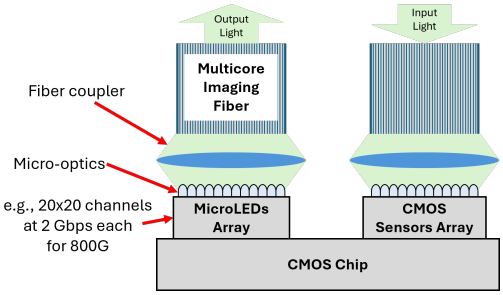


Figure 3: MOSAIC's high-level WaS architecture and key components.

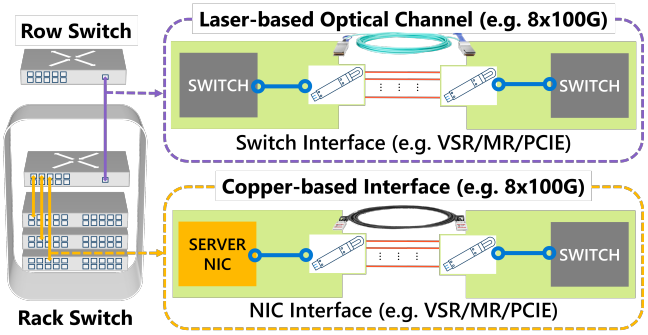


Figure 4: Today's network architecture comprising optical (top) and copper (bottom) links, using NaF architecture.

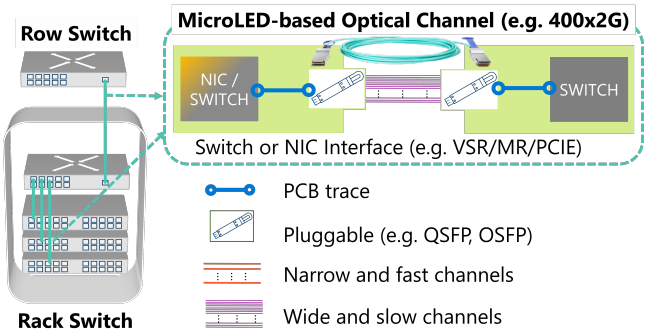


Figure 5: MOSAIC can seamlessly replace existing optical and copper cables without any hardware changes.

array can accommodate over 400 channels in 1 mm^2, enabling ultra-dense solutions with a simple packaging design when combined with MOSAIC's compact multicore imaging fiber (§3.3). Finally, unlike lasers, which require temperature control and active wavelength stabilization, microLEDs are intrinsically more robust and due to their array nature it is easy to add redundant channels to further enhance reliability. In the following sections, we describe how MOSAIC leverages microLEDs to create a WaS interconnect that combines long reach, low power, and high reliability.

3 MOSAIC Design

In this section, we describe the key components of our proposed link technology and highlight the main differences compared to mainstream, laser-based optical cables. We first start with a high-level overview and then we discuss each individual component in detail.

3.1 Overview

Fig. 3 illustrates the key building blocks of our technology. MOSAIC adopts a WaS architecture, i.e., it uses a large number of parallel channels, each operating at relatively low data rate per channel, 2 Gbps in our prototype (§6), using microLEDs as transmitters (§3.2). To scale to speeds of 800 Gbps and higher, we use a grid architecture. Simplistically, we could assume to just provision as many microLEDs as the ratio between the target link speed and the channel rate (e.g., 400 microLEDs in a 20×20 grid for 800 Gbps, at 2 Gbps per microLED). However, we show in §4 that introducing a few spare channels or *overprovisioning* can greatly improve link reliability as well as reducing the electronic complexity and power consumption with only marginal impact on overall power and cost.

Unlike lasers used for communication, which operate in the infrared range, microLEDs operate in the visible range (400 nm–700 nm). This is advantageous because it allows the use of low-cost CMOS sensors as receivers, similar to those found in mobile phone cameras (§3.2). A potential downside of using a large number of channels though is the increase in fiber count. In fact, if we were to naively use a separate fiber per channel as is the case today, the additional cost and complexity of managing such a large bundle would be unsustainable. Therefore, in MOSAIC, we depart from traditional single-core fiber in favor of multicore imaging fiber (§3.3) and we use custom micro-optics to efficiently couple light into it (§3.4). Transmitting at low speeds significantly reduces the electronics complexity, as no DSP is needed. We take advantage of this to design a power-efficient electronic backend, which relies on simple ON-OFF (NRZ) modulation scheme and low-speed analog equalization (§3.5).

In the design of MOSAIC, we ensured that all these components can fit within the form factor and support the same electrical interface as today's network links, enabling full compatibility with current network architectures and hardware. Modern data center networks use a multi-tier structure where copper cables connect servers to rack switches and AOCs link racks to row switches (Fig. 4), although due to the copper scalability challenges discussed in §2, in the near future we expect optical links to be used within the rack too. Both copper and optical links rely on the same standard pluggable connectors (e.g., QSFP, OSFP [54]) and electrical interface (e.g., PCIe, VSR/MR [57, 63]), ensuring deployment flexibility as both copper and optical cables can coexist.

We adopted a similar approach in MOSAIC and followed the same standards (§6), thus enabling replacing today's optical and copper links with MOSAIC ones, without requiring any modifications to switches or network interface cards (NICs) (Fig. 5). This provides a practical solution to scale to next generations of network speeds while maintaining full compatibility with current network architectures. At the same time, however, MOSAIC also unlocks new possibilities for novel and more efficient network topologies and protocols, which we discuss in §8.

3.2 MicroLEDs and CMOS sensors

MicroLEDs are functionally similar to standard LEDs used for illumination but they are significantly smaller (ranging from a few μm to tens of μm, compared with the mm scale of standard LEDs). The smaller size combined with higher efficiency and lifetime compared to OLED devices used in today's displays makes them very attractive for next-generation devices, especially portable ones such as augmented/virtual reality (AR/VR) head-mounted displays and smartwatches [44].

Blue and green microLEDs are fabricated on Gallium Nitride (GaN) wafers, while red microLEDs typically use aluminum gallium indium phosphide (AlGaInP) although recently red GaN microLEDs have also been developed. Compared to lasers, microLEDs exhibit a much simpler structure because they emit light through *spontaneous* rather than *stimulated* emission, i.e., light is generated by simple recombination of electrons with holes without requiring the use of any cavities like in lasers. This reduces manufacturing costs and increases reliability (including resilience to temperature variations). Further, unlike lasers, microLEDs have no lasing threshold, enabling operating at very low power levels. However, microLEDs also introduce two main downsides compared to lasers. First, the light beam generated by microLEDs is not collimated and it covers most of the emitting surface. This makes it harder to couple into the fiber, requiring the development of a new micro-optics design, specifically tailored for microLED emitters (§3.4). Second, microLEDs emit light with a much broader spectrum (tens of nm versus sub-pm) than lasers. This is particularly challenging for communication over fiber due to the impact of *chromatic dispersion*. This occurs because the propagation speed of light within a medium depends on its specific wavelength and, hence, if a signal comprises different wavelengths (i.e., it has a broad spectrum), each component will travel at a different speed, distorting the signal. We discuss the implications of this phenomenon and our proposed solution in §3.5.

A key advantage of operating in the visible range is the ability to use CMOS sensors (or silicon photodetectors), which are functionally equivalent to the ones found in mobile phone cameras, although operating at higher frequency (1-2 GHz as opposed to cameras' 120 Hz rates). This provides two key benefits. First, it enables leveraging the very mature and proven CMOS ecosystem and technology, which results in lower costs. Second, since they share the same CMOS technology as the receiver-side electronic backend (§3.5), they allow for tighter integration, including monolithic design, i.e., having a single silicon die with both the analog electronic components and the photodetector array, which further reduces costs (because it leads to fewer dies and avoids the need for complex packaging) as well as power (due to the tighter integration and shorter electrical traces).

3.3 Multicore Imaging Fiber

Imaging fibers are mass-produced and commercially available for medical applications (e.g., endoscopy) and illumination but they are typically not used for communication. They can comprise up to 10,000 cores per fiber. This is important because it enables multiplexing many MOSAIC channels within a *single* fiber, greatly simplifying packaging and deployment, and reducing costs.

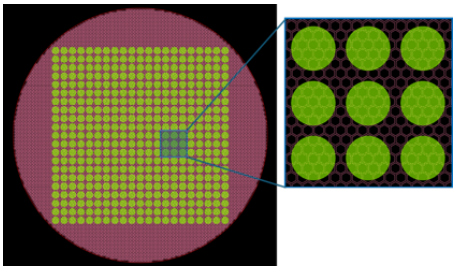


Figure 6: Example of a multicore imaging fiber with 1,000s of cores and 400 channels (left). Each channel is mapped onto multiple cores (right).

In principle, we could implement a 1:1 mapping between each fiber core and a microLED. In our design, however, given the abundance of available cores, we found it more beneficial to map a single microLED onto multiple fiber cores (see Fig. 6). This approach significantly relaxes alignment accuracy requirements, thus reducing overall complexity and cost.

Another advantage of using a single imaging fiber, as opposed to a bundle of discrete fibers, is that an imaging fiber is fabricated in a single process. This fabrication approach ensures that transmission characteristics such as optical loss and chromatic dispersion remain highly uniform across all cores within the same fiber. It also means that all cores have nearly identical lengths. Combined with the relatively slow per-channel data rate, this results in negligible channel-to-channel skew. For example, even assuming an extremely large length mismatch of 1 cm, with a light propagation speed in fiber of 5 ns/m, the resulting delay difference is only 50 ps. This corresponds to just 10% of the bit period (0.5 ns at 2 Gbps), which can be easily tolerated.

3.4 Micro-optics

A major disadvantage of using microLEDs compared to traditional lasers is that the former are *Lambertian* emitters, i.e., as shown in Fig. 7 (left), they emit across a hemisphere rather than generating a collimated spot as lasers do. This makes it harder to couple into the fiber without compromising on coupling efficiency. Additionally, the *Lambertian* beam shape in a multi-channel setup can lead to inter-channel crosstalk, as light from one microLED may couple into adjacent channels within the array.

To address this issue, initially we experimented with using standard micro-lens array (MLA) as depicted in Fig. 7 (center). However, while MLAs helped to improve coupling efficiency, they were still unable to capture much of the light. Therefore, we developed a novel custom lens design that leverages the principle of total internal reflection (TIR). TIR lenses are functionally similar to those used in torches and consist of a two-component micro-optics design, as shown in Fig. 7 (right). This design traps light within the lenses and achieves more than 2$\times$ higher coupling efficiency than MLAs. An important feature of these lenses is that despite their unconventional design, they are compatible with wafer-scale, high-throughput and low-cost manufacturing using nano-imprinting lithography [15].

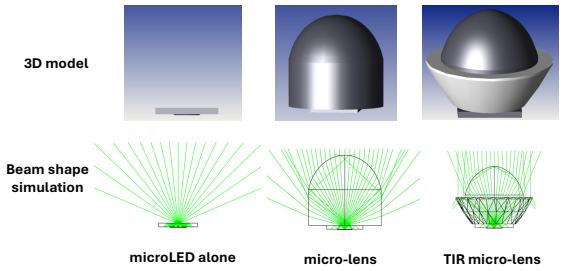


Figure 7: Ray-tracing simulation of microLED emitted light with no lenses (left), standard micro-lens (center), custom TIR micro-lens (right).

3.5 Electronics

In today's optical links, the electronic backend is a critical contributor to overall power consumption (§5), largely due to the presence of complex DSP circuitry required to compensate for transmission impairments at 100 Gbps speeds and beyond.

In MOSAIC, instead, we take advantage of the low channel speeds to drastically simplify the electronic design and achieve a very low-power solution. We adopted the following three design principles. First, each channel employs a simple non-return-to-zero (NRZ) coding scheme, consisting of just two levels (ON and OFF). Compared to the more complex 4-level signaling scheme (PAM-4) prevalent in today's links, NRZ requires a lower signal-to-noise ratio (SNR), has less stringent linearity requirements, and does not require expensive digital-to-analog (DAC) or analog-to-digital (ADC) converters. Due to the lower SNR requirements, no additional forward error correction (FEC) logic is needed, which further reduces complexity. Second, due to lower channel speeds, MOSAIC does not require any DSP logic but it only relies on *analog* equalization blocks to compensate for transmission impairments. Finally, by leveraging the ability to overprovision channels, we dedicate some of them to transmit the clock signal, thus avoiding the need for a full-fledged clock-and-data-recovery (CDR) circuit at the receiving end. We discuss this in more detail in §4.

4 Channel Overprovisioning

A key benefit of the WaS architecture is that *overprovisioning* redundant channels is relatively inexpensive, since the cost per channel represents only a small fraction of the total. MOSAIC leverages this overprovisioning to enhance reliability (§4.1) and reduce power consumption (§4.2 and §4.3).

4.1 Fault tolerance

Each channel in MOSAIC targets a Bit Error Rate (BER) before applying forward error correction (FEC) specified by the layer-2 protocols used in data centers, such as Ethernet and InfiniBand. For example, IEEE standard for Ethernet [23] specifies a pre-FEC BER of 2×10^{-4}. More specifically in the Ethernet physical layer (PHY), MOSAIC wide-and-slow conversion is effectively applied in the physical medium attachment (PMA) sublayer, and the physical coding sublayer (PCS) handles the FEC encoding and decoding. Since PCS sits higher in the PHY stack than PMA, all the independent channels including the redundant ones in MOSAIC are still under the FEC protection. Hence, the layer-2 link will expose the same post-FEC error characteristics and post-FEC BER guarantee as using the original narrow-and-fast architecture.

MOSAIC further leverages extra channels to increase link reliability. It maintains a small set of channels as hot spares, so when a channel is determined to have failed, it is replaced by one of these hot-spare channels. However, such hot sparing, by itself, is insufficient as any failure will lead to a short downtime.

To achieve near-zero downtime upon channel failure, MOSAIC adopts a two-layer approach that combines hot spares with a lightweight, power-efficient error correction code (ECC) such as Hamming coding across all data channels. This enables single-channel failure masking and localization, while triggering rapid switchover to a spare channel—all without exposing a link failure or exceeding pre-FEC BER thresholds. Notably, MOSAIC's ECC scheme operates at lower latency and overhead than layering an additional FEC on top of the existing link-layer FEC.

For each transmission cycle, MOSAIC transmits payload bits over k data channels, augmented by n redundancy channels. The k channels are partitioned into blocks of b channels; for each block, p parity bits are generated and transmitted via p additional channels, yielding a total of k+n channels, where $n = \frac{k}{b} \cdot p$. To correct a single-bit error per block, Hamming coding requires $2^p > b$+p. For example, with a channel rate of 2 Gbps, an 800 Gbps link would require $k = 400$ data channels. Assuming $b = 40$, $p = 6$ parity bits per block would be required to satisfy the above formula, leading to $n = 60$ additional channels and 460 channels in total (i.e., 15% overhead).

We evaluate the effectiveness of this design through both empirical experiments and large-scale simulations (§7), demonstrating that MOSAIC achieves reliability comparable to copper links.

4.2 In-band Control Plane

Overprovisioning also enables physically separated *in-band* control channels to support link operations (e.g., link training and negotiation, and network telemetry) without impacting data channel throughput. An example of how MOSAIC leverages these additional channels is to forward the clock signal. A common challenge in endpoint communication is ensuring the receiver's clock is synchronized with the transmitter's to accurately sample incoming signals. Traditionally, transceivers achieve this using a clock-and-data-recovery (CDR) circuit at the receiver end. This circuit extracts the clock signal from the incoming data stream using a phase-locked loop (PLL), a process that consumes significant power and area. In contrast, in MOSAIC, the transmitter directly sends its clock signal through a control channel, allowing the receiver to use this clock signal without a power-hungry CDR. This approach results in substantial power and area savings. Implementing this in conventional, NaF transceivers would be costly; for example, adding an extra channel to a 4-laser transceiver would increase the channel count, and thereby the cost, by 25%, whereas adding one channel to a 400-channel array would only result in a 0.25% increase.

4.3 Power Proportionality

Clock forwarding in MOSAIC not only eliminates the need for a CDR circuit but it can also contribute to reducing average power

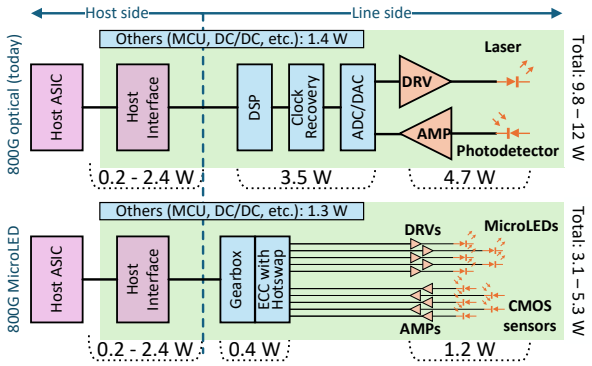


Figure 8: Power breakdown for one end of the link for today's optical links (top) and MOSAIC links (bottom).

consumption. Current CDR-based transceivers must continuously transmit data, even during idle periods using idle frames, leading to unnecessary power usage. By leveraging clock forwarding, MOSAIC avoids the need for constant transmission. Consequently, during idle periods, it is possible to deactivate some channels, thus aligning power consumption with actual data transmission. The transmitter in MOSAIC utilizes a simple FIFO (First-In, First-Out) queue for outgoing packets. When this queue is empty, channels are progressively deactivated. As the queue fills, channels are reactivated as needed, either until all are operational or the queue is emptied. An in-band protocol communicates these status changes to the receiver, ensuring smooth operation. This approach could be particularly beneficial for AI inference workloads, which are read-dominated and, hence, exhibit significant asymmetry in utilization across different links.

5 Power Consumption Analysis

In this section, we present our estimates for MOSAIC power consumption and compare it against today's optical links. We do not consider copper links as they are passive devices so their power consumption is essentially zero. In our analysis, we use a 10 m, 800 Gbps AOC as reference [2] because it is the latest generation for which specifications have been confirmed. We also extrapolate how power scales for future generations.

Power analysis. We report our power comparison in Fig. 8. Since the power consumption is the same for both ends of the link, for ease of illustration, we focus on only one end of the link (i.e., half of the total link power). We start by considering the power breakdown for the mainstream baseline [37]. The host interface is responsible for driving the signal from the optical module to the host (i.e., the switch ASIC or the NIC) over the PCB trace (see Fig. 5). This is independent of the optical technology used (e.g., laser or microLEDs) and it is mostly a function of the trace length. This power can vary from 2.4 W for pluggable modules to 0.2 W for co-packaged optics (CPO), in which the optical module is located next to the host itself (see CPO discussion in §9). Next, we have the digital backend (DSP, CDR, and ADC/DAC), which consumes 3.5 W. The analog frontend, comprising laser drivers (DRV) and receiver amplifiers (AMP), and lasers account for 4.7 W (photodetectors' power consumption is negligible). Finally, the on-board microcontroller unit (MCU) and DC/DC conversion contribute 1.4 W. In total, the aggregate power for today's optical links is 9.8-12 W depending on the host interface used (24 W for the full AOC including both ends).

In contrast, the power of MOSAIC's optical link is 3.1-5.3 W (10.6 W for the full AOC), i.e., 56-68% lower than mainstream baseline. The digital backend only consumes 0.4 W due to its lack of complex functions such as DSP, ADC/DAC, and CDR (§3.5). It only has simple gearboxing and lightweight failure protection. The analog frontend and microLEDs are also significantly lower power than their mainstream counterparts due to the lower speeds and the lower power consumption of microLEDs, amounting to 1.2 W. Finally, we include 1.3 W for the remaining MCU and DC/DC conversion (the figure is slightly lower than mainstream because DC/DC is proportional to the module power).

Scaling to 1.6 Tbps and beyond. Looking ahead, we expect the absolute power difference between mainstream and MOSAIC to grow at each generation due to the challenges of continuing to scale bandwidth per channel. While 1.6 Tbps optical links are not commercially available yet, initial guidance from leading manufacturers suggests a power consumption of 23-25 W per transceiver [6]. In contrast, MOSAIC can scale out to higher speeds by doubling the number of channels every generation, which would lead to 10.6 W per transceiver. In fact, we expect that as microLEDs further mature, it should be possible to achieve even lower power consumption.

Cost. A detailed cost analysis is harder to prepare and it is outside the scope of this work. Cost breakdowns are typically not publicly available and final prices depend on complex business arrangements and volumes. However, beside operational expenditure (*OpEx*) reduction due to lower power consumption, we speculate that the MOSAIC innovations also result in lower fabrication cost and, hence, ultimately lower capital expenditure (*CapEx*). In particular, we envision three main cost reduction vectors. First, the lack of advanced DSP or ADC/DAC functionality means that the electronic backend does not need to be manufactured with small process node (e.g., 5 nm or 7 nm) as with today's optical links. Second, the microLED array is significantly cheaper to fabricate and to integrate with a CMOS die. Third, the use of imaging fiber with thousands of cores simplifies alignment. Finally, the ability to overprovision channels (§4.1) also helps to improve overall yield and, hence, to reduce costs.

6 Prototype

To validate the design of MOSAIC, we implemented an end-to-end prototype comprising 100 channels, each capable of transmitting at 2 Gbps. In the following, we first review the key building blocks of our prototype and then we summarize how this design can be extended to build an 800 Gbps link in a standard QSFP form factor.

100-channel prototype. We partnered with microLED and CMOS suppliers to fabricate bespoke 10×10 array microLED and CMOS sensor dies, which we wire-bonded onto a printed circuit board (PCB) as shown in Fig. 9 (a). The microLEDs require no structural changes compared to those used for displays, except for the need to control each pixel individually. For the receiver array, we worked with a CMOS supplier to fabricate the array shown in Fig. 9 (b). To improve coupling efficiency, we manufactured a set of custom

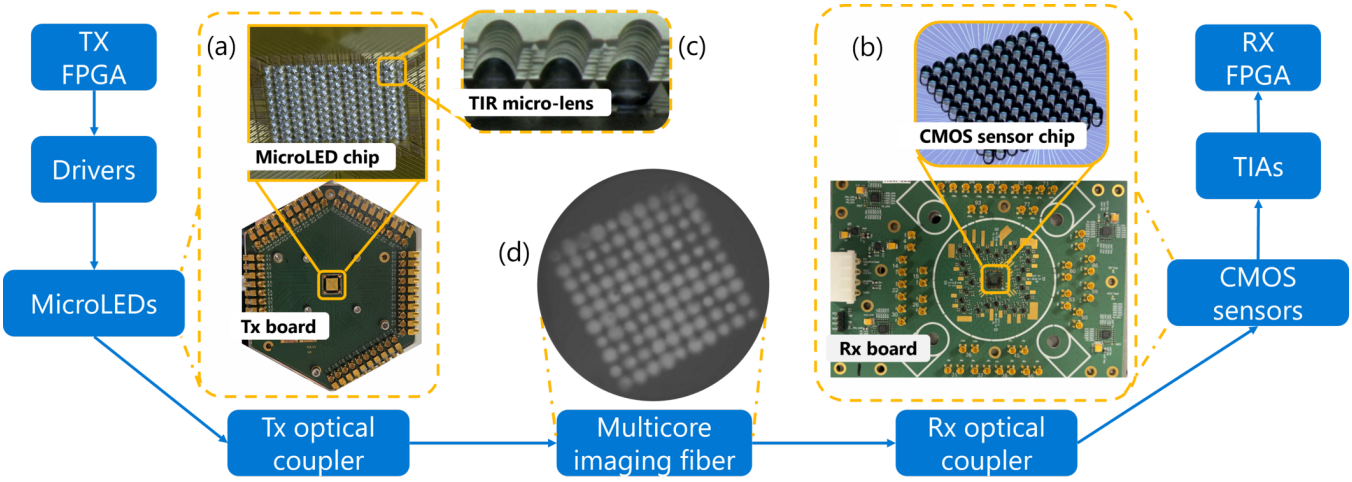


Figure 9: Prototype setup. (a) transmitter and (b) receiver PCB with fabricated microLEDs and CMOS sensors (insets). (c) Fabricated TIR micro-lens array. (d) 100-channel image at the output of the imaging fiber.

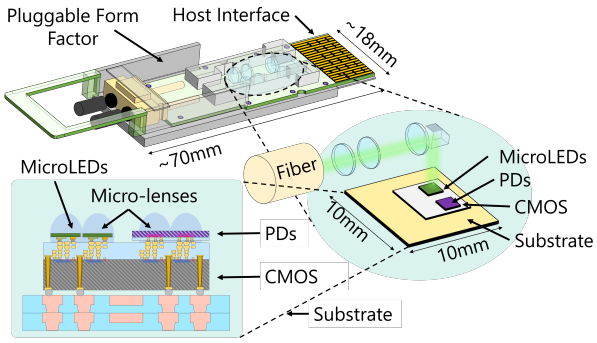


Figure 10: An illustration of the MOSAIC pluggable module (top) and the cross-sectional view of the chip packaging (bottom left).

micro-lenses on top of both microLEDs and CMOS sensor array, using the TIR design presented in §3.4.

On the transmitter side, we employed one HTG-940 FPGA [20] to emulate our electronic backend (§3.5). The FPGA generates the streams of data, using ON-OFF (NRZ) signaling, feeding a set of discrete drivers that modulate the microLEDs. The microLED light is emitted through the micro-lenses and then coupled into the imaging fiber using a combination of discrete lenses. On the receiver side, the light from the fiber is coupled into the CMOS sensor array, also leveraging both discrete lenses and the micro-lenses. The CMOS sensor array converts the optical signal back to the electrical signal that is further amplified by discrete transimpedance amplifiers (TIAs) before entering the receiver FPGA board. In our experiments (§7.1), following standard industry practices, we used pseudorandom binary sequences (PRBS) to measure bit error rate (BER). In addition, we have also connected FPGA boards to two server NICs to send Ethernet and InfiniBand traffic over our prototype. This is possible because MOSAIC is protocol-agnostic and operates at the physical layer, without any connection termination or traffic inspection.

Pluggable module design. In our testbed prototype, we had to compromise on the channel count and performance due to several prototyping constraints, e.g., wire bonding and discrete, bulky lenses and electronics. For the production module, we can take advantage of miniaturization and integration to avoid these limitations. We envision the design sketched in Fig. 10, whose feasibility has been confirmed by our suppliers. Compared to our prototype, this presents several advantages, resulting in better performance and efficiency. First, the use of integrated lenses and custom fiber couplers drastically improves coupling efficiency and launching conditions, which results in lower modal dispersion. Second, all drivers and TIAs are included in a single CMOS chip. Both microLEDs and CMOS sensor arrays can be vertically bonded on top of the chip. This configuration improves overall performance by *i)* considerably shortening the length of electrical traces and by *ii)* enabling the use of smaller microLEDs with smaller pitch by avoiding the wire bonding pitch constraints. Our analysis indicates that 10 μm microLEDs are sufficient to meet our transmission target while allowing more than 460 channels to be packed within a single fiber. This would be sufficient to realize an 800 Gbps transceiver, including more than 10% redundant channels. Higher speed transceivers (e.g., 1.6 Tbps or 3.2 Tbps) could be realized by using the same microLED sizes but with smaller pitches, multiple fibers (today's 800 Gbps AOCs already use 16 fibers) and/or by increasing modulation speed (e.g., to 4 or 8 Gbps per microLED).

7 Evaluation

In the following, we first demonstrate MOSAIC's performance using our 100-channel prototype (§6), and then we evaluate the behavior at scale through large-scale simulations.

7.1 Prototype Experiments

For our prototype experiments, we consider the same setup shown in Fig. 9. The FPGA on the transmit side generates PRBS and sends it over the MOSAIC prototype. Upon receipt of the data, the FPGA

connected to the CMOS sensor array compares the received stream with the expected PRBS to measure the bit error rate (BER), which we use as a key metric of interest throughout this analysis. Our target is to achieve a BER $< 2 \times 10^{-4}$, which is the FEC threshold adopted by Ethernet and InfiniBand standards for a link to be considered error-free [23]. Unless otherwise specified, we consider a data rate per channel of 2 Gbps and a transmission distance of 20 m. In §7.2, we estimate that MOSAIC can achieve longer distances (up to 50 m) by avoiding the limitations of our current prototype. Due to the complexity of manually wiring 100 drivers and TIAs, in our transmission experiments we only used 25 channels at a time but any subset of the 100 microLEDs could be used.

System performance. We start our analysis by measuring the individual BER of each channel. Fig. 11 shows the cumulative distribution of the measured BER for the 25 channels. *All* of them are below the FEC threshold with a median BER $< 2 \times 10^{-8}$. This demonstrates that despite the limitations of our current prototype (especially the use of large microLEDs, wire bonding, and discrete components), we can still meet the desired performance. While most of the channels have very low BER (BER $\leq 1 \times 10^{-6}$), the worst BERs (albeit still below the FEC threshold) are experienced by the channels at the edge of the 2D channel array (BER $\leq 4 \times 10^{-5}$). This is expected because the loss of the imaging fiber is 1 dB higher for the cores at the edge as their light confinement is lower than that of the central ones. While this has an impact on the prototype BER, we do not anticipate this will be a concern for our envisioned transceiver because by having much smaller microLEDs we can ensure that we do not utilize the edge cores as discussed in §7.2.

Transmission speed and distance. We evaluated the impact of varying channel rates from 1.3 Gbps to 2 Gbps on BER as the fiber length increases from 10 m to 30 m. Fig. 12 demonstrates that our prototype can sustain 2 Gbps transmission over 20 m with BER $< 10^{-6}$, i.e., more than two orders of magnitude below the FEC threshold. At 30 m, the BER at 2 Gbps slightly exceeds the FEC threshold, requiring a reduction to 1.6 Gbps to meet FEC requirements. This is due to the limitations of our current prototype, e.g., wire bonding and discrete electronics and optics, and we expect a production-quality pluggable module (Fig. 10) to achieve greater BER margin thanks to tighter optoelectronic integration and improved microLEDs and coupling (§6). Indeed, as discussed in §7.2 (Fig. 15), simulations indicate that the pluggable module should sustain 2 Gbps transmission over 50 m with BER $< 10^{-6}$.

The results in Fig. 12 further show that decreasing the data rate or transmission distance significantly improves BER. This improvement is due to reduced chromatic dispersion at lower speeds and shorter distances. While laser sources typically have very narrow linewidths (<1 pm), microLEDs emit over a much broader spectrum (>10 nm), making chromatic dispersion more significant. At 1.3 Gbps per channel and below, we observe error-free transmission (BER $< 10^{-12}$) up to 20 m (and up to 10 m for 2 Gbps).

These observations suggest that, by lowering channel speed and/or link distance, MOSAIC can be deployed even in scenarios without host-side FEC. The trade-off is that maintaining the same aggregate speed would require more channels (e.g., at 1.3 Gbps per channel, an 800 Gbps MOSAIC transceiver would require 616 channels), which

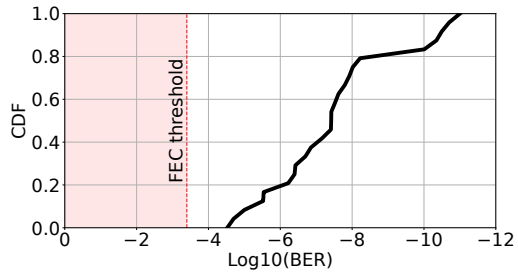


Figure 11: Cumulative distribution of BER across 25 channels (20 m distance and 2 Gbps per channel).

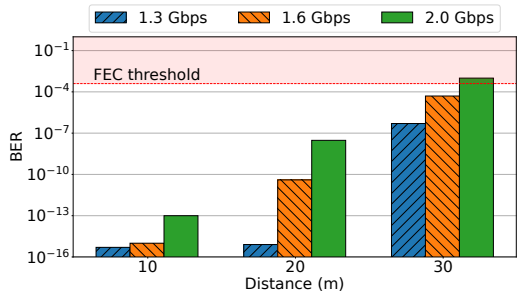


Figure 12: BER for different transmission speeds per channel and distances. Our prototype meets FEC requirements when transmitting up to 2 Gbps per channel over 20 m (or 1.6 Gbps at 30 m).

could be achieved by reducing the microLED pitch or using larger-diameter fiber. In return, NIC and switch chips could save silicon area and power, and avoid the latency overhead of FEC (∼100 ns). As discussed above, we expect the pluggable module to overcome current limitations, making error-free transmission feasible at higher per-channel rates and/or longer distances.

Fault tolerance. After validating MOSAIC's BER performance, we evaluate the effectiveness of the dual-layered fault-tolerance approach described in §4.1. In addition to experiments with the 100-channel MOSAIC prototype, we set up a live 10 Gbps Ethernet link over a 10 m span using a 12-channel MOSAIC prototype connecting two network interface cards. This setup allocated six channels for data transmission (∼ 1.7 Gbps per channel) to support the 10 Gbps data rate.

According to the Hamming coding requirements in §4.1, protecting $b = 6$ data bits requires at least $p = 4$ parity bits for single-error correction. In practice, our FPGA IP implementation provides single-error correction and double-error detection (SECDED) [1], which necessitates one additional parity bit. Thus, the prototype uses $b = 6$ data channels and $p = 5$ parity channels. This relatively high overhead is a consequence of the short codeword (b+p = 11) used in this experiment, dictated by prototype limitations. In a production system, longer codewords (e.g., b+p = 46 as discussed in §4.1) significantly reduce the required number of additional channels, yielding only a 15% overhead in our example.

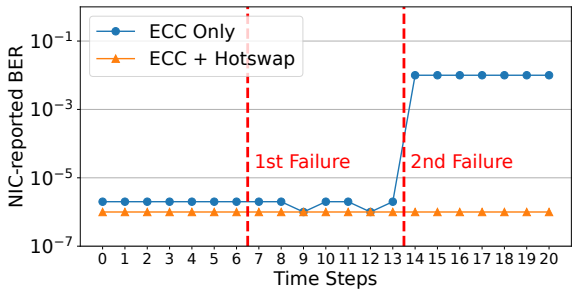


Figure 13: BER monitored from NIC running 10 Gbps Ethernet with two channel failures.

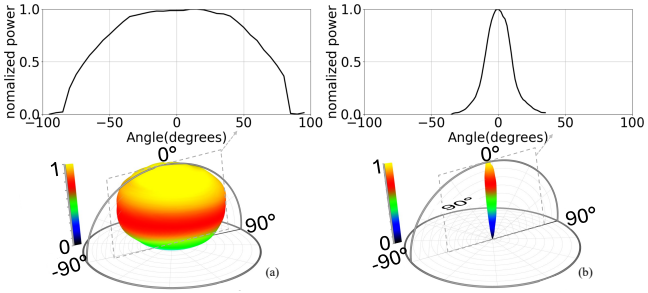


Figure 14: Measured angular beam shape before (a) and after (b) printing the TIR micro-lenses.

We set up an experiment in which we artificially failed two channels in real time. We compare the behavior of MOSAIC (*ECC + Hotswap*) with the system using only ECC but no hotswapping (*ECC Only*). The results in Fig. 13 show that if only ECC is used, the BER remains stable after the first failure but it cannot protect against the second failure. The transceiver would need to be replaced after the second failure. If we only use hotswapping without ECC, we will need extra channel-failure detection logic and cannot prevent a link failure from the first failure. In contrast, when both ECC and hotswapping are used, the BER is not impacted even after the second failure. The reason is that ECC is able to protect against the failure happening since time step 7, and the hotswapping has quickly occurred during the time between time step 7 and 14.

Micro-optics. In the prototype, the functionality of the micro-optics was experimentally validated by assessing coupling efficiency and directionality using goniometer measurements. The output power of the microLED was measured at various detection angles to visualize the beam shape. Fig. 14 illustrates the measured beam shape: (a) before the TIR lens print, and (b) after the TIR lens print. Without the TIR micro-lens, the microLED emitted light in a $\pm 90^\circ$ cone, creating significant challenges for coupling to the fiber and leading to high levels of crosstalk, as light from one microLED could easily interfere with adjacent ones. After printing the TIR micro-lenses, the beam shape was successfully collimated into a $\pm 12^\circ$ cone, resulting in improved fiber coupling and suppressed crosstalk between channels.

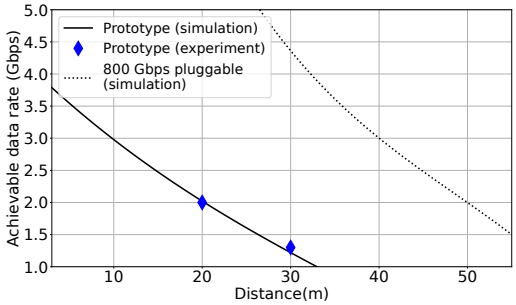


Figure 15: Channel data rate versus distance simulations to achieve BER $< 10^{-6}$ for our current prototype and for 800 Gbps pluggable system. The experimental data are based on the measurements in Fig. 12.

7.2 Simulation Results

We now focus on assessing the behavior of MOSAIC at scale, targeting the 800 Gbps link with 460 channels described at the end of §6. First, we evaluate how to achieve longer transmission distances and higher data rates. Second, we evaluate the impact of reliability with hundreds of channels. Finally, we evaluate whether the pitch and size of microLEDs combined with micro-optics are adequate to ensure high coupling efficiency with tolerable crosstalk.

Distance and speed scalability. As discussed in §7.1, the reach and achievable speed of a MOSAIC link are dictated by the characteristics of the microLED and CMOS sensors, with noise and chromatic dispersion as the primary limiting factors. Improving transmission speed and distance will require advancements in both microLED and CMOS sensor performance. For microLEDs, manufacturers expect improvements in efficiency, directionality, and spectral width. For the CMOS sensors, tighter integration with CMOS chips (e.g., TIAs) should significantly enhance link sensitivity compared to our prototype. Taking these anticipated improvements into account, Fig. 15 shows simulation results for the achievable data rate and transmission distance with an 800 Gbps pluggable module, assuming a target BER $< 10^{-6}$ to ensure sufficient margin with respect to the FEC threshold. For reference, we also include simulated prototype performance cross-validated against the experimental measurements in Fig. 12. The results indicate that a MOSAIC pluggable module could achieve 2 Gbps per channel over 50 m (or more than 8 Gbps per channel for distances up to 10 m).

Reliability. We simulate the effectiveness of MOSAIC's dual-layered reliability design at large scale. First, we evaluate the resilience when relying on ECC code alone. Similar to the baseline "ECC-only" in Fig. 13, this can only tolerate up to 2% of failed channels (results not shown for space reasons). Next, we turn our attention to evaluating the combination of ECC and hotswapping. We opted for failures in time (FIT) as the metric of interest since this is often used in the industry. A FIT is defined as one failure in a billion hours. As a reference point, typically short-reach optical links exhibit FIT values of few 100s while passive electrical cables have very low values of FIT (< 10). In the graph, we report estimated FIT values for an 800 Gbps link, assuming both a conservative FIT of 1 per microLED and a more typical FIT of 0.1 per microLED. Fig. 16 plots the

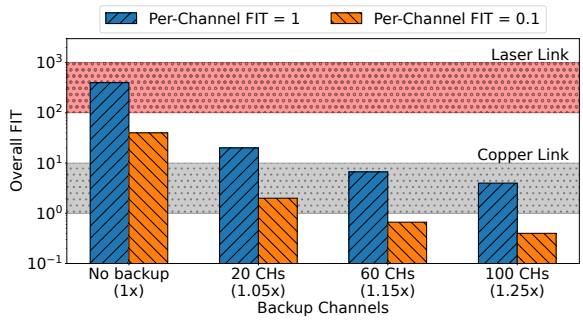


Figure 16: The overall FIT of an 800 Gbps MOSAIC link for per-channel FIT = 1 and 0.1. Shaded areas represent the typical FIT values observed for laser-based optical links and copper links.

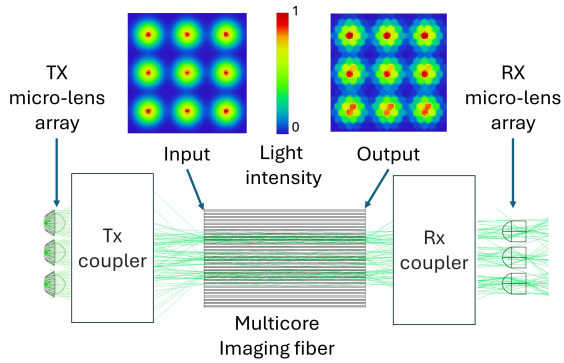


Figure 17: Optical simulation of an 800 Gbps link with 460 channels (3×3 subset shown for ease of visualization).

results for a MOSAIC instantiation with 400 active data channels and an arbitrary number of redundant channels, ranging from zero ("No backup") to 100 ("1.25x"). Not surprisingly, if no redundant channels are present in the conservative setting, then the FIT matches or exceeds that of mainstream optical links because one microLED failing out of 400 channels is sufficient to consider the link failed. On the other hand, even just 5% redundant channels are sufficient to bring the FIT below 20. This is already one order of magnitude better than mainstream optical links. If additional redundant channels are included, or if we use more reliable microLED channels, the FITs become even lower, matching the failure rates of electrical cables. Since the marginal cost of redundant channels is very small, this provides a practical way to achieve very high reliability without impacting costs.

Coupling efficiency and crosstalk. Next, we want to verify that when packing 460 channels per fiber, we still ensure that *i)* enough light is coupled into the fiber and that *ii)* crosstalk is negligible. We use Zemax, a commercial ray-tracing engine and de facto standard in the industry. Zemax enables modeling of light rays and quantifying the fraction of light emitted by the microLEDs that is captured by the CMOS sensors after passing through the fiber (*coupling efficiency*), as well as the amount of light that spills into neighboring pixels (*crosstalk*). For ease of visualization, we present a 3×3 scaled-down version of the analysis in Fig. 17. The insets in the figure show that the beams are clearly separated with virtually zero crosstalk. Further, we could verify that the received optical power is comparable to that observed in the 100-channel prototype, indicating that there will be enough margin for an 800 Gbps link. Higher bandwidth could be achieved by further optimization of the source, fiber and detection arrays to conceivably scale out the number of channels to 800 and beyond.

8 Discussion

MOSAIC combines the best properties of copper and optical links, enabling networks that are long-reach, low-power, and highly reliable, and can scale to future bandwidth demands. Historically, each 10× increase in network bandwidth (and corresponding reduction in latency) has driven a new era of distributed computing, from FTP and email in the 1970s to today's epoch of machine learning and resource disaggregation [64]. By enabling the next step-change in bandwidth, MOSAIC could not only unlock recently proposed architectures but also enable the community to explore entirely new designs for networks, compute, memory and clusters. In this section, we discuss some of these broader implications. We hope that by highlighting these new directions, this work will spur future innovation at the intersection of networks and systems.

Network design. The limited reach of copper links imposes significant constraints on network architecture and topology. For example, top-of-rack (ToR) switches are usually deployed because copper cables cannot span longer distances. Similarly, while 3D torus topologies are suboptimal in terms of bisection bandwidth, they are often used in high-performance computing (HPC) clusters (e.g., Google TPU cluster [26] or Amazon Trainium [48]) due to their compatibility with short copper interconnects.

MOSAIC challenges this status quo by providing an interconnect with the power and reliability of copper links but with a reach of tens of meters. This enables network designs that were previously infeasible. For instance, eliminating ToR switches becomes practical, allowing servers to connect directly to row or end-of-row switches [12]. This reduces both network latency and hardware costs, while improving reliability by removing the ToR as a single point of failure. Traditional optical links negated these gains due to their higher power consumption, cost and reliability overheads. Additionally, MOSAIC makes fully non-blocking topologies more viable, potentially simplifying congestion control protocols (e.g., [10, 16, 34]). Longer-reach links also make advanced topologies, such as multi-dimensional torus, dragonfly, and hypercubes [21], practical, as designers are no longer constrained by short-reach copper or the high cost and complexity of current optical solutions. Overall, MOSAIC expands the design space for custom, application-optimized networks.

GPU design. The exponential growth of AI/ML workloads and the slowdown of Moore's Law have pushed GPU vendors toward multi-die solutions and ever-larger package sizes (e.g., TSMC CoWoS [58]). For example, the upcoming NVIDIA Rubin Ultra GPU will feature four dies per package [35], AMD MI450X supports up to eight [27], and Cerebras WSE-3 integrates a full wafer-scale chip [18]. These trends are driven by the need for high-speed, low-power die-to-die interconnects, which are currently only possible via millimeter-scale, on-package copper traces. This approach increases manufacturing

complexity and cost, and is emerging as a fabrication bottleneck for AI accelerator production [49]. By providing a fast, reliable, and low-latency optical fabric, MOSAIC enables new possibilities in GPU design. For example, it could facilitate the disaggregation of large, complex multi-die packages into smaller, single-die "LiteGPUs" interconnected by a low-power, microLED-based optical network [7].

Memory design. Similar packaging constraints also impact memory design, leading to a dependence on expensive 3D-stacked high-bandwidth memory (HBM) to maximize capacity and bandwidth density. However, scaling HBM is increasingly difficult due to the physical limits of individual DRAM layer scaling and the growing complexity and thermal challenges associated with deeper stacks [53]. Thus, GPU memory remains capped (e.g., 192GB for NVIDIA B200 GPU [39]), which limits utilization, especially for I/O-intensive inference workloads [50].

Memory disaggregation, i.e., decoupling compute and memory across nodes [13], has long been proposed as a solution to these scaling challenges, but adoption has been hampered by the lack of interconnects that are simultaneously low-power, low-latency, and highly reliable. Latency is especially critical for memory links, where even modest delays can stall compute cores. Current optical links rely on strong FEC and DSP, adding up to 100 ns of latency, which would more than double typical HBM access times. In contrast, because MOSAIC requires neither FEC nor DSP, it incurs only a few nanoseconds of latency, while maintaining low power and high reliability. This makes it a promising candidate to finally enable off-package memory without being constrained by on-package area. The benefits are twofold: *i)* increasing the memory capacity available to GPUs and *ii)* reducing the reliance on expensive HBM and complex 3D stacking, thus creating new opportunities for using novel, more efficient memory technologies [29, 60].

Cluster design. The extended reach and low cost and power of MOSAIC has significant implications for the design of AI/ML clusters as well. Currently, tensor parallelism and other scaling techniques are constrained by the limited bandwidth and short reach of electrical links (§2). By enabling larger GPU clusters with high-bandwidth low-latency connections, MOSAIC has the potential to accelerate training and inference as well as improve resource utilization by reducing fragmentation [47]. This could also enable rethinking commonly-used parallelization strategies [5] and collective optimizations [31] in networks where bandwidth is abundant. Furthermore, in conjunction with resource disaggregation, MOSAIC could make possible a *physical* form of "elastic computing", where resources are dynamically aggregated at runtime to match workload requirements, driving greater efficiency and flexibility.

Limitations and future work. MOSAIC intentionally leverages technologies from the consumer space, like microLEDs and imaging fibers, to make WaS optical links practical and deliver its performance benefits. However, adapting the manufacturing lines of these technologies for data center environments–and developing the associated ecosystem–poses unique challenges. Further, while MOSAIC opens up a range of promising opportunities for rethinking network, GPU, memory and cluster design, realizing these visions requires much more work to co-optimize the link and system architecture technology, including advances in several areas like packaging, deployment, system integration, and reliability at scale.

9 Related Work

Silicon Photonics. Silicon Photonics refers to the ability to manufacture some optical components (e.g., modulators or mux/demux but not lasers) in silicon, using processes compatible with CMOS fabrication [9, 28, 55]. This has attracted much interest from companies that were already invested in the CMOS ecosystem, e.g., GlobalFoundries or TSMC. While this technology can enable some cost savings because of CMOS ecosystem, overall it still relies on a NaF architecture and, hence, it still suffers from the same issues in terms of power, scalability, and reliability that we discussed in §2.

Co-packaged Optics (CPO). In this paper, we have focused on pluggable transceivers as these are the predominant solution adopted in the industry. Pluggables provide high flexibility because they can be selected independently from the NIC/switch vendor, enabling horizontal integration across different suppliers. On the negative side, though, they require long electrical traces to transmit the signal from the host die (e.g., NIC or GPU) to the front panel. This increasingly consumes a nontrivial amount of power (§5). CPO can circumvent this issue by integrating optical transceivers directly onto the same package as the host die [9, 33], possibly saving up to 25-30% of the power according to recent industry estimates [62]. MOSAIC is fully compatible with such arrangement. In fact, as shown in Fig. 8, if CPO is adopted, the benefits for MOSAIC should be even higher because it can take advantage of the low data rate of chip-to-chip interconnect to directly modulate microLEDs without requiring high-speed conversion as is the case for incumbent technology.

MicroLED-based communication. While microLEDs have primarily been developed for display applications, there have been a few proposals in the literature exploring their use for free-space communication [11] and short-reach chip-to-chip links [4, 51, 52]. In contrast, MOSAIC targets much longer reach over fiber (up to 50 m) to enable rack-to-rack connectivity in data centers. This required addressing a different set of challenges related to fiber impairments (such as chromatic and modal dispersion) and coupling losses, which in turn required novel microLED optimizations and system design choices. MOSAIC further introduces mechanisms for high reliability, improved alignment tolerance, and a new micro-lens design for efficient fiber coupling. Finally, while prior work has mostly focused on single-channel demonstrators operating over free-space or very short waveguides, we evaluate MOSAIC with a 100-channel prototype over up to 30 m of fiber.

10 Conclusions

In this paper, we introduced MOSAIC, a novel optical link technology that breaks today's fundamental trade-off between reach, power, and reliability. By leveraging a WaS architecture with microLEDs, it achieves long reach, low power, and high reliability while remaining fully compatible with existing network architectures and offering a practical path for scalability to future network generations.

In the past, step changes in network technologies have enabled entirely new, and often unforeseen, classes of applications and workloads. We aspire to trigger the next wave of innovation with MOSAIC-based data center networks. While we have provided some preliminary examples in this paper, we hope this work will stimulate further discussion and innovation within the community.

Acknowledgments

We would like to thank our colleagues at Microsoft Research for their support of the MOSAIC project over the years. In particular, we thank Neeltje Berger, Miguel Castro, Nathanaël Cheriere, James Clegg, Doug Kelly, Teresa LaScala, Dushyanth Narayanan, Aditya Nori, Jacob Nelson, Ed Nightingale, Francesca Parmigiani, Dan Ports, John Romualdez, Adam Smith, Lex Story, Jonathan Westcott, Charles Whittaker, Xingbo Wu, and our former intern Alberto Pepe. We are also grateful to Doug Burger, Peter Lee, and Abigail Sellen from the MSR leadership team for their ongoing guidance. This work was made possible through close partnership with the Azure Networking, Azure Hardware Systems and Infrastructure (AHSI), and Microsoft 365 teams. We especially appreciate the support and feedback from Rahul Agarwal, Rich Baca, Christian Belady, Charlie Boecker, Rani Borkar, Dave Bragg, Doug Carmean, Adrian Caulfield, Howard Chen, Saurabh Dighe, Eyran Eylon, Binbin Guan, Ashwin Gumaste, Ram Huggahalli, Fotini Karinou, Salman Khaleghi, Jim Kleewein, Gopi Kumar, Erica Lan, Yingo Lin, Ashley Llorens, Dave Maltz, Dimitry Melts, Daniel Mohaghegh, Claire Szu Ma, Sharad Mehrotra, Tony Pearson, Saravan Rajmohan, Jesus Rios, Victor Rühle, Mark Russinovich, Winston Saunders, Andrew Silverman, Sriram Srinivasan, Marc Tremblay, Franco Tu, Matt Tuggle, Jaster Yen, and Yawei Yin. We also acknowledge the valuable partnership and support of our suppliers throughout this work. Finally, we thank our shepherd Andrew Moore and the anonymous SIGCOMM reviewers for their detailed feedback and suggestions.

References

[1] AMD. 2017. ECC v2.0 LogiCORE IP Product Guide (PG092). https://docs.amd.com/v/u/en-US/pg092-ecc

[2] Arista Networks 2025. 10M Compatible 800G QSFP-DD Active Optical Cable. https://www.fs.com/uk/products/229643.html?attribute=102708&id=3721202

[3] MicroLED Industry Association. 2023. MicroLED Smartwatch Displays in 2023, LED Cost Analysis. https://www.microledassociation.com/wp-content/uploads/2023/01/MicroLED-smartwatch-whitepaper-2023-01.pdf.

[4] N. Bamiedakis, X. Li, J. J. D. McKendry, E. Xie, R. Ferreira, E. Gu, M. D. Dawson, R. V. Penty, and I. H. White. 2015. Micro-LED-based guided-wave optical links for visible light communications. In *ICTON*.

[5] Tal Ben-Nun and Torsten Hoefler. 2019. Demystifying Parallel and Distributed Deep Learning: An In-depth Concurrency Analysis. *Comput. Surveys* 52, 4 (2019).

[6] Broadcom 2025. Broadcom Extends 200G/lane DSP PHY Leadership for Next-Generation AI Infrastructure. https://www.broadcom.com/company/news/product-releases/62986

[7] Burcu Canakci, Junyi Liu, Xingbo Wu, Nathanael Cheriere, Paolo Costa, Sergey Legtchenko, Dushyanth Narayanan, and Ant Rowstron. 2025. Good things come in small packages: Should we build AI clusters with Lite-GPUs?. In *HotOS*.

[8] Reza Chaji. 2024. Why MicroLED is Poised to Revolutionize Wearable Technology. https://www.microled-info.com/why-microled-poised-revolutionize-wearable-technology.

[9] Qixiang Cheng, Chen Sun, Mark T. Wade, Yunsup Lee, Mark J. Byrd, Jason S. Orcutt, Rajeev J. Ram, Vladimir Stojanović, and Milos Popović. 2018. Recent Advances in Optical Technologies for Data Centers: A Review. *Optica* 5, 11 (2018).

[10] Inho Cho, Keon Jang, and Dongsu Han. 2017. Credit-Scheduled Delay-Bounded Congestion Control for Datacenters. In *SIGCOMM*.

[11] Ricardo X. G. Ferreira, Enyuan Xie, Jonathan J. D. McKendry, Sujan Rajbhandari, Hyunchae Chun, Grahame Faulkner, Scott Watson, Anthony E. Kelly, Erdan Gu, Richard V. Penty, Ian H. White, Dominic C. O'Brien, and Martin D. Dawson. 2016. High Bandwidth GaN-Based Micro-LEDs for Multi-Gb/s Visible Light Communications. *IEEE Photonics Technology Letters* 28, 19 (2016).

[12] FS.com 2022. Top of Rack and End of Row: What's the Difference? https://www.fs.com/blog/top-of-rack-and-end-of-row-whats-the-difference-2934.html

[13] Peter X. Gao, Akshay Narayan, Sagar Karandikar, João Carreira, Sangjin Han, Rachit Agarwal, Sylvia Ratnasamy, and Scott Shenker. 2016. Network Requirements for Resource Disaggregation. In *OSDI*.

[14] Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, and Kurt Keutzer. 2024. AI and Memory Wall. *IEEE Micro* 44, 3 (2024).

[15] Gregory Haley. 2023. Nanoimprint Finally Finds Its Footing. https://semiengineering.com/nanoimprint-finds-its-footing-in-photonics/.

[16] Mark Handley, Costin Raiciu, Alexandru Agache, Andrei Voinescu, Andrew W. Moore, Gianni Antichi, and Marcin Wójcik. 2017. Re-architecting Datacenter Networks and Stacks for Low Latency and High Performance. In *SIGCOMM*.

[17] Reece Hayden and Paul Schell. 2024. *Opportunities and Challenges for Compute Express Link (CXL)*. Technical Report. ABI Research. https://computeexpresslink.org/wp-content/uploads/2024/11/CR-CXL-101_FINAL.pdf

[18] Congjie He, Yeqi Huang, Pei Mu, Ziming Miao, Jilong Xue, Lingxiao Ma, Fan Yang, and Luo Mai. 2025. WaferLLM: A Wafer-Scale LLM Inference System. In *OSDI*.

[19] John L. Hennessy and David A. Patterson. 1990. *Computer Architecture: A Quantitative Approach*. United States.

[20] HitechGlobal 2024. HTG-940: Xilinx Virtex UltraScale+ QUAD FMC+ Development Platform. https://www.hitechglobal.com/Boards/UltraScale+QuadFMC+.htm

[21] Torsten Hoefler. 2016. Network Topologies for Large-Scale Compute Centers: It's the Diameter, Stupid!. In *IEEE High-Performance Interconnects (HOTI)*.

[22] Jensen Huang. 2024. GTC March 2024 Keynote. https://www.youtube.com/live/Y2F8yisiS6E?feature=shared&t=2867.

[23] IEEE 2022. IEEE 802.3ck-2022. https://standards.ieee.org/ieee/802.3ck/7322/

[24] Tetsuya Iizuka. 2015. CMOS Technology Scaling and Its Implications. In *Digitally-Assisted Analog and Analog-Assisted Digital IC Design*, Xicheng Jiang (Ed.). Cambridge University Press, Cambridge, UK.

[25] Intel Corporation. 2021. *Accelerating Innovation Through AIB*. Technical Report. Intel Corporation. https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/accelerating-innovation-through-aib-whitepaper.pdf Accessed: 2024-02-02.

[26] Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. 2023. TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings. In *ISCA*.

[27] Sarfraz Khan. 2025. AMD Instinct MI400 Spotted in Latest Patches, Will Feature Up to 8 Chiplets on Dual Interposer Dies. Wccftech. https://wccftech.com/amd-instinct-mi400-spotted-feature-up-to-8-chiplets-on-dual-interposer-dies/

[28] Mohammad Khani, Manya Ghobadi, Mohammad Alizadeh, Zihuai Zhu, Meir Glick, Keren Bergman, and Amin Vahdat. 2021. SiP-ML: High-Bandwidth Optical Network Interconnects for Machine Learning Training. In *SIGCOMM*.

[29] Sergey Legtchenko, Ioan Stefanovici, Richard Black, Ant Rowstron, Junyi Liu, Paolo Costa, Burcu Canakci, Dushyanth Narayanan, and Xingbo Wu. 2025. Storage Class Memory is Dead, All Hail Managed-Retention Memory: Rethinking Memory for the AI Era. In *HotOS*.

[30] Elaine Liew, Taka Aki Okubo, Toshio Sudo, Toshihiro Hosoi, Hiroaki Tsuyoshi, and Fujio Kuwako. 2014. Signal transmission loss due to copper surface roughness in high-frequency region. In *IPC APEX EXPO*.

[31] Xuting Liu, Behnaz Arzani, Siva Kesava Reddy Kakarla, Liangyu Zhao, Vincent Liu, Miguel Castro, Srikanth Kandula, and Luke Marshall. 2024. Rethinking Machine Learning Collective Communication as a Multi-Commodity Flow Problem. In *SIGCOMM*.

[32] Marvell 2024. Industry's 1st 1.6T AEC DSP for Accelerated Infrastructure Copper Connections. https://www.marvell.com/content/dam/marvell/en/company/media-kit/marvell-alaska-a-1-6t-dsp-for-aec-media-deck.pdf

[33] Cyriel Minkenberg, Rajagopal Krishnaswamy, Aaron Zilkie, and David Nelson. 2021. Co-packaged datacenter optics: Opportunities and challenges. *IET Optoelectronics* 15, 2 (2021).

[34] Behnam Montazeri, Yilong Li, Mohammad Alizadeh, and John Ousterhout. 2018. Homa: A Receiver-Driven Low-Latency Transport Protocol Using Network Priorities. In *SIGCOMM*.

[35] Timothy Prickett Morgan. 2025. *Nvidia Draws GPU System Roadmap Out To 2028*. Technical Report. NextPlatform. https://www.nextplatform.com/2025/03/19/nvidia-draws-gpu-system-roadmap-out-to-2028/

[36] Manuel Mota. 2023. UCIe: An Open Standard Interface for Chiplet-Based Multi-Die Systems. In *Chiplet Summit 2023*. Synopsys.

[37] Radhakrishnan Nagarajan, Agustin Martino, Damian A. Morero, Lenin Patra, Christian Lutkemeyer, and Mario A. Castrillón. 2024. Recent Advances in Low-Power Digital Signal Processing Technologies for Data Center Applications. *Journal of Lightwave Technology* 42, 12 (2024).

[38] NVIDIA 2024. NVLink and NVLink Switch. https://www.nvidia.com/en-us/data-center/nvlink/

[39] NVIDIA Corporation 2024. NVIDIA Blackwell Architecture. https://resources.nvidia.com/en-us-blackwell-architecture

[40] NVIDIA Corporation 2024. NVIDIA GB200 NVL72. https://nvdam.widen.net/s/wwnsxrhm2w/blackwell-datasheet-3384703

[41] NVIDIA/Mellanox 2025. Compatible 1.6T OSFP Close Top InfiniBand XDR Passive Direct Attach Copper Twinax Cable. https://www.broadcom.com/company/news/product-releases/62491

[42] The Open Compute Project (OCP). 2023. Bunch of Wires (BoW) PHY Specification. https://www.opencompute.org/documents/bow-specification-v2-0d-1-pdf.

[43] The Open Compute Project (OCP). 2023. OpenHBI Specification Version 1.0. https://www.opencompute.org/documents/odsa-openhbi-v1-0-spec-rc-final-1-pdf.

[44] Peter J. Parbrook, Brian Corbett, Jung Han, Tae-Yeon Seong, and Hiroshi Amano. 2021. Micro-Light Emitting Diode: From Chips to Applications. *Laser & Photonics Reviews* 14, 8 (2021).

[45] Dylan Patel, Wega Chu, Chaolien Tseng, Myron Xie, Jeremie Eliahou Ontiveros, and Daniel Nishball. 2024. GB200 Hardware Architecture – Component Supply Chain and BOM. SemiAnalysis. https://semianalysis.com/2024/07/17/gb200-hardware-architecture-and-component/

[46] Dylan Patel and Daniel Nishball. 2024. *100,000 H100 Clusters: Power, Network Topology, Ethernet vs InfiniBand, Reliability, Failures, Checkpointing*. Technical Report. SemiAnalysis. https://semianalysis.com/2024/06/17/100000-h100-clusters-power-network/

[47] Dylan Patel and Daniel Nishball. 2024. *Nvidia's Optical Boogeyman – NVL72, Infiniband Scale Out, 800G & 1.6T Ramp*. Technical Report. SemiAnalysis. https://semianalysis.com/2024/03/25/nvidias-optical-boogeyman-nvl72-infiniband/

[48] Dylan Patel, Daniel Nishball, and Reyk Knuhtsen. 2024. *Amazon's AI Self Sufficiency: Trainium2 Architecture & Networking*. Technical Report. SemiAnalysis. https://semianalysis.com/2024/12/03/amazons-ai-self-sufficiency-trainium2-architecture-networking

[49] Dylan Patel, Myron Xie, and Gerald Wong. 2023. AI Capacity Constraints – CoWoS and HBM Supply Chain. SemiAnalysis. https://semianalysis.com/2023/07/05/ai-capacity-constraints-cowos-and/

[50] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient generative LLM inference using phase splitting. In *ISCA*.

[51] Bardia Pezeshki, Farzad Khoeini, Alex Tselikov, Robert F. Kalman, Cameron Danesh, and Emad Afifi. 2022. LED-array based optical interconnects for chip-to-chip communications with integrated CMOS drivers, detectors, and circuitry. In *Optical Interconnects*.

[52] Bardia Pezeshki, Suresh Rangarajan, Alex Tselikov, Emad Afifi, Ivan Huang, Jeff Pepper, Sarah Zou, Howard Rourke, Rowan Pocock, Alasdair Fikouras, Farzad Khoeini, Vahid Mirkhani, Steve Novak, and Rob Kalman. 2024. 304 channel MicroLED based CMOS transceiver IC with aggregate 1 Tbps and sub-pJ per bit capability. In *OFC*.

[53] Timothy Prickett Morgan. 2024. We Can't Get Enough HBM, Or Stack It Up High Enough. The Next Platform. https://www.nextplatform.com/2024/11/06/we-cant-get-enough-hbm-or-stack-it-up-high-enough/

[54] QSFP-DD MSA 2017. Accelerating 400GbE Adoption with QSFP-DD. http://www.qsfp-dd.com/wp-content/uploads/2017/03/QSFP-DD-whitepaper-15.pdf

[55] G. T. Reed and A. P. Knights. 2004. *Silicon Photonics: An Introduction*. Wiley.

[56] Lei Shan. 2024. *The Bunch of Wires (BoW) – An Open-Source Physical Interface Enabling Chiplet Architectures*. Technical Report. Ampere Computing. https://eps.ieee.org/images/files/TC_The_Bunch_of_Wires_rev3.pdf

[57] Debendra Das Sharma. 2022. The History of PCI IO Technology: 30 Years of PCI-SIG Innovation. Webinar. https://pcisig.com/sites/default/files/files/30%20Years%20of%20PCI-SIG%20Innovation%20Webinar_Final%20Slides.pdf

[58] Anton Shilov. 2025. TSMC mulls massive 1000W-class multi-chiplet processors with 40X the performance of standard models. Tom's Hardware. https://www.tomshardware.com/tech-industry/tsmc-mulls-massive-1000w-class-multi-chiplet-processors-with-40x-the-performance-of-standard-models

[59] Priyank Shukla. 2022. Short-Reach Interconnects for the Emerging Multi-Die System Era. In *Solid-State Circuits Chapter*. Synopsys. https://www.ieeetoronto.ca/wp-content/uploads/2022/12/Short-Reach-Interconnect-for-the-Emerging-Multi-Die-System-Era.pdf.

[60] Stanford. 2024. *DAM: Differentiated Access Memory Systems and Applications*. Technical Report. Stanford. https://dam.stanford.edu/assets/Stanford_DAM_2_Pages_2024.pdf

[61] The Information 2024. Nvidia Customers Worry About Snag With New AI Chip Servers. https://www.theinformation.com/articles/nvidia-customers-worry-about-snag-with-new-ai-chip-servers.

[62] Anthony Torza. 2023. Cisco Demonstrates Co-packaged Optics (CPO) System at OFC 2023. https://blogs.cisco.com/sp/cisco-demonstrates-co-packaged-optics-cpo-system-at-ofc-2023.

[63] Nathan Tracy, Gary Nicholl, Cathy Liu, Mike Li, Ed Frlan, and Steve Sekel. 2020. 112 Gbps Electrical Interfaces – An OIF Update on CEI-112G. In *OFC*.

[64] Amin Vahdat. 2020. Coming of Age in the Fifth Epoch of Distributed Computing: The Power of Sustained Exponential Growth. https://www.youtube.com/watch?v=27zuReojDVw&t=16s.

[65] Zhou Wang, Shijie Zhu, Xinyi Shan, Zexing Yuan, Zeyuan Qian, Xinyi Lu, Yi Fu, Kui Tu, Hui Guan, Xugao Cui, and Pengfei Tian. 2022. Red, green and blue InGaN micro-LEDs for display application: temperature and current density effects. *Opt. Express* 30, 20 (2022).