Slow-And-Wide Transceivers Shouldn't Be Color Blind

Lukas Röllin
roellinl@ethz.ch
ETH Zürich

Sushovan Das
susdas@ethz.ch
ETH Zürich

Benjamin Hoffman
bhoffman@ethz.ch
ETH Zürich

Laurent Vanbever
lvanbever@ethz.ch
ETH Zürich

Abstract

Modern AI interconnects face hard trade-offs: copper links are efficient but have a short reach, optical transceivers are long-reach but more costly and energy-hungry. Similar trade-offs occur in the switching domain: packet switching is flexible but power-hungry, while circuit switching is efficient but coarse-grained. Recently, slow-and-wide transceivers have closed the gap between copper and traditional optics by exposing many low-rate optical channels, but limit them to a single logical point-to-point link.

To address this, we propose ChromaTile, a potential extension of slow-and-wide transceivers with multicolor transmission and passive color steering. It would enable channels to be independently allocated across destinations simultaneously, creating a novel switching abstraction, *i.e.*, *optical micro-circuits*. We outline the optical components that could be leveraged for the data plane, present one possible control-plane algorithm for collision-free channel assignment, and discuss use cases. Our analysis suggests that ChromaTile can extend the scale-up domain beyond tightly packed racks and reduce scale-out network infrastructure requirements — using ChromaTile with 8 colors reduces the number of switches by ~ 40% while improving the network latency by one hop.

As a position paper, we introduce optical micro-circuits as a new switching abstraction and hope to spark broader collaboration across academia and industry to evaluate and realize this vision.

CCS Concepts

• **Networks** → **Physical links**; **Data center networks**.

Keywords

Optical Switching, AI Infrastructure, Slow-and-Wide Transceiver

ACM Reference Format:
Lukas Röllin, Sushovan Das, Benjamin Hoffman, and Laurent Vanbever. 2026. Slow-And-Wide Transceivers Shouldn't Be Color Blind. In *ACM SIGCOMM Workshop on Hot Topics in Optical Technologies and Applications in Networking (HotOptics '26), August 17–21, 2026, Denver, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3789240.3822612

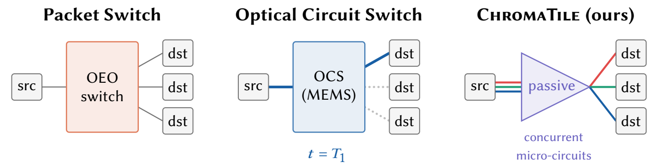

Figure 1: ChromaTile uses concurrent optical micro-circuits rather than electronic forwarding (left) or point-to-point coarse-grained optical circuits (middle).

1 Introduction

Modern AI clusters require communication fabrics for both scale-up and scale-out domains that simultaneously provide high bandwidth, low latency, and extended physical reach, all within tight power constraints. In *scale-up* domains, GPUs exchange latency-sensitive synchronization, collective, and memory-coherence traffic, while *scale-out* networks demand high aggregate bandwidth across thousands of endpoints with minimal hop count. Meeting these requirements is becoming increasingly difficult as network speeds continue to increase rapidly [9].

Current approaches. Existing interconnect technologies expose fundamentally different tradeoffs (Table 1).

NVLink-like copper fabrics provide low latency and high energy efficiency, but their limited reach of a few meters [6] forces GPUs to remain physically close, leading to very dense rack deployments with 72 GPUs and very complex cooling and power requirements of over 100 kW per rack [13]. As AI workloads continue to push scale-up domains to connect more hosts, packet-switched Ethernet and optical circuit switching (OCS) are increasingly moving towards the accelerator fabric.

Packet-switched Ethernet provides highly flexible any-to-any communication over longer distances when using optical transceivers, but incurs substantial optical-electrical-optical (OEO) conversion overhead, switch latency, and power consumption of ~ 20W [6, 12] at every hop. However, this flexibility often exceeds what AI workloads require: communication across dynamic endpoint subsets rather than arbitrary any-to-any patterns.

Optical circuit switching avoids packet processing overheads, but exposes communication through coarse-grained circuits: endpoints receive high bandwidth only during scheduled circuit windows, forcing traffic to either wait for reconfiguration (~ 50ms) [31, 35] or traverse indirect paths [2, 5]. Such point-to-point communication is too rigid for AI workloads. These trends underscore the need for a new interconnect design combining moderate flexibility with low latency, high bandwidth, long reach, and high energy efficiency.

Slow-and-wide optical transceivers. We believe that recent works on slow-and-wide optical transceivers, *e.g.*, MOSAIC [6] and Avicena [4], open up new opportunities. Instead of using a small number of high-speed optical lanes, they expose hundreds of low-speed optical channels per port using microLED arrays and imaging fibers. This approach already promises significantly lower power usage (> 50%) [6] than conventional optical transceivers while providing longer reach than copper. However, we argue that the real opportunity is broader: the large number of low-speed channels creates a useful switching granularity between packet switching and coarse-grained optical circuit switching.

ACM ISBN 979-8-4007-2467-1/26/08
https://doi.org/10.1145/3789240.3822612

Table 1: Trade-offs between switching technologies.

	Latency	Power	Reach	Downtime
NVLink (copper cable)	+	+	-	+
Ethernet (optical transceiver)	-	-	+	+
OCS	~*	+	+	-
ChromaTile (our proposal)	+	+	+	+

Optical Micro-Circuits. Our key observation is that these channels should not just be static but act as *independently steerable optical micro-circuits*. Rather than allocating an entire high-speed circuit to a single source-destination pair, channels can be allocated to connect multiple endhosts simultaneously. This new abstraction falls between the two extremes: more concurrent and flexible than time-multiplexed optical circuits, while avoiding per-packet electronic forwarding overheads (Fig. 1); thus, it aligns naturally with AI communication patterns.

This work: ChromaTile. As a potential realization of this vision, we propose ChromaTile, a color-steered slow-and-wide switching substrate that extends each microLED channel with multiple transmission colors and uses passive optical steering, enabled by micro-lens grating arrays [28], to route different colors toward different endpoints. Changing a channel's destination only requires selecting the emitted color, avoiding the downtime associated with reconfiguration in conventional optical circuit switches. The result is a *fabric of many parallel low-rate optical connections* whose bandwidth can be dynamically allocated across endpoints.

This new paradigm combines low-latency communication with reduced switching overhead and power consumption, while still enabling longer-reach connectivity and flexible bandwidth allocation. The design provides benefits in both scale-up (*e.g.*, extended physical rack domain) and scale-out networks (*e.g.*, flexible multi-plane fabrics), for example, reducing the number of switches in the topology by ~ 40%.

Key contributions. In summary, our key contributions are:

(1) **ChromaTile:** We propose to leverage the large number of channels in slow-and-wide optics to introduce *optical micro-circuits*, a middle ground between packet and coarse-grained circuit switching.
(2) **Data-plane:** We outline a ChromaTile data plane that combines multicolor microLED transmission with passive optical color steering.
(3) **Control-plane:** We sketch one possible solution of a control plane algorithm providing collision-free channel assignments from traffic matrices using Birkhoff-von Neumann decomposition.
(4) **Evaluation and potential:** We evaluate the promise of this approach through decomposition analysis and scale-up/scale-out case studies demonstrating potential reductions of ~ 40% in network switches.

*OCS latency is low, however, adapting circuits is comparatively slow.

Beyond proposing ChromaTile as one possible realization, this position paper advocates for optical micro-circuits as a new switching abstraction that better matches emerging AI communication requirements. While we outline the opportunities enabled by this design point, significant work remains to evaluate its feasibility, explore alternative realizations, and build practical prototypes. We hope this work motivates collaboration across networking, systems, and photonics communities to validate and realize this vision.

2 Motivation

This section examines the challenges of AI clusters and motivates how existing switching paradigms fall short.

2.1 Challenges in AI Clusters

Large-scale AI workloads increasingly combine highly concurrent communication patterns with growing cluster scale, pushing both scale-up and scale-out networks to their limits.

Scale-up communication. Modern accelerator systems increasingly depend on large scale-up domains connecting tens to potentially hundreds of GPUs [10]. These systems must support low-latency collective communication among many endpoints simultaneously [33], often under highly dynamic communication demands. Currently, systems rely heavily on copper interconnects, which provide low latency and low power [22], but also have limited reach at high speeds. This forces GPUs into dense rack deployments with increasingly complex cooling and power-delivery requirements. Extending communication domains beyond a rack with optical transceivers, on the other hand, significantly increases both power consumption [13] and switching overhead.

Scale-out communication. As communication increasingly extends to more GPUs, maintaining locality in large AI clusters with hybrid parallelisms [11, 15, 21, 26, 29] becomes difficult. But, building fully non-oversubscribed fabrics at this scale is expensive [9, 25]. Recent proposals suggest splitting the network into multiple planes to reduce switch radix requirements and improve network depth [10, 18, 34]. However, bandwidth partition across planes is still static: an 800 G host interface may be divided into four fixed 200 G connections. These systems rely on complex fine-grained load balancing to utilize the available bandwidth [3]. Under failures, where a ToR or network plane is unavailable, the bandwidth assigned to that plane is effectively lost, even if spare capacity in other planes is available.

2.2 Why Existing Paradigms Fall Short

Existing interconnects expose fundamental tradeoffs between flexibility, latency, power, and communication granularity.

Packet switching: flexible but expensive. Packet-switched fabrics provide flexible any-to-any communication over long distances. However, each switch hop incurs OEO conversion, packet forwarding, and buffering, inflating both latency and power consumption [6, 12]. These overheads become increasingly prominent in large-scale accelerator fabrics spanning multiple switching layers [9, 25]. In summary, packet switching may provide substantially more flexibility than the workload actually requires while still paying the associated forwarding and power costs at every hop.

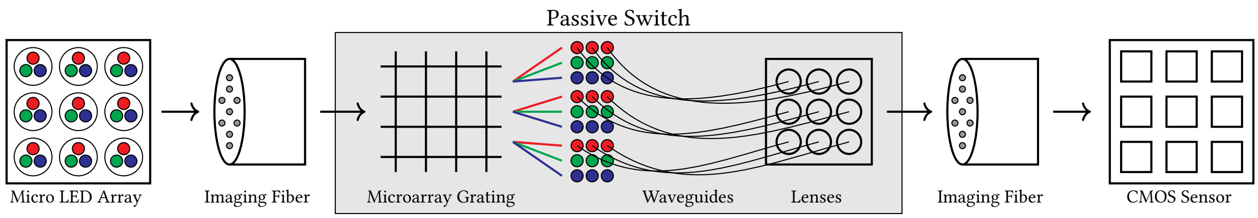

Figure 2: The ChromaTile data plane. Each endpoint uses a multicolor slow-and-wide transceiver. Passive color-steering routes colors per channel to distinct destinations, enabling simultaneous connections across endpoints.

Optical circuit switching: efficient but coarse-grained. Optical circuit switches avoid packet-forwarding overhead by establishing circuits between endpoints. Recent traffic-agnostic [2, 5] and ML-aware [27, 35] approaches rely on coarse-grained optical circuits and time-division multiplexing, connecting only a limited set of endpoint pairs at a time. Communication to other destinations must either wait for future configurations or traverse indirect paths, creating a throughput vs. latency trade-off [2]. Optical switches also suffer from reconfiguration downtime, during which communication is disrupted as it transitions between configurations. In summary, coarse-grained optical circuits are efficient but problematic for workloads with many concurrent, dynamic communication pairs, such as expert parallelism [19].

3 A New Paradigm: Optical Micro-Circuits

Recent slow-and-wide transceivers like MOSAIC [6] use microLED arrays and imaging fibers to expose hundreds of low-speed optical channels, providing the substantial channel-level parallelism that traditional narrow-and-fast optics lack. Our primary insight is that, instead of treating these channels as one indivisible high-speed circuit, subsets of channels can be independently allocated across multiple destinations simultaneously. Conceptually, these channel groups behave as *optical micro-circuits*: coarser than individual packets, yet more flexible than conventional optical circuits. This enables concurrent communication across multiple peers while dynamically allocating bandwidth at channel granularity.

Realizing such a system requires solving two key challenges. First, channels must be switched efficiently without incurring the coarse reconfiguration overheads of traditional OCS. As the ChromaTile data plane, we propose using multicolor microLED transmission with passive optical color steering, which can enable lightweight channel switching without mechanical circuit reconfiguration (§4). Second, endpoints must coordinate channel assignments while adapting to changing traffic demands. Unlike packet switching, where arbitration occurs at packet granularity, ChromaTile allocates channel resources directly across communicating endpoints. We sketch one possible control-plane algorithm that computes collision-free channel allocations from traffic matrices (§5).

4 ChromaTile Data Plane

We propose the ChromaTile data-plane (Fig. 2) which consists of two key modules: (i) a slow-and-wide optical transceiver with color-selectable transmission (§ 4.1) and (ii) a passive optical switch with color steering capability (§ 4.2). We also highlight the benefits over physical subchannels (§ 4.3) and seek community input on the data plane's feasibility (§ 4.4).

4.1 Multicolor Slow-and-Wide Transceiver

The ChromaTile transceiver extends a Mosaic-like slow-and-wide transceiver with the ability to independently select the transmission color of each channel. Importantly, this modification primarily affects the transmitter side, while the receiver remains unchanged.

Transmitter. On the transmitting side, each channel is extended with multiple possible LED colors instead of only one. This can be realized either by using multiple colored subpixels per channel, similar to RGB display pixels, or by using tunable microLEDs [23] capable of dynamically changing wavelength. The transmitter electronics select both the channel and emitted color, where the selected color determines the destination after passive optical steering (§4.2). Conceptually, color therefore acts as a lightweight routing primitive embedded directly into the transmission process [20].

The number of colors defines the radix of ChromaTile as each color corresponds to a destination port. Previous work [20] shows that roughly 8-color microLEDs without overlapping spectra can be used within the visible spectrum. Reducing the spectral width of microLEDs (*e.g.*, by using quantum dots [17]) or using colors outside the visible spectrum both improve the radix. Additionally, recent work [1] and our own analysis in § 6 demonstrate the usefulness of even low-radix switches.

Receiver. The receiver-side camera-like photodetector arrays are agnostic to transmitted wavelength and convert incoming photons into electrical signals – color discrimination occurs only when optical filters are applied. Therefore, receivers do not require destination-specific optical hardware, allowing the same design to operate across all wavelengths.

Minimal transceiver changes. We expect the modifications in the ChromaTile transceiver module to have a limited impact on manufacturing complexity. Commercial microLED fabrication already supports multiple colors per pixel in display technologies [24, 30]; ChromaTile largely repurposes such existing capabilities for communication. Moreover, the changes are primarily required on the transmitter, while the receiver remains unchanged.

4.2 Passive Color Steering

The second key component of ChromaTile is a passive optical steering fabric that routes channels based on transmitted wavelength, i.e., based on the color of the transmitted light. Logically,

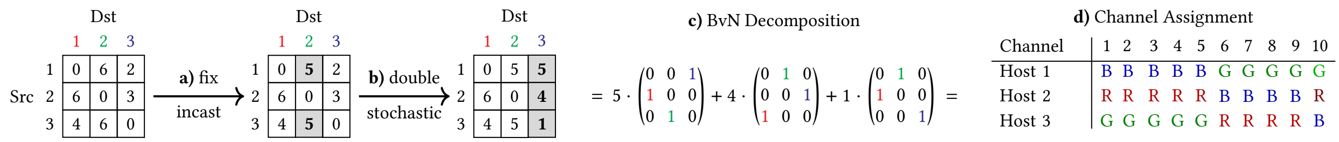

Figure 3: The 3×3 traffic matrix is first transformed into a doubly stochastic matrix and then decomposed into permutation matrices, which are then mapped onto physical channels (10 in total).

the switch performs two operations: (i) separating incoming channels by color, and (ii) recombining channels destined for the same endpoint into outgoing imaging fibers. We acknowledge that, as future work, a more in-depth feasibility analysis is required.

Color separation. The first stage acts as an array of miniature prisms, one for each incoming channel. Different transmitted colors are spatially separated and redirected towards different outputs. While steering hundreds of channels independently may appear challenging, similar structures already exist in multi-channel spectrometers [28]. These systems use micro-lens grating arrays to separate wavelengths across thousands of parallel optical paths. Prior work demonstrates arrays with more than 2000 micro-lens elements, which is sufficient for the channel counts envisioned in ChromaTile. Each of those micro-lens elements separates colors along only one dimension, but because it is a 2D array of such lenses, it creates the rows of colors shown in Fig. 2. Those micro-lens elements have been shown to work well over a wide wavelength range, from at least 400 *nm* to 700 *nm* [28].

Channel recombination. After wavelength separation, channels corresponding to the same destination are redirected towards the appropriate outgoing imaging fiber. This routing could be implemented using glass waveguides [8, 32] or by splitting an imaging fiber row by row. Finally, a lens stage focuses the redirected channels back into the outgoing imaging fiber connected to the destination transceiver. Conceptually, this stage performs the inverse operation of the initial wavelength separation.

Passive switching. As we observe, the ChromaTile steering module is entirely passive. Changing the destination of a channel only requires selecting a different emitted color at the transmitter; the optical fabric itself never reconfigures. Thus, ChromaTile avoids the reconfiguration downtime unlike MEMS-based optical fabrics [31].

4.3 Why Not More Physical Subchannels?

An alternative design would be to physically subdivide each imaging-fiber channel into many smaller, static subchannels, each connected to a different destination. We instead use *wavelength selection* because it preserves the physical channel geometry while introducing flexibility in destination selection. Increasing the number of physical subchannels would significantly complicate alignment, cabling, and optical isolation. As channels become smaller and more densely packed, crosstalk between the channels and fabrication complexity become increasingly difficult to manage.

Just as with conventional slow-and-wide transceivers, ChromaTile does not require every channel to function perfectly, since spare channels can be allocated without much overhead [6]. This should improve manufacturability and increase yield.

4.4 Call to Arms: Data Plane Feasibility

While we outline the opportunities enabled by this new paradigm of optical switching, many questions remain open. We invite the community to help with validating many aspects of our proposal.

First, the passive optical component requires a more thorough feasibility analysis, especially given the additional complexity of operating over a wide wavelength range. More detailed computations and simulations of loss and crosstalk within this passive optical component are essential to demonstrate the design's feasibility. Other important aspects include a more concrete analysis of the signal propagation through different physical channels (waveguides), and the recombination within the passive optical component, where channels from many different angles are combined into a single output channel.

Using broad-spectrum light sources such as microLEDs, the maximum achievable radix of ChromaTile is another important consideration, as it directly determines the range of use cases the design can support (§6). An important open question is how many distinct colors can be accommodated within the operating window of the passive optical component.

5 ChromaTile Control Plane

Assuming that the previously proposed ChromaTile data plane is feasible, this would enable many independently steerable low-rate optical channels, allowing concurrent connections among endpoints (Fig. 1). However, the endpoints must select channel colors such that collisions are avoided while adapting assignments to changing traffic demands. More concretely, if multiple endpoints simultaneously choose colors that route to the same destination on the same channel index, collisions occur at the receiver. The control plane objective is to compute collision-free channel assignments while efficiently utilizing the available channels. We give a sketch of one possible control-plane algorithm.

The endpoints can exchange communication demands using a lightweight out-of-band control channel. One or more spare slow-and-wide channels can be reserved for this purpose. Each endpoint periodically communicates the number of channels it requests to every other endpoint. These demands are aggregated into a traffic matrix T, where entry T_{ij} represents the number of channels requested from source i to destination j. Based on this traffic matrix, the controller computes a collision-free channel assignment and distributes the resulting color schedule back to the endpoints.

5.1 Collision-Free Channel Assignment

A collision-free channel assignment can be naturally represented by a permutation matrix; it corresponds to a valid communication pattern in which each source communicates with exactly one

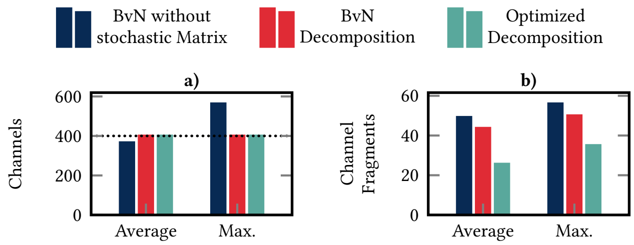

Figure 4: Our optimized decomposition reduces the number of channel fragments while *always* staying below the number of available channels.

destination and each destination receives from exactly one source. Given a traffic matrix, the Birkhoff-von Neumann (BvN) decomposition can express the demand as a weighted combination of such collision-free patterns.

Spatial interpretation of BvN. Prior reconfigurable optical networks [14, 27] interpret the decomposition temporally: each permutation matrix corresponds to a circuit configuration activated over time. In ChromaTile, the interpretation is fundamentally different. Rather than using permutation matrices as temporal schedules, we instantiate them spatially across the physical channels. Fig. 3d shows the channel assignment for the resulting decomposition in Fig. 3c. The coefficient 5 means that 5 channels are allocated according to the corresponding permutation matrix.

This distinction is important. Traditional optical circuit switches rely on time-multiplexing because only a small number of circuits can be active at any given time. In contrast, ChromaTile uses its large number of slow channels to realize many communication patterns simultaneously.

5.2 ChromaTile-Aware Implementation

Applying BvN decomposition directly to ChromaTile introduces several practical considerations. First, arbitrary traffic matrices must be preprocessed into a decomposable doubly stochastic form. Second, different ways of doing the decompositions can produce significantly different levels of channel fragmentation, motivating fragmentation-aware heuristics.

Constructing doubly stochastic matrices. BvN decompositions work best with doubly stochastic matrices, in which the sum of each row and column must not exceed the number of available channels. Intuitively, this constraint ensures that no endpoint requests more channels than physically exist.

However, arbitrary traffic matrices do not always satisfy this condition. Such a scenario can arise during traffic incasts where many sources simultaneously communicate with the same destination. For example, consider the input traffic matrix of Fig. 3. Source hosts 1 and 3 both require 6 channels each to destination host 2 (i.e., total 12 channels), which exceeds the available channels per host (i.e., 10). To handle this situation, we proportionally scale the requests until all column sums satisfy the channel constraints (Fig. 3a), similar to what congestion control would do in the same situation.

Next, the values are increased until every row and column sum equals exactly the total number of available channels, creating a doubly stochastic matrix (Fig. 3b). At this point, the Birkhoff-von Neumann theorem guarantees that the doubly stochastic matrix can be decomposed into a weighted set of permutation matrices [7] (Fig. 3c). These permutation matrices can then be mapped directly onto physical channels (Fig. 3d).

Fig. 4a shows the number of channels needed to satisfy the traffic matrix. The evaluation is run on 10,000 randomly generated 8-color traffic matrices with a total of 400 channels per host. We observe that without first creating a doubly stochastic matrix, the decomposition cannot always satisfy the traffic demand with the available channels (dotted line). By first extending the traffic matrix to a doubly stochastic matrix, we can guarantee that the decomposition uses exactly the number of available channels.

Reducing channel fragmentation. In principle, many different valid decompositions may exist for the same traffic matrix. Ideally, we would like to minimize the number of permutation matrices that are used in the decomposition to reduce channel fragmentation. However, finding the minimum decomposition is NP-hard.

Instead, we design a greedy heuristic that iteratively selects the permutation matrix with the largest coefficient. This heuristic tends to generate larger contiguous channel groups and reduces fragmentation in the resulting assignments. For example, in Fig. 3d, host 2's channel assignment has 5 red channels (index 1 – 5), 4 blue channels (index 6 – 9), and 1 red channel again (index 10); meaning 3 contiguous channel groups or fragments. Fig. 4b shows that our optimized decomposition heuristic effectively reduces the number of channel fragments by almost half (same scenario as Fig. 4a).

Runtime. While our current implementation has a runtime of a few milliseconds on a CPU, Chronos [27] has demonstrated that BvN decompositions can be implemented on hardware with a runtime below 1 μs given an input size of 64×64. Selecting a different microLED takes nanoseconds, enabling ChromaTile to allocate channels quickly enough for modern AI workloads that require reconfigurations at the millisecond scale [16].

6 Architectural Use Cases

ChromaTile would enable several new architectural opportunities across both scale-up and scale-out AI communication fabrics. In this section, we discuss how ChromaTile could be used to: (i) extend the communication domains beyond tightly packed racks, (ii) provide flexible bandwidth allocation across multiple network planes, and (iii) reduce the number of switches and transceivers required in large-scale AI fabrics.

6.1 Extending Scale-up Domains

ChromaTile could be used to extend the physical reach of scale-up domains while simultaneously providing low-latency dynamic bandwidth allocation without needing additional switching elements between communicating GPUs.

Expert parallelism. In case of expert parallelism, subsets of GPUs repeatedly exchange `all_to_all_v` traffic. Unlike traditional optical circuit switching systems that cycle through circuit schedules over time, ChromaTile can maintain simultaneous connectivity

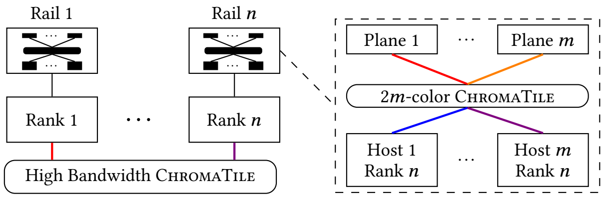

Figure 5: ChromaTile can be used for both the scale-up and scale-out network at the same time.

across many communicating endpoints, enabling low-latency communication between experts without waiting for future circuit configurations. This is especially important for inference workloads where end-to-end latency directly impacts responsiveness.

Demand-driven bandwidth. Scale-up domains without any switching usually resort to static full-mesh connectivity. Instead of uniformly partitioning bandwidth across all GPU pairs, ChromaTile dynamically allocates channels according to communication demand. As a result, communication patterns such as all_to_all_v, rings or trees can utilize substantially more bandwidth than under static full-mesh assignment, where each endpoint pair is limited to $\sim \frac{1}{n-1}$ of the total bandwidth, and many links remain underutilized.

6.2 Flexible Multi-Plane Fabrics

ChromaTile acts as a lightweight layer between hosts and ToR switches, enabling more flexible multi-plane scale-out network topologies. Unlike traditional multi-plane architectures (*e.g.*, P-Fattree), ChromaTile retains more of the flexibility of a unified network fabric and does not need complex fine-grained load balancing to extract available bandwidth.

Resilience. Dynamic channel assignment is particularly useful when facing failures or temporary traffic imbalances in the network. In conventional multi-plane networks, a ToR or network plane failure directly reduces the bandwidth available to all attached hosts, as it is statically pinned to the failed plane. With ChromaTile, endpoints can dynamically redirect channels to still-available planes, preserving host bandwidth unless the remaining aggregate capacity becomes the bottleneck.

Low-latency host groups. Hosts connected to the same ChromaTile domain can communicate with low latency without traversing additional packet-switching stages. This can be strategically exploited for synchronization-heavy traffic or cache-coherent communication among these hosts.

6.3 Reducing Network Layers

As ChromaTile enables flexible communication across multiple planes, larger scale-out network topologies can be constructed with fewer switching stages. A $2m$-color ChromaTile allows the network to be partitioned into m planes, but crucially supports dynamic bandwidth allocation across them (Fig. 5). A total of $2m$ colors per ChromaTile are needed since there are m planes and m hosts that need to be connected. For the same number of stages, more endpoints can be accommodated than in a single unified network plane. This is important for latency-sensitive AI communication,

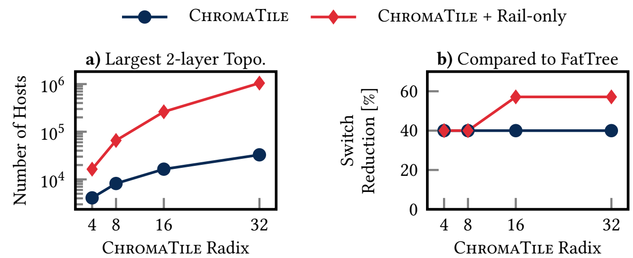

Figure 6: ChromaTile connects more hosts with lower hop count alongside reducing the number of switches.

as every hop adds forwarding latency. Fewer stages also reduce the number of required switches and transceivers. Fig. 5 shows ChromaTile being used both for scale-up and scale-out. Leveraging the high-bandwidth domain with a Rail-only topology [34] reduces the overall network size further.

Quantifying savings. Fig. 6a shows the number of hosts connected by 64-port switches and ChromaTile radix while only using 2-stage FatTrees (2048 hosts). Fig. 6b shows the network switch savings w.r.t. a single unified 3-stage FatTree. ChromaTile can reduce the number of switches by 40% (Fig. 6b) while connecting more hosts (Fig. 6a) with a smaller hop count. By combining ChromaTile in both scale-out and scale-up with a Rail-only topology, the number of hosts served by 2-stage networks grows even faster. At a ChromaTile radix of 16 and above, using the ChromaTile + Rail-only configuration (Fig. 5) even reduces the number of stages from four to two. Additional discussion can be found in § A.1.

7 Discussion & Outlook

The primary contribution of this work is to identify an untapped opportunity in slow-and-wide transceivers and outline the associated benefits. In particular, concurrent optical micro-circuits make this approach fundamentally different from traditional optical circuit switching, enabling communication among multiple hosts simultaneously, without relying on time-multiplexed circuit schedules. Broadly, ChromaTile suggests a path towards combining the benefits of optical switching with those of slow-and-wide transceivers. Since the large number of channels makes traditional optical circuit switching too costly. But as power continues to increase with each generation for both electrical switching and optical transceivers, clearly, alternatives are needed.

Still, open questions remain regarding the manufacturability, packaging, and system design of ChromaTile, *e.g.*, building a prototype to demonstrate ChromaTile's feasibility. On the control-plane side, improved algorithms could further reduce channel fragmentation or improve runtime under rapidly changing traffic demands. A more comprehensive evaluation of how best to deploy ChromaTile across both scale-up and scale-out scenarios, *e.g.*, low-radix switch deployments for AI workloads [1], also remains future work.

Acknowledgments

We thank the anonymous reviewers for their insightful feedback and the shepherd for the helpful guidance. This work was supported by the ETH Future Computing Laboratory (EFCL), financed by a donation from Huawei Technologies.

References

[1] Daniel Amir, Ori Cohen, Jakob Krebs, and Mark Silberstein. 2026. ACOS: Arrays of Cheap Optical Switches. https://arxiv.org/abs/2602.17449v1.

[2] Daniel Amir, Nitika Saran, Tegan Wilson, Robert Kleinberg, Vishal Shrivastav, and Hakim Weatherspoon. 2024. Shale: A Practical, Scalable Oblivious Reconfigurable Network. In *Proceedings of the ACM SIGCOMM 2024 Conference (ACM SIGCOMM '24)*. Association for Computing Machinery, New York, NY, USA, 449–464. doi:10.1145/3651890.3672248

[3] Joao Araujo, Alex Chow, Mark Handley, Ryder Lewis, Christoph Paasch, Jitendra Padhye, Michael Papamichael, Greg Steinbrecher, Amin Tootoonchian, Lihua Yuan, S. Anantharamu, Abhishek Dosi, Mohit Garg, Mahdieh Ghazi, Torsten Hoefler, Deepal Jayasinghe, Jithin Jose, Abdul Kabbani, Guohan Lu, Yang Wang, K. Doddapaneni, Murali Garimella, Vipin Jain, Yanfang Le, H. Nagulapalli, S. Narayanan, Rong Pan, Rathina Sabesan, Raghava Sivaramu, Rip Sohan, Eric Davis, Dragos Dumitrescu, Mohan Kalkunte, Bhaswar Mitra, Guglielmo Morandin, Adrian Popa, Costin Raiciu, Eric Spada, John Spillane, Niranjan Vaidya, Aviv Barnea, Idan Burstein, Elazar Cohen, Yamin Friedman, Noam Katz, Masoud Moshref, Yuval Shpigelman, Shahaf Shuler, Shy Shyman, and Sayantan Sur. 2026. Resilient AI Supercomputer Networking Using MRC and SRv6. doi:10.48550/ARXIV.2605.04333

[4] Avicena demonstrates first microLED based Transceiver IC in 16nm finFET CMOS for chip-to-chip communications at ECOC 2023 – Avicena [n. d.]. Avicena Demonstrates First microLED Based Transceiver IC in 16nm finFET CMOS for Chip-to-Chip Communications at ECOC 2023 – Avicena. https://avicena.tech/avicena-demonstrates-first-microled-based-transceiver-ic-in-16nm-finfet-cmos-for-chip-to-chip-communications-at-ecoc-2023/.

[5] Hitesh Ballani, Paolo Costa, Raphael Behrendt, Daniel Cletheroe, Istvan Haller, Krzysztof Jozwik, Fotini Karinou, Sophie Lange, Kai Shi, Benn Thomsen, and Hugh Williams. 2020. Sirius: A Flat Datacenter Network with Nanosecond Optical Switching. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '20)*. Association for Computing Machinery, New York, NY, USA, 782–797. doi:10.1145/3387514.3406221

[6] 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, and Paolo Costa. 2025. Mosaic: Breaking the Optics versus Copper Trade-off with a Wide-and-Slow Architecture and MicroLEDs. In *Proceedings of the ACM SIGCOMM 2025 Conference*. ACM, São Francisco Convent Coimbra Portugal, 234–247. doi:10.1145/3718958.3750510

[7] Haibin Chen, Liqun Qi, Louis Caccetta, and Guanglu Zhou. 2019. Birkhoff-von Neumann Theorem and Decomposition for Doubly Stochastic Tensors. *Linear Algebra Appl.* 583 (Dec. 2019), 119–133. doi:10.1016/j.laa.2019.08.027

[8] Gligor Djogo, Amir Rahimnouri, and Peter R. Herman. 2025. Laser Structured Optical Interposer for Ultra-dense Vertical Coupling of Multi-core Fibers to Silicon Photonic Chip. arXiv:2512.01972 [physics.optics] doi:10.48550/arXiv.2512.01972

[9] Adithya Gangidi, Rui Miao, Shengbao Zheng, Sai Jayesh Bondu, Guilherme Goes, Hany Morsy, Rohit Puri, Mohammad Riftadi, Ashmitha Jeevaraj Shetty, Jingyi Yang, Shuqiang Zhang, Mikel Jimenez Fernandez, Shashidhar Gandham, and Hongyi Zeng. 2024. RDMA over Ethernet for Distributed Training at Meta Scale. In *Proceedings of the ACM SIGCOMM 2024 Conference (ACM SIGCOMM '24)*. Association for Computing Machinery, New York, NY, USA, 57–70. doi:10.1145/3651890.3672233

[10] Yibo Guo, William M. Mellette, Alex C. Snoeren, and George Porter. 2022. Scaling beyond Packet Switch Limits with Multiple Dataplanes. In *Proceedings of the 18th International Conference on Emerging Networking EXperiments and Technologies (CoNEXT '22)*. Association for Computing Machinery, New York, NY, USA, 214–231. doi:10.1145/3555050.3569141

[11] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2019. GPipe: Efficient Training of Giant Neural Networks Using Pipeline Parallelism. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Number 10. Curran Associates Inc., Red Hook, NY, USA, 103–112.

[12] Romain Jacob, Lukas Röllin, Jackie Lim, Jonathan Chung, Maurice Béhanzin, Weiran Wang, Andreas Hunziker, Theodor Moroianu, Seyedali Tabaeiaghdaei, Adrian Perrig, and Laurent Vanbever. 2025. Fantastic Joules and Where to Find Them. Modeling and Optimizing Router Energy Demand. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25)*. Association for Computing Machinery, New York, NY, USA, 48–62. doi:10.1145/3730567.3732920

[13] Patrick Kennedy. 2024. This Is the NVIDIA DGX GB200 NVL72. https://www.servethehome.com/this-is-the-nvidia-dgx-gb200-nvl72/.

[14] Yiran Lei, Dongjoo Lee, Liangyu Zhao, Daniar Kurniawan, Chanmyeong Kim, Heetaek Jeong, Changsu Kim, Hyeonseong Choi, Liangcheng Yu, Arvind Krishnamurthy, Justine Sherry, and Eriko Nurvitadhi. 2026. FAST: An Efficient Scheduler for All-to-All GPU Communication. In *23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26)*. USENIX Association, Renton, WA, USA, 2515–2531.

[15] Wenxue Li, Xiangzhou Liu, Yuxuan Li, Yilun Jin, Han Tian, Zhizhen Zhong, Guyue Liu, Ying Zhang, and Kai Chen. 2024. Understanding Communication Characteristics of Distributed Training. In *Proceedings of the 8th Asia-Pacific Workshop on Networking (APNet '24)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3663408.3663409

[16] Xudong Liao, Yijun Sun, Han Tian, Xinchen Wan, Yilun Jin, Zilong Wang, Zhenghang Ren, Xinyang Huang, Wenxue Li, Kin Fai Tse, Zhizhen Zhong, Guyue Liu, Ying Zhang, Xiaofeng Ye, Yiming Zhang, and Kai Chen. 2025. MixNet: A Runtime Reconfigurable Optical-Electrical Fabric for Distributed Mixture-of-Experts Training. In *Proceedings of the ACM SIGCOMM 2025 Conference (SIGCOMM '25)*. Association for Computing Machinery, New York, NY, USA, 554–574. doi:10.1145/3718958.3750465

[17] Zhaojun Liu, Chun-Ho Lin, Byung-Ryool Hyun, Chin-Wei Sher, Zhijian Lv, Bingqing Luo, Fulong Jiang, Tom Wu, Chih-Hsiang Ho, Hao-Chung Kuo, and Jr-Hau He. 2020. Micro-Light-Emitting Diodes with Quantum Dots in Display Technology. *Light: Science & Applications* 9, 1 (May 2020), 83. doi:10.1038/s41377-020-0268-1

[18] William M. Mellette, Alex C. Snoeren, and George Porter. 2016. P-FatTree: A Multi-Channel Datacenter Network Topology. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*. ACM, Atlanta GA USA, 78–84. doi:10.1145/3005745.3005746

[19] Qingkai Meng, Hao Zheng, Zhenhui Zhang, ChonLam Lao, Chengyuan Huang, Baojia Li, Ziyuan Zhu, Hao Lu, Weizhen Dang, Zitong Lin, Weifeng Zhang, Lingfeng Liu, Yuanyuan Gong, Chunzhi He, Xiaoyuan Hu, Yinben Xia, Xiang Li, Zekun He, Yachen Wang, Xianneng Zou, Kun Yang, Gianni Antichi, Guihai Chen, and Chen Tian. 2025. Astral: A Datacenter Infrastructure for Large Language Model Training at Scale. In *Proceedings of the ACM SIGCOMM 2025 Conference (SIGCOMM '25)*. Association for Computing Machinery, New York, NY, USA, 609–625. doi:10.1145/3718958.3750521

[20] Driss Mouloua, Michael Martin, Miguel Beruete, Christophe Jany, Karim Hassan, and Thierry Baron. 2025. Exploring Strategies for Performance Enhancement in Micro-LEDs: A Synoptic Review of III-V Semiconductor Technology. *Advanced Optical Materials* 13, 6 (2025), 2402777. doi:10.1002/adom.202402777

[21] Deepak Narayanan, View Profile, Aaron Harlap, View Profile, Amar Phanishayee, View Profile, Vivek Seshadri, View Profile, Nikhil R. Devanur, View Profile, Gregory R. Ganger, View Profile, Phillip B. Gibbons, View Profile, Matei Zaharia, and View Profile. 2019. PipeDream. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*. ACM, Huntsville, ON, Canada, 1–15. doi:10.1145/3341301.3359646

[22] NVIDIA. [n. d.]. NVLink & NVLink Switch for Advanced Multi-GPU Communication. https://www.nvidia.com/en-us/data-center/nvlink/.

[23] Q-Pixel develops smallest full-color pixel and demonstrates first 10,000PPI full-color micro-LED display [n. d.]. Q-Pixel Develops Smallest Full-Color Pixel and Demonstrates First 10,000PPI Full-Color Micro-LED Display. https://www.semiconductor-today.com/news_items/2023/dec/q-pixel-0612123.shtml.

[24] Longheng Qi, Peian Li, Xu Zhang, Ka Ming Wong, and Kei May Lau. 2023. Monolithic Full-Color Active-Matrix Micro-LED Micro-Display Using InGaN/AlGaInP Heterogeneous Integration. *Light: Science & Applications* 12, 1 (Oct. 2023), 258. doi:10.1038/s41377-023-01298-w

[25] Kun Qian, Yongqing Xi, Jiamin Cao, Jiaqi Gao, Yichi Xu, Yu Guan, Binzhang Fu, Xuemei Shi, Fangbo Zhu, Rui Miao, Chao Wang, Peng Wang, Pengcheng Zhang, Xianlong Zeng, Eddie Ruan, Zhiping Yao, Ennan Zhai, and Dennis Cai. 2024. Alibaba HPN: A Data Center Network for Large Language Model Training. In *Proceedings of the ACM SIGCOMM 2024 Conference (ACM SIGCOMM '24)*. Association for Computing Machinery, New York, NY, USA, 691–706. doi:10.1145/3651890.3672265

[26] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory Optimizations toward Training Trillion Parameter Models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '20)*. IEEE Press, Atlanta, Georgia, 1–16.

[27] Sundararajan Renganathan and Nick McKeown. 2025. Chronos: Prescheduled Circuit Switching for LLM Training. In *Proceedings of the 2nd Workshop on Networks for AI Computing (NAIC '25)*. Association for Computing Machinery, New York, NY, USA, 89–97. doi:10.1145/3748273.3749210

[28] Shuonan Shan, Jingwen Li, Peiyuan Liu, Qiaolin Li, Xiaohao Wang, and Xinghui Li. 2023. A Microlens Array Grating for Miniature Multi-Channel Spectrometers. *Sensors* 23, 20 (Oct. 2023). doi:10.3390/s23208381

[29] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv:1909.08053 [cs.CL] doi:10.48550/arXiv.1909.08053

[30] Jiahao Song, Jingjing Jiang, Minhua Li, Qianxi Zhou, Linyue Meng, Ke Sun, Siyuan Cui, Kuosheng Wen, Sheng Liu, and Shengjun Zhou. 2025. Recent Advances in Nitride-Based Micro-LEDs for Next-Generation Display. *Advanced Physics Research* 4, 6 (2025), 2400190. doi:10.1002/apxr.202400190

[31] Michal Stepanovsky. 2019. A Comparative Review of MEMS-Based Optical Cross-Connects for All-Optical Networks From the Past to the Present Day. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2928–2946. doi:10.1109/COMST.2019.2895817

[32] Dezhi Tan, Zhuo Wang, Beibei Xu, and Jianrong Qiu. 2021. Photonic Circuits Written by Femtosecond Laser in Glass: Improved Fabrication and Recent Progress in Photonic Devices. *Advanced Photonics* 3 (March 2021), 024002–024002. doi:10.1117/1.AP.3.2.024002

[33] Sai Krishna Vemuri, Ajay Joshi, Vijay Kumar Motagi, Rohit Sindhu, Venkata Ravi Shankar Jonnalagadda, Amandeep Singh, Praveen Kumar Rukmangada, Sachin Shaw, and Vishal Tanna. 2025. Communication-Centric UALink and NVLink Comparison for Large Scale AI Training. In *2025 IEEE 32nd International Conference on High Performance Computing, Data and Analytics Workshop (HiPCW)*. IEEE, Hyderabad, India, 68–74. doi:10.1109/HiPCW66559.2025.00020

[34] Weiyang Wang, Manya Ghobadi, Kayvon Shakeri, Ying Zhang, and Naader Hasani. 2024. Rail-Only: A Low-Cost High-Performance Network for Training LLMs with Trillion Parameters. In *2024 IEEE Symposium on High-Performance Interconnects (HOTI)*. IEEE, Virtual, 1–10. doi:10.1109/HOTI63208.2024.00013

[35] Weiyang Wang, Moein Khazraee, Zhizhen Zhong, Manya Ghobadi, Zhihao Jia, Dheevatsa Mudigere, Ying Zhang, and Anthony Kewitsch. 2023. TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. USENIX Association, Boston, MA, USA, 739–767.

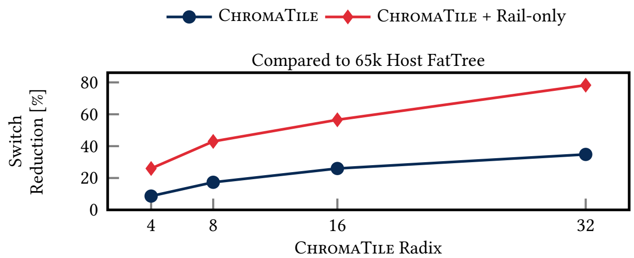

Figure 7: When looking at the interpolated formula, on average, the 8-color version of ChromaTile needs around 20% fewer switches.

A Appendix

This section provides additional context on evaluating the number of switches needed for a FatTree.

A.1 FatTree Interpolated Formula

In practice, the number of switches in a FatTree follows a step function. Once a radix-defined capacity is reached, adding another host requires a discrete "jump" in infrastructure by introducing an additional layer. Where this step occurs depends on many factors, such as the switch radix used or the number of hosts, and should be optimized on a case-by-case basis. To get rid of those arbitrary jumps in our analysis, we use an interpolated version of the formula that allows us to compare the radix of ChromaTile over a wide range, without arbitrary steps introduced by picking a specific number of hosts or switch radix. Intuitively, this interpolated formula presents the average number of switches that can be saved. Eq. (1) shows the interpolated Formula.

$$\begin{aligned} L &= \frac{\ln\left(\frac{hosts}{2}\right)}{\ln\left(\frac{radix}{2}\right)} \\ Switches &= \left\lceil (2L-1)\cdot\left(\frac{radix}{2}\right)^{L-1} \right\rceil \end{aligned} \tag{1}$$

Fig. 7 shows the average number of switches that can be saved when using the interpolated formula. On average the 8-color version of ChromaTile is able to reduce the number of switches needed by around 20%. With the ChromaTile + Rail-only version roughly doubling that number. For higher-radix versions, the benefits begin to flatten out, but there is still a benefit in using more colors.