Shedding light on AI

Making AI clusters for large language model training that are more reliable with photonics

Ana Gonzalez

Conceptual vision of large-scale AI infrastructure with optical circuit switches

Training large language models (LLMs) is critical for advancing AI, but scaling these models requires efficient and reliable infrastructure to mitigate failures. This article explores fault tolerance strategies employed by Google, Meta, and Alibaba, focusing on their approaches to managing failures during LLM training. It compares techniques such as check-pointing, redundancy, and network reconfiguration through Optical Circuit Switches, highlighting their impact on training efficiency.

We support the idea that Optical Circuit Switching (OCS), particularly with microsecond-range reconfiguration, emerges as a key solution, enabling rapid topology adjustments to bypass failures minimizing costly checkpoint recovery. This approach reduces downtime, supports dynamic workloads, and enhances scalability for hyperscale AI clusters.

Language model vs. cloud

General cloud computing differs significantly from LLM training in terms of traffic patterns and fault tolerance:

- Cloud computing generates millions of flows, resulting in high network entropy. These flows are continuous and typically utilize less than 20 % of NIC capacity.
- LLM training, on the other hand, produces few but periodically bursty flows, leading to low entropy and high utilization – sometimes reaching full NIC capacity.
- Additionally, LLM training is a synchronous process, where all GPUs collaborate to complete a series of distributed tasks. A failure in any task can delay or crash the entire training process, making LLM training more sensitive to faults than traditional cloud computing.

Current data centers running LLMs rely on Electrical Packet Switches (EPSs) with static wired topologies. These are designed to handle arbitrary communication patterns but are oblivious to actual traffic and optimized only for worst-case scenarios. They lack failure recovery strategies, such as the ability to rewire around faults.

Reliability challenges

Reliability is a fundamental challenge in operating large-scale machine learning (ML) infrastructures due to:

- Increased likelihood of hardware failure at scale, with training clusters continuously growing.
- Gang scheduling semantics, where all tasks of a parallel ML job (e.g., training across multiple GPUs) must run simultaneously or not at all.

When a failure occurs, LLM training relies on checkpoints to recover. However, this requires storage and incurs high overhead, often rolling back training by several hours – leading to financial loss and reduced productive runtime.

- Another challenge is identifying defective nodes in highly interconnected systems, where a single failure can cascade and obscure the root cause.

Meta [1] details how workloads are managed on their large-scale ML research clusters (RSC-1 with 16,000 GPUs and RSC-2 with 8,000 GPUs). Users submit jobs composed of many tasks, each running on the GPUs of a node. The scheduler attempts to co-locate tasks based on the physical network topology. A single task failure can trigger a complete job reallocation.

This motivates fault tolerance strategies such as checkpointing and redundancy for gang scheduling. While checkpointing allows recovery from a saved state, it introduces unproductive scheduled time due to restart and overhead.

Meta also notes potential improvements in the network fabric, suggesting that resilience could be enhanced by enabling topology reconfiguration to route around failures. Additionally, failure identifi-

cation requires real-time telemetry, prompting a redesign of the network.

Alibaba HPN [2] introduces a two-tier, dual-plane architecture with dual Top-of-Rack (ToR) switches to improve fault tolerance. While this design mitigates ToR failures and hash polarization, it has drawbacks: It only addresses ToR failures. Each host requires nine NICs connected to dual ToRs, increasing deployment complexity and testing overhead. Scalability becomes challenging for hyperscale clusters.

Now, let's use photonics!

In a previous work, Meta describes TopoOpt [3], a direct-connect fabric for deep neural network training. This strategy optimizes distributed training across computation, communication, and network topology.

TopoOpt uses reconfigurable telescent patch panel optical switches to create dedicated partitions for each training job. ToR switches connect to the optical layer, forming a direct-connect topology. The robotic patch panel physically reconfigures fibers, achieving up to 3.4× faster DNN training at lower cost. When a fiber fails, TopoOpt can temporarily use a link dedicated to Model Partitioning traffic to recover an AllReduce ring. In the case of permanent failures, TopoOpt reconfigures the topology by swapping ports to restore the failed connection.

However, reconfiguration takes seconds to minutes, limiting real-time adaptability. TopoOpt assumes static traffic patterns between iterations, which is unsuitable for models with dynamic communication needs (e.g., GNNs, MoE). A single link failure in an AllReduce ring disrupts sequential communication, and the slow reconfiguration cannot quickly restore optimal topology.

Let's use faster photonics

Google introduced its MEMS-based Optical Circuit Switch (OCS) [4], offering millisecond-range reconfiguration for large-scale AI clusters. Unlike Meta's TopoOpt, which relies on slow robotic patch panels, Google's OCS dynamically reconfigures topologies to route around link failures.

Built on Palomar MEMS switches, OCS achieves up to 3× better system availability compared to static fabrics. It maintains high bandwidth for bursty, high-utilization traffic in GPU/TPU superpods and eliminates hash polarization through optical transparency.

By supporting dynamic workloads like GNNs and MoE models, OCS reduces reliance on costly checkpointing and minimizes unproductive scheduled time.

Let's use even faster photonics

The most common training patterns involve periodic topology changes. However, as the number of model parameters surges, more efficient fine-grained solutions – such as model parallelism and data parallelism – must be considered for LLM training, and these approaches introduce dynamic topologies.

In early work, KDDI [5] developed a nanosecond switching architecture based on a 2×2 electrically controllable Faraday rotator. The authors argue that fast optical switching, with at least microsecond-level reconfiguration time, is necessary to support the rapidly varying topologies involved in distributed training jobs. While MEMS optical switches enable topological flexibility, they cannot reach the required speed.

Silicon Photonics (SiPh) offers several advantages for developing fast, reconfigurable Optical Circuit Switches (OCS) for distributed LLM training. These advantages stem from solid-state integration, compact design, and flexible processing capabilities. Key benefits include:

- Gain control by aggregating Optical Semiconductor Amplifiers (OSAs), which are currently disaggregated components but may be integrated on-chip in the future.
- Fast reconfiguration of connections in the microsecond range using thermo-optical phase shifters, and scalability to nanosecond range using electro-optical phase shifters.
- Real-time telemetry via on-chip Si/Ge photodetectors that monitor signal amplitude and can localize failures.
- Increased cost efficiency, which scales dramatically with volume due to shared manufacturing processes with microelectronics.
- Enhanced reliability, for the same reasons as above.

At iPronics, we have developed and launched the first SiPh OCS (**Fig. 1**) with a reconfiguration time of less than 100 µs. It provides per-channel gain control by incorporating disaggregated OSAs. iPronics' technology is based on high-performance, programmable Photonic Integrated Circuit (PIC) building blocks designed in-house.

Our development efforts have focused on:

- Improving performance such as reducing losses in PIC building blocks to enable low-impact OSA design.
- Increasing circuit density, enabling the creation of strictly non-blocking, high-radix OCS.
- The developed OCS is software-controlled, ready for deployment in AI cluster control planes, and supports additional functions such as multicast capabilities.

Conclusions

This article highlights a growing trend toward the adoption of Optical Circuit Switches (OCS) by large-scale clusters for training heavy models like LLMs using distributed patterns. This shift is driven by the high cost of current reliability-enhancing mitigations, posing a significant challenge for cloud service providers.

As Meta notes, "over 90 % of jobs use less than one server but account for less than 10 % of GPU time" [1], while Alibaba reports that "a fault in LLM training can cost 20× more than in general cloud computing" [2]. Cloud providers are actively seeking solutions to improve efficiency for large models. OCS – especially those based on solid-state technology with microsecond reconfiguration times – emerge as promising solutions to meet these demands.

[1] *A. Kokolis* et al., Revisiting Reliability in Large-Scale Machine Learning Research Clusters, 2025 IEEE International Symposium on High-Performance Computer Architecture

[2] *K. Qian* et al., Alibaba HPN: A Data Center Network for Large Language Model Training, 2024 ACM SIGCOMM

[3] *W. Wang* et al., TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs, arXiv:2202.00433v3 (2022)

[4] *R. Urata* et al., Mission Apollo: Landing Optical Circuit Switching at Datacenter Scale, arXiv:2208.10041 (2022)

[5] *C. Wang* et al., Modoru: Clos Nanosecond Optical Switching for Distributed Deep Training, Journal of Optical Communications and Networking **16**, A40 (2024)

Author

Dr. Ana Gonzalez,
VP Business Development at iPronics

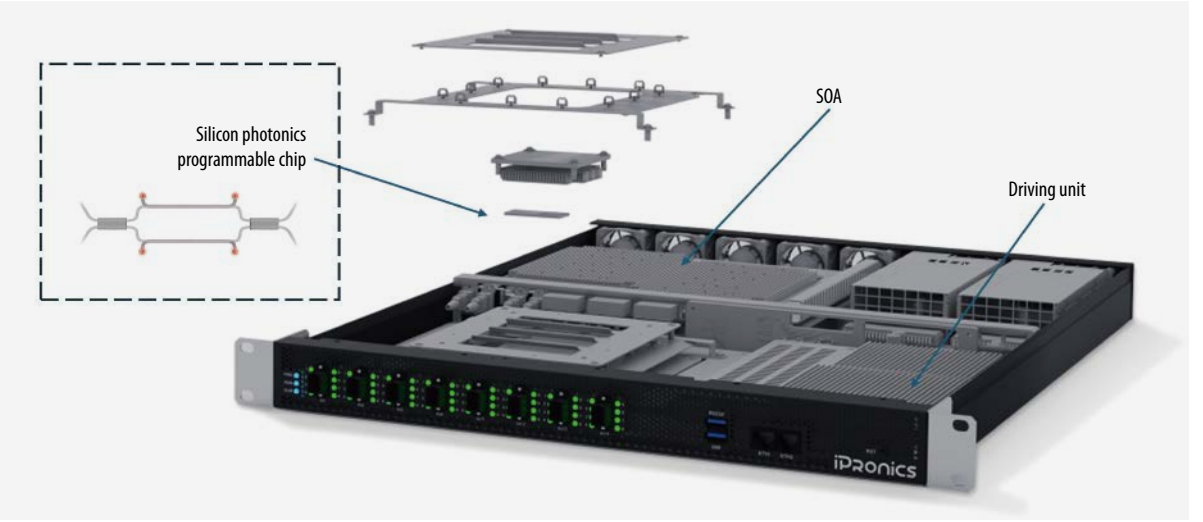

Fig. 1 iPronics Optical Networking Engine: 32 radix fast reconfiguration OCS, with incorporated gain control and telemetry, design for AI cluster applications. The picture highlights the SiPh chip, the OSA and driving unit. Inset: detail of the photonic unit cell, a 2×2 MMI controlled by a thermo-optical phase shifter.