

# Optical Network Resource Management for Large-Scale AI Training Clusters

Macaria Iniguez

Department of Computer Science, University of New Hampshire, Durham, NH, USA.  
macariamail@unh.edu

Sigmar Schweitzer

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.  
sigmar.schweitzer272@uab.edu

## Abstract

Large-scale artificial intelligence training clusters increasingly depend on optical networking to sustain the bandwidth, latency, and synchronization requirements of distributed model training. This paper examines optical network resource management as a socio-technical problem that spans physical-layer transmission, switching architectures, scheduler design, operational governance, energy sustainability, robustness, and fairness. The analysis begins by characterizing training workloads, including collective communication, parameter synchronization, pipeline and tensor parallelism, and the sensitivity of convergence to tail latency. It then evaluates optical resource management architectures, including hybrid electrical-optical fabrics, optical circuit switching, wavelength switching, and software-defined control planes, with attention to structural trade-offs among reconfiguration latency, port count, buffering, circuit utilization, and failure isolation. The paper argues that resource management for AI clusters cannot be reduced to bandwidth provisioning because optical reconfiguration interacts with job placement, checkpointing, collective algorithms, and failure recovery. Governance mechanisms must therefore coordinate admission control, topology-aware scheduling, tenant fairness, observability, and environmental accounting. The discussion also considers deployment constraints in multi-tenant and geo-distributed environments, the risks of overfitting infrastructure to a narrow class of model architectures, and the policy implications of concentrating advanced optical capacity in a small number of institutions. The conclusion proposes a research agenda around adaptive control, disaggregated optical fabrics, energy-aware scheduling, robust circuit restoration, and transparent reporting of infrastructure performance and sustainability.

## Keywords

optical networks; AI training clusters; resource management; data center interconnects; scheduling; sustainability; governance; robustness; fairness; policy.

## 1. Introduction

Large-scale artificial intelligence training has become a defining workload for modern computing infrastructure. Training frontier models requires thousands of accelerators to operate as a coordinated system, exchanging gradients, activations, optimizer states, and checkpoints across a fabric that must remain predictable under intense bursts of collective communication. The computational cost of training is substantial, but the communication cost is often the limiting factor in achieved scale and efficiency. As model parallelism, data

parallelism, and pipeline parallelism are combined, the network ceases to be a passive interconnection medium and becomes an active determinant of convergence time, resource utilization, and failure recovery. Research on tail latency in large-scale services has long demonstrated that rare slow components can dominate user-visible performance, and this insight applies with even greater force to synchronized training, where a single delayed participant can stall an entire collective operation [1].

The architectural response to these demands has been a gradual movement from purely electrical switching toward hybrid and optical fabrics. Optical interconnects offer high bandwidth, low loss, and potentially lower energy per bit over distance, but they also introduce new management problems. Optical circuit switches, wavelength-selective elements, and reconfigurable add-drop multiplexers operate on different time scales and with different failure modes than packet switches. A resource manager for an optical AI cluster must decide not only which flows receive bandwidth but also when circuits are established, how long they persist, which topology is presented to the scheduler, and how reconfiguration interacts with ongoing training jobs. These decisions are further complicated by the fact that distributed training frameworks already impose their own communication patterns through pipeline schedules, sharded optimizers, and collective algorithms [2], [3], [4].

This paper argues that optical network resource management for large-scale AI training clusters should be understood as a problem of coordinated infrastructure governance rather than isolated bandwidth allocation. The central challenge is to align physical-layer reconfigurability with application-level synchronization, operational policy, and sustainability constraints. The analysis draws on distributed training systems, data center network architectures, optical switching research, and policy-oriented studies of computing infrastructure. It examines how different parallelism strategies create distinct traffic matrices, why collective operations are especially sensitive to latency and jitter, and how optical resource management can either alleviate or amplify these sensitivities. It also considers fairness among tenants, robustness under failure, energy accounting, and the broader policy consequences of building specialized optical infrastructures for AI.

## **2. Workload Characteristics and Network Demands**

Distributed training workloads are not monolithic. Data-parallel training replicates the model across workers and exchanges gradients after each step, producing periodic all-reduce or reduce-scatter operations. Pipeline parallelism partitions layers across stages and sends activations and gradients between stages, creating bursty point-to-point communication that is sensitive to pipeline bubbles. Tensor parallelism splits individual layers across devices, generating frequent all-reduce and all-gather operations with stringent latency requirements. Mixture-of-experts models introduce all-to-all communication as tokens are routed to different experts. Frameworks such as Horovod, PyTorch Distributed, Megatron-LM, ZeRO, and GShard each embody different assumptions about network behavior, and their performance depends on whether the fabric can support the resulting communication patterns efficiently [3], [4], [5], [7], [8].

The temporal structure of training traffic is particularly important for optical resource management. Gradients are exchanged in synchronized phases, while pipeline stages communicate in overlapping but uneven bursts. Checkpointing creates large, bursty writes to storage that may traverse the same fabric as training traffic. Compiler and runtime optimizations can overlap computation and communication, but they cannot eliminate the underlying need for low-latency, high-bandwidth exchanges. Quantization and gradient

compression reduce message sizes, yet they introduce trade-offs in convergence and do not remove the need for predictable scheduling [6]. Consequently, an optical network manager must account for both average throughput and instantaneous congestion, because training progress is often determined by the slowest collective or the most delayed pipeline stage.

Tail latency and jitter are especially consequential in this setting. A collective operation completes only when all participating workers have contributed, so a small number of delayed messages can idle a large fraction of the cluster. This is a structural property of synchronous training, not merely an implementation artifact. Optical fabrics can reduce latency by shortening paths and avoiding electrical buffering, but they can also introduce reconfiguration delays, wavelength contention, and polarization or synchronization effects that are unfamiliar to conventional packet-switched data centers. The design of resource management must therefore consider the distribution of completion times, not just mean utilization. A fabric that performs well on average but occasionally stalls a collective may be inferior to a slightly lower-bandwidth fabric with more predictable service.

The workload also evolves over time. Early stages of training may be more compute-bound, while later stages may become more communication-bound as gradient exchange dominates. Hyperparameter tuning, data preprocessing, and evaluation introduce different traffic patterns that may coexist with training. In multi-tenant clusters, several jobs with different parallelism strategies may share the same optical fabric, creating a composite traffic matrix that changes on scheduling and checkpointing time scales. Resource management must therefore be adaptive rather than static. It must observe job structure, communication phases, and tenant priorities, and it must decide when to reconfigure optical paths, when to keep circuits in place, and when to rely on electrical switching for fine-grained multiplexing.

### **3. Optical Resource Management Architectures**

Optical resource management architectures for AI clusters can be grouped into several broad families. Hybrid electrical-optical designs use a high-radix electrical packet-switched network for fine-grained traffic and an optical circuit-switched network for large, predictable flows. Helios is an early exemplar of this approach, combining electrical pod switches with optical circuit switches to connect pods and reduce the cost and power of the core [9]. OSA similarly explores optical switching architectures with flexibility and reconfiguration, showing how wavelength routing can support high-capacity data center communication [10]. These designs demonstrate that optical circuits can offload bulk flows, but they also reveal the difficulty of matching circuit setup times to application requirements.

A second family relies on wavelength switching and tunable transceivers. In such architectures, endpoints or edge switches tune to specific wavelengths to establish logical connections through passive or semi-passive optical elements. The advantage is high capacity and potentially low latency, but the management challenge is wavelength assignment, contention resolution, and rapid tuning. Recent work on low-latency wavelength-switched and clock-synchronized intra-data center interconnects using hollow core fiber illustrates how physical-layer advances can reduce propagation delay and synchronization overhead, thereby changing the feasible design space for tightly coupled training traffic [17]. Such advances do not eliminate the need for resource management; rather, they shift the bottleneck toward control, scheduling, and failure handling.

A third family uses optical circuit switches controlled by a software-defined network controller. Jupiter evolving at Google is a prominent example of transforming a large data

center network with optical circuit switches and software-defined networking, showing that optical reconfiguration can be integrated into production operations at scale [12]. Projector explores agile reconfigurable data center interconnects, emphasizing the importance of matching reconfiguration granularity to traffic patterns [13]. These systems suggest that optical resource management should be topology-aware and application-aware, because the value of a circuit depends on whether it carries a long-lived flow or a transient burst. If circuits are established too slowly, training jobs may finish before the topology adapts. If they are established too aggressively, the control plane may thrash and destabilize the network.

The structural trade-offs among these architectures are substantial. Optical circuit switching offers high bandwidth and low per-bit energy for large flows, but it has limited support for fine-grained multiplexing and may require guard times or buffering during reconfiguration. Wavelength switching can be faster than mechanical or micro-electromechanical switching, but it requires tunable components and careful spectrum management. Hybrid designs preserve the flexibility of packet switching while offloading predictable traffic, but they add complexity in routing, monitoring, and fault isolation. A resource manager must decide how much traffic to place on optical circuits, how to route around failed or degraded optical elements, and how to prevent optical reconfiguration from violating service-level objectives of unrelated jobs.

From a systems perspective, the optical control plane is not merely a lower-layer concern. It determines the effective topology seen by distributed training schedulers. If a scheduler assumes a fixed fat-tree topology while the optical layer presents a time-varying circuit topology, placement decisions may become suboptimal or even incorrect. Conversely, if the optical layer exposes reconfiguration capabilities to the scheduler, it may be possible to align circuit setup with collective communication phases. This alignment requires interfaces between job runtimes, cluster schedulers, and network controllers. It also requires a governance model that defines who can trigger reconfiguration, how conflicts are resolved, and what performance guarantees are offered to tenants.

#### **4. Scheduling, Control, and Governance**

Scheduling in optical AI clusters operates at multiple levels. At the job level, a cluster scheduler decides which jobs run concurrently, how many accelerators each receives, and where they are placed. At the communication level, a collective library chooses algorithms and message sizes. At the network level, an optical controller decides circuit allocation, wavelength assignment, and path routing. These layers are often designed independently, but their interactions determine overall efficiency. A placement decision that ignores optical topology may create long electrical paths for traffic that could have used a direct circuit. A collective algorithm that assumes full bisection bandwidth may perform poorly if optical circuits are oversubscribed or frequently reconfigured.

Governance mechanisms are needed to coordinate these layers without creating brittle central control. One approach is hierarchical control, in which the cluster scheduler reserves optical capacity for a job based on its communication profile, and the optical controller manages circuits within that reservation. Another approach is market-based or priority-based arbitration, in which jobs bid for optical circuits according to deadline, tenant priority, or energy budget. A third approach is predictive control, in which telemetry and machine learning are used to forecast communication phases and preconfigure circuits. Each approach has trade-offs. Hierarchical control is predictable but may underutilize capacity. Market-based

control can improve efficiency but may disadvantage smaller tenants. Predictive control can be proactive but risks errors when workloads deviate from historical patterns.

Fairness is a central governance concern. Large AI training jobs often require thousands of accelerators and can dominate an optical fabric for hours or days. If optical circuits are allocated on a first-come, first-served basis, smaller jobs may experience starvation or unpredictable latency. If allocation favors the largest jobs, the cluster may achieve high throughput on a few models while reducing diversity of research and experimentation. Fairness in this context is not limited to equal bandwidth; it includes equal access to low-latency paths, predictable reconfiguration opportunities, and transparent information about congestion and failures. Policy frameworks for shared infrastructure should therefore define service classes, maximum reservation durations, and preemption rules that balance efficiency with equitable access.

Observability is a prerequisite for fair and effective governance. Operators need to know which flows use optical circuits, how often reconfiguration occurs, which jobs experience tail latency, and how much energy is consumed per training step. Without such telemetry, resource management becomes reactive and opaque. Model cards and similar documentation practices offer a broader analogy: just as model reporting can improve accountability by disclosing training data, evaluation, and limitations, infrastructure reporting can improve accountability by disclosing network topology, scheduling policies, energy sources, and failure modes [18]. In the optical domain, this might include circuit utilization, reconfiguration latency distributions, wavelength contention, and the carbon intensity of the energy used during training windows.

## **5. Robustness, Fairness, and Sustainability**

Robustness in optical AI clusters must address both component failures and systemic fragility. Optical transceivers, amplifiers, switches, and fibers can degrade or fail. A single misconfigured circuit switch or a contaminated fiber connector can disrupt a large training job. Because training jobs are synchronized, a failure in one optical path can cause a collective operation to hang or time out, wasting accelerator time and energy. Robust resource management must therefore include fast failure detection, circuit restoration, and graceful degradation to electrical paths. It must also consider correlated failures, such as power events that affect multiple optical elements or cooling failures that force cluster-wide throttling. The goal is not merely to restore connectivity but to preserve training progress through checkpointing, elastic reconfiguration, or job migration.

Fairness and robustness interact in subtle ways. When the network is degraded, operators may prioritize large jobs to protect sunk investment, but this can disadvantage smaller tenants and reduce long-term innovation. Alternatively, strict fairness may lead to frequent preemption and thrashing, which harms all jobs. A balanced policy might reserve a portion of optical capacity for best-effort traffic, guarantee minimum service levels for all tenants, and allow large jobs to use excess capacity only when it does not violate guarantees. Such policies require accurate accounting of optical resource usage and clear definitions of entitlement. They also require dispute resolution mechanisms, because tenants may disagree about whether performance degradation was caused by network management, application behavior, or external factors.

Sustainability is another critical dimension. Data centers already consume significant electricity, and AI training adds demanding accelerator and networking loads. Studies of data

center energy use have shown that efficiency improvements can be offset by growth in demand, making it important to measure and manage total consumption rather than assume that efficiency gains are sufficient [14]. Work on energy and policy considerations for deep learning has highlighted the carbon footprint of model training and the need for transparency about energy sources and training duration [15]. Optical networks can reduce energy per bit relative to electrical interconnects over distance, but they also add control overhead, cooling requirements, and manufacturing impacts. Resource management should therefore incorporate energy-aware scheduling, carbon-aware placement, and lifecycle accounting rather than treating optical capacity as environmentally neutral.

There are also concerns about the concentration of AI infrastructure. Advanced optical networks are expensive to design, deploy, and operate, which may limit participation to a small number of well-resourced institutions and corporations. This concentration affects research agendas, access to compute, and the ability to scrutinize model development. Policy responses might include shared national research facilities, transparent procurement standards, open benchmarks for optical network performance, and requirements for reporting energy and bandwidth usage. The technical community can contribute by designing resource management systems that are modular, interoperable, and auditable, so that smaller organizations can participate without rebuilding an entire infrastructure stack. Sustainability, fairness, and robustness are therefore not peripheral concerns; they shape the legitimacy and long-term viability of large-scale AI infrastructure.

## **6. Deployment and Operational Considerations**

Deploying optical resource management in real AI clusters involves practical constraints that are often absent from conceptual designs. Cabling, patch panels, power distribution, and cooling must accommodate optical transceivers and switches. Operational teams need tools to diagnose optical faults, measure signal quality, and validate reconfiguration without disrupting running jobs. Software-defined networking controllers must integrate with existing cluster schedulers, monitoring systems, and security policies. In multi-tenant environments, isolation is essential: one tenant's optical reconfiguration should not degrade another tenant's traffic, and telemetry should not leak sensitive information about model architecture or job size. These requirements imply that optical resource management is as much an operations and governance problem as an engineering problem.

The pace of AI model evolution creates additional uncertainty. A network optimized for dense transformer training may perform poorly for mixture-of-experts models, spiking neural networks, reinforcement learning, or multimodal training. If the optical fabric is too specialized, it may become a stranded asset. A more resilient strategy is to design reconfigurable and disaggregated optical resources that can be repurposed across workload classes. This means investing in flexible spectrum allocation, modular switching, and software abstractions that hide physical-layer details while exposing useful capabilities such as latency, bandwidth, and reliability. It also means avoiding overfitting the control plane to a single collective library or parallelism strategy.

Supply chain and standardization issues matter as well. Optical components are produced by a relatively small number of vendors, and interoperability between transceivers, switches, and controllers is not guaranteed. Open standards and multi-vendor ecosystems can reduce lock-in and improve resilience, but they may also slow innovation if consensus is difficult. Operators must balance the benefits of tightly integrated proprietary systems against the risks of dependence and limited competition. For AI training clusters, where performance

requirements are extreme, this balance is particularly delicate. A governance framework should include procurement policies that encourage interoperability, security auditing, and long-term maintainability.

## **7. Policy and Cross-Domain Implications**

The policy implications of optical resource management extend beyond the data center. If AI training capacity becomes a strategic resource, then access to high-performance optical networks may influence national competitiveness, scientific research, and democratic oversight. Governments and funding agencies may need to support shared infrastructure, open testbeds, and benchmarks that allow diverse institutions to experiment with optical AI fabrics. They may also need to set standards for energy reporting, data protection, and fair access. Because optical networks are capital-intensive, public investment can shape whether the benefits of AI infrastructure are broadly distributed or concentrated in a few firms and regions.

Cross-domain comparisons can illuminate these issues. High-performance computing has long managed specialized interconnects, batch schedulers, and fair-share policies, and its experience with scientific workloads offers lessons for AI clusters. Cloud computing has developed multi-tenant isolation, service-level agreements, and cost models that can inform optical resource governance. Telecommunications has deep expertise in optical transport, restoration, and spectrum management, though its time scales and traffic patterns differ from those of synchronized AI training. Bringing these communities together is essential, because no single discipline owns the problem. Optical engineers understand physical-layer constraints, systems researchers understand scheduling and runtime behavior, and policy scholars understand incentives and accountability.

The future of optical AI infrastructure may involve greater disaggregation. Compute, memory, and storage could be separated and connected through reconfigurable optical fabrics, allowing resources to be composed dynamically for different training jobs. This vision promises flexibility and efficiency, but it also intensifies the need for resource management. If accelerators, memory pools, and storage nodes are connected through optical circuits, then every job placement becomes a network design problem. Failure detection, security isolation, and fairness become even more complex. Research should therefore proceed not only on faster optical components but also on control abstractions, formal guarantees, and governance models that can handle this complexity.

## **8. Conclusion**

Optical network resource management for large-scale AI training clusters is a systems problem with technical, operational, and policy dimensions. The workloads are synchronized, bursty, and sensitive to tail latency, while optical fabrics offer high bandwidth and low energy per bit but introduce reconfiguration delays, wavelength contention, and unfamiliar failure modes. Effective management requires coordination among job schedulers, collective communication libraries, optical controllers, and tenant governance policies. It also requires robustness mechanisms that preserve training progress under failure, fairness mechanisms that prevent starvation and concentration, and sustainability mechanisms that account for energy and carbon across the infrastructure lifecycle. The required reference illustrates how physical-layer advances can reduce latency and synchronization overhead, but such advances must be embedded in a broader control and governance architecture. Future research should develop adaptive, auditable, and interoperable optical resource management systems that can support diverse AI workloads without sacrificing robustness, fairness, or environmental responsibility.

## References

1. Dean, J., & Barroso, L. A. (2013). The tail at scale. *Communications of the ACM*, 56(2), 74–80. <https://doi.org/10.1145/2408776.2408794>
2. Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N. R., Ganger, G. R., Gibbons, P. B., & Zaharia, M. (2019). PipeDream: Generalized pipeline parallelism for DNN training. *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, 1–15. <https://doi.org/10.1145/3341301.3359646>
3. Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.
4. Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). ZeRO: Memory optimizations toward training trillion parameter models. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16. <https://doi.org/10.1109/SC41405.2020.00024>
5. Sergeev, A., & Del Balso, M. (2018). Horovod: Fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799*.
6. Alistarh, D., Grubic, D., Li, J., Tomioka, R., & Vojnovic, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30, 1709–1720.
7. Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., & Chintala, S. (2020). PyTorch Distributed: Experiences on accelerating data parallel training. *Proceedings of the VLDB Endowment*, 13(12), 3005–3018. <https://doi.org/10.14778/3415478.3415530>
8. Lepikhin, D., Lee, H., Xu, Y., Krikun, M., Shazeer, N., & Chen, Z. (2021). GShard: Scaling giant models with conditional computation and automatic sharding. *International Conference on Learning Representations*.
9. Farrington, N., Porter, G., Radhakrishnan, S., Bazzaz, H. H., Subramanya, V., Fainman, Y., Papen, G., & Vahdat, A. (2010). Helios: A hybrid electrical/optical switch architecture for modular data centers. *ACM SIGCOMM Computer Communication Review*, 40(4), 339–350. <https://doi.org/10.1145/1851182.1851223>
10. Chen, K., Singla, A., Singh, A., Ramachandran, K., Xu, L., Zhang, Y., Wen, X., & Chen, Y. (2014). OSA: An optical switching architecture for data center networks with unprecedented flexibility. *IEEE/ACM Transactions on Networking*, 22(2), 498–511. <https://doi.org/10.1109/TNET.2013.2253120>
11. Kachris, C., & Tomkos, I. (2012). A survey on optical interconnects for data centers. *IEEE Communications Surveys & Tutorials*, 14(4), 1021–1036. <https://doi.org/10.1109/SURV.2011.122111.00069>
12. Poutievski, L., Mashayekhi, O., Tikhomirov, A., Rojas-Cessa, R., Mokhtari, M., & Vahdat, A. (2022). Jupiter evolving: Transforming Google's datacenter network via optical circuit switches and software-defined networking. *Proceedings of the ACM SIGCOMM 2022 Conference*, 66–85. <https://doi.org/10.1145/3544216.3544265>

13. Ghobadi, M., Mahajan, R., Phanishayee, A., Devanur, N., Kulkarni, J., Ranade, G., Blanche, P. A., Rastegarfar, H., Kilper, D., & Brendel, R. (2016). Projector: Agile reconfigurable data center interconnect. *Proceedings of the 2016 ACM SIGCOMM Conference*, 216–229. <https://doi.org/10.1145/2934872.2934911>
14. Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, 367(6481), 984–986. <https://doi.org/10.1126/science.aba3758>
15. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. <https://doi.org/10.18653/v1/P19-1355>
16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
17. Zhou, Z., Dzieciol, H., Clark, K., Luo, Y., Richardson, D., Poletti, F., ... & Liu, Z. (2023). Low-latency wavelength-switched clock-synchronized intra-data center interconnects enabled by hollow core fiber. *Optics Express*, 31(15), 24739-24748.
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
19. Barroso, L. A., Clidaras, J., & Hölzle, U. (2013). *The datacenter as a computer: An introduction to the design of warehouse-scale machines*. Morgan & Claypool Publishers. <https://doi.org/10.2200/S00516ED2V01Y201306CAC024>
20. Greenberg, A., Hamilton, J., Maltz, D. A., & Patel, P. (2008). The cost of a cloud: Research problems in data center networks. *ACM SIGCOMM Computer Communication Review*, 39(1), 68–73. <https://doi.org/10.1145/1496091.1496103>