

Adaptive Bandwidth Allocation for Optical Interconnects in Cloud Computing Infrastructure

Agnolo Lettiere

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.

lettieri1982@unr.edu

Jamal Salomonsson

Department of Computer Science, University of Central Florida, Orlando, FL, USA.

salomonsson2005@ucf.edu

Fainche Perri

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.

fainche.perri612@colostate.edu

Fergus McCay

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

mccay975@unh.edu

Abstract

Cloud computing infrastructure is increasingly constrained not by computation alone but by the capacity, latency, and energy behavior of the interconnect fabric that binds servers, storage, accelerators, and tenants into a coherent system. Optical interconnects have become central to this fabric because they offer high bandwidth, low loss, and distance-insensitive transmission, yet their static provisioning often lags behind the bursty and heterogeneous demands of cloud workloads. This paper examines adaptive bandwidth allocation for optical interconnects as a system-level problem that spans physical-layer flexibility, network control, resource governance, and sustainability. It argues that adaptive allocation should be understood less as a narrow optimization of link utilization and more as a socio-technical capability that shapes fairness, robustness, deployment economics, and environmental accountability. The discussion analyzes architectural foundations, including wavelength switching, optical circuit switching, and hybrid electrical-optical topologies, before turning to control-plane requirements, telemetry, and learning-based management. It then considers structural trade-offs among latency, capacity, energy, and cost, and examines governance questions that arise when bandwidth becomes dynamically allocated across tenants, services, and regions. The paper also explores robustness, fairness, and sustainability implications, using cross-domain comparisons from wide-area traffic engineering and warehouse-scale computing. It concludes that adaptive optical bandwidth allocation must be designed with explicit policy choices, transparent operational metrics, and resilient fallback mechanisms if cloud infrastructure is to remain efficient, equitable, and environmentally credible.

Keywords

optical interconnects; adaptive bandwidth allocation; cloud computing; data center networks; software-defined networking; sustainability; fairness; governance.

1. Introduction

Cloud computing has evolved from a collection of virtualized servers into a planetary-scale infrastructure in which computation, storage, and communication are tightly coupled [1]. In this setting, the interconnect fabric is no longer a passive utility. It determines how quickly distributed applications can coordinate, how effectively storage tiers can be shared, and how gracefully services can absorb failures or demand surges. Electrical interconnects have historically dominated short-reach data center networking, but their scaling limits in bandwidth density, reach, and energy per bit have made optical interconnects increasingly attractive [2]. Optical links can carry enormous aggregate capacity over distances that would otherwise require complex electrical regeneration, and they enable architectural separation between switching logic and transmission media [3]. Hybrid electrical-optical designs have shown that optical resources can be introduced selectively where they provide the greatest benefit, rather than replacing all electrical links at once [4].

Recent experimental work has demonstrated that wavelength-switched, clock-synchronized intra-data center interconnects using hollow core fiber can reduce latency while maintaining flexible optical paths [5]. Such developments are significant because they suggest that optical bandwidth can be reconfigured on time scales relevant to cloud workloads, not merely on the provisioning cycles of physical infrastructure. Nevertheless, adaptive bandwidth allocation remains difficult in practice. Cloud traffic is bursty, multi-tenant, and often unpredictable; different services demand different combinations of latency, throughput, jitter, and isolation. A fabric that reallocates bandwidth aggressively may improve utilization but risk instability, while a fabric that remains conservative may preserve predictability but waste expensive optical capacity. The central argument of this paper is that adaptive bandwidth allocation for optical interconnects should be treated as a socio-technical design problem. It involves not only algorithms and control loops but also governance, fairness, sustainability, and operational policy.

2. Architectural Foundations of Optical Interconnects in Cloud Infrastructure

The architectural foundation of optical cloud interconnects begins with topology. Traditional data center networks often use hierarchical electrical switching, with access, aggregation, and core layers arranged to provide high bisection bandwidth and predictable path diversity [6]. As cloud scale increased, operators adopted Clos-like fabrics and centralized control to manage large numbers of servers and links [7]. Optical interconnects can be inserted at multiple points in this hierarchy. They may replace long-reach links between pods, augment core switching with optical circuit switches, or provide reconfigurable paths between racks and disaggregated resource pools. Each placement changes the control problem. Replacing a fixed electrical link with a reconfigurable optical path introduces new opportunities for capacity matching, but it also introduces switching latency, wavelength contention, and the need for synchronization between endpoints.

Wavelength division multiplexing is one of the main mechanisms that make optical bandwidth allocation flexible. Multiple wavelengths can share a single fiber, and each wavelength can be routed independently through optical components such as arrayed waveguide gratings, reconfigurable optical add-drop multiplexers, and wavelength-selective switches. Silicon microring resonators and related photonic devices have been investigated for dense, energy-efficient wavelength routing and modulation [13]. However, the practical flexibility of these devices depends on thermal stability, fabrication variation, and control complexity. Optical circuit switching offers another form of adaptability by establishing high-capacity paths for relatively long durations [10]. Such circuits are well suited to elephant

flows, backup traffic, or bulk data movement, but they are less effective for small, latency-sensitive messages. Hybrid architectures therefore often combine packet-switched electrical paths with circuit-switched optical paths, allowing each traffic class to use the medium that best matches its temporal and spatial characteristics.

The architecture also determines where adaptation can occur. If optical bandwidth is allocated at the rack level, the control plane may aggregate traffic from many servers and make coarse-grained decisions. If allocation occurs closer to the server or accelerator, the system can respond to finer-grained demand but must manage many more endpoints and may incur higher overhead. Disaggregated memory and accelerator pools further complicate the picture because they create high-bandwidth, low-latency dependency chains that cross traditional server boundaries [12]. In these environments, adaptive optical interconnects are not simply links between switches; they are part of a resource fabric that must coordinate bandwidth with compute placement, storage access, and scheduling policy. The architectural choice therefore shapes how much freedom the control plane has, how quickly it can react, and what kind of fairness or isolation guarantees can be offered to tenants.

3. Adaptive Bandwidth Allocation as a Control and Resource Management Problem

Adaptive bandwidth allocation is fundamentally a control problem. The system must observe demand, infer future needs, decide how to assign wavelengths, circuits, or time slots, and then enforce those decisions across optical and electrical domains. Software-defined networking provides a useful conceptual foundation because it separates forwarding from control and allows centralized or logically centralized decision-making [19]. In wide-area cloud networks, software-driven traffic engineering has demonstrated that centralized control can improve utilization while maintaining service objectives [8]. Similar principles apply inside data centers, but the timescales are shorter and the traffic matrix is more volatile. A control plane for optical interconnects must therefore combine fast local reactions with slower global optimization.

Telemetry is the first requirement. Without accurate measurements of link utilization, queue occupancy, flow completion time, and application-level service level objectives, adaptive allocation becomes speculative. However, exhaustive telemetry can itself consume network and computation resources. A practical design must choose which signals to collect, at what granularity, and for how long. It must also distinguish between transient bursts that should not trigger reconfiguration and sustained shifts that justify changing optical paths. In multi-tenant clouds, telemetry has privacy and governance implications because traffic patterns can reveal tenant behavior, workload cycles, or business relationships. Aggregation, differential privacy, and contractual limits on measurement may be necessary to prevent operational visibility from becoming a form of surveillance.

Control policies can be rule-based, optimization-based, or learning-based. Rule-based systems are transparent and predictable, but they may fail to capture complex interactions among workloads. Optimization models can incorporate capacity, latency, and energy constraints, but they often require assumptions about demand that are difficult to validate. Deep reinforcement learning has been explored for resource management in data centers and networks because it can learn policies from experience [16]. For example, learning-based traffic optimization has been proposed to manage datacenter-scale traffic and improve application performance [17]. Experience-driven networking similarly uses learned models to adapt to changing conditions [18]. These approaches are promising, but they introduce challenges of safety, explainability, and operational trust. A learned policy that reallocates

optical bandwidth in unexpected ways could create instability or violate tenant isolation. Therefore, adaptive allocation should include guardrails, fallback modes, and human-auditable policy constraints.

4. Structural Trade-offs in Capacity, Latency, and Resource Efficiency

The most visible trade-off in adaptive optical bandwidth allocation is between capacity utilization and latency predictability. Optical circuits can provide very high throughput, but establishing or switching them takes time. Wavelength switching may be fast in advanced systems, yet it is not instantaneous. If the control plane switches paths too frequently, it may spend more time reconfiguring than transmitting, and applications may experience jitter or transient loss. If it switches too slowly, it may fail to capture short-lived opportunities to relieve congestion. The design challenge is to match reconfiguration granularity to traffic behavior. Bulk transfers, checkpointing, and backup jobs can tolerate slower adaptation, while distributed machine learning synchronization and financial trading workloads may require much tighter latency bounds.

A second trade-off concerns energy and cost. Optical transmission can be more energy-efficient per bit over distance than electrical alternatives, but the supporting components, lasers, cooling, and control electronics also consume power. Adaptive allocation can reduce energy by powering down unused wavelengths or routing traffic over fewer active elements [14]. However, frequent reconfiguration may increase control overhead and component wear. The environmental benefit of adaptive optics depends on the entire lifecycle, including manufacturing, deployment, maintenance, and decommissioning. Cloud operators increasingly face pressure to report energy use and carbon intensity, and data center energy estimates have been recalibrated as workloads and efficiency measures change [15]. Adaptive bandwidth allocation should therefore be evaluated not only by throughput gains but also by net energy and carbon effects.

A third trade-off involves modularity and vendor diversity. Optical interconnects often rely on proprietary management interfaces, specialized transceivers, and vendor-specific switching hardware. This can slow innovation because operators may struggle to integrate multi-vendor optical domains into a single adaptive control plane. Open interfaces and standardized telemetry can reduce lock-in, but they may also constrain performance optimization. The structural choice is whether to prioritize best-of-breed optical performance or composable, policy-driven infrastructure. In cloud environments, where tenants expect portable services and operators expect efficient resource pooling, the balance often shifts toward openness and predictable interfaces, even if peak efficiency is slightly lower.

5. Deployment, Governance, and Operational Policy

Deployment of adaptive optical interconnects is not merely a technical rollout. It is an organizational and contractual process. Cloud providers must decide which services receive adaptive bandwidth, how tenants are charged, and what guarantees accompany dynamic allocation. If bandwidth is allocated dynamically, a tenant may receive excellent performance at one moment and degraded service at another. Without clear policy, this variability can be perceived as unfair or unreliable. Governance mechanisms should specify minimum guarantees, burst allowances, priorities, and remedies when service objectives are missed. They should also define how decisions are logged and how tenants can audit or contest allocation outcomes.

Operational policy must address failure modes. Optical components can fail, wavelengths can drift, and control planes can misconfigure paths. A robust system needs graceful degradation, such as falling back to electrical paths or statically provisioned optical circuits when adaptation is unavailable. It also needs isolation between tenants so that one workload cannot monopolize reconfigurable bandwidth or cause cascading congestion. In regulated sectors, such as finance and healthcare, adaptive allocation may need to respect data residency, auditability, and nondiscrimination requirements. These constraints may limit how freely bandwidth can be moved across regions or tenants, even when the optical fabric could technically support it.

Governance also extends to sustainability reporting. If an operator claims that adaptive optics reduce energy consumption, the claim must be supported by measurement boundaries, assumptions, and verification methods. Otherwise, efficiency gains may be overstated or shifted to other parts of the infrastructure. Policy instruments such as internal carbon pricing, energy-aware scheduling, and transparent sustainability dashboards can align adaptive allocation with environmental goals. Yet such instruments must be designed carefully to avoid perverse outcomes, such as moving traffic to regions with cleaner electricity but higher latency, thereby harming service quality without achieving meaningful system-level reductions.

6. Robustness, Fairness, and Sustainability Implications

Robustness in adaptive optical interconnects requires more than redundancy. It requires the ability to detect anomalies, isolate faults, and reconfigure around failures without violating service commitments. Because optical paths may carry aggregated traffic from many tenants, a single failure can affect a large population. The control plane must therefore maintain a model of dependencies and risk. It should avoid creating single points of failure in the adaptation logic itself. Distributed decision-making, stable fallback configurations, and conservative default policies can reduce the chance that a mis-optimization becomes a widespread outage.

Fairness is especially challenging in multi-tenant clouds. Static bandwidth allocation offers predictable shares, while adaptive allocation may improve aggregate efficiency at the cost of individual variability. A fair system might guarantee minimum bandwidth, cap burst usage, or allocate according to weighted priorities. However, fairness is not only a technical metric; it reflects business relationships, regulatory obligations, and social expectations. A cloud provider may prioritize emergency services, paid premium tiers, or internal control traffic. These choices should be explicit rather than hidden in an optimization objective. Otherwise, adaptive bandwidth allocation can reproduce or amplify existing inequalities in access to digital infrastructure.

Sustainability adds another dimension. Data centers consume significant electricity, water, and land, and their environmental footprint is a matter of public concern [15]. Optical interconnects can contribute to sustainability by reducing energy per bit and enabling longer-lived, higher-capacity infrastructure [14]. Adaptive allocation can further help by matching active capacity to demand. Yet the rebound effect is real: if bandwidth becomes cheaper and more available, applications may consume more of it, increasing total energy use. A sustainability-oriented design should therefore combine adaptive optics with workload scheduling, carbon-aware routing, and demand management. It should also consider embodied carbon in optical transceivers and switches, not just operational electricity.

7. Case Illustrations and Cross-Domain Comparisons

Wide-area traffic engineering offers a useful comparison. Software-driven WAN systems have shown that centralized optimization can improve utilization while respecting service priorities [8]. Globally deployed systems have demonstrated that dynamic path selection can operate at scale, but they also require careful failure handling and policy constraints [9]. These lessons transfer to intra-data center optical fabrics, where the control loop is faster but the number of flows and the sensitivity to latency are greater. In both domains, the hardest part is not computing an optimal assignment but making the system stable, auditable, and fair under real operational uncertainty.

Optical circuit switching in data centers provides another comparison. Reconfigurable optical interconnects can create high-capacity paths for demanding traffic, but they require coordination with packet-switched networks to avoid bottlenecks at the edges [10]. Hybrid architectures have explored this coordination by using optical circuits for long-lived flows and electrical packets for short-lived traffic [4]. Similarly, agile reconfigurable data center interconnects have experimented with dynamic topology changes to match traffic demand [11]. These cases suggest that adaptive bandwidth allocation works best when it is embedded in a broader traffic management architecture rather than treated as an isolated optical function.

Cloud computing research also offers lessons about multi-tenancy and resource abstraction. Cloud platforms must balance utilization, isolation, and cost across diverse customers [20]. Adaptive optical bandwidth allocation inherits this challenge. If tenants can observe or influence allocation, they may game the system. If they cannot, they may distrust it. The design must therefore provide appropriate visibility, predictable contracts, and mechanisms for redress. In this sense, optical bandwidth is not just a physical resource; it is a governed service with economic and social meaning.

8. Future Directions and Research Agenda

Future research should pursue several interrelated directions. First, control algorithms must become more robust to distribution shift and adversarial conditions. Learning-based methods are attractive, but they need safety constraints, interpretability, and formal fallback guarantees. Second, telemetry must become more efficient and privacy-preserving. Fine-grained measurement is valuable, but it should be designed with tenant consent, aggregation, and data minimization in mind. Third, optical and electrical control planes should converge around open interfaces that allow multi-vendor operation without sacrificing performance. Fourth, sustainability metrics should be integrated into adaptive allocation at the design stage, not added as an afterthought. This requires lifecycle assessment, carbon-aware scheduling, and transparent reporting.

Another direction is the integration of adaptive optics with disaggregated and heterogeneous computing. As accelerators, memory pools, and specialized processors become more distributed, the interconnect must support dynamic affinity between computation and data. Adaptive bandwidth allocation could help by creating temporary high-capacity paths for data-intensive phases and releasing them when demand subsides. However, this requires coordination with schedulers and application runtimes. Without such coordination, optical adaptation may optimize the network while leaving the application bottleneck elsewhere. Finally, policy research should examine how adaptive bandwidth allocation affects competition, privacy, and digital equity. Cloud infrastructure is increasingly a foundation for public services, and its governance choices have consequences beyond the data center.

9. Conclusion

Adaptive bandwidth allocation for optical interconnects in cloud computing infrastructure is a promising but demanding capability. It can improve utilization, reduce latency, and align network capacity with dynamic workload needs, but it also introduces complexity, variability, and new governance challenges. The physical layer provides flexible wavelengths, circuits, and hybrid paths, while the control plane must decide when and how to reconfigure them without undermining stability or fairness. Structural trade-offs among capacity, latency, energy, and cost mean that no single optimization objective is sufficient. Deployment and policy choices determine whether adaptation serves tenants equitably or simply maximizes provider efficiency. Robustness requires fallback modes and failure-aware design. Sustainability requires lifecycle thinking and awareness of rebound effects. Ultimately, adaptive optical bandwidth allocation should be treated as a socio-technical infrastructure problem. Its success will depend not only on photonic and algorithmic advances but also on transparent governance, operational discipline, and a commitment to fair and sustainable cloud services.

References

1. Barroso, L. A., Hölzle, U., & Ranganathan, P. (2018). *The datacenter as a computer: Designing warehouse-scale machines* (3rd ed.). Springer.
2. Kachris, C., & Tomkos, I. (2012). A survey on optical interconnects for data centers. *IEEE Communications Surveys & Tutorials*, 14(4), 1021–1036.
3. Benner, A. F., Ignatowski, M., Kash, J. A., Kuchta, D. M., & Ritter, M. B. (2005). Exploitation of optical interconnects in future server architectures. *IBM Journal of Research and Development*, 49(4/5), 755–775.
4. 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.
5. 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.
6. Al-Fares, M., Loukissas, A., & Vahdat, A. (2008). A scalable, commodity data center network architecture. *ACM SIGCOMM Computer Communication Review*, 38(4), 63–74.
7. Singh, A., Ong, J., Agarwal, A., Anderson, G., Armistead, A., Bannon, R., ... & Vahdat, A. (2015). Jupiter rising: A decade of Clos topologies and centralized control in Google's datacenter network. *ACM SIGCOMM Computer Communication Review*, 45(4), 183–197.
8. Hong, C. Y., Kandula, S., Mahajan, R., Zhang, M., Gill, V., Nanduri, M., & Wattenhofer, R. (2013). Achieving high utilization with software-driven WAN. *ACM SIGCOMM Computer Communication Review*, 43(4), 15–26.
9. Jain, S., Kumar, A., Mandal, S., Ong, J., Poutievski, L., Singh, A., ... & Vahdat, A. (2013). B4: Experience with a globally-deployed software defined WAN. *ACM SIGCOMM Computer Communication Review*, 43(4), 3–14.

10. Poutievski, L., Mashayekhi, O., Tariq, M., Zeghal, R., Li, H., O'Keefe, R., ... & Vahdat, A. (2022). Jupiter evolving: Transforming Google's datacenter network via optical circuit switches and software-defined networking. In *Proceedings of the ACM SIGCOMM 2022 Conference* (pp. 66–85). ACM.
11. Ghobadi, M., Gaudette, J., Gunning, P., & Rajan, D. (2016). Projector: Agile reconfigurable data center interconnect. In *Proceedings of the 2016 ACM SIGCOMM Conference* (pp. 216–229). ACM.
12. Rumley, S., Bahadori, M., Polster, R., Hammond, S. D., Calhoun, D. M., Wen, K., Rodrigues, A., & Bergman, K. (2017). Optical interconnects for extreme scale computing systems. *Parallel Computing*, 64, 65–80.
13. Bogaerts, W., De Heyn, P., Van Vaerenbergh, T., De Vos, K., Kumar Selvaraja, S., Claes, T., Dumon, P., Bienstman, P., Van Thourhout, D., & Baets, R. (2012). Silicon microring resonators. *Laser & Photonics Reviews*, 6(1), 47–73.
14. Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, 367(6481), 984–986.
15. Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. *Nature*, 561(7722), 163–166.
16. Mao, H., Alizadeh, M., Menache, L., & Kandula, S. (2016). Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks* (pp. 50–56). ACM.
17. Chen, L., Lingys, J., Chen, K., & Liu, F. (2018). AuTO: Scaling deep reinforcement learning for datacenter-scale automatic traffic optimization. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication* (pp. 191–205). ACM.
18. Xu, Z., Tang, J., Yin, C., Wang, Y., & Xue, G. (2018). Experience-driven networking: A deep reinforcement learning based approach. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications* (pp. 1871–1879). IEEE.
19. Kreutz, D., Ramos, F. M. V., Veríssimo, P. E., Rothenberg, C. E., Azodolmolky, S., & Uhlig, S. (2015). Software-defined networking: A comprehensive survey. *Proceedings of the IEEE*, 103(1), 14–76.
20. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1), 7–18.