Implementation of Robotic Optical Fiber Switches for Network Diagnostics and Other Data Center Use Cases

Alan Gibbemeyer, Anthony Kewitsch, Robert Shine and Ramiro Voicu
Telescent, Inc., 16832 Red Hill Ave., Irvine, CA 92606 USA
shine@telescent.com

Abstract: Optical fiber switches can add diagnostic capability to maintain performance in high-speed optical link, reduce power and cost in hyperscale front-end networks, and improve the training speed in machine learning clusters. © 2022 The Authors.

1. Introduction and Use Cases

While new hyperscale data centers are constantly being built, older facilities remain in use. Although compute, storage and switch hardware can be refreshed during a technology upgrade cycle, the existing fiber infrastructure is re-used and, in some instances, can be over a decade old. Links can have from 1 to 6 hops through manual patch panels within data centers, may have been disconnected and reconnected multiple times and may have degraded performance in terms of increased loss or reflections. The transition to higher-speed transceivers combined with higher-density modulation formats such as PAM4 increases performance requirements, reduces link budgets, and increases susceptibility to multi-path interference (MPI). Since record keeping is notoriously poor and subject to human errors, many links have unknown optical performance. The sheer number of links in a hyperscale data center with tens of thousands of servers makes manual characterization daunting. Replacing the manual patch panels with a reconfigurable optical switch allows diagnostic equipment to be inserted in the link, allowing testing to be done automatically and remotely.

While the above discussion centered on fiber connectivity issues within a data center, many of the same issues occur between data centers. Since construction and installation are a significant cost of fiber deployment, networks today are built using high-count fiber cables which can contain from 864 to 3,456 fibers per cable. These high count strands are expected to be consumed over many years, with only a portion of the fibers being lit initially. Monitoring these fibers is a challenge due to the distributed nature of the network, again creating a need for remote and automated fiber testing and reconfiguration.

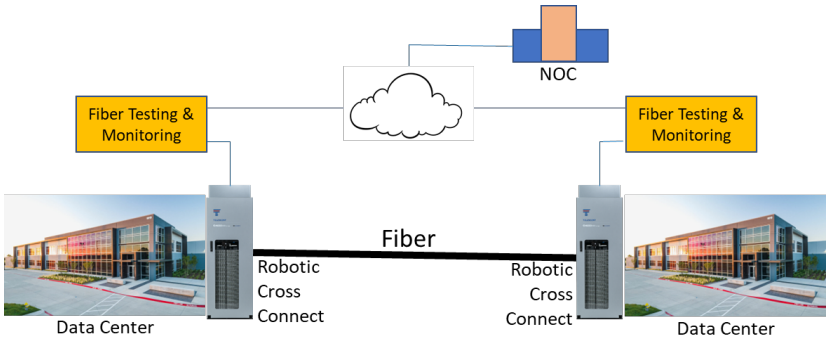

Fig. 1. Use of a robotic cross-connect to insert diagnostic equipment into network (inside or outside fiber plant) to test performance.

In addition to the need for automated testing for fiber networks, recent results highlight the value of all-optical switches in hyperscale datacenter fabrics. Google recently published impressive results discussing how they evolved their typical multi-stage Clos-based electrical switch architecture to a direct-connect topology using optical circuit switches within the machine aggregation blocks in hyperscale data centers [1]. Reported advantages of this approach include significantly improved restriping speed, greater flexibility in the network architecture and bit-rate independent switching allowing the use of multiple transceiver generations. Additional benefits include a 30% reduction in cost and a 41% reduction in power for the networking fabric [2]. Since the main use of the optical switch was to manage data center expansions, it was found that block level topology reconfigurations more often than every few weeks yielded limited benefits [1].

Beyond this use in the front-end Clos network, optical fiber switches can also improve performance of machine learning (ML) clusters. As the datasets and xPU cluster sizes have increased for training, worker communication has become a significant bottleneck. Even with just a few tens of parallel workstations, the amount of time required for worker communication exceeds that required for computing [3]. Reconfiguring the communication links between xPUs using optical switches can offer significant benefits, for example when employing model parallelism to run different ML algorithms. For model parallelism, the data exchange between workers is not uniform across the cluster but is predictable and stable during the full training run. There are workers that exchange more data between iterations and optimizing the bandwidth between these high-load workers can improve the overall speed of the training run considerably. Researchers have demonstrated a 3 to 4x improvement in training time through optimization of the bandwidth between workers [4].

2. Technology Options

The desire for an all-optical circuit switch goes back decades and many technology options have been pursued. These include free-space options such as MEMS and piezo-based devices as well as all-fiber systems using robotics to reconfigure the fiber. Challenges in the implementation of the technologies included the additional optical losses, scaling to high-port count and cost as well as the ability to scale manufacturing and reliability to meet the hyperscale data center requirements [2].

In addition, most of the prior technologies were an NxN switch that required deployment of the complete system, raising the initial deployment cost. For many of the use cases above, an upgradeable system where the number of ports deployed match the current need would also advantages. Even beyond an upgradeable system, an asymmetric deployment can offer additional benefits. This is the case for the high-count fiber deployment where only a portion of the fibers may be used initially. Having the ability to connect diagnostic equipment to any of the fibers while limiting the initial deployment cost is extremely beneficial in this application.

3. All-Fiber Robotic Cross Connect

This all-fiber cross-connect system consists of a fiber interconnection volume bounded at opposite ends by parallel planes separated by a short distance (Fig. 1). Within this volume, a large number (100's to 1000's) of optical fiber strands linking inputs to outputs intermix. The first plane coincides with an input terminal array, where the connectorized optical fiber strands internal to the cross-connect are interfaced with external fiber patch cords, and the second plane coincides with an intermediate port array internal to the cross-connect, through which these same optical fiber strands pass to the output connector array. While often this system is configured in a symmetrical NxN configuration, this design offers a unique advantage of an asymmetrical MxN configuration. This asymmetrical configuration offers value in applications such as in multi-tenant data center meet-me-rooms or in dark fiber management of high-count fiber deployments where the number of active fibers may be a small percentage of the total fibers connected to the system.

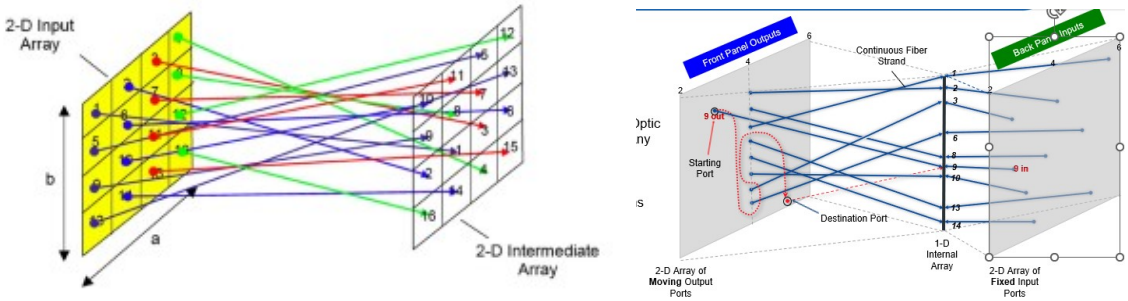

Fig. 2. Schematic diagram showing back panel to front panel fiber connectivity (left) with a simplified routing path shown around other fiber strands (right).

For arbitrary reconfiguration of the fiber matrix, an algorithm based on the theory of knots, braids and strands (KBS) has been developed [5]. This enables one end of a fiber within the interior of the input array to be maneuvered so that this fiber weaves through the interconnect volume without entangling with other fibers. This allows any fiber optic strand to be arbitrarily reconfigured to a new output port regardless of the configuration of

the surrounding strands or prior configuration state of the system. This KBS algorithm enables a simple pick-and-place robot to execute an unlimited number of arbitrary reconfigurations while linearly scaling to thousands of ports, rather than the onerous N^2 scaling of crossbar array switches.

A physical embodiment of the robotic patch panel has been developed with up to 1,200 duplex connectors per rack. Further increases in port count can be achieved through the use of smaller connectors such as SN-type connectors, smaller diameter fibers or multi-core fibers. The optical performance of this robotic optical switch is identical to a manual patch panel with 2 UPC connectors and a 3 meter length patch cord. The typical insertion loss through the system is 0.3 dB, with a maximum insertion loss of 0.5 dB. The return loss is consistent with UPC polished single mode connections (-50 dB). Low-low is maintained throughout the life of the system through the use of an automated cleaning process used for all reconfigurations. APC connectors are available as an option in the system with improved optical return loss of -65dB if required.

The reliability of the system is based on its latching design and therefore equals the reliability of a manual patch panel. If the robotic system loses power for any reason, all connections remain active. Also, since all the mechanical elements of the system are outside of the fiber matrix, any failed component can be replaced without affecting existing traffic on any interconnection. The reliability of the system has been tested through multiple customer trials and has been deployed in production networks. An engineering system has been running continuously for over 3 years and has performed over 320,000 reconfigurations.

The Telescent system has been deployed in production networks, including a nationwide service provider of lit and dark fiber services. For this deployment, the system was installed in the asymmetric configuration to allow testing of all 864 fibers in the fiber cable while initially requiring only 1 robotic patch panel module containing 96 ports, reducing the initial capital expense by 50%. The system also included an OTDR (Exfo Fiber Guardian 750) which allowed remote monitoring and testing of the fiber.

While the robotic OCS offers low loss, high reliability and an 8x higher port count in the current configuration than the MEMS OCS which greatly reduces the complexity of implementation, one limitation of the robotic OCS is the switching speed. The robotic switch will take from 40 seconds to 4 minutes to make a new connection. However, the use case for the MEMS OCS discussed above is managing data center expansion which occurs on the order of weeks which is very manageable with the robotic OCS [1].

In summary, robotic optical fiber switches find many use cases in data centers and fiber optic networks. Applications include monitoring and diagnostics of high-speed fiber networks particularly as the link requirements become more stringent, managing data center expansion, and improving performance in machine learning clusters. While the robotic approach does not offer the speed achievable with MEMS technology, the robotic system offers lower optical loss, much higher port count and significantly lower cost while offering the speed required for managing capacity expansions in hyperscale data centers. With these benefits, it is anticipated that optical fiber switches would be used increasingly as data centers transition to new architectural designs.

[1] L. Poutievski et. al. “Jupiter Evolving : Transforming Google’s Datacenter Network via Optical Circuit Switches and Software-Defined Networking,” SIGGCOMM’22 August 22-26, 2022 Amsterdam, Netherlands.

[2] R. Urata et. al., “Mission Apollo: Landing Optical Circuit Switching at Datacenter Scale,” SIGGCOMM’22 August 22-26, 2022 Amsterdam, Netherlands.

[3] C. Li et. al., “1-Bit Lamb: Communication Efficient Large Scale Large-Batch Training with Lamp’s Convergence Speed,” https://arxiv.org/pdf/2104.06069.pdf

[4] W. Wang et. al., “TOPOOPT: Optimizing the Network Topology for Distributed DNN Training,” USENIX Symposium on Networked Systems Design and Implementation (**NSDI** '22) April 4-6, 2022, https://doi.org/10.48550/arXiv.2202.00433

[5] Anthony Kewitsch, “Large Scale, All-Fiber Optical Cross-Connect Switches for Automated Patch-Panels,” J. Lightwave Tech. Vol. 27, No. 15 (August 1, 2009).