Apollo: Large-Scale Deployment of Optical Circuit Switching for Datacenter Networking

Ryohei Urata, Hong Liu, Kevin Yasumura, Erji Mao, Jill Berger, Xiang Zhou, Cedric Lam, Roy Bannon, Darren Hutchinson, Daniel Nelson, Leon Poutievski, Arjun Singh, Joon Ong, and Amin Vahdat
Google LLC

Abstract: In this paper, we describe Apollo, to the best of our knowledge, the world's first large-scale production deployment of optical circuit switches (OCSes) for datacenter networking. We review the underlying hardware technologies including the design of our internally developed OCS and WDM transceivers.
OCIS codes: (060.0060) Fiber optics and optical communications; (060.4250) Networks

1. Introduction

Over the past few decades, hyperscale datacenters have enabled new, global-scale Internet services, from web search, e-commerce, social media, to the public cloud. Datacenter networking provides the interconnectivity to scale the underlying compute functions powering these services. While industry best practices have focused on pure packet solutions employing Clos topologies as the basis for large-scale datacenter networks [1], the research community has developed a number of potentially game changing network designs around the premise of incorporating OCSes into the datacenter network [2]. OCS technologies hold a number of benefits relative to electrical packet switches (EPSes), being data rate and wavelength agnostic, low latency, and energy efficient. However, the lack of the required optical technologies at datacenter volumes and scale inhibited adoption. In this paper, we present the design and implementation of Apollo, which we believe is the first large-scale deployment of optical circuit switching for datacenter networking. Apollo enables topology engineering, fabric expansion, and rapid tech refreshes for right-sized, lowest cost networks utilizing the most modern interconnect and EPS technologies.

The Apollo OCS hardware platform consists of a home-grown, internally developed OCS (Palomar), customized wavelength-division-multiplexed (WDM) optical transceiver technology, and optical circulators to enable bidirectional links through the OCS. Apollo serves as the backbone of all of our datacenter networks, having been in production for nearly a decade and delivering substantial improvements to the cost efficiency, power consumption, and application performance of our infrastructure [3].

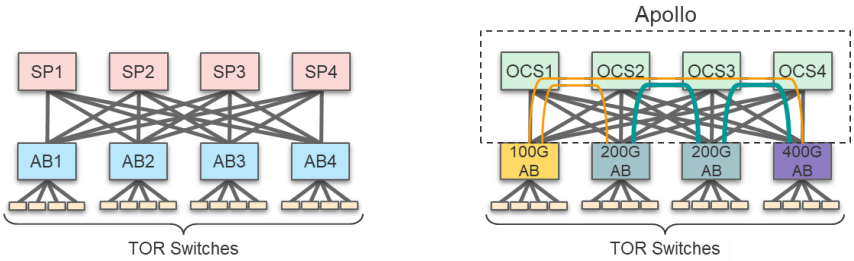

Fig. 1: a) Traditional datacenter network comprising EPSes only. b) Apollo-based datacenter network, with ABs connecting directly through OCSes. Different colored ABs consist of different speed/generation optical transceivers and switch ASICs. Colored lines illustrate the logical connections between ABs, with different colors indicating different speeds.

2. Apollo-based Datacenter Network

Fig. 1a) illustrates a traditional datacenter network, with spine blocks (SPs) connecting aggregation blocks (ABs) [1]. All blocks are EPS-based, composed of interconnect (copper and optical transceivers) and switch ASICs. Fig. 1b) illustrates the evolved datacenter network incorporating the Apollo OCS layer. The Apollo layer replaces all SPs for significant cost and power savings through elimination of the electrical switches and optical interfaces that are used to implement the spine layer while adding new capabilities not possible with EPSes alone:

Topology Engineering: When moving towards spine-free architectures, with the elimination of routing capability by the spine, reconfiguration of network topology for each specific application and corresponding traffic patterns is desirable in addition to traditional electrical-switch-level traffic engineering. As an example,

reconfiguration of the fabric connectivity allows maximization of inter-AB bandwidth in the event of an increase in long-lived traffic demand (i.e., elephant flows) between two particular ABs or multiple ABs.

Fabric Expansion: With expansion capability, the size of the network fabric can be augmented incrementally, allowing an initial number of ABs to be deployed with additional ABs added to the fabric as needed (pay as you grow). For each expansion event, Apollo automates the necessary large-scale reconfiguration/re-striping of interconnectivity between ABs through software control of the OCSes.

Rapid Tech Refresh: The above expansion capability leads to the possibility of mixing ABs based on different speed/bandwidth generations of optical transceivers and switch ASICs, as illustrated in Fig. 1b). Interoperability between heterogeneous ABs is ensured through the compatibility of optical transceiver specifications across multiple generations. This enables faster introduction of new technology, as well as incremental tech refresh. The OCS-based expansion and interconnect interoperability thus enable optimally sized datacenter networks incorporating the latest and lowest cost per bandwidth optical/networking technologies.

3. Apollo hardware components: OCS, WDM transceivers, Circulators

We developed three critical hardware components, the OCS, WDM transceivers, and optical circulators, to realize a cost-effective, large-scale optical switching layer. WDM optics maximize the efficiency and usage of OCS ports. Single mode operation is needed for compatibility and scaling of the OCS technology and the reach to support a physically large infrastructure. Furthermore, circulators are coupled to the optical transceivers to operate links in a bidirectional manner [3]. Full duplex communication is thus achieved for each single strand of fiber and OCS port, doubling the effective OCS radix and halving the required number of fibers and OCS ports. With the OCS, circulator, and single mode fiber being largely data rate agnostic, they can be used across multiple generations of networking and interconnect to amortize their costs.

Palomar OCS: Fig. 2a) shows the high-level optical design and operation principles of the Palomar OCS. The input/output optical signals enter the optical core through two-dimensional (2D) fiber collimator arrays that consist of an NxN fiber array and 2D lens array. The optical core consists of two sets of Micro-Electro-Mechanical-Systems (MEMS) mirror arrays. Each inband optical signal traverses through a port in each collimator array and two MEMS mirrors, as indicated by the green line. Mirrors are actuated and tilted to steer the optical signal to a corresponding input/output collimator fiber. The entire end-to-end optical path is broadband and reciprocal, for data rate agnostic and bidirectional communication across the OCS. Superposed with the inband signal path, a monitoring channel (thick red arrows) assists with tuning of the mirrors. Each MEMS array is injected with this 850nm light. The reflected monitor signals are received at a matching camera module. Control hardware/firmware utilizes the camera image feedback to optimize MEMS actuation to minimize signal loss. A pair of injection/camera modules controls each MEMS array.

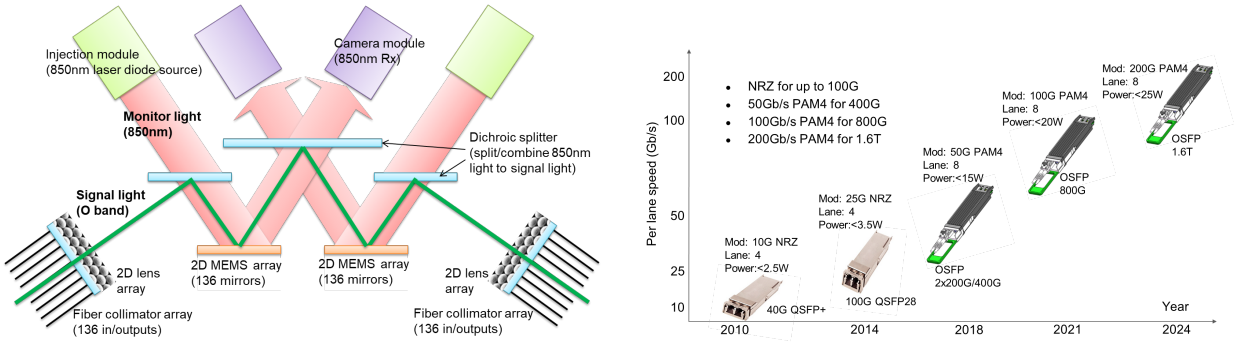

Fig. 2: a) Illustration of Palomar OCS optical core and design. b) Google WDM interconnect review/roadmap.

Use of a single camera per MEMS array significantly simplifies the mirror control scheme in comparison to conventional approaches which require individual monitoring and/or photodetector hardware per mirror. This design choice was critical to realizing a low-cost, manufacturable OCS solution. The above design yields a non-blocking, 136x136 OCS with bijective, any-to-any input to output port connectivity. The maximum power consumption of the entire system is 108W, which is a fraction of the power of an EPS system with the same switch capacity. With the above hardware architecture, tens of thousands of 136x136 duplex port OCS were manufactured and deployed. In

terms of performance, insertion losses are <2dB for all NxN permutations of connectivity, including a tail in the distributions due to splice and connector loss variation. Return loss is typically -46dB, with a nominal spec of <-38dB. The stringent return loss requirement stems from the use of bidirectional communication along each optical path, as any single reflection superposes directly on top of the main optical signal to degrade signal-to-noise ratio.

WDM transceivers: As bandwidth requirements scaled in the datacenter, adoption of WDM was critical to improve the cabling efficiency and reach needed for a large and ever-growing compute infrastructure. This pushed the need to optimize performance, cost, and develop an industry ecosystem befitting the datacom application. Fig. 2b) reviews the corresponding WDM single mode interconnect technologies developed over the past decade [4]. From 40Gb/s QSFP+ to the latest 800Gb/s OSFP, the pluggable transceiver bandwidth has grown by 20x, with continuous improvements in cost, energy efficiency, and density. Notable technology directions and innovations for application to Apollo are:

Optical Interference Mitigation (OIM): Circulator-based bidirectional links impose unique physical impairments as reflections along the path cause multi-path-interference (MPI) and in-band optical crosstalk. Tight specification of return loss of interfaces and transmitter extinction ratio, dynamic range, and relative-intensity noise (RIN) are necessary to support such links at scale. For high-speed IC/electrical technologies, we migrated from analog-based clock-and-data recovery (CDR) solutions to digital signal processing (DSP)-based ASICs starting from 50Gb/s per-lane PAM4 modulation. The DSP ASIC not only provided a more robust, scalable solution, it also allowed implementation of digitally-based algorithms that mitigate the optical interference impairments.

Optical link budget: To support the higher loss budget due to the OCS and circulators, we started with transceiver designs emphasizing high transmitter power (directly modulated lasers (DMLs)), and low optical component (wavelength mux/demux) and packaging losses (opto-mechanical design). In addition to standard forward error correction (FEC) implemented at the host, a concatenated (inner) FEC was embedded into the DSP ASIC to further increase link budget at higher data rates.

Backward compatibility: One additional and critical element of the interconnect development was the requirement for backward compatibility of the optical transceiver technology. The current generation transceiver must support a superset of transmitter and receiver performance of all previous generations. This includes optical performance specifications (transmitter power, receiver sensitivity/overload), wavelength grid compatibility (CWDM4), and a data path capable of running at the various line rates.

Circulators: Before we drove high volume use in the datacenter, application of circulators was mostly in the C-band wavelength range with fairly limited volumes. Use of proper optical coatings and optical re-design allowed extension of circulator capability to O-band/CWDM4 wavelength operation. Reducing return loss and enhancing directivity were critical, as corresponding stray light is effectively equivalent to having a reflection in the link.

4. Conclusion

In this paper, we motivated the use of optical switching for datacenter networking with our previous work on datacenter networks and industry-wide proposals outlining a variety of use cases and the advantages of optical switching for such applications. We then presented Apollo: an optical switching layer for datacenter networks at scale, comprising key hardware components of the OCS, WDM optical transceivers, and circulators. These capabilities combined have allowed us to achieve: 1) the lowest cost networks at scale, with pay-as-you-grow capability, 2) fastest adoption of the latest generation optics and networking that we develop, and 3) dynamic network reconfiguration capabilities for highest efficiency and/or performance.

Acknowledgments

The authors would like to acknowledge the Apollo and Palomar teams, Platforms, Net Infra, Google Global Networking (GGN), Global Capacity Delivery (GCD), Datacenter Operations, Site Reliability Engineering (SRE), Program Management Organization (PMO), and industry/vendor partners for making Apollo a reality at scale.

References

[1] A. Singh, et al., “Jupiter rising: A decade of clos topologies and centralized control in Google’s datacenter network,” ACM SIGCOMM 2015.
[2] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, “The emerging optical data center,” OFC 2011.
[3] L. Poutievski, et al., “Jupiter evolving: Transforming Google’s datacenter network via optical circuit switches and software-defined networking,” ACM SIGCOMM 2022.
[4] H. Liu, et al., “Evolving requirements and trends in datacenter networks,” Springer handbook of optical networks. Springer, 2020.