A High-Radix, Low-Latency Optical Switch for Data Centers

Dan Alistarh, Hitesh Ballani, Paolo Costa
Microsoft Research

Adam Funnell, Joshua Benjamin,
Philip Watts, Benn Thomsen
University College London

ABSTRACT

We demonstrate an optical switch design that can scale up to a thousand ports with high per-port bandwidth (25 Gbps+) and low switching latency (40 ns). Our design uses a broadcast and select architecture, based on a passive star coupler and fast tunable transceivers. In addition we employ time division multiplexing to achieve very low switching latency. Our demo shows the feasibility of the switch data plane using a small testbed, comprising two transmitters and a receiver, connected through a star coupler.

CCS Concepts

•Networks → Bridges and switches; Data center networks;

Keywords

Optical switching; TDMA; WDM;

1. INTRODUCTION

Today's datacenter networks are built using low radix electrical packet switches. They scale by arranging the switches in a multilayer topology (e.g., folded Clos or Fat Tree) [1,3]. However using low radix switches as a building block means that, in large datacenters, a very high number of switches, cables and transceivers are needed. This increases network cost, power and complexity. Lots of cables are particularly problematic as server NICs are upgraded from 10 Gbps to 25 and 100 Gbps — links throughout the network need to be expensive optical cables, and all switches need optoelectronic transceivers to convert optical signals to electronic and back at each hop [5].

These problems can be alleviated through higher radix switches supporting 1000s of ports per switch. In large datacenters, operators already use high-radix chassis switches at higher levels of the hierarchy. Internally, these switches use low radix switching chips connected in a Clos topology. Such chassis switches can economically scale to ∼250 ports. However, building a *single* electrical switching chip with a high radix and high per port bandwidth is hard because of limitations on bandwidth at the edge of the chip, mostly because of power constraints [2]. This is not expected to improve in the near future as the ITRS roadmap predicts only a modest increase in the chip pin count and per-pin bandwidth over the next decade [2].

We thus target an all-optical datacenter network built using high-radix optical switches. We will demonstrate a switch design that can scale to a thousand ports with high per-port bandwidth (25 Gbps+) and low switching latency (40 ns, best case). The switching latency here is two to three orders of magnitude less than that of the 2D MEMS-based wavelength selective switches underlying recent hybrid network designs [5,8]. The reduced switching latency means we can switch at the granularity of a few 64 byte packets, and avoids the need for two separate networks.

2. DESIGN

While traditional electrical switches have a switching chip that dynamically connects the input and output ports, we distribute the switching functionality among the nodes connected by the switch. This allows the switch to have a very simple optical core. Overall, our design is based on two key technologies.

First, the switch's core comprises a passive optical device called *star coupler* [4]. A star coupler simply replicates an optical signal on any of its input ports to all output ports. This creates a cost and power-efficient "broadcast and select" network providing one hop connectivity between all nodes connected to it. Second, we use fast tunable transceivers that can tune to a wavelength in 200 ns [6].

The broadcast nature of the star coupler means that, by default, only a single node can transmit across it at any given time. To achieve packet switch like functionality across it, we allow many node pairs to communicate across the coupler simultaneously by combining wavelength and time division multiplexing. This, as we explain below, is achieved through the tunable transceivers.

2.1 Switch components

The switch comprises nodes connected to a star coupler. A "node" can be a server or even a Top-of-Rack switch. Thus the switch can be used as a modular building block in data centers, to connect servers or other switches too. Below we briefly describe the switch design, including the data and control plane setup.

Wavelength multiplexing. A star coupler is colourless, i.e., many wavelengths can be multiplexed through it. An optical fiber can support 160 wavelengths using the C and L-band with 50 GHz spacing between the wavelengths. We equip each node with a wavelength tunable transceiver, and periodically change the wavelength that a transceiver is tuned to. It takes <200 ns to re-tune a transceiver during which period it is not operational. To amortize the tuning latency, the transceivers are tuned at most once every 2 μs, thus imposing a throughput overhead of less than 10%. We call the period after which transceivers are re-tuned an *epoch*.

SIGCOMM '15 August 17-21, 2015, London, United Kingdom

ACM ISBN 978-1-4503-3542-3/15/08.

DOI: http://dx.doi.org/10.1145/2785956.2790035

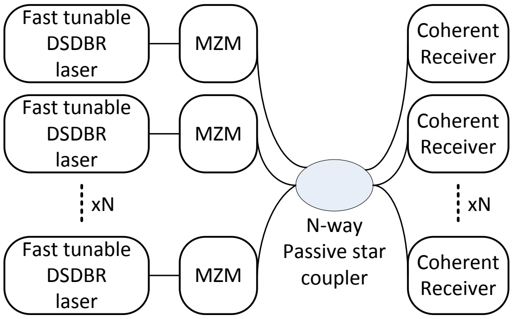

Figure 1: Hardware layout of the switch's data plane, where MZM is Mach-Zehnder modulator and *N* is the total number of nodes.

Timeslot multiplexing. Since the number of wavelengths across the fiber is significantly less than the number of node pairs, simply relying on dynamic wavelength multiplexing would result in high switching latency. To reduce the switching latency, we use time division multiplexing. Each epoch is divided into many fixed-length *slots*. Multiple node pairs are assigned the same wavelength and thus, can communicate with each other in the same epoch, but never in the same slot. In our prototype, there are 50 slots per epoch, so each slot is 40 ns long. This is the best case switching latency. Such timeslot multiplexing relies on fine-grained time synchronization across the nodes, and the ability of a transceiver to send short bursts of an optical signal at a pre-specified time.

Switch Capacity. To keep transceiver cost low, we use interleaved bipolar line coding at 25 Gbps. Each switch port thus has a bandwidth of 25 Gbps. With 160 wavelengths, the total capacity of the switch is 3.64 Tbps (after accounting for tuning overhead). We can scale up the switch capacity in two ways: *i)* through better modulation which allows each port to have 100 Gbps+ bandwidth, and *ii)* by using more than one star coupler.

Switch Scale. The number of ports supported by a star coupler is dictated by its splitting loss. Specifically, the power lost when transmitting through the coupler is given by $L_{dB} = 3 \text{ dB} \times \log_2 N$, where N is the number of ports. We are using a (DSP free) coherent receiver with a sensitivity of -27 dBm. Assuming a transmit power of 6 dBm, and system and coupling losses of 3 dB means that we have a total optical loss budget of 30 dB. Thus, we can support a 1,000 ports — $3 \text{ dB} \times \log_2 N \leq 30$ dB means that $N \leq 1000$.

2.2 Switch data plane

The switch's data plane is shown in Figure 1. Each node has a transceiver with transmit and receive functionality. The transmitter part contains a fast tunable laser modulated by a Mach-Zehnder Modulator (MZM). We use DS-DBR lasers that have been shown to tune and stabilize within 200 ns [6]. The transmitter is connected to a passive star coupler through an optical fiber. A clock signal using an out of band 1.3 um channel is distributed to all nodes over the same star coupler for synchronisation across the entire network.

The receive part is a DSP-free coherent receiver with a fast tunable local oscillator. This provides full wavelength selectivity. The local oscillator at the receiver is tuned to a single wavelength at the start of each epoch, and can receive all data on that wavelength for the duration of the epoch. Furthermore, this design ensures that the transmitter and receiver are independently tunable.

2.3 Switch control plane

The switch control plane, at the beginning of each epoch, assigns wavelengths to transceivers on all nodes based on the expected traffic demand. We use past traffic statistics and queuing information at the end hosts to estimate the expected traffic for the next epoch. The control plane also determines a timeslot schedule for the epoch, i.e., for each wavelength, determine which pair of nodes that have been tuned to the wavelength is allocated each timeslot.

We are developing an algorithm for this wavelength and timeslot assignment problem with three main design goals: high switch utilization, bounded worst-case performance and freedom from starvation. These goals typically underlie scheduling algorithms even in traditional low radix switches [7]. Beyond this, we want the algorithm to be simple enough to be implemented in hardware.

3. DEMONSTRATION DETAILS

The goal of the demo is to demonstrate the practical feasibility of the switch data plane. In summary, we will show *i)* the ability of our prototype to tune the lasers to a different wavelength within 200 ns and hence, achieve WDM, *ii)* the effectiveness of TDM to achieve low switching latency, and *iii)* the scalability of our approach.

The demonstration comprises two experiments using a small testbed consisting of two transmitter nodes and a single receiver node, all connected to a star coupler. We use an arbitrary waveform generator connected to the transmitter to generate traffic and an oscilloscope connected to the receiver to verify that the signal received matches the modulated input data.

In the first experiment, we tune the lasers of one of the transmitter and the receiver on a common wavelength and we start generating traffic between the two. Then, we re-tune both nodes to a different wavelength λ and show that within 200 ns the receiver can receive traffic again.

In the second experiment, instead, we tune all three nodes on the same wavelength and we configure the transmitters to send back-to-back using alternate 20-ns time slots. This shows the feasibility of TDM at a very fine-grain time scale as required by our design in order to achieve low switching latency.

In both experiments, we artificially introduce the appropriate transmission loss and interfering channels to emulate the conditions of our target scale of 500-1,000 nodes.

4. REFERENCES

[1] M. Al-Fares, A. Loukissas, and A. Vahdat. A Scalable, Commodity Data Center Network Architecture. In *SIGCOMM*, 2008.

[2] N. Binkert, A. Davis, N. P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, and J. H. Ahn. The role of optics in future high radix switch design. In *ISCA*, 2011.

[3] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: A Scalable and Flexible Data Center Network. In *SIGCOMM*, 2009.

[4] Q. Li, S. Rumley, M. Glick, J. Chan, H. Wang, K. Bergman, and R. Dutt. Scaling Star-coupler-Based Optical Networks for Avionics Applications. *IEEE/OSA Journal of Optical Communications and Networking*, 5(9), 2013.

[5] H. Liu, F. Lu, A. Forencich, R. Kapoor, M. Tewari, G. M. Voelker, G. Papen, A. C. Snoeren, and G. Porter. Circuit Switching Under the Radar with REACToR. In *NSDI*, 2014.

[6] R. Maher, D. Miller, S. Savory, and B. Thomsen. Fast Wavelength Switching 112 Gb/s Coherent Burst Mode Transceiver for Dynamic Optical Networks. In *ECOC*, 2012.

[7] N. McKeown. The iSLIP Scheduling Algorithm for Input-queued Switches. *IEEE/ACM Trans. Netw.*, 7(2), 1999.

[8] G. Porter, R. Strong, N. Farrington, A. Forencich, P.-C. Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat. Integrating Microsecond Circuit Switching into the Data Center. In *SIGCOMM*, 2013.