High Port Count Hybrid Wavelength Switched TDMA (WS-TDMA) Optical Switch for Data Centers

Adam Funnell[1], Joshua Benjamin[1], Hitesh Ballani[2], Paolo Costa[2], Philip Watts[1], Benn C. Thomsen[1]
[1] *Optical Networks Group, Dept of Electronic & Electrical Eng, UCL (University College London), London, WC1E 7JE, UK*
[2] *Microsoft Research, Station Road, Cambridge, CB1 2FB, U.K.*
adam.funnell.13@ucl.ac.uk

Abstract: A WS-TDMA optical switch fabric scalable to 1024 ports is demonstrated. Fast tunable (<200ns) transceivers using bipolar encoded transmission and coherent reception enable wavelength switching and TDMA over a passive star network for low latency packet level switching.
OCIS codes: (060.4259) Networks, packet-switched; (060.4263) Networks, star.

1. Introduction

Current data center networks are built using hierarchical tree structures of low port count electrical switches [1]. The resulting layered topologies are complicated, expensive and energy intensive with high and variable network latencies. Using fewer but higher port count switches in flatter architectures will reduce network complexity, cost and latency, by reducing the total number of switches and packet hops required in a given network. Thus an ideal future switch for data centers has a large port count (>1000 ports) with a high per-port bitrate (>10Gb/s), providing consistently low latency whilst being able to flexibly share a large total bandwidth across all ports.

Increases to the port count and per-port bitrates of electronic switches are limited by integrated circuit pin density and front panel bandwidth constraints. Considering optics, passive switch designs using MEMS [2,3] or WDMA/TDMA PONs can scale to large numbers of high bitrate ports, but such systems are not rapidly reconfigurable to optimally share the total switch bandwidth. Fast switching technologies using active optical elements (SOAs and MZIs) exist, however such switches currently only scale to 128 ports in integrated form [4]. Designs based on Arrayed Waveguide Grating Routers (AWGRs) are highly scalable [5], but laser tuning times dominate the packet latencies and the colored nature of AWGRs does not allow the decoupling of time and wavelength domains for flexibility.

To provide maximum flexibility and make the best use of all available bandwidth, the system design presented here uses a passive fibre star coupler to directly connect a large number of 10Gb/s tunable optical transceivers (up to 1024), for low latency, single-hop communication between any two nodes. Previous works have studied the use of star couplers with both WDM [6,7], and combined time and wavelength division multiplexing (TWDM) [8,9], for the routing of data flows. However, this work is to our knowledge the first to employ hybrid wavelength switched (WS-)TDMA over a star coupler with both the transmitters and receivers independently and rapidly tunable for fast switch reconfigurability. Moreover, this system requires no active components beyond commercially available tunable lasers, Mach-Zehnder modulators and DSP-free coherent receivers. This work presents the design and evaluation of the data plane for a low latency, high port count switch, suitable for connecting servers or top-of-rack switches.

2. System Design

The system (fig. 1a) comprises N transmitters, each of which consists of a fast tunable laser bipolar modulated by a Mach-Zehnder modulator (MZM) and coupled into one side of an N-way passive star coupler. N coherent receivers are connected to the other side of the passive coupler, each with a fast tunable local oscillator (LO) for full wavelength selectivity and high sensitivity. Each coherent receiver subsystem (fig. 1b) consists of a polarization diverse 90° optical hybrid, with balanced photodetectors providing an electrical output of the received electric field I and Q components of each polarization. After band-pass filtering (BPF), the 4 electrical signals are independently squared, and summing these 4 resulting signals gives a unipolar NRZ signal. A matched filter and decision circuit recovers binary data.

All of the N transmitters and receivers can be tuned to any wavelength on the ITU 50GHz grid within 200ns to establish a connection that is maintained for an "epoch" of 2μs before retuning. However, within the C-band this only provides 80 wavelengths through the switch at any given time. Using WDM alone thus limits the number of nodes

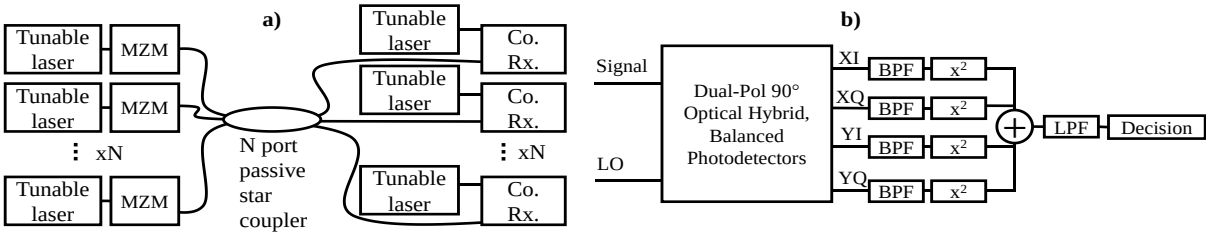

Fig. 1: a) Physical design. MZM = Mach-Zehnder modulator, N = total transceivers and Co. Rx. = Coherent Receiver; **b)** Co. Rx. subsystem.

that can send data in each epoch to N or 80, whichever is smaller. For fine-grained sharing of the total switch bandwidth and to reduce the latency in establishing a connection between any two nodes, TDMA is also used.

Each epoch is further split into T timeslots with up to M transmitters sharing the same wavelength during an epoch. The LO of each coherent receiver is also tuned to a single wavelength for the duration of an epoch, in order to receive all T timeslots on that wavelength during that epoch. Although up to M transmitters may be tuned to the same wavelength, only one transmitter per wavelength is modulated during each TDMA timeslot (M≤T). The remaining (M-1) transmitters which would otherwise interfere with the data channel are extinguished by biasing the MZM at its null point, reducing the power of each interfering channel by the extinction ratio of the modulator.

The bit-rate overhead lost to laser tuning time is minimized through the use of fast tunable DSDBR lasers [10] as transmitter sources and as LOs at each coherent receiver. These lasers have been shown to tune and stabilize within 200ns [11] (only a 10% overhead on each 2µs tuning epoch). However, by only allowing such a short tuning time, no time is available for precise frequency correction, and tuning alignment between the transmitters and LO is only guaranteed within a ±500MHz range. These spectral offsets between the signal, LO and any unmodulated transmitters (even if reduced in power by the MZM extinction ratio) at the same wavelength will give rise to interference terms in the spectrum of the final received signal in the region ±500MHz around the notional optical carrier frequency. If not removed, these interference terms vastly degrade BER performance for any M>1.

To allow for appropriate spectral filtering at the receiver, interleaved bipolar line coding (IBLC) [12] is applied at the transmitter, which spectrally shapes the data creating a null in the modulated spectrum at ±500MHz around the central optical frequency. This technique also ensures that there is no optical carrier (assuming a perfect modulator), and is preferable to 8b10b encoding or similar codes since it incurs no overhead. All signals from unmodulated interfering transmitters are expected to lie in this central null region, and the band-pass filter in fig. 1b is thus designed to remove all spectral content within ±500MHz of the nominal carrier, along with high frequency noise outside the data bandwidth. This filtering, along with the squaring and summing of each signal, can be entirely performed in the electrical RF domain. This removes the need for receiver DSP, reducing cost and power consumption.

3. Experimental Setup

As shown in fig. 2, fast tunable DSDBR lasers were used as a transmission source (Tx1) and as an LO for a coherent receiver. Tuning currents were supplied to each DSDBR laser by 250MS/s arbitrary waveform generators (AWGs) to allow fast wavelength switching. A second transmitter (Tx2) consisted of two external cavity laser sources (ECLs) and was used to emulate two nodes assigned to different wavelengths, λ_1 (1555.26nm) and λ_2 (1562.56nm). Tx1 and Tx2 were modulated using MZMs, with 10GS/s AWGs providing the 10Gb/s IBLC electrical data signals.

A further pair of unmodulated ECLs provided interference channels at λ_1 and λ_2, to emulate up to (M=11) additional transmitters on the switch. All sources were joined by a 4x1 star coupler to a single receiver. Optical attenuators were used to emulate the losses of a higher port count coupler.

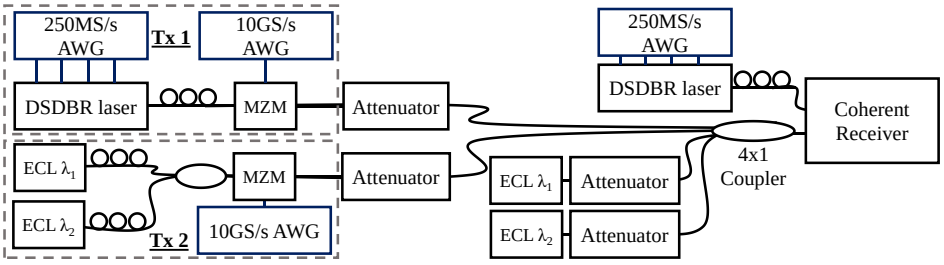

Fig. 2: Experimental setup: AWG=Arbitrary Waveform Generator, MZM=Mach-Zehnder Modulator, ECL=External Cavity Laser.

4. Results

The power sensitivity of each receiver limits the maximum number of transceivers (N) supported by each star coupler. Fig. 3a shows the achievable BER as a function of the received signal power for a single data channel. This BER was inferred from the Q-factor of a continuous random bit sequence of 2^20 bits, using only Tx1 (DSDBR laser) constantly at λ_1 without interfering ECLs. To achieve a BER of 10^{-12}, a minimum received power of -24.5dBm is required. The minimum power lost by each channel in an ideal N-port star coupler ($P_{lost(dB)}$) is $P_{lost(dB)} = 3 \times \log_2(N)$. For +6dBm transmit power, a received power of -24dBm allows $P_{lost(dB)}$=30dB, supporting N=1024 nodes at 10Gb/s.

To determine the number of transmitters that can be tuned to the same wavelength during each epoch with only the extinction ratio of the modulator to reduce their power (by 12.3dB in this experiment), a single unmodulated ECL of variable power was added to Tx1 also at λ_1. Simulations showed that a single interfering channel at a higher power has a greater effect on BER than multiple lower power interfering channels spread over the ±500MHz frequency range, verifying this experimental technique. After converting the absolute power of the ECL to an equivalent number of transmitters extinguished by MZMs, fig. 3b shows the BER of Tx1 at a received power of -24.5dBm as the number

of additional transmitters at λ_1 is increased. A small BER penalty is observed for up to 10 interfering channels at the same wavelength, thus allowing at least 11 nodes to be assigned to the same wavelength per epoch.

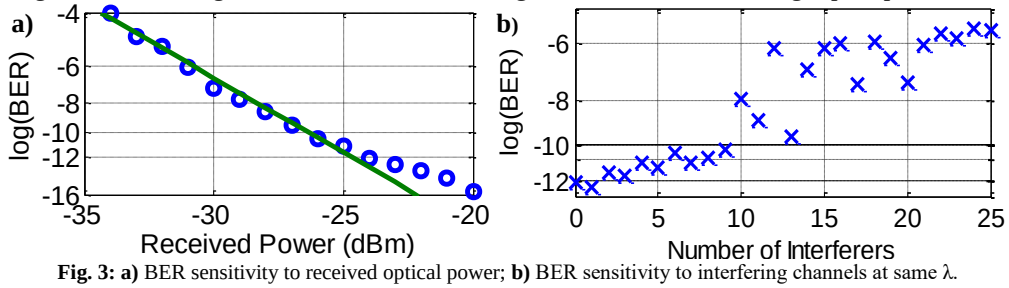

Fig. 3: a) BER sensitivity to received optical power; **b)** BER sensitivity to interfering channels at same λ.

To demonstrate the entire system, the DSDBR lasers (Tx1 and receiver LO) were switched between λ_1 and λ_2 each 2μs epoch (with tuning time consistently below 200ns), whilst Tx2 modulated ECL sources at both λ_1 and λ_2 simultaneously. The two transmitters were time-division multiplexed with data in alternate timeslots from each transmitter, with T=50 slots per epoch. Each 40ns slot contained 399 bits of random data, with 1 bit guard spacing, minimising TDMA overhead. The attenuators were set to emulate the loss from N=1024 ports in a star coupler, and the unmodulated ECL powers were set to emulate 10 additional transmitters on each λ. Fig. 4 shows: a) 2 full epochs captured at the coherent receiver; b) 200ns λ switching between epochs; c) TDMA with 1-bit guards; and d) an eye diagram of error-free performance of the switch over a whole timeslot.

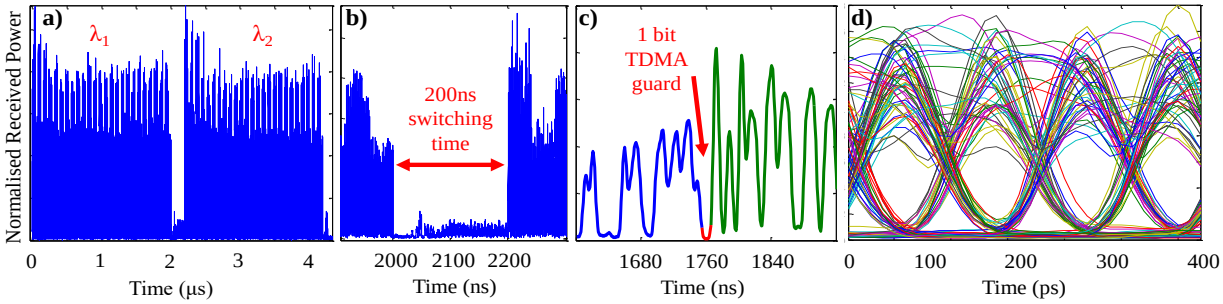

Fig. 4: a) 2 full epochs; **b)** detail of 200ns λ switching time; **c)** detail of TDMA guard**; d)** eye diagram for a single timeslot.

5. Conclusions

A 1024 port optical switch data plane combining wavelength switching and TDMA over a passive fibre star coupler for packet level reconfigurability has been demonstrated. Fast tunable transceivers and coherent reception allow wavelength switching at nanosecond timescales, whilst bipolar line coding and RF filtering support TDMA at packet level for low latency switching and fine grained sharing of switch capacity. Experiments verify single-hop communication at a 10Gb/s BER of 10^{-12}, with up to 10 transceivers sharing a single wavelength using TDMA.

6. Acknowledgements

The support of Oclaro in supplying DSDBR lasers for use in this work is gratefully acknowledged.

7. References

[1] A. Singh et al., "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", SIGCOMM '15, p183-197 (2015)
[2] N. Farrington et al, "Helios: a hybrid electrical/optical switch architecture for modular data centres", SIGCOMM '10
[3] N. Farrington et al., "A 10 μs Hybrid Optical-Circuit/Electrical-Packet Network for Datacenters", Proc. OFC, OW3H.3 (2013)
[4] Q. Cheng et al., "Demonstration of the feasibility of large-port-count optical switching using a hybrid Mach–Zehnder interferometer–semiconductor optical amplifier switch module in a recirculating loop", Opt. Lett. 39 (18), p5244-5247 (2014)
[5] Y. Yin et al., "LIONS: An AWGR-Based Low-Latency Optical Switch for High-Performance Computing and Data Center", Selected Topics in Quantum Electronics, IEEE Journal of, 19 (2), pp.3600409 (2013)
[6] C-J. Chae et al., "Hybrid Optical Star Coupler Suitable for Wavelength Reuse", Phot. Tech. Lett. 10 (2), p279 (1998)
[7] Q. Li et al., "Scaling Star-Coupler-Based Optical Networks for Avionics Applications", J. Opt. Commun. Netw. 5 (9), p945 (2013)
[8] P.J. Wan et al., "TWDM single-hop lightwave networks using multiple fixed transceivers at each station", Local Computer Networks, Proc. 21st IEEE Conf. (1996)
[9] S.-K. Lee et al., "Hypercube interconnection in TWDM optical passive star networks", Massively Parallel Processing Using Optical Interconnections, Proc. of the Second International Conf., (1995)
[10] A. J. Ward et al., "Widely tunable DS-DBR laser with monolithically integrated SOA: Design and performance", J. Quant. Electron. 11, p149-156 (2005)
[11] R. Maher et al., "Fast Wavelength Switching 112Gb/s coherent burst mode transceiver for dynamic optical networks", Proc. ECOC, Tu.3.A.2, Amsterdam (2012)
[12] A. Croisier, "Introduction to pseudoternary transmission codes," IBM J. Res. Dev. 14, p354-367 (1970)