Hybrid Wavelength Switched–TDMA High Port Count All-Optical Data Centre Switch

Adam C. Funnell, *Student Member, IEEE,* Kai Shi, *Member, IEEE,* Paolo Costa, Philip Watts, *Member, IEEE,* Hitesh Ballani, Benn C. Thomsen, *Member, IEEE*

***Abstract*—The physical layer, data plane design of an all-optical network switch capable of scaling to 1024 ports at 25 Gb/s per port is presented and experimentally evaluated. Fast tuning DSDBR lasers modulated with line-coded bipolar data allow combined wavelength switching and TDMA to provide packet switch-like functionality with over 2 Tb/s of total switch bandwidth. A passive fibre star coupler core with high sensitivity, DSP-free coherent receivers creates a low complexity, easily upgradeable building block for data centre networks.**

***Index Terms*—Optical fiber communication, Optical packet switching, Time division multiplexing, Wavelength division multiplexing.**

I. Introduction

As worldwide internet connectivity to homes and businesses increases, IT services such as large scale data storage, intensive data processing and interactive web applications are moving from local computers located on users' premises into vast data centres, with many hundreds of thousands of servers in a single building. This presents challenges not just for long distance data transfer between the end users and data centres, but also for intra-data centre communications where high bandwidth, low latency channels are critical; over 75% of data traffic remains inside the data centre [1] and total traffic is doubling every 12-15 months [2].

Conventional network architectures for data centres, such as folded Clos topologies, use hierarchical layers of low port count electronic switches to create a switching fabric reaching many servers [2]. This is energy intensive and the high number of hops between switches results in high and unpredictable latencies [3]. Current high port count electrical switches are formed from an internal Clos topology of low radix chips, and only small increases in future total pin count per chip and bandwidth per pin are expected due to power constraints at each pin [4]. This makes a single, high port count electronic switch challenging.

Additionally, as the bandwidths of links between electronic switches increases, the distance over which electrical signals can travel is limited by frequency dependent losses in the cabling. Optical links are not distance limited in this way and are scalable to higher bitrates than equivalent electrical networks, meeting expected future demands. For this reason, inter-rack links in large data centres are generally optical. However, the large number of electrical-optical-electrical conversions required for links between low radix electrical switches dominates data centre network cost, suggesting a holistic approach to optical switching for data centres is required, targeting fewer but higher port count switches.

Optical solutions to data centre networking to date include devices based on electro-mechanical switches such as MEMS [5], with slow switching speeds. PON technologies have also been explored [6], however these cannot run at packet switching speeds without optical buffering [7]. Active optical elements such as Semiconductor Optical Amplifiers (SOAs) or Mach-Zehnder Interferometers (MZIs) can be used for fast optical switching, yet to date, integrated devices have only been demonstrated for low port counts [8]. Arrayed Waveguide Grating Routers (AWGRs) can scale to large port counts [9], but are inflexible in wavelength allocation and manipulation resulting in either a high number of components required for wavelength conversion and routing or poor performance under variable data centre traffic patterns [10].

Previously we demonstrated a high port count, low latency, all-optical switch using a combination of wavelength switching and TDMA over a star coupler [11]; however the high loss of the star coupler limited transmission to 10 Gb/s per port, and severe signal BER impairments were observed when TDMA introduced multiple transmitters at the same wavelength. In this work we show that the addition of receiver based equalisation to increase sensitivity enables 25 Gb/s per port while maintaining over 1000 ports per switch and also enables up to 26 transmitters to share a wavelength using TDMA. This work presents detailed experimental characterisations of: 25 Gb/s receiver sensitivity to this signal format including a simple equaliser; the first characterisation of the number of transmitters tolerable per wavelength in this TDMA format; error-free reception in a DSP-free coherent

A. C. Funnell is with the Optical Networks Group, Department of Electrical and Electronic Engineering, University College London, Torrington Place, WC1E 7JE, UK.

P. Watts was with the Optical Networks Group, Department of Electrical and Electronic Engineering, University College London, Torrington Place, WC1E 7JE, UK. He is now with ARM Ltd., Fulbourn Road, Cambridge, CB1 9NJ , UK.

P. Costa and H. Ballani are with Microsoft Research, Station Road, Cambridge, CB1 2FB, UK.

K. Shi and B. C. Thomsen were with the Optical Networks Group, Department of Electrical and Electronic Engineering, University College London, Torrington Place, WC1E 7JE. They are now with Microsoft Research, Station Road, Cambridge, CB1 2FB, UK.

system with full loss budget and TDMA emulation; and time-resolved characterisation of fast wavelength switching for channels across the entire optical C-band. The combined results demonstrate the feasibility of the proposed all-optical data centre switch, capable of scaling to 1024 ports with each node operating at 25 Gb/s, and using all 89 optical C-band wavelengths to optimally share an aggregate switch throughput of over 2 Tb/s.

II. Switch Design

Figure 1 shows the high level physical layer switch design. Instead of an extended fabric formed from many small switches, this network provides the functionality of a single, high port-count switch with direct connectivity between any two nodes by connecting all transmitters and receivers together through a star coupler. This greatly reduces latency by ensuring single-hop connectivity between all nodes. In practice each node could be either an individual server, or an electrical switch aggregating traffic from a multi-server rack.

In order to provide maximum throughput across the star coupler, fast switching WDM is used. Tunable lasers at each transmitter can select any of 89 possible wavelengths on the ITU 50 GHz grid within the optical C-Band, so that up to 89 transmitters, one per wavelength, can simultaneously transmit with minimal cross-talk. The colourless nature of the star coupler broadcasts all transmitters to all receivers, and each receiver selects a single wavelength to receive by tuning the local oscillator of the coherent receiver to match the desired channel. To minimise the time lost to laser tuning, both the transmitter and receiver LO lasers at all nodes simultaneously retune every 2 μs (defined here as an Epoch), with tuning complete within 200 ns. This gives a network reconfiguration overhead of 10%, with total switch throughput of over 2 Tb/s.

If fast switching WDM alone is used for routing within the switch, only 89 transmitters can transmit across the coupler during each Epoch. However, this may result in high latencies when serving connection requests from 1000 or more nodes. For increased granularity in sharing the switch bandwidth, TDM is also used on each wavelength. TDM breaks up each Epoch into individual timeslots which transmitters can request for data transmission.

Each transmitter comprises a fast tunable DSDBR laser and an optical Mach-Zehnder modulator (MZM) for 25 Gb/s bipolar modulation, as shown in figure 2. The DSDBR laser

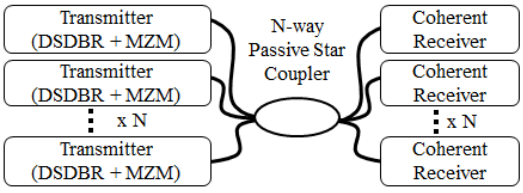

Fig. 1. High level diagram of the switch design. All transmitters and receivers are directly connected in a single hop star network.

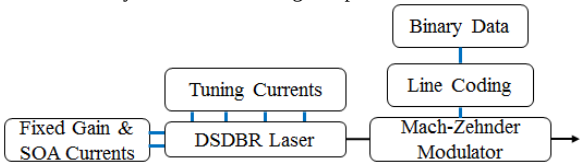

Fig. 2. Transmitter design, showing DSDBR, tuning and gain current supplies, and input signal flow.

used in this work is rapidly tunable through current injection to 8 individual front gratings for mode selection, a rear grating for coarse wavelength control, and a phase section for fine adjustment [12]; the gain and SOA currents are held constant for fast wavelength stabilisation.

Each coherent receiver uses a DSDBR laser as local oscillator (LO) for fast selectivity of received wavelengths, as shown in figure 3. Coherent reception also provides increased receiver sensitivity compared to direct detection, allowing a larger star coupler power split ratio to support more nodes. The receiver hardware comprises a standard dual polarisation optical hybrid, with balanced photodiodes providing electrical outputs of the in-phase and quadrature signals of each polarisation. It is necessary to use a dual polarisation optical hybrid to make the receiver polarisation independent. As the coherent receiver is only used for wavelength selectivity and increased sensitivity over direct detection when receiving intensity modulated signals, only intensity information is recovered.

Figure 3 shows the receiver signal processing: each electrical signal is first independently passed through an initial band pass filter with a 1 GHz-22.5 GHz pass range. This is followed by a squaring operation for intensity detection, before a summing circuit combines all 4 signals. A final 17.5 GHz low pass filter is applied before a decision circuit recovers binary data. Although in this work these functions were performed offline in DSP, passive hardware band-pass filters and signal multipliers can carry out these operations in the electrical RF domain, making receiver DSP unnecessary. This reduces receiver complexity, cost and energy consumption. Practical implementations could use RF frequency multipliers and power combiners, readily available with passbands of 1-20 GHz. Passive analogue equalisers running at 20 GHz have also been demonstrated elsewhere, including clock recovery [13]. This system is also designed to operate without FEC, relying on error-free transmission (10^{-12} BER or better).

By using a combination of the above techniques, although coherent receiver hardware is necessary to provide wavelength selectivity and polarisation diversity, no coherent receiver DSP is required and error-free data can be returned directly from passive electronics following the coherent receiver photodiode outputs. DSP to recover data from full coherent transmission is estimated to use 20W of the 32W available in CFP format transceivers [14]; this DSP-free receiver could therefore operate with at most 37.5% of the power consumption of a comparable fully coherent device. A simplified receiver [15] could be used to provide similar levels of performance, but either at the loss of polarisation diversity or a power budget penalty due to optical component losses, constraining the construction of the fibre network core.

Between the transmitters and receivers, a passive star coupler is used to combine all transmitted signals, and to split this combined transmission identically to every receiver [16].

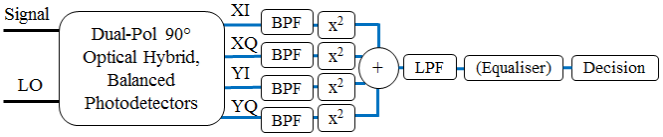

Fig. 3. Coherent receiver design, showing electrical filtering, squaring and summing, followed by equalisation and decision for data recovery.

This device is purely passive as all switching functionality has been moved to the nodes at the edges of the switch; future switch upgrades and repairs are easier with a passive core.

To allow TDM with short guard intervals whilst maintaining transmitter wavelength stability, external shuttering of transmitters is not possible during timeslots in which they are not allocated data transmission rights. For example, reverse biasing the SOA integrated on the DSDBR lasers could attenuate the laser output by approximately 40 dB, however the SOA switching time is limited by the capacitance of the SOA section and the carrier dynamics to several 100ps. In addition the resulting thermal change due to the high current flow causes wavelength drifts over timescales greater than the 40 ns timeslots, leading to poor BER performance [17]. Instead, only the MZMs are held in the "off" state (i.e. transmitting zeros), reducing transmitter output power by the MZM extinction ratio (14.2 dB), providing blanking on a bit rate timescale which does not have any thermal effects on output wavelength.

The finite extinction ratio of the MZM results in some unmodulated power from each laser always passing through the star coupler to all receivers, which creates a signal after coherent reception at the beat frequency between the LO and the unmodulated carrier frequencies tuned to nominally the same channel wavelength. The difference in the LO and transmitter carrier frequencies arises from the small frequency offset (consistently <1 GHz) between the independent transmitter and receiver LO lasers; this cannot be corrected in real time due to only allowing 200 ns for retuning. However, signals arising from these offsets can be removed using high pass filtering at the receiver, provided that the modulated signal from the allocated transmitter is spectrally shaped such that it contains no frequency components below 1 GHz. Figure 4 shows an example of this interference at the Signal-LO frequency offset and ideal filtering around it (where the filtering is performed by the 1 GHz-22.5 GHz band pass filter described above).

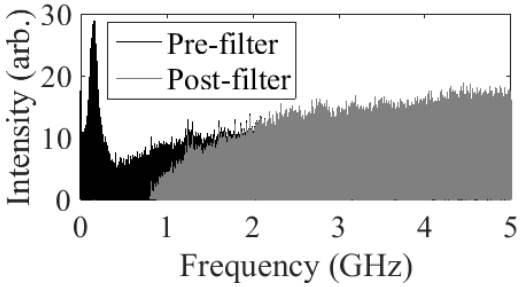

Fig. 4. Received data following summing and squaring, shown before and after a high-pass filter removing the beating interference term due to Signal-LO frequency offset.

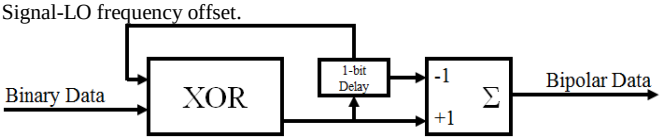

Fig. 5. Bipolar line coding for a DC balanced signal – note this coding is interleaved and applied independently on even and odd data bits to provide further spectral shaping.

In this work Interleaved Bipolar Line Coding (IBLC) [18] is used since it is a DC balanced code providing spectral suppression of a modulated signal for ±1 GHz around DC. As no data modulation is carried in this spectral region, a high pass filter at the receiver can remove any interference arising from Signal-LO offsets of up to ±1 GHz without impact on the data (the value of ±1 GHz is a hard limit imposed by the spectral shaping properties of the IBLC, which only removes data spectral components below this value). IBLC requires no overhead and is simple to implement in hardware (as in figure 5) or software, minimising impact on network throughput and pre-processing requirements respectively. The IBLC is also advantageous as it is directly decoded at the receiver when using square law detection. IBLC coding first splits the data into even and odd binary bit streams, before each bit stream is encoded with alternating positive and negative values for each binary "1", regardless of the number of binary "0"s between them. This provides a fully DC balanced code with bounded digital sum variation of two. The even and odd bit streams are then re-interleaved, resulting in spectral nulls around DC and the symbol rate. Figure 6 shows example transmitted three-level IBLC data and an eye diagram of how this maps back to

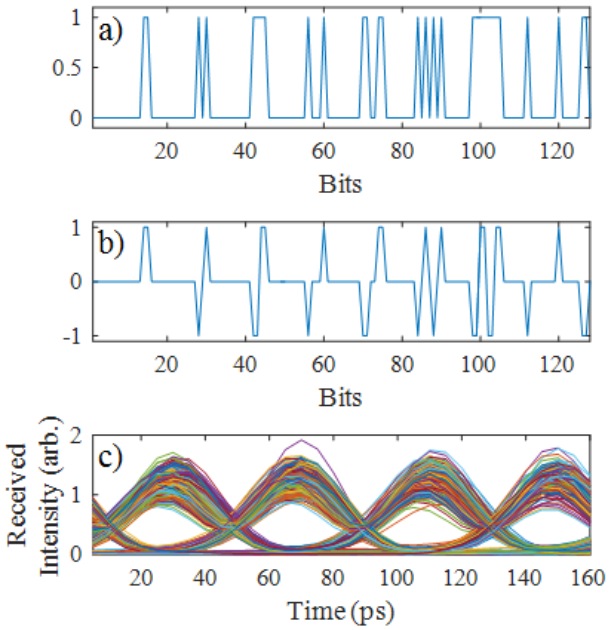

Fig. 6. a) Original binary NRZ data at transmitter; b) IBLC encoding of the same NRZ data at transmitter; c) Eye diagram at the receiver showing received data returned to two-level NRZ form.

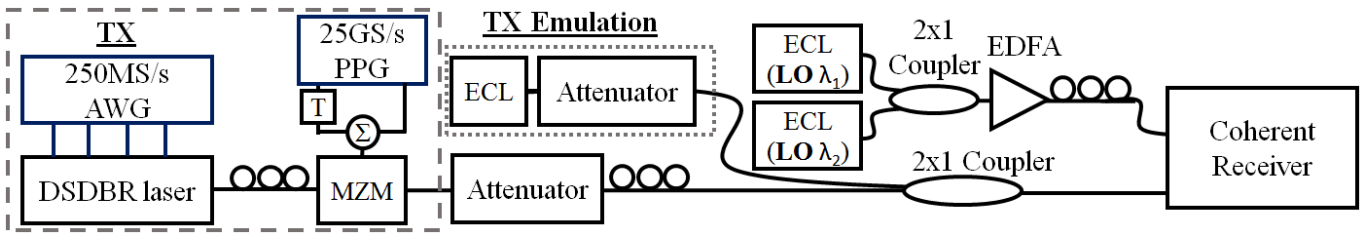

Fig. 7. Full experimental setup. AWG = Arbitrary Waveform Generator; PPG = Pulse Pattern Generator; MZM = Mach-Zehnder Modulator; ECL = External Cavity Laser.

two-level NRZ at the receiver.

The scheduling and synchronisation across the switch is beyond the scope of this work, which presents only the data plane. However, in a practical implementation, a second star coupler can be used to provide communication between each node and a central wavelength and timeslot allocation controller, and to synchronise all nodes. Additional star couplers can operate in parallel to reduce the oversubscription ratio, although the current mean case of 12:1 for this system is comparable to practical data centre traffic requirements [19]. Real-time implementations of receiver clock recovery have been shown elsewhere within 31 ns at 25 Gb/s, comparable to the laser tuning times in this work [20].

To compare the performance of this system to conventional electronic packet switched data centre networks two metrics are considered: power consumption and latency.

Considering network power consumption from the transceiver outwards (i.e. not including NIC functionality prior to the interface), a top-range electrical switch capable of 1152 ports at 25 Gb/s has a total power consumption of 8700 W [21]. For connectivity to and from the nodes attached to this switch, 25 Gb/s active optical cables (AOCs) in the SFP28 form factor are available, each at a power of 1W [22], making a total power consumption of 9724W to reach 1024 ports with an electrical switch. In this all-optical network design only the transceivers themselves require power; the central star coupler element is completely passive. However, the additional complexity of a fast tunable transmitter and receiver compared to the fixed wavelength active cable described above will result in higher power consumption (due to the required fast DACs and additional photodiodes), and it is estimated that a suitable transceiver for this system would consume 5.6 W. For 1024 nodes, this is a total power of 5734 W, a reduction of 41% compared to the state of the art electrical alternative.

To compare the latency of this network to a currently available electrical network, consider building a 1024 port network from electrical switches, each of 256 ports at 25 Gb/s. Up to three hops are required to reach all ports, and at each switch buffering may be required. Switches operate with a shared memory buffer across all ports, so for uniform traffic across many ports the latency performance is greatly reduced due to queuing latencies, with a worst case of milliseconds queuing at each switch when a standard 4 MB buffer at each port is filled at 25 Gb/s. Latency increases dramatically with both port count per switch and network load, particularly at higher layers of a Clos structure of switches [23]. In this switch, there is no central buffering and the latency through the data plane from transmitter to receiver is dictated only by the speed of light. While the network scheduler may add latency (not the focus of this work, but practical scheduling algorithms for this design are under development [24]), there is room for considerable scheduling delay before this system would show the same performance, particularly in terms of worst case tail latency, as an electrical network reaching a high port count.

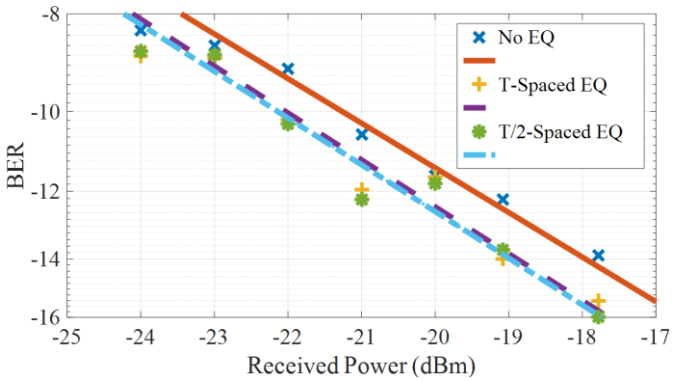

Fig. 8. Receiver BER sensitivity to received optical power for a continuous 25 Gb/s PRBS IBLC data stream.

III. Experimental Results

Figure 7 shows the experimental setup. As a data transmitter (TX), the wavelength switching of a single DSDBR laser was controlled using 250 MSample/s Arbitrary Waveform Generators (AWGs). The laser signal was externally modulated by a MZM with pre-coded IBLC data signals from a 25 GS/s Pulse Pattern Generator (PPG). A single data TX polarisation was used, aligned for equal power arriving on each polarisation photodiode pair at the receiver. This polarization alignment represents the worst case in terms of receiver performance. A pair of ECLs were used as LOs at the receiver (only 1 ECL enabled in section III.A and III.B), and a further ECL with optical attenuator was used in experiment III.B, emulating additional transmitters.

A. *Receiver Power Sensitivity*

To characterise the receiver power sensitivity, a continuous 25 Gb/s stream of a 2^{20} PRBS pattern was encoded into IBLC format and modulated onto the TX DSDBR laser, held at a constant wavelength and connected via a variable optical attenuator to the coherent receiver. A single LO was kept at a constant power of +10 dBm, and the second ECL in figure 7 was not enabled. The attenuator was used to vary the signal power entering the coherent receiver, emulating the losses of a star coupler. No additional noise loading is required, since a star coupler is entirely passive.

System performance was measured using 8x over sampling to recover the optimum sampling instant and provide ideal phase synchronisation (processed offline but can be implemented in hardware through receiver sampling phase adjustment). BER performance was measured in two scenarios: with no further signal processing before a simple decision circuit to recover binary data; and with an adaptive

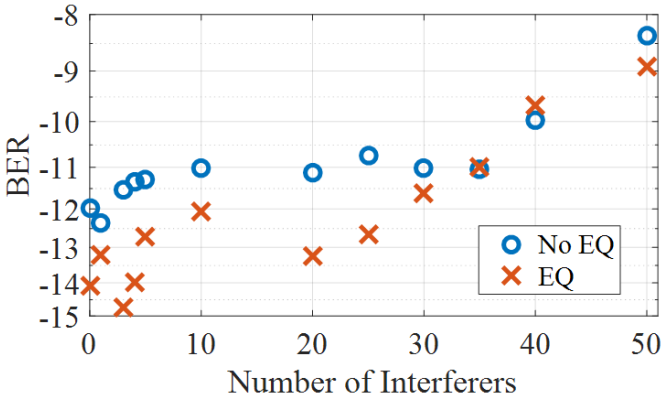

Fig. 9. Receiver BER sensitivity to additional unmodulated "interferers" at the same wavelength for a continuous 25 Gb/s PRBS IBLC data stream.

digital equaliser for enhanced filtering. Two equalisers were tested for each sensitivity measurement: for bit period T, a T/2 fractionally spaced equaliser, requiring oversampling in the receiver; and a T spaced equaliser as found in common off-the-shelf electrical transceiver modules. Both equalisers were 6-tap decision-directed equalisers, implemented offline in this work. Adding a real-time equaliser to each receiver need not result in adding receiver DSP; it is implementable in real time using dedicated low complexity chips as found in standard commercial receivers or in an analogue circuit [13].

Figure 8 shows the receiver power sensitivity, where the received BER is inferred from the Q-factor calculated using the swept-threshold technique [25]. For a BER of 10^{-12} (considered error-free in the Ethernet standard and thus required in this FEC-free system), a received signal power of -19.6 dBm is required at the coherent receiver, without using an equaliser. Given the loss of 30 dB in a 1024 port splitter, and transmitter output power of +14 dBm [12], the loss budget of 33.4 dB is sufficient to allow a 1024 way split (30 dB), with coupling and manufacturing losses (3.4 dB). It can be seen that there is little benefit in oversampling at the receiver to use a T/2 spaced equaliser – this provides only a 0.1dB sensitivity improvement at 10^{-12} over a T-spaced equaliser operating at line rate. The addition of an equaliser provides a gain in receiver sensitivity of 1.2 dB to -20.8 dBm – this is not sufficient to increase the splitting ratio further (assuming a waveguide splitting structure based on 2x2 blocks, each incurring 3 dB loss), but enhances the tolerable manufacturing loss budgets through increased system margin.

B. Maximum Simultaneous Transmitters per Wavelength

To characterise the maximum number of nodes M that can share a wavelength during each Epoch, an unmodulated ECL was added to the IBLC modulated DSDBR TX, at the same nominal channel wavelength. The received power from the modulated transmitter was held at $P_{RX} = -20dBm$. The power P_{int} of the additional ECL was attenuated to match the power of $M-1$ additional transmitters each with modulators held in the zero state and attenuated by the MZM extinction ratio $(ER = 14.2dB)$ i.e. $P_{int} = (M-1)\times(-20dBm - ER)$. Simulations show that several unmodulated interfering signals exactly aligned in wavelength have a greater effect on BER performance than several unmodulated signals of slightly different frequencies (within the ±1 GHz range). Increasing the power of this single interfering transmitter therefore emulates the worst case scenario, of increasing the power from unmodulated transmitters at precisely the same wavelength.

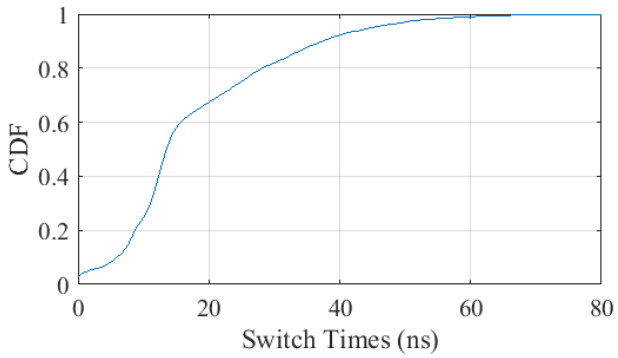

Fig. 10. Cumulative Distribution Function (CDF) of laser switch times between all possible pair combinations of the 89 available ITU 50 GHz grid channels.

Figure 9, again analysed from a full 2^{20} PRBS pattern capture, shows that without using an equalizer, the BER performance of the switch is increased above 10^{-12} for any number of additional transmitters per wavelength. When using a T/2 spaced equaliser, the BER is not degraded beyond 10^{-12} for up to 25 additional channels at the same wavelength. This is due to the equaliser providing optimal sampling, increasing high frequency content and reducing low frequency content arriving from laser frequency drift. Applying this equaliser provides much greater flexibility, enabling the allocation of up to 26 transmitters to the same wavelength in a given Epoch, allowing more transmitters to be granted TDM timeslots on each wavelength in any given Epoch. For greater than 25 unmodulated channels the BER decreases beyond the error-free limit, due to the spectral filtering no longer being sufficient to remove the interference signal power caused by the offset between the signal and LO lasers.

There is little benefit to allowing more than 26 transmitters per wavelength. For 26 equal timeslots in a 2 μs epoch each slot would carry 240 bytes, comparable to the median packet size in data centre workflows [26]. In practice shorter timeslots could be used but only to allow finer granularity sharing of the available bandwidth and variable packet size between 64 bytes and 6000 bytes, by allowing nodes to request multiple adjacent timeslots as required. This would permit optimal matching of the distribution of packet sizes resulting from real applications [26].

C. Fast Wavelength Switching Characterisation

To demonstrate the fast switching performance of the system, a full characterisation was made of the switching speed between all accessible wavelength channels of a DSDBR laser. Three separate characterisations were performed such that the total system performance is verified through their combination: the time taken to reach 90% of steady state intensity when switching from one wavelength channel to another; the time taken for laser frequency to stabilise within ±1 GHz of the target wavelength after switching; and the time elapsed before error-free data can be received following a switch event.

A single unmodulated DSDBR laser was switched repeatedly between each pair out of all possible pair combinations of 89 ITU grid channels, and passed through an attenuator into the signal port of a coherent receiver. A commercially available integrated tunable laser assembly was used as a static LO into the same coherent receiver. To measure the switching time, initially the LO was tuned to match the channel that the signal switched away from, before changing to match the target wavelength channel. The coherent receiver signal outputs were summed and squared as described in section II to return an intensity signal, which was processed for edge detection to find a sharp discontinuities in intensity. Edge detection thresholds were chosen to locate the times where the intensity at the original wavelength fell to

below 10% of the steady state intensity, and where the new target wavelength reached 90% of the steady state intensity.

Figure 10 shows a cumulative distribution of wavelength switching time measurements for all available laser channels. The median laser switching time is 12 ns, and 90% of laser channel switches complete within 40 ns. However some channel switches take noticeably longer than others due to large changes in the current applied to the rear grating. Additionally, oscillations can be observed in the laser's optical frequency response which can cause the desired lasing mode to be reached briefly, before the laser mode jumps to a wavelength far from the target before returning to the desired channel. In these measurements tuning times are only reported for the time taken to reach the final steady state with no further mode hops. Figure 10 shows that it is possible to tune between any of the channels in the C-band in less than 90 ns, verifying the switch system design.

The frequency variation and drift over time can also be measured using a time domain windowed Fast Fourier Transform of the coherent receiver outputs. Figure 11 shows the measured offset frequency as a function of time, demonstrating 3 switch events between ITU 50 GHz grid channels 46 and 44 (i.e. 100GHz total switch distance). The frequency offsets between the DSDBR and each LO are reduced to within ±1 GHz after an average of 113 ns, and after a further 50 ns settling period remain in a stable ±500 MHz

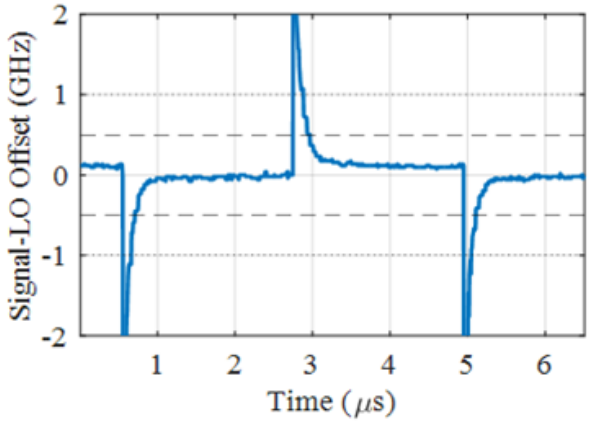

Fig. 11. Time resolved offset between switching Signal and static Local Oscillator frequencies over three switching events when switching by 100GHz.

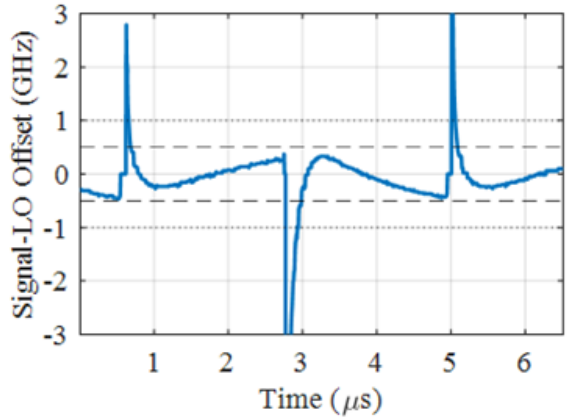

Fig. 12. Time resolved offset between switching Signal and static Local Oscillator frequencies over three switching events when switching by 4.25THz

range for the remainder of each Epoch. Figure 12 shows the time-resolved frequency offset when switching between ITU grid channels 2 and 87 (i.e. 4.25THz total switching distance). The frequency offset still settles within 147 ns, but is no longer stable and drifts over the remainder of the Epoch duration, due to thermal recovery from the larger change in switching current in the rear section. These instabilities may cause issues for higher complexity and coherent modulation formats, however they should not affect the performance of the amplitude only modulation as used in this work.

The DSDBR laser permits full use of the C-band and wavelength tuning is not restricted to any particular grid; it should therefore be possible to reduce the grid spacing in future to allow elastic grids and thus flexibility in modulation format to match demands. Given the total wavelength drift seen in figure 12 is around 1 GHz, it is imperative that sufficient guard bands are allowed to avoid cross-talk from adjacent channel drift; for example, in this IBLC modulation which is filtered at 22.5 GHz it is essential that no smaller grid spacing than 25 GHz should be used.

To examine how this switching directly impacts on the data integrity, a continuous stream of 25 Gb/s IBLC data was modulated onto an optical signal that switched repeatedly between a pair of wavelengths on the ITU Grid. This signal passed through an optical attenuator into the coherent receiver, ensuring -20.5 dBm signal power at the receiver, shown in section III. A. to ordinarily be received error free. As shown in figure 7, both LOs were enabled into the receiver, with

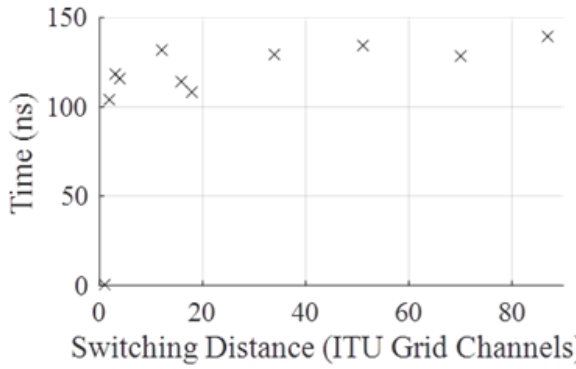

Fig. 13. Switching time measured by the time taken after a switching event to recover error-free data, for a range of switching distances across the C-band.

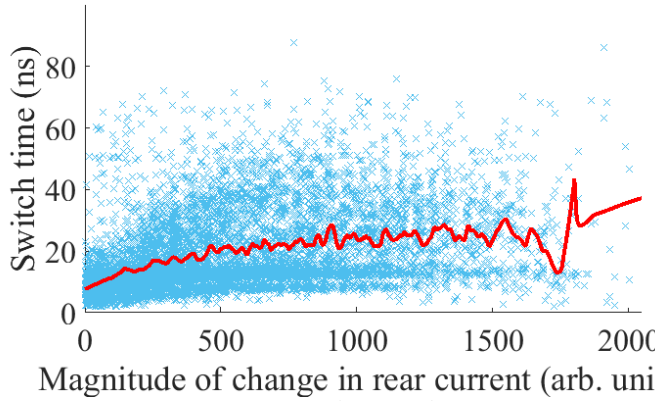

Fig. 14. The switching time recorded for each of the 89 x 89 wavelength switches compared to the magnitude of the rear current change for each switch (maximum ear current change is 2048 units).

wavelengths set to match the pair of channels being switched between by the signal. No emulation of other transmitters at the same wavelength was used. During a switch event, errors are observed in the data stream captured at the receiver, and the duration of this burst of errors before returning to error-free reception was recorded as the switching time.

Figure 13 shows the recorded durations of bursts of errors i.e. the times taken to return to error-free data reception after a switching event. The subset of wavelength switches in figure 13 is a selection of all possible wavelength switches, including full range changes of tuning current in the front, rear and phase sections, and switches between wavelength pairs spaced from 1 to 88 ITU 50 GHz grid channels apart. By selecting test cases with these properties, the performance in figure 13 is representative of all accessible channels of the DSDBR laser. In all cases error-free reception was observed within 200 ns of the switch event, verifying switch performance as per the design outlined above. For switching between adjacent channels, requiring only small changes to the phase and rear tuning currents of the laser, the tuning time was zero as no errors were observed in the received data. There is no direct correlation between switching time and switch distance in wavelength; this is due to the largest factor influencing tuning time being the large (up to 60 mA) changes in switching current in the rear grating, which do not directly correlate with wavelength switching distance when crossing multiple laser modes. Figure 14 shows this for all 89 x 89 recorded switches; in general the lowest switch times (below 8 ns) are only observed for changes in rear current below 600 units (out of a maximum 2048 unit swing). There is a general positive correlation between the magnitude of the rear current change (for either positive or negative swing in current) and the wavelength switch time.

IV. Conclusions

The physical layer data plane of a high port count all optical switch has been demonstrated, capable of scaling up to 1024 nodes per switch with each node operating at 25 Gb/s. Fast tunable transmitters and LOs with high sensitivity coherent receivers allow scalability to high port counts while maintaining low latency through guaranteed single-hop links; transmitter line coding and receiver equalisation enables packet switch-like functionality. Combining wavelength switching with TDMA (with experimental demonstration of the up to 26 transmitters per wavelength) provides highly flexible, low latency bandwidth allocation. The resulting network forms a single high port count switch for future data centre networks.

References

[1] Cisco, “Cisco Global Cloud Index: Forecast and Methodology, 2015-2020,” (Cisco, 2015), Available: http://www.cisco.com/go/cloudindex.

[2] A. Singh et al., “Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google’s Datacenter Network,” in *SIGCOMM ’15 Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, 2015, pp. 183-197.

[3] K. Chen et al., “Survey on routing in data centers: insights and future directions,” *IEEE Network*, vol. 25, no. 4, pp6-10, 2011.

[4] N. Binkert et al., “The role of optics in future high radix switch design,” in *38th Annual International Symposium on Computer Architecture (ISCA)*, 2011, pp. 437-447.

[5] N. Farrington et al, “Helios: a hybrid electrical/optical switch architecture for modular data centres,” in *SIGCOMM ’10 Proceedings of the 2010 ACM Conference on Special Interest Group on Data Communication*, 2010, pp. 339-350.

[6] N. Farrington et al., “A 10μs Hybrid Optical-Circuit/Electrical-Packet Network for Datacenters,” in *Optical Fiber Communication Conference 2013*, 2013, pp. OW3H-3.

[7] S. J. Ben Yoo, “Switching Technologies for the Future Photonic Internet,” *J. Lightwave Technol.*, vol. 24, no. 12, pp. 4468-4492, 2006.

[8] 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,” *Optics Letters*, vol. 39, no. 18, pp. 5244-5247, 2014.

[9] R. Proietti et al., “A Scalable, Low-Latency, High-Throughput, Optical Interconnect Architecture Based on Arrayed Waveguide Grating Routers,” *J. Lightwave Technol.*, vol. 33, no. 4, pp. 911-920, 2015.

[10] S. Di Lucente et al., “Numerical and experimental study of a high port-density WDM optical packet switch architecture for data centres,” *Optics Express*, vol. 21, no. 1, pp. 263-269, 2013.

[11] A. Funnell et al., “High Port Count Hybrid Wavelength Switched TDMA (WS-TDMA) Optical Switch for Data Centers,” in *Optical Fiber Communication Conference 2016*, 2016, pp. Th2A.54.

[12] A. J. Ward et al., "Widely tunable DS-DBR laser with monolithically integrated SOA: Design and performance," *J. Quant. Electron.*, vol. 11, pp. 149- 156, 2005.

[13] C. F. Liao and S. I. Liu, “A 40 Gb/s CMOS serial-link receiver with adaptive equalization and clock/data recovery,” *IEEE J. Solid-State Circuits*, vol. 43, no. 11, pp. 2492–2502, 2008.

[14] R. A. Griffin, “InP-Based High-Speed Transponder,” in *Optical Fiber Communication Conference 2014*, 2014, pp. W3B.7.

[15] M. Presi et al., “Low Cost Coherent Receivers for UD-WDM NRZ Systems in Access Networks,” in *ICTON 2014*, 2014, pp. Mo.C3.1.

[16] E. G. Rawson et al., “Bitaper star couplers with up to 100 fibre channels,” *Electronics Letters*, vol. 15, no. 14, pp. 432-433, 1979.

[17] B. Puttnam et al. “Burst mode operation of a DS-DBR widely tunable laser for wavelength agile system applications,” in *Optical Fiber Communications Conference 2006*, 2006, pp. OWI86.

[18] A. Croisier, “Introduction to pseudoternary transmission codes,” *IBM J. Res. Dev.*, vol. 14, pp. 354-367, 1970.

[19] A. Roy et al., “Inside the Social Network’s (Datacenter) Network,” in *SIGCOMM ’15 Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pp. 123–137, 2015.

[20] A. Rylyakov et al., “A 25Gb/s burst-mode receiver for rapidly reconfigurable optical networks,” in *Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf.*, vol. 58, pp. 400–401, 2015

[21] Arista Networks, Inc. 7500R Series Data Center Switch Router Data Sheet, June 2017. [Online]. Available: https://www.arista.com/assets/data/pdf/Datasheets/7500RDataSheet.pdf

[22] Cisco Systems, Inc. Cisco 25GBASE SFP28 Modules Data Sheet, April 2016. [Online]. Available: http://www.cisco.com/c/en/us/products/collateral/interfaces-modules/transceiver-modules/datasheet-c78-736950.pdf

[23] J. Cao et al., “Per-packet Load-balanced, Low-Latency Routing for Clos-based Data Center Networks,” in *CoNEXT '13 Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, pp. 49-60, 2013.

[24] J. Benjamin et al., “A High Speed Hardware Scheduler for 1000-port Optical Packet Switches to Enable Scalable Data Centers,” accepted for presentation at *Hot Interconnects 2017; IEEE25th Annual Symposium on High-Performance Interconnects*, 2017.

[25] N. S. Bergano et al., “Margin measurements in optical amplifier system,” *IEEE Photon. Tech. Lett.*, vol. 5, pp. 304-306, 1993.

[26] A. Roy et al., “Inside the Social Network’s (Datacenter) Network,” ,” in *SIGCOMM ’15 Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, 2015, pp. 123-137.