System Demonstration of Nanosecond Wavelength Switching with Burst-mode PAM4 Transceiver

Kai Shi, Sophie Lange, Istvan Haller, Daniel Cletheroe, Raphael Behrendt, Benn Thomsen, Fotini Karinou, Krzysztof Jozwik, Paolo Costa, Hitesh Ballani*

Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, U.K.
*t-kashi@microsoft.com

Keywords: Optical Switches, Fast Tuneable Lasers, Burst-mode PAM4 Receiver, Data Centres

Abstract

We demonstrate a wavelength switching PIC whose switching time (0.912 ns) is independent of the tuning range, and an optical switching system (50Gbps PAM4) using the PIC and fast burst-mode channel equalization, achieving 3.84 ns switching latency.

1. Introduction

Optical switches have the potential of significantly lowering data-centre (DC) network cost and power by reducing the number of expensive and power-hungry transceivers that are required by today's electrically-switched network to transmit packets over optical links. Also, optical switches can provide lower and predictable latency due to the lack of packet buffers. This has led to many different technologies for optically circuit-switched DC networks [1, 2], with switching times ranging from milliseconds to nanoseconds. Nanosecond-granularity switching is critical to support modern DC workloads such as key-value stores and memory disaggregation, which are dominated by small packets (>90 % of packets less than 576 bytes [3]). We conducted simulations to understand the impact of optical switching time using realistic workloads modelled after production DC workloads. Fig. 1 shows the overhead in terms of the flow completion time (FCT) as we vary the switching time compared to today's electrically-switched network. The analysis indicates that a switching time lower than 5 ns is needed to achieve a performance comparable to today's networks.

Among the physical technologies to achieve fast switching, tuneable lasers and arrayed waveguide grating routers (AWGRs) [4, 5, 6, 7, 8, 9] are particularly appealing for a DC deployment due to the lack of active elements in the core, resulting in a simpler and more robust design. Using lasers that can tune across the C-band, architectures that can scale to thousands of endpoints (servers or racks) in a DC have been proposed [9]. State-of-the-art tuneable lasers, however, exhibit tuning times that increase with the tuning range; the maximum tuning time has been reduced to 80 ns with off-the-shelf tuneable lasers [10] and to 5 ns with a custom design [11]. This exceeds the bounds derived from our simulation study. Furthermore, fast tuning is necessary but not enough for nanosecond switching at the system-level. It also requires a burst-mode transceiver that, unlike PON transceivers [12], can accommodate nanosecond switching at high bandwidth while still being amenable for a practical and low-power implementation.

In this paper, we present*: i)* a fast tuneable laser that decouples the switching time from the wavelength tuning range by separating the lasing from wavelength selection. The design can achieve ultra-fast tuning between any pair of wavelengths in the C-band. We implemented this design on an InP photonic integrated circuit (PIC) using a generic foundry (Jeppix), and demonstrate a switching time of 0.912 ns; using discrete components with better drive circuitry, we further reduced the switching time to <600 ps, an order of magnitude lower than the best result that we are aware of [11], *ii)* a system demonstration of an optically switched data centre that uses an optimized 50 Gbps (25 GBaud PAM4) burst-mode transceiver, coupled with our fast switching source, to achieve an end-to-end switching time of 3.84 ns, including the header and guard band overhead, with a receiver sensitivity of -8.1 dBm at the 5e-5 forward error correction (FEC) threshold. To achieve fast channel equalization, the transceiver "caches" the parameters previously learnt.

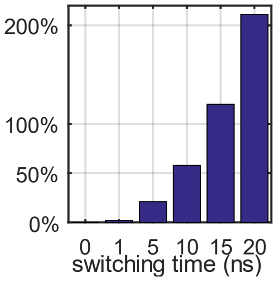

Fig. 1 FCT overhead vs. optical switching time

2. Wavelength Tuneable Source (WTS)

Electrically tuned semiconductor lasers typically comprise a gain section and tuneable grating sections, which are controlled through current injection. However, the electrical carrier dynamics in the tuning sections generate several intermediate modes during a switching event, which perturbate the cavity and increase the switching time [13]. Usually the larger the wavelength jump, the longer the time it takes to settle to the destination wavelength. AWG-based lasers have been proposed to eliminate the intermediate modes and achieve a switching time that is mostly wavelength-independent by using an array of semiconductor optical amplifiers (SOAs) to both provide gain and select the wavelength of interest [14]. However, the switching time is limited to a few ns due to the relaxation oscillation of the cavity when the SOAs are switched on/off.

In this paper, instead, we reduce the switching time and make it independent of the wavelength span by *disaggregating* the tuneable laser, i.e., by separating the light generation from the wavelength selection. This eliminates the perturbations due to the intermediate modes as well as the relaxation oscillation. The light generation component can be implemented using a multiwavelength source, e.g., through an integrated array of lasers or a comb device. For the wavelength selection, instead, we opted for the design depicted in Fig. 2 (a). It uses an array

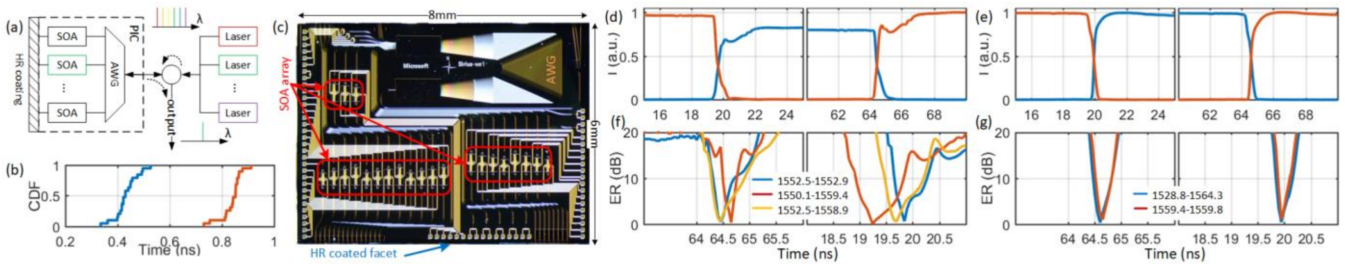

Fig. 2 (a) Schematic of the WTS using a wavelength selector on a PIC (dotted area), (b) CDF of the wavelength selection/deselection time of all the 19 SOAs on the PIC, (c) Microscope image of the PIC, (d-e) Intensity of the source and the destination channel during forward (left) and backward switching (right) using the PIC and discrete components, (f-g) Dynamic ER of forward (left) and backward switching (right) using the PIC and discrete components.

of SOAs combined with an AWG that is used as a wavelength multiplexer and demultiplexer simultaneously. This is achieved by applying a high-reflection (HR) coating at one facet to reverse the light path, and an optical circulator at the input of the wavelength selector to separate the input and the output light. This design simplifies the wavelength alignment between the source and selector, and reduces the size of the chip. When switching from wavelength *A* to wavelength *B*, the SOA connecting to the output port of the AWG for *A* is switched off while the one corresponding to B is turned on. As the two operations can occur in parallel, the switching time is determined by the slower one. We realized this approach on an 8×6 mm InP PIC (see Fig. 2 (c)). The chip contains 23 SOAs with varying lengths from 300 to 450 µm. Four of them are used for debugging and testing while the remaining 19 are connected to the output of the AWG, which is designed by Bright Photonics with a free spectral range of 33×50 GHz.

To evaluate our fast tuneable source, we inject our chip with a pair of tuneable lasers (ITLA), which we use as the light generation component. The PIC is placed on a probe station and driven by two high speed RF probes, terminated with ~50 Ω. We first measure the 90%-10% rising and falling edge of the signal at the output of designed PIC when each of the 19 SOAs is switched on/off. The intensity of the switched optical signal is then detected by a 50 GHz photodetector (PD) and the electrical waveform is captured by an oscilloscope. The cumulative distribution function (CDF) of the rise and fall time for each SOA switching signal is shown in Fig. 2 (b). The worst-case value measured was 527 ps and 912 ps, respectively. Since the switching time depends on the slower of these two operations, these results indicate that it is possible to achieve a sub-nanosecond switching time. Further, since the SOA switching time is largely independent of the specific wavelength within the device's gain spectrum, this can be achieved regardless of the source and destination wavelength. To measure the wavelength switching time, an optical bandpass filter (OBF) is used to discriminate the source and destination wavelength during a switching event [13]. The filtered signal is plotted in Fig. 2 (d) and the extinction ratio (ER) for the three channel pairs is shown in Fig. 2 (f) with switching times equal or lower than 930 ps at the 10-dB threshold. In Fig. 2(d) we can observe that the SOAs on the chip exhibit some degree of over- and under-shoot. This is mainly due to a ringing effect caused by the impedance mismatch between the RF probe and the SOAs on the chip and the parasitic effects of the electrical traces on the PIC. To verify this, we replicate the setup using discrete components. With this setup, as it can be observed by looking at the plots in Fig. 2 (e), the signals originating from the discrete SOAs are much cleaner, which resulted in a switching time of ~300 ps at 10 dB threshold (see Fig. 2(g)), an order of magnitude lower than the state-of-the-art results for tuneable lasers [11]. For both cases, there is no variation of the switching time at different wavelength spans.

A drawback of our approach is that the light generation component needs to produce multiple wavelengths simultaneously, e.g., through a laser array. This can increase its cost and power compared to existing tuneable lasers. One way to amortize this overhead is to share the tuneable laser across multiple transmitters at each source (server or rack switch). The degree of sharing is limited by the link budget. While this introduces the constraint that all transmitters at a source need to operate on the same wavelength at any instance, recent network architectures for fast optical switching without explicit scheduling [15, 16] already use a fixed schedule whereby transmitters at any given source can indeed use the same wavelength that switches periodically.

3. System Demonstration

Fig. 3 (a) shows the system testbed we use to demonstrate end-to-end performance. To account for the sharing of the wavelength tuneable source (WTS) across multiple transmitters, we introduce a 6-dB attenuator emulating a 4-way split. The optical spectra of all the 19 channels from our PIC are superimposed in the inset in Fig. 3 (a). A micro erbium-doped fibre amplifier (µEDFA), also shared across the transmitters, is used to boost the signal power to 18.5 dBm before going through the attenuator. The light is then modulated using a 20 GHz Lithium Niobite modulator biased at its quadrature point. The signal is generated by an arbitrary waveform generator (DAC) sampling at 60 GSample/s. Due to the hardware limitations, such as the number of the DAC channels, only one transceiver is set up. The modulated signal is connected to a single input of a 90-channel AWGR and can be routed to different output ports. We designed a fast switching PCB board to drive the WTS; it can support up to four SOAs, hence four output ports of the AWGR are combined with a 4×1 coupler so that we can receive the bursts presented at different output of the AWGR with various wavelengths using a single receiver. To compensate for the additional insertion loss, a second amplification is used after the coupler (this would not be required in a real system).

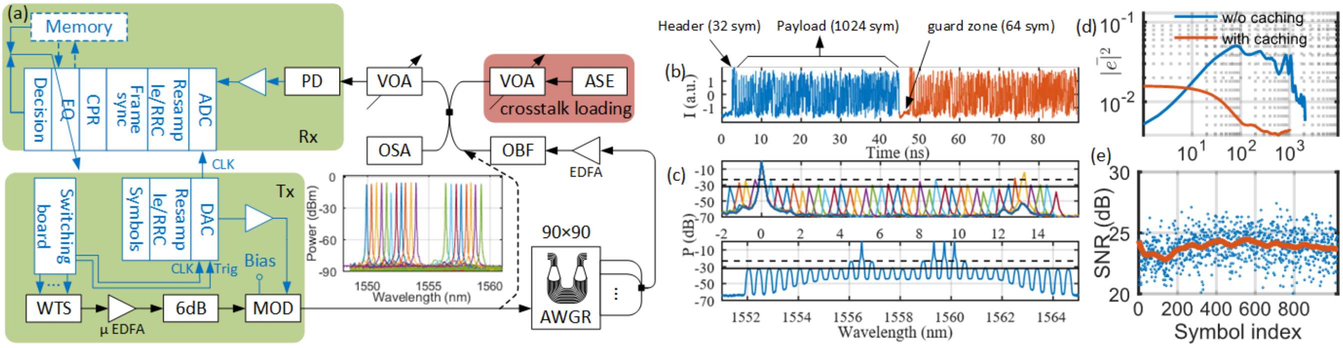

Fig. 3 (a) System setup, (b) Captured burst waveforms, (c) Measured (top) and emulated (bottom) AX/NX crosstalk, (d) Learning curve of the equalizer with and without caching. (e) SNR variation inside the burst

One limitation of the above setup is that since only one input port of the AWGR is loaded, it does not accurately indicate the crosstalk from the other ports. This is more significant when the side mode power from the WTS is high. We counter this by adding a crosstalk loading stage after the output of the AWGR to replicate the crosstalk introduced by the side modes originating from of the adjacent and non-adjacent crosstalk (AX/NX) of the AWG on the PIC. To model this, we measured the power of all the side modes at the output of the PIC by keeping the wavelength of one of the two input ITLAs constant and sweeping the other from -1.6nm to 38.8 nm (over a full FSR). The results in Fig. 3(c) (top) show that the PIC exhibits an AX of 23 dB and an NX of 32 dB. We used these values to generate the crosstalk profile and add it into the routed signal, obtaining the spectrum shown in Fig. 3 (c) (bottom). Finally, a 50 GHz PD after a variable optical attenuator (VOA) is used to convert the signal into the electrical domain for offline digital signal processing (DSP). The structure of each burst is illustrated in Fig. 3 (b), captured by a real-time oscilloscope (ADC), sampling at 160 GSample/s (GS/s). The symbol rate of the signal is 25 GBaud. The header of the burst contains 32 symbols of NRZ signal in the form of [+B +B -B +B], which can be used for both frame synchronization and as a training sequence. The payload of the burst contains 1024 symbols of PAM4 signal while the guard zone between each burst is 64 symbols. The 2.56 ns guard zone is conservatively chosen to guarantee the WTS finishes switching with relatively constant amplitude fluctuation as shown in Fig.3 (b). The total end-to-end switching time is 3.84 ns, including guard zone and header. The captured signal is immediately resampled to 1.25 samples per symbol (31.25 GS/s) after the ADC. The ADC and DAC are locked onto the same 100 MHz clock and the DSP is only responsible for the clock phase recovery (CPR). A key novelty of our DSP implementation is the technique used to achieve fast equalizer convergence. In a conventional least mean square (LMS) equalizer, it takes a few thousand symbols for the equalizer to converge as shown in Fig. 3 (d) (blue line). This makes it unsuitable for burst DC workloads. To overcome this issue, we propose to “cache” the equalizer coefficients. They can be obtained and stored during the system boot up, where the equalizer is running in data-aided (DA) mode. After that, the corresponding coefficients of each arrived burst can be pre-loaded with the cached value before switching to decision directed (DD) mode. As we can see in Fig. 3 (d) (red line), the equalizer converges within only a few tens of symbols with the coefficients caching. The signal-to-noise ratio (SNR) for each symbol within a burst is shown in Fig.3 (e). We can see that there is no degradation of the SNR at the beginning of each burst. Fig. 4 shows the bit error rate (BER) measured after the DSP for each burst. All the BER measurements are calculated over more than 1,000 bursts, rotating the wavelengths in a round-robin fashion. As a baseline, we also show the BER when no switching is performed. The results show that the received power penalty between switching and non-switching is negligible at the FEC threshold for DC transceivers (5e-5), which demonstrates the caching of equalizer is suitable for an ultra-short burst mode receiver. The penalty introduced by the crosstalk due to limited side mode suppression ratio (SMSR) is found to be less than 0.5 dB. There is a 1.5 to 3.5 dB penalty between WTS using discrete components and the PIC. This is due to the dynamic electrical and thermal effects of the integrated SOAs compared to the discrete ones.

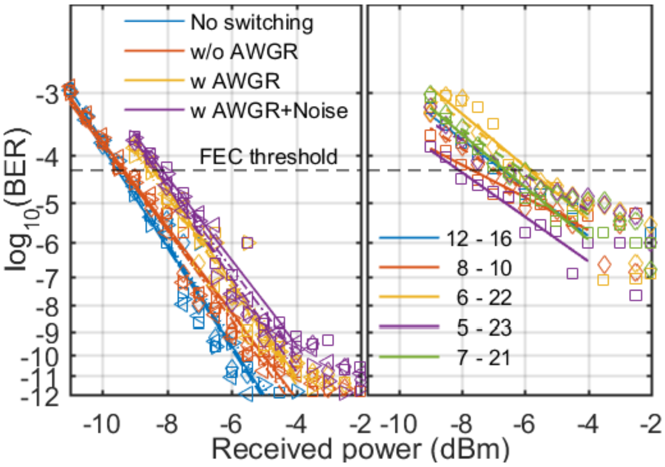

Fig. 4 System performance for the WTS using (a) discrete components (b) PIC (without AWGR and noise loading) for different channel transitions as indicated by the legend.

4. Conclusion

We showed that by disaggregating the lasing and wavelength selection, it is possible to achieve sub-nanosecond wavelength switching independent of the tuning range. We implemented our design using an InP PIC and combined it with a burst-mode 50Gb/s PAM4 receiver that uses a novel caching technique to achieve fast channel equalization, demonstrating a system-level switching time of 3.84 ns.

5. References

[1] H. Ballani, P. Costa, I. Haller, K. Jozwik, K. Shi, B. Thomsen and H. Williams, "Bridging the Last Mile for Optical Switching in Data Centers," in *Optical Fiber Communication Conference (OFC'18)*, 2018.

[2] Q. Cheng, M. Bahadori, M. Glick, S. Rumley and K. Bergman, "Recent advances in optical technologies for data centers: a review," *Optica*, vol. 5, no. 11, 2018.

[3] K. Clark et al., "Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network," in *ECOC*, 2018.

[4] T. Segawa, S. Matsuo, T. Kakitsuka, et. al., "All-optical wavelength-routing switch with monolithically integrated filter-free tunable wavelength converters and an AWG," *Optics Express*, vol. 18, no. 5, pp. 4340-4345, 2010.

[5] S. C. Nicholes, M. L. Masanovic, B. Jevremovic, et. al., "An 8×8 InP Monolithic Tunable Optical Router (MOTOR) Packet Forwarding Chip," *Journal of Lightwave Technology*, vol. 28, no. 4, pp. 641-650, 2010.

[6] N. Terzenidis, M. Moralis-Pegios, G. Mourgias-Alexandris, et. al., "High-Port and Low-Latency Optical Switches for Disaggregated Data Centers: The Hipoλaos Switch Architecture," *Journal of Optical Communications and Networking*, vol. 10, no. 7, pp. 102-116, 2018.

[7] Wang Miao, Fulong Yan, and Nicola Calabretta, "Towards Petabit/s All-Optical Flat Data Center Networks Based on WDM Optical Cross-Connect Switches with Flow Control," *Journal of Lightwave Technology*, vol. 34, no. 17, pp. 4066-4075, 2016.

[8] R. Proietti, G. Liu, X. Xiao, S. Werner, P. Fotouhi and S. Yoo, "FlexLION: A Reconfigurable All-to-All Optical Interconnect Fabric with Bandwidth Steering," in *CLEO*, 2019.

[9] Z. Cao, R. Proietti and S. J. B. Yoo, "Hi-LION: Hierarchical large-scale interconnection optical network with AWGRs," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 1, 2015.

[10] A. C. Funnell, K. Shi, P. Costa, et. al., "Hybrid Wavelength Switched-TDMA High Port Count All-Optical Data Centre Switch," *Journal of Lightwave Technology*, vol. 35, no. 20, pp. 4438-4444, 2017.

[11] J. E. Simsarian, et al., "Less than 5-ns wavelength switching with an SG-DBR laser," vol. 18, no. 4, pp. 565-567, Photonics Technology Letters.

[12] M. D. Santa, et al., "25Gb/s PAM4 Adaptive Receiver Equalisation Requirements for Burst-Mode Transmission Systems," in *ECOC*, 2016.

[13] Y. Yu and R. O'Dowd, "Fast intra-modal and inter-modal wavelength switching of a high-speed SG-DBR laser for dynamic wavelength routing," *Optical and Quantum Electronics*, vol. 33, no. 6, pp. 641-652, 2001.

[14] M. J. R. Heck et al., "Monolithic AWG-based Discretely Tunable Laser Diode With Nanosecond Switching Speed," *Photonics Technology Letters*, vol. 21, no. 3, pp. 905-907, 2009.

[15] V. Shrivastav, A. Valadarsky, H. Ballani, P. Costa, K. S. Lee, H. Wang, R. Agarwal and H. Weatherspoon, "Shoal: A Network Architecture for Disaggregated Racks," in *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI'19)*, 2019.

[16] W. M. Mellette, R. McGuinness, A. Roy, A. Forencich, G. Papen, A. C. Snoeren and G. Porter, "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network," in *Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*, 2017.