Fast and Uniform Optically-Switched Data Centre Networks Enabled by Amplitude Caching

Thomas Gerard[1,*], Kari Clark[1], Adam Funnell[2], Kai Shi[3], Benn Thomsen[3], Philip Watts[1], Krzysztof Jozwik[3], Istvan Haller[3], Hugh Williams[3], Paolo Costa[3] and Hitesh Ballani[3]

(1) Optical Networks Group, Dept. Electronic & Electrical Engineering, University College London, London, WC1E 7JE, UK.
(2) Multidisciplinary Engineering Education, University of Sheffield, 32 Leavygreave Rd, Sheffield, S3 7RD, UK.
(3) Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, UK.

**uceetmh@ucl.ac.uk*

Abstract: We propose amplitude caching to optically equalise burst mode traffic without delay stages. Through a fast, optically-switched system prototype, we demonstrate burst-mode penalties can be mitigated to within 0.4 dB at the KR4 HD-FEC level.

1. Introduction

Data centres are fast approaching a networking bottleneck, where their incumbent electrically-switched networks may struggle to keep pace with the exponentially growing demand for bandwidth. This looming crunch has renewed interest in fast optical circuit switching (OCS), which can be used to create flat, energy-efficient, low-latency networks [1, 2]. A wide variety of OCS architectures have been proposed based on wavelength routing, space switching, or a combination thereof [2], but all OCS systems face the common challenges of varying optical loss and phase across the network paths. These effects can introduce baseline wander and symbol-missampling at the receiver and are particularly hard to account for in sub-microsecond granularity switching systems. However, the data centre represents a unique, closed environment in which the loss and phase changes between source-destination pairs are relatively stable; hence, instead of discovering the phase and gain offsets through complex circuits at burst-granularity, they can be remembered or cached and applied at either end. Previously, we have demonstrated that fast clock phase recovery can be achieved by creating a synchronous data centre network that applies clock phase caching (CPC) to phase-align all transceivers with sub-bit precision [6]. In this work, we extend this principle to 'amplitude caching', in which optical equalisation is applied prior to transmission (or reception) to mitigate baseline wander in fast burst-mode transceivers. We combine amplitude and phase caching in a generic space-and wavelength switched testbed to create a power- and phase-uniform OCS system, for the first time. By testing our system with real time 25 Gbps on-off keying (OOK) burst data, we demonstrate low-penalty burst switching that permits the creation of highly scalable, flat, energy efficient, OCS data centre networks.

2. Concept of Amplitude Caching

Previously, optical equalisers have been proposed that use feed-forward semiconductor optical amplifier (SOA) gain control circuits to normalise the received optical power of burst-mode detectors [5,6]. However, these methods require additional SOAs, burst headers and delay stages that are incompatible with the fast amplitude recovery required by bursty data centre traffic. In this work, we exploit the deterministic path-loss of a flat data centre network to introduce 'amplitude caching', in which optical equalisation is applied at the transmitter. The principle of operation is shown in Fig. 1 which considers an N node network intra-connected through an $N \times N$ optical switch. Each node's optical transmitter is followed by a device capable of fast optical power control - our prototype uses SOAs, commonly included in optical transmitters, though fast optical attenuators or modulators could also be used. A two-stage process is used to achieve optical equalisation. The calibration stage is shown in Fig. 1(a), in which the equalisation of Node 0 is considered. Nodes 1 to N consecutively transmit dummy-data to Node 0, which measures the varying received optical powers. Node 0 then transmits feedback to Nodes 1 to N, informing them by how they should attenuate their output to match the received burst with the lowest power. Nodes 1 to N log this information in a (N-1) entry lookup table (LUT). This process is repeated for all nodes in the network until all LUTs are populated. The operation stage is shown in Fig. 1(b). When transmitting to Node 0, Nodes 1 to N use the cached amplitude values to vary the current applied to the SOA. This compensates for the varying path loss across the network path and the optical switch, and guarantees uniform-intensity burst-to-burst reception at Node 0. Optical path loss is not expected to vary significantly over time; however, the amplitude cache values can be periodically updated using in-band [1] or out-of-band [3] communication schemes. Amplitude caching, therefore, represents a low complexity method for achieving fast optical equalisation throughout a data centre network without requiring complex receiver-side gain control circuits or delay stages.

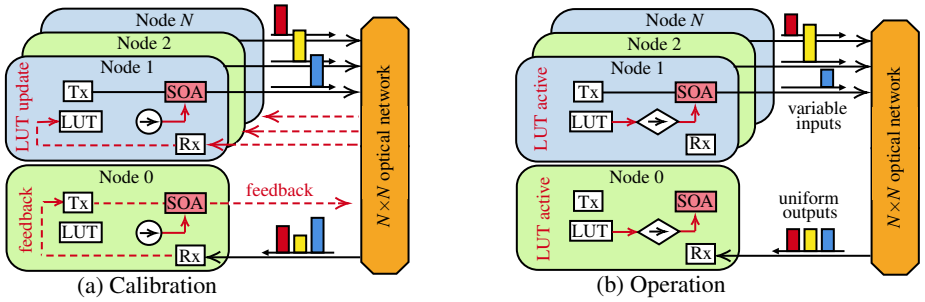

Fig. 1. Concept of amplitude caching applied to a flat $N \times N$ optical network. (a) The optical loss of all intra-connections are pre-calibrated and cached in a lookup table (LUT) in each node. (b) During normal operation, nodes vary their launch power using the LUTs to optically equalise the received signal.

3. Experimental Setup

Amplitude caching is a general technique that can be used in both space and/or wavelength switched networks; to demonstrate this, our setup uses a combination of both. Two space- and wavelength-switching nodes, Node A and Node B, were constructed using off-the-shelf components as shown in Fig. 2(a), and arranged into the the experimental setup shown in Fig. 2(b). A field programmable gate array (FPGA) was used to generate and receive real time data and act as a unified control-hub to the nodes' optical subsystems. Each node contained a semiconductor tuneable laser (TL) supporting in excess of 96 C-band 50 GHz wavelength channels, modulated by a 20 GHz intensity modulator. The modulator was biased at the quadrature point and driven with RF-amplified 25 Gb/s OOK data using the FPGA's inbuilt transmitter (TxA for Node A and TxB for Node B). An optical broadcast and select (OB&S) stage was emulated using an optical power coupler with outputs passed to two discrete semiconductor optical amplifiers (SOAs) supporting 69 nm of bandwidth with typical characteristics of 7 dB noise figure, 20 dB gain, and 10 dBm saturation power. These SOAs were used for both space-switching and amplitude caching. Space switching was achieved by driving one SOA in the 'on' state while the other was driven with a small negative current, creating an 'off' state with $>$ 30 dB extinction. Greater OB&S orders were emulated using a variable optical attenuator (VOA) before the optical coupler. The four OB&S outputs (Node A SOA0 and SOA1, and Node B SOA0 and SOA1) were passed to a 32×32 arrayed waveguide grating router (AWGR) for wavelength routing. TL wavelengths and AWGR input ports were selected to ensure the four input lines mapped to just two optical AWGR outputs, where were each connected to an avalanche photodiode (APD) with typical responsivity of 2 A/W in the C-band. The APD output was passed to the FPGA receiver (RxA for Node A and RxB for Node B), which performed real time sampling with an adaptive CDR circuit before making a hard decision on the received data. The measured bits were downloaded for offline analysis in MATLAB.

4. Results and Discussion

Space- and wavelength optical burst switching was demonstrated by setting each node's FPGA to instruct the TL and SOAs to play out a repeating train of four optical bursts. Each burst was made up of a $2^9 - 1$ pseudo-random bit sequence (PRBS) played out on repeat to emulate a 4 kB packet persisting for 1.3 μs at 25 Gb/s. Between each burst, the SOAs were set to the 'off' state, the TL wavelengths reconfigured, and the new SOAs set to 'on', requiring a last-bit to first-bit guard band of 42 ns. This can be reduced to 3.84 ns by using novel tuneable sources [4]. Fig. 2(c)(i) shows the output of the APD in Node B while receiving four sequential optical bursts from both nodes, measured using an 80 GS/s oscilloscope. The received optical power varies significantly from burst to burst; this is attributed to wavelength-dependent TL power, non-ideal coupling losses, and variable

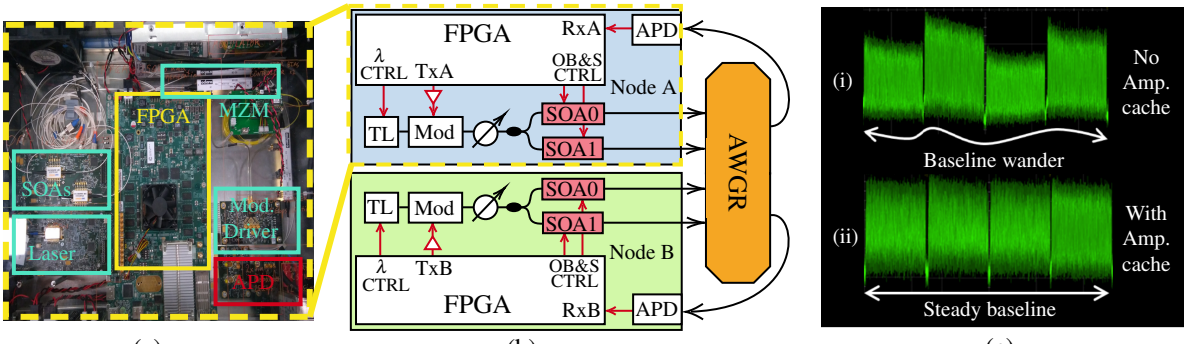

Fig. 2. (a) Space- and wavelength-switching node. (b) Two nodes perform fast OCS connections by switching the TL wavelength and active SOA. (c) APD output in Node B (i) without and (ii) with amplitude (Amp.) caching applied. Each burst represents a unique wavelength/SOA state from both TxA and TxB.

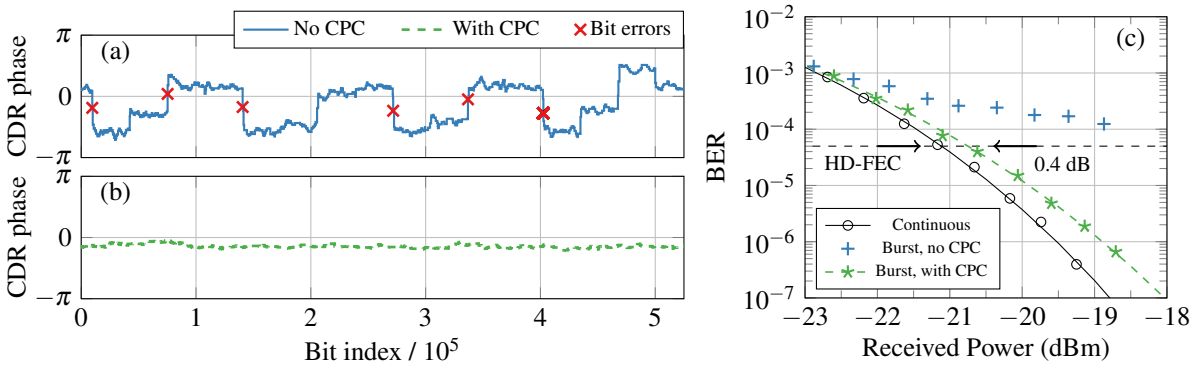

Fig. 3. Adaptive CDR phase in RxA when receiving burst mode data (a) without, and (b) with clock phase caching (CPC) enabled. Bit error locations are shown as crosses. (c) Associated BER curves.

AWGR insertion loss. The power variation is seen to introduce baseline wander to the AC-coupled APD output. To correct this, we apply amplitude caching as described in section 2 using the 'on' SOA in each node's OB&S stage for fast optical equalisation. The APD output with amplitude caching applied is given in Fig. 2(c)(ii), which shows that the burst-to-burst power variation have been suppressed to < 0.5 dB and that baseline wander has been removed. For the switch combinations tested in the work, up to a 2 dB reduction in transmitter launch power was required to correct for varying path loss, chiefly caused by variable AWGR insertion loss.

With amplitude caching applied, the burst mode data was received using the real time FPGA receivers. The inbuilt CDR circuits were used to adaptively adjust the real time sampler, correcting for phase-misalignments between TxA and TxB. Fig. 3(a) shows the adaptive CDR phase of RxA across 15 bursts, which is observed to make sudden, significant updates between bursts. This is because the FPGA transmitters are not phase matched and, therefore, the CDR unit in RxA must correct for phase misalignment after the burst arrival. Complete alignment can take 100's of ns to complete, and during this period non-ideal sample timing can introduce bit errors. The location of measured bit errors are overlaid on Fig. 3(a), and show a strong correlation with locations in which large CDR phase updates are required. To create a uniform phase response at each receiver we apply clock phase caching (CPC) [3], which exploits the FPGA's ability to phase shift the transmitted data prior to modulation. As with amplitude caching, the FPGA in each node was provided with an (N-1)-entry lookup table that contains the phase offsets required to correct for the misalignment associated with every receiver in the network. CPC was turned on for both TxA and TxB and the CDR phase remeasured at RxA. Fig. 3(b) shows that, with CPC on, phase variance has been reduced by > 20 dB and the associated bit errors have been removed.

The advantage of a uniform power-and phase OCS system is shown by the bit error rate (BER) curves in Fig. 3(c). Here we target a BER better than 5×10^{-5} for use with the KR4 Reed-Solomon hard decision forward error correction code (HD-FEC), overhead 2.7%, widely used in coded optical data centre systems. The data without phase caching exhibits a noise floor around a BER of 1×10^{-4}, preventing error free recovery during burst mode operation. With amplitude and phase caching applied this noise floor is removed, allowing our system to achieve low BERs better than 1×10^{-6} (limited by our FPGA's capture size). HD-FEC ready performance is achieved at an average received optical power of -20.7 dBm, a penalty of just 0.4 dB with respect to our system when transmitting continuous data (TxA transmitting on 1560.67 nm through SOA0 to RxA). This result shows that the application of amplitude and phase caching eliminates the bandwidth limitations of receiver-side recovery methods, enabling low-penalty OCS. CPC has been shown to fully mitigate burst errors to a BER $< 1 \times 10^{-12}$; therefore, the 0.4 dB of penalty observed here is attributed to residual SOA intensity variation on the rising edge. This can be corrected using intelligent SOA drive systems [7].

5. Conclusions

In this system demonstration, we apply amplitude and clock phase caching to create, for the first time, a power- and phase-uniform fast OCS system. The transmitter-side techniques are shown to push error floors to below a BER of 10^{-6}, mitigating real-time burst mode penalties to within 0.4 dB at the KR4 HD-FEC level. These general techniques can benefit any optically switched system, and can be used to create scalable flat data centre networks.

This work was supported by TRANSNET (EP/R035342/1) and Microsoft Research through its PhD scholarship programme. The authors thank Prof. P. Bayvel for useful discussion.

References

1. H. Ballani *et al.*, "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching", Proc. SIGCOMM, 2020.
2. Q. Cheng. *et al.*, "Photonic switching in high performance datacenters", OE **26**(12), 2018.
3. K. Clark *et al.*, "Synchronous subnanosecond clock and data recovery for optically switched data centres using clock phase caching", Nature Electronics **3**, 426–433, 2020.
4. K. Shi *et al.*, "System demonstration of nanosecond wavelength switching with burst-mode PAM4 transceiver", ECOC pdp, 2019.
5. E. Aw *et al.*, "Layered control to enable large scale SOA switch fabric", Proc. ECOC, Th1.2.5, 2006.
6. B. Thomsen *et al.*, "Optically equalised 10Gb/s NRZ digital burst-mode receiver for dynamic optical networks", OE **15**(15), 2007.
7. C. Parsonson *et al.*, "Optimal control of SOAs With artificial intelligence for sub-nanosecond optical switching", JLT **38**(20), 2020.