# Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network

Kari Clark(1), Hitesh Ballani(2), Polina Bayvel(1), Daniel Cletheroe(2), Thomas Gerard(1), Istvan Haller(2), Krzysztof Jozwik(2), Kai Shi(2), Benn Thomsen(2), Philip Watts(1), Hugh Williams(2), Georgios Zervas(1), Paolo Costa(2), Zhixin Liu(1)

(1) Optical Networks Group, Dept. of Electronic & Electrical Engineering, UCL (University College London), London, WC1E 7JE, U.K. kari.clark.14@ucl.ac.uk

(2) Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, U.K.

**Abstract** We demonstrate a clock and data recovery technique that achieves <625ps locking time for 25.6Gb/s-OOK and show its robustness under worst-case data centre temperature variation. The locking time was improved by 12×, making nanosecond optical switching viable in data centres.

## Introduction

Optical switches could address many of the challenges facing the electronically-switched networks used in cloud data centres (DCs) [1,2]. They can reduce cost by eliminating expensive and power-hungry transceivers responsible for opto-electronic conversions at each network switch and they can significantly reduce latency by eliminating network buffers.

However, modern DC workloads require rapid switch reconfiguration. We analysed network traces from a large-scale production cloud service and show the packet size distribution in Fig. 1a. Over 34% of the packets comprise less than 128 bytes, which means switching every 11 ns (with 100 Gb/s ports) while 97.8% of the packets are 576 bytes or less. Similarly, over 91% of the packets generated by Facebook's in-memory cache are 576 bytes or less [3].

A few optical switching technologies can be reconfigured in nanoseconds and, hence, could efficiently support small packets [2]. However, a key challenge to making nanosecond-optical switching practical is ultra-fast, burst-mode clock and data recovery (CDR) [1]. Unlike electronic switches, optical switches create momentary physical links between transmitter-receiver pairs every time they are reconfigured, incurring CDR locking time when a new momentary link is created. The minimum time between packets, which limits network throughput, is the sum of the optical switching time and the CDR locking time.

Commercially-available CDR technologies can take up to hundreds of nanoseconds for CDR locking with state-of-the-art research prototypes achieving 8 ns [4]. Thus, even if the optical switching takes a nanosecond, the minimum gap between packets will still be much higher due to CDR locking time. To quantify the impact of this, in Fig. 1b we estimate the application throughput for the packet size distribution shown in Fig. 1a as we vary the CDR locking time, *t*. Assuming 100 Gb/s ports (4×25 Gb/s) and 1 ns optical switching time, for *t*=8 ns the throughput is only 66.6%, i.e., more than one third of the available bandwidth is wasted. In contrast, more than 92% throughput can be achieved if we can reduce the CDR locking time to sub-nanosecond.

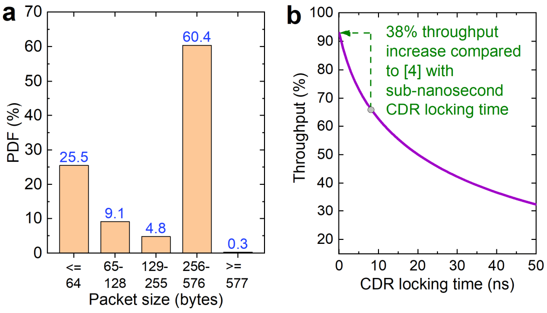

**Fig. 1:** (a) Packet size distribution and (b) Throughput against CDR locking time

To overcome the throughput bottleneck due to CDR locking time, we propose *clock phase caching,* a practical and inexpensive CDR technique (using standard CDR hardware) that leverages unique characteristics of cloud data centres to achieve sub-nanosecond locking time. We note that optical switching requires network nodes to be frequency-synchronised so that their transmissions can be scheduled to avoid conflicts. Such synchronisation can be achieved in a controlled environment like a DC using existing mechanisms for robust control plane distribution of a reference clock [5,6]. CDR locking is then simplified as a receiver only needs to determine the phase of incoming data. However, even with frequency synchronisation, finding the correct phase offset between communicating nodes can still take more than 40 ns (see Fig. 4). To minimise the locking time, we exploit the observation that the received phase is relatively constant across multiple transmissions between the same (transmitter, receiver) pair, only slowly drifting with temperature. Thus, we can "cache" or store the correct clock phase to be used for transmissions between the pair. At the start of a new packet, we shift (either at transmitter or receiver side) the clock phase to match the phase value measured during the last transmission.

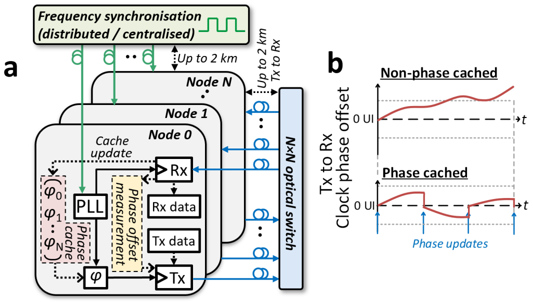

**Fig. 2:** (a) Synchronised optically switched data centre architecture, (b) Operational principle of phase caching.

In this paper, we demonstrate the clock phase caching technique in a 25.6 Gb/s real-time data centre network (DCN) testbed, comprising a nanosecond optical switch with two transmitters and one receiver, connected through 2×2 km of fibre located in a thermally controlled chamber to study the impact of worst-case temperature changes. We show error-free DCN interconnection with <625 ps locking time in a 6-hour continuous measurement.

## DCN architecture and operational concept

Fig. 2a shows the architecture of an optically switched DCN and the operating principle of our CDR technique. The nodes could be servers or rack switches. The architecture leverages existing techniques [5,6] to achieve robust, distributed frequency synchronisation at DC scale. In our testbed, we use a central clock source. With transmit-side clock phase caching, each transmitter in an N-node DC stores a set of (N-1) phase values corresponding to each receiver. Every node must exchange a packet with all other nodes at DC start-up to populate the caches and then periodically to keep all caches updated. Just prior to the transmission of a new packet, the node shifts its transmitter clock phase to the cached value to correct for the clock phase offset along the Tx-Rx path. This ensures that the phase of every packet arriving at each receiver is aligned with a sufficiently small offset as shown in Fig. 2b. This eliminates the need for the receiver to recover the phase at the start of packets, achieving sub-nanosecond locking time.

A key challenge for our approach is that the relative clock phase offset between a (Tx, Rx) pair drifts due to temperature changes. For example, standard single mode fibre (SMF-28) has a thermal coefficient of ~37 ps/km/°C [7]. A change of 1°C temperature will result in about 37 ps fibre time of flight change for 1 km of SMF-28, corresponding to almost one unit interval (UI) for a 25 Gb/s data centre link. We monitored the inlet and exhaust temperature for a rack in a production DC with evaporative cooling over 228 days, and the largest temperate variation observed was 9°C; even assuming this happens across adjacent 5 minute measurement intervals, it translates to a rate of change of 0.03 °C/s. This is also consistent with the recommendations for industrial DC design [8]. As shown in Fig. 5b, a slow phase cache update rate of 20 Hz is sufficient to cope with even triple this worst-case temperature change for a real DC environment.

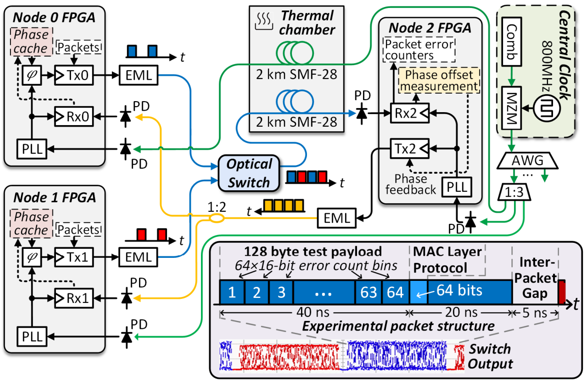

**Fig. 3:** Proof-of-concept experiment. Inset: Oscilloscope output from switch and experimental packet structure.

## Experimental setup

Fig. 3 shows our experimental testbed. We use three field programmable gate arrays (FPGAs), Xilinx VCU108, to emulate three DC nodes. The first two nodes, Node 0 and Node 1, transmit 128-byte on-off-keying (OOK) modulated packet payloads at 25.6 Gb/s embedded in 60-ns packets, via electronic modulated lasers (EML), to Node 2 through a 2×2 LiNbO$_3$ optical switch. After the switch, alternate packets from Node 0 and Node 1 propagate through 2.0 km SMF-28 before reaching Node 2. Every time a phase update is required, Node 2 measures the phase offset using the FPGA digital phase-interpolator CDR and subsequently sends the phase offset values back to Node 0 and Node 1 via the link shown in orange, which emulates duplex interconnection. We shift transmitter clock phase using the transmitter phase interpolator in the FPGA. All optical signals were detected by photo detectors with transimpedance amplifiers.

For frequency synchronisation, an 800-MHz clock signal was modulated onto a 25-tone frequency comb (25 GHz spacing) by a 1×2 Mach-Zehnder modulator. After modulation, each tone has about 0 dBm power, which is 11 dB higher than the minimum required optical power for clock distribution in our setup. In principle our clock distribution system can therefore support frequency synchronisation for 512 DC nodes with AWG and passive splitters. Up to 10,000 nodes can be supported using a wide-band frequency comb [9]. The optical clock was transmitted through a 2.1 km SMF-28 to Node 0 and was intentionally attenuated to -11 dBm to emulate the optical clock splitting in a cloud DC.

Our packet structure, shown as inset in Fig. 3, contains 3 De Bruijn sequences of $2^9$ length. A MAC layer protocol (64-bits) is embedded in the

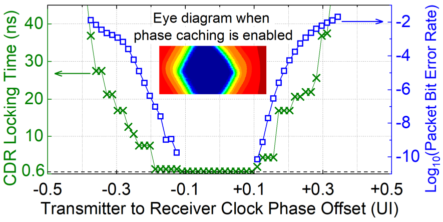

**Fig. 4:** Impact of transmitter to receiver clock phase offset on CDR locking time with clock frequency synchronisation.

3rd sequence for frame alignment, clock phase offset feedback and packet identification. The 2nd and 3rd sequences are used to measure clock phase offset (only when a clock phase update is required). To assess the performance of phase caching, the 1st and 2nd sequences are divided into 64×16-bit bins, and the number of bit errors falling in each bin is recorded in real-time over 60 s intervals. CDR locking time was calculated as the first bin in the packet with a BER of $<10^{-10}$, if all following bins also have BER of $<10^{-10}$.

To study varying environmental conditions observed in production data centres, we placed the 2 km of clock and data fibre in a thermally controlled chamber. When switched on, the temperature begins at 25 °C and increases approximately linearly between 30 and 50 °C at a rate of 0.11 °C/s, over triple the worst-case rate of temperature change in our measurement study.

## Results and discussion

We first investigate the dependency of locking time on the initial phase offset between the transmitters and the receiver. As shown in Fig. 4, <625 ps locking times were achieved when the phase offset is less than ±0.1 UI when starting clock recovery, because every 16-bit bin of the payload has a BER $<10^{-10}$. A larger initial phase requires a longer locking time and results in bit errors at the start of packets that increases the packet BER. The required locking time is over 40 ns (equivalent to the length of a 128-byte packet) when the initial phase offset is >~±0.3UI. This indicates that the use of only frequency synchronisation is insufficient for ultra-fast CDR.

Fig. 5a shows that our technique is stable over a 6-hour measurement period with ambient temperature variation. Fig. 5a (1) shows the measured ambient temperature and Fig. 5a (2) total phase offset, which shows the total applied correction for clock phase misalignment between the two transmitters and the receiver. Over the 6-hour period, the clock phase drifted by more than four UI. Fig. 5a (3) shows the measured locking time and compares the case with and without phase caching. Phase caching corrects the phase offset of both transmitters at a rate of 20 Hz to compensate for the changes in fibre time of flight. *No bit errors were observed with phase caching enabled for six hours* with a locking time of less than 625 ps and $3.4\times10^{14}$ bits received (blue line in Fig. 5a (3)). When phase caching was disabled, we observed a quick degradation in BER due to temperature drift, resulting in an increase in locking time to over 40 ns within 4 minutes (green cross markers).

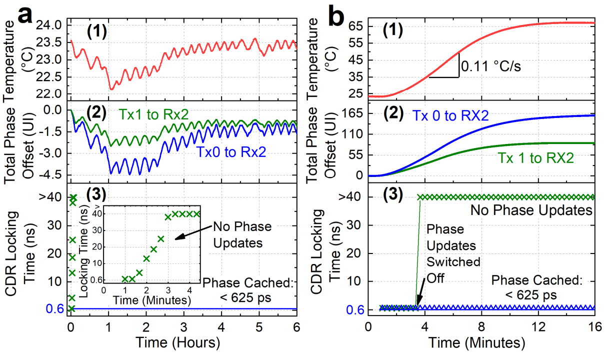

**Fig. 5:** (a) Long-term stability and (b) impact of temperature. No errors were recorded with phase caching enabled.

Fig. 5b shows the performance of our clock phase caching technique under emulated worst-case DC temperature variation. Fig. 5b (1) and Fig. 5b (2) show the measured temperature in the temperature chamber and the resulting total phase shifts. The 0.11 °C/s increase of temperature resulted in a 16 ps/s change in fibre time of flight across 2×2 km of SMF-28. Fig. 5b (3) shows the measured locking time with phase caching enabled at a rate of 20 Hz. Even with the rapid change of temperature, no errors were observed, and the locking time was <625 ps. When clock phase caching was switched off, we observed a loss of clock phase alignment in less than one second due to the temperature induced change in fibre time of flight.

## Conclusion

We demonstrated clock phase caching, a CDR technique that achieves <625 ps locking time for 25.6 Gb/s OOK in a real-time optical switching testbed. Phase caching exploits the peculiarities of DCs such as limited temperature variation and geographical size, and lowers locking time by an order of magnitude, thus making nanosecond optical switching viable in DCs. The stability and robustness of our technique was demonstrated across six hours of error-free optical switching. Phase caching can be easily implemented on top of off-the-shelf transceivers, enabling burst-mode, optically-switched operation using a standard 100Gb/s transceiver (4×25 Gb/s lanes).

We acknowledge financial support from Microsoft, Inphi Inc. and EPSRC grants EP/R035342/1 and EP/R041792/1. Eblana photonics provided lasers used in this work.

## References

[1] H. Ballani et al., OFC’18, W1C.3, 2018.
[2] F. Testa and L. Pavesi, eds., Springer, 2017.
[3] Q. Zhiang et al., IMC’17, 2017.
[4] A. Cevrero et al., OFC’18, 2018.
[5] M. Lipiński et al., ISPCS’11, 2011.
[6] ITU-T G.8262/Y.1362, 2015.
[7] R. Slavík et al., Sci. Rep., 5(15447), 2015.
[8] ASHRAE TC9.9, 2016.
[9] B.P.P.Kuo et al., J. Lightw. Technol., 31(21), 2013.