Synchronous Sub-Nanosecond Clock and Data Recovery for Optically-Switched Data Centres using Clock Phase Caching

Kari A. Clark[1,*], Daniel Cletheroe[2], Thomas Gerard[1], Istvan Haller[2], Krzysztof Jozwik[2], Kai Shi[2], Benn Thomsen[2], Hugh Williams[2], Georgios Zervas[1], Hitesh Ballani[2], Polina Bayvel[1], Paolo Costa[2,†], and Zhixin Liu[1,‡]

[1]Department of Electronic and Electrical Engineering, University College London, London, United Kingdom. WC1E 7JE
[2]Microsoft Research
[*]kari.clark.14@ucl.ac.uk
[†]paolo.costa@microsoft.com
[‡]zhixin.liu@ucl.ac.uk

Abstract

The rapid growth of data transferred within data centres, combined with the slowdown in Moore's Law, creates challenges for the future scalability of electronically-switched data-centre networks. Optical switches could offer a future-proof alternative and photonic integration platforms have been recently demonstrated with nanosecond-scale optical switching times. End-to-end switching time, however, is currently limited by the clock and data recovery time, which typically takes microseconds, removing the benefits of nanosecond optical switching. Here we show a *clock phase caching* technique that can provide clock and data recovery times of under 625 ps (16 symbols at 25.6 Gb/s). Our approach uses the measurement and storage of clock phase values in a synchronised network to simplify clock and data recovery versus conventional asynchronous approaches. We demonstrate our technique using a real-time prototype with commercial transceivers and validate its resilience against temperature variation and clock jitter, based on measurements from a production cloud data centre.

Main

The rate of data transmitted between servers within data centres has rapidly increased over the last few years [1], driven by cloud adoption and data-intensive cloud workloads such as data analytics and machine learning. Cloud providers have been able to accommodate this fast growth by relying on Moore's Law for networking: every two years the electronic switch integrated circuits (ICs) double their bandwidth at same cost and power. The long-term sustainability of this trend, however, is being questioned by two upcoming challenges: Firstly, similar to processor ICs, scaling transistor density on electronic switch ICs is fundamentally limited by power dissipation as few-nm transistor sizes are approached [2]. Secondly, electronic high-speed serial transceiver data rates are predicted to be hard to scale beyond 112 Gb/s due to the steep increase in dielectric loss when operating at high frequencies [3, 4]. Consequently, increasing the aggregate switch capacity will require a proportional increase in the number of serial transceivers surrounding the chip, resulting in greater power density and packaging complexity.

Although continued bandwidth scaling in the near future could be supported by architectural optimisations such as co-packaged optics [5], preserving cost neutrality in the medium-to-long term appears very challenging. This uncertainty has motivated research in optical switches as a viable alternative to electronic switches [6]. Optical switches simply redirect the incoming signals onto output ports without any optical/electronic conversion or digital processing and, hence, they do not suffer from the limitations of transistor or transceiver technology. They could, therefore, provide a future-proof solution for bandwidth scaling within the data centre [7].

Optical switches would need to support nanosecond-granularity switching, because it is critical to supporting modern data-centre workloads such as key-value stores, which are dominated by small packets [8]. Fast optical switches with nanosecond optical switching time have been demonstrated using different techniques, including Mach-Zehnder modulators (MZMs) [9], semiconductor optical amplifiers (SOA) [10], and tuneable lasers [11]. However, the overall end-to-end switching latency also includes the *clock and data recovery (CDR) locking time* in addition to the optical switching time. *CDR locking time* is incurred because, unlike the continuous point-to-point interconnections between electronic packet switches, a new physical layer link is established every time an optical switch is reconfigured to transmit a new data packet. The receiver therefore needs to recover a new clock every time this occurs. In commercially-available devices (e.g. phase-interpolator-based CDR modules), the locking time is typically in the order of microseconds [12] as they require comparing the phase of a reference clock to the phase of the embedded data clock using an early-late vote counter. Therefore, the overall end-to-end switching latency would be limited by the CDR locking time rather than by the nanosecond switching time.

Long end-to-end switching time greatly reduces the network utilisation, defined as the percentage of time that the network is used to send data (as opposed to time spent re-configuring the network or spent for CDR locking). This occurs because the data transmitted in data-centre networks is dominated by small data packets, and therefore removes the benefits of optical switching. To quantify the impact of CDR locking time when switching data-centre workloads, we analysed network traces from a cloud production service. We investigated the impact on network utilisation as we varied the CDR locking time. As detailed below, we conclude that sub-nanosecond CDR locking time (under 16 symbols at 25.6 Gb/s) is needed for a data centre to achieve over 90% network utilisation.

CDR locking time in 1-2 symbols has been shown using gated voltage-controlled oscillator (GVCO) based CDRs [13, 14]. However, there are two significant fundamental implementation issues associated with this approach: *i)* the need for a dedicated gated oscillator per channel leads to high power consumption, large silicon area and possible crosstalk between oscillators, and *ii)* an inability to operate at multiple data-rates without large performance degradation, which is important for data-centre operation [15]. All-optical clock recovery such as optical-injection locked lasers have also been investigated [16] but these have limited stability for practical application in data centres. These implementation issues are not present for clock phase interpolator-based CDRs, but commercial implementations typically have CDR locking times in the order of microseconds [12], with state-of-the-art research prototypes still limited to approximately 325 symbols or longer [17], due to clock phase position metastability when there is a 0.5 symbol phase offset between the initial CDR phase and incoming data [15, 17]. This metastability problem arises from CDRs needing to be able to lock to any arbitrary incoming clock phase, which in turn arises from not establishing frequency and phase synchronisation between transmitters and receivers connected to an optical switch.

In this article, we report a technique that can achieve sub-nanosecond locking time through the measurement and storage of clock phase values in a synchronised network, to simplify clock and data recovery versus conventional asynchronous approaches, using commercial off-the-shelf transceivers. Optical switches lack packet buffers, and so they require the network transmissions to be accurately synchronised at a nanosecond timescale to avoid conflicts [18], which, in turn, requires that the network endpoints must have their clocks frequency synchronised. Network-wide frequency-synchronisation can be achieved in a controlled environment such as a data centre using well-known mechanisms for robust control-plane distribution of a reference clock [19, 20]. This partly simplifies the CDR task as only the clock phase (rather than both frequency and phase) needs to be recovered. Nevertheless, even with a known clock frequency, clock phase recovery can still take up to hundreds of nanoseconds [8].

However, the clock phase shift across multiple transmissions between the same (transmitter, receiver) pair is relatively constant in a data-centre environment, only slowly changing with temperature due to fibre time-of-flight change, which occurs due to fibre length expansion and fibre refractive index change [21]. Therefore, we propose 'caching' or storing the correct clock phase shift to be used when transmitting between a given transmitter to receiver pair, so that the received clock phase is always constant irrespective of which transmitter is communicating with the receiver.

To account for the shifting of the correct clock phase values as the data-centre temperature changes, we designed a low-overhead clock phase update mechanism that all end-points periodically execute. The clock phase and synchronisation achieved by clock phase caching is analogous to the clock synchronisation of transistors across integrated circuits but on distance scales of up to 2 km. Clock phase updates are required

at data-centre scale because temperature change across 2 km of optical fibre leads to clock delay change on the order of 80 picoseconds per C [21]. Across data-centre temperature ranges of up to 40 C [22] and data centre distance scales of 2 km, this, in the worst case, leads to clock phase changes on the order of nanoseconds, requiring clock phase updates to maintain clock phase synchronisation. In contrast, in integrated circuits, temperature change causes negligible clock delay change [23], and so periodic clock phase updates are not required.

Clock phase caching technique. A data-centre architecture using clock phase caching is illustrated in Fig. 1a. We consider a typical data-centre cluster, comprising hundreds of racks (48 servers per rack), and propose to interconnect all the electronic top-of-rack (ToR) switches through an all-optical fabric. Some of these switches can be used as IP gateways to provide inter-cluster connectivity and enforce routing policies, e.g. firewalls. Each rack receives a synchronised clock (Fig. 1a), which is used to drive the transceiver logic. This clock could be distributed from multiple synchronised sources, so it is tolerant to clock device failure as represented by the dotted lines in Fig. 1a [19].

A clock phase cache is located within each transceiver, containing a set of values (one value per receiver) corresponding to the phase shifts that need to be applied to the synchronised clock before each packet is sent from its transmitter. At start-up every transmitter exchanges a packet with all receivers connected through the optical switch in the data centre. Every receiver then measures the phase offset of the received packet and feeds back this information to the transmitter to populate its phase cache. This process is then repeated periodically to account for clock phase drifts due to temperature variation. A single clock phase update between a transmitter and receiver is shown in Fig. 1b.

This technique allows sub-nanosecond CDR locking time to be consistently achieved across a wide range of temperature and jitter variation while only requiring a slow update frequency (10 Hz). The additional logic required to measure the phase offset, store the phase values, and apply the phase shift (orange components in Fig. 1a) could be embedded in the switch die or in the transceiver itself. In our prototype implementation the clock phase caching is performed at the transmitter side. However, clock phase caching can also be performed at the receiver side (see Supplementary Information Section 4). This would avoid the need to send the updates back to the transceivers, but it would require the receiver to know the source of traffic ahead of time.

In our approach, we distribute an optical clock via a control plane so that the clock frequency is synchronised. Although optical fibre-based clock synchronisation, including clock phase and frequency transfer between nodes, has been used for metrology [24] and optical time domain division multiplexing [25, 26], it has not been investigated for burst mode data communications. Furthermore, our method for measuring and controlling clock phase is implemented on a digital CDR module and clock phase interpolator respectively. The capability of phase tracking is therefore not limited by optical or electronic delay lines, avoiding the cost, large size and limited clock phase tracking range associated with these devices.

Impact of CDR locking time on network utilisation. To assess the impact of CDR locking time for cloud data-centre networks, we analysed network traces from a cloud production service [8]. As shown in Fig. 2b, over 34% of the packets comprise less than 128 bytes, which means switching every 11 ns (with 100 Gb/s ports) while 98% of the packets are equal or smaller than 576 bytes [8]. This is consistent with a similar study from Facebook, which found that 91% of the packets generated by their in-memory distributed cache service are 576 bytes or less [27].

As we illustrate in Fig. 2a, with small packets, even if re-configuring the optical switch takes only 1 ns, long locking time leads to low network utilisation. The network utilisation decreases as the CDR locking time increases because an increasing proportion of time is used for CDR locking rather than data packet reception. For example, assuming a 25.6 Gb/s transceiver, sending a 64-byte packet only takes 20 ns. Using a commercially-available CDR with 100 ns locking time, the overall network utilisation would be less than 17%. In Fig. 2c, we plot the results of a study to evaluate the impact of CDR locking time using the packet distribution in Fig. 2b, assuming 4 25 Gb/s transceivers. As shown in the graph, reducing the CDR locking time from the state-of-the-art CDR locking time for phase interpolator CDRs of 325 symbols (12.7 ns at 25.6 Gb/s) [17] to sub-nanosecond would result in a 1.54 improvement in network utilisation.

Experimental demonstration of clock phase caching. To demonstrate the feasibility of our phase-caching technique and evaluate its performance, we built a prototype system using three field programmable gate arrays (FPGAs) and investigated the performance of our CDR technique in a 2 1 optical switching network. As shown in Fig. 1c, the first two nodes, Node 0 and Node 1, transmit 128-byte on-off-keying (OOK) modulated packet payloads embedded in 60-ns packets at 25.6 Gb/s, via externally modulated lasers (EMLs), to Node 2 through a 2 2 $LiNbO_3$ optical switch with a 200 ps switching time. After the switch, alternate packets from Node 0 and Node 1 propagate through 2 km of SMF-28 and are attenuated to -10.5 dBm before reaching Node 2 to emulate power budgets of intra-data-centre transmission standards [28].

The clock phase cached transceiver architecture shown in Fig. 1a is implemented in the three FPGAs. As shown in Fig. 1b, Node 2 periodically measures the clock phase offset using the FPGA CDR module and subsequently sends the clock phase offset values back to Node 0 and Node 1 via the link shown in yellow, which emulates duplex interconnection. We then shift the transmitter clock phase using the clock phase interpolator in the transmitter. An 800 MHz reference clock was modulated onto a 24-channel frequency comb and distributed to all three nodes for frequency synchronisation, emulating distributed frequency synchronisation techniques such as White Rabbit [29] and Sync-E [19].

To study the impact of varying environmental conditions observed in production data centres, we placed 2 km of SMF-28 for the data path and 2 km of SMF-28 for the clock path 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. We configured the system such that the signals propagate in opposite directions in the data and clock fibres. Variation in temperature therefore causes the phase shift in the two fibres to be additive, resulting in an overall phase shift at the receiver equal to 4-km SMF-28 (about 160 ps/ C), allowing us to investigate the worse-case rate of clock phase shift in synchronous intra-data-centre interconnection. This contrasts with the case where clock and data signals propagate in the same direction, where balanced changes in both fibres due to temperature variation would cause the phase shifts to cancel at the receiver.

To demonstrate the reliability of our technique, we first carried out a measurement over a 48 hour period in a laboratory environment where the temperature fluctuates within 5 C. During this long-term stability test, we emulated cooling failure by turning off the laboratory air conditioner (Fig. 3a(i)), which led to a 2 C temperature rise at 17 hours followed by a slow room temperature increase during the 17th and 23rd hours. After that we turned on the air conditioner and cooled the room back to 18:5 C. Our transceiver continuously monitored the phase shift of each transmitter, as shown in Fig. 3a(ii). The CDR locking time was determined by finding the first error free 16-bit bin by averaging 9:2 10^8 packets over 60 s, as detailed in the Methods section. When clock phase caching was enabled, the system was error-free over the whole 48 hours, resulting in a CDR locking time of under 16 symbols (under 625 ps), as shown in Fig. 3a(iii). When phase caching was disabled, we observed a quick degradation in BER due to temperature drift, resulting in an increase in CDR locking time to over 40 ns within 4 minutes.

Fig. 3b shows the performance of our clock phase caching technique under emulated worst-case rate-of-change of DC temperature. Fig. 3b(i) and Fig. 3b(ii) 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. 3b(iii) shows the measured CDR locking time with phase caching enabled at a rate of 10 Hz. Even with the rapid change of temperature, no errors were observed, and the CDR locking time was under 16 symbols (under 625 ps). When clock phase caching was initially ran, and then switched off during the 0.11 C/s rate of temperature increase, we observed a loss of clock phase alignment in less than one second due to the temperature induced change in fibre time-of-flight.

The rate of clock phase drift is proportional to the rate of temperature change. This impacts the stability of our CDR technique which requires a sufficient rate of re-calibration of clock phases for all transceiver pairs to maintain a CDR locking time of under 16 symbols (under 625 ps). To investigate the required phase update rate for different rate of temperature change, we employed an electronically-controlled tuneable optical delay line to emulate the temperature induced change of time-of-flight across 2 2 km of SMF-28. The optical delay line is driven with triangular waveforms of different slopes to emulate different rates of temperature change from 0.01 to 0:45 C/s.

As we show in Fig. 4, the critical rate of temperature change at which CDR locking time begins to increase to greater than under 16 symbols (under 625 ps) is proportional to the rate of clock phase updates. The proportionality constant, k, represents the clock phase update rate required to cope with the worst-case rate-of-change of temperature within the data-centre and is 27:0 C^{-1} in our system.

To evaluate the worst-case rate-of-change of temperature within data centres, 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 recommendations for industrial DC design [22]. Based on Fig. 4b, a clock phase update rate of 1 Hz is sufficient to cope with this worst-case rate-of-change of temperature.

To demonstrate the jitter tolerance of our CDR technique, we separately applied different amplitudes of 1 MHz sinusoidal jitter and white noise jitter to the 800 MHz reference clock source. For both types of jitter, increasing the clock jitter amplitude eventually degrades the CDR locking time. The 2 2 km fibres were kept in an insulated chamber to minimise the impact of ambient temperature variation. In Fig. 5, we measure the CDR locking time against RMS jitter amplitude for 1 MHz sinusoidal jitter (Fig. 5a) and white noise jitter (Fig. 5b). The CDR locking time was under 16 symbols (under 625 ps) when the RMS jitter is equal to or less than 5.6 ps. Standard reference oscillators have jitter of a few hundred femtoseconds, ensuring sufficient performance for our CDR technique for intra-data-centre interconnection. We also show the simulated effect of sinusoidal and white noise jitter on an NRZ-OOK signal.

Estimated scalability of clock phase caching. Our clock phase caching technique must be scalable to support hyper-scale data-centre interconnection. Based on the experimental results, we estimate the worst-case network utilisation overhead resulting from clock phase caching at different data-centre distance scales. An estimate of the worst-case optical switch network utilisation resulting overhead from clock phase caching is given by $D(\frac{dT}{dt})kN(t_{meas} + t_{update})$=2 km, where D is the maximum distance between the servers and/or switches, $\frac{dT}{dt}$ is the worst-case rate-of-change of data-centre temperature, k is the proportionality constant from Fig. 4b of 27.0 C ^{1}, N is the number of nodes and t_{meas} is the time taken to measure the clock phase offset between each transmitter to receiver pair as detailed in the Methods section and t_{update} is the length of time each transmitter spends transmitting a packet to carry each phase update value.

We estimate the worst-case optical switch network utilisation overhead in Fig. 6 of clock phase caching, based on our measured worst-case rate-of-change of data-centre temperature of 0.03 C/s, a time taken for each clock phase update per transmitter to receiver pair of 2.08 µs as detailed in the Methods section and a time taken to transmit each update packet of 0.065 µs.

The estimated worst-case network utilisation overhead resulting from clock phase caching is only 1.7% for 10,000 data-centre end-points (servers or switches) separated by a maximum of 2 km. If 10,000 top-of-rack switches, each networking 48 servers (480,000 servers in total), were connected by an optically-switched fabric, our clock phase calibration technique would allow all possible transmitter to receiver pairs to communicate with each other with under 16 symbols (under 625 ps) CDR locking time with only a 1.7% optical switch worst-case network utilisation overhead. For 10,000 end-points, a traditional asynchronous CDR would need a clock recovery time of 25 symbols to have an equivalent network utilisation to clock phase caching.

Further scalability improvements can be made by using optical fibre that experiences a smaller change in time-of-flight due to temperature change. For example, the change of the time-of-flight in hollow core fibre is 20 times less than that in SMF-28 [21], which can potentially lead to an overhead reduction by a factor of 20, or scaling up of the number of supported nodes at the same overhead by a factor of 20. Specially coated fibre with 3-4 times lower thermal sensitivity is also commercially available for overhead reduction or scaling up the number of clock synchronised nodes [30].

Conclusions

We proposed and demonstrated a clock phase caching technique that enables under 625 ps CDR locking time for optically-switched data-centre networks. Through a real-time prototype network, we show that the clock phase caching technique can tolerate a rate-of-change of temperature of 0.11 C/s, which is 3 times greater than the worst-case conditions in production cloud data centres, as well as a root-mean-square jitter tolerance of 0.135 symbols. Our technique allows for an increase of data-centre network utilisation to more than 90%. We estimate that it can be scaled to support 10,000 data-centre end-points (servers or switches) for realistic worst case data-centre environmental temperature variation and fibre lengths with a worst-case small 1.7% overhead on network utilisation.

Our clock phase caching technique could be of value in the various scientific and engineering applications that require high-throughput burst-mode transmission, and also in communities that require synchronisation. For example, time-division multiplexing-based passive optical networks (PON) could potentially benefit from the synchronisation technology that we have demonstrated [31]. The optical network units in PONs can be synchronised in similar manner to minimise the required inter-packet gap and data-detection latency, crucial for latency-sensitive applications such as the Internet of Skills, virtual reality and gaming [32]. Another example lies in the clock synchronisation for quantum key distribution systems (QKD), which require the time gate synchronised to the photon arrival to identify the quantum signals correctly [33]. Our clock phase caching technique can track the drift of the time-of-flight in transmission links without the need for paired fibre, significantly reducing the complexity of QKD clock networks.

Methods

Optical clock distribution. We achieved frequency synchronisation of the three FPGA nodes by modulating an 800 MHz reference clock onto a 24-tone optical frequency comb with 25 GHz spacing with an Mach-Zehnder modulator (MZM). The frequency comb used was an optoelectronic comb source consisting of a 27-dBm CW light source emitting at 1555 nm, a phase modulator and an MZM [34]. Both the phase modulator and the MZM were driven by a 25 GHz RF source, generating 25-GHz spacing comb 2 dB power flatness. The generated comb was modulated and subsequently amplified to 18 dBm by an EDFA, yielding an average power of about 3 dBm per tone. These optical clock signals were filtered by an array waveguide grating (AWG) and detected by 18-GHz photodiodes with trans-impedance amplifiers (TIA). In our experiment, the lowest receiver power required for each modulated tone was -11 dBm. We thus emulate the use of a 1:8 optical splitter by attenuating the optical power from -0.5 dBm (power per tone after AWG) to -11 dBm, indicating that our current system can optically synchronise 192 nodes from a single clock source. In practice, relatively low-speed photodiodes (e.g. 5 GHz bandwidth, which were unavailable at the time of the experiment) may provide higher sensitivities because of their slightly higher responsivity (e.g. about 1A/W). This number of nodes can be easily increased to 3072 if we use an high power EDFA to amplify the modulated clock to 30 dBm. The clock signal between the AWG and Node 1 additionally travels through 2 km of single mode optical fibre (SMF-28). Noting that the clock distribution does not necessitate the use of a frequency comb, conventional laser arrays with 50 GHz or 100 GHz channel spacing could also be used. Nevertheless, a frequency comb provides a compact source with well-defined wavelength spacing which may ease the thermal and wavelength management in a clock synchronised network with potentially low operating expenses. In addition, wideband and high OSNR frequency combs such as the parametric comb [35] and thin-film $LiNbO_3$ [36] can potentially allow the system to scale to more than 10,000 nodes.

Transmitter data modulation. The externally modulated lasers shown in Fig. 1c consist of 1550 nm carriers modulated with data using 35-GHz bandwidth electro-absorption modulators.

Optical switch control. The focus of this work is on demonstrating that clock phase caching can reduce CDR locking time to sub-nanosecond in optical switches. To minimise experimental complexity, we used a 2 2 optical switch, both transmitters are configured to output packets continuously with no central or edge scheduling, and frame alignment of packets from the two transmitters is performed manually. The 2 2 optical switch, which is an 18 GHz $LiNbO_3$ MZM, was driven by a 130-ns period (two 60 ns packets plus 5 ns inter-packet gap) square wave clock signal from a GTY transceiver (18 GHz bandwidth) from FPGA Node 1.

Experimental packet structure. Our experimental packet structure used to measure CDR locking time, shown in Supplementary Fig. 1S, contains 3 De Bruijn sequences of 2^9 length. A MAC layer protocol (64-bits) is embedded in the 3rd sequence for frame alignment, clock phase offset feedback and packet identification. In a full system demonstration of clock phase caching, clock phase measurement packets (such as those used in our experimental demonstration) would be scheduled and transmitted periodically between standard Ethernet packets. In our implementation, for the purposes of simplifying implementation complexity since the aim of our experimentation is to demonstrate the proof-of-principle of clock phase caching, we allowed all packets transmitted to be potentially used for both clock phase measurement and data transmission,

avoiding the need to implement a packet scheduler. The 2nd and 3rd sequences in the packets are used to measure clock phase offset when a clock phase update is required.

CDR locking time measurement. To assess the performance of phase caching, the 1st and 2nd sequences act as the 128 byte payload and are divided into 64 16-bit bins. The number of bit errors falling in each bin across all packets arriving at the receiver is recorded in real-time over 1 s intervals. As shown in Supplementary Fig. 2S, CDR locking time was calculated as the first bin in the packet with a BER of under 10^{-10}, if all following bins also have BER of under 10^{-10}. BER was calculated by summing errors collected across all 64 bins.

Transmitter clock phase shift. The clock phase interpolator consists of a standard clock phase rotator that takes a half rate clock from the output of the FPGA transceiver LC-Tank quad PLL, and outputs one clock phase selected from 128 evenly spaced phase steps split across 2 symbols (resolution of 64 steps per symbol). The selected clock phase is then used to drive the serial clock of the parallel in serial out (PISO) converter in the FPGA transmitter. The phase interpolator selected phase output is controlled by a PIPPM control circuit built into the Xilinx GTY transceiver. This control circuit allows the transmitter clock phase to be shifted by up to $15/64^{\text{th}}$ of a symbol every transmitter parallel clock cycle, which is 400 Mhz in our proof-of-concept experiment. Within the 5 ns interpacket gap, immediately prior to packet transmission to a receiver, the phase interpolator selected phase is changed to a clock phase value equivalent to the cached clock phase value for communication with the receiver. Shifting by half a symbol requires two FPGA parallel clock cycles which takes 5 ns total, hence leading to a required 5 ns gap, followed by a further $2/64^{\text{th}}$ of a symbol shift at the beginning of the next packet. The $2/64^{\text{th}}$ of a symbol shift, if required, could alternatively be performed at the end of the previous packet. Shifting by greater than half a symbol is achieved by shifting clock phase in the reverse direction by up to half a symbol. The gap could be reduced to sub-nanosecond by directly using the clock phase cached value to change the phase interpolator selected phase, bypassing the PIPPM control circuit. This would remove the phase shift delay associated with the PIPPM control circuit, thereby enabling the phase of the transmitter clock to switch to any arbitrary phase in sub-nanosecond time, which would fall within the 1 ns required interpacket gap required to allow for optical switching time.

Measurement of transmitter to receiver clock phase o set. For each clock phase update, transmitter to receiver clock phase offset was measured by averaging the FPGA receiver raw CDR phase across the 2nd and 3rd packet sequences, and then averaging the clock phase offset of 32 packets separated by 1 μs, taking a total of 2.08 μs per update (including inter-packet gaps). The receiver CDR phase interpolator spans 128 evenly-spaced phase steps split across 2 symbols (resolution of 64 phase steps per symbol). The raw CDR phase is output at a rate of 400 MHz, and this phase is equal to the clock phase shift applied to the receiver reference clock to keep it aligned with the bit edges in incoming data, which is achieved within the CDR by using a phase interpolator to measure the ratio between the number of times the clock edge occurs before and after the data edge in 64 bit intervals, and shifting the receiver reference clock such that this ratio equals 50%. A single packet is sent back to the transmitter node along the return link to update the clock phase offset at the transmitter once the clock phase offset has been calculated at the receiver. No oversampling is used. The built-in Xilinx GTY transceiver receiver-side continuous time linear equaliser (CTLE) filter is used.

Experimental emulation of rate-of-change of temperature. To evaluate the clock phase update rate required to maintain a CDR locking time under 16 symbols (under 625 ps) for a given rate of change of temperature across 2 2 km single mode optical fibre, we introduced a 166 ps, voltage-controlled free space optical delay line between the AWG and the clock input of Node 1. We drove the fibre delay line with a saw-tooth waveform to experimentally emulate the rate of change of phase that occurs for a given rate of change of temperature in SMF-28. The saw-tooth waveform frequency was swept through a series of values equal to that theoretically experienced by a loose tube single mode fibre of 4 km length with a temperature coefficient of delay of 40.6 ps/(km C) at a series of different rate of change of temperatures as shown in Fig. 4b. 10^{11} bits per 16-bit bin were collected for each rate-of-change of temperature. To avoid the need to manually re-perform time synchronisation of the two transmitters, we biased the switch into the crossbar state such that packets only from Node 1 arrive at the receiver. To prevent the receiver CDR remembering the clock phase of incoming data between successive packets, we reset the receiver CDR phase within the

5 ns inter-packet gap.

Clock jitter generation. Jitter was applied to the electronic clock source shown in Fig. 1c by frequency modulating the 800 MHz clock source with a 1 MHz sinusoidal and white noise voltage waveforms of different amplitudes. The peak-to-peak sinusoidal jitter applied to the clock was calculated after the clock output and before the MZ modulator using an radio frequency spectrum analyser by measuring the ratio of the power of the 800 MHz clock tone to the power of its 1 MHz side-tones. The RMS jitter amplitude of the clock after frequency modulation with white Gaussian noise jitter was measured using a digital communications analyser. 10^{11} bits per 16-bit bin were collected for each applied jitter amplitude. To prevent the receiver CDR remembering the clock phase of incoming data between successive packets, we reset the receiver CDR phase within the 5 ns inter-packet gap.

Estimation of network utilisation in Fig. 2c. To estimate the impact of the CDR locking time on the overall network utilisation, we used an event-based network simulator that we developed in-house, which we cross-validated against real data-centre networks. The simulator models a network consisting of nodes interconnected by an optical switch. The line-rate of each node to/from the optical switch was 4 25 Gb/s. We generated a synthetic workload by randomly selecting the payload size for each packet from the distribution shown in Fig. 2b and by selecting sources and destinations with a uniform random distribution. As soon as a node finishes sending a packet, a new packet is generated and a new destination is selected. If a source-destination path is not already set up, before the packet payload can be received, the optical switch needs to be reconfigured (*optical switching time*) and the receiver CDR has to lock onto the new incoming signal (*CDR locking time*). We set the optical switching time to 1 ns, based on recent advances in optical switching devices [9, 10, 11], and we varied the CDR locking time from 0 to 400 ns (corresponding to a range from 0 to 1,000 symbols). To generate Fig. 2c, we recorded the network utilisation (measured as the ratio of the time in which packet payloads were sent divided by the total simulation time) as we increased the CDR locking time.

Data availability statement

The data that support the figures within this paper are available from the UCL Research Data Repository (DOI: 10.5522/04/12043224), which is hosted by FigShare.

References

[1] Singh, A. *et al.* Jupiter rising: A decade of clos topologies and centralized control in Google's datacenter network. *ACM SIGCOMM Computer Communication Review* **45**, 183–197 (2015).

[2] Markov, I. Limits on fundamental limits to computation. *Nature* **512**, 147 (2014).

[3] Dorren, H. *et al.* Challenges for optically enabled high-radix switches for data center networks. *J. Lightwave Technol.* **33**, 1117–1125 (2015).

[4] Ghiasi, A. Large data centers interconnect bottlenecks. *Opt. Express* **23**, 2085–2090 (2015).

[5] Krishnamoorthy, A. *et al.* From Chip to Cloud: Optical Interconnects in Engineered Systems. *J. Lightwave Technol.* **35** (2017).

[6] Testa, F. & Pavesi, L. *Optical Switching in Next Generation Data Centers* (Springer International Publishing, Basel, Switzerland, 2017).

[7] Ballani, H. *et al.* Bridging the Last Mile for Optical Switching in Data Centers. In *2018 Optical Fiber Communication Conference (OFC)* (OSA, 2018).

[8] Clark, K. *et al.* Sub-nanosecond clock and data recovery in an optically-switched data centre network. In *2018 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, 2018).

[9] Chen, C. P. *et al.* Programmable Dynamically-Controlled Silicon Photonic Switch Fabric. *J. Lightwave Technol.* **34** (2016).

[10] Cheng, Q., Wonfor, A., Wei, J. L., Penty, R. V. & White, I. H. Low-energy, high-performance lossless 8x8 SOA switch. In *2015 Optical Fiber Communication Conference (OFC)* (OSA, 2015).

[11] Shi, K. *et al.* System Demonstration of Nanosecond Wavelength Switching with Burst-mode PAM4 Transceiver. In *2019 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, 2019).

[12] Xilinx. Virtex UltraScale Architecture Data Sheet: DC and AC Switching Characteristics. Tech. Rep., Xilinx (2015).

[13] Banu, M. & Dunlop, A. E. Clock recovery circuits with instantaneous locking. *Electron. Lett.* **28**, 2127–2130 (1992).

[14] Terada, J. *et al.* A 10.3 Gb/s burst-mode CDR Using a $\Delta\Sigma$ DAC. *IEEE J. Solid-State Circuits* **43**, 2921–2928 (2008).

[15] Rylyakov, A. *et al.* A 25 Gb/s Burst-Mode Receiver for Low Latency Photonic Switch Networks. *IEEE J. Solid-State Circuits* **50**, 3120–3132 (2015).

[16] Yang, Y., Wen, Y. J., Nirmalathas, A., Liu, H. F. & Novak, D. Optical clock recovery at line rates via injection locking of a long cavity Fabry-Pérot laser diode. *IEEE Photonics Technol. Lett.* **16**, 1561–1563 (2004).

[17] Ozkaya, I. *et al.* A 56Gb/s burst-mode NRZ optical receiver with 6.8 ns power-on and CDR-Lock time for adaptive optical links in 14nm FinFET CMOS. In *2018 IEEE International Solid-State Circuits Conference (ISSCC)*, 266–268 (IEEE, 2018).

[18] Bostica, B., Burzio, M., Gambini, P. & Zucchelli, L. Synchronisation issues in optical packet switched networks. In Prati, G. (ed.) *Photonic Networks*, 362–376 (Springer London, London, 1997).

[19] G.8261. Timing and synchronization aspects in packet networks (2008).

[20] Lee, K. S., Wang, H., Shrivastav, V. & Weatherspoon, H. Globally synchronized time via datacenter networks. In *Proceedings of the 2016 ACM SIGCOMM Conference*, 454–467 (ACM, 2016).

[21] Slavík, R. *et al.* Ultralow thermal sensitivity of phase and propagation delay in hollow core optical fibres. *Sci. Rep.* **5**, 15447 (2015).

[22] ASHRAE TC9.9. Data Center Networking Equipment–Issues and Best Practices (2016).

[23] Soleimani, S., Afzali-Kusha, A. & Forouzandeh, B. Temperature dependence of propagation delay characteristic in FinFET circuits. *Proceedings of the International Conference on Microelectronics, ICM* 276–279 (2008).

[24] Ma, L.-S., Jungner, P., Ye, J. & Hall, J. L. Delivering the same optical frequency at two places: accurate cancellation of phase noise introduced by an optical fiber or other time-varying path. *Opt. Lett.* **19**, 1777–1779 (1994).

[25] Lord, A., Blank, L. C., Boggis, J. M., Bryant, E. & Stallard, W. A. Theory of Control Mechanism for an Optically Time-Division-Multiplexed System. *Electron. Lett.* **24**, 2011–2012 (1988).

[26] Ellis, A. D., Widdowson, T., Phillips, I. D. & Pender, W. A. High speed OTDM networks employing electro-optic modulators. *Trans. Inst. Electron., Inf. Commun. Eng., Sect. E* **E81-C**, 1301–1308 (1998).

[27] Zhang, Q., Liu, V., Zeng, H. & Krishnamurthy, A. High-resolution Measurement of Data Center Microbursts. In *Internet Measurement Conference (IMC)* (2017).

[28] 802.3bm. IEEE Standard for Ethernet - Amendment 3: Physical Layer Specifications and Management Parameters for 40 Gb/s and 100 Gb/s Operation over Fiber Optic Cables (2015).

[29] Lipínski, M., Włostowski, T., Serrano, J. & Alvarez, P. White rabbit: a PTP application for robust sub-nanosecond synchronization. In *IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication* (2011).

[30] Bousonville, M. *et al.* New phase stable optical fiber. In *Beam Instrumentation Workshop 2012 (BIW2012)*, 101–103 (2012).

[31] G.989.2. 40-Gigabit-capable passive optical networks 2 (NG-PON2) (2015).

[32] Aijaz, A., Dohler, M., Aghvami, A. H., Friderikos, V. & Frodigh, M. Realizing the tactile Internet: Haptic communications over next generation 5G cellular networks. *IEEE Wireless Communications* **24**, 82–89 (2016).

[33] Takesue, H. *et al.* Quantum key distribution over a 40-dB channel loss using superconducting single-photon detectors. *Nat. Photonics* **1**, 343 (2007).

[34] Torres-Company, V. & Weiner, A. M. Optical frequency comb technology for ultra-broadband radio-frequency photonics. *Laser Photonics Rev.* **8**, 368–393 (2014).

[35] Kuo, B. P.-P., Myslivets, E., Alic, N. & Radic, S. Wavelength multicasting via frequency comb generation in a bandwidth-enhanced fiber optical parametric mixer. *J. Lightwave Technol.* **29**, 3515–3522 (2011).

[36] Zhang, M. *et al.* Broadband electro-optic frequency comb generation in a lithium niobate microring resonator. *Nature* **568**, 373 (2019).

Acknowledgements

We acknowledge financial support from Microsoft, Inphi Inc., and EPSRC grants EP/R041792/1 and EP/R035342/1. Royal Society Paul Instrument Fund PIF/R1/180001. Eblana Photonics provided the lasers used in this work. We thank P. Watts for helpful discussion at the early stages of the work and E. Vonhof for her assistance in the generation of the figures.

Author Contributions

K.C., P.C., H.B., I.H., K.J., B.T., H.W. and T.G. conceived of the concept of clock phase caching, which was later refined with help from K.S., D.C. and Z.L.. K.C. and Z.L. conceived of and constructed the experimental setup. K.C. implemented clock phase caching in the experiment. K.C. designed and implemented all FPGA hardware code. K.C. led the experiment and collected all experimental results, supervised by Z.L., with support provided by P.C., H.B., B.T., P.B. and G.Z.. P.C. and H.B. collected data-centre traffic and performed the optical switch network utilisation analysis. K.C., Z.L., P.C. and H.B. wrote and revised the manuscript. All authors discussed the results and commented on the manuscript.

Corresponding Authors

Correspondence to Kari A. Clark, Paolo Costa or Zhixin Liu.

Competing Interests

A patent application, entitled "Phase Caching for Fast Data Recovery", was filed by Microsoft Technology Licensing, LLC with the U.S. Patent and Trademark Office on 27th October 2017, on the technology described in this paper. This patent is currently pending, with patent number US20190132112A1.

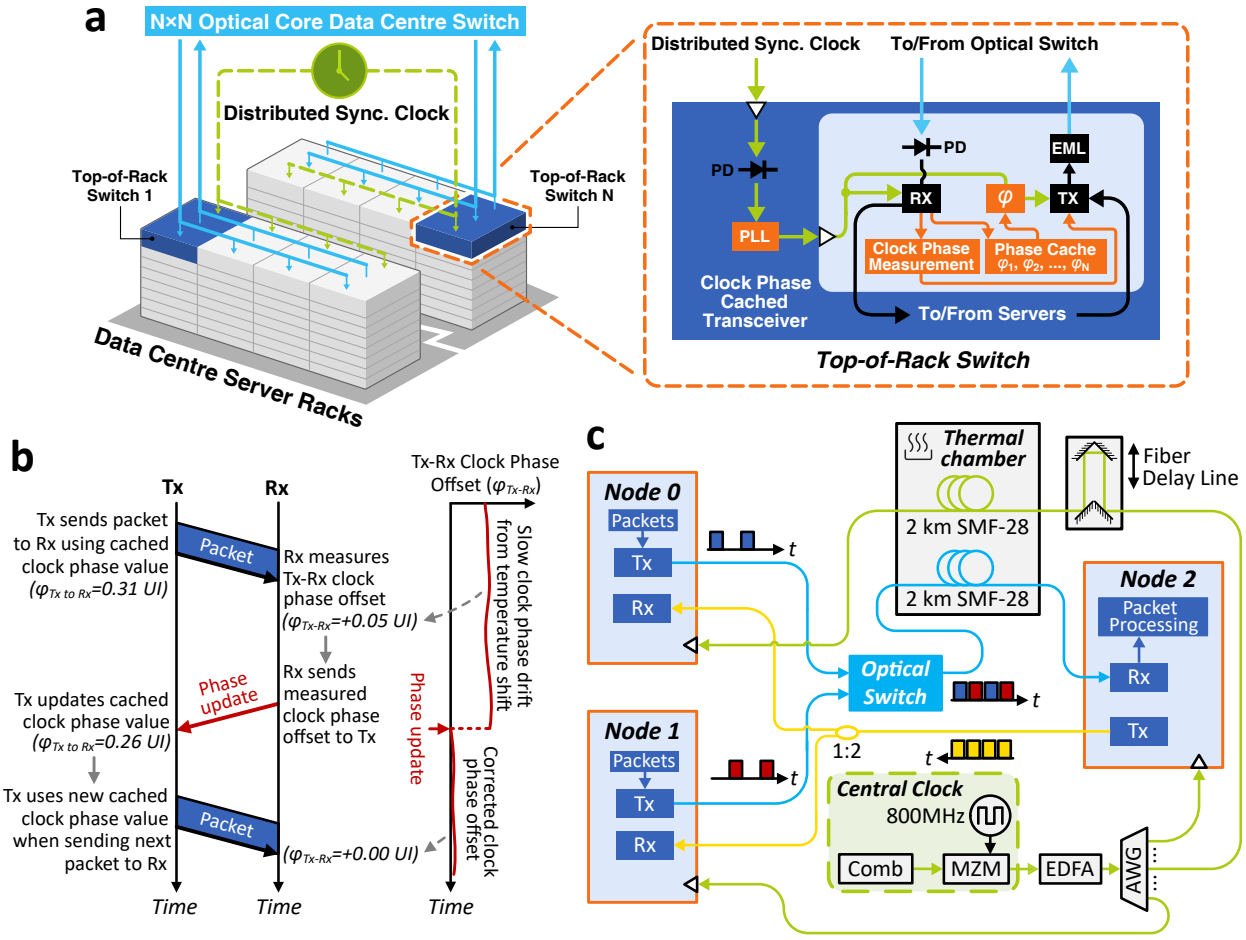

Figure 1: **Clock phase caching architecture, operational principle and demonstration.** **a**, Clock frequency synchronised optically-switched data-centre architecture, using our clock phase caching technique to enable sub-nanosecond CDR locking time. **b**, Operational principle of clock phase caching, showing one clock phase update for one transmitter-receiver pair. In clock phase caching, clock phase updates are performed at a slow rate across all transmitter-receiver pairs. This aligns the clock phase of all data packets arriving at each receiver irrespective of origin, simplifying sub-nanosecond clock and data recovery. UI, unit interval, equal to one symbol period. **c**, Proof-of-concept experimental demonstration of clock phase caching operating on a 2-to-1 optical switch. 2 km clock fibre and 2 km data fibre is placed in a thermal chamber. The signals in these fibres counter-propagate to study the effect of worst-case rates-of-change of clock phase due to temperature on clock phase caching. PLL, phase locked loop; Tx, transmitter; Rx, receiver; ′, clock phase interpolator; EML, externally modulated laser; MZM, Mach-Zehnder modulator; PD, photodiode; EDFA, erbium-doped fibre amplifier; AWG, arrayed-waveguide grating.

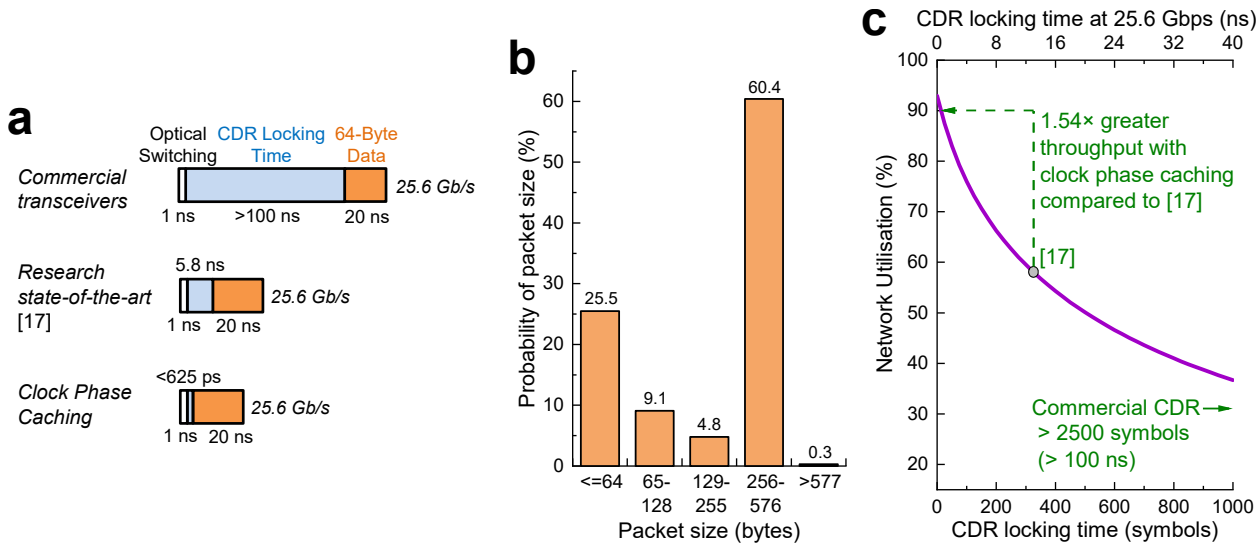

Figure 2: **Increase of data centre network utilisation enabled by clock phase caching.** **a**, Receiver overhead caused by different CDR techniques when receiving minimum size (64-Byte) data packets. **b**, Small packet dominated distribution of packet size in a measured production cloud data-centre traffic pattern. **c**, Gain in data-centre network utilisation from using clock phase caching versus long CDR locking time when handling the traffic pattern shown in **b** (assuming 1 ns optical switching time).

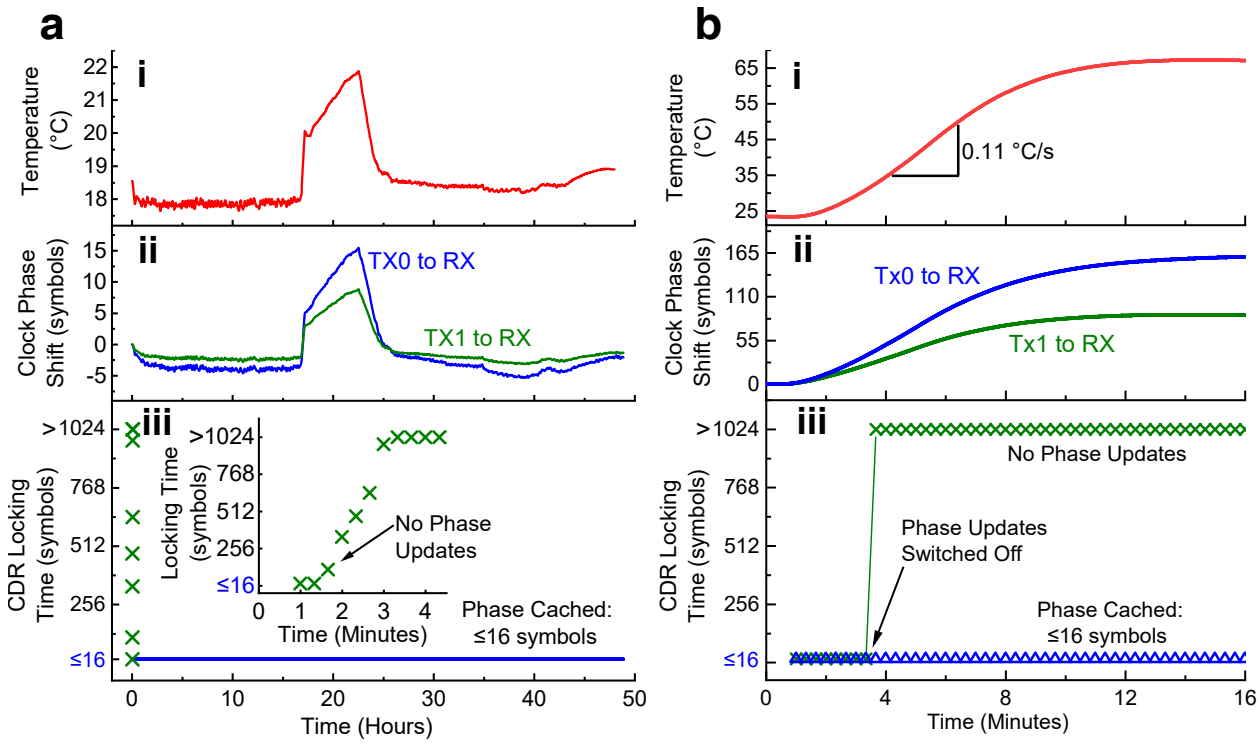

Figure 3: **Demonstration of the stability of clock phase caching.** **a**, Stability over 48 hours of measurement: **i**, Recorded ambient temperature, with the air conditioner switched off between off between the 17th hours and 23rd hours; **ii**, Recorded receiver clock phase shift for packets originating from each transmitter; **iii**, Receiver CDR locking time. Blue line, with clock phase caching; Green crosses, no clock phase caching. **b**, Stability of clock phase caching under a rapid 0.11 C/s rate of temperature change, which is over 3 greater than the worst-case rate of change of temperature we observed in a production cloud data centre: **i**, Temperature within thermally controlled chamber; **ii**, Recorded clock phase shift for packets originating from each transmitter; **iii**, Receiver CDR locking time. Blue triangles, with clock phase caching; Green crosses, clock phase caching turned off during the rapid temperature shift. A straight-line fit is shown as a guide to the eye.

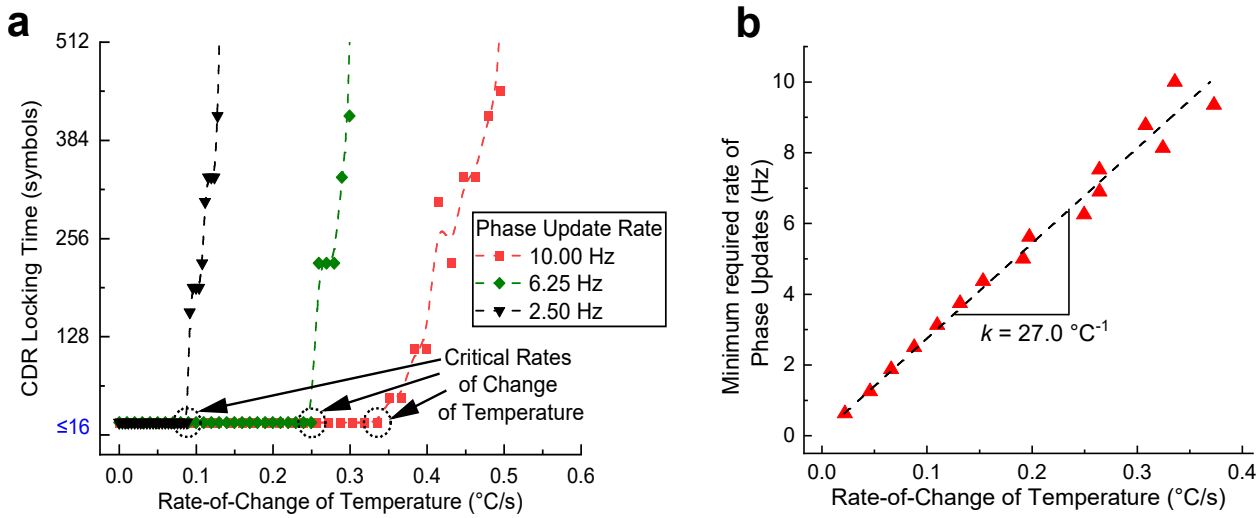

Figure 4: **Impact of rate-of-change of temperature on clock phase caching.** **a**, CDR locking time for different rates of temperature and clock phase cache update rates across 2 km of SMF-28 clock fibre and 2 km of SMF-28 data fibre. In all cases, the rate of clock phase updates is small (<10 Hz). A B-spline fit is shown as a guide to the eye. **b**, Minimum required rate of clock phase updates to achieve 625 ps (16 symbol) CDR locking time for different rates of temperature change for the fibre length described in **a**. Clock phase updates are required at a rate of 27.0 Hz for each 1 C/s of rate-of-change of temperature (note that the worst-case rate of change of temperature that we observed in a production cloud data centre was 0.03 C/s). A linear regression fit (dashed line) was used to calculate this proportionality constant.

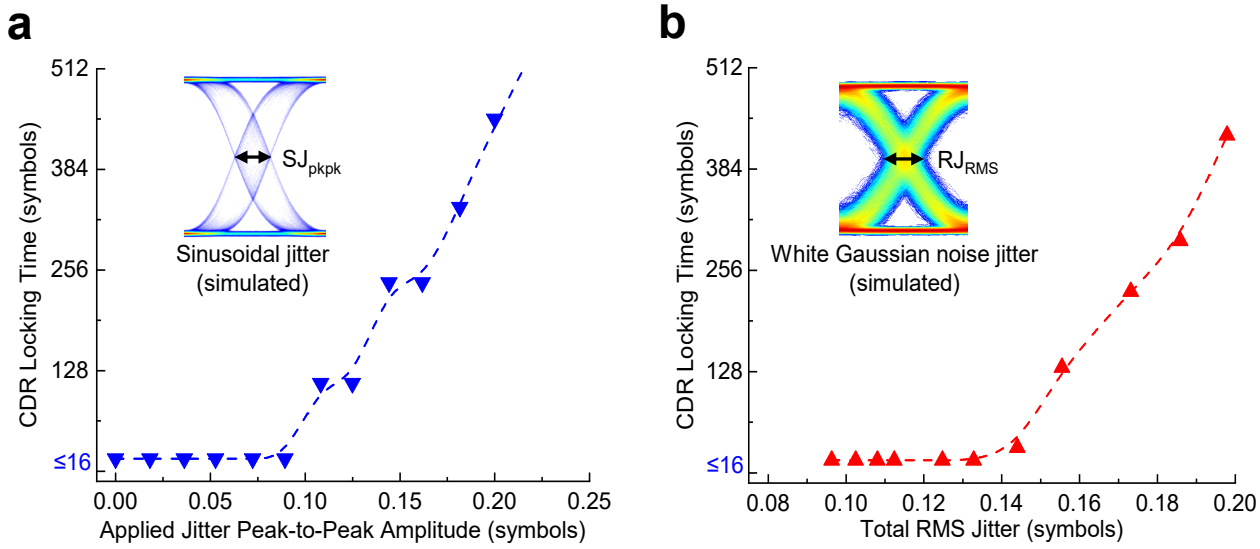

Figure 5: **Impact of clock jitter on CDR locking time.** **a**, Impact of 1 MHz sinusoidal jitter on CDR locking time. The tolerance to applied sinusoidal jitter is 0.090 symbols, corresponding to approximately 3.5 ps at the 25.6 Gb/s data rate used in our experiment. A B-spline fit is shown as a guide to the eye. **b**, Impact of white Gaussian noise jitter on CDR locking time. The RMS jitter tolerance is 0.135 symbols, corresponding to approximately 5.2 ps RMS jitter at 25.6 Gb/s. A B-spline fit is shown as a guide to the eye.

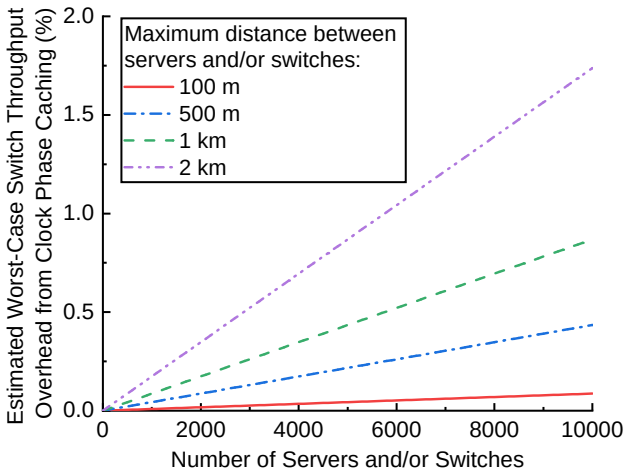

Figure 6: **Estimated worst-case optical switch overhead from clock phase caching technique.** The overhead resulting from clock phase caching is directly proportional to the number of clock phase cached end-points connected to an optical switch. Clock phase caching supports all intra-data-centre distance scales with small network utilisation overhead from performing clock phase cache updates. At the largest intra-data-centre distance scale, with 10,000 servers or switches interconnected by an optical switch with a maximum of 2 km optical fibre between end-points, the estimated overhead from clock phase caching is only 1.7%.