Modeling The Performance of the Clock Phase Caching Approach to Clock and Data Recovery

Kari Clark, *Member, IEEE, Member, OSA,* Zhixin Liu, *Senior Member, IEEE, Senior Member, OSA*

(Invited Paper)

***Abstract*—Optical switching could enable data center networks to keep pace with the rapid growth of intra-data center traffic, however, sub-nanosecond clock and data recovery time is crucial to enabling optically-switched data center networks to transport small packet dominated data center traffic with over 90% efficiency. We review the clock-synchronized approach to clock and data recovery, which enables sub-nanosecond switching time in optically switched networks. We then introduce an analytical model to mathematically explore the operation of clock phase caching, and use this model to explore the impact of factors such as fiber temperature, clock jitter and symbol rate on the BER and clock and data recovery locking time performance of the clock phase caching approach, as well as their impact on scalability. Using commercial data center parameters matching those used in our previous experimental research, we find that our analytical model provides estimates that closely match our previous experimental results, validating its use for making predictions of the performance of clock phase cached systems.**

***Index Terms*—data center networks, optical switching, clock synchronization, clock and data recovery, clock phase caching**

I. Introduction

Optical switching has attracted significant attention in recent research on data centre networks (DCNs) as they promise to be a viable route for the further scaling of hyperscale data centers, so that DCNs can keep pace with the fast growth of machine-to-machine traffic [1]. In optically switched networks, a data packet is transmitted via a momentary optical path through an optical switch, created when two network nodes communicate with each other. At the receiver side, the clock signal of the transmitted optical signals must be recovered before the data can be correctly sampled. The time taken for this process to complete is called the clock and data recovery (CDR) locking time. Due to the dominance of small packets in DCN traffic [2], [3], the CDR locking time must be less than one nanosecond to achieve high (e.g. >90%) network throughput [3]. Such fast CDR locking time imposes one of the main challenges for optically-switched, small-packet dominated DCNs, as well as for any optically switched network that requires small and stable end-to-end transmission latency.

In conventional asynchronous networks, the clock is embedded in the transmitted signals and extracted from the received optical signals. To achieve fast CDR, gated voltage controlled oscillator (GVCO) CDRs have demonstrated CDR locking times of 1 to 2 symbols [4]. However, GVCO CDRs have high power consumption and poor jitter rejection characteristics [5]. Alternatively, time domain oversampling CDRs have demonstrated a CDR locking time of 8 ns [6]. However, they have high circuit complexity and high power consumption because they require a high frequency clock to drive the data sampler [5]. A preferred approach would be to use digital phase interpolator CDRs, which are widely used in commercial transceivers due to their merits of high stability, small silicon area (thus low cost) and low power consumption [5]. However, digital phase interpolator CDRs suffer from metastability, which slows the CDR phase movement and limits the CDR locking time to more than 100s of nanoseconds [3]. To achieve sub-nanosecond CDR locking time, digital phase interpolator CDRs must avoid the initial CDR phase randomly falling within the CDR metastable clock phase region. This, in-turn, requires clock frequency and clock phase synchronization, which cannot be achieved in asynchronous networks.

In contrast to asynchronous networks, synchronous networks offer deterministic end-to-end latency so that data packets can be transmitted within scheduled time slots, minimizing congestion and buffering [1]. Clock signals can be distributed to top-of-rack switches and the transceivers using Sync-E [7] or White Rabbit [8] techniques, through control plane fibers necessary for scheduling. With optical clock synchronization established, CDR modules in each transceiver only need to track the slow change of clock phase, occurring due to change of fiber time-of-flight as temperature varies. It is desirable to reduce the rate of clock phase tracking, or even remove it, if the temperature-induced clock phase drift can be significantly reduced, reducing transceiver power consumption as well as the network overhead [3].

We introduced clock phase caching, which enables sub-nanosecond CDR locking time for clock synchronized optically-switched DCNs by leveraging synchronization to achieve clock phase synchronization for all transmitters and receivers connected to a frequency synchronized optical switch [9]. We also showed that fibers of low thermal sensitivity, such as hollow core fibre (HCF) that offers 20× smaller thermal coefficient of delay versus single-mode optical fibre (SMF-28), can be used to significantly increase the distance scale of clock phase caching approach or reducing the network overhead of optically switched DCN with sub-nanosecond CDR locking time [9].

However, current work on clock phase caching have been purely experimental, and although analytical models of other approaches to CDR exist, such as for oversampling CDRs [6], [10], bang-bang phase interpolator CDRs [11] and for GVCO CDRs [10], no equivalent has been explored for the clock

K. Clark and Z. Liu are with the Department of Electrical and Computer Engineering, UCL, London, United Kingdom, e-mail: kari.clark.14@ucl.ac.uk, zhixin.liu@ucl.ac.uk

Manuscript received August 15, 2021; revised Month Day, Year.

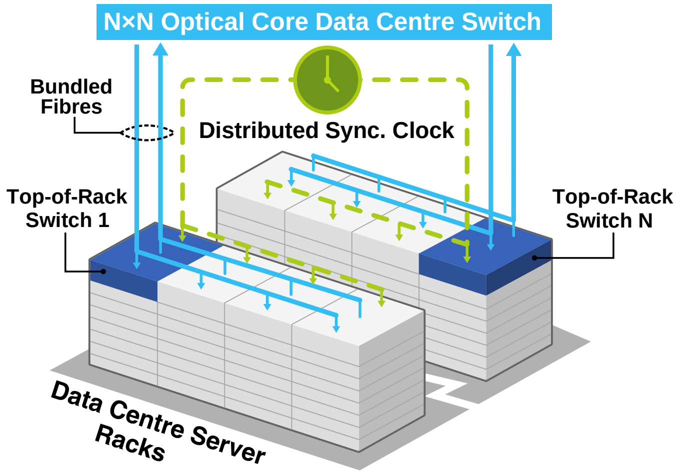

Fig. 1. An example synchronized optical core switch interconnecting N top-of-rack switches in a data center, a key application where clock phase caching can be used to minimize CDR locking time to sub-nanosecond.

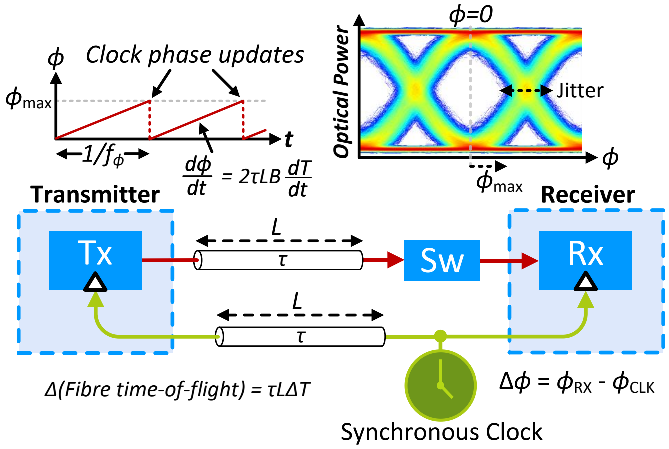

Fig. 2. Factors impacting performance in a clock phase cached transmitter (Tx) to receiver (Rx) pair. The worst-case clock phase offset magnitude in a single worst-case transmitter (Tx) to receiver (Rx) path through the optical switch (Sw) where the clock and data signals entirely counter propagate is shown. ϕ, clock phase offset; ϕ_{max}, maximum clock phase cached clock phase offset; $d\phi/dt$, rate-of-change of clock phase; τ, fiber thermal coefficient of delay; L, fiber length; B, symbol rate; dT/dt, rate-of-change of temperature; ΔT, change in temperature; $\Delta\phi$, clock phase difference; ϕ_{RX}, clock phase of received data; ϕ_{CLK}, phase of clock input to receiver data sampler.

phase cached approach to CDR. Here, we introduce an analytical model of the bit error rate (BER) and CDR locking time performance of clock phase cached links. Using this model, we explore the operation of clock phase caching analytically, while discussing the factors that affect BER and CDR locking time performance of clock phase cached links, and assess the impact of these factors on the scalability of the approach.

II. Principle of Clock Phase Caching

Consider an optical switch that interconnects N network nodes, which each have a transmitter (Tx) and a receiver (Rx), as shown in Fig. 1. Each of the N transmitters connects to N receivers through the optical switch, including its own associated receiver. This results in N^2 different Tx to Rx paths through the optical switch, one of which is shown in Fig. 1. Each of these Tx to Rx paths normally has a unique and changing clock frequency and phase offset, which in asynchronous CDR must be recovered for each packet transmitted through the optical switch. Clock phase caching instead removes these offsets for all paths through the optical switch, ensuring that all packets arrive at a given Rx arrive with the same clock frequency and phase irrespective of the origin Tx, which simplifies the CDR locking process. The clock frequency offsets are removed by frequency synchronizing all network nodes (green dashed line in Fig. 1). The clock phase offsets are removed by measuring all phase offsets (at a slow rate of about 1 to 10 Hz), 'caching' the phase offsets at the transmitter (or receiver), and applying a clock phase shift before the transmission of each packet along a Tx to Rx path with a phase interpolator, which cancels out the clock phase offset of that Tx to Rx path [3].

III. Factors Impacting Performance in a Clock Phase Cached Transmitter to Receiver Pair

In this section, we will introduce and model the main factors that together determine the performance of clock phase cached links. The performance of a clock phase cached Tx to Rx pair depends on four main factors: the received signal quality (i.e. signal to noise ratio or quality factor); the relative clock jitter between the sampling clock and the received signal; the clock phase update rate; and the movement of the CDR as it corrects for the residual initial clock phase offset of each packet, which are illustrated in Fig. 2 and Fig. 4. We define the CDR locking time as the time taken since the beginning of packet reception for the BER to fall below a threshold. We use a threshold BER of 10^{-10}, matching our previous experimental work [3], [9]. The four factors in this section interact in the following manner: the received signal quality and relative clock jitter determine the maximum tolerable clock phase offset before the BER at the beginning of incoming packets degrades to worse than 10^{-10}; the clock phase update rate then must be high enough to ensure that the worst-case clock phase offset is smaller than this maximum tolerable clock phase offset; finally, the movement of the CDR determines the rate of improvement of BER in incoming packets.

A. Received signal quality

The BER of a received non-return to zero on-off keying (NRZ-OOK) signal depends on the signal current amplitude of the zero and one levels, I_0 and I_1 respectively, as well as the root mean square noise for the 0 and 1 levels, σ_0 and σ_1 respectively, and the decision point, I_D, which all depend on the sampling clock phase offset, ϕ. The BER as a function of clock phase offset, ϕ, assuming an equal probability of receiving a 0 or a 1, is [12]:

$$\text{BER}(\phi) = \frac{1}{4}\left(\text{erfc}\left(\frac{I_1(\phi) - I_D(\phi)}{\sqrt{2}\sigma_1(\phi)}\right) + \text{erfc}\left(\frac{I_D(\phi) - I_0(\phi)}{\sqrt{2}\sigma_0(\phi)}\right)\right) \quad (1)$$

where $\operatorname{erfc}(z)$ is the complementary error function, defined as [13]:

$$\operatorname{erfc}(z) \triangleq 1 - \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2)dt \tag{2}$$

In intra-data center transmission, the optical extinction ratio may be small. Accounting for a finite optical extinction ratio, $I_1(\phi)$ and $I_0(\phi)$ can be defined as:

$$I_1(\phi) \triangleq (I_{\text{max}} - I_{\text{min}})v_1(\phi) + I_{\text{min}} \tag{3}$$

$$I_0(\phi) \triangleq (I_{\text{max}} - I_{\text{min}})v_0(\phi) + I_{\text{min}} \tag{4}$$

where $v_1(\phi)$ and $v_0(\phi)$ are the unit amplitude pulses for the 1 and 0 levels respectively, I_{max} is the maximum received signal current and I_{min} is the minimum received signal current.

Then consider that the extinction ratio of the received optical signal, r_e, is defined as the ratio between the maximum and minimum received optical powers, P_{max} and P_{min} respectively, $P_{\text{max}} = r_e P_{\text{min}}$ [12], that the average received power, P_{avg}, is the mean of P_{max} and P_{min}, i.e. $P_{\text{avg}} = \frac{1}{2}(P_{\text{max}} + P_{\text{min}})$ [12], and that the received signal current amplitude, I, is given by $I = RP$ [12], where R is the photodiode responsivity and P is the received optical power. Using these relationships, expressions for the maximum and minimum signal currents, I_{max} and I_{min} can be obtained:

$$I_{\text{max}} = \frac{2RP_{\text{avg}}r_e}{r_e + 1} \tag{5}$$

$$I_{\text{min}} = \frac{2RP_{\text{avg}}}{r_e + 1} \tag{6}$$

We then define the unit amplitude pulses for the 1 and 0 levels, $v_1(\phi)$ and $v_0(\phi)$ respectively, as non-return to zero (NRZ) Gaussian pulses derived from a Gaussian filtered square NRZ pulse [14]:

$$v_1(\phi) \triangleq \frac{1}{2}\operatorname{erf}\left(\frac{2\pi f_c(\phi + \frac{1}{2})}{B\sqrt{2\ln 2}}\right) - \frac{1}{2}\operatorname{erf}\left(\frac{2\pi f_c(\phi - \frac{1}{2})}{B\sqrt{2\ln 2}}\right) \tag{7}$$

$$v_0(\phi) \triangleq 1 - v_1(\phi) \tag{8}$$

where B is the symbol rate and f_c is the filter bandwidth.

We use an NRZ Gaussian pulse shape to minimize the mathematical complexity of our analytical model, allowing for closed form expressions while still allowing the impact of transceiver bandwidth on pulse shape to be accounted for in our model. However, note that in practice transmitter pulse shapes are rarely Gaussian, and alternatives, such as those derived using a 4th order Bessel-Thomson filter [14], could be used to model signals generated from practical transmitters (e.g. using Mach-Zehnder modulators (MZMs), electro-absorption modulators (EAMs)).

The signal amplitude of the 1 level in terms of clock phase ϕ, $I_1(\phi)$, can then be obtained by substituting (5) and (6) into (3), and the signal amplitude of the 0 level in terms of clock phase ϕ, $I_0(\phi)$, can be obtained by substituting (5) and (6) into (4):

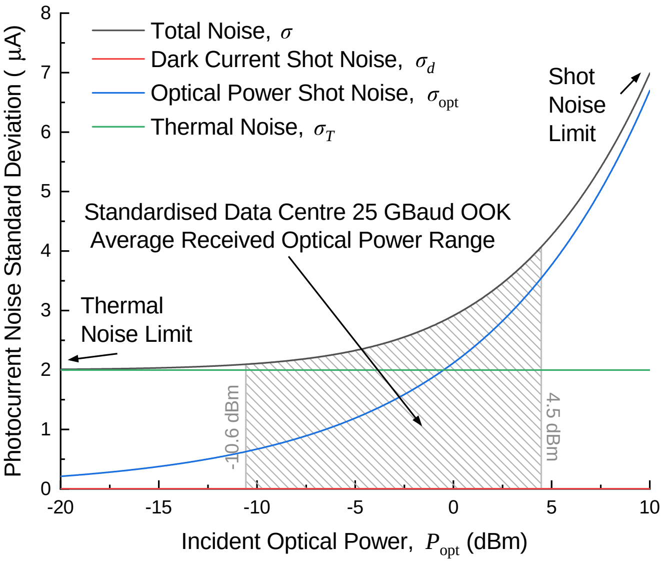

Fig. 3. Total photocurrent noise for a typical photoreceiver consisting of a PIN photodiode and a TIA, designed for 25 GBaud OOK signal reception in an intra-data center link. The standardized average received optical power range of -10.6 to 4.5 dBm in data center 25 GBaud transmission is shown [15].

$$I_1(\phi) = 2RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)v_1(\phi) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{9}$$

$$I_0(\phi) = 2RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)v_0(\phi) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{10}$$

We will now derive the noise currents as a function of clock phase for the 1 and 0 levels, $\sigma_1(\phi)$ and $\sigma_0(\phi)$ respectively. In an intensity modulated direct detection (IM-DD) system, the total noise current in a photoreceiver, σ, is given by [12]:

$$\sigma = \sqrt{\sigma_T^2 + \sigma_s^2 + \sigma_I^2} \tag{11}$$

where σ_T is the root mean square of the thermal noise, σ_S is root mean square of the shot noise and σ_I is the root mean square of the relative intensity noise. We will derive the noise current assuming that the relative intensity noise is negligible, as this greatly reduces the mathematical complexity of our model.

In typical data center receivers that comprise a PIN photodiode followed by a transimpedance amplifier (TIA), thermal noise is dominated by the TIA's amplifier noise figure, and the thermal noise current of the photoreceiver is therefore well approximated by the TIA's input referred noise, σ_{ni}:

$$\sigma_T^2 = \sigma_{ni}^2 \tag{12}$$

The shot noise is proportional to the sum of the signal current, I_p, and the PIN photodiode dark current, I_d, and is given by:

$$\sigma_s^2 = 2q(I_p + I_d)f_c = 2q(RP_{\text{opt}} + I_d)f_c \tag{13}$$

where q is the elementary charge and f_c is the photoreceiver bandwidth.

Fig. 3 illustrates the contribution of the above noise currents to overall photoreceiver noise, assuming a typical data center photoreceiver designed for 25 GBaud OOK signal reception with a responsivity, R, of 0.80 A/W [16]; a photodiode dark current, I_d, of 2 nA [16]; a TIA input referred noise, σ_{ni}, of 2 μA [17]; and a TIA-limited photoreceiver bandwidth of 17.5 GHz [17]. Fig. 3 illustrates that the contribution of the dark current to overall noise is negligible at all received optical powers, that thermal noise greatly dominates shot noise at the standardized minimum received optical power of -10.6 dBm, and that the received optical power contribution to shot noise contributes significantly to overall noise at incident optical powers greater than approximately -7.5 dBm.

Our modeling will primarily focus on the performance of clock phase caching at the standardized minimum average received data center optical power of -10.6 dBm [15], as the signal to noise ratio at this average received optical power is minimized. This leads to the worst-case for the performance of clock phase caching, matching the conditions used in our previous experimental investigations into clock phase caching [3]. However, we will include both the contribution of thermal noise and the received optical power contribution to shot noise in our model so that it may be used to predict the performance of clock phase caching for any standardized data center average received optical power.

The noise currents of the 1 and 0 levels as a function of clock phase, $\sigma_1(\phi)$ and $\sigma_0(\phi)$ respectively, are then given by:

$$\sigma_1(\phi) = \sqrt{2qI_1(\phi)f_c + \sigma_{ni}^2} \tag{14}$$

$$\sigma_0(\phi) = \sqrt{2qI_0(\phi)f_c + \sigma_{ni}^2} \tag{15}$$

Finally, to obtain the BER as a function of clock phase, we first assume that the photoreceiver does not optimize the decision point, I_D, to minimize BER, and instead samples at the average current halfway between the 0 and 1 levels of $I_D = RP_{avg}$. We then substitute this expression for I_D, as well as the expressions for $\sigma_1(\phi)$ and $\sigma_0(\phi)$ given in (14) and (15), into (1), resulting in the following final expression modeling the impact of clock phase offset on BER:

$$\begin{aligned}\text{BER}(\phi) =& \frac{1}{4}\Bigg(\text{erfc}\left(\frac{I_1(\phi) - RP_{avg}}{\sqrt{4qI_1(\phi)f_c + 2\sigma_{ni}^2}}\right) \\ &+ \text{erfc}\left(\frac{RP_{avg} - I_0(\phi)}{\sqrt{4qI_0(\phi)f_c + 2\sigma_{ni}^2}}\right)\Bigg)\end{aligned} \tag{16}$$

where $I_1(\phi)$ and $I_0(\phi)$ are the signal currents for the 1 and 0 levels as a function of clock phase, given by (9) and (10) respectively.

B. Relative jitter between sampling clock and received data

In a clock phase cached system, the position of the sampling clock with respect to the incoming data is subject to clock jitter. Increased jitter causes more samples to be taken in a region of the pulse shape with worse BER leading to performance degradation. As clock phase caching only compensates for the slow drift of the clock phase (about 10 Hz caching rate in our demonstration), it cannot compensate for clock phase noise above the clock phase update rate. Therefore, the integrated phase noise from about 1 to 10 Hz (depending on clock phase update rate) to about 10 MHz (the clock filter bandwidth) determines the overall jitter.

In a receiver that is not clock phase cached and therefore has no clock phase offset resulting from fiber temperature change, the BER after applying jitter, $\text{BER}_{\text{post-jit}}$ can be obtained by integrating the jitter probability distribution function (PDF) with the unjittered BER, $\text{BER}_{\text{pre-jit}}$ [14]:

$$\text{BER}_{\text{post-jit}} = \int_{-\infty}^{\infty} \text{PDF}_{\text{jit}}(\alpha)\text{BER}_{\text{pre-jit}}(\alpha)\, d\alpha \tag{17}$$

where α is the clock phase offset induced by timing jitter.

In a receiver that is clock phase cached, and therefore does have a clock phase offset, ϕ, resulting from fiber temperature change, the impact of jitter may be modeled by extending (17) to a cross correlation, which is indicated by the 'star' symbol:

$$\text{BER}_{\text{post-jit}}(\phi) = (\text{PDF}_{\text{jit}} \star \text{BER}_{\text{pre-jit}})(\phi) \tag{18}$$

$$\text{BER}_{\text{post-jit}}(\phi) = \int_{-\infty}^{\infty} \text{PDF}_{\text{jit}}(\alpha)\text{BER}_{\text{pre-jit}}(\phi+\alpha)\, d\alpha \tag{19}$$

Assuming Gaussian jitter, i.e. that the jitter timing variation follows a Gaussian distribution with root mean square (rms) sigma, σ_{jit}, to maximize model simplicity, assuming that all jitter is random with no deterministic contribution, the BER as a function of clock phase offset, ϕ, after accounting for the impact of jitter then is given by:

$$\text{BER}_{\text{post-jit}}(\phi) = \int_{-\infty}^{\infty} \frac{1}{\sigma_{\text{jit}}B\sqrt{2\pi}} e^{-\frac{\alpha^2}{2\sigma_{\text{jit}}^2 B^2}} \text{BER}_{\text{pre-jit}}(\phi+\alpha)\, d\alpha \tag{20}$$

where B is the symbol rate of the received data and σ_{jit} is the Gaussian jitter magnitude.

C. Clock phase update rate

Each clock phase cached Tx to Rx pair has an associated clock phase update rate, f_ϕ, which is the rate at which clock phase updates are performed for that pair of nodes. Each clock phase update for a given Tx to Rx pair compensates for changes in fiber time-of-flight changes that have occurred since the last clock phase update for that pair, as illustrated for packets arriving at a single receiver from two different transmitters in Fig. 4.

For each link, there will be a maximum tolerable clock phase offset, ϕ_{max}, above which the BER at the beginning of packet reception degrades to below a critical threshold (e.g. 10^{-10} as used in our previous experimental work [3] based on IEEE 802.3 Ethernet standardization [15]). This maximum tolerable clock phase offset is dependent on the receiver signal quality and clock jitter magnitude. Better receiver signal quality, and/or a smaller clock jitter magnitude, result in a larger maximum tolerable clock phase offset. However, the magnitude of the worst-case clock phase offset that occurs for a given clock phase cached pair is dependent on series of factors: the fiber length used to interconnect the clock phase cached pair, the thermal coefficient of delay of that fiber,

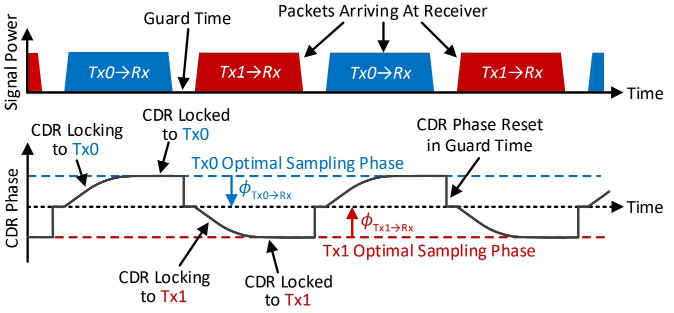
Signal Power
Guard Time
Packets Arriving At Receiver
Tx0→Rx
Tx1→Rx
Tx0→Rx
Tx1→Rx
Time
CDR Phase
CDR Locking to Tx0
CDR Locked to Tx0
Tx0 Optimal Sampling Phase
$\phi_{Tx0\rightarrow Rx}$
CDR Phase Reset in Guard Time
$\phi_{Tx1\rightarrow Rx}$
Time
Tx1 Optimal Sampling Phase
CDR Locking to Tx1
CDR Locked to Tx1

Fig. 4. Packets arriving at a receiver, Rx, from two transmitters, Tx0 and Tx1. A small residual initial sampling offset remains for each of the transmitter to receiver pairs, $\phi_{\mathrm{Tx0}\rightarrow\mathrm{Rx}}$ and $\phi_{\mathrm{Tx1}\rightarrow\mathrm{Rx}}$, as a consequence of time-of-flight changes that have occurred since the last clock phase update for each pair. The CDR locks to each small residual initial sampling offset as each new packet arrives. The CDR phase is reset between each packet to minimize the worst-case magnitude of clock phase that the CDR must travel.

the rate of temperature change that fiber is subjected to, the symbol rate of data transmitted through the link and the rate at which clock phase updates are performed.

To derive the required rate of clock phase updates to cope with a given maximum tolerable clock phase offset, first consider that the magnitude of the time-of-flight (ToF) change, Δ(ToF), in an optical fiber subjected to a change of temperature, ΔT, is dependent on the thermal co-efficient of that fiber, τ (about 40 ps/km/°C for SMF-28 [18]) and the length of that fiber, L:

$$\Delta(\mathrm{ToF}) = \tau L \Delta T \tag{21}$$

For example, if 1 km of SMF-28 is subjected to a 1 °C temperature increase, the time-of-flight through that fiber will increase by approximately 40 ps. Furthermore, if that fiber is subjected to a linear temperature change, dT/dt, and we consider that the change in clock phase offset, $\Delta\phi$, as a function of change of time-of-flight is dependent on the symbol rate, B, and is given by $\Delta\phi = B\Delta(\mathrm{ToF})$, then the rate-of-change of clock phase through that fiber, $d\phi/dt$, is:

$$\frac{d\phi}{dt} = \tau L B \frac{dT}{dt} \tag{22}$$

If we assume that the fiber is subjected to an approximately linear temperature change over the short 100 ms to 1 s timescale between each clock phase update occurring at 1 to 10 Hz, the worst-case time-of-flight change then occurs just before each clock phase update, or when $dt = 1/f_\phi$. The time-of-flight through both the clock and data fibers contributes to the magnitude of the overall time-of-flight change at the receiver. As illustrated in Fig. 2, this time-of-flight change is maximized when the receiver is adjacent to the clock source and switch, with the transmitter located as far as possible away from both of these, causing the time-of-flight increase through the clock and data fibers (each of length L) to be entirely additive, leading to a factor of 2 increase in the worst-case time-of-flight change. Together, the required rate of clock phase updates, f_ϕ, to maintain a worst-case clock phase offset of below $\phi_{\max}$ is then:

$$f_\phi \approx \frac{2\tau L B}{\phi_{\max}} \frac{dT}{dt} \tag{23}$$

D. Clock and data recovery circuit clock phase movement

In a clock phase cached link, upon arrival of each data packet with a small residual clock phase offset due to temperature shift since the last clock phase update, the receiver CDR will shift clock phase to the ideal sampling point for that packet. This results in an improvement of BER as a function of time since the packet first started to be sampled. This subsection will model this CDR behavior.

The clock phase movement, $d\phi$, generated during each phase measurement interval, dt, of an ideal bang-bang phase interpolator CDR is given by:

$$d\phi = \begin{cases} pdt, & \text{if } \phi < 0 \\ -pdt, & \text{if } \phi > 0 \\ 0, & \text{if } \phi = 0 \end{cases} \tag{24}$$

where p is the CDR proportional gain constant.

Given an initial clock phase offset, ϕ_0, (25) can used to show the linear progression of the CDR phase as a function of the time since the beginning of packet reception, t, which continues until $\phi = 0$ is reached:

$$\phi(t, \phi_0) = \begin{cases} \phi_0 - pt, & \text{if } 0 < \phi_0 < 0.5 \text{ and } \phi_0 - pt > 0 \\ \phi_0 + pt, & \text{if } -0.5 < \phi_0 < 0 \text{ and } \phi_0 + pt < 0 \\ 0, & \text{if } \phi_0 = -0.5,\ 0 \text{ or } 0.5 \end{cases} \tag{25}$$

where $\phi = \pm 0.5$ is half a symbol from the ideal sampling point at $\phi = 0$, at which the behavior of the CDR repeats.

Fig. 5 illustrates the clock phase movement of an ideal bang-bang phase interpolator CDR for different initial clock phase offsets, ϕ_0, as described by (25).

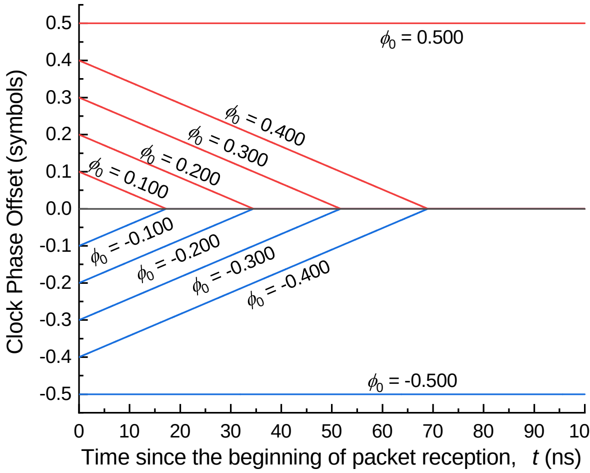

Fig. 5. Clock phase movement of an ideal bang-bang phase interpolator CDR with a proportional gain, p, of 5.8 symbols/μs, during the process of locking to an incoming data packet, for a series of initial clock phase offsets, ϕ_0.

In practice clock jitter slows the clock phase movement rate as $\phi = 0$ is approached, linearizing the clock phase movement rate at clock phase offsets close to $\phi = 0$ [19]. The clock phase movement is also slowed at initial clock phases close to $\phi = \pm 0.5$, which causes the metastability effect leading to long CDR locking time in bang-bang phase interpolator CDRs. The effect of clock jitter on CDR phase movement rate will not be modeled here to minimize model complexity, but it is however modeled and explored in detail in Chapter 5 of [20]. In this model, we assume that the behavior of the CDR is ideal with no impact from jitter, and that the residual clock phase offset therefore shifts towards $\phi = 0$ at a rate of $d\phi/dt = -p$ for positive initial clock phase values and at a rate of $d\phi/dt = p$ for negative initial clock phase offsets, and then remains at $\phi = 0$ once $\phi = 0$ is reached.

IV. Analytical Modeling of Clock Phase Cached Link Performance

Based on the modeling described in section III, we show the performance of a clock phase cached link using an analytical approach, using parameters that match those used in commercial DCNs and those used in our experimental work, which are listed in Table I.

TABLE I
Data center, photodiode and TIA properties matching those used in commercial DCNs and in our experimental work.

Data center / Photodiode / TIA property	**Value**
SMF-28 thermal coefficient of delay, τ	40 ps/km/°C [18]
Length of clock and data fiber, L	2 km [15]
Symbol rate, B	25 GBaud [15]
Photodiode responsivity, R	0.80 A/W [16]
Average received optical power, P_{opt}	-10.5 dBm [15]
TIA input referred noise, σ_{ni}	2 μA [17]
TIA bandwidth, f_c	17.5 GHz [17]
Gaussian jitter magnitude, σ_{jit}	2.4 ps [3]
CDR proportional gain, p	5.8 symbols/μs [20]
Extinction ratio, r_e	2.51 (4 dB [15])

We numerically analyzed the BER and CDR locking time performance of clock phase caching using the following steps:

1) A receiver pulse shape as a function of clock phase offset was generated using (7), (8), (9), (10) and (16) with the receiver performance parameters given in Table I.
2) This pulse shape was then cross-correlated with a Gaussian jitter PDF using (20) to generate a look-up-table of BER as a function of clock phase offset, which includes the impact of Gaussian jitter.
3) A series of initial clock phase values between -0.5 and 0.5 symbols were used with (25) to generate clock phase offset values as a function of time since the beginning of packet reception, t, and initial clock phase offset, ϕ_0.
4) These clock phase offset values were then used with the look-up-table of BER as a function of pulse shape and jitter generated in Step 2) to generate BER as a function of time since the beginning of packet reception, t, and initial clock phase offset, ϕ_0.
5) The BER as a function of time since the beginning of packet reception was then used to calculate the time taken for BER to fall below 10^{-10}. This threshold BER matches the threshold BER used to determine CDR locking time in our previous clock phase caching experiments [3], based on IEEE 802.3 Ethernet standardization [15].

Fig. 6 and Fig. 7 show BER as a function of initial clock phase offset and time since the beginning of packet reception. Fig. 8 then shows the impact of the progression of BER at different initial clock phase offsets, ϕ_0, on CDR locking time. These figures illustrate that CDR locking time can be minimized to sub-nanosecond using clock phase caching if the worst-case initial clock phase offset, ϕ_{max}, which occurs just before each clock phase update occurs, is constrained to be less than approximately 0.103 symbols. Note that if longer CDR locking times are tolerable (which would not be typical for data center traffic patterns as they are typically dominated by small packets, requiring sub-nanosecond CDR locking time for high optical switch throughput [3]), Fig. 8 shows that the maximum tolerable initial clock phase offset would in that case be larger, which would in-turn require a lower rate of clock phase updates to maintain.

Given a worst-case rate-of-change of data center temperature, dT/dt, of 0.03 °C/s and the parameters listed in Table I matching those in our previous experiment [3], (23) can be used to calculate that the estimated minimum rate-of-clock phase updates to maintain an initial clock phase offset of

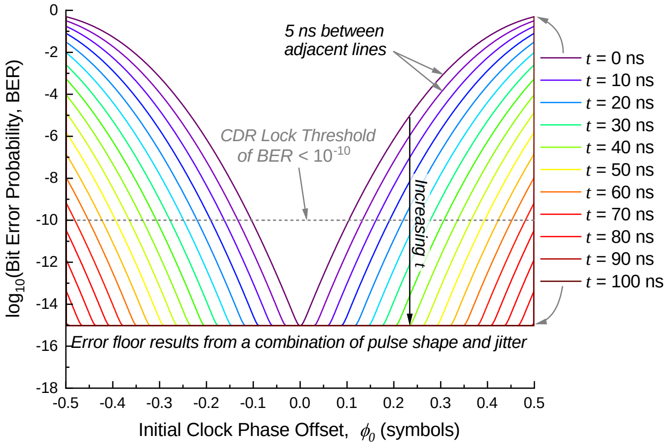

Fig. 6. Bit error probability as a function of initial clock phase offset, ϕ_0, for a series of different times since the beginning of packet reception, t. Moving vertically downwards from one colored line to the nearest adjacent line represents an increase in time since packet reception of 5 ns.

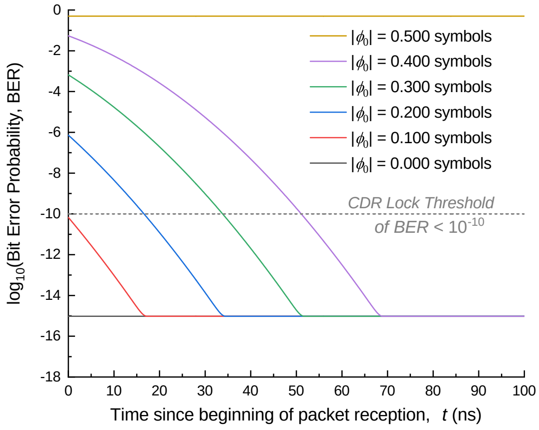

Fig. 7. Bit error probability as a function of time since the beginning of packet reception, t, for different initial clock phase offsets, ϕ_0. These initial clock phase offsets occur as changes in temperature causes the fiber time-of-flights within each clock phase cached Tx to Rx pair to change.

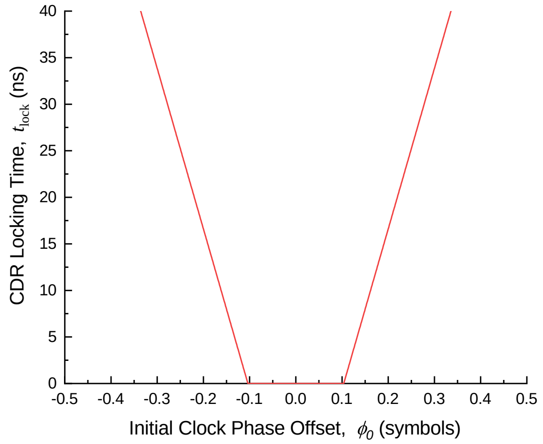

Fig. 8. Increase of CDR locking time beyond instantaneous for initial phase offsets above approximately ±0.103 symbols, using a threshold BER for calculating the CDR locking time of 10^{-10} and the modeled data used to generate Fig. 7 and Fig. 6.

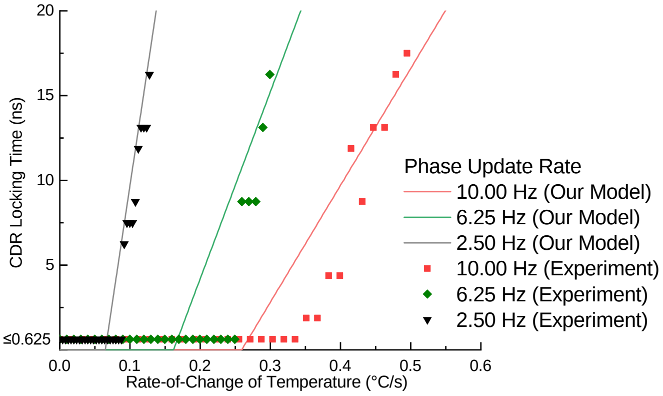

Fig. 9. Impact of rate-of-change of temperature on clock phase caching for different clock phase update rates. The solid lines show the CDR locking time estimated by our analytical model, and the red, green and black points show our previous experimental results from [3].

0.103 symbols in a worst-case data center environment is 1.2 Hz. This value is a good approximation to our experimentally determined minimum required clock phase update rate of 0.81 Hz, which was calculated by multiplying the experimentally determined proportionality constant of 27.0 °C^{-1} from Fig. 4b of [3] by our measured worst-case rate-of-change of temperature in a data center of 0.03 °C/s [3].

In Fig. 9, we compare our experimental results from [3] with the CDR locking time estimated from our model for different rates-of-change of temperature, dT/dt, and clock phase update rate, f_ϕ, by using (23), the clock phase values in Fig. 8 and the distance, L, and fiber thermal coefficient of delay parameters, τ, listed in Table I. We again show that our model well approximates our previous experimental results, which justifies that our analytical model can be used to perform good estimations of the performance of clock phase cached links using data center operational parameters.

V. Scalability of Clock Phase Cached Systems

The impact of the rate of clock phase updates is that it determines scalability, as each clock phase update for each Tx to Rx pair introduces a time period where the pair is forced to communicate so that the clock phase for that pair can be measured and then updated, even if they do not have data to exchange between them. This time period introduces a worst-case overhead on throughput caused by clock phase caching, which is directly proportional to the number of transmitters that can communicate with each receiver. An estimate of the worst-case network overhead caused by clock phase caching, o_{max}, can be calculated from:

$$o_{\text{max}} = N f_\phi (t_{\text{meas}} + t_{\text{update}}) \tag{26}$$

where N is the number of end-points (servers or switches) connected to the optical switch, f_ϕ is the rate of clock phase updates, t_{meas} is the time a receiver takes to perform each clock phase measurement and t_{update} is the time taken to send each update from a receiver back to a transmitter. A Tx to Rx pair that experiences a greater rate of change of clock phase will require a greater rate of clock phase updates. To estimate the impact of different data center and interconnect parameters on worst-case estimated clock phase caching overhead, (26) can be combined with (23) by substituting for the clock phase update rate, f_ϕ:

$$o_{\text{max}} \approx \frac{2\tau LBN}{\phi_{\text{max}}} \frac{dT}{dt} (t_{\text{meas}} + t_{\text{update}}) \tag{27}$$

where τ is the optical fiber thermal coefficient of delay (TCD), L is the length of the clock and data fibers, B is the symbol rate, ϕ_{max} is the maximum tolerable clock phase offset before the initial packet BER increases to above the CDR locking threshold and $\frac{dT}{dt}$ is the worst-case rate-of-change of data center temperature.

For a real data center environment with a worst-case measured rate of temperature change of 0.03 °C/s, with 2 km clock and 2 km data SMF-28 between each pair, transmitting 25.6 Gb/s NRZ-OOK packet signals measured from our experiment at minimum data center received optical powers of -10.5 dBm, we found experimentally that a clock phase update rate of 0.81 Hz was sufficient to track the worst-case clock phase change [3]. With a time taken to perform each clock phase update, t_{meas}, of 2.08 μs and a time taken to transmit each clock phase update, t_{update}, of 0.065 μs from our experimental work [3], (26) can be used to estimate a worst-case overhead of 1.7% for 10,000 optical switch end-points in an optically switched DCN. Using the clock phase update rate of 1.2 Hz derived from our analytical model with (26) gives an estimated worst-case overhead of 2.5% for 10,000 optical switch end-points, again a good approximation to our previous experiment-derived estimate.

This worst-case overhead can be lowered by reducing the optical fiber TCD, τ, and therefore reducing the required rate of clock phase updates, f_ϕ, by the same factor, by using low TCD fibers such as hollow core fiber (20 times smaller TCD and therefore 20 times smaller required f_ϕ and worst-case overhead than SMF-28 [9]). This overhead can also be potentially reduced by using homogeneous multicore fibre

(MCF), which features a low thermal coefficient of skew of about 40 fs/km/°C between different cores [21]. Further modeling of system performance and scalability in a clock phase cached optical switch using SMF-28 and HCF may be found in Chapters 6 and 7 respectively of [20].

VI. Discussion

In addition to the change of fiber time-of-flight primarily explored in this paper, the change of temperature in DCNs also leads to wavelength drift of optical components such as the lasers and wavelength multiplexers/demultiplexers. For solitary semiconductor lasers, the operational wavelength typically changes by about 0.1 - 0.2 nm/°C, resulting in a wavelength drift of up to 8 nm when operating uncooled. This relatively large wavelength drift causes the optical impairments to change, such as a change of loss, optical filtering shape and dispersion, even for coarse wavelength division multiplexed (CWDM) systems. Although these impairments are negligible for low baud rate (e.g. 25 GBd) NRZ-OOK signals, they become prominent as DCN interconnects evolve to higher baud rates (e.g. to above 50 GBd) and to higher order modulation formats such as 4-level pulse amplitude modulation (PAM-4) [22], [23]. Techniques such as equalizer caching could be used for improved system flexibility and reliability [22].

Another possible approach to mitigate the impact of this effect is to disaggregate the light sources from the switches, using tuneable lasers and wavelength demultiplexers/multiplexers for switching [24]. However, the thermal crosstalk between tuneable lasers on the same photonic circuit may cause unwanted wavelength shifts that degrade system performance [25]. In addition, demultiplexer/multiplexers may perform poorly in an unpredictable thermal environment, causing degraded performance or link failure due to thermally dependent wavelength drift [22], [24].

Furthermore, as the symbol rate increases from 25 GBd to 50 GBd and beyond, the source clock jitter becomes the main limiter of system performance and must be minimized to achieve stable and low BER [26]. By characterizing the BER performance of a 51.2-GBd clock synchronized optically switched system, we have shown that a low phase noise source clock is required to ensure sufficient high system performance. Indeed, clock phase caching must work with other novel technologies for future optically switched data center networks.

The analytical model presented here was developed for cross-validation with our previous experimental results, and so was designed to model NRZ-OOK signals as used in our previous experiments [3]. Our analytical model could be extended to model PAM-4 and other higher-order amplitude modulated signals by incorporating filter responses that affect the eye width and height of the different signal levels.

VII. Conclusion

We reviewed our recent results on sub-nanosecond CDR enabled by clock synchronization and clock phase caching for optically-switched DCNs. We developed an analytical model that can be used to make predictions of the BER and CDR locking time performance of clock phase cached links, by analyzing the key factors that affect clock phase cached optical link performance. We show that our analytical model allows for good estimations of the performance of clock phase cached links to be made, and show that the proposed CDR scheme can be scaled to support 10,000 node DCNs. Furthermore, we note that more than 20 times lower clock phase caching overhead can be achieved using HCF. We also emphasize that low phase noise source clocks and equalizer status caching should be considered for high baud rate (e.g. >50 GBd) interconnects.

Data availability statement

The data that supports the figures within this paper are available from the UCL Research Data Repository (DOI: 10.5522/04/17057804), which is hosted by FigShare.

Acknowledgments

The authors acknowledge financial support from Microsoft, Inphi Inc. and EPSRC grants EP/R041792/1, EP/V051377/1 and EP/R035342 and the Royal Society Paul Instrument Fund PIF/R1/180001.

References

[1] H. Ballani, P. Costa, R. Behrendt, D. Cletheroe, I. Haller, K. Jozwik, F. Karinou, S. Lange, K. Shi, B. Thomsen *et al.*, "Sirius: A flat datacenter network with nanosecond optical switching," in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 2020, pp. 782–797.

[2] Q. Zhang, V. Liu, H. Zeng, and A. Krishnamurthy, "High-resolution measurement of data center microbursts," in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 78–85.

[3] K. A. Clark, D. Cletheroe, T. Gerard, I. Haller, K. Jozwik, K. Shi, B. Thomsen, H. Williams, G. Zervas, H. Ballani *et al.*, "Synchronous subnanosecond clock and data recovery for optically switched data centres using clock phase caching," *Nature Electronics*, vol. 3, no. 7, pp. 426–433, 2020.

[4] L.-C. Cho, C. Lee, C.-C. Hung, and S.-I. Liu, "A 33.6-to-33.8 Gb/s burst-mode cdr in 90 nm cmos technology," *IEEE journal of solid-state circuits*, vol. 44, no. 3, pp. 775–783, 2009.

[5] A. Rylyakov, J. E. Proesel, S. Rylov, B. G. Lee, J. F. Bulzacchelli, A. Ardey, B. Parker, M. Beakes, C. W. Baks, C. L. Schow *et al.*, "A 25 Gb/s burst-mode receiver for low latency photonic switch networks," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 12, pp. 3120–3132, 2015.

[6] B. J. Shastri and D. V. Plant, "5/10-Gb/s burst-mode clock and data recovery based on semiblind oversampling for PONs: Theoretical and experimental," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 16, no. 5, pp. 1298–1320, 2010.

[7] ITU-T, "G.8261 Timing and synchronization aspects in packet networks," 2008, Accessed: 14-01-2021. [Online]. Available: {https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-G.8261-201908-I!!PDF-E&type=items}

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

[9] K. A. Clark, Y. Chen, E. R. N. Fokua, T. Bradley, F. Poletti, D. J. Richardson, P. Bayvel, R. Slavík, and Z. Liu, "Low thermal sensitivity hollow core fiber for optically-switched data centers," *Journal of Lightwave Technology*, vol. 38, no. 9, pp. 2703–2709, 2020.

[10] C. Mélange, B. Baekelandt, J. Bauwelinck, P. Ossieur, T. De Ridder, X.-Z. Qiu, and J. Vandewege, "Burst-mode CDR performance in long-reach high-split passive optical networks," *Journal of lightwave technology*, vol. 27, no. 17, pp. 3837–3844, 2009.

[11] H. Adrang and H. Miar-Naimi, "Modeling of jitter in bang-bang CDR with fourier series analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 1, pp. 3–10, 2012.

[12] G. P. Agrawal, *Fiber-Optic Communication Systems*, 4th ed. New Jersey, USA: John Wiley & Sons, 2010.

[13] A. Jeffrey, *Handbook of Mathematical Formulas and Integrals*. San Diego, CA, USA: Academic Press, Inc., 1995.

[14] S. Bottacchi, *Noise and Signal Interference in Optical Fiber Transmission Systems: An Optimum Design Approach*. Chichester: John Wiley & Sons, 2008.

[15] IEEE, "802.3-2018 IEEE Standard for Ethernet 802.3-2018," 2018, Accessed: 14-01-2021. [Online]. Available: https://standards.ieee.org/standard/802_3-2018.html

[16] Beijing Lightsending Technologies Ltd, "25G High Speed InGaAs PIN photodiode," p. 2, 2017, Accessed: 14-08-2021. [Online]. Available: http://www.lightsensing.com/upfile/2017-06/20170622853.pdf

[17] Texas Instruments, "ONET2804TLP Low-Power, 28-Gbps, 4-Channel Limiting TIA," pp. 1–29, 2017, Accessed: 14-08-2021. [Online]. Available: https://www.ti.com/lit/ds/sbas796/sbas796.pdf?ts=1595588247747

[18] R. Slavík, G. Marra, E. N. Fokoua, N. Baddela, N. V. Wheeler, M. Petrovich, F. Poletti, and D. J. Richardson, "Ultralow thermal sensitivity of phase and propagation delay in hollow core optical fibres," *Scientific reports*, vol. 5, no. 1, pp. 1–7, 2015.

[19] J. Lee, K. S. Kundert, and B. Razavi, "Analysis and modeling of bang-bang clock and data recovery circuits," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1571–1580, 2004.

[20] K. A. Clark, "Clock synchronisation assisted clock and data recovery for sub-nanosecond data centre optical switching," Ph.D. dissertation, UCL (University College London), 2021.

[21] R. S. Sohanpal, K. A. Clark, B. J. Puttnam, Y. Awaji, N. Wada, P. Bayvel, and Z. Liu, "Clock and data recovery-free data communications enabled by multi-core fiber with low thermal sensitivity of skew," *Journal of Lightwave Technology*, vol. 38, no. 7, pp. 1636–1643, 2020.

[22] Z. Hu, Z. Zhou, C. C.-K. Chan, and Z. Liu, "Equalizer state caching for fast data recovery in optically-switched data center networks," *Journal of Lightwave Technology*, vol. 39, no. 17, pp. 5362–5370, 2021.

[23] B. Buscaino, E. Chen, J. W. Stewart, T. Pham, and J. M. Kahn, "External vs. integrated light sources for intra-data center co-packaged optical interfaces," *Journal of Lightwave Technology*, vol. 39, no. 7, pp. 1984–1996, 2020.

[24] B. Buscaino, B. D. Taylor, and J. M. Kahn, "Multi-Tb/s-per-fiber coherent co-packaged optical interfaces for data center switches," *Journal of Lightwave Technology*, vol. 37, no. 13, pp. 3401–3412, 2019.

[25] M.-C. Lo, Z. Zhou, S. Pan, G. Carpintero, and Z. Liu, "Characterisation of thermal crosstalk-induced wavelength shift in monolithic InP dual DFB lasers PIC," in *Integrated Photonics Platforms: Fundamental Research, Manufacturing and Applications*, vol. 11364. International Society for Optics and Photonics, 2020, p. 113641U.

[26] Z. Zhou, K. Clark, C. Deakin, P. Laccotripes, and Z. Liu, "Clock synchronized transmission of 51.2 GBd optical packets for optically switched data center interconnects," in *2021 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE, 2021, pp. 1–3.

Kari A. Clark Kari A. Clark (S'18–M'21) received the B.Sc M.Phys degree in physics from the University of Warwick, Coventry, U.K., in 2013. He joined the Optical Networks Group at University College London (UCL) in 2014 to pursue the Ph.D. in optical communications. His doctoral research focused on using central optical clock synchronization to achieve sub-nanosecond clock and data recovery times in optically switched data center networks. In 2016, he completed a 9-month internship at Microsoft Research Cambridge, applying his research in an industrial context. In 2018, he was the overall winner of the EPSRC Connected Nation Pioneers competition, and in 2019, he won the Bronze in Engineering at STEM for BRITAIN 2019. In 2021, he was awarded the Ph.D. degree in optical communications, and he then joined the Optical Networks Group at University College London (UCL) as a Research Fellow where he continues to perform research focusing on clock synchronized optical networks Dr. Clark is a Member of OSA.

Zhixin Liu Zhixin Liu (M'12–SM'17) received the B.Eng. degree in information engineering and the B.B.A. degree in business administration from Tianjin University, Tianjin, China, in 2006, the M.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009, and the Ph.D. degree in information engineering from the Chinese University of Hong Kong, Hong Kong, in 2012. He joined the Optoelectronics Research Centre, University of Southampton, Southampton, U.K., in 2013. In 2016, he joined the Department of Electronics and Electrical Engineering at UCL. His research focuses on exploring analog and optical signal processing techniques for high performance communication systems, including high-speed direct modulation, frequency comb, photonic-assisted data conversion, and low-latency data communications. Dr. Liu is a Senior Member of OSA.