Clock-Synchronized Clock and Data Recovery to Enable Sub-Nanosecond Optically-Switched Networks

Zhixin Liu and Kari A. Clark
Optical Networks Group, Department of Electronic and Electrical Engineering,
University College London, London, UK WC1E 7JE
Author e-mail address: zhixin.liu@ucl.ac.uk

Abstract: We review the clock-synchronized approach to clock and data recovery, which enables sub-nanosecond switching time in optically switched networks, and explore the impact of factors such as temperature and jitter on performance and scalability. © 2021 The Authors

1. Introduction

Optical switching has attracted significant attention in recent research on data center networks (DCN) as it promises a viable route for the further scaling of hyperscale data center networks, 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 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 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 is extracted from the received optical signals. To achieve fast CDR, gated voltage-controlled oscillator (GVCO) CDRs have demonstrated CDR locking time in 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 synchronisation, 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 or White Rabbit techniques, through control plane fibers necessary for scheduling. With optical clock synchronisation, 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, which would reduce the power consumption of transceivers.

In this paper, we review clock phase caching, which enables sub-nanosecond CDR locking time for clock synchronized optically-switched DCNs. We also model the bit-error-ratio (BER) performance of a clock phase cached link and discuss the factors that affect its BER performance, size and distance scalability.

2. 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. 1a. Each of the N transmitters connects to N receivers through the optical switch, resulting in N^2 different Tx to Rx paths through the optical switch, one of which is shown in Fig. 1b. Each of these Tx to Rx paths normally has a unique 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, simplifying the CDR process. The clock frequency offsets are removed by frequency synchronizing all network nodes. The clock phase offsets are removed by measuring all phase offsets (at a slow rate of ≈1 to 10 Hz), 'caching' the phase offsets at the transmitter, 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].

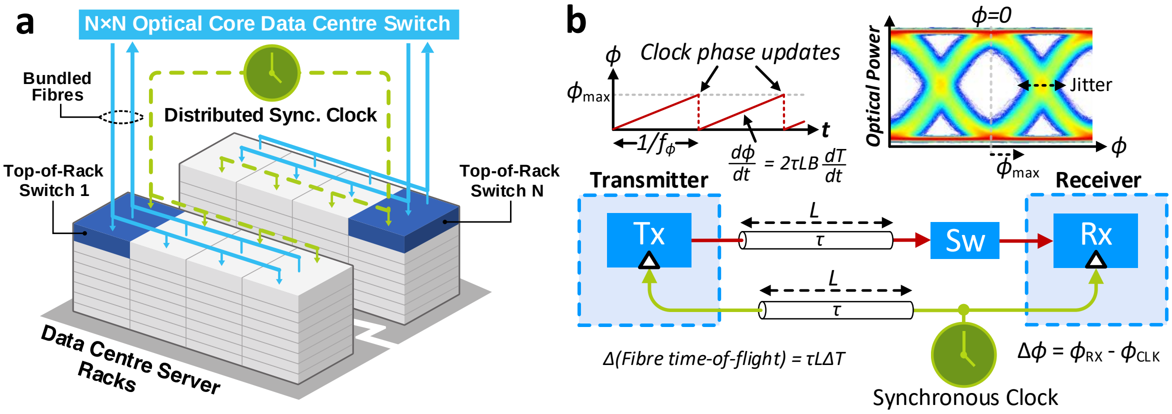

Fig. 1. a) Synchronized optical core data center switch interconnecting *N* top-of-rack switches; b) Illustration of clock phase variation in a single worst-case transmitter (Tx) to receiver (Rx) path through the optical switch (Sw) where the clock and data signals entirely counter propagate.

3. System performance and scalability

The system performance of a clock phase cached system depends on received signal quality as well as the parameters that influence the magnitude of the initial clock phase offset. Fig.1b shows the factors that affect a point-to-point clock phase cached link. When a packet initially arrives at a clock phase cached receiver, a small unremoved clock phase offset remains, which occurs due to fiber time-of-flight changes that have occurred since the last clock phase update, as well as due to clock jitter. This small clock phase offset may lead to a higher BER at the beginning of packet sampling, which results from sampling offset from the optimum sampling point. We define the time taken since the beginning of packet sampling for the BER to fall below some threshold (e.g. 10^{-10}) as CDR locking time. The worst-case BER in a clock phase cached system, $\mathrm{BER}_{\mathrm{max}}$, can be estimated from the worst-case clock phase offset, $\phi_{\mathrm{max.}}$ Assuming a 1st order Gaussian pulse shape and added white Gaussian noise, one can obtain:

$$\mathrm{BER}_{\mathrm{max}} = \frac{1}{2}\mathrm{erfc}\left(\frac{Q_0}{\sqrt{2}}\left(2e^{-\ln(2)(2\phi_{\mathrm{max}})^2} - 1\right)\right), \qquad \phi_{\mathrm{max}} = \frac{2\tau L}{f_\phi}\frac{dT}{dt}B$$

where Q_0 is the *Q*-factor at $\phi_{\mathrm{max}} = 0$, f_ϕ is the clock phase update rate, dT/dt is the rate of change of temperature, τ is the fiber thermal coefficient of delay (TCD, about 40 ps/km/°C for SMF28 [7]), L is the length of fiber and B is the symbol rate. The factor of 2 accounts for the worst-case where the clock and data signals entirely counter propagate. The worst-case magnitude of the initial clock phase offset must be small to minimize the impact of clock phase offset on BER (e.g. <0.1 symbols in [3]). To ensure this, the rate of clock phase updates must be sufficiently fast to compensate for changes in fiber time of flight that occur due to worst-case rates of temperature change in the data center environment. Each clock phase update introduces network overhead, which in-turn determines scalability. An estimate of the worst-case network overhead caused by clock phase caching, o_{max}, can be calculated from $o_{\mathrm{max}} = Nf_\phi(t_{\mathrm{meas}} + t_{\mathrm{update}})$, where N is the number of end-points (servers or switches) connected to the optical switch, 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 small worst-case throughput overhead of 1.7% was estimated to result from clock phase caching for 10,000 nodes in an optically switched DCN, assuming 2 km clock and 2 km data SMF-28 transmitting 25.6 Gb/s NRZ-OOK packet signals measured from our experiment and worst-case measured data center temperature change of 0.03° C/s [3]. The network overhead can be reduced by reducing the TCD, τ, by using low TCD fibers such as hollow core fiber (20 times lower than SMF28) [8]. It can also be potentially reduced by using homogenous multicore fibers (MCF), which features a low thermal coefficient of skew of about 40 fs/km/°C [9]. A comprehensive analytical model of system performance and scalability in a clock phase cached optical switch may be found in [10].

4. Discussion: other factors that affect the system performance

In addition to the fiber TCD and the temperature change rate in DCN, the performance of a clock phase cached link is also affected by the jitter of the transceivers and temperature-dependent optical impairments. The jitter (or clock phase noise) of the transceivers arises from two sources – a) the jitter of the distributed synchronized source clock and b) the transceiver PLL that locks to the optically-distributed clock.

- *Source clock jitter*

Generally, in comparison to asynchronous networks, clock synchronized systems offer lower transceiver jitter because high performance oscillators can be used to reduce phase noise in the low frequency region. By measuring the phase noise of an optically synchronized clock, we show that the low frequency region (from DC to the locking bandwidth of the transceiver PLL) of the phase noise is dominated by the source clock whilst the high frequency noise (e.g. 2 k – 10 MHz) is dominated by the output phase noise of the PLL [11]. Since 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 10 Hz. Therefore, the integrated phase noise from 10 Hz to about 10 MHz (the clock filter bandwidth) determines the overall jitter. As the baud rate increase from 25 GBd to 50 GBd and beyond, the source clock jitter must be minimized to achieve stable and low BER performance.

- *Temperature induced change of Optical impairments*

In addition to the change of fiber time-of-flight, 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 leads to a change of the optical impairments, 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) OOK signals, they become prominent as the DC interconnects evolve to higher baud rate (e.g. >50 GBd) and higher order modulation formats (e.g. PAM4) [12]. One possible approach to mitigate this impact is to disaggregate the light sources from the switches and use tuneable lasers and wavelength demultiplex/multiplex for switching [13]. However, the thermal cross talk between tuneable lasers on the same photonic circuit may cause unwanted wavelength shifts that degrade system performance [14]. In addition, multiplexer/multiplexers may perform poorly in an unpredictable thermal environment, causing degraded performance or link failure due to the thermal dependent wavelength drift [12, 13].

4. Conclusion

We review our recent results on fast CDR enabled by clock synchronization and clock phase caching for optically switched DCN. We analyze the factors that affect the BER performance of a clock phase caching based optical link and show that the proposed CDR scheme can be scaled to support 10,000 node DCN. 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.

The authors acknowledge the EPSRC TRANSNET (EP/R035342/1) and PhotoDAC (EP/R041792/1) projects, as well as Microsoft, for their funding and technical support.

References

[1] H. Ballani et al., "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching," in Proc. *ACM*, 782–797 (2020).
[2] Q. Zhang et al., "A. High-resolution measurement of data center microbursts," in Proc. Internet Measurement Conference (IMC) 78–85 (ACM, 2017).
[3] K. A. Clark et al., "Synchronous sub-nanosecond clock and data recovery for optically switched data centres using clock phase caching," *Nat. Electron*., **3,** 426–433 (2020).
[4] L.C. Cho et al., "A 33.6-to-33.8 Gb/s Burst-Mode CDR in 90 nm CMOS Technology," IEEE J. Solid-State Circuits, **44**, 775-783 (2009).
[5] A. Rylyakov et al., "A 25 Gb/s burst-mode receiver for low latency photonic switch networks," IEEE J. Solid-State Circuits, **50**, 3120-3132 (2015).
[6] B. J. Shastri, D. V. Plant, "5/10-Gb/s Burst-Mode Clock and Data Recovery Based on Semiblind Oversampling for PONs: Theoretical and Experimental," IEEE J. Sel. Top. Quantum Electron., **16**, 1298-1320 (2010).
[7] R. Slavik et al., "Ultralow thermal sensitivity of phase and propagation delay in hollow core optical fibres," *Sci. Rep*. **5,** 1-7 (2015).
[8] K. A. Clark et al., "Low Thermal Sensitivity Hollow Core Fiber for Optically-Switched Data Centers," *IEEE/OSA J. Lightwave Technol*. **38,** 2703-2709 (2020)
[9] R. S. Sohanpal et al., "Clock and Data Recovery-Free Data Communications Enabled by Multi-Core Fiber With Low Thermal Sensitivity of Skew," *J. Lightwave Technol*. **38**, 1636-1643 (2020).
[10] K. A. Clark, "Clock Synchronisation Assisted Clock and Data Recovery for Sub-Nanosecond Data Centre Optical Switching", PhD Thesis, University College London (2020).
[11] 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," submitted to *Optical Fiber Communication Conference* (2021).
[12] Z. Hu, Z. Zhou, C.K. Chan, and Z. Liu, "Equalizer Status Caching for Fast Data Recovery in Optically-Switched Data Center Networks," submitted to *J. Lightwave Technol.* (2021).
[13] B. Buscaino, B. D. Taylor, and J. M. Kahn, " Multi-Tb/s-per-Fiber Coherent Co-Packaged Optical Interfaces for Data Center Switches," *IEEE/OSA J. Lightw. Technol*., **37**, 2703- 2709 (2019)
[14] M. 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, p. 113641U (2020).