MD4.3
16.30 – 16.45

Burst-Mode Clock and Data Recovery for Optically Interconnected Data Centers

Bhavin J. Shastri, and David V. Plant
Photonic Systems Group, Dept. of Electrical and Computer Eng., McGill University, Montreal, QC H3A 2A7, Canada
shastri@ieee.org

***Abstract*—We propose a novel burst-mode clock/data recovery (BM-CDR) architecture for optical data center applications. Our design is based on a hybrid topology of a CDR (feedback) and clock phase aligner (feed-forward) utilizing multi-phase clocks.**

I. Introduction

Data centers or large clusters of servers are currently being aggressively deployed in a number of institutions to harness petaflops of computational power and petabytes of storage in a cost-efficient manner [1]. Consequently, there exists a world-wide research interest in designing such large data centers for optimally supporting various applications including scientific computing, financial analysis, data analysis and warehousing, and large-scale network services.

Data centers in general follow a tiered architecture in which network devices (switches or routers) are organized into two or three layers. The highest layer—core tier—is at the root of the tree, whereas the lowest layer—edge tier—is at the leaves of the tree. Between these layers, an aggregation tier may exist when the number of devices is large. The need for highly-specialized ASICs is undeniable [2], with clock and data recovery (CDR) being a critical function in backplane routing and chip-to-chip interconnects. The data received on the aggregation and edge node links is inherently bursty [3] with asynchronous phase steps $|\Delta\varphi| \leq 2\pi$ rad, that exist between the consecutive k^{th} and $(k+1)^{\text{th}}$ packet. This inevitably causes conventional CDR circuits to lose pattern synchronization leading to packet loss. Preamble bits can be inserted at the beginning of each packet to allow the CDR feedback loop enough time to settle down and thus acquire lock. However, the use of a preamble introduces overhead, reducing the effective throughput and increasing delay. Consequently, to deal with bursty data, these nodes require a burst-mode CDR (BM-CDR). The most important characteristic of the BM-CDR is its phase acquisition time which must be as short as possible. In this paper, we present a novel BM-CDR architecture based on a hybrid topology; that is, a combination of feedback and feed-forward.

II. Novel BM-CDR Architecture

A block diagram of the proposed BM-CDR is shown in Fig. 1. The BM-CDR is composed of a phase-tracking CDR and a clock phase aligner (CPA). The CDR senses data D_{in}, and generates a synchronized clock CK, with a voltage-controlled oscillator (VCO) in a phase-locked (feedback) loop (PLL). The phase and frequency of CK is compared to D_{in} in

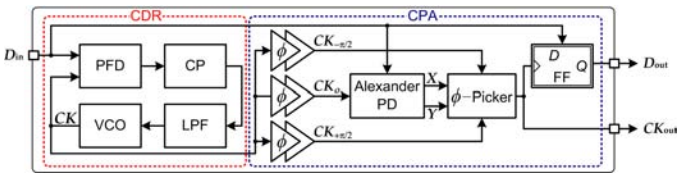

Fig. 1. BM-CDR architecture.

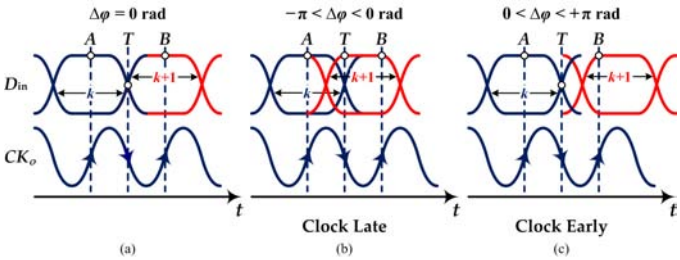

Fig. 2. (a) Three-point sampling scheme; (b) and (c) early-late waveforms.

the phase/frequency detector (PFD), generating an error signal that is passed through the charge pump (CP) and the low-pass filter (LPF) to set the voltage required by the VCO to oscillate at the frequency of interest.

Burst-mode functionality is obtained with the CPA which utilizes multi-phase clocks and a phase picking algorithm based on an "early-late" detection principle. This CPA is based on a feed-forward topology, and comprises of phase (ϕ-) shifters, an Alexander PD, a ϕ-picker, and a D flip-flop (D-FF). The ϕ-shifters utilize the clock recovered by the CDR CK, to provide multiple clocks: CK_o, $CK_{-\pi/2}$, and $CK_{+\pi/2}$, with low skew and different phases: 0 rad, $-\pi/2$ rad, and $+\pi/2$ rad, respectively, with respect to CK. Next, an Alexander PD [4] which inherently exhibits *bang-bang* (binary) characteristics is used to strobe the data waveform D_{in}, with consecutive clock CK_o edges, at multiple points in the vicinity of expected transitions [see Fig. 2(a)], resulting in three data samples: previous bit A, current bit B, and a sample of the current bit at the zero crossing T. Depending on the phase difference between the consecutive packets, the PD aided by these samples, $X \equiv T \oplus B$ and $Y \equiv A \oplus T$, can determine the location of the clock edge with respect to the data edge as follows: (a) if $A \neq T = B$ ($X \downarrow$, $Y \uparrow$) $\Rightarrow$ CK_o lags D_{in}—is late—when $-\pi < \Delta\varphi < 0$ rad [see Fig. 2(b)]; (b) if $A = T \neq B$ ($X \uparrow$, $Y \downarrow$) $\Rightarrow$ CK_o leads D_{in}—is early—when $0 < \Delta\varphi < +\pi$ rad [see Fig. 2(c)]; (c) if $A = T = B$ ($X \downarrow$, $Y \downarrow$) $\Rightarrow$ no data transition is present due to consecutive identical digits (CIDs); and (d) if $A = B \neq T$ ($X \uparrow$, $Y \uparrow$) $\Rightarrow$ no decision is possible. The clock

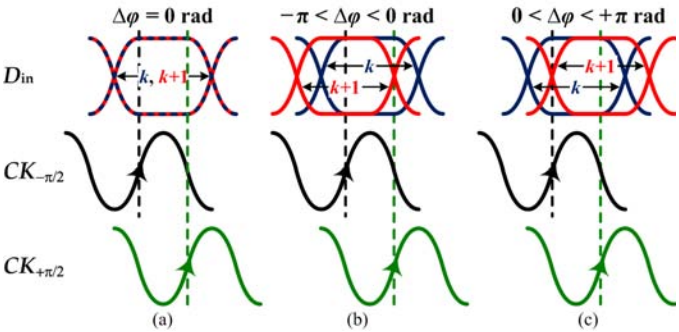

Fig. 3. CPA phase picking algorithm.

CK_o early-late information (X and Y) together with the two multi-phase clocks, $CK_{-\pi/2}$ and $CK_{+\pi/2}$, is provided to the ϕ-picker. The idea then behind the phase picking algorithm is depicted with the aid of eye diagrams in Fig. 3. When there is no phase difference between the consecutive packets, $\Delta\varphi = 0$ rad, either of the clocks, $CK_{-\pi/2}$ and $CK_{+\pi/2}$, will correctly sample the data bits of the phase shifted $(k+1)^{\text{th}}$ packet [see Fig. 3(a)]. This is also true for an antiphase step $\Delta\varphi = \pm\pi$ rad—not shown as this is a modulo-π process. For a phase step $-\pi < \Delta\varphi < 0$ rad, clock $CK_{+\pi/2}$ will sample the bits on or close to the transitions of the data eye, whereas clock $CK_{-\pi/2}$ will correctly sample the data [see Fig. 3(b)]. Similarly for a phase step $0 < \Delta\varphi < +\pi$ rad, clock $CK_{-\pi/2}$ will sample the bits on or close to the transitions, whereas clock $CK_{+\pi/2}$ will correctly sample the data [see Fig. 3(c)]. That is, regardless of any phase step, there will be at least one clock, either $CK_{-\pi/2}$ or $CK_{+\pi/2}$, that will yield an accurate sample. The ϕ-picker then selects the most accurate clock CK_{out}, from these two possibilities for driving the D-FF to retime the data; that is, sample the noisy data, yielding an output D_{out} with less jitter. The foregoing concepts on the Alexander PD and the ϕ-picker are summarized in Table I, leading to the circuit topology in Fig. 4.

III. Hardware Implementation

The BM-CDR is being implemented for operation at 10 Gb/s. The main building blocks include a CDR from Centellax (Part #TR1C1-A) and a CPA built by integrating individual chips from Hittite Microwave on a custom designed printed circuit board (PCB). More specifically, the PCB is populated with three 4-bit digital ϕ-shifters (Part #HMC543), an Alexander PD comprised of four D-FFs (Part #HMC673LC3C) and two XOR gates (Part #HMC671LC3C), and a ϕ-picker comprised of an AND gate (Part #HMC672LC3C) and a 2:1 selector (Part #HMC748LC3C).

IV. Results

Fig. 5 shows the simulated BER performance of the CDR and BM-CDR as a function of the phase step between two consecutive packets, for a zero preamble length. As expected the worst-case phase steps for the CDR are $\pm\pi$ rad because these represent the half-bit periods, and therefore the CDR is sampling exactly at the edge of the data eye, resulting in a BER ~ 0.5. At relatively small phase shifts (near 0 or 2π rad), we can easily achieve error-free operation, BER $< 10^{-10}$, because the CDR is almost sampling at the middle of each data bit. For the proposed BM-CDR we achieve error-free operation for any phase step $|\Delta\varphi| \leq 2\pi$ rad, allowing for instantaneous phase acquisition.

TABLE I
CPA (Alexander PD and ϕ-Picker) Logic

Data Condition	CK_o	XY	CK_{out}
$-\pi < \Delta\varphi < 0$ rad	Late	↓↑	$CK_{-\pi/2}$
$0 < \Delta\varphi < \pi$ rad	Early	↑↓	$CK_{+\pi/2}$
$\Delta\varphi \in \{0, \pm\pi$ rad$\}$ or CIDs	×	↑↑, ↓↓	$CK_{-\pi/2,+\pi/2}$

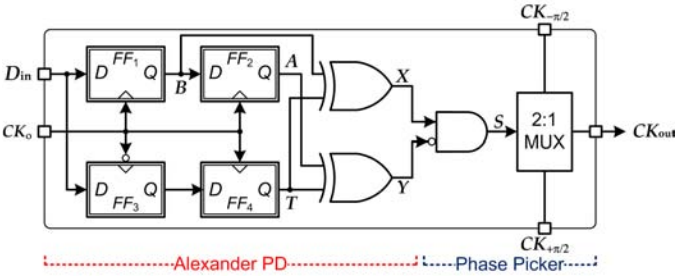

Fig. 4. Hardware implementation of Alexander PD and ϕ-picker.

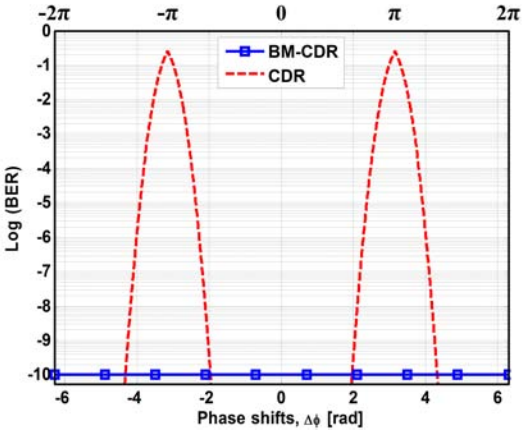

Fig. 5. BER performance of the CDR and BM-CDR (for zero preamble length) versus phase step.

V. Conclusion

Recently, BM-CDRs that achieve instantaneous phase acquisition have been demonstrated for multi-access networks [5], [6]. These BM-CDRs are based on time oversampling techniques requiring electronics operating at twice or thrice the aggregate bit rate resulting in wasted power. They also have the knowledge of a predefined unique delimiter (start of packet) that they exploit as a signature for the phase picking algorithm. In contrast, our work based on space sampling, uses electronics operated *at* the bit rate with *no a priori* knowledge of the delimiter, leading to more efficient power consumption and being truly modular across application testbeds, respectively.

References

[1] M. Al-Fares, et. al, *SIGCOMM Comput. Commun. Rev.*, **38**(4), 2008.
[2] N. Farrington, et. al, in *Proc. IEEE Symp. on High Performance Interconnects*, pp. 93–102, 2009.
[3] T. Benson, et. al, *SIGCOMM Comput. Commun. Rev.*, **40**(1), 2010.
[4] J. D. H. Alexander, *Electronic Lett.*, **11**(22), 1975.
[5] B. J. Shastri, et. al, *IEEE J. Sel. Topics Quantum Electron.*, **16**(5), 2010, to appear.
[6] B. J. Shastri, et. al, *J. Opt. Commun. and Netw.*, **2**(1), 2010.