Clock and Data Recovery over Optical Links and Networks

A MEng Project Final Report

Ammar Bin Shaqeel Ahmed
ammar.ahmed.16@ucl.ac.uk
16080322

University College London

Supervisor:
Dr. Georgios Zervas

Secondary Assessor:
Dr. Domaniç Lavery

May, 2020

Contents

Chapter 1

Introduction

Bandwidth demands in data centers have been doubling every 12-15 months. For data center providers to keep pace with the increased demand (at the same price point) network switches have had to double their capacity while staying at roughly the same cost [1]. However this trend seems to be coming to an end for two reasons. The first is a predicted increase in the rate of growth of demand, due to trends like hardware accelerated programming and dis-aggregated workloads. The second is because electrical switches are predicted to reach a limit due to the physical limits on the pin density of ball grid arrays (BGA) [2].

For these reasons optical switching is being explored, as it has the potential to overcome many of these problems. Optical switches do not require opto-electrical (OEO) conversion, and hence the number of expensive and power hungry transceivers required is reduced. Furthermore, as buffering is not needed, the latency of the optical switches is much lower. Lastly, they do not use electronics for switching, thus bypassing the aforementioned physical limit [2].

In data centers much of the traffic that is transmitted between servers is in the form of small data packets, with 97.8% of packets being 576 bytes or less [3]. With 100 Gb/s ports this means that switching should take place on the order of hundreds of nanoseconds.

When data is transmitted without a clock signal, the clock has to be regenerated at the receiver before the data can be decoded - this is known as clock and data recovery (CDR). This introduces latency, and requires the use of training sequences and other techniques to ensure that the CDR circuit can lock correctly. In optical switches physical links are created between each transceiver-receiver pair, hence each time the switch is reconfigured, the CDR must re-lock to the new link. This means that the network throughput is limited by the sum of the optical switching time and the CDR locking time. The CDR locking time can be hundreds of nanoseconds in the worst case and tens of nanoseconds in the best case [4]. Assuming an optical switching time of 1 nanosecond, it is evident that the CDR locking time acts as bottleneck that can drastically reduce the throughput [5].

In a source synchronous system the clock is transmitted alongside the data, removing the CDR locking time. This would remove the bottleneck, theoretically increasing the throughput.

Chapter 2

Theoretical Basis

2.1 Background Theory

Clock and Data Recovery

Commonly a serial data stream is sent over a channel without a clock signal. Clock recovery is the process of extracting timing information from a serial data stream. This timing data is then used to decode the received data stream. This process is known as Clock and Data Recovery (CDR).

In this application a bang-bang CDR circuit is used. In this case transitions in a received data signal are counted as "early" or "late" as compared with a local clock. The clock can then be adjusted based on the local transitions[6].

Source Synchronous System

In a source synchronous system a clock signal is provided alongside the data signal. This allows us.

2.2 Literature Review

[5] outlines how CDR circuits are a limiting factor in optical switching and proposes a method of overcoming this (phase caching). Through phase caching they were able to demonstrate sub nanosecond locking times, improving the locking time 12x.

In [7], [8], and [9], the white rabbit project is discussed. A white rabbit system provides sub-nanosecond synchronisation accuracy. To achieve this, accurate measurements of the link delay between the nodes of the network must be calculated. While instructive, the method may not be directly applicable to the project, as in a White Rabbit system, all the nodes are locked to the same frequency. Hence the link delay can be calculated by having a node receive a clock signal from another node, then return the same signal. The link delay can then be calculated by comparing the phase offset of the two signals.

[10] described an optical source synchronous system. It describes how choosing the correct wavelength for the clock can minimise the modal cross-talk. Furthermore, in conjunction with [11] it was quite instructive in describing how source synchronous systems are able to track correlated jitter between clock and data channels, and how system performance can be degraded by channel slew between clock and data channels.

[12] further explored reducing the modal crosstalk by proposing an architecture with re-configurable clock and data paths, thus allowing the user to chose the optimal lane for the sensitive clock for each photonic interconnect. This may not be needed however, as each transmitter should have a fixed data characteristic.

[13] and [14] describe fixed latency links. There is a possibility that we will be unable to completely bypass the CDR. In this case if we can organise the system to have fixed latency, we could force the CDR to the appropriate fixed phase. Thus the circuit would then be able to have a stable source-synchronous link.

[15], [16] describe an intellectual property (IP) that allows the high speed serial transceivers to be used at much lower data rates. This was initially of interest because it would have been easier to demonstrate a working system with lower data rates. However as this is an extra IP used in conjunction with the transceivers it may not be useful for the project.

[17] this presentation describes a system where the phase of a transceiver on Xilinx board is kept stable over resets. While this was done primarily on the transmitter side (we are interested in the receiver side) this still shows that fixing the phase of the transceiver is possible in some cases.

Chapter 3

System Overview and Objectives

To demonstrate the efficacy of a source synchronous system we use a single receiver that receives data from two different channels. The two channels would transmit a pseudorandom binary sequence (PRBS), from a single source. At set intervals the source alternates the channel over which it transmits. Hence the overall effect is that the channels would optically transmit bursts of data at non-overlapping intervals. If the receiver is successfully able to receive the full sequence, then the source synchronous system would be working correctly. An overview of the system is shown in Figure 3.1.

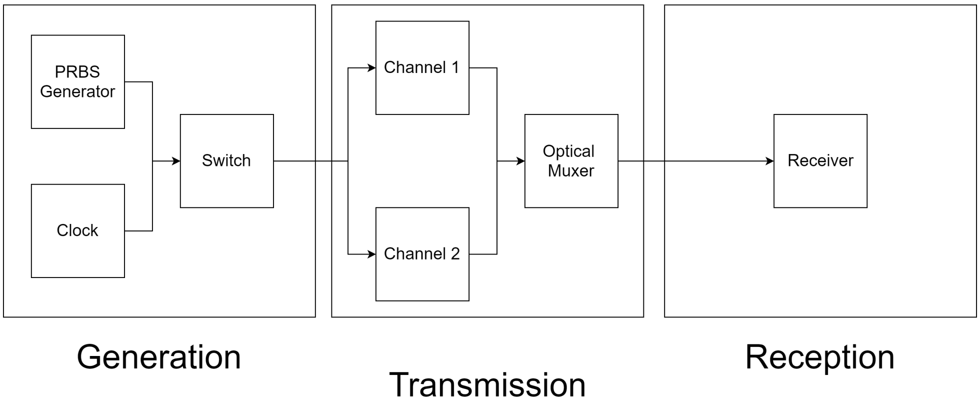

Figure 3.1: Overview of System

The overall objective is to demonstrate successful burst source synchronous communication. If successful, we would then be able to compare the throughput of the system with a similar system that utilised a CDR circuit.

To accomplish this the following components are needed:

- A burst mode PRBS generator
- An optical switch
- Source synchronous PRBS checker

Chapter 4

Implementation and Results

In this section we cover the implementation of the project and the results. The hardware used was the Virtex UltraScale+ FPGA VCU118 Board.

Figure 4.1: VCU118 Board

The project used the high-speed parallel to serial GTY transceiver built into the VCU118 board. We looked to modify the functionality of a basic implementation of the transceiver.

The basic implementation is as follows: There is a PRBS generator which feeds wordbits to the transceiver. The transceiver receives these bits, serializes them, then transmits them

Chapter 5

Conclusion

Bibliography

[1] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano *et al.*, "Jupiter rising: a decade of clos topologies and centralized control in google's datacenter network," *Communications of the ACM*, vol. 59, no. 9, pp. 88–97, 2016.

[2] H. Ballani, P. Costa, I. Haller, K. Jozwik, K. Shi, B. Thomsen, and H. Williams, "Bridging the last mile for optical switching in data centers," in *2018 Optical Fiber Communications Conference and Exposition (OFC).* IEEE, 2018, pp. 1–3.

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

[4] X. Chen, S. Chandrasekhar, G. Raybon, S. Olsson, J. Cho, A. Adamiecki, and P. Winzer, "Generation and intradyne detection of single-wavelength 1.61-tb/s using an all-electronic digital band interleaved transmitter," in *Optical Fiber Communication Conference Postdeadline Papers.* Optical Society of America, 2018, p. Th4C.1. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2018-Th4C.1

[5] K. Clark, H. Ballani, P. Bayvel, D. Cletheroe, T. Gerard, I. Haller, K. Jozwik, K. Shi, B. Thomsen, P. Watts *et al.*, "Sub-nanosecond clock and data recovery in an optically-switched data centre network," in *2018 European Conference on Optical Communication (ECOC).* IEEE, 2018, pp. 1–3.

[6] J. Alexander, "Clock recovery from random binary signals," *Electronics letters*, vol. 11, no. 22, pp. 541–542, 1975.

[7] J. Serrano, M. Lipinski, T. Wlostowski, E. Gousiou, E. van der Bij, M. Cattin, and G. Daniluk, "The white rabbit project," 2013.

[8] P. Moreira, P. Alvarez, J. Serrano, I. Darwezeh, and T. Wlostowski, "Digital dual mixer time difference for sub-nanosecond time synchronization in ethernet," in *2010 IEEE International Frequency Control Symposium.* IEEE, 2010, pp. 449–453.

[9] P. Moreira, J. Serrano, T. Wlostowski, P. Loschmidt, and G. Gaderer, "White rabbit: Sub-nanosecond timing distribution over ethernet," in *2009 International Symposium on Precision Clock Synchronization for Measurement, Control and Communication.* IEEE, 2009, pp. 1–5.

[10] C. Williams, B. Banan, G. Cowan, and O. Liboiron-Ladouceur, "A source-synchronous architecture using mode-division multiplexing for on-chip silicon photonic interconnects," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 22, no. 6, pp. 473–481, 2016.

[11] A. Ragab, Y. Liu, K. Hu, P. Chiang, and S. Palermo, "Receiver jitter tracking characteristics in high-speed source synchronous links," *Journal of Electrical and Computer Engineering*, vol. 2011, p. 5, 2011.

[12] C. Williams, D. Abdelrahman, X. Jia, A. I. Abbas, O. Liboiron-Ladouceur, and G. E. Cowan, "Reconfiguration in source-synchronous receivers for short-reach parallel optical links," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2019.

[13] K. Chen, H. Chen, W. Wu, H. Xu, and L. Yao, "Optimization on fixed low latency implementation of the gbt core in fpga," *Journal of Instrumentation*, vol. 12, no. 07, p. P07011, 2017.

[14] X. Liu, Q.-x. Deng, B.-n. Hou, and Z.-k. Wang, "High-speed, fixed-latency serial links with xilinx fpgas," *Journal of Zhejiang University SCIENCE C*, vol. 15, no. 2, pp. 153–160, Feb 2014. [Online]. Available: https://doi.org/10.1631/jzus.C1300249

[15] P. Novellini and G. Guasti, *Dynamically Programmable DRU for High-Speed Serial I/O*, Xilinx.

[16] P. Novellini, G. Guasti, and A. Di Fresco, *Clock and Data Recovery Unit based on Deserialized Oversampled Data*, Xilinx.

[17] Eduardo Mendes, "Xilinx transceiver study," URL: https://indico.cern.ch/event/717613/contributions/2948664/attachments/1637690/2613637/hptd_fixed_phase_xcvr_04_18_eduardo_mendes.pdf, 2018.