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

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc blandit tellus id lectus egestas, sit amet finibus diam eleifend. Nulla nulla felis, hendrerit vel mi ac, tempus varius libero. Proin lobortis sapien eget malesuada aliquet. Nulla ac eleifend velit. Suspendisse feugiat magna convallis, consectetur leo a, ultrices ex. Aliquam suscipit mi eu tempor porttitor. In facilisis eget massa ac iaculis. Sed dolor lacus, scelerisque in diam vitae, finibus aliquam lorem. Donec eu leo eu lorem sodales varius. Donec dictum mauris sed ornare porttitor. Nunc imperdiet eu elit non hendrerit. Sed velit lectus, lacinia sed dolor vitae, faucibus eleifend felis. Aliquam ac ex enim. Aliquam erat volutpat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque pulvinar sollicitudin mattis.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc blandit tellus id lectus egestas, sit amet finibus diam eleifend. Nulla nulla felis, hendrerit vel mi ac, tempus varius libero. Proin lobortis sapien eget malesuada aliquet. Nulla ac eleifend velit. Suspendisse feugiat magna convallis, consectetur leo a, ultrices ex. Aliquam suscipit mi eu tempor porttitor. In facilisis eget massa ac iaculis. Sed dolor lacus, scelerisque in diam vitae, finibus aliquam lorem. Donec eu leo eu lorem sodales varius. Donec dictum mauris sed ornare porttitor. Nunc imperdiet eu elit non hendrerit. Sed velit lectus, lacinia sed dolor vitae, faucibus eleifend felis. Aliquam ac ex enim. Aliquam erat volutpat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque pulvinar sollicitudin mattis.

Duis vitae malesuada tellus, malesuada tincidunt sem. Nam dapibus mattis lorem, sagittis egestas risus ullamcorper a. Etiam finibus, libero non hendrerit scelerisque, nisl nisi blandit lacus, id laoreet tortor turpis sit amet sapien. Nullam non lacus ut tortor consequat auctor. Vivamus congue, massa in suscipit imperdiet, dui neque ultricies purus, ut fermentum sapien ante vel urna. Sed at blandit ligula. Sed non enim ante. Integer volutpat vestibulum leo, eu malesuada diam lacinia vel. Nulla ultrices leo eu cursus feugiat. Morbi semper nisl ex. Etiam volutpat consectetur sodales. Vestibulum facilisis dapibus enim, ac rutrum massa blandit vehicula. Aenean nec lectus varius, congue risus at, volutpat purus. Duis id rhoncus felis.

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 pin density [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). The time taken for the local clock to "lock" to the data stream, adds latency. 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 - which 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

Here we go deeper into the theory of certain elements of the system.

Bang-Bang CDR

Commonly a serial data stream is sent over a channel without a clock signal. Clock and Data Recovery (CDR) is the process of extracting timing information from a serial data stream, then using it to decode the received data stream. A CDR circuit has two primary functions. The first is to extract a clock based on the input data, and the second is to resample the data.

To extract the clock from the data, a local clock is generated, then is adjusted as "early" or "late" when compared with the incoming data signal [6]. We can think of this as a control system, as shown in Figure 2.1.

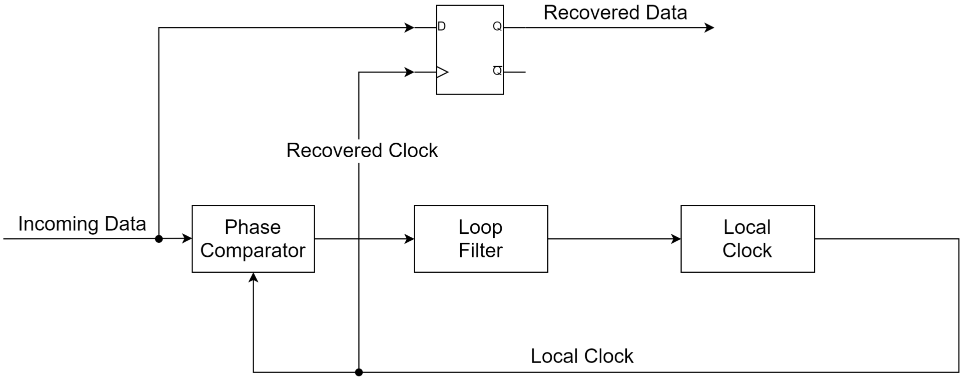

Figure 2.1: Basic CDR design

Phase detectors can be divided into two types, linear (where the output has a linear relationship to the input) and binary or bang-bang phase detectors (where the output is either positive or negative). Binary phase detectors are more commonly used in digital CDR circuits [7]. An example of one is the Alexander detector [8] which gives out a high D0+ and a low D0- if the clock lags and vice-versa if the clock leads, as shown in Figure 2.2.

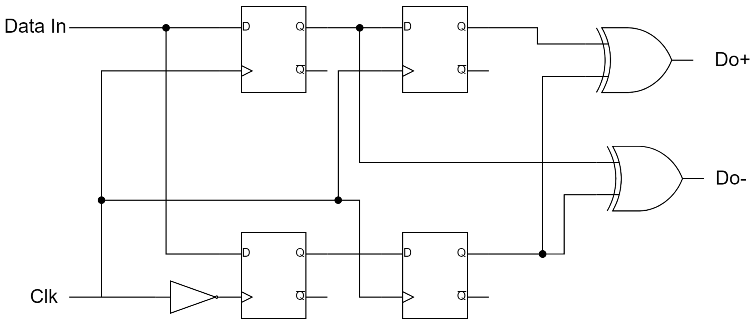

Figure 2.2: Alexander Phase Detector

Pseudorandom Binary Sequence

A pseudorandom binary sequence (PRBS) is a sequence of bits that appears to be random. However as it is generated using a deterministic algorithim, it can be replicated if the inital conditions are the same.

A common practical implementation of PRBS generation uses linear-feedback shift registers. As an example, a PRBS-4 sequence could be generated by using a 4 bit register. We seed the register with a non-zero number, then tap two bits of the register as an input. We then shift the contents of the register, taking the last bit as an output and the new bit as an input, as illustrated in Figure 2.3.

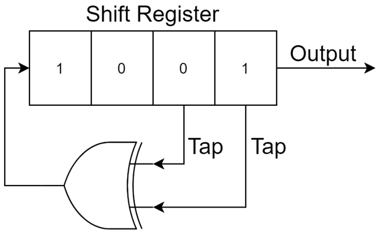

Figure 2.3: Shift Register Implementation

The full operation can be seen in Table 2.1. As 0000 cannot appear (the value of the register would never change) we see that for a register of size N, the bitsequence is $2^N - 1$ bits long.

Source Synchronous System

In a source synchronous system a clock signal is provided alongside the data signal, as shown in Figure 2.4. This has the advantage of not needing a CDR circuit. Furthermore as both the clock and the data come from the same device any jitter will be similar across both signals and can likely be ignored [9]. A downside is that there will be crossing of clock domains at the reciever as the transmitted clock will not be synchronous with the clock domain of the receiving device.

Cycle	Input	Shift Register				Output
0	1	1	0	0	1	1
1	0	1	1	0	0	0
2	1	0	1	1	0	0
3	0	1	0	1	1	1
4	1	0	1	0	1	1
5	1	1	0	1	0	0
6	1	1	1	0	1	1
7	1	1	1	1	0	0
8	0	1	1	1	1	1
9	0	0	1	1	1	1
10	0	0	0	1	1	1
11	1	0	0	0	1	1
12	0	1	0	0	0	0
13	0	0	1	0	0	0
14		0	0	1	0	0

Table 2.1: Shift Register Operation

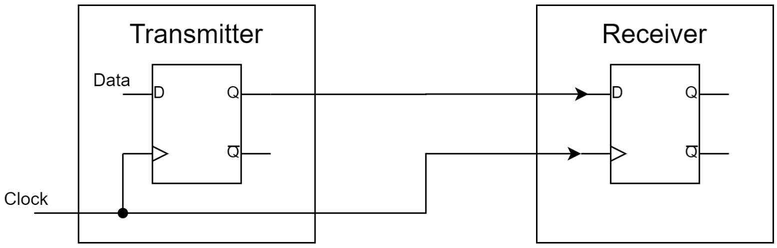

Figure 2.4: Source Synchronous System

2.2 Literature Review

[5] outlines how CDR circuits are a limiting factor in optical switching and proposes a method of phase caching to overcome this. Here the phase of the local clock in relation to the data is cached. The PRBS data is pregenerated (written to memory) and is sent in short bursts with a known sequence at the end. When the data arrives it is then written to memory and then processed. The phase caching improved locking time on switching by 12 times.

In [10], [11], and [12], 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 is not 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.

[13] 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 [9] it describes 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.
[14] 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.

[15] and [16] describe fixed latency links. In the event we were unable to bypass the CDR, it may be possible to organise the system to have a fixed latency, then force the CDR to the appropriate fixed phase. Thus the circuit could thus have a much reduced CDR lock time.

[17], [18] describe an Xilinx intellectual property 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 did not turn out to be useful for the project.

[19] this presentation describes a system where the phase of a transceiver on Xilinx board is kept stable over resets. While this was done on the transmitter side it shows that fixing the phase of the transceiver is possible.

Chapter 3

Proposed System and Objectives

3.1 Proposed System Overview

To demonstrate the efficacy of a source synchronous system we propose a single pseudorandom binary sequence (PRBS) source that optically transmits over two channels to a single reciever. If transmission is alternated the effect is that the reciever would recieve bursts of data from two different channels. If the full PRBS sequence is recieved then the system would be working correctly.
An overview of the system is shown in Figure 3.1.

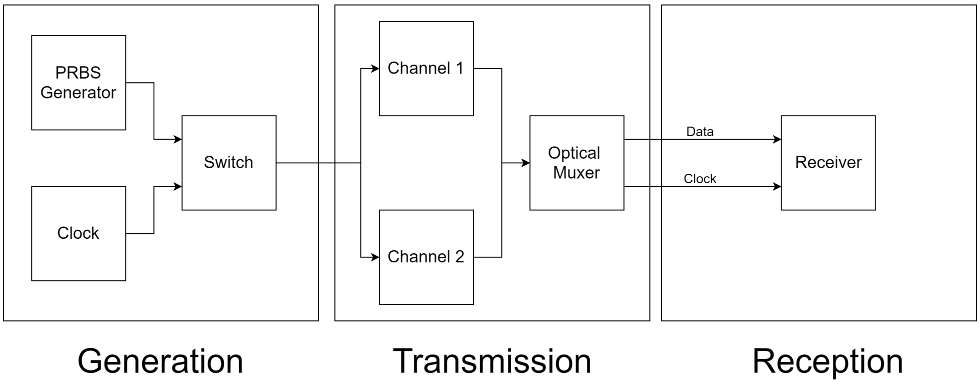

Figure 3.1: Overview of System

3.2 Objectives

The overall objective is to demonstrate successful burst source-synchronous communication for comparison with a system that uses a CDR. Overall we can break down the project to the following sub-objectives:

- Burst mode PRBS transmission over two channels alongside clock
- Transmit data optically and mux the two channels together
- Source synchronous reception of PRBS data

Chapter 4

Implementation and Results

In this section we cover the implementation of the project and the results. As outlined in the Objectives section we can divide the tasks into three main parts: generation, transmission, and reception. In this project we looked at using a FPGA board for the generation and reception of the PRBS data. Hence the overall design is of a board in a loopback configuration as shown in Figure 4.1.

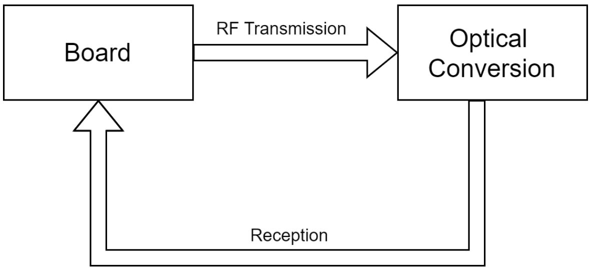

Figure 4.1: Loopback Configuration

4.1 Generation and Reception

4.1.1 Hardware

To generate and receive PRBS data the VCU118 board was used. The transmission and reception of the data was handled by the onboard high-speed parallel to serial GTY transceivers in conjuction with a Si5345 external clock (as the board is not able to generate an internal clock to the needed precision). The full details of the setup can be found in the appendix.

Figure 4.2: VCU118 Board

Figure 4.3: Si5345 Clock

4.1.2 PRBS Generation

We looked to modify the functionality of a basic implementation of the transceiver. In the basic implementation a PRBS generator is is fed to the transceiver channel, through a wrapper.

Todo: image of prbsgen wrapped and passed to transceiver

The PRBS module was unchanged from the default with the exception of reduced the length of the PRBS sequence from PRBS31 (2.1 billion bits) to PRBS7 (511 bits) for ease of checking. There were two variations of the PRBS generation module that were developed.

Burst Mode over Single Channel

Here we modified the PRBS generation module further and set it to output zeros if the output flag was disabled. In combination we placed a 2 bit register inside the wrapper, which on overflow toggles the enable flag. This has the effect of causing the PRBS module to output a sequence interspersed with zeros.

Todo: burst mode modification

Switching Between Two Channels

The main modification was to change the PRBS wrapper to feed the two different outputs. The PRBS generator module was unchanged. Using a 2 bit register which on oveflow alternated between which of the outputs the PRBS data was sent to, with the other output being sent zeros. This had the overall effect of having the whole sequence be sent over two different channels.

Todo: two channel switch (showing module with two outputs)

4.1.3 PRBS Checking

In normal operation the transceiver would parallise the serial data, and then pass the data to the PRBS checking module.

Todo: image of transceiver -> prbs module

Normal Checking

In the case where two transmitters were muxed together and were sent to a single receiver, it should not have been necessary to change the behaviour of the PRBS checking module. However we were unable to check as the optical transmission was not implemented.

Burst Mode Checking

For burst mode checking there are some issues as there are periods when the incoming bitstream is all zeros. The PRBS checker module takes the incoming data as a seed to calculate the next expected sequence. If zeros are provided then this interferes with module (as the next expected word will be calculated based on zeros). To compensate this we added a register to the wrapper that would not pass zeros to the checker module. In a basic simulation of PRBS generation to PRBS checker this worked correctly, but when passed to through the transceiver, the checker module would throw errors. We were not able to determine why.

Source Synchronous Reception

The final step would have been to run the reciever source synchronously. The transceiver did not allow much flexibilty here. We attemped to do this was by disabling the CDR and using the same clock to drive the reciever and transmitter. However the link here was not stable, and we were unable to get phase readouts that may have allowed us to modify the phase of the incoming clock. Overall this was unsuccessful.

4.2 Optical Transmission

This part of the project was not completed as the labs were closed before we were able to test it. However some hardware was prepared, and is described in the following sections.

4.2.1 SOA Board

4.2.2 Heatsink and Mount

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] S. Y. Sun, "An analog pll-based clock and data recovery circuit with high input jitter tolerance," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 2, pp. 325–330, 1989.

[7] H. Zhang, S. Krooswyk, and J. Ou, "Chapter 4 - link circuits and architecture," in *High Speed Digital Design*, H. Zhang, S. Krooswyk, and J. Ou, Eds. Boston: Morgan Kaufmann, 2015, pp. 163 – 198. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124186637000046

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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

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

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

[19] 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.