UNIVERSITY OF CALIFORNIA SAN DIEGO

Improving the Burst-mode Capability of Commercial Transceivers

A thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Electrical Engineering (Electronic Circuits and Systems)

by

Jason Kelley

Committee in charge:

Professor George Papen, Chair
Professor Joseph Ford
Professor Shayan Mookherjea

2020

The Thesis of Jason Kelley is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to extend my most sincere thanks to Professor George Papen. Without his guidance, feedback, and resources this thesis absolutely would not have been possible.

I also thank Dr. Alex Forencich for providing his support with test equipment and practical knowledge.

I thank Professor Joseph Ford for his words of wisdom and giving me an initial push to take on this work. His dry email humor was a sight for sore eyes.

I thank LateX and Overleaf for their incredible efforts in figure and table numbering.

A final thank you to Leah Searl for her support and dinner company.

Chapter 2, in part, has been submitted and accepted for publication and oral presentation at Optical Interconnects Conference 2020. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 2 is coauthored with Forencich, Alex. The thesis author was the primary author of this chapter.

ABSTRACT OF THE THESIS

Improving the Burst-mode Capability of Commercial Transceivers

by

Jason Kelley

Master of Science in Electrical Engineering (Electronic Circuits and Systems)

University of California San Diego, 2020

Professor George Papen, Chair

The thesis presents methods to enable commercial off-the-shelf (COTS) datacom transceivers for use in optical circuit-switching applications. These applications require fast or (burst-mode) locking characteristics. This goal requires research in two areas. First, the burst-mode characteristics of (COTS) datacom transceivers were characterized as a function of the mean received power, the power offset between the channels being switched, and the transition time of the optical switch. Second, a high-speed board that can be used for variable optical attenuation was developed and tested to mitigate the power offset. This substantially improves the burst-mode lock times. The culmination of this work is a calibration procedure that can improve the locking characteristics of COTS transceivers used in a circuit-switched optical network.

Chapter 1

Introduction

Over the last few decades, there has been an explosion in digital data generation and usage across all commercial and industrial sectors. Around the early 2000's, the popularization of PC's for home and work purposes accelerated digital data generation dramatically. More recently, big cloud-based data processing applications—particularly those involving machine learning—have resulted in the the rate of digital data generation doubling approximately every year [7]. For efficient cloud-based services, all of this data must be stored, transmitted, and processed efficiently. Data centers form the computational backbone of these cloud-based operations, porting data to and from servers for processing. Many of these inter-server data links are handled by networks of that use 5-9 optical transceivers per multi-hop connection. Existing server transceiver technology is now commonly operating at 100Gb/s with future architectures being developed for 400 Gb/s transcievers.

Currently, the most common datacenter network topology is a multi-layer "Fat-Tree" topology as it can scale out the network bandwidth in an incremental manner. Using this topology, most modern datacenter networks are built from traditional electrical packet switches. Electrical packet switches route packets through networks in a distributed, scalable fashion.

However, packet switches are running into serious scaling challenges because the rate of

increase in the aggregate bandwidth within a datacenter is increasing much faster than the rate of increase of the switch ASIC bandwidth. Moreover, the number of ports on a single switch ASIC is limited. This means that larger networks require more layers of switches in order to provide sufficient ports and sufficient bandwidth.

These issues have motivated research into other kinds of datacenter architectures. One alternative network solution is to use optical circuit switching (OCS). Optical circuit switches operate by making and breaking connections at the physical layer. As a result, the operation of such a switch is independent of the line rate and modulation format, enabling switching of very large bandwidths per port. In OCS data center architectures, there is potential for the switch itself to be the only layer, provided the switching speed between network configurations is fast enough [5].

However, in a circuit-switched network, the connections between end points in the network must be coordinated so that data follows the correct paths through the switches. These connections must be quickly re-established at the physical layer after an optical switching event. However, the existing standards for commercial datacom transceivers (COTs) were not explicitly designed for an interruption of the light path at the physical layer. Instead, COTs transceivers are designed for point-to-point links for which the link is established only once. For this case, the receiver initially locks onto a data stream and maintains that link over the lifetime of the transceiver. Once a link is established, it is kept ”alive” irrespective of any data transmission by sending idle packets over the channel so that a transceiver never sees a break in the data stream. The design and traditional operation of these optical transceivers naturally poses a problem when integrated with an OCS element: the switch interrupts the data stream over switching time.

This means that OCS architectures require burst-mode functionality in the optical receiver so that it can quickly lock onto the incoming data stream after a switching event. Burst-mode optical receivers are not by any means a new technology. There are a variety of published and tested optical receiver designs capable of well under microsecond lock times [8, 9]. Moreover,

burst-mode receivers of this sort have been employed for PON applications since the early 2000's [6]. However, burst-mode functionality is not explicitly specified in existing datacom standards because all existing datacenter architectures use point-to-point optical links. For example, several of the COTS modules that were characterized are based on the PSM4 multi-source agreement (MSA) [1]. This agreement does not explicitly require the module to have a burst-mode capability. Nevertheless, there is considerable interest in continuing the use of PSM4 or other current COTS datacom transceivers in OCS applications because the technology is already a mature, widely employed technology and is therefore cost effective compared to a custom burst-mode transceiver designed to work with OCS. The burst-mode characteristics of COTS datacom transceivers have not been explicitly studied or measured before because that is not a requirement in current data centers. If they are to be employed in an OCS system, then their burst-mode characteristics must be more well-defined.

The first objective of this thesis is to accurately characterize the burst-mode characteristics of several COTS transceivers at 10Gb/s and 25Gb/s line rates utilizing a time-domain Bit-Error Rate (BER) measurement method. The second objective is to derive and validate methods to reduce the variability of the locking time, which is a strong function of the difference in received power during a switching transition [6]. This method involves developing and testing a circuit board that can equalize the power during a switching event on a microsecond time scale using a variable optical attenuator (VOA).

The thesis is organized as follows. Chapter 2 will discuss the experimental setups used for characterizing the burst-mode characteristics of several commercial single and multi mode datacom transceivers. The results will be described in some detail and the characteristics of each transceiver over a nominally standard set of test conditions will be compared. In Chapter 3, a solution to improve the burst-mode characteristics of one or more of these transceivers will be proposed. Hardware that was developed to solve this issue will be described and results of hardware tests will be presented. Chapter 3 will contain a brief conclusion of the results of

this study. Note that part of this thesis wad included in a paper that has been accepted for oral presentation at Optical Interconnects 2020.

Chapter 2

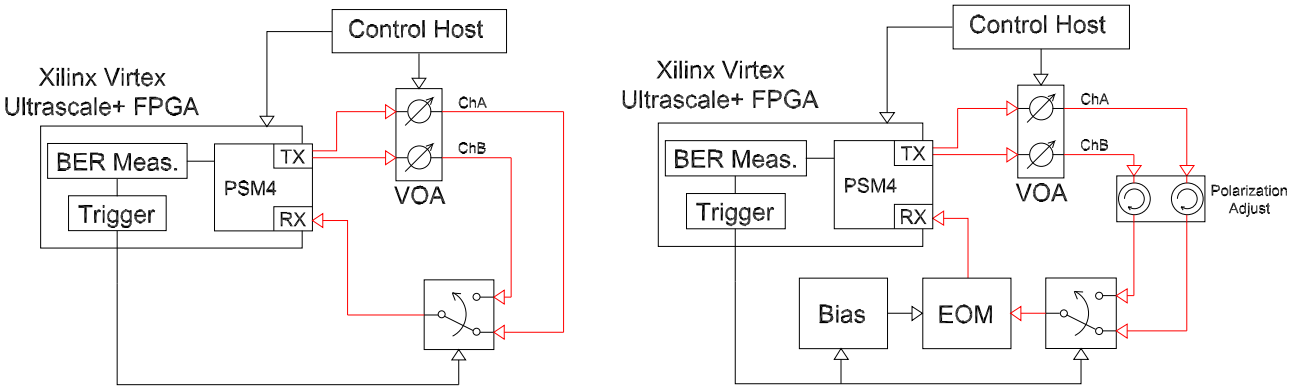
Burst-Mode Transceiver Characteristics

This chapter presents measurements of the locking-time characteristics of commercial transceivers as a function of both the total link attenuation and the difference in the received power after a switching event.

2.1 Measurement Setups

In order to assess the burst-mode capabilities of datacom transceivers, BER locking characteristics need to be measured in time-resolved manner. Due to measurement equipment restrictions, two setups were devised. These setups are shown in Figure 2.1.

At the core of each setup is a Xilinx Virtex Ultrascale+ FPGA Dev board. The development boards house the on chip and peripheral capabilities to support a multitude of optical transceivers. The FPGA was imaged with a modified version of the BER measurement hardware in Corundum [4], making the FPGA function in the setup as a BER measurement device. In this method, errors in received bits are counted and grouped in temporal bins on the order of a microsecond at line rates up to 25Gb/s. Burst-mode locking characteristics are measured by tracking the BER of a single receiver when the link is subjected to differing conditions before, during, and after a switching event. The development of this method was not part of this thesis.



(a) Short Dead Time Setup

(b) Variation Dead Time Setup

Figure 2.1: Measurement setups used in the characterization of transceivers. (a) shows a setup used to measure the locking characteristics of transceivers for very short dead times. This setup is capable of measuring over a large range of mean attenuations and offsets due to the low loss. (b) shows the setup used to simulate switching dead times during the measurements of locking characteristics. The electo-optic modulator and polarizers added significant loss to the setup and restricted the useable attenuation levels during measurements.

Setup A (shown in Figure 2.1a), simply consists of the FPGA and transceiver, an SPDT optical switch, and a calibrated VOA. This setup was designed to measure locking characteristics as a function of the power difference between two channels. Two channels, denoted 'ChA' and 'ChB' were broken out of the transceiver and passed through the VOA. One channel was selected by the switch to be fed back to the receiver. The switch transitions were controlled on an automated basis by the FPGA to correspond with measurement intervals. Depending on whether the measurements were being done on a single or multi-mode transceiver, the switch would be exchanged for either a single or multimode model. The VOA was used to provide power offset and/or equalization between the TX channels depending on what measurement was being taken. The VOA was manually characterized and software calibrated. The VOA and FPGA were controlled by a host PC, automating measurements over several hours. Due to the short transition time of the optical switch, Setup 2.1a could only be used to gather "heat map" data for an assumed 1 μs dead time. That dead time corresponds to the temporal resolution used for the measurements done with this setup. However, the low loss of Setup 2.1a provides a good

look into how the locking characteristics change as a function of power since BER data could be collected over a wide range of channel offsets about the mean channel attenuation while still remaining within operational limits of the transceiver. As will be discussed shortly, Setup 2.1b can only support a narrow range of of power offsets, but has the capability of being able simulate different switching dead times.

Setup B (shown in Figure 2.1b) contains the same components as setup A with the addition of an electro-optic modulator (EOM), its bias, and polarization rotators. Depending on the design of a switch, there may be a period where no signal signal allowed through to the receiver. Setup B was designed to measure this period, called a "dead time". The EOM was included to simulate the effect of an optical switch's dead time. By adjusting the bias point in sync with the measurement cycles of the FPGA, the EOM could completely cutoff the channel guiding power to the receiver for a configurable length of time. During the dead time, the switch would be triggered to switch the selected channel. Because the EOM transitivity is dependant on input polarization, polarization rotators were included to minimize loss. Additionally, the EOM has a substantial amount of loss which leads to the transceiver under test to generate too many errors. Moreover, Setup 2.1b is only capable of measuring single mode modules because of the use of the EOM. All of these factors contribute to the decision to develop two independent test setups to characterize the performance of a COTs transceiver over a wide range of operating points.

The configuration of the hardware in the setup depended on the transceiver module. The single-mode setups ran at 1310nm and the multimode setups ran at 850nm, both industry standard operating wavelengths. The single mode switch was the Agiltron NSSW-123111332, which had a 76ns switching speed, 0.4dB insertion loss, and -28dB crosstalk. The multimode switch was the Agiltron NSMS-128115332, which had a 76ns switching speed, 1.7dB insertion loss, and -18dB crosstalk. The EOM was rated for 10Gb/s and had an extinction ratio of 12dB and an insertion loss of 7dB. The VOA was a Dicon GP600 model with two single mode and two multi-mode

attenuator channels. The single mode channels could be adjusted from 0.5dB to 30dB attenuation and the multi-mode channels could be adjusted from 1dB to 30dB attenuation.

2.2 Results

A small subset of the collected data for the single and multimode transceivers is presented in the form of a "heat map", with the y axis being the power offset ratio $10\log_{10}(P_A/P_B)$ between the two channels and the x axis being time. Mean channel attenuation values were chosen to be representative of values seen in an optical rotor switch. To quantitatively compare the transceivers locking characteristics, a channel was considered "locked" after a switching event when the BER was below 10^{-8} and was considered "error free" when the BER was below 10^{-10}. The lock times to 10^{-8} BER for each value of power offset were plotted next to the heat map. Lock times over 40 μs were truncated for more visibility in the plot. All data was taken using Setup 2.1a. The only difference was that single mode or multimode hardware was inserted as appropriate and the on-module clock-data recovery circuit (CDR) was turned off for the 25Gbps measurements.

A brief summary of the observed characteristics is be given the next two subsections. A summary of the locking characteristics for each transceiver is provided in Table 2.1.

2.2.1 Single-Mode Transceivers

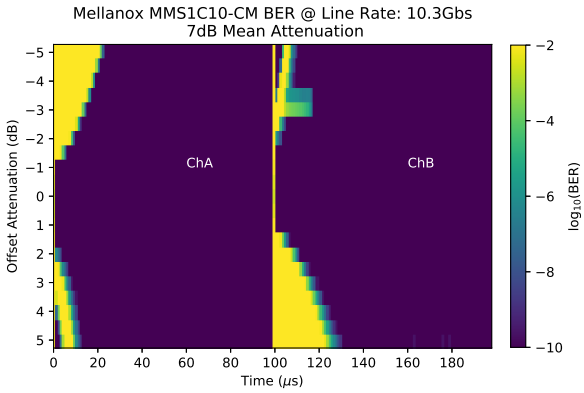
Figure 2.2 shows the performance of a Mellanox MMS1C10-CM, which costs about $ 1000 and has a rated range of 500m. The power on the receiver was recorded to be 126 μW at 7dB mean attenuation and 63 μW at 10dB mean attenuation. As the mean channel attenuation increases, the lock times get generally longer. Both heat maps show a smooth, RC-like increase in lock time as the magnitude of the power offset gets larger. That is not unexpected, since many older and lock time insensitive applications employ optical receiver receiver circuits with AC coupling between gain stages [6]. Due to the presence of these AC coupling capacitors, the

threshold detection circuits must wait until the RC circuits settle before they can begin to lock to the new data stream, which leads to an a smooth, analog-esque rise in the lock time with increasing power offsets. Lock times on transitions from a higher to lower power channel, as seen in the transitions from ChA to ChB on the lower half the heat maps, are generally longer than lock times on transitions from low to higher power, as seen in the transitions from ChB to ChA on the upper half of the heat maps. That behavior is unsurprising as the lower power is more difficult for the transceiver to detect. The increase in errors seen at the highest attenuation is due to the signal received by the transceiver being below the minimum incoming power required for error-free operation. The $B \rightarrow A$ transition in Figure 2.2 starts at 0 μs, and the $A \rightarrow B$ transition starts at about 100 μs. The locking time of the transceiver is measured with respect to these values for each transition.

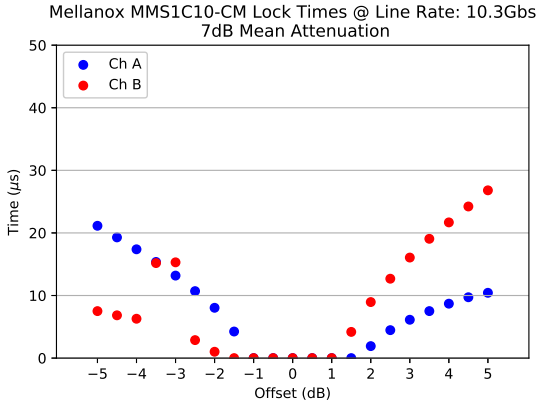
The same Mellanox module was also measured at a line rate of 25 Gbps over the same physical channel as with the 10 Gbps data rate. As expected, the lock times are generally longer due to the overall lower power levels per bit. There is a substantial increase in errors on ChA for the low power offset settings. This effect is unique to this transceiver's operation at 25Gbps. It may be due to a second order or nonlinear threshold detection method causing an excessive settling time or ringing for this large absolute power step. The extracted lock times show an increase in sensitivity compared to operation at 10 Gbps.

The next set of measurements, shown in Figure 2.4, are for the FiberStore QSFP-PLR4-40G, which is about $ 300, and is designed for 10 km. It is rated for a maximum line rate of 10 Gbps. At 10 dB channel attenuation, the power incoming to the receiver was measured to be 63 μW, and at 7 dB channel attenuation, the received power was 126 μW. Lock times for this transceiver are fairly uniform and short for both mean attenuations of 7dB and 10dB.

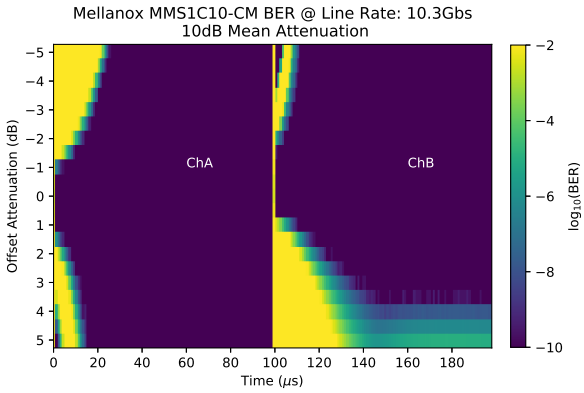
The FiberStore QSFP+-PIR4-40G is another 10 Gbps maximum transceiver. Figure 2.5 shows its BER heat maps and extracted lock times. It costs $220 and has a rated range of 1.4km. In constast to the first two transcievers, this transceiver exhibited significant ringing in the BER



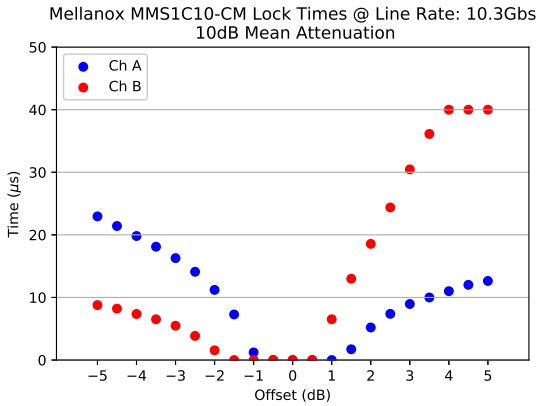
(a) BER heat map for 7dB channel attenuation.



(b) Extracted lock times for 7dB channel attenuation.



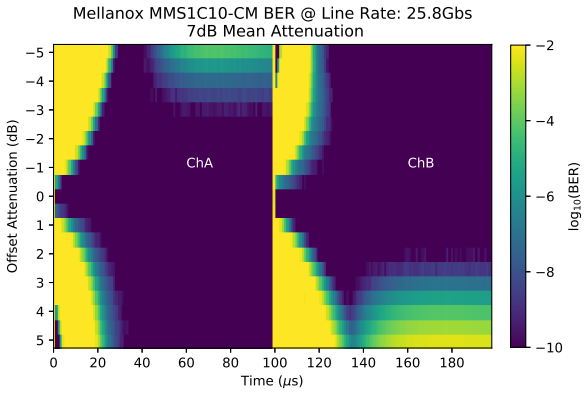
(c) BER heat map for 10dB channel attenuation.



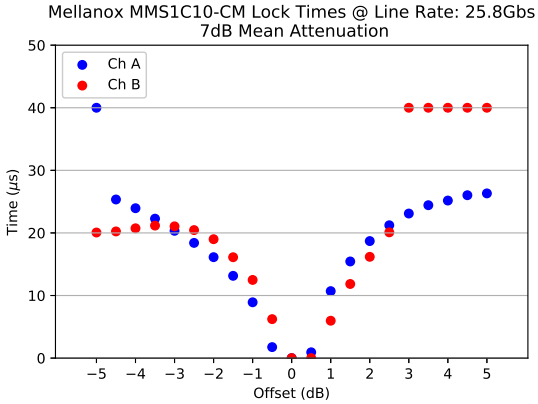
(d) Extracted lock times for 10dB channel attenuation.

Figure 2.2: Mellanox MMS1C10-CM BER measurements at a line rate of 10.3 Gbps. There is a smooth increase in the lock time as a function of power offset for this transceiver. The additional attenuation generally increases the lock times. The attenuation also increases their sensitivity to a power offset.

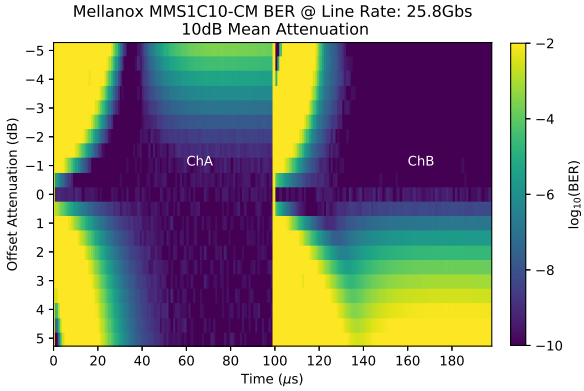
directly after the transition as the transceiver is attempting to adjust its threshold. This was seen in the BER data for the transceiver in Figure 2.4 as well, though it was far less pronounced. Despite these quirks, this transceiver seems to lock quickly over a large range of power offsets, and is a suitable candidate for further studies for power offset mitigation.



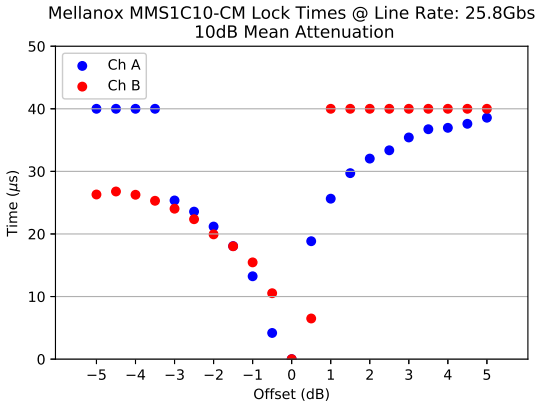
(a) BER heat map for 7dB channel attenuation.



(b) Extracted lock times for 7dB channel attenuation



(c) BER heat map for 10dB channel attenuation.

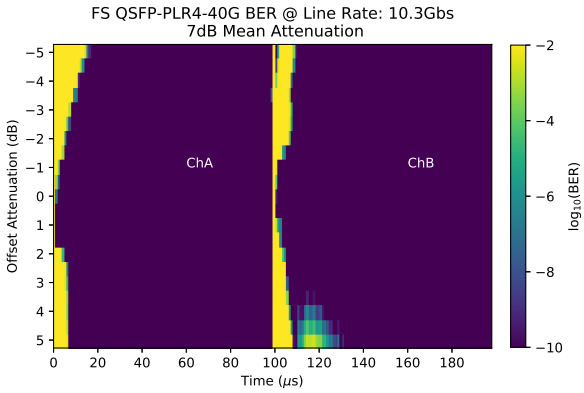


(d) Extracted lock times for 10 dB channel attenuation.

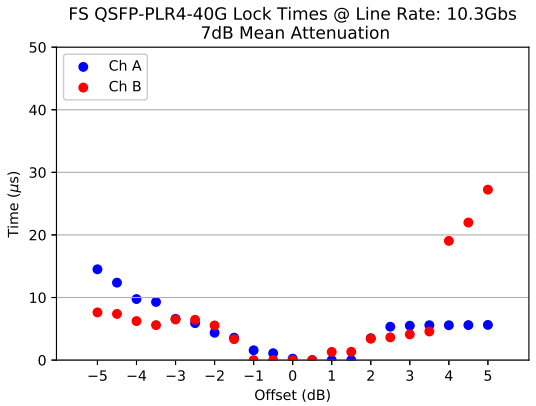
Figure 2.3: Mellanox MMS1C10-CM BER measurements at a line rate of 25.8 Gbps. The shape and trends are similar to the 10 Gbps case (Figure 2.2), however the error-rate floor is reached much earlier, as expected, and it is much more sensitive to power offset.

2.2.2 Multi-Mode Transceivers

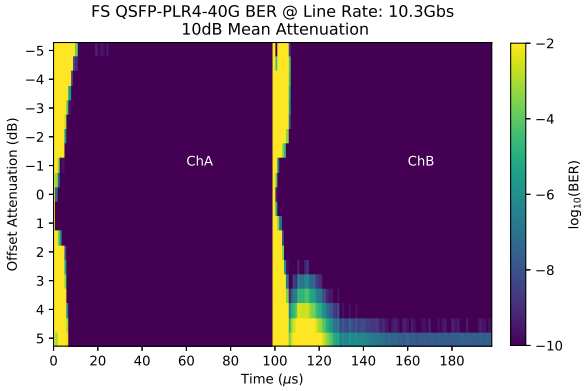
Multi mode transceivers are also of interest in designing data centers due to their low cost as compared to their single mode cousins. However, multi mode data links typically have a smaller link margin than similar single mode links, restricting their range to a maximum of around 600 m. That is due to the difficulty of designing functional hardware that can deal with how all the modes present in the fiber will propagate through the system. Despite these added



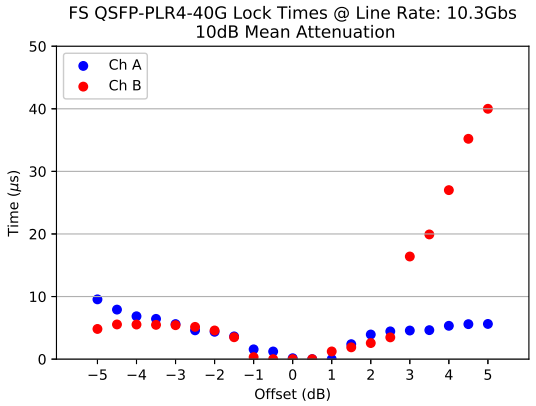
(a) BER heat map for 7 dB channel attenuation.



(b) Extracted lock times for 7 dB channel attenuation.



(c) BER heat map for 10 dB channel attenuation.

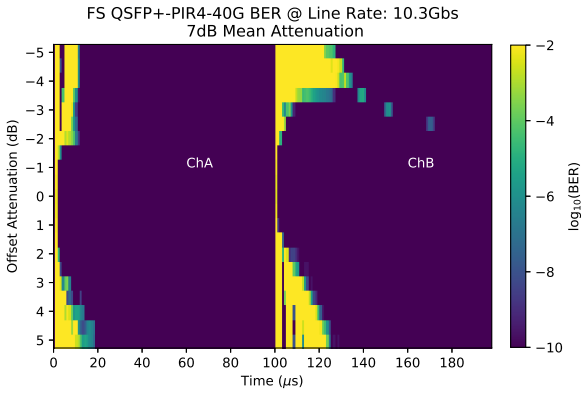


(d) Extracted lock times for 10 dB channel attenuation.

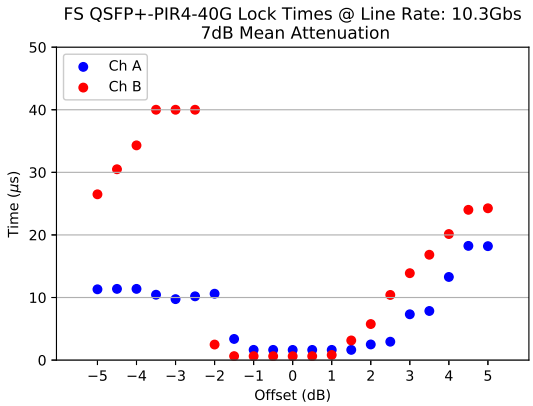
Figure 2.4: FiberStore QSFP-PLR4-40G BER measurements at a line rate of 10.3Gbps. Lock times are very short and have a low sensitivity to power offset. There is also a low sensitivity to power offset ratio.

requirements, multi mode transceivers have an attractive cost point.

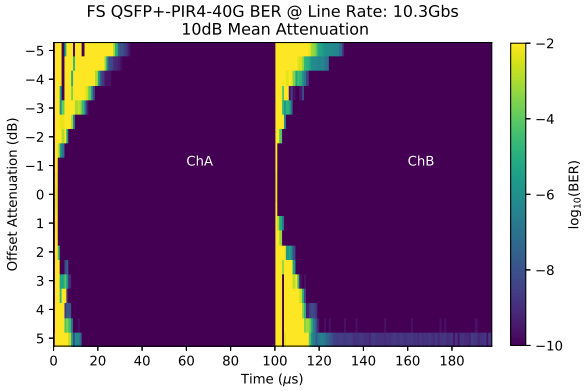
The first of three multi mode transceivers examined was the FTLC9555REPM, made by Finisar. This module costs about $260 and has a rated range of 100m. Heat maps and extracted lock times were gathered for mean channel attenuations of, again, 7 dB and 10 dB at a line rate of 10 Gbps and are shown in Figure 2.6. The multimode hardware used in the test setup was generally more lossy than the single-mode hardware, which placed a limit on the mean attenuation of about 4 dB. The power offset range for the 7 dB case was thus reduced to $\pm$ 3 dB from $\pm$ 5 dB.



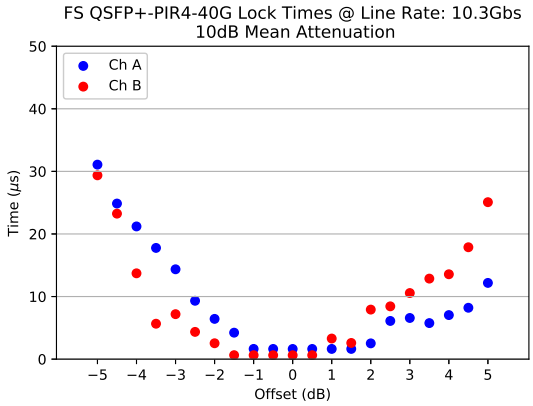
(a) BER heat map for 7 dB channel attenuation.



(b) Extracted lock times for 7 dB channel attenuation.



(c) BER heat map for 10 dB channel attenuation

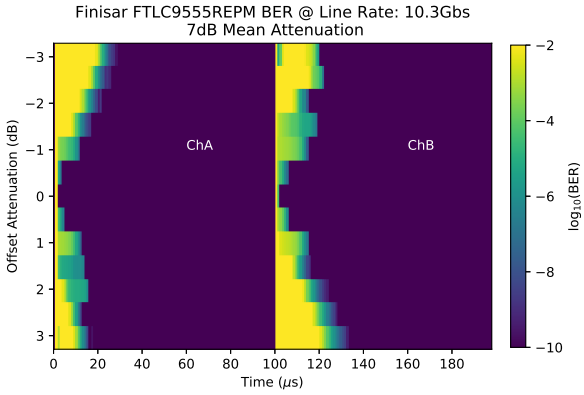


(d) Extracted lock times for 10 dB channel attenuation.

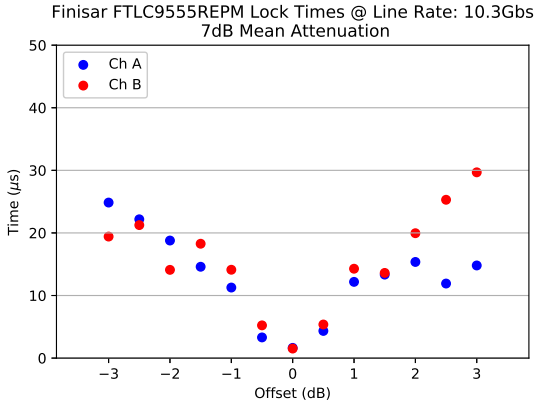
Figure 2.5: FiberStore QSFP+-PIR4-40G BER measurements at a line rate of 10.3 Gbps. Lock times are very short and have allow sensitivity to power offset. There is also a low sensitivity to power offset ratio.

The offset range for 10 dB channel attenuation could be kept at $\pm$ 5 dB. The FTLC9555REPM's locking characteristics at 7 dB mean attenuation is rather choppy, but smooths considerably at lower power levels, as demonstrated by the 10dB channel mean attenuation results.

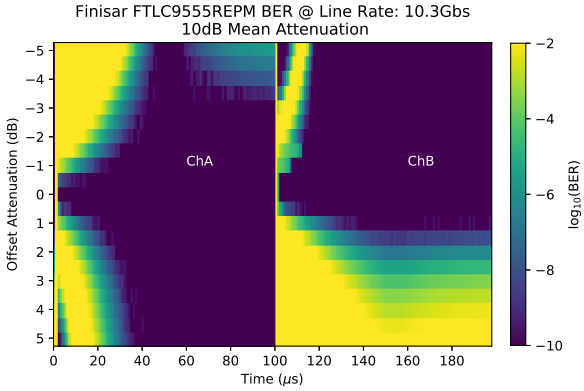
The last FiberStore transceiver measured was the QSFP28-SR4. This transceiver costs $99 and has a rated range of 100 m. This transceiver is capable of 25 Gbps line rate operation, but was measured here at 10 Gbps, and 7dB mean attenuation. The power at the receiver was about 147 μW. Despite the low power levels, this transceivers exhibited exceptionally low lock



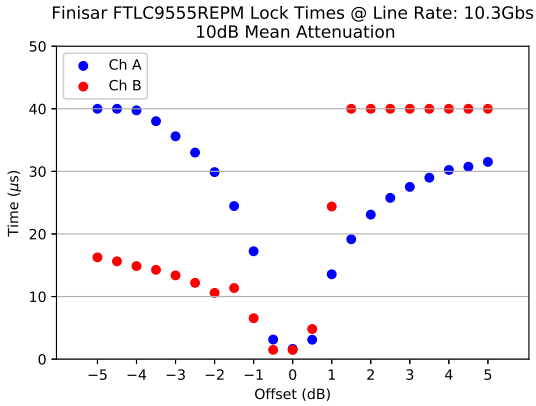
(a) BER heat map for 7 dB Channel attenuation.



(b) Extracted lock times for 7 dB channel attenuation



(c) BER heat map for 10 dB channel attenuation.

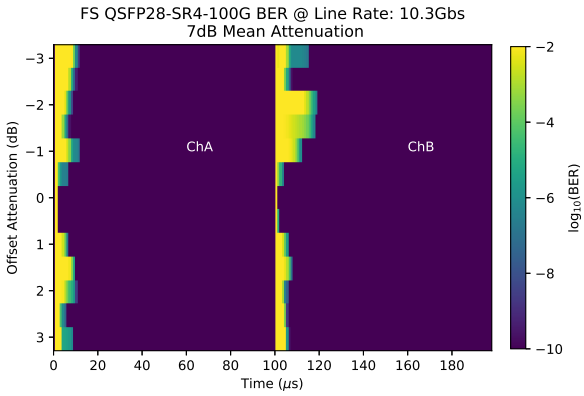


(d) Extracted lock times for 10 dB channel attenuation.

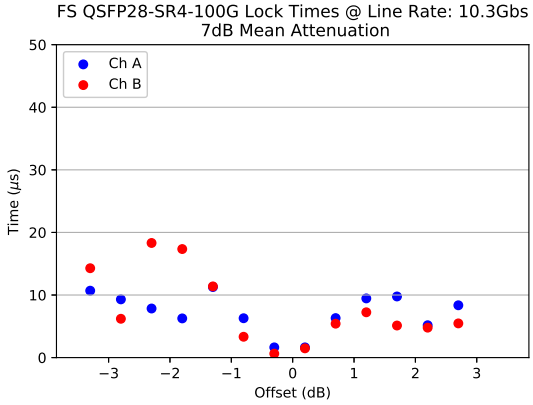
Figure 2.6: Finisar FTLC9555REPM BER measurements at a line rate of 10.3 Gbps. The offset range at 7 dB mean attenuation is reduced because of the larger losses in the VOA and switch. Lock times degrade significantly with a change in mean attenuation because the module is very close to its RX error-rate floor at 10 dB mean attenuation. Locking characteristics look sporadic at 7 dB mean attenuation, possibly indicating some nonlinear threshold detection that can only be seen at larger power levels.

times across the entire range of power offsets measured at both 7 dB and 10 dB mean attenuation. The performance difference between this transceiver and the Finisar FTLC9555REPM, which costs three times as much, is striking. This highlights the importance of actually measuring the characteristics of these transceivers. Burst-mode requirements are not listed in the specifications of these devices, so cost and listed specifications of the model provide no indication as to the

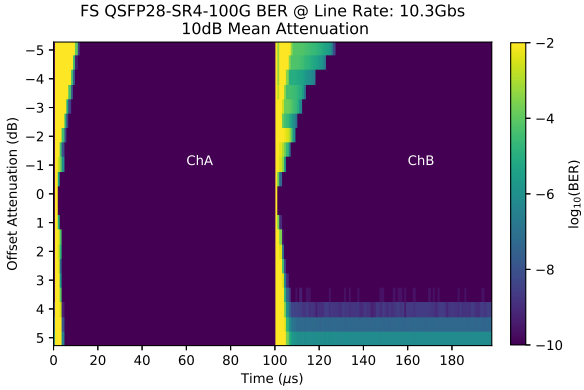
burst-mode performance of a transceiver.



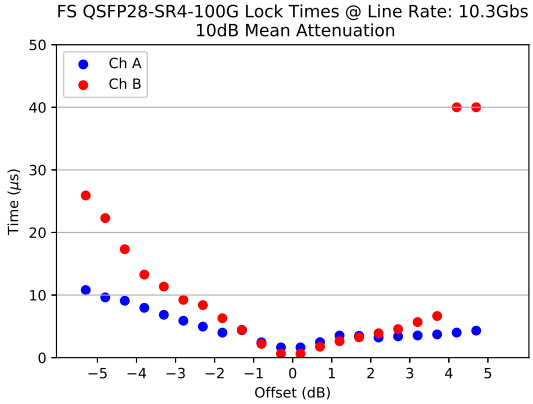
(a) BER heat map for 7 dB channel attenuation.



(b) Extracted lock times for 7 dB channel attenuation.



(c) BER heat map for 10 dB channel attenuation.

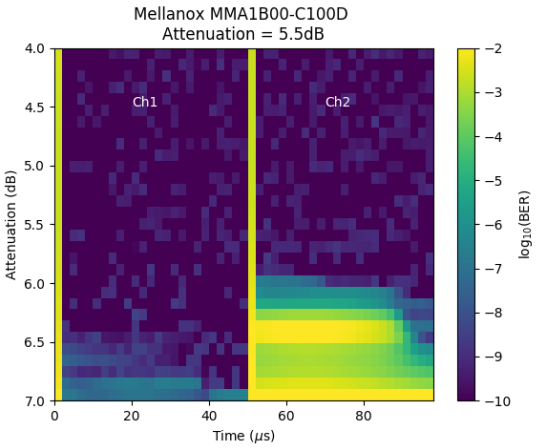


(d) Extracted lock times for 10 dB channel attenuation.

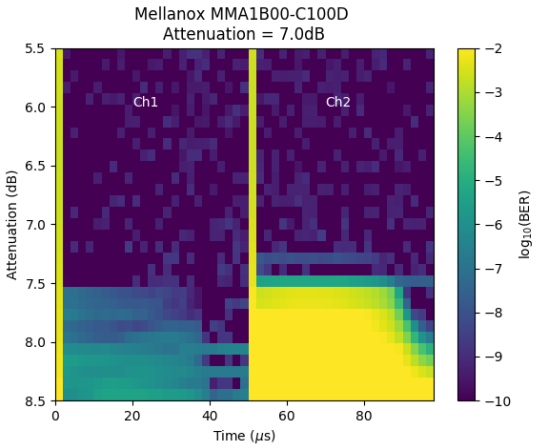
Figure 2.7: FiberStore QSFP28-SR4 BER measurements at a line rate of 10.3 Gbps. Lock times are very short and have a low sensitivity to power offset. There is a strikingly low sensitivity to power offset ratio for both of the tested attenuations.

Finally, there is the interesting case of the Mellanox MMA1B00-C100D, another multi-mode transceiver. This transceiver runs at about $300 and is rated for a range up to 100m. As seen in Figure 2.8, this transceiver has extremely quick lock times over a large range as a long as the power offset between the two channels remains within certain bounds. Outside of these bounds however, the transceiver does not display a gradual roll-off in lock speed. The loss in performance is abrupt and drastic. It can be seen clearly in Figure 2.8a that the transceiver uses

some kind of stepped threshold detection technique, where fast locking is guaranteed over small range of power offsets, but one out of that range, the lock speed is relatively long and somewhat random as the threshold makes broad leaps in order get in the appropriate range to detect a new set point.



(a) BER heat map for 7 dB channel attenuation.



(b) BER heat map for 8 dB channel attenuation.

Figure 2.8: BER heat maps at (a) 7dB and (b) 8dB channel attenuations. Past a certain power offset, and only for positive power offsets, the locking time jumps from near instant to exceptionally long. A change of 1 dB mean attenuation noticeably alters the pattern of locking to a degree that is uncharacteristic of any of the other transceivers.

2.2.3 Summary of Transceiver Characteristics for Short Dead Times

The key characteristics of each of these transceivers is summarized in Table 2.1. It can be seen these results that there is a wide range of algorithms and circuits that were used for threshold detection in these devices. These results are for a single, randomly selected transceiver and may vary slightly from transceiver to transceiver. In some initial tests, the variance between transceivers of the same model proved to be rather small when the receivers were subject to the same incident power levels. The variation in transmitted power across all four channels of each transceiver could vary quite a bit depending on the transceiver model. For example, the output power of all four channels of 10 FiberStore QSFP+-PIR4-40G transceivers were measured

directly using an optical power meter and were seen to vary as per the specification by as much as 3dB. By contrast, all four of the Finisar FTLC9555REPM transmitter output powers were within 1% of each other. By that example alone, it is fairly obvious that depending on the manufacturer or even transceiver model, there can be large performance variations for these devices that could be very detrimental in a system requiring uniformity in performance of multiple devices. The potential variation in output power between channels of a transceiver alone could account for substantial variation in the locking times.

These large variations in power led to work on how to mitigate these variations. That is the topic of the next chapter.

Table 2.1: Summarized burst-mode characteristics of various transceivers

Transceiver	λ (nm)	Line Rate (Gbps)	Mean Att. (dB)	RX Power @ 10dB Att. (dBm)	Lock Window (dB)
MMS1C10-CM	1310	10.3	7	-8.9	4.3
			10	-11.9	2.9
		25.8	7	-8.9	1.8
			10	-11.9	0.9
QSFP-PLR4-40G	1310	10.3	7	-9.0	6.9
			10	-12.0	7.7
QSFP+-PIR4-40G	1310	10.3	7	-9.3	3.0
			10	-12.3	5.3
FTLC9555REPM	850	10.3	7	-6.8	1.5
			10	-9.8	1.4
QSFP28-PIR4-100G	850	10.3	7	-7.9	2.3
			10	-10.9	6.8
MMA1B00	850	10.3	7	-7.7	-
			9	-9.7	-

2.2.4 Dead Time Measurements

In order to understand a how the inclusion of a particular transceiver will affect the system level performance of an OCS system, the effect of a switching dead time must be examined. The setup of Figure 2.1b was used to characterize a Mellanox MMS1C10-CM at 25Gbps. It was seen

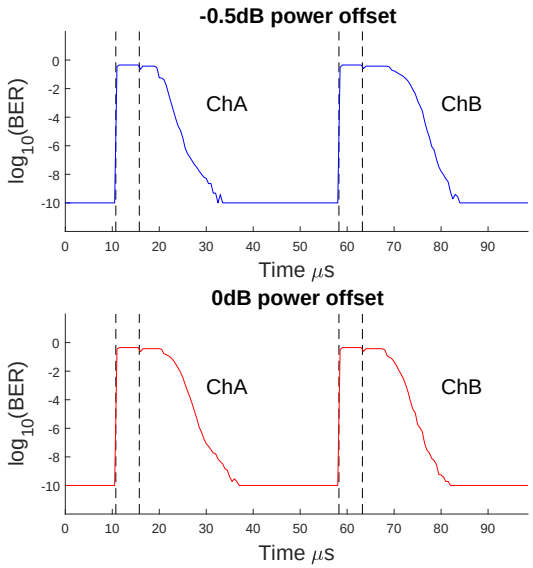
in the previous characterizations of locking characteristics for short dead times that the locking speed of COTS transceivers is dependent highly on the difference in power between two channels. The case of the Mellanox MMS1C10-CM operating at 25Gbps was exceptionally sensitive while still exhibiting well-behaved trends on lock time. For this reason, it can serve as a excellent case study of the effect of switching dead time on lock characteristics.

Presented in Figure 2.9 are the BER measurements for this module for switching dead times of 5 μs, 10 μs, 20 μs, and 40 μs. The channel was set to a mean attenuation of 7dB (power at the receiver given in table 2.1 and power offsets of 1dB and 1.5dB. When this data was collected, and 1.5dB power offset approximately equalized the power between the channels. Extracted lock times to 10^{-8} BER and the percent difference between the lock times of each power offset for a certain dead time setting are presented in Table 2.2.

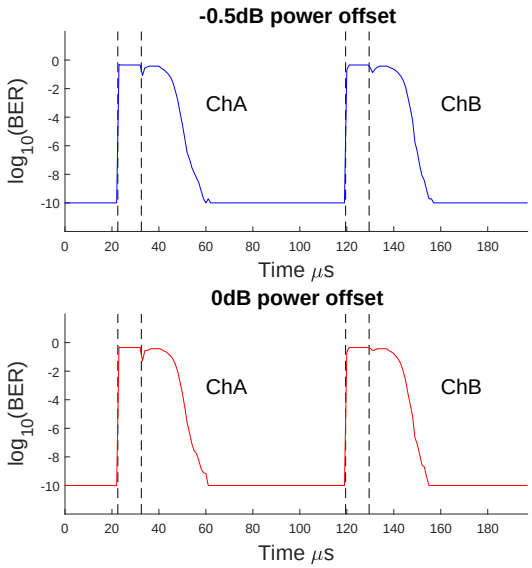
Table 2.2: Extracted lock times for the Mellanox MMS1C10-CM at various dead times power offsets. As the dead time increases, the difference in lock time between two power offsets decreases dramatically.

Dead time (μs)	Power Offset (dB)	Lock Time B→A (μs)	Lock Time A→B (μs)	|Δ Lock Time| (μs)
5	-0.5	13.77	17.70	3.93
	0	17.01	15.13	1.88
10	-0.5	23.83	23.40	0.43
	0	24.97	22.74	2.23
20	-0.5	35.08	31.73	3.35
	0	34.02	31.60	2.42
40	-0.5	>40	36.57	-
	0	>40	36.84	-

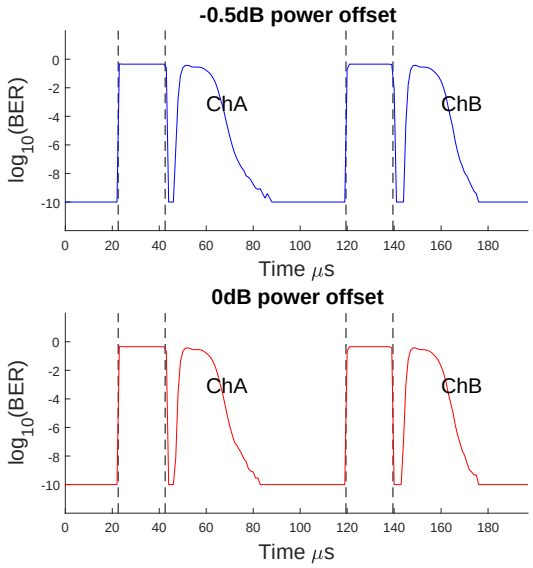
From the trend of difference in lock times for each case of Table 2.2, a larger switching dead time time makes this transceiver markedly less sensitive to power offsets in addition to substantially increasing the time it takes for a transceiver to lock. There is a substantial difference in lock time between a 5 μs and 10 μs dead time. The difference between a 20 μs and 40 μs dead time is negligible. This is likely because the transceiver will quickly lose memory of the characteristics of the previously received data stream when no light is incident on the receiver. All



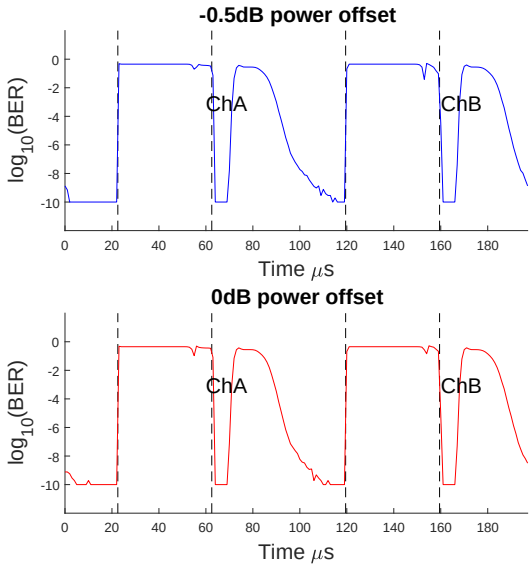
(a) BER heat map for 7 dB channel attenuation.



(b) BER heat map for 8 dB channel attenuation.



(c) BER heat map for 7 dB channel attenuation.



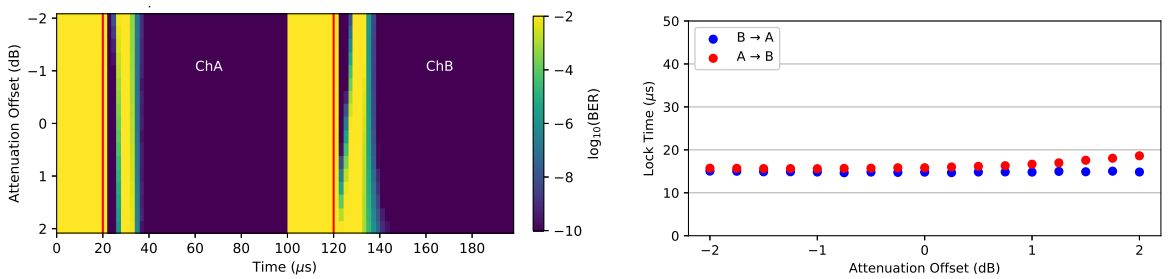
(d) BER heat map for 8 dB channel attenuation.

Figure 2.9: BER measurements over time for the Mellanox MMS1C10-CM at 10Gbps for mean channel attenuations of 7 and 8dB and power offsets of -0.5dB and 0dB. BER was measured for dead times of (a) 5 μs, (b) 10 μs, (c) 20 μs, and (d) 40 μs.

receivers will exhibit this characteristic since they clearly all have a dynamic method determining threshold settings depending on the incoming data stream. In addition to being able to determine threshold settings on detection of a link, the threshold setting method needs to be able to adjust

during operation to account for any drift in the link characteristics.

This can be expanded to see locking characteristics over a larger power offset range, forming heat maps, and extracting lock times per offset settings as was done in the previous sections. Figures 2.10 and 2.11 show heat maps and extracted lock times to 10^{-8} BER for line rates of 10Gbps and 25Gbps, respectively.

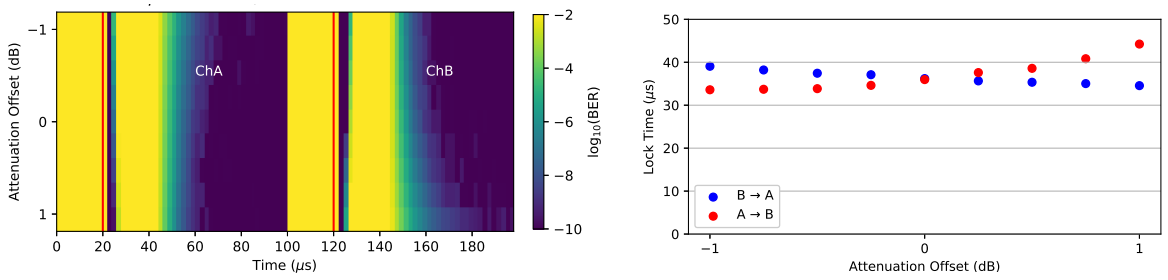


(a) BER Heat Map

(b) Extracted Lock Times

Figure 2.10: (a) BER heat map for Mellanox MMS1C10-CM at 10Gbps and a 20 μs dead time. Measurements taken for 9dB mean attenuation and a power offset range spanning ±2dB. (b) Extracted lock times for the accompanying heat map.

In contrast to the results of Figure 2.2, Figure 2.10 shows a decrease in the sensitivity in lock times over this power offset range. For very short dead times, it was seen that the transceiver would lock extremely quickly for power offsets under about ±1.5dB. With a 20 μs dead time, there is a near constant 16 μs lock time with a very weak dependence on the offset power.



(a) BER Heat Map

(b) Extracted Lock Times

Figure 2.11: (a) BER heat map for Mellanox MMS1C10-CM at 25Gbps and a 20 μs dead time. Measurements taken for 9dB mean attenuation and a power offset range spanning ±1dB. (b) Extracted lock times for the accompanying heat map.

Similarly, for a very short dead times this transceiver showed a strong dependence on

power offset at 25 Gbps. With the inclusion of the 20 μs dead time, that dependence has been flattened considerably and the minimum lock times are above 30 μs.

From these results, it can be concluded that a substantial switching dead time desensitizes the locking characteristic of a transceiver to power offsets between the channels. The data presented here is a case study of the Mellanox MMS1C10-CM. As demonstrated by the varied results of other sections, transceivers need to be measured on an individual basis in order to understand that model's response to switching dead times.

Chapter 2 is coauthored with Forencich, Alex. The thesis author was the primary author of this chapter.

Chapter 3

Equalization of Power Offsets during a Switching Event

3.1 High Speed VOA and drivers

The measurements presented in the previous chapter showed that the locking time is a strong function of the power difference between the two channels being switched as well as the dead time. The combination of these two effects and their effect on a selected transceiver must be well understood by a system designer to produce a repeatable and well-behaved product.

This chapter discusses the development of a fast multi-channel variable optical attenuator that can compensate for the power offset that occurs during a typical switching event. Switching dead time is determined by the switch, so its affect on the switching time cannot be easily compensated in a COTs device. The power offset that occurs during a switching event can be mitigated by placing a fast variable optical attenuators on each of the switch ports. Dynamic gain equalization like this is a known technique used to homogenize the characteristics of optical links and improve system performance [2]. Knowing the power offset, attenuators can be synchronized to the switching events, loading a new (pre-determined) attenuation setting during the switching

deadtime so as to produce no power offset during a switching event. For this purpose, an eight channel optical attenuator made by Kotura (now owned by Mellanox) were chosen and an accompanying driver board was designed and fabricated. The attenuators are rated to have a 40 dB attenuation range at 1550nm and an optical rise time of 1 μs when switching from 20 dB to 0 dB attenuation. A simplified electrical schematic of a driver board is shown in Figure 3.1. The VOA optical element, from an electrical drive point of view, is a diode. Each diode is operated in current mode by the circuit shown in Figure 3.2. The regulator circuits that control the current function as transconductance amplifiers when operated with a high side load. These regulator circuits were configured to provide a large signal gain of $50\frac{mA}{V}$. The control voltage signal for these regulators is generated from an 8 channel, 12-bit DAC with a maximum control voltage of 2.048 V, given by a voltage reference internal to the DAC. This sets the output current maximum at 40 mA with 12 bit resolution, or about 10 μA per set. These DACs showed a full swing rise and fall time of 750 ns, with the current sources accurately following the output DAC waveforms. The circuits are powered from a single 12 V supply and consume about 55 mA quiescent current at no load. A 3 line serial interface (SPI) was used to communicate with the DACs and set the output currents. The VOAs were mounted on Zero Insertion Force (ZIF) packages outfitted with a DB15 connector as an electrical interface to the diodes. The top and bottom layouts are shown in Figure 3.3. The assembled boards are pictured in Figure 3.4.

As shown in the schematic, there is no microcontroller or other control device present on the board. Instead, the SPI and other digital control interfaces from the DAC chip arew broekn out to a connector so that any microcontroller can be used. This feature enables greater system-level flexibility because of the ability to test and try different controller solutions. We also note that during board bring-up, a small error was present in the original wiring. This error was corrected by lifting the affected pin of the DAC chip and jumping it to the appropriate connection point.

Seen in Figure 3.5 are the optical rise and fall times of the VOA when driven with these circuits over a large 3 dB to 20 dB range. It should be noted that the rise and fall times are

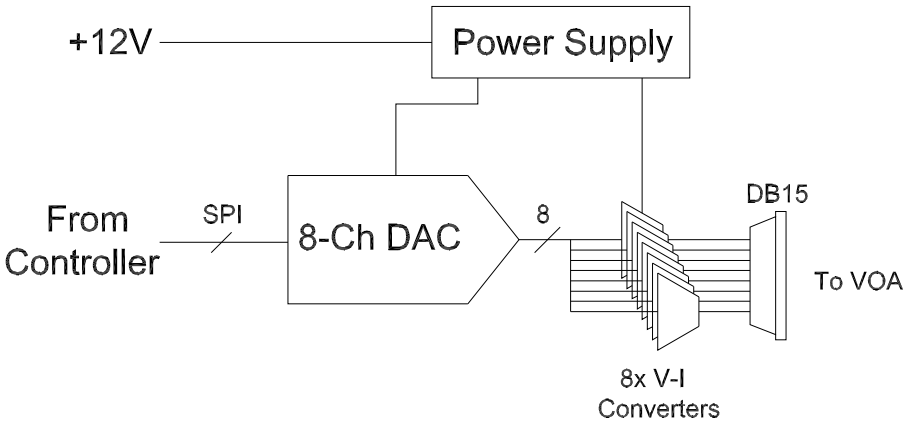


Figure 3.1: VOA Driver board block diagram. A control unit can adjust and simultaneously update all 8 DAC outputs over a SPI channel. Each of the 8 DAC outputs are fed through a transconductance amplifier in order to control the VOA channels. The entire board is powered from an external 12V supply.

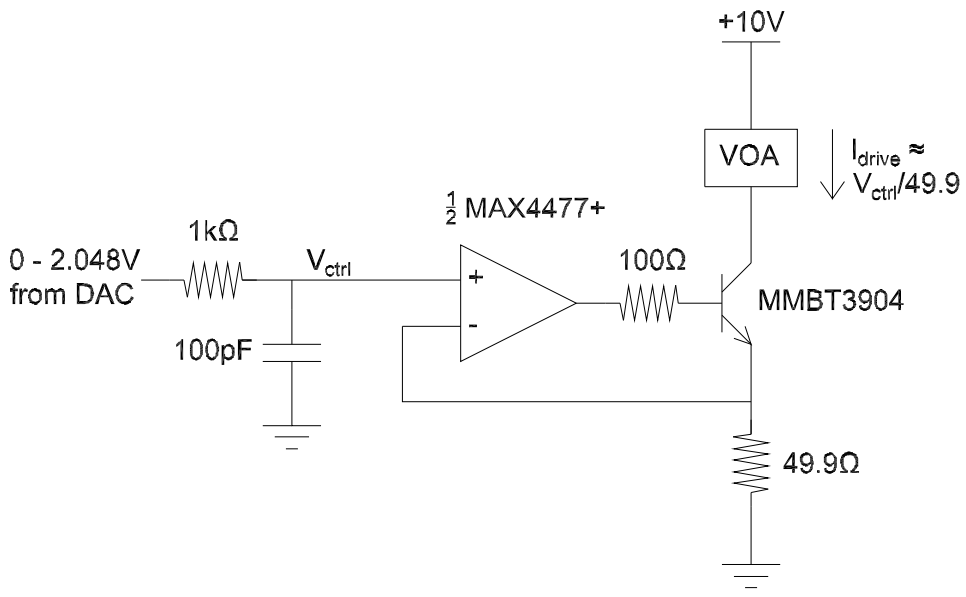
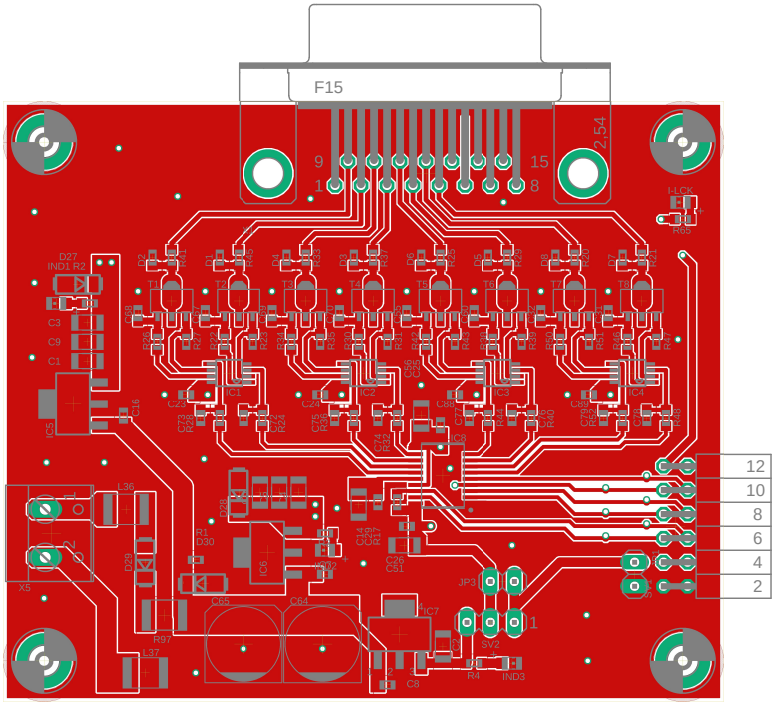
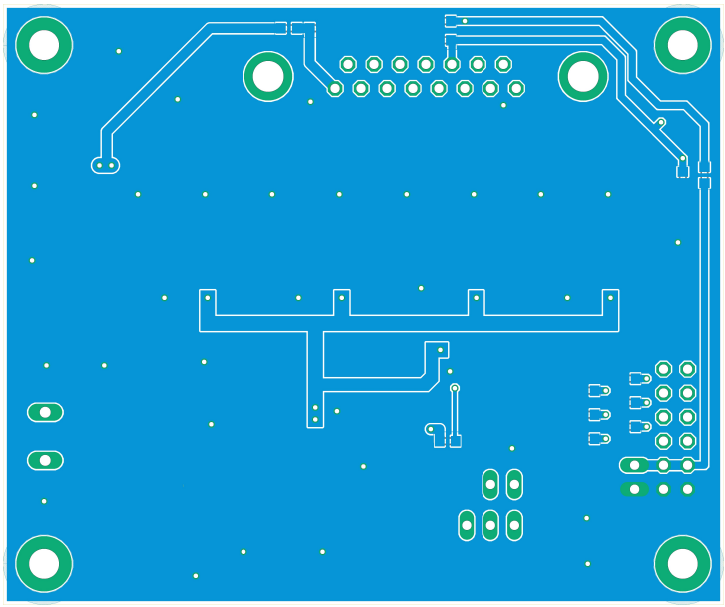


Figure 3.2: Transconductance amplifier circuit used to drive the VOA diodes. This is a low-side current source circuit, though is more commonly used as a voltage regulator.

dependent on the initial and final values of a transition. This is due to the slew rate of the chosen opamp and nonlinearity of the BJT. Although the feedback and high loop gain ensure the steady



(a) Top side layout.



(b) Bottom side layout.

Figure 3.3: Layout of (a) top and (b) bottom layers, done in Autodesk Eagle.

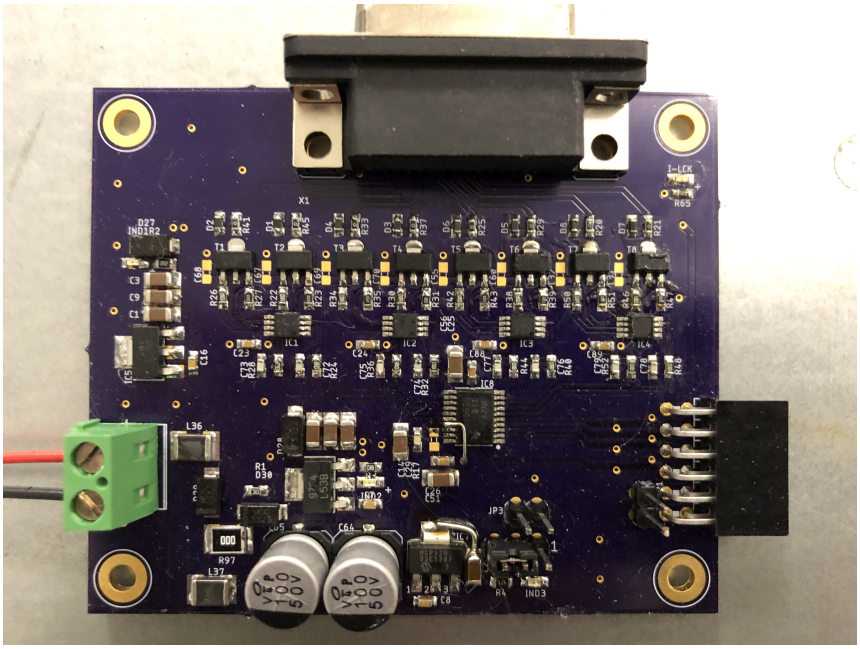
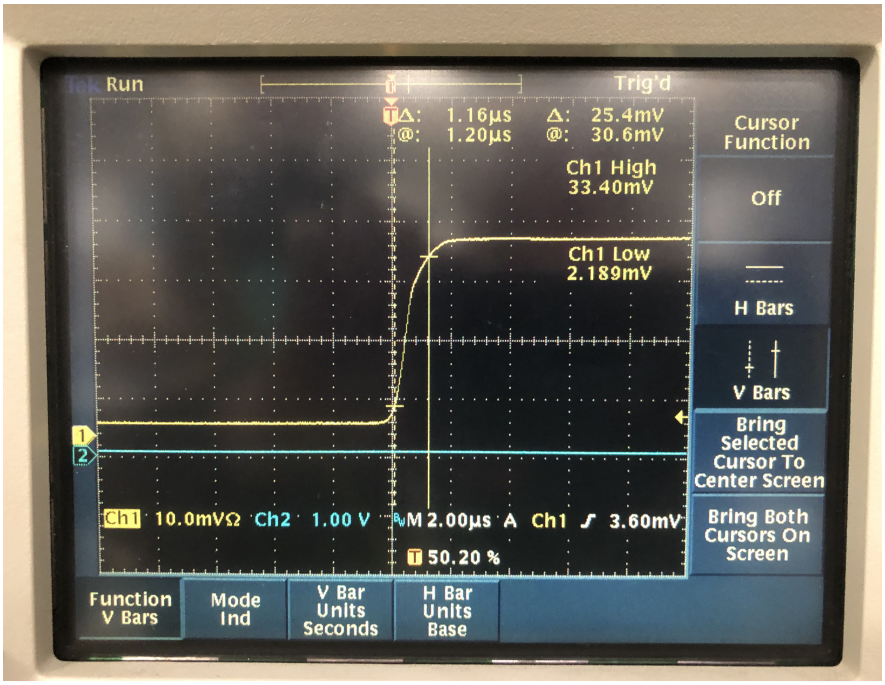


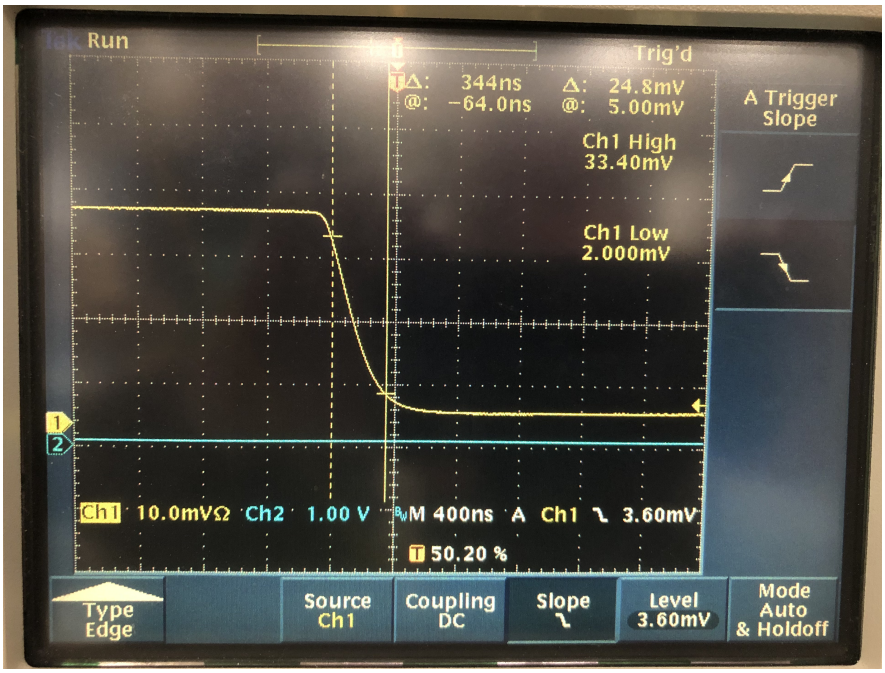
Figure 3.4: Assembled VOA board. This is a 2-layer board with 1oz copper weight fabricated on 62mil Kingboard 6167F FR4. PCB fabrication was done by OSHPark using their standard 2-layer process.

state output current as a function of steady state input voltage is exceptionally linear, the fast transitions move over an inherently nonlinear transfer function causes the rise and fall times to vary with initial and final values. It was observed that the rise and fall time are longer for larger swings, which is in agreement with how slew rate limiting and changing BJT β would affect the transitions.

Using these VOA drive boards, two BER heat maps were collected and gathered in order to confirm that the fast VOA module did indeed act only as an attenuator. The mean channel attenuation was set to 10 dB and the power offset was slewed over a small range. In one case, shown Figure 3.6a, the VOA was omitted in order to get a baseline characteristic to compare to. Then, the VOA was inserted into the measurement setup shown in Figure 2.1a after the switch. These results are shown in Figure 3.6b. The transceiver under test for this is the Mellanox MMS1C10-CM. There is a slight offset between the plots due to the difficulty of accurately equalizing power between the two channels, but the locking characteristics are nearly



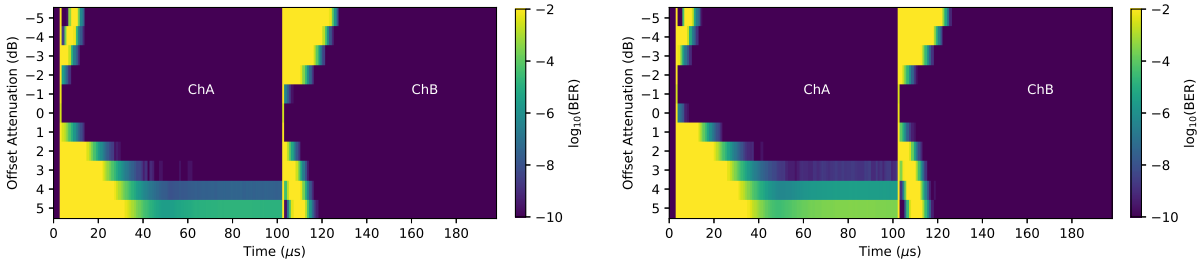
(a) Optical rise time



(b) Optical fall time

Figure 3.5: Measured optical (a) rise and (b) fall time of the VOA when slewed from 3 to 20 dB attenuation using the designed driver boards.

identical, confirming that the VOA does act only as an attenuator and does not alter the locking characteristics of a transceiver beyond the added loss.



(a) Without VOA

(b) With VOA inserted, but no equalization applied.

Figure 3.6: BER heat maps at a nominal channel attenuation of 10dB for (a) no fast VOA included and (b) with the fast VOA included.

3.2 Results

The fast VOA, driver board, and a microcontroller were then used in a modified test setup shown in Figure 2.1b. The modified setup includes the fast VOA assembly and microcontroller after the switch and before the EOM. The full setup is depicted in Figure 3.7. In this setup, the microcontroller was preloaded with the DAC codes that would equalize the two channels. The microcontroller was programmed to preload the DAC with the upcoming attenuation settings, then trigger the load DAC logic input during the switching dead times. The timing to accurately load the DAC was derived from a trigger signal supplied to the microcontroller by the FPGA. The light from the output of the EOM was measured using a fast detector instead of the module receiver. The results for non-equalized and equalized power output are shown in Figure 3.8. A dead time of 40 μs was chosen for this demonstration for clarity, but the setup is capable of equalizing the two channels for dead times of just under 5 μs.

Finally, two modules, the Mellanox MMS1C10-CM and Fiberstore QSFP+-PIR4-40G, were measured using the setup of Figure 3.7. Both modules were both measured at a line rate of 10.3 Gbps. The power imbalance between the channels inherent in the setup and in the module Tx

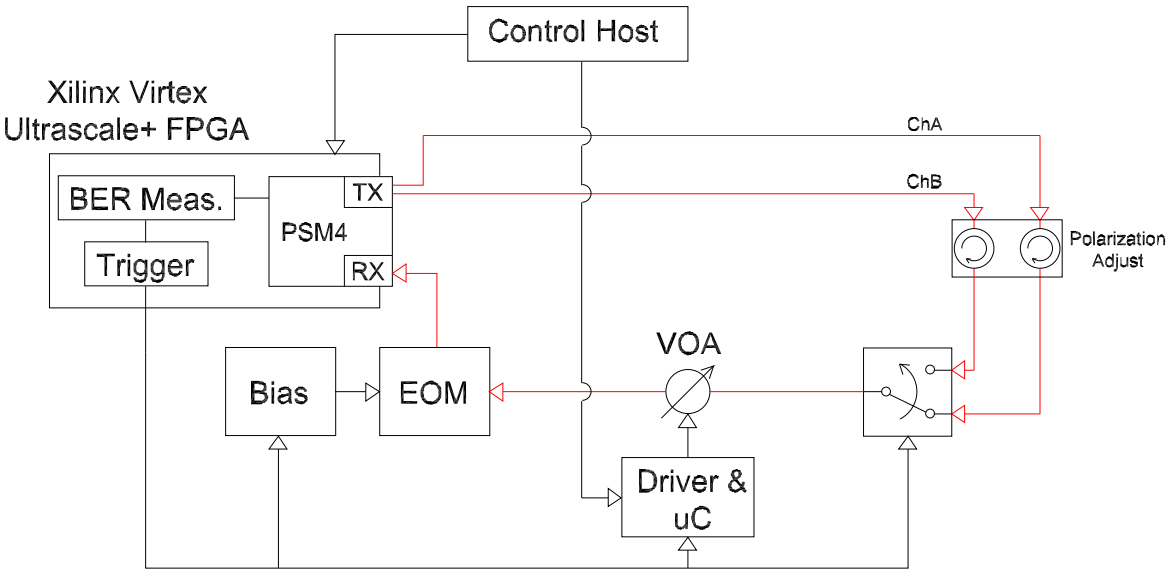
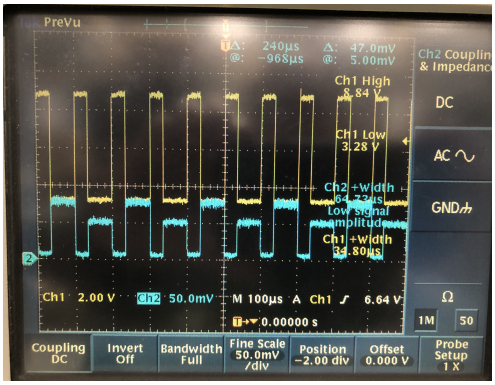
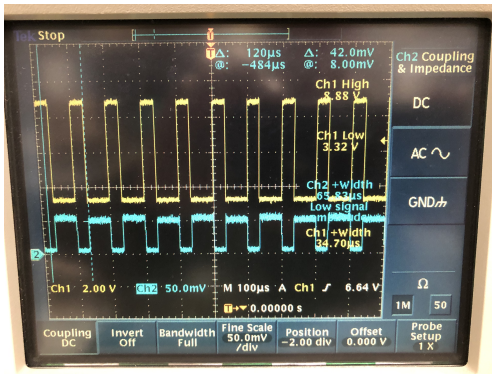


Figure 3.7: Measurement setup including the EOM to simulate switching deadtimes and the fast VOA to equalize the power during a switching event.



(a) Unequalized optical output power.



(b) Equalized optical output power.

Figure 3.8: (a) Received optical power for a switching event with no equalization applied. These are traces for a power offset of approximately 1.5dB. The yellow waveform indicates the 40 μs dead time, with "highs" representing time interval for which the optical power incident on the receiver is reduced by about 12 dB (about a factor of 16). The blue waveform shows the optical power levels at the receiver. A switching event happens during the deadtime. (b) Same as (a), but with equalization applied.

outputs caused the two measurements to have different power offsets. Because of the additional loss incorporated by the fast VOA, the achievable offset was limited and thus only a single BER measurement was recorded for no equalization and with equalization. The BER measurements

for Mellanox module for cases of no equalization and with equalization for dead times of 5 μs and 10 μs are shown in Figure 3.9. The dashed, black lines show the start and end of the dead time interval. Measurements using the Fiberstore with a slightly different power offset are shown in Figure 3.10. The lock times and lock time differences for all cases for both modules are shown in Table 3.1.

The figure show a noticeable difference between the locking characteristics of the two modules. The Fiberstore module rings during its locking procedure, whereas the Mellanox module does not. The Fiberstore module BER is seen to be ringing during the deadtime in all cases.

To understand this behavior, the optical power was measured after the BER measurement for each case of Figure 3.10. These power measurments are shown in Figure 3.11. Figure 3.11a and 3.11b correspond to the measurements of Figure 3.10a and Figure 3.11c and 3.11d correspond to the measurements of Figure 3.10b. It can be seen in all cases that the optical power was interrupted to a degree that should result in *high BER* over the duration of the dead time. This phenomena was observed in several transceivers. This effect was also seen with the Mellanox MMS1C10-CM in Figure 2.9d.

Upon further analysis, it was concluded that the effect of the error rate detector reporting a *decreasing error rate* during the deadtime is an artifact of how the bit errors are measured. [3] The BER in this setup is measured using a linear feed-forward shift register implemented on an FPGA. This method uses this register to align an incoming data stream to a known PRBS pattern and then determine errors by predicting what the next bits in the incoming sequence should be based on the known PRBS pattern. This design was chosen because a linear feed-forward shift is faster than a feedback shift register and thus can resolve bit errors on time scales as short as a few microseconds. However, a downside of this method is that it fails to report the correct BER when a long string of logical zeros are received. The reason it fails is because checking for a bit error using this method is essentially using a large collection of XOR gates. When a long string of logical zeros is received, the predicted bits also become zero, leading to no recorded

errors. Whenever the threshold of the transceiver drops far below the mean light level, the module outputs all zeros and the BER measurement algorithm falsely reports no recorded errors.

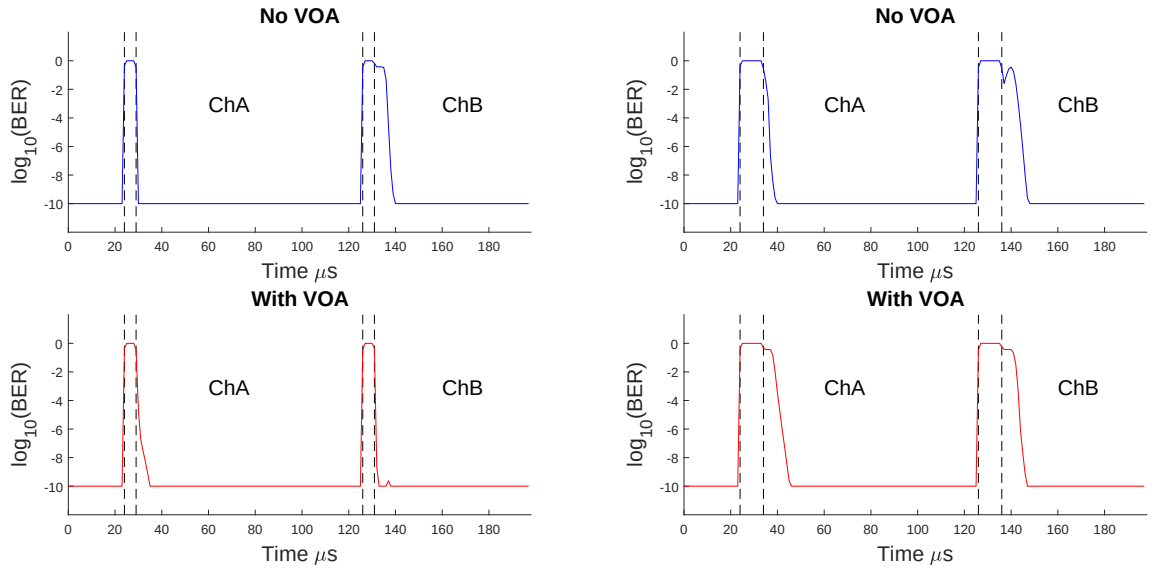
The actual dynamics of the threshold/detection circuit determines exactly how this occurs. The FiberStore module was seen to ring during its locking sequence even for small dead times, indicating that the ringing behavior is characteristic of how this module implements threshold detection. Its threshold likely rings during the longer dead times in a way that a long sequence of zeros is seen at the BER detection circuit. This leads to the BER algorithm reporting no errors when in fact the error rate approaches 1/2.

Based on previous results shown in Figure 2.9, one would expect an improvement in the lock time for a 5 μs dead time. At 10 μs however, we would expect the lock times to be longer, but have the same locking pattern for each transition. The results in Figures 3.9b and 3.10bstrongly support this. The difference in locking behavior is likely due to a decay of the transceiver's memory over time when receiving no optical power. At 10 μs of dead time, the receiver is likely forgetting where the threshold was set to a degree that it needs to re-lock after every switching event. For this case, compensating for the power offset will not improve the lock time. It can be seen, however, that the locking characteristics with equalized power levels are essentially identical for both transitions.

A potential improvement that could merit investigation is ramping the optical attenuation through or around a switching transition in a way that the threshold detection circuits could track. That way, ringing and excess resulting excess settling time could be reduced. This would require a redesign of the VOA driver board to accommodate additional filtering and pulse shaping circuits that could tailor the optical rise time wave form for optimum threshold detection transient performance.

From these results, we conclude that the shorter the switching dead time, the more likely power equalization can improve the lock time in COTS transceivers. Dynamic power equalization is a useful technique both for low dead time switches to drive down lock times and improve

system performance, as well as to make the lock characteristics between switching configurations consistent and predictable.



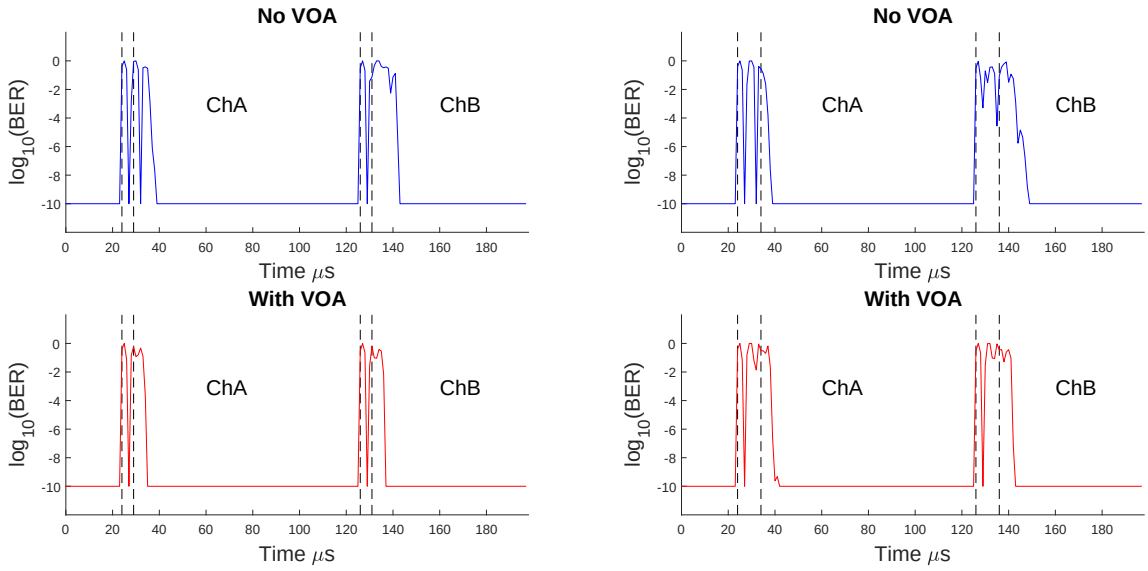
(a) 5 μs dead time

(b) 10 μs dead time.

Figure 3.9: Mellanox MMS1C10-CM locking characteristic over time with the blue trace representing no channel power equalization with the VOA and red including equalization.(a) shows locking from ChA to ChB and back with a 5 μs switching dead time and an inherent channel power offset of 1.7 dB. (b) shows the same test setup situation, except for a switching dead time of 10 μs.

Table 3.1: Extracted lock times and changes in lock time between a case of equalized and unequalized channels for dead times of 5 μs and 10 μs. Positive changes in lock time indicate the new lock time is faster than without the VOA.

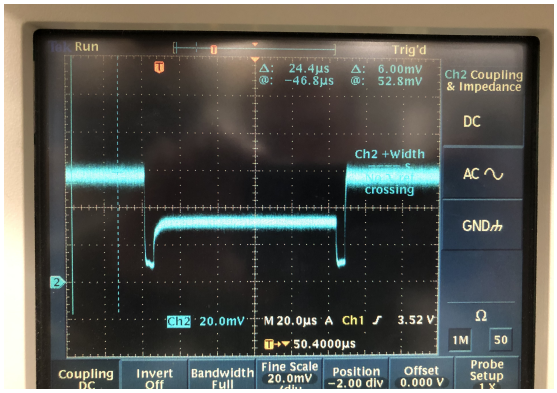
Module	Dead Time (μs)	Transition	Lock Time (no VOA) (μs)	Lock Time (VOA) (μs)	Change (μs)
MMS1C10-CM	5	B → A	1.79	4.61	-2.82
		A → B	8.25	1.92	6.33
	10	B → A	4.75	10.67	-5.92
		A → B	10.99	10.21	0.78
QSFP+-PIR4-40G	5	B → A	3.33	3.76	-0.43
		A → B	8.50	3.75	4.75
	10	B → A	5.60	8.12	-2.52
		A → B	8.68	8.51	0.17



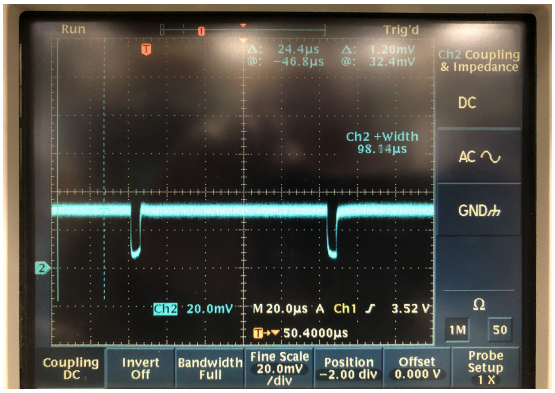
(a) 5 μs dead time

(b) 10 μs dead time

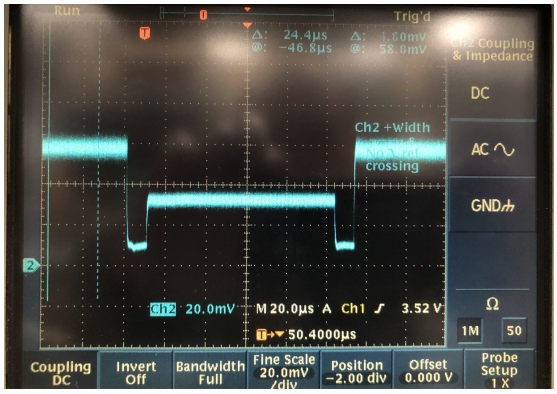
Figure 3.10: FiberStore QSFP+-PIR4-40G locking characteristic over time with the blue trace representing no channel power equalization with the VOA and red including equalization.(a) shows locking from ChA to ChB and back with a 5 μs switching dead time and an inherent channel power offset of 3 dB. (b) shows the same test setup situation, except for a switching dead time of 10 μs.



(a) 5 µs dead time, no VOA



(b) 5 µs dead time, with VOA



(c) 10 µs dead time, no VOA

(d) 10 µs dead time, with VOA

Figure 3.11: Power waveforms originating from the FiberStore QSFP+-PIR4-40G and measured from the output of the EOM using a high speed optical power detector. Each trace shows a complete interruption of the optical power during the dead time events. There is some degree of asymmetry in the rise/fall times times depending on the initial optical power level due to the driver circuit. (a) Shows the optical power waveform for a 5 µs dead time and no VOA equalization while (b) is for a 5 µs dead time with VOA equalization. (c) Shows the optical power waveform for a 10 µs dead time and no VOA equalization while (d) is for a 10 µs dead time with VOA equalization.

Chapter 4

Conclusion

Over the past decade, there has been an interest in OCS architectures for data centers. Accelerating rates of digital data generation require increasing levels of link bandwidth and and efficiency in order to maintain a smooth processing and transfer of data. Due to the unique demands of OCS compared to conventional Fat Tree architectures, several existing technologies are being evaluated. A new demand that OCS systems place on COTS components is the need for burst-mode functionality because of the nature of a physical switching event. In conventional data center architectures, optical links are established and kept alive throughout their entire lifespan. OCS architectures inherently break this link during their switching events in order to reroute data, which forces transceivers in the system to re-establish the link quickly in order to maintain a relatively efficient system. Modern datacom transceivers were not designed explicitly to have this functionality, and new technology develop would be time consuming and costly. So, several COTS transceivers of several price points, origin, and specification were subjected to tests to characterize their ability to function in an OCS environment in order to assess the technology readiness level of current COTS datacom transceivers for OCS applications.

During and around a switching event, there can be several variations in the link parameters that can lead to the transceivers taking an unacceptably long time to establish a new link. The time

that the receiver takes to establish a new link and receive data at lower than 10^{-8} bit error rate (BER) is called the "lock time". Lock time is a critical parameter in the performance of a switched datacom network since any time that the transceiver spends locking to a data stream corresponds to an effective reduction in the link line rate. The lock time of the links, as observed in this thesis, is primarily dependant on the power incident on the receiver as the receiver circuitry needs time to find a new threshold. Transceivers in general have very different locking characteristics from model to model, even within the same company. Current standards on transmitter output power are quite loose, allowing up to 3dB variation between transmitters on the same physical device. Another link variation which can be introduced by the optical switch characteristics is the effect of an optical switch's "dead time", where the dead time is defined as a period where no readable power or recognizable data is incident on a module receiver due to the physical mechanism by which the switch transitions from one connection to another. In addition to these system level variances, transceiver modules vary in basic architecture depending on the type and vendor. The combination of all these factors can create an unacceptable level variability in a switch datacom system.

It was found through measurement of several single and multimode transceivers at line rates of 10Gbps and 25Gbps that maintaining the power level between switching events from one link to another minimized the lock time for short dead times and locking characteristic uniformity. For lock times of at least 5 μs and under, equalizing power incident on the receiver for any switching configuration reduced locking times dramatically and showed a great uniformity in locking characteristics. For dead times of at least 10 μs, lock times generally increased, but locking characteristics maintained uniformity compared to a case with no equalization.

The results of this thesis show that many COTS transceivers have a potential to be used in burst-mode applications such as OCS, but careful characterization of their locking characteristics is important. As OCS technology matures and switching speeds become faster, power equalization will become more important as a means to achieve reliable system performance. A power

equalization approach can be used to avoid the cost and overhead of developing new burst-mode transceivers by leveraging the extensive existing datacom transceiver technologies.

References

[1] Psm4 multiple source agreement. http://psm4.org/.

[2] T. E. Darcie, P. F. Driessen, M. Osusky, and W. Lin. Optical network control overlay using silicon voa arrays. *IEEE Photonics Technology Letters*, 17(2):513–515, Feb 2005. ISSN 1941-0174. doi: 10.1109/LPT.2004.839788.

[3] A. Forencich. Private Communication, 2020.

[4] A. Forencich, A. C. Snoeren, G. Porter, and G. Papen. An open source 100G NIC. In *28th IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM 20)*, Fayetteville, AR, May 2020. IEEE.

[5] W. M. Mellette, R. Das, Y. Guo, R. McGuinness, A. C. Snoeren, and G. Porter. Expanding across time to deliver bandwidth efficiency and low latency. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 1–18, Santa Clara, CA, Feb. 2020. USENIX Association. ISBN 978-1-939133-13-7. URL https://www.usenix.org/conference/nsdi20/presentation/mellette.

[6] X. Z. Qiu, X. Yin, J. Verbrugghe, B. Moeneclaey, A. Vyncke, C. V. Praet, G. Torfs, J. Bauwelinck, and J. Vandewege. Fast synchronization 3r burst-mode receivers for passive optical networks. *Journal of Lightwave Technology*, 32(4):644–659, Feb 2014. ISSN 0733-8724. doi: 10.1109/JLT.2013.2285594.

[7] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai,

B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, and A. Vahdat. Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network. SIGCOMM '15. ISBN 978-1-4503-3542-3. doi: 10.1145/2785956.2787508. URL http://doi.acm.org/10.1145/2785956.2787508.

[8] J. Terada, K. Nishimura, S. Kimura, H. Katsurai, N. Yoshimoto, and Y. Ohtomo. A 10.3 gb/s burst-mode cdr using a δσ dac. *IEEE Journal of Solid-State Circuits*, 43(12):2921–2928, Dec 2008. ISSN 0018-9200. doi: 10.1109/JSSC.2008.2006229.

[9] X. Yin, J. Put, J. Verbrugghe, J. Gillis, X. Z. Qiu, J. Bauwelinck, J. Vandewege, H. G. Krimmel, and M. Achouche. A 10gb/s burst-mode tia with on-chip reset/lock cm signaling detection and limiting amplifier with a 75ns settling time. In *2012 IEEE International Solid-State Circuits Conference*, pages 416–418, Feb 2012. doi: 10.1109/ISSCC.2012.6177071.