Design of a DAC-based Cryo-CMOS 51.2 Gb/s PAM4 Wireline Transmitter

Niels Fakkel

TU Delft

Design of a DAC-based Cryo-CMOS 51.2 Gb/s PAM4 Wireline Transmitter

by

Niels Fakkel

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday October 14, 2021 at 12:00 AM.

Student number:	4483774	
Project duration:	September 1, 2020 – October 14, 2021	
Thesis committee:	dr. M. Babaie,	TU Delft, supervisor
	dr. F. Sebastiano,	TU Delft
	Prof. dr. L.C.N. de Vreede,	TU Delft

TU Delft

Abstract

With the current advancements in quantum computing moving more circuitry into the cryogenic chamber there is a need for high-speed connectivity between the cryogenic and room temperature environment. Studies conducted to date have achieved high-speed links of multiple Gb/s utilizing a single CMOS chip at room temperature, yet a Cryo-CMOS wireline transmitter addresses a new topic in high-speed cryogenic electronic design necessary for the functioning and scale-up of quantum computers. This thesis entails the design of A DAC-based Cryo-CMOS 51.2 Gb/s PAM4 Wireline Transmitter. Overall, the proposed design is meant to demonstrate a high-speed signal generated by a Cyro-CMOS chip can be send through a cable from a cryogenic environment and received at room temperature. Requirements have been set up based on the measurement of the cable channel and simulation results showed these could be met with the designed circuitry. The system consists of a low-speed 16-to-1 serializing structure, a high-speed 4-to-1 Multiplexer, and a 6-bit (4b binary, 2b unary) CML DAC. The design is finished and taped out in 40-nm technology, however the chip is still in fabrication, so the results are based on simulation data only, in future research measurements will verify the working of the chip.

Contents

Acknowledgements

I wish to express my sincere gratitude to my supervisor dr. Masoud Babaie for his guidance and advice throughout this thesis project. In addition I would like to thank my advisors Bishnu Patra and Mohsen Mortazavi for their technical help and effort. Furthermore I would like to thank the members of the coolgroup and especially the fellow master students for their support and online fun during covid times. Also I want to thank my family and parents for supporting me and keeping me motivated all year long. Finally I would like to thank dr. Fabio Sebastiano and Prof. Leonardus de Vreede for taking part in the thesis committee.

1 Introduction

1.1. Wireline Communication

In general, a wireline communication system consists of a Transmitter (TX), a propagation medium(the channel) and a Receiver (RX), see Figure 1.1. The TX employs a Multiplexer (MUX) in order to serialize parallel data streams. The Phase-Locked Loop (PLL) generates the clock necessary for the MUX. The driver should deliver sufficient voltage swing to the channel. Then an equalizer will compensate for the channel's imperfections; this can be done both on TX and RX sides. The RX site incorporates a Clock and Data Recovery (CDR)) circuit to recover the clock and retime the received data, a Demultiplexer (DMUX) then de-serializes the data back into separate streams.

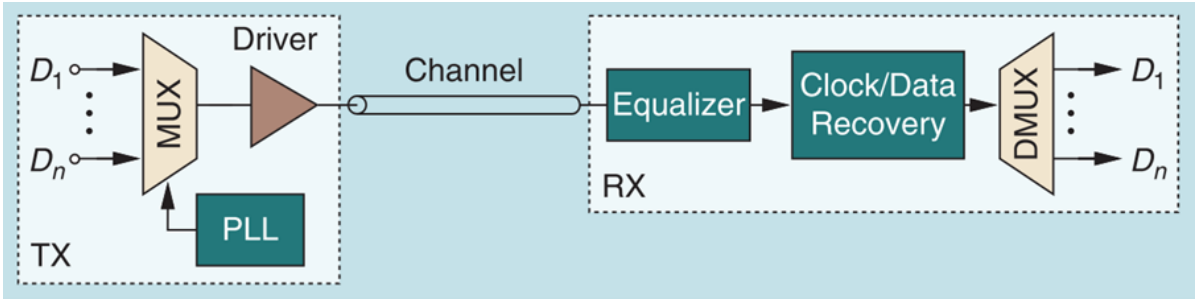


Figure 1.1: Generic wireline communication system [1].

1.1.1. Trend in Generic Wireline Communication

Wireline communication has already been around for a long time, in the 1830s "the telegraph" by Claude Chappe was used for more than one and a half centuries. Alexander Graham Bell invented the telephone in 1870, but it took up to 1960 that IBM decided to send data over this line with the fax machine. Soon after came the first Digital Subscriber Line (DSL) in 1979 and Ethernet in 1990, with a data rate of 10 Mb/s [1]. Nowadays, the speed of datelines has improved significantly both through the advancements in consumer products like graphic cards requiring high video bandwidths and commercial internet servers transferring more and more data through backplanes and cables. Soon internet cable speeds of 400 GbE with 8x 50 Gb/s per channel will become a reality, and the current state-of-the-art already reaches 224 Gb/s over a single serial link using 4-level pulse-amplitude-modulation (PAM4) [2].

1.1.2. Application to Quantum Computer

The trend in quantum computers is to move the initial processing into the cryogenic chamber, so fewer cables are required to transfer qubit information out to the processing environment. The number of quantum bits is increasing up to thousands and thereby the amount of readout information as well. To prevent a large cluster of copper cabling necessary, the idea is to serialize all information within the 4-K environment and transmit it all to the room temperature processing environment using a single

wireline system. This way, the heat losses through the amount of copper cabling can be minimized and the readout system will become more scalable. An overview of such an integrated quantum readout system is given in figure 1.2. The location of the High-speed TX is indicated in the right top.

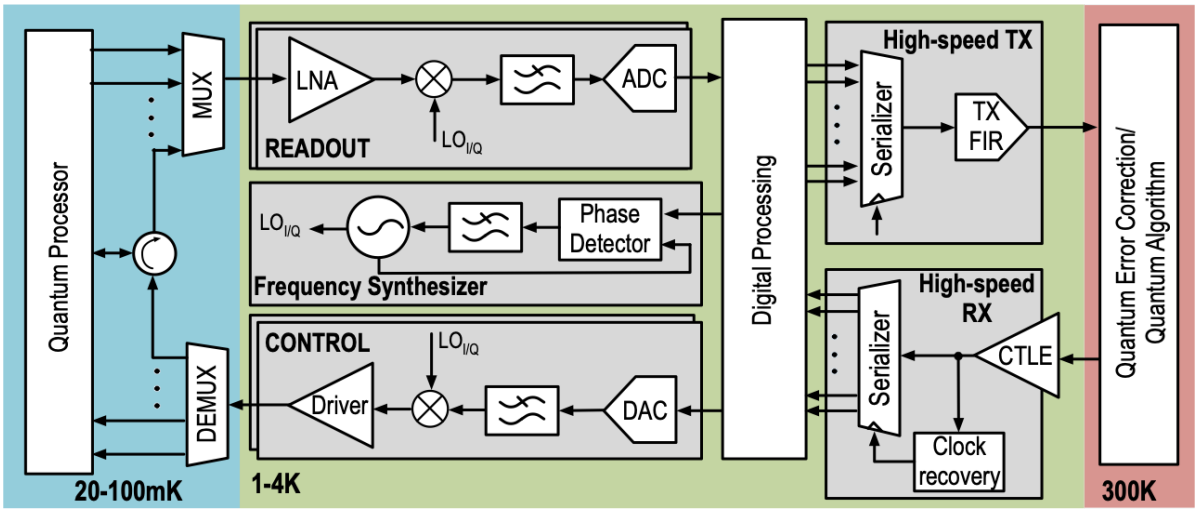


Figure 1.2: Cryogenic quantum control and readout system [3].

1.2. Challenges

Designing a high-speed wireline transmitter in CMOS technology for a cryogenic environment will have some limitations. The challenge is to transmit the highest possible data rate over the available channel without losing data integrity.

1.2.1. Technology Scaling

Technology scaling was dominated by the trend of the digital industry towards smaller size transistors. The advantage of smaller-sized transistors in digital circuitry is the reduced parasitic capacitance, decreased gate delay, increased device density, and reduced supply voltage, leading to lower power consumption [4]. Since a wireline transmitter is mostly limited by the maximum achievable speed of retimers and Multiplexers, which are mostly digital circuits, technology scaling has a good influence on the performance. The most state-of-the-art designs make use of 10-nm FinFET technology to reach extremely high speeds [2]. Although there are some challenges going down in size mainly in the ability to drive the wireline, for example, the lower supply voltage and lower breakdown voltage limit the total achievable output swing [5]. The design proposed in this thesis is made in 40-nm CMOS technology, aiming for a speed of 51.2 Gb/s PAM4 is pushing the limits of this technology.

1.2.2. Channel

Not only the transistor scaling, but also the interfacing with the technology can limit the performance. In particular, wireline communication requires a large bandwidth channel, since the transmitted signal is ideally a rectangular wave. It makes use of the complete bandwidth from DC to at least 85% of the data frequency of the signal, Table 2.1. A large loss in frequency over the complete channel means a direct degradation of the received signal. The pads in 40-nm technology introduce a parasitic capacitance in the signal path, decreasing the bandwidth, and the required bond wires will introduce a parasitic inductance in series with the signal path. Also, interfaces such as PCB traces, connectors, cable and in cryogenic applications, sealed connectors to the measurement chamber will limit the total system bandwidth.

1.2.3. Cryogenic CMOS Technology

The CMOS technology has to operate at cryogenic temperatures and preferably also at room temperature to verify functionality. This cryogenic chip design is referred to as Cryo-CMOS [6]. At cryogenic temperatures, CMOS devices will still work, however, the behaviour is different and the available device

models used in the foundry software are no longer valid. It is, therefore, difficult to design accurately, using regular room temperature models. The design has to account for some margins. One problem is the increased threshold voltage at cryogenic temperatures, in analog design, this could lead to headroom issues [7]. In digital design, the increased threshold voltage will slow down the circuitry, mainly in pass gate structures where the voltage swing is limited. Lastly, device mismatch increases at cryogenic temperatures [8], setting more stringent requirements on, for example current source sizing for the Digital-to-Analog Converter (DAC).

1.3. Thesis Objectives

The objective of this thesis is to implement a Cryo-CMOS wireline transmission system. Therefore, this thesis presents the design of A DAC-based Cryo-CMOS 51.2 Gb/s PAM4 Wireline Transmitter.

Due to the current advancements in quantum computing moving more circuitry into the cryogenic environment there is a need for high-speed connectivity between the cryogenic environment and room temperature. The proposed design will be meant to prove a high-speed signal generated by a Cyro-CMOS chip can be sent through a cable from a cryogenic environment and received at room temperature. The design will be pushing the limits of what is possible in 40-nm technology and cryogenic environment. The tape out for this design is finished, but still in production, therefore the conclusions are based on simulation data only.

1.4. Thesis Outline

The thesis is ordered as follows. First in Chapter 2 the background information for wireline transmitters is given and the requirements of the system are set. Different state-of-the-art architectures are compared and the system overview is given in Chapter 3. Following up are 4 sections that go into detail about the design of each part of the system. Chapter 4 gives an overview of the clock generation circuitry. Chapter 5 introduces the 400 Mbps retimer and latchless 16:1 Multiplexer design. Chapter 6 discusses the design of a high-speed 6.4 Gb/s retimer and 4:1 Multiplexer to reach 25.6 Gb/s. Chapter 7 will break down the design of the 6b current-steering DAC and drivers in detail. Finally a conclusion is given in Chapter 8.

1.5. Original Contributions

The main novelty in this design is the design and application of a wireline transmitter in Cryo-CMOS, and other contributions are listed below:

- Analysis of available wireline transmitter architectures Chapter 3
- Design and implementation of a low power multiplexing structure Chapter 5
- Design and implementation of a high-speed retimer and 4:1 Multiplexer Chapter 6
- Comprehensive analysis for deriving DAC parameters from wireline transmitter requirements Chapter 7
- Design and implementation of a high-speed current steering DAC Chapter 7

2

Wireline Transmission Background

In this chapter the background literature for designing a cryogenic wireline transmitter is discussed to find out the system requirements.

This chapter is structured as follows, Section 2.1 introduces three possible data formats for wireline communication. Section 2.2 shows the Power Spectral Density of these data formats and their bandwidths, Section 2.3 calculates the required nyquist frequency and signal bandwidth. All characteristics from an eye diagram are analysed in Section 2.4. The statistical calculations to find the Bit Error Rate are done in Section 2.5. Section 2.6 defines the Ratio of Level Mismatch. In Section 2.7 the channel is analysed, possible equalisation techniques are discussed in Section 2.8, then a few link simulations are shown in Section 2.9. Lastly based on all the background knowledge and definitions the requirements are set in Section 2.10.

2.1. Data Formats

There are multiple data formats which could be used to send over the channel. For high speed ($> 10\,Gb/s$) data over short ($< 100m$) channels, only a few existing solutions remain feasible, namely: non-return-to-zero (NRZ), 4-level pulse-amplitude-modulation (PAM4) and Duobinary.

2.1.1. NRZ

NRZ is the most common way to send over data and the easiest to implement. This protocol makes use of only 2 levels which are either above zero or below. However, as the data rate increases to over 20 Gb/s, also a system bandwidth higher than 20 GHz is required. In the wireline systems, both the channel and the power efficiency of the circuitry are bandwidth limited, thus demanding new data formats. However, up to the bandwidth limitation, an NRZ signal can achieve an outstanding Bit Error Rate (BER) and signal integrity, due to the simple receiver structure and the NRZ data that can be amplified without considering linearity [9].

2.1.2. PAM4

PAM4 makes use of 4 levels and hence sends 2 bits/symbol, therefore, it only requires half the system bandwidth of NRZ, which makes it favorable for high-speed links. However, this comes at a cost as the distance between the different signals decreases by 3×, there is an inherent 9.5 dB SNR loss. To compensate this loss in SNR a larger output swing is required. In the receiver front-end, however, the PAM4 signal may need linear amplification to correctly receive the different levels. Moreover, the data is more difficult to recover, and retiming flipflops are mandatory, thus complicating clock recovery and costing high power consumption. In order to get similar BER performance as NRZ also Forward Error Correction might be required.

2.1.3. Duobinary

Duobinary modulation can theoretically achieve a data rate twice the channel bandwidth by introducing Inter-symbol Interference (ISI) in a controlled matter such that it can be canceled at the receiver. An advantage of this method is that by already putting a delay at the transmitter, the protocol inherently

incorporates the channel loss as part of the overall response[10]. This way, the protocol already accounts for the channel loss by a part and thus requires less equalization (amplification). Unfortunately, the complexity of the transceiver architecture tends to limit the BER performance in reality.

2.2. Power Spectral Density

The most limiting factor in high-speed wireline communication is often the bandwidth of the channel. A common measure for expressing the required bandwidth is Power Spectral Density (PSD). The PSD can be calculated in different ways depending on the form of the signal. For a rectangular wave, the PSD of the three data formats can be expressed as follows [11]:

$$\begin{aligned} \mathrm{PSD}_{\mathrm{NRZ}} &= \left| \frac{\sin\left(\pi f T_{\mathrm{Snr}}\right)}{\pi f T_{\mathrm{Snrz}}} \right|^2 \\ \mathrm{PSD}_{\mathrm{PAM4}} &= \left| \frac{\sin\left(\pi f T_{\mathrm{Spam4}}\right)}{\pi f T_{\mathrm{Spam}}4} \right|^2, \quad T_{\mathrm{Spam4}} = 2T_{\mathrm{Snrz}} \\ \mathrm{PSD}_{\mathrm{DB}} &= \left| \frac{\sin\left(\pi f 2T_{\mathrm{sdb}}\right)}{\pi f 2T_{\mathrm{db}}} \right|^2, \end{aligned} \tag{2.1}$$

where T_{Snrz} is the symbol time of an NRZ waveform, which can be calculated as the inverse of the symbol rate $T_{\mathrm{Snrz}} = 1/R_{\mathrm{Snrz}}$. Since PAM4 has twice as many bits per symbol the symbol time will be equal to $T_{\mathrm{Spam4}} = 2/R_{\mathrm{Snrz}}$. A more natural representation of the PSD is assuming it will have a Raised-Cosine Pulse Shape. For a roll-off factor of $\beta = 1$ the PSD can be expressed as:

$$\begin{aligned} \mathrm{PSD}_{\mathrm{NRZ}} &= \begin{cases} \left|\frac{1}{2}\left(1+\cos\left(\pi f T_{\mathrm{Snrz}}\right)\right)\right|^2 & \text{if } f \le \frac{f_{\mathrm{bitrate}}}{2} \\ 0 & \text{if } f > \frac{f_{\mathrm{bitrate}}}{2} \end{cases}, \\ \mathrm{PSD}_{PAM4} &= \begin{cases} \left|\frac{1}{2}\left(1+\cos\left(\pi f T_{\mathrm{Spam}\ 4}\right)\right)\right|^2 & \text{if } f \le \frac{f_{\mathrm{bit\ rate}}}{4} \\ 0 & \text{if } f > \frac{f_{\mathrm{bit\ rate}}}{4} \end{cases}, \\ \mathrm{PSD}_{\mathrm{DB}} &= \begin{cases} \left|\frac{1}{2}\left(1+\cos\left(\pi f T_{S\mathrm{db}}\right)\right)\cos\left(\pi f T_{\mathrm{Sdb}}\right)\right|^2 & \text{if } f \le \frac{f_{\mathrm{bit\ rate}}}{2} \\ 0 & \text{if } f > \frac{f_{\mathrm{bit\ rate}}}{2} \end{cases} \end{aligned} \tag{2.2}$$

The last approach to the PSD is assuming it will have a Sinc Pulse form which can be expressed as Raised-cosine with $\beta = 0$:

$$\begin{aligned} \mathrm{PSD}_{\mathrm{NRZ}} &= \begin{cases} 1 \text{ if } f < \frac{f_{\mathrm{bit\ rate}}}{2} \\ 0 \text{ if } f > \frac{f_{\mathrm{bit\ rate}}}{2} \end{cases} \\ \mathrm{PSD}_{\mathrm{PAM4}} &= \begin{cases} 1 \text{ if } f < \frac{f_{\mathrm{bit\ rate}}}{4} \\ 0 \text{ if } f > \frac{f_{\mathrm{bit\ rate}}}{4} \end{cases} \\ \mathrm{PSD}_{\ \mathrm{DB}} &= \begin{cases} \left|\frac{1}{2}\left(1+\cos\left(\pi f T_{\mathrm{sd}\ b}\right)\right)\right|^2 & \text{if } f \le \frac{f_{\mathrm{bit\ rate}}}{2} \\ 0 & \text{if } f < \frac{f_{\mathrm{bit\ rate}}}{2} \end{cases} \end{aligned} \tag{2.3}$$

The resulting spectra for all levels are plotted in figure 2.1.

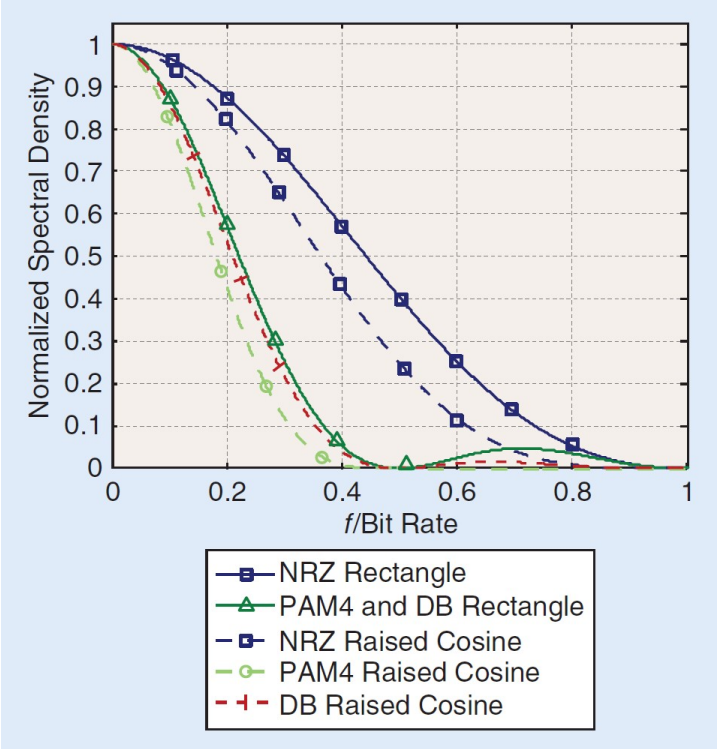


Figure 2.1: The PSD of NRZ, PAM4, and duobinary, with rectangular and a raised-cosine pulse shapes (roll-off $\beta = 1$) [11].

Quantifying the minimum required bandwidth from the PSD can be done by selecting the frequency at which the cumulative spectral power contains 90% of the bit energy. This point can be found by integrating equations (2.1), (2.2), (2.3) and setting each equal to 90%. This results in the percentages showing how concentrated the data format is to DC, see table 2.1. From the different pulse shapes in this table can be concluded PAM4 is most concentrated around DC, then Duobinary and last NRZ. So effectively PAM4 will require the smallest bandwidth in order to operate.

Pulse Shape	NRZ ($\%f_b$)	Duobinary ($\%f_b$)	PAM4 ($\%f_b$)
Rectangle	85	43	43
Raised cosine ($\beta = 1$)	47	28	23
Sinc	45	30	23

Table 2.1: A comparison of the required bandwidth of 90% energy for NRZ, duobinary, and PAM4 with respect to the bit rate f_b.

2.3. Nyquist Frequency

A better way to compare the different modulation data formats is the Nyquist frequency (f_N), defined from sampling theory as the highest frequency minimally required in the waveform to receive the transmitted stream without unwanted Inter-symbol Interference. An approach related to the signal bandwidth can be made by taking the fastest single frequency component in the signal necessary to transmit the data pattern [11]. An overview of all highest possible eye heights by means of fundamental sine waves at Nyquist is shown in Figure 2.2. For NRZ, this is at half the bit rate $f_N = f_b/2$ and for PAM4 at quarter the bit rate $f_N = f_b/4$. For Duobinary, the signal theoretically is sampled the same rate as NRZ, however, when looking at the fundamental tone required effectively duobinary equals one-third of the bit rate $f_N = f_b/3$. Another thing noticeable is that both NRZ and PAM4 make use of the full signal swing, while duobinary essentially only uses two-thirds.

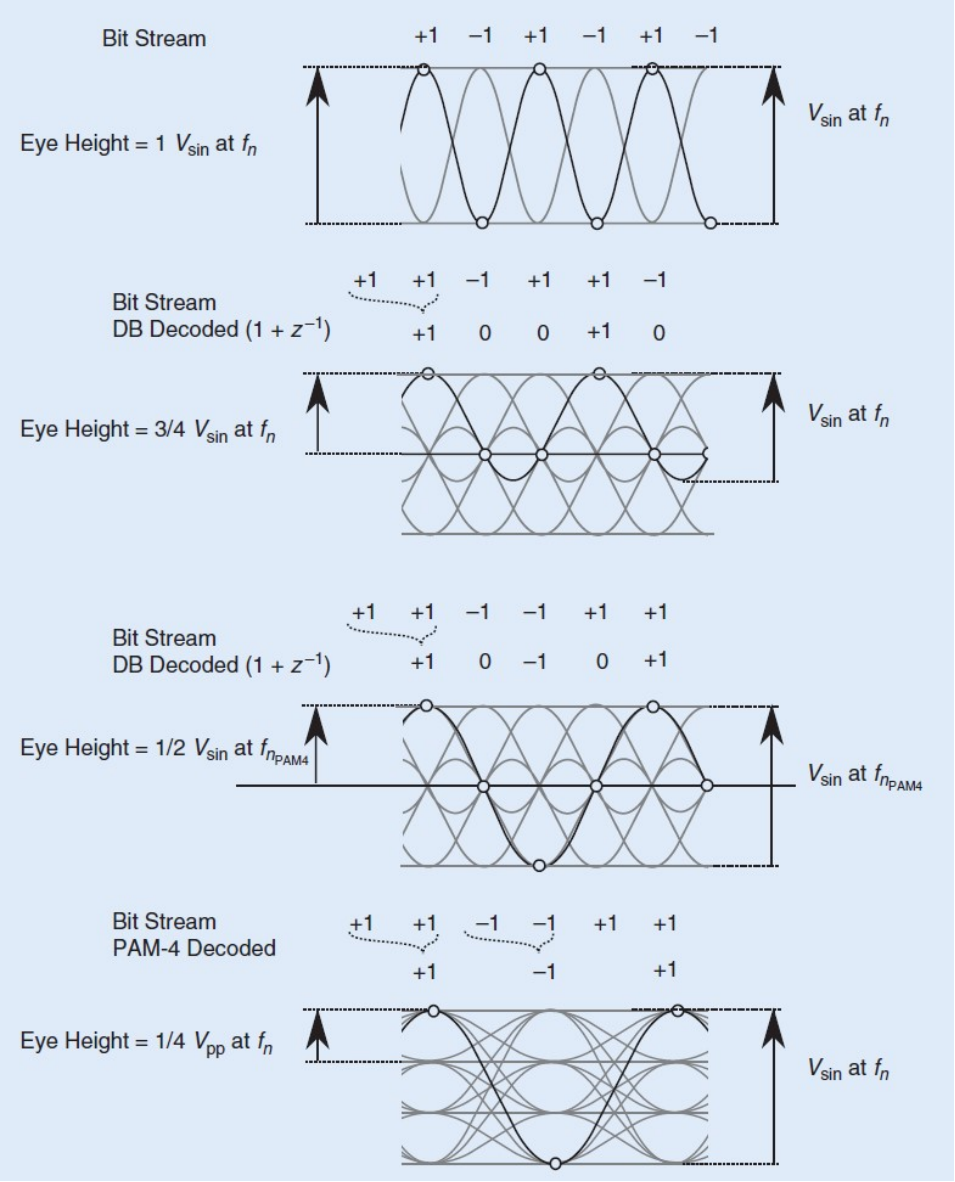


Figure 2.2: A comparison of the highest possible eye heights by means of fundamental sine waves at the Nyquist frequencies that occur with specific data patterns [11].

2.4. Eye Diagram

A common way to characterize a wireline transmitter is by using an Eye Diagram to estimate transmitter parameters, including noise, jitter, and BER. An eye diagram is constructed by overlapping all different bit transitions over time into a single image. An example of the construction of the NRZ Eye Diagram is shown in Figure 2.3. The width of one eye is 1 Unit Interval (1 UI) and is equal to the duration of one single symbol.

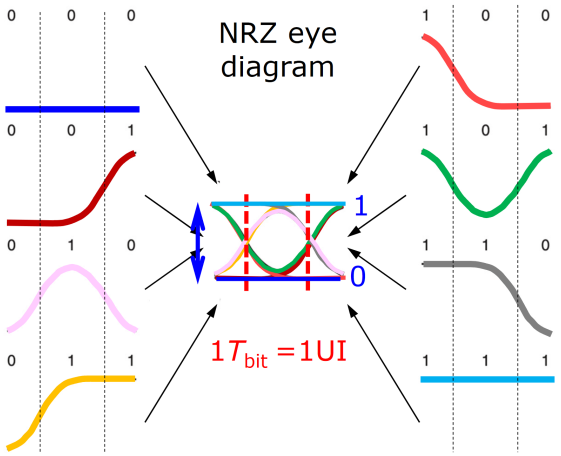


Figure 2.3: NRZ eye construction.

Figure 2.4 shows an example of a measured NRZ eye diagram at the transmitter side. The quality of the eye can be defined by the eye opening, both vertically (in amplitude) and horizontally(in time). At the receiver side, data points ideally must be sampled in the midpoint of the eye. Hence, the larger the eye opening, the bigger the chance of sampling the correct bit. The peak-to-peak swing of this figure is the eye amplitude and can be expressed as:

$$V_{pp} = V_{top} - V_{bot} \tag{2.4}$$

However the signal is not ideal and there will be some noise present in both amplitude and time domain. This could cause a bit to be received erroneously at the receiver side. In order to calculate the chance of a bit error occurring, the noise can be represented as a Gaussian distribution with zero mean and written as the probability density function (PDF) of n(t):

$$P_n = \frac{1}{\sigma_n \sqrt{2\pi}} \exp \frac{-n^2}{2\sigma_n^2}, \tag{2.5}$$

where σ_n is the rms noise. The Gaussian noise amplitude is represented as σ_{top} and σ_{bot} in Figure 2.4.

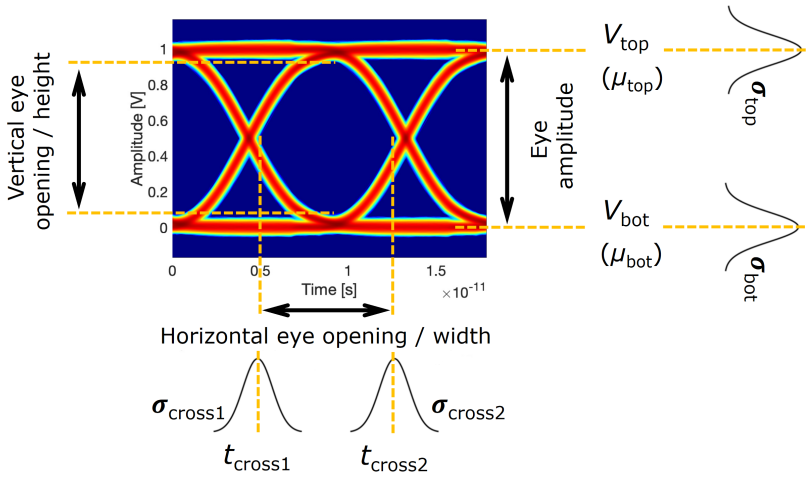


Figure 2.4: NRZ eye diagram characteristics

2.4.1. Vertical Eye Opening

The vertical eye opening of a data waveform is a good measure to display the influence of crosstalk and noise. For different modulation data formats, there can be a penalty in eye height and thereby directly

the SNR of the system. In order to calculate the penalty, one can look at the maximum amplitude of the sinewave at the Nyquist frequency occurring in the bit stream, see Figure 2.2. The penalty is then the ratio between this maximum amplitude at the Nyquist frequency and the eye height, which can be extracted from the transfer function [11].

For NRZ, the penalty is 0 dB since the vertical eye height is equal to the maximum voltage swing of the sine wave at Nyquist. PAM4, however, has a ratio of one-third of the maximum signal swing, resulting in a penalty of $20log(1/3) = -9.5\,dB$. For duobinary, the Nyquist frequency component has a smaller amplitude than the total voltage swing, as shown in Figure 2.2. Actually, the loss presented in the channel is used to partially generate the duobinary waveform meaning it is not necessary to compensate for a certain amount of loss. Therefore, one could say the penalty of duobinary is equal to $20log(3/4) = -2.5dB$. The full-swing frequency component of duobinary is however equal to that of PAM4. So, duobinary will have a similar relative reduction in eye height as PAM4, and initially will start with a penalty of -6 dB. A comparison of the three data formats with the maximum eye height as a function of the channel loss at half the bit rate is plotted in Figure 2.5.

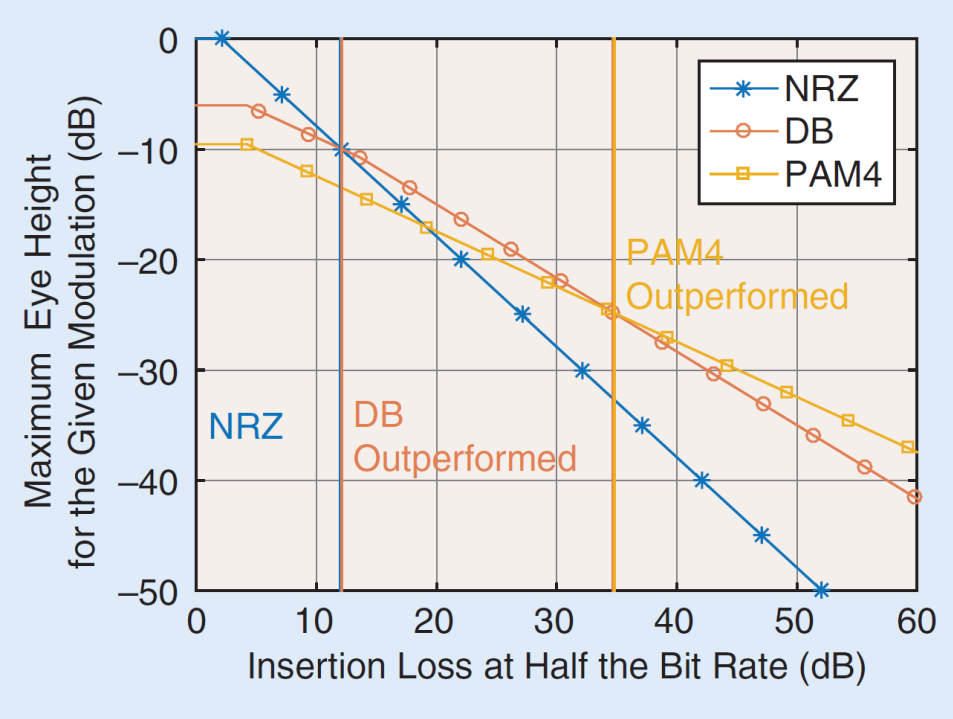


Figure 2.5: The maximum eye height as a function of the channel loss at half the bit rate, assuming the channel has a linear (on a log scale) loss profile [11].

When the channel loss at half the bit rate increases to 12 dB, duobinary outperforms NRZ in terms of eye height. A further increase of loss to 34 dB results in larger eyes for PAM4.

2.4.2. Horizontal Eye Opening

The last important measure is the Horizontal eye opening, since a bit error can also occur in the time domain. The width of the eye depends on the data format, channel bandwidth, channel skew, transmitter jitter and jitter from the CDR. The UI of PAM4 is inherently twice that of NRZ since it transmits two bits per symbol. However, this is not true in reality, due to the multilevel modulation format, the actual maximum PAM4 eye width is 0.66 UI. Figure 2.6 shows the eye widths of all three data formats. Comparing the Horizontal Eye opening based on the maximum eye width is not a completely fair comparison, because the Signal-to-Noise Ratio (SNR) at the edges might be too low for the receiver to detect. Therefore, middle eye width is a better representation, since this is located at the threshold where the receiver will slice the bit. This results in a middle eye width of 1 UI for NRZ, 0.54 UI for PAM4, 0.66 UI for duobinary due to it is diamond shape.
Deterministic jitter and random jitter from the spectrum of the transmitted signal often scale inversely

with the baud rate. Hence, a PAM4 signal will have inherently more absolute jitter than NRZ or duobinary at the same bit rate. Also, the amount of possible transitions can cause jitter or worse phase offsets in the CDR. PAM4 has 16 different transitions of which four have a skewed crossing which is more difficult for the CDR in the receiver [12]. Duobinary has 7 transitions, also requiring advanced CDR techniques, while NRZ uses only 4 transitions with a wide variety of available CDRs systems available.

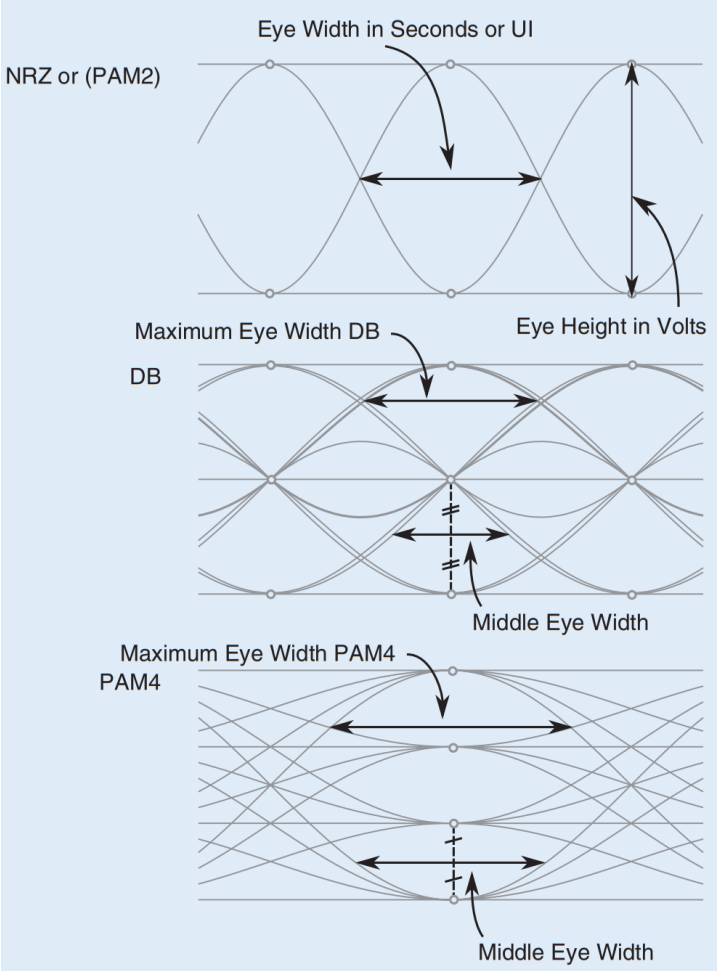


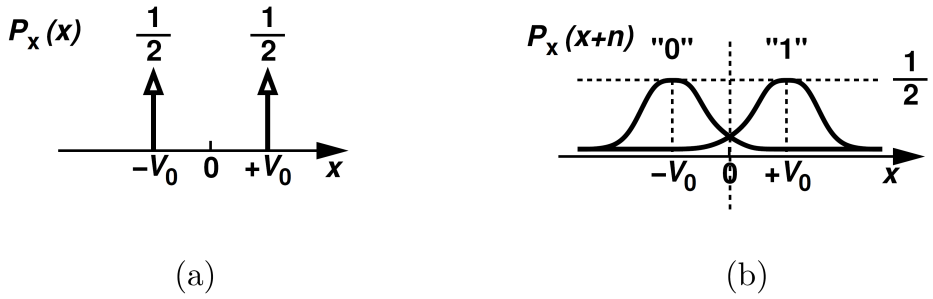
Figure 2.6: The construction of an eye diagram and the different definitions for eye height and eye width [11].

2.5. BER

The measure to define how many bits of the wireline system get received erroneously is the Bit Error Rate (BER). Looking in the vertical domain of the eye diagram, an approximation of the BER can be calculated for both NRZ and PAM4.

2.5.1. BER of NRZ

Assuming an NRZ signal x(t), with data consisting of both ONE and ZERO bits in equal probabilities, the PDF of a signal without noise can be represented as a combination of two pulses at $x = -V_0$ and $x = +V_0$ having each a weight of 1/2, see figure 2.7(a). With adding the additive noise, the PDF of n(t) and x(t) convolve with a combination of two Gaussian distributions as shown in Figure 2.7(b).



(a) (b)

Figure 2.7: PDF of (a) noiseless NRZ signal, and (b) noisy NRZ signal [13].

The BER is the probability that the bottom **"0"** bit $-V_0 + n(t)$ falls in the region above the threshold value referenced as 0 in 2.7(b). In addition to the probability of the top **"1"** bit $V_0 + n(t)$ falling below 0.

The probability of **"0"** falling in the **"1"** zone can be expressed as:

$$P_{0\to1} = \frac{1}{2}\int_{0}^{+\infty} \frac{1}{\sigma_n\sqrt{2\pi}} \exp\frac{-(\mu+V_0)^2}{2\sigma_n^2} d\mu \tag{2.6}$$

The probability of **"1"** falling in the **"0"** zone:

$$P_{1\to0} = \frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sigma_n\sqrt{2\pi}} \exp\frac{-(\mu-V_0)^2}{2\sigma_n^2} d\mu \tag{2.7}$$

Then using the Q-function defined as:

$$Q(x) = \int_{x}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\frac{-\mu^2}{2} d\mu. \tag{2.8}$$

Both probabilities can be rewritten as:

$$P_{tot} = P_{0\to1} + P_{1\to0} = \frac{1}{2}Q\left(\frac{V_0}{\sigma_n}\right) + \frac{1}{2}Q\left(\frac{V_0}{\sigma_n}\right) = Q\left(\frac{V_0}{\sigma_n}\right) \tag{2.9}$$

The BER of the NRZ signal is then equal to this total probability.

$$BER_{NRZ} = Q\left(\frac{V_0}{\sigma_n}\right) = Q\left(\frac{V_{pp}}{2\sigma_n}\right) \tag{2.10}$$

2.5.2. BER of PAM4

The BER of a PAM4 signal can be represented in a similar fashion as with NRZ, only now there are 4 possible levels that occur in equal probabilities with a weight of 1/4. The representation of a noiseless PAM4 signal with signal levels at $-V_0, -V_0/3, +V_0/3, +V_0$ and thresholds at $-\frac{2}{3}{}_0, 0, +\frac{2}{3}V_0$ are shown in Figure 2.8a. If noise is introduced on all 4 levels, it will result in a plot with 4 Gaussian distributions. For simplicity, only a binary encoding is chosen, meaning the bit order of levels is **"00"**, **"01"**,**"10"**,**"11"**, the noisy version is plotted in 2.8b.

(a) noiseless PAM4 PDF. (b) noisy binary code PAM4 PDF.

Figure 2.8: PDF of PAM4 signal

Since for PAM4 there are 2 bits per symbol, multiple bit errors could occur in one UI. Let us first look at the probability that a bit error occurs when **"00"** should be detected. This would mean either a **"01"** or **"10"** is received causing 1 bit error, or **"11"** causing 2 bit errors. The bit error probability can be written as:

$$\begin{aligned} P_{b,00\to others} &= P_{00\to 01} + P_{00\to 10} + 2P_{00\to 11} \\ &= P_{00\to others} + P_{00\to 11} \\ &= \frac{1}{4}Q\left(\frac{V_0}{3\sigma_n}\right) + \frac{1}{4}Q\left(\frac{5V_0}{3\sigma_n}\right). \end{aligned} \tag{2.11}$$

Then the probability symbol **"01"** is received mistakenly has a bit error probability of

$$\begin{aligned} P_{b,01\to others} &= P_{01\to 00} + 2P_{01\to 10} + P_{01\to 11} \\ &= P_{00\to others} + P_{01\to 10} \\ &= 2\times\frac{1}{4}Q\left(\frac{V_0}{3\sigma_n}\right) + \frac{1}{4}Q\left(\frac{V_0}{3\sigma_n}\right) - \frac{1}{4}Q\left(\frac{V_0}{\sigma_n}\right) \\ &= \frac{1}{2}Q\left(\frac{V_0}{3\sigma_n}\right) + \frac{1}{4}Q\left(\frac{V_0}{3\sigma_n}\right) - \frac{1}{4}Q\left(\frac{V_0}{\sigma_n}\right). \end{aligned} \tag{2.12}$$

Due to symmetry in the PDF the probabilities of $P_{b,01\to others} = P_{b,10\to others}$ and $P_{b,00\to others} = P_{b,11\to others}$ are equal. Thus, the total bit error probability is:

$$\begin{aligned} P_{b,tot} &= P_{b,00\to \text{ others}} + P_{b,01\to \text{ others}} + P_{b,10\to others} + P_{b,11\to \text{ others}} \\ &= \frac{3}{2}Q\left(\frac{V_0}{3\sigma_n}\right) + \frac{1}{2}\left[Q\left(\frac{5V_0}{3\sigma_n}\right) + Q\left(\frac{V_0}{3\sigma_n}\right) - Q\left(\frac{V_0}{\sigma_n}\right)\right] \\ &= 2Q\left(\frac{V_0}{3\sigma_n}\right) - \frac{1}{2}\left[Q\left(\frac{V_0}{\sigma_n}\right) - Q\left(\frac{5V_0}{3\sigma_n}\right)\right] \end{aligned} \tag{2.13}$$

Ignoring the small last term the BER can be written as:

$$\begin{aligned} BER_{PAM4} = P_{b,tot} &\approx 2Q\left(\frac{V_0}{3\sigma_n}\right) \\ &= 2Q\left(\frac{V_{pp}}{6\sigma_n}\right) \end{aligned} \tag{2.14}$$

The Q-function is a monotonic decreasing function and can be plotted for both NRZ and PAM4, see Figure 2.9. Since the term inside the Q function decides the decay, PAM4 requires a much larger swing to reach similar BER levels as NRZ. However, the maximum swing is very limited on chip due to hardware headrooms. This is the reason the BER standard [14] is lower for PAM4 at 10^{-6} compared to NRZ at 10^{-15}. The BER shortcoming can be compensated by implementing FEC.

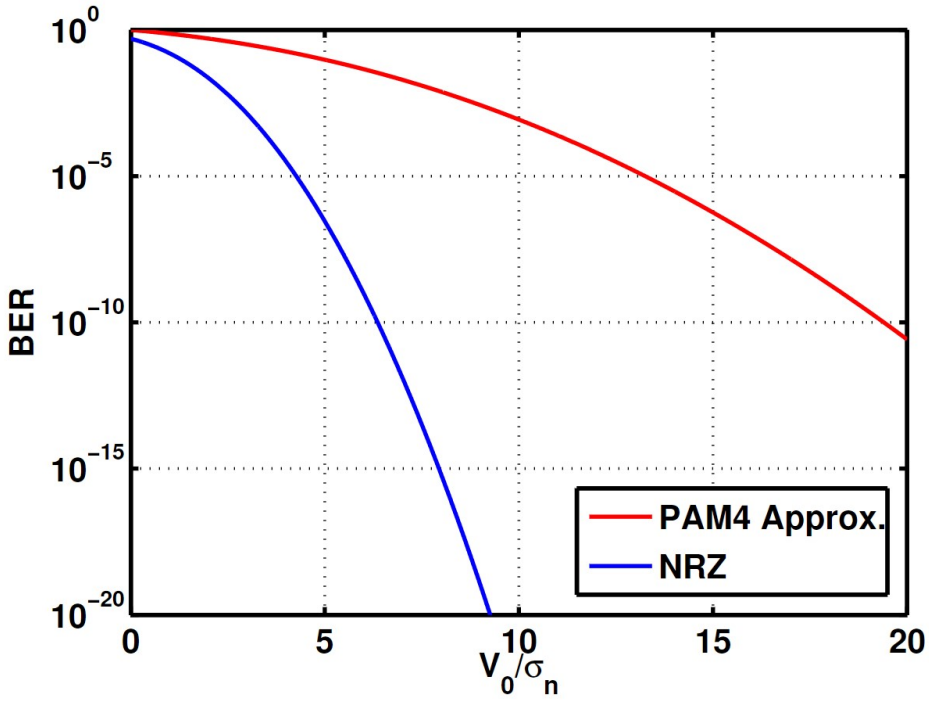


Figure 2.9: BER of PAM4 and NRZ [13].

2.6. RLM

For data formats with more than 2 levels like PAM4 the linearity of the system will have a large influence on the reception of these distinct levels. One measure to evaluate the linearity of the waveform is the Ratio of Level Mismatch. This is a ratio between the minimum signal level and total swing. For the minimum signal level or the smallest swing between adjacent symbol levels. Looking at the PAM4 eye diagram in figure 2.10 the minimum signal level S_{min} can be calculated taking the minimum of the differences between the 4 distinct levels.

$$S_{min} = min(V_3 - V_2, V_2 - V_1, V_1 - V_0) \tag{2.15}$$

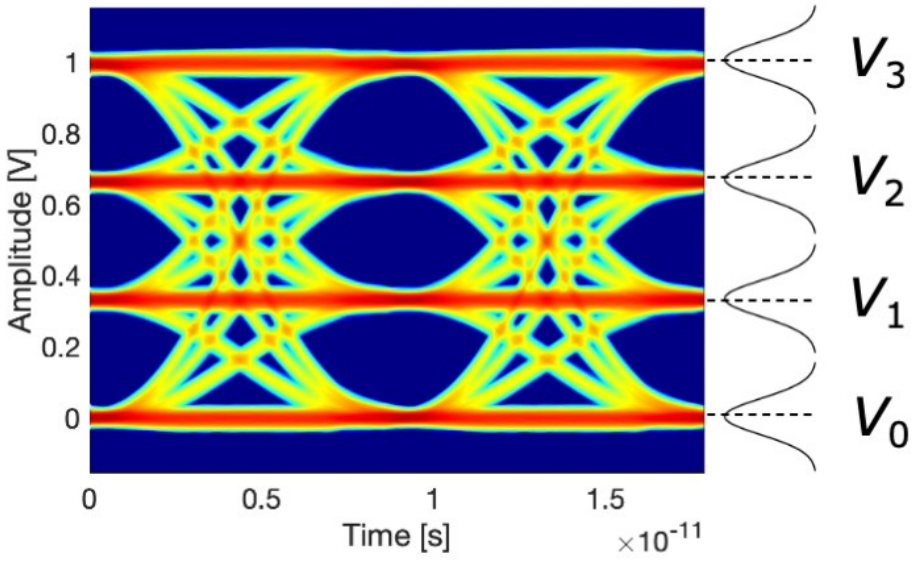


Figure 2.10: PAM4 eye diagram with distinct levels

Then dividing the minimum swing S_{min} by the total swing will give the RLM:

$$RLM = \frac{3 \cdot S_{min}}{V_3 - V_0} \tag{2.16}$$

The long-range PAM4 standard, as stated in the Common Electrical I/O (CEI) agreement requires an RLM of at least 0.95 [14].

2.7. Channel

The channel is the full path between the TX chip and the RX. This includes contemporary backplane materials, connectors and the wire itself. The path has a low-pass behavior and thus cannot provide sufficient bandwidth for high-speed data transmissions, encouraging research on signal processing and/or data coding to overcome the poor channel properties. When the losses at higher frequencies are very high, using a PAM4 signal would be more viable than NRZ since the required bandwidth is halved.

2.7.1. Channel analysis

Copper traces on FR4 substrates suffer from both skin effect and dielectric loss, thus leading to the existence of reflections in the channel and attenuation at high frequencies. These microstrip lines can be modeled to match the characteristic impedance to minimize reflections and be routed on top of the board to minimize the skin effect. However, the loss is still significant and needs to be considered in the design.

Connectors differ a lot in size and characteristics, and each type will have a different frequency response. This should also be considered in the design. When cascading the S-parameter chain with a connector having a different cut-off frequency and scattering parameters than the other traces, this complicates the overall behavior of the channel.

Lastly, the actual transmission line or cable is not ideal. For example, let us take a section of RG-58/U coaxial cable and measure its characteristics per meter: R = 0.354 Ω/m, C = 102.4 pF/m, L = 252 nH/m [15].

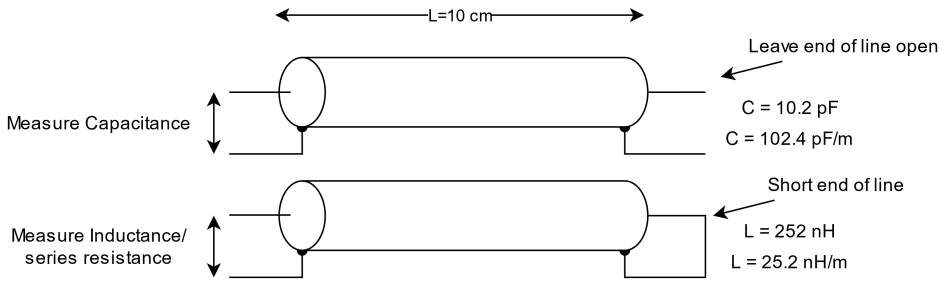


Figure 2.11: Measuring RG-58/U cable characteristics[15]

Assuming the transmission line is ideal and has no resistive loss, it will still have a characteristic impedance of $Z_0 = (252nH/102.4pF)^{1/2} = 50\Omega$, and hence behave as a 50 Ω load. However, in reality, the transmission line will be lossy as it has a finite series resistance. This will add both attenuation(loss) and distortion to the ideal transmission line model and is frequency dependent.

$$H_x(\omega) = e^{-X[(R+j\omega L)(G+j\omega C)]^{\frac{1}{2}}}, \tag{2.17}$$

Where X = cable length [m], R = series Resistance[Ω/m], L = series inductance[H/m], C = parallel capacitance[F/m], G =parallel conductance [mhos/m] and H = complex function of the transmission line. However, at higher frequencies, other effects start to play a role. The most critical is the skin effect, which causes more attenuation (more loss) at higher frequencies. The reason for the skin effect is because as the frequency increases, the currents start to flow more near the surface, effectively reducing the area of copper and thus increasing the series resistance.

Another physical phenomenon is the proximity effect which causes opposing currents in adjacent wires to draw toward one another. At high frequencies, the proximity effect will also redistribute the current density to the inside, causing an effective larger effective resistance.
Lastly, dielectric loss could play a role at higher frequencies, which is the loss in heat absorbed by a dielectric material in the presence of changing electric fields. At high-frequency designs, it might therefore be preferred to use ceramic substrates like alumina over FR-4.
The effective resistance can now be recalculated assuming these effects.

$$R(f) = \sqrt{R_{DC}^2 + R_{AC}^2} = \sqrt{R_{DC}^2 + \frac{(2.16 \cdot 10^{-7}) \cdot (f \cdot p_r)}{\pi D}} \tag{2.18}$$

Filling this resistance back into the transmission loss equation $-X[(R + j\omega L)(G + j\omega C)]^{\frac{1}{2}}$ will give a frequency response in propagation coefficient shown in Figure 2.12.

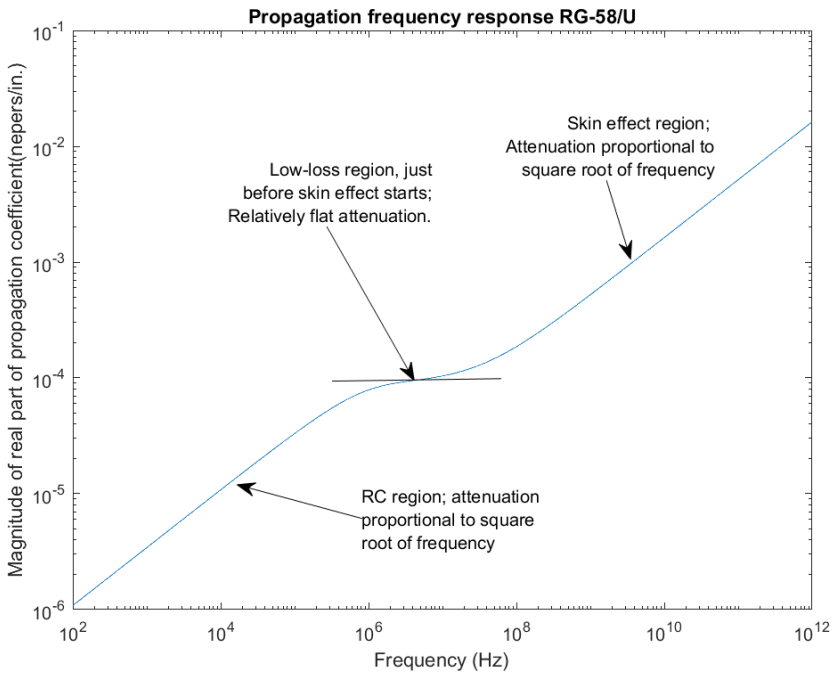


Figure 2.12: Propagation Coefficient RG-58/U cable

At high frequencies, the skin effect is the most limiting factor in the frequency response. The skin effect is dependent on the cable length X, which means the bandwidth of the channel is seriously degraded at longer cable lengths. In order to visualize this, the calculated frequency response for a number of cable lengths is plotted in Figure 2.13.

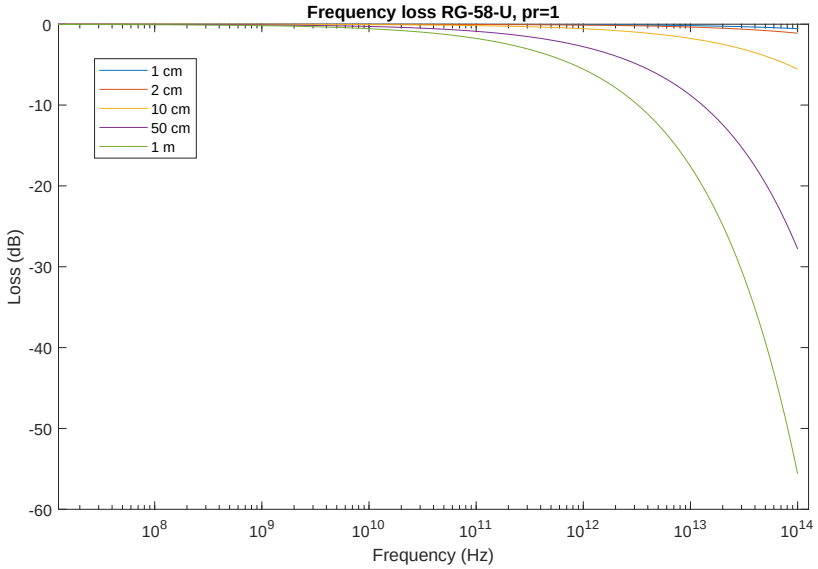


Figure 2.13: Frequency response RG-58/U cable pr=1.

Since at maximum 0.5 dB loss is permitted at the knee frequency of the data, for very high data rates over longer length channels, the loss is inevitable. This effect must therefore actively be canceled using equalization techniques. An interesting point for the application between a cryogenic environment and room temperature is that the cable will most likely also be cooled down. Assuming pure copper is used in the transmission line, this means at 4-K the copper resistivity factor pr will decrease by a factor of 100 compared to room temperature, see Figure 2.14.

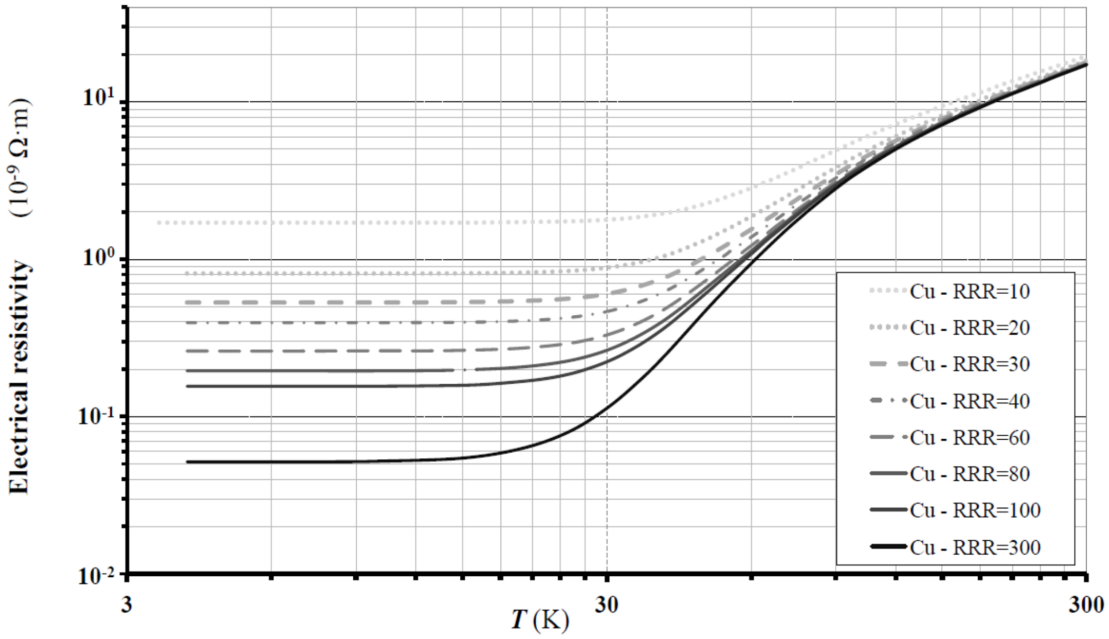


Figure 2.14: Electrical resistivity ρ of different coppers. Impurities and crystallographic defect content are indicated by the RRR. Electrical resistivities of purer coppers are indicated by darker solid lines, whereas those of less pure coppers are plotted with lighter grey dotted lines [16].

This would significantly improve the channel response, see Figure 2.15. It is however questionable if this is realistic, since only a small part of the actual cable will be cooled down.

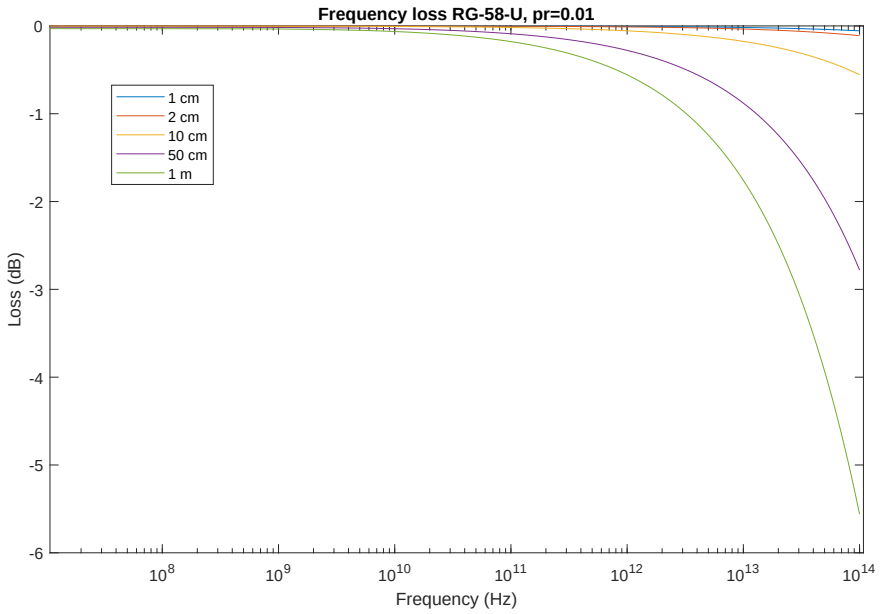
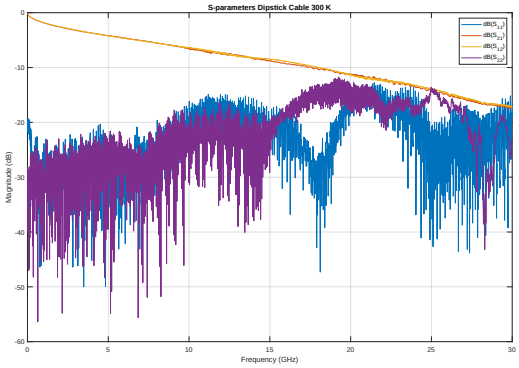


Figure 2.15: Frequency response RG-58/U cable pr=0.01.

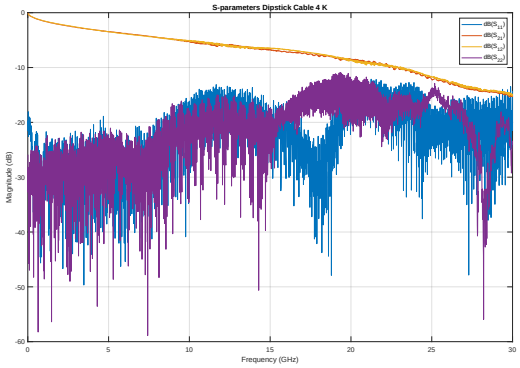
2.7.2. Channel Measurement

Two different channels have been measured to see the influence of channel characteristics when cooling down to cryogenic temperatures.

The first measurement was done using a dipstick, with two high-frequency cables interconnected at the bottom. Such so actually two cables are measured, in order to get a realistic overview of the single channel, the S12 and S21 have been halved in dB. The resulting S-parameters are plotted in Figure 2.16a for Room temperature and Figure 2.16b cooled down to cryogenic temperature.



(a) Channel measurement at room temperature.



(b) Channel measurement at cryogenic temperatures.

Figure 2.16: Channel measurements Dipstick configuration.

The result shows that the difference in reflection parameters S11 and S22 is not very much different when cooled down. However, the transmission parameters S12 and S21 are reduced by a few dB over the whole bandwidth. Figure 2.17 shows a clearer comparison between the transmission parameter S12 at each temperature.

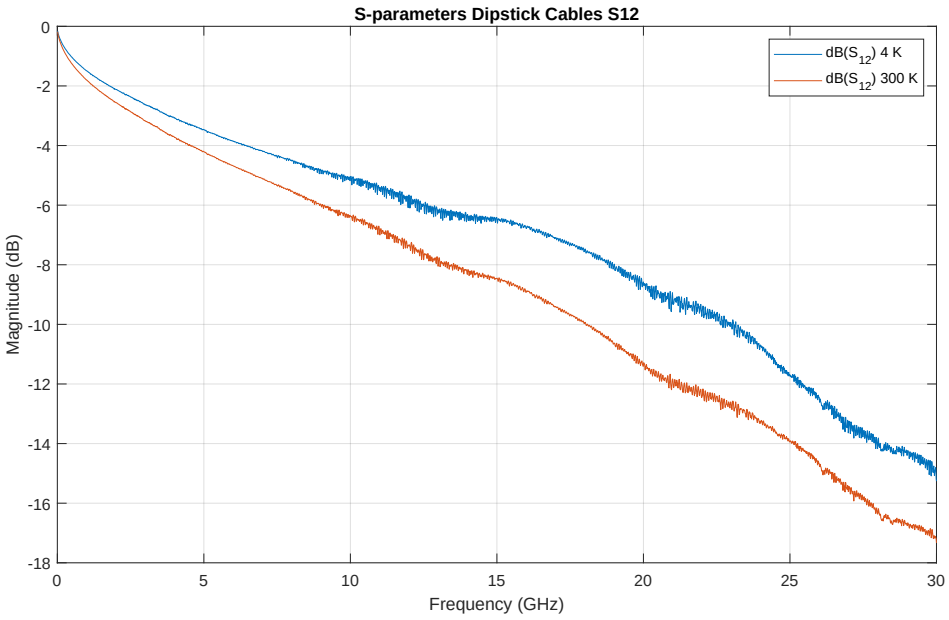


Figure 2.17: Dipstick channel S12 measurement comparison, at 4-K and 300-K.

The cable that finally needs to be driven will be a cryogenic high-frequency cable located in a dilution refrigerator. This cable has also been measured, and the resulting S-parameters are shown in figure 2.18. Note that there is a 6 dB attenuator which causes the baseline to start at -6 dB. Due to the better quality of the cable, the attenuation will be about 12 dB at 25 GHz when cooled down.

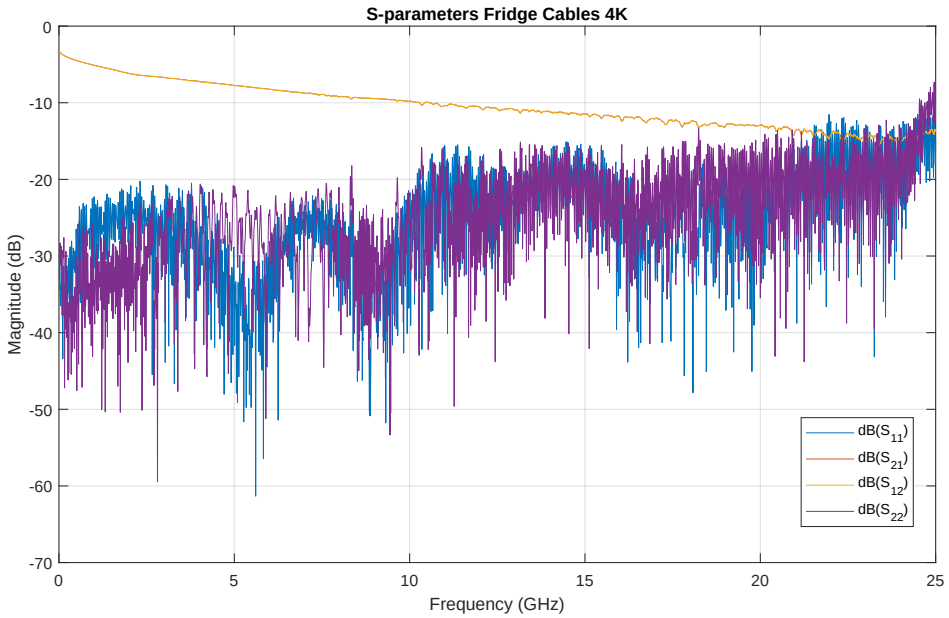
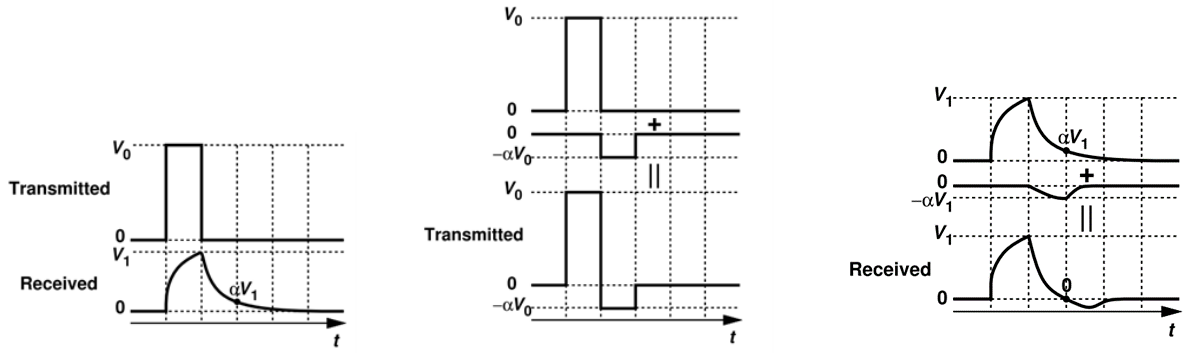


Figure 2.18: Dilution fridge channel S-parameter measurements, with 6 dB attenuation, at 4-K.

2.8. Equalization

2.8.1. Transmitter Pre-Emphasis

Since the signals are inevitably degraded by the low-pass behavior of the channel, the transmitter can already partially compensate for the channel loss at the cost of dynamic range. Another way to view the low-pass distortion in the time domain is when a single ONE is transmitted. Due to the channel, the pulse rises slowly and reaches a lower level, making it hard for the slicer at the receiver side to detect ONE. Moreover, the pulse entails a long tail that lasts for several UIs, disturbing upcoming signals. The problem is illustrated in the time domain in Figure 2.19a.



(a) Without de-emhasis (b) With de-emphasis (c) With de-emphasis received

Figure 2.19: Single transmitted pulse through low-pass channel in time domain [13]

In order to compensate for this behavior, a delayed and inverted scaled copy of the signal is sent, so the signal is brought down faster, as shown in Figure 2.19b. Therefore, the interference for the next bit is reduced, see figure 2.19c. This technique is called Feed Forward Equalisation (FFE), or better explained as de-emphasis; the attenuation of low-frequency components. In the frequency domain, this equalization technique shows a high-pass response up to the Nyquist frequency, which ideally compensates the low-pass channel. The frequency response of a 2-tap FFE is shown in Figure 2.20.

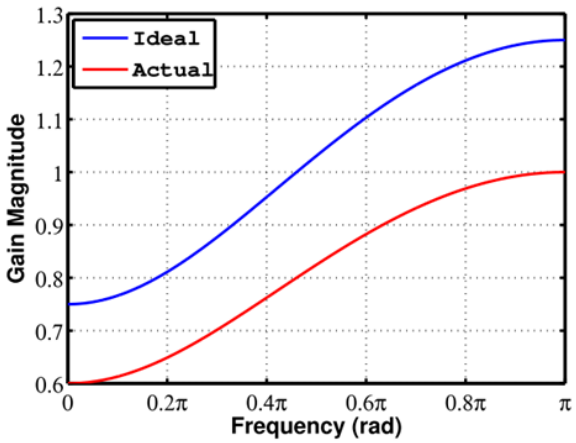


Figure 2.20: FFE frequency response [13].

The amount of de-emphasis can be adjusted if a steeper response is required. However, since the high boosting ratio will result in a penalty of significant attenuation on the DC component, we cannot sacrifice too much swing for boosting and make the transmitted signals more vulnerable to noise, reflections and crosstalk. The typical boosting ratio of the transmitter pre-emphasis is about 5-6 dB. Adding more taps will increase the ability to shape the output to the channel and thereby increase the SNR and hence BER. However, traditional voltage mode driver power scales up quickly with resolution (and the number of taps). For very lossy channels (>20 dB), where 8-taps or more are required a Digital-to-Analog Converter (DAC) based TX would be an option [17].

2.8.2. Transmitter Inductive peaking

In order to minimize the degradation of the digital signal, the bandwidth and rise time should be optimized. To deliver the required current, the transmitter front-end inevitably exploits large transistors presenting a large input capacitance to the preceding stage. Also self-loads arising from multiple branches in the Multiplexer branches limit the bandwidth. Lastly, the output ESD protection adds capacitance to the output. In order to obtain the best eye performance a large bandwidth is required. To be able to reach this bandwidth with a large parasitic capacitance at the output node inductive peaking proves useful. Figure 2.21 shows the equivalent circuit of multiple inductive peaking techniques.

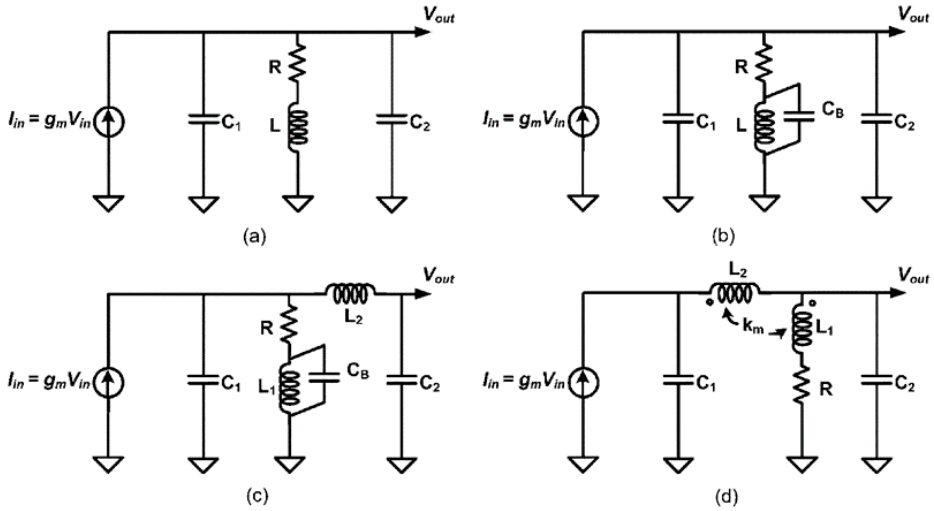


Figure 2.21: Equivalent circuit schematics of (a) shunt-peaked, (b) bridge-shunt-peaked, (c) bridged-shunt-series-peaked and (d) asymmetric T-coil peaked amplifiers [18].

For the shunt peaked amplifier, the equivalent voltage gain can be written as

$$A_V = -g_m \frac{R + sL}{1 + sRC + s^2LC} \tag{2.19}$$

This function is a typical second-order low-pass response where the q factor can be chosen. The inductor value is typically chosen according to the equation:

$$L = mR^2C \tag{2.20}$$

The factor m can be optimized to achieve a maximally flat-amplitude frequency response (MFA), 0 dB of peaking is achieved when choosing m=0.41. However, for high-speed interconnects, the m factor can also already be chosen higher up to m = 0.707, resulting in 1.5 dB peaking. This way, some of the high-frequency channel loss can already be compensated. In some applications, it is preferred to keep the shape as close to a square as possible, thereby a maximally flat envelope delay response (MFED) is the best option with m=0.33.

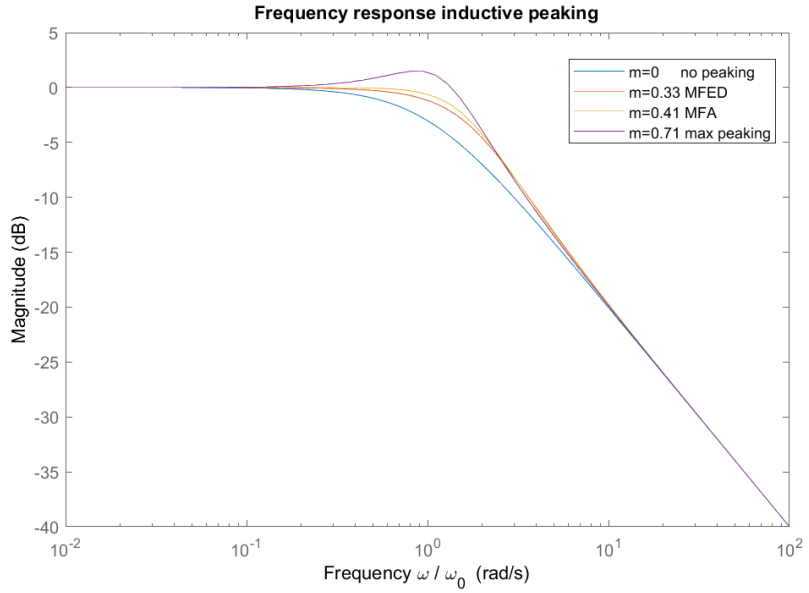


Figure 2.22: Frequency response inductive-peaking.

From Figure 2.22 can be seen, a larger value for m results in a higher Q-factor of the circuit and hence more overshoot in the frequency response.

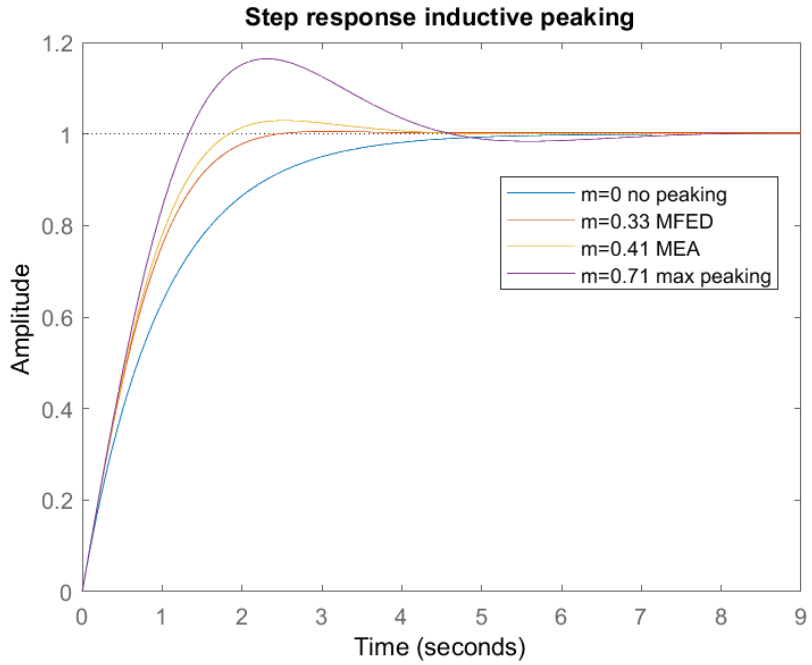


Figure 2.23: Step response inductive-peaking.

Looking at the step response in Figure 2.23, the maximum peaking response has the fastest rise time, but the MFED has the fastest settling time. MFED is preferred for PAM4 applications, since the settling time is the fastest and it has the smallest overshoot minimizing distortion to the eye diagram.

Double-inductor peaking such as the bridge-shunt-peaked Figure 2.23(c) and bridge-shunt-series-

peaked Figure 2.23(d), is a technique where the inductor physically separates C1 and C2. This method is optimal for larger capacitance ratios. For example, if the ESD load capacitance is very large compared to a small output driving capacitance. The transfer function is shown below:

$$A_v(s) = \frac{-g_m R\left(1+\left(\frac{1}{m_1}\right)\frac{s}{\omega_0}+\left(\frac{k_m}{m_1}\right)\frac{s^2}{\omega_0^2}\right)}{1+\frac{s}{\omega_0}+\left(\frac{1+k_B}{m_1}+\frac{1-k_C}{m_2}\right)\frac{s^2}{\omega_0^2}+\left(\frac{k_B}{m_1}+\frac{k_C(1-k_C)}{m_2}\right)\frac{s^3}{\omega_0^3}+\left(\frac{(k_C+k_B)(1-k_C)}{m_1 m_2}\right)\frac{s^4}{\omega_0^4}+\left(\frac{k_B k_C(1-k_C)}{m_1 m_2}\right)\frac{s^5}{\omega_0^5}} \tag{2.21}$$

The transformer or T-coil inductive peaking is advantageous if the drain parasitic capacitance is small compared to the total load capacitance, also it uses a smaller area. The transfer function is given by:

$$A_v(s) = -g_m R\frac{1+\left(\frac{1}{m_1}+\frac{k_m}{\sqrt{m_1 m_2}}\right)\frac{s}{\omega_0}}{1+\frac{s}{\omega_0}+\left(\frac{1}{m_1}+\frac{k_C}{m_2}+\frac{2k_C k_m}{\sqrt{m_1 m_2}}\right)\frac{s^2}{\omega_0^2}+\left(\frac{k_C(1-k_C)}{m_2}\right)\frac{s^3}{\omega_0^3}+\left(\frac{k_C(1-k_C)(1-k_m^2)}{m_1 m_2}\right)\frac{s^4}{\omega_0^4}} \tag{2.22}$$

With both methods, there are more variables to play with in order to design either for bandwidth or flat response. If it is done correctly, the T-coil can reach even faster settling times than the shunt peaking circuit.

2.8.3. Receiver CTLE

The receiver's Continuous time linear equalizer (CTLE) has a frequency response that compensates for the channels response. If designed properly, it will result in a relatively flat frequency response, thereby restoring the signal in the original form. There are many different forms; active boosting the amplitude, or passive attenuating lower frequencies. An active design can improve SNR, but might result in nonlinear behavior, which is problematic for PAM4. A common implementation is using a RC-degenerated differential pair, where the frequency response inverts the low-pass channel. The capacitor often can be adjusted for different channels.

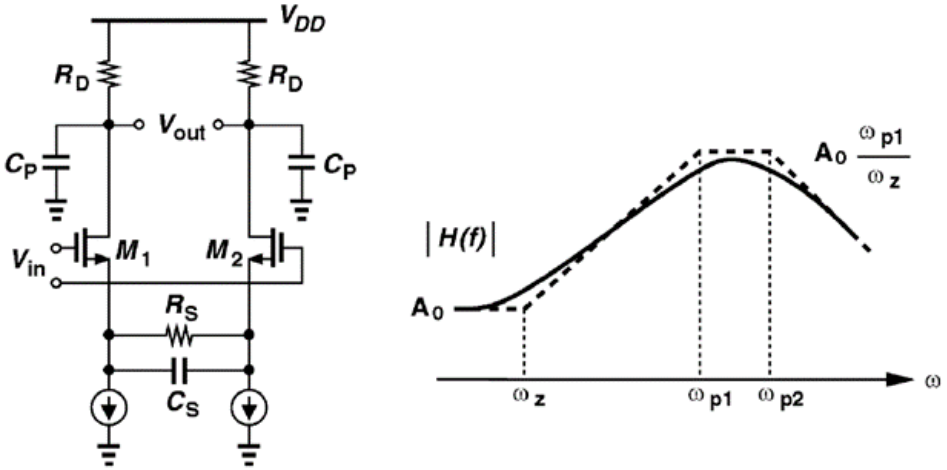


Figure 2.24: CTLE [19].

2.8.4. Receiver DFE

Decision feedback equalizer (DFE) uses an infinite impulse response (IIR) structure where the sum of all past decisions are adjusted by the DFE coefficients and used to minimize the errors at the target symbol levels. Due to the IIR structure, the DFE can correct a large amount of ISI with relatively short tap lengths for channel characteristics. Furthermore, the decisions are free of noise, so the DFE can compensate the channel ISI without amplifying noise levels. An example structure is shown below.

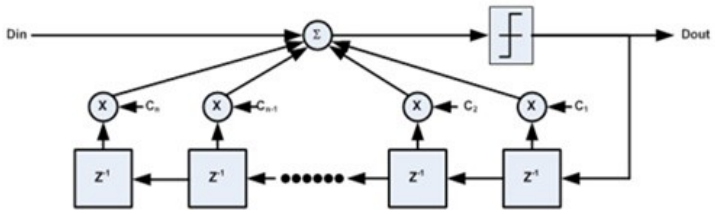


Figure 2.25: DFE [20].

2.8.5. FEC

The last technique is Forward Error Correction (FEC) [21]. This is not so much a channel equalization technique but is essential in serial links where the data rate exceeds 25 Gb/s. The problem with high-speed serial links is the limited swing leading to a low SNR and thus a higher chance of a bit error. For PAM4 this is even worse, since the 9.5 dB loss in SNR complemented with non-linear eye levels often results in a large BER. FEC can reduce the BER in the code domain, by adding extra redundancy bits to reduce the chance of an error. Therefore, by implementing forward error correction, the BER can be improved at the cost of effective bitrate.

2.9. Link Simulation

The channel imperfection limits the signal bandwidth and BER due to reflections. In order to get insight into the channel response, a S-parameter simulation was run using an example file from ADS. The signal path goes from die to package to PCB to a connector followed by a Coax Cable and then to the receiver. Each element in this chain has its own S-parameter block defining its Reflection/Transmission characteristics in the frequency domain. An example of the complete system in ADS is shown in Figure 2.26. Here, the S-parameters of a complete wireline system from the TX package, PCB traces, connectors, transmission line and receiving connector are evaluated.

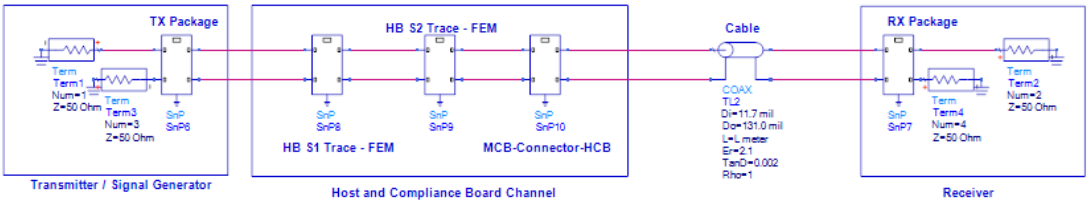


Figure 2.26: ADS model of a wireline system.

The virtual model of the coaxial line is varied from 0 to 10 meters to see the influence of the skin effect on the bandwidth of the system. The result is shown in Figure 2.27. Due to the skin effect, the transfer parameters S21 and S12 will become steeper the longer the cable. This means at the design frequency of about 26 GHz each meter longer cable will result in about 9 dB loss. It is thus key to keep the cable as short as possible.

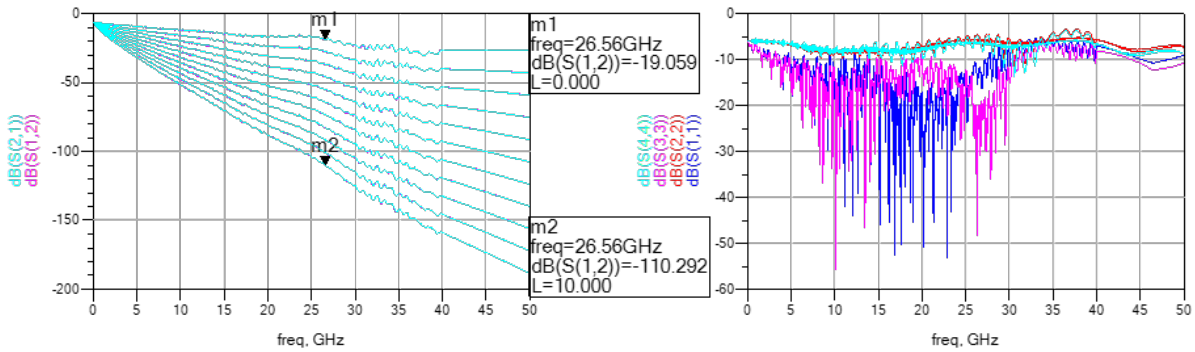
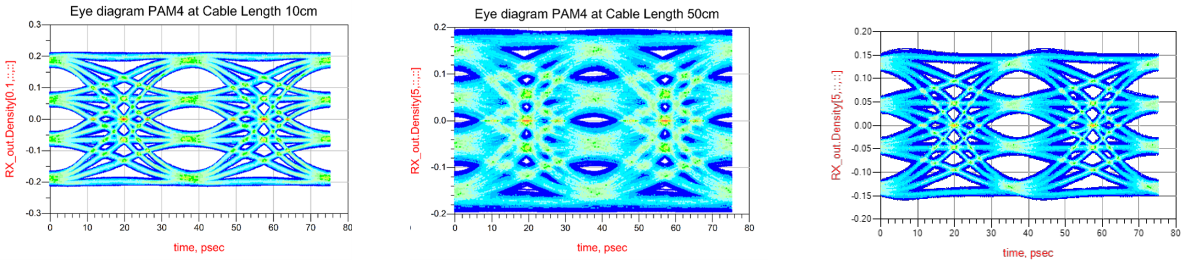


Figure 2.27: ADS simulation of Wireline varied from 0 to 1 meter.

Simulation of the complete wireline can be done using I/O Buffer Information Specification - Algorithmic Modeling Interface (IBIS-AMI) models that are available as example in ADS. The IBIS-AMI models define the channel response of the Transmitter and Receiver chips, so they can be simulated and the eye diagram plotted like in Figure 2.28. The complete system including variable Coax cable can be simulated, for example at 10 cm Figure 2.28a and at 50 cm Figure 2.28b. At 50 cm the eye diagram is almost completely closed. One useful feature in IBIS-AMI simulations is the ability to alter the equalization on both transmitter and receiver side. This could improve the eye opening significantly when chosen correctly. For example, when using 5.5 dB CTLE the performance over the cable length already improves at 50 cm. Note that although the eye opening has increased from almost none to a few mV, however the total swing decreased from 200mV to 150mV due to the de-emphasis.



(a) No de-emphasis 10 cm Coax (b) No de-emphasis 50 cm Coax (c) 5.5dB de-emphasis 50 cm Coax

Figure 2.28: IBIS-AMI simulation eye diagrams of example wireline system.

2.10. Requirements

The requirements are defined in this section. The cable loss has been analysed in section 2.7, since the loss is comparable to a long-range CEI channel the requirements are mostly based on this PAM4 standard [14].

The requirements are summarized in table 2.2. The minimum and maximum values are based on the PAM4 standard [14]. The Goal listed is the value which this wireline transmitter design aims to reach.

Characteristic	Symbol	Min.	Goal	Max.	Unit
Feed Forward Equalisation Taps	FFE_{taps}	2	3	-	-
Data Rate	R_b	36	51.2	58	Gb/s
Baud Rate	$BAUD$	18	25.6	29	Gsym/s
Bit Error Rate(pre-FEC)	BER	-	-	1e-6	-
Signal-to-Noise-and-Distortion-Ratio (SNDR)	$SNDR$	31	32.5	-	dB
Maximum Output Differential Voltage	$V_{PPd_{MAX}}$	-	800	1200	mVppd
Relative Level Mismatch	RLM	0.95	-	-	%
Uncorrelated Bounded High Probability Jitter	T_{UBHPJ}	-	-	.05	UIpp
Uncorrelated Unbounded Gaussian Jitter	T_{UUGJ}	-	-	.01	UIrms
Power	P_{tot}	-	300	500	mW
Insertion Loss	IL	-	$15_{@14GHz}$	$30_{@14GHz}$	dB
Operating Temperature	T	4	-	300	Kelvin

Table 2.2: Wireline transmitter requirements.

2.11. Conclusion

This chapter set out to analyze the background knowledge and definitions for wireline communication and set the requirements for this Cryo-CMOS design. Multiple data formats have been considered including NRZ, duobinary and PAM4. Analysing the actual coaxial wire channel measurements at cryogenic temperatures lead to a small improvement of a few dB compared to room temperature. However the loss is still significant as could be seen in link simulations at high data rates, meaning designing a PAM4 capable transmitter would be most viable. Based on this, the design requirements for the PAM4 wireline transmitter have been set.

3

Transmitter Architecture

This chapter means to compare different PAM4 transmitter architectures and define the architecture for this design. Section 3.1 compares recent architectures on their performance. Section 3.2 compares FFE implementations. Section 3.3 then defines this design system architecture.

3.1. State-of-the-art Comparison

In this section some of the most recent PAM4 transmitter designs are being compared. Although the purpose of all transmitters is the same, the results vary a lot between the transmitters. The main difference is the choice in architecture. The architecture differentiates between the design in FFE implementations and output driver choices.

The digital architecture introduced in [13] employs a simple multiplexing structure without using any latches, thereby keeping the power consumption extremely low, see Figure 3.1. However, no FFE is implemented in the design and since the last bit is held for only 2 quarter-rate cycles, there is only the possibility to expand to a 2-tap FFE. Also, the swing is limited to only $0.63\,V_{ppd}$, which might not be good enough for a high-loss channel. Moreover, this makes the design very sensitive to errors in the quarter rate clock, so it might be good to add a correction algorithm to make sure the duty cycles are 25%.

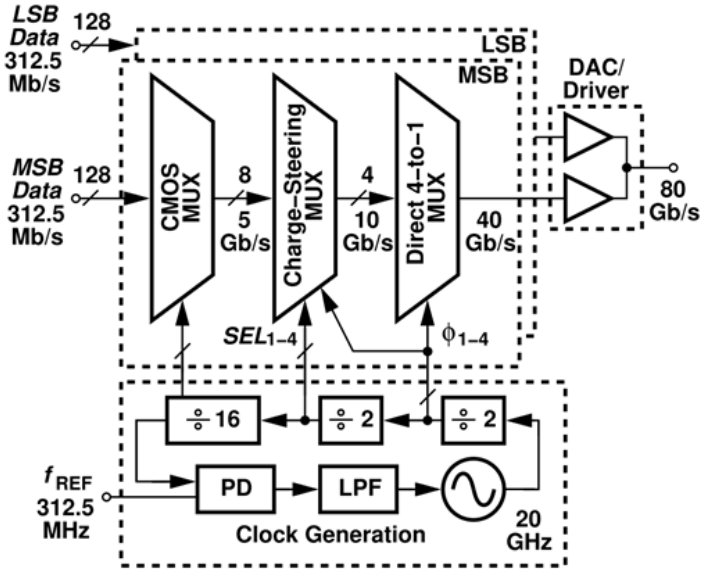


Figure 3.1: Block diagram of transmitter [13].

The architecture introduced in [22] differentiates itself with a coarse and fine level 4-tap FFE. The

coarse and fine FFE is implemented by adjusting the output impedance of each SST driver cell individually with reference generators. Another thing to note is that in this design also latches in the 4:1 MUX are mitigated by adopting an automatic phase alignment technique, improving power efficiency and robustness. The Source Series Terminated driver (SST) drivers are well-known to be more power-efficient than their Current Mode Logic driver (CML) counterparts due to their lower termination power. However, to achieve precise control of the FFE tap weights, the SST drivers require several driver segments connected in parallel. This requirement results in high dynamic power consumption due to higher clock and data distribution requirements associated with the increased number of SST segments.

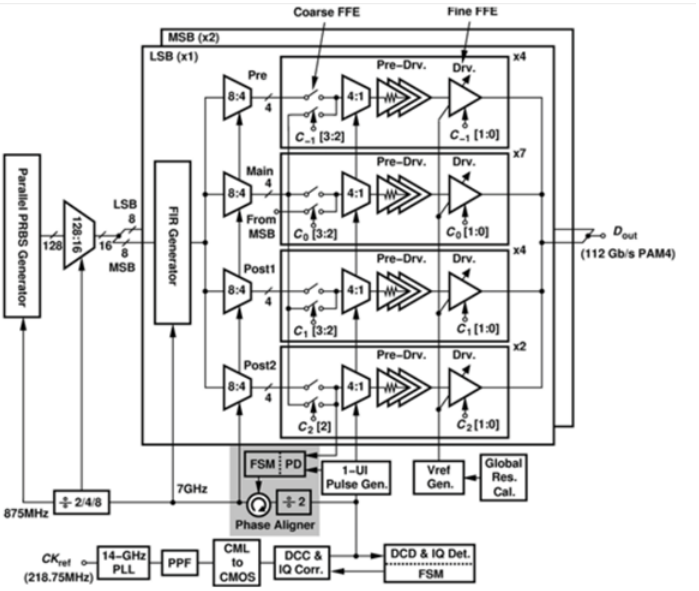


Figure 3.2: Block diagram of transmitter [22].

[23] was one of the first 112 Gb/s papers and set the standard for PAM4. The architecture implements a 3-tap FFE by having 3 slices at the output, both implementing a direct 4:1 MUX and output driver. Before each slice, the data for both MSB and LSB paths are multiplexed to 4 UI and retimed before going into the output slices.

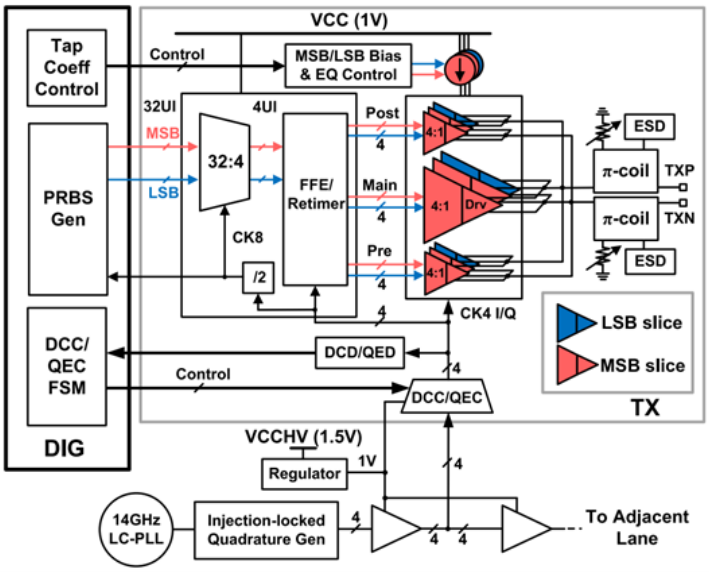


Figure 3.3: Block diagram of transmitter [23].

[24] implements a fractional 2-tap FFE, meaning the delay time of the FFE is not 1 UI, but can be adjusted by a factor α. This way, there is another variable to adjust the equalization, thereby improving the compensation without amplifying noise. The tap delay is adjusted in the clock path utilizing a course-fine capacitor array-based delay cell.

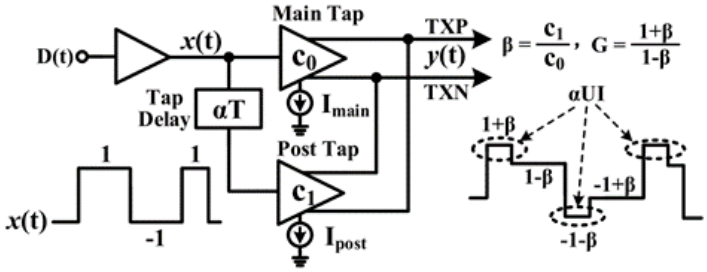


Figure 3.4: Block diagram of transmitter [24].

[25] does an excellent job in both implementing a 3-tap FFE, while keeping the power consumption very low. The architecture uses only a few latches for delay compensation. Another difference in this design is the tailless CML driver, compared to the traditional differential-pair current-mode driver, the maximum voltage stress across the output transistors is remarkably reduced, allowing a higher supply voltage to be used without requiring thick oxide or cascode devices.

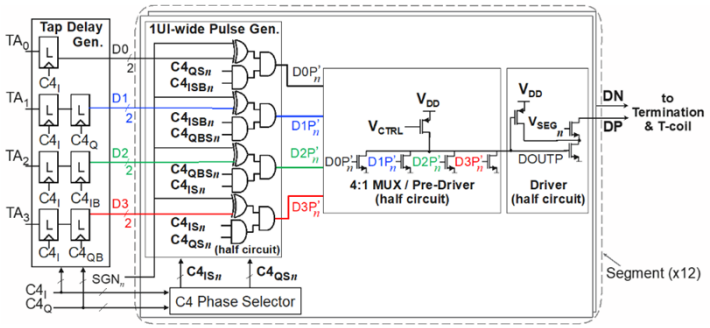


Figure 3.5: Block diagram of transmitter [25].

At last, [17] uses a completely different architecture approach, see Figure 3.6, implementing an 8b SST DAC makes it flexible to implement any equalization technique digitally. In this case, an 8-tap digital FIR is implemented, which could be well utilized in compensating more complex channels. The performance of the SST DAC elements is also reasonable, only due to the higher number of digital multiplexing circuitry required for driving all separate DAC elements. The total efficiency is 2.6 pJ/bit.

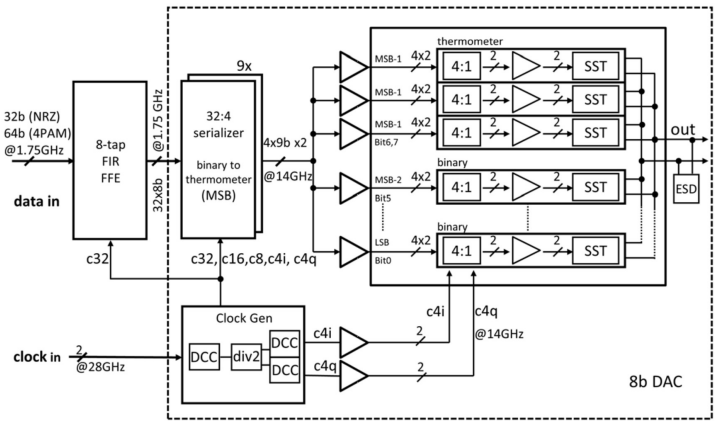


Figure 3.6: Block diagram of transmitter [17].

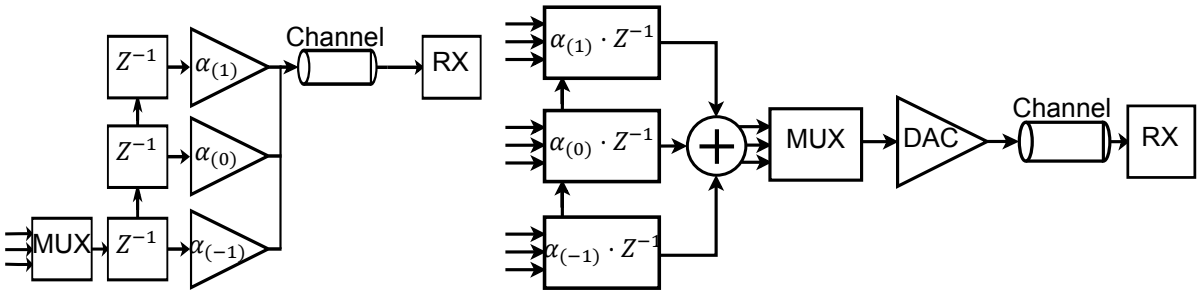
The result of the comparison is shown in Table 3.1. It can be seen [13] does an excellent job at achieving a high energy efficiency including PLL, however, it does not include a FFE. Therefore [24] would be a more viable option being both energy efficient and having sufficient swing and bandwidth for the cable loss.

Paper	[13]	[22]	[2]	[24]	[25]	[17]
Technology [nm]	45	40	10	65	14	14
Data Rate [Gb/s]	80	112	112	112	128	112
FFE	None	4-tap	3-tap	2-tap(fractional)	3-tap	8-tap FIR DAC
Output driver	CML	SST	CML	CML	CML(tailless)	SST
Supply [V]	1	1/1.2	1/1.5	1.2	0.95/1.2	0.95
Max. Output Vpp,d [V]	0.63	1	0.75	1.2	1	-
RMS jitter [fs]	205	208	154	-	-	-
Power [mW]	44.1*	436*	232	243	166*	-
RLM [%]	99	97.7	98.5		99	-
Efficiency [pJ/bit]	0.55*	3.89*	2.07	2.17	1.3*	2.6
Area [mm2]	0.1	0.56	0.0302	0.694	0.048	0.095

Table 3.1: Comparison PAM4 transmitters *including PLL power.

3.2. FFE Architectures

In general, a wireline transmitter consists of a serializer mostly consisting of retimers and Multiplexers, a Feed Forward Equalisation (FFE) implementation to compensate the channel, and lastly, a driver able to provide enough swing so it can be received at the other side of the line. The main architectural difference in transmitter design is the implementation of the Feed Forward Equalisation (FFE). The first option to implement the FFE is by means of a delaying copies with adjustable level α of the output signal after serializing, as shown in Figure 3.7a [22], [23], [25]. The difficulty in this application is that the delayed copy needs to work at the desired output data rate. This makes the structure not only hard to design but also merely flexible in the number of taps and accuracy of the α value. Therefore a better option would be to implement the FFE in the digital domain, as shown in Fig. 3.7b [17]. This DAC-based implementation does pre-emphasis before serializing in the digital domain thereby allowing for a wide variety of possibilities. The DAC can be separately designed to provide enough resolution to deliver the right compensation for the channel. In terms of power consumption, one could say a DAC-based implementation requires more power since there is more digital logic required. However, most digital logic is happening at a low frequency, and finally, both implementations need to drive the same load, which consumes most of the power. The DAC implementation is, therefore, the preferred option.



(a) Delay based (b) DAC based

Figure 3.7: 2-tap FFE implementations in transmitter architecture.

3.3. Transmitter Architecture

The complete System overview is shown in Figure 3.8. At the bottom right, the clock generation circuit will generate the necessary clock frequencies and phases employing dividers from an external 12.8 GHz input. On the left, the incoming low-speed 400 Mbps data is digitally generated by a synthesized SRAM. The digital clock can be selected from one of the available phases. This incoming data

will then be retimed and serialized 16× to 6.4 Gb/s data signals utilizing DFFs and a 16:1 Multiplexer. Next, this 6.4 Gb/s data-stream is retimed and non-overlapping 25% duty-cycle data pulses are generated by making use of the clock phases Φ_{1-4}. Then, a CML-based 4:1 Multiplexer will serialize this data into a single 25.6 Gb/s differential stream. This will then be amplified by a CML based DAC driver to generate the output signal. The first 4 bits from the bottom are binary encoded where for each driver the amount of I_{LSB} cells is doubled, the top 2 bits are unary encoded and therefore require 6 cells each consisting of $8xI_{LSB}$ current sources.

3.4. Conclusion

In conclusion, different state-of-the-art wireline transmitters have been compared on their performance and the architecture of the transmitter has been decided. Based on an analysis of the different FFE architectures, a DAC-based option seems most viable for this application. To keep the power consumption, low the best option is to combine the low power digital Multiplexer of [13] with a flexible DAC FFE architecture like in [17]. This leads to the system implementation shown in Figure 3.8 to be the architecture for this design.

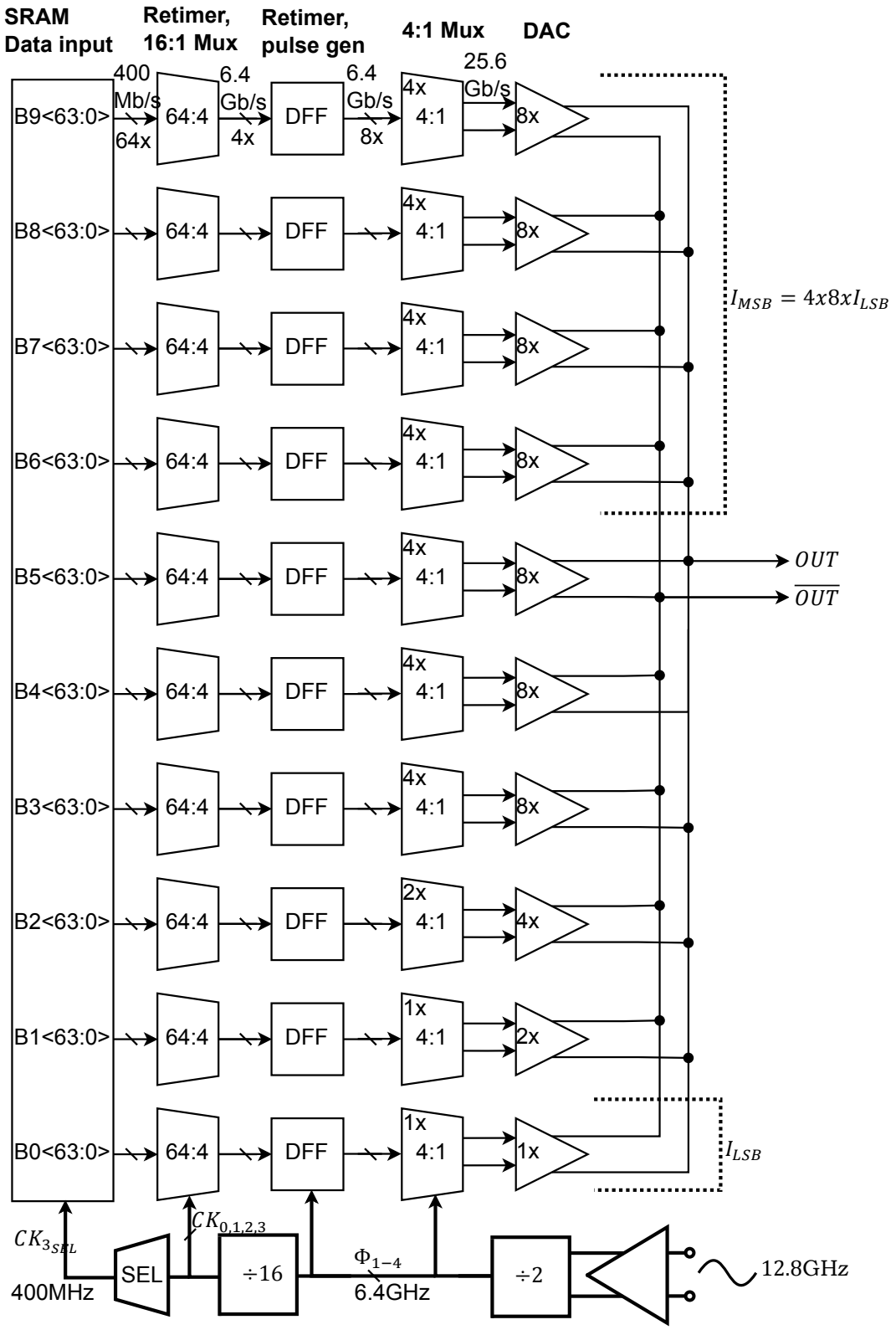


Figure 3.8: 6bit DAC-based wireline transmitter system overview.

4

Clock generation

The goal of the clock generation circuit is to generate all the required quadrature clock phases at the right frequencies, for the multiplexing structures to work. The 4 overlapping quadrature 6.4 GHz clock phases need to be carefully designed since jitter or clock skew will directly influence the output eye diagram. The other phases required for the low-speed Multiplexer have more slack, but should still be checked on functionality. Section 4.1 introduces the clock generation architecture. Section 4.2 goes into detail about the divide-by-2 circuit design. Section 4.4 shows the layout of the clock generation circuitry. Section 4.5 will discuss the simulation results.

4.1. Architecture

The clock generation architecture should be able to deliver the necessary overlapping quadrature clock phases for the multiplexing and retiming structures used in the system. The multiplexing structure has been specifically designed to work with all 50 % duty cycle clock phases, to keep the divider design simple and symmetrical to the multiplexing structure. The reason for this choice is that each divider will inherently have some delay, meaning the lower frequency clock phases $CK3$ will always be lagging behind higher frequency clock phases like $CK1$. For the latchless 16:1 Multiplexer design in Chapter 5, this is an advantage since it would fail and select the wrong data if the rising edge of a fast clock like $CK1$ would be leading the falling edge of a $CK3$, meaning data that is multiplexed would not be available anymore. The clock generation architecture is shown in Figure 4.1. A 12.8 GHz external differential reference clock will come in from the left, where it will be terminated and buffered to drive the first divider. The first divider is a CML latch-based divide-by-two implementation converting the differential 12.8GHz incoming clock to 4x 6.4GHz IQ phases. The next dividers are normal latch-based back-to-back circuits dividing down 16× to generate all 45^o quadrature clock phases at 3.2 GHz, 1.6 GHz, 800 MHz and 400 MHz. What is not shown in this figure is that all clocks $CK_{0,1,2,3}$ are also complementary, since every $\div 2$ has a complementary IQ output the $\overline{CK_{0,1,2,3}}$ can also be used in the multiplexing structure. Lastly, two-phase selectors have been designed to be able to select both the 6.4 GHz Φ_{1-4} phases for the high-speed retimer and the 400 MHz $CK3_{<0-7>}$ for the digital SRAM.

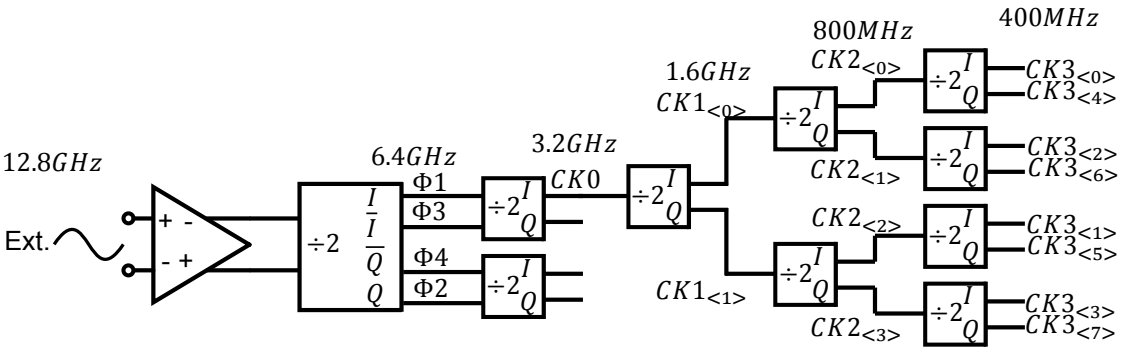
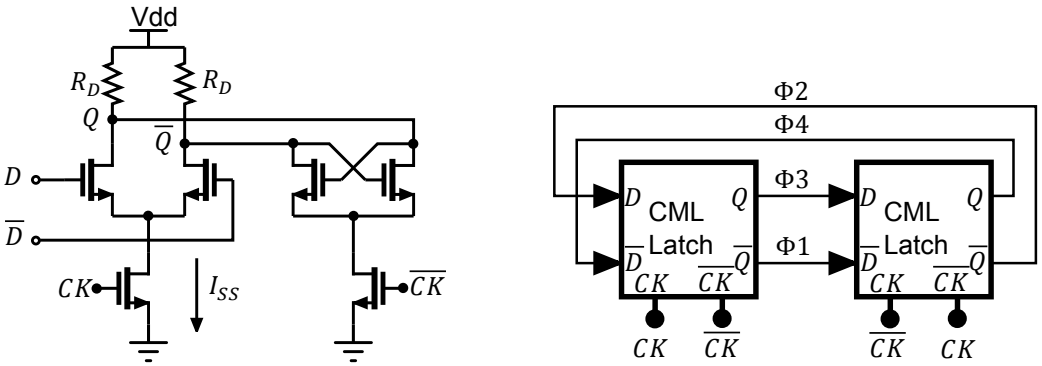


Figure 4.1: Clock generation architecture.

4.2. Dividers

In this design, only divide-by-2 circuits are used whose main purpose is to half the frequency and generate the quadrature phases necessary for the multiplexing structure. Most frequency dividers employ one or more latches, the design choices depend on parameters such as the maximum speed, power consumption, and the amount of clocked transistors [26]. This design employs two different divider structures. These structures have been used earlier in a test buffer of a PLL and proven to work at cryogenic temperatures [27], [28].

The first divide-by-2 has to run at a high frequency and divide the 12.8 GHz into the 4 6.4 GHz clock phases $\Phi_1 - \Phi_4$. The chosen topology here is a CML static latch, although it consumes high power, it can reliably operate at high frequencies. The CML latch is shown in Figure 4.2a, the two main advantages of this circuit are: (1) a moderate voltage swing is used equal to $I_{SS} \cdot R_D$, allowing fast transitions, and (2) only NMOS devices are used in the data and clock paths minimizing input load capacitance. In the divide-by-2 structure shown in Figure 4.2b, two CML latches are connected in a master-slave configuration to generate the required 50% quadrature clock phases $\Phi_1 - \Phi_4$.

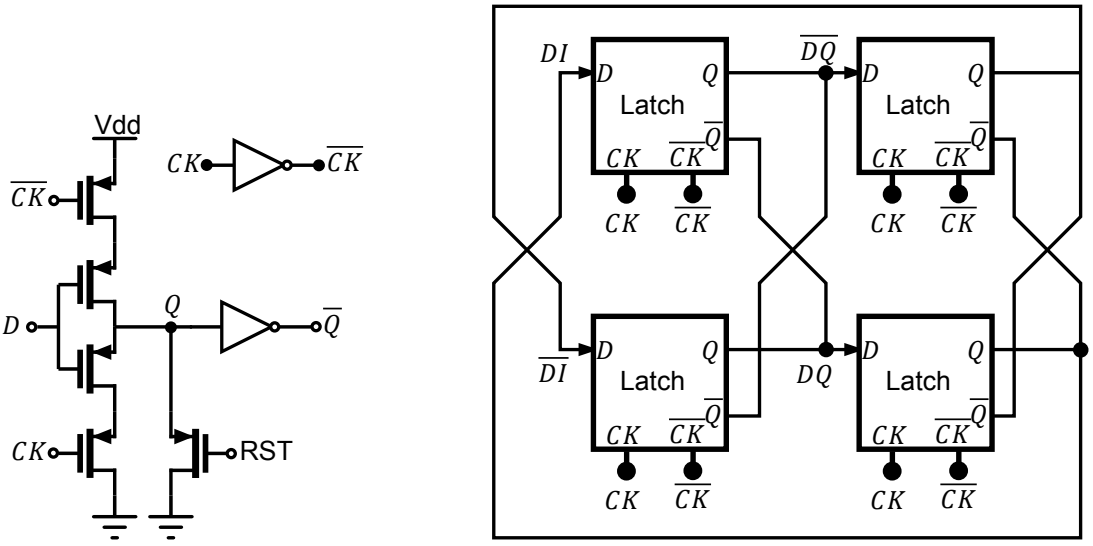


(a) CML Latch

(b) CML Divider

Figure 4.2: Static CML divide-by-2 structure.

The other divide-by-2 structure in this design is used for all other divisions from 6.4 GHz down to 400 MHz. The chosen latch is a dynamic ”Clocked CMOS” (C2MOS) latch, which can operate at reasonably high speeds while having a lower power consumption than the CML topology offering an attractive solution for all dividers in this design. The C2MOS latch is shown in Figure 4.3a, the first branch will latch the input D on the Q node when CK is high, the second branch inverter will generate the complementary $\overline{Q}$. The complete C2MOS divide-by-2 structure is shown in 4.3b, where 4 latches are used to generate the quadrature phases in a complementary master-slave structure. A disadvantage of this structure is the meta-stability at startup, due to the input and output nodes of the C2MOS latches being undefined. Therefore, an asynchronous reset transistor has been added to make sure all quadrature phases can be reset at startup.



(a) C2MOS Latch

(b) C2MOS Divider

Figure 4.3: Dynamic C2MOS divide-by-2 structure.

4.3. Input Buffer

The input buffer circuit is shown in figure 4.4. The termination resistors will match the incoming clock signal to $R_T = 50\,\Omega$. Coupling capacitors will block DC signals and pass the high-frequency clock into the back-to-back inverters. The back-to-back inverters will generate a full rail-to-rail complementary clock signal. This rail-to-rail signal is then amplified by two inverters to drive the input of the first clock divider.

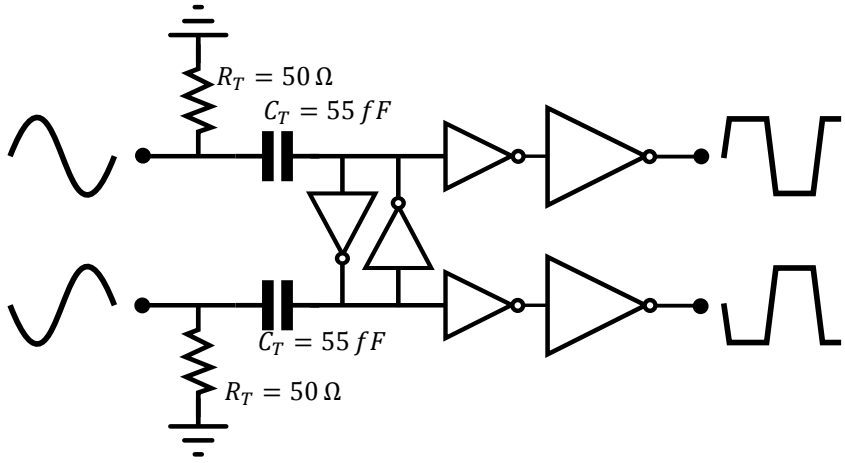


Figure 4.4: Input buffer circuit.

4.4. Layout

The layout of the clock generator is shown in Figure 4.5. The external 12.8 GHz differential clock input will come in from the pads on top. Then this will be terminated on-chip, using 50Ω unsilicided poly resistors. The input buffer will then convert the input signal to a rail-to-rail clock signal. The CML divide-by-2 circuit is directly below and will divide the input clock into 4 quadrature phases $\Phi_1 - \Phi_4$. These 4 phases then go directly down to the phase selector, so the path to the Multiplexers is as short as possible. The Phase Selector make use of the same Multiplexers as will be later discussed in Chapter 5 to select a phase. Φ_2 and Φ_4 are used as input for the C2MOS dividers on the right to generate all other required clock phases. Another selector on the bottom can select one phase for the SRAM. After each divider and in between long lines, large buffers have been added to drive lines and connected circuitry.

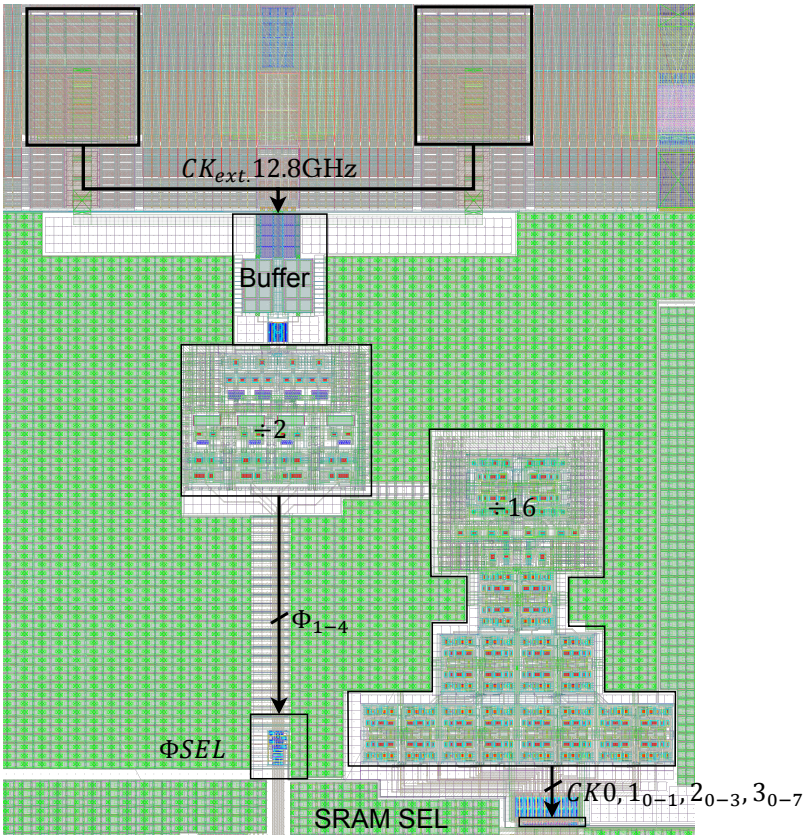


Figure 4.5: Clock generation architecture.

4.5. Results

A simulation of the post-layout RC extraction is run with an ideal external clock signal, the resulting phase outputs are shown in Figure 4.6. Note that only the positive phases are shown; there is also a complementary version. This means that in total there are 34 quadrature phases being generated to be used in this system.

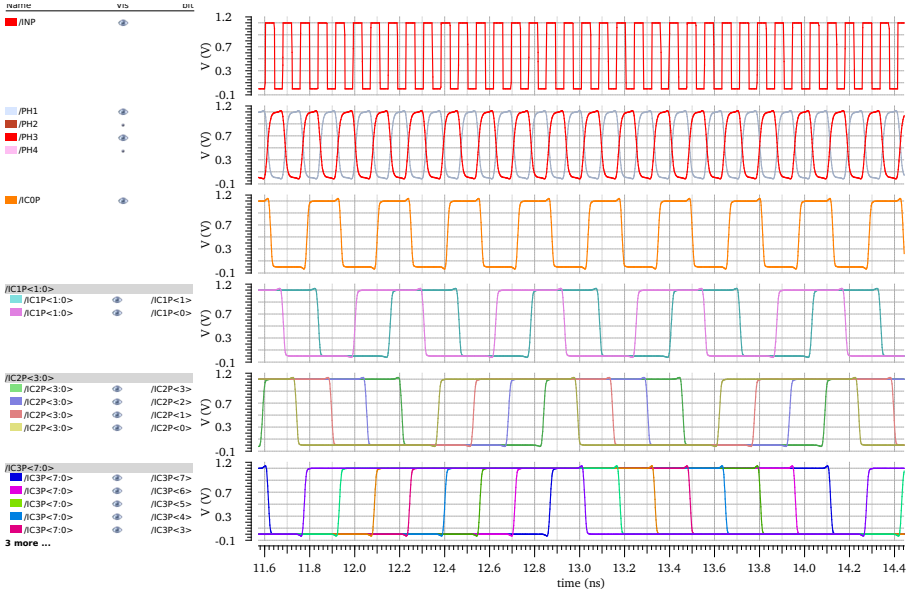


Figure 4.6: Extracted simulation clock dividers, resulting clock phases.

Any phase noise on the 4 6.4 GHz phases $\Phi_1 - \Phi_4$ will directly influence the output eye diagram. Simulating the phase noise spectrum of this big circuit was not possible due to server limitations, so an approximation of the peak-to-peak random jitter was simulated. A 25 ns transient simulation was run with transient noise from 10 kHz to 100 GHz. The simulated eye diagram of the Φ_1 output is shown in Figure 4.7, measuring the peak-to-peak jitter at the middle crossing gives a total of $J_{pp} \approx 300\,fs$.

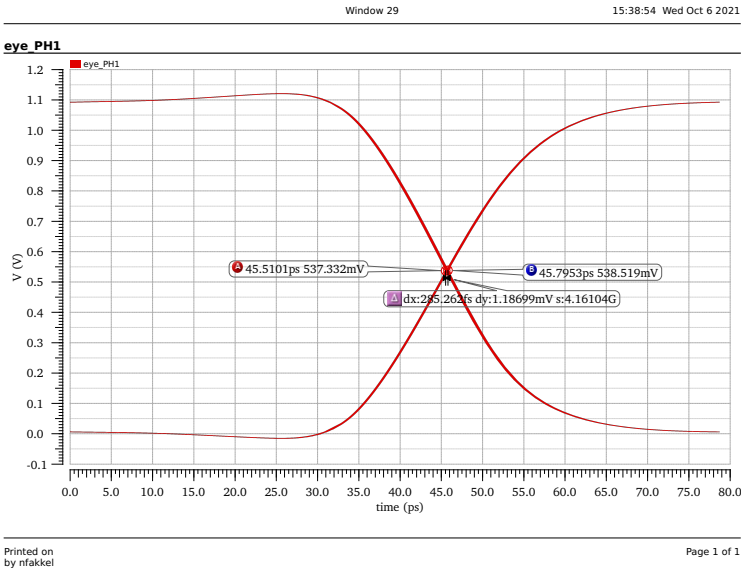


Figure 4.7: Extracted simulation CML divider output Φ_1, to measure peak-to-peak random jitter.

A breakdown of the power consumption of the clock generator is shown in table 4.1. It can be seen that most power is consumed by the high-frequency CML divider, which was expected. Another thing to notice is that in total the clock consumes more power than all other systems in the chip together, mainly due to oversizing the input buffer and CML divider for jitter performance and adding many buffers to drive the long lines.

Part	Current @1.1V
Input buffer	5.35 mA
CML Divider $\Phi_1 - \Phi_4$	15.43mA
C2MOS Divider $CK0$	3.347 mA
C2MOS Dividers $CK1,2,3$	4.245 mA
Clock buffers $CK0,1,2,3$	2.436 mA
SRAM selector $CK3_{SRAM}$	19.52 uA
Phase selector	736.5 nA
Total	29.41 mA

Table 4.1: Power consumption Clock.

4.6. Conclusion

The clock generation was meant to generate all quadrature clock phases for the structure with sufficient clock jitter, skew, and bandwidth performance. Those goals were achieved. The architecture was mainly based on divide-by-2 circuits generating quadrature clock outputs. A CML divider was chosen for the important high-speed clock phases to minimize jitter, that would be visible on the output eye. The other dividers made use of a C2MOS latch division structure having a lower power consumption, but still able to deliver the quadrature performance. The layout has been kept simple and symmetric to minimize clock skew. Results show the peak-to-peak jitter requirements can be reached, RMS jitter simulations could not be run.

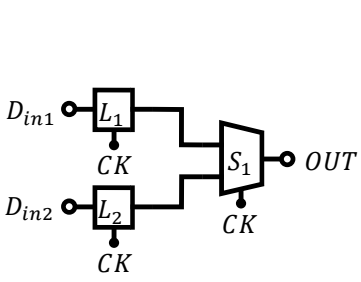
5

Low speed 16:1 Retimer and Multiplexer

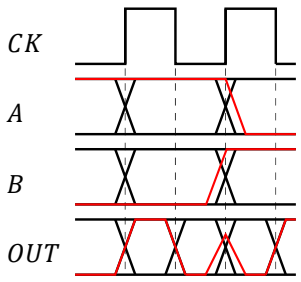
This section will discuss the design of the almost latch-less 16-to-1 Multiplexer used in the Wireline transmitter.
First, the latch-less Multiplexer architecture is introduced in Section 5.1. The 2-to-1 Multiplexer used in this architecture is shortly evaluated in Section 5.2. The layout approach to minimize the size and clock skew is shown in Section 5.3. Following up, the issues of this design are addressed in Section 5.4 and the extracted simulation results are shown in Section 5.5.

5.1. Architecture

As shown in section 3.3, the 16:1 Multiplexer is responsible for serializing the low-speed 400 Mbps data signals from the SRAM memory to high-speed 6.4 Gb/s output signals. The complete system employs 10 separate DAC elements for the 4-bit binary, 2-bit unary(2+4) DAC. For each DAC element, there is a separate serializing slice, each multiplexing $\frac{25.6\,\text{Gb/s}}{400\,\text{Mbps}} = 64\times$, this means in total there are $10 \times 64 = 640$ input bits. The low-speed multiplexing structure is responsible for serializing these 640 bits $\frac{6.4\,\text{Gb/s}}{400\,\text{Mbps}} = 16\times$, meaning a total of $\frac{640}{16} = 40$ 16-to-1 serializers are needed. A conventional implementation of a 16-to-1 Multiplexer combining 2-to-1 Multiplexers, would require 15 2-to-1 Multiplexers each, so $40x15 = 600$ in total. A common way to synchronize the data before each 2-to-1 Multiplexer is done using 2 latches or DFFs, as shown in Figure 5.1a. Using this structure would require at least $600 \times 2 = 1200$ latches, which would be very power consuming. Another problem with this architecture is the possible glitches that could appear when L_1, L_2 and S_1 transition on the same clock input, but there is a delay mismatch resulting in the data not being ready before selected by S_1. This could happen due to clock jitter, clock skew, or even a delay difference between rise and fall times. Figure 5.1b illustrates such a simultaneous transition in red, due to L_2 leading and L_1 lagging, a large glitch appears at the output of the Multiplexer S_1. Even if the timing would be extremely accurate, a glitch, although smaller, could still appear.



(a) Circuit



(b) Timing Diagram

Figure 5.1: Basic 2-latch structure.

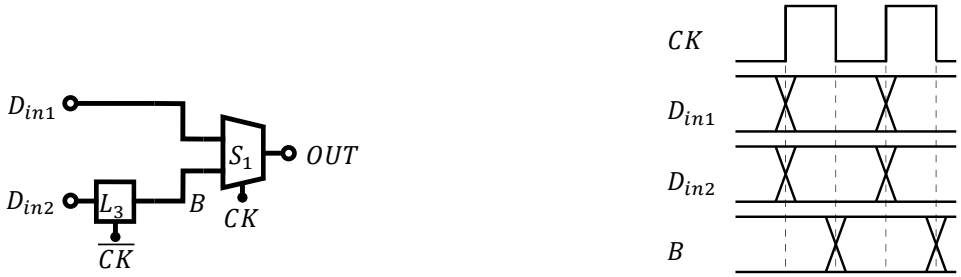
A conventional solution shown in Figure 5.2a would be to keep the first two latches L_1 and L_2 syn-

chronizing the data to block any glitches from the preceding multiplexing stage. Then adding an extra latch L_3 to avoid any input change when the clock is selected A. Another purpose of L_3 is to make sure there are no simultaneous transitions or “glitches”. Figure 5.2b shows the same transition in red, only now due to point B being delayed half a clock cycle no simultaneous transition appears, and the glitch is suppressed. Unfortunately, adding this extra latch would increase the power consumption, since now in total 600x3=1800 latches are needed.

(a) Circuit (b) Timing Diagram

Figure 5.2: Conventional 3-latch structure.

When the input signals Din1 and Din2 are known and well-controlled, L_1 and L_2 can be omitted. This is assumed that D_{in1} and D_{in2} change on one edge of the clock and settle before the next edge. For example, if the previous Multiplexer in the structure uses the same clock. Now, this trick cannot be done directly after the SRAM since there the data is undefined. However, removing L_1 and L_2 after the first multiplexing stage where this signal is defined would be beneficial for power. See Figure 5.3, only L_3 is still necessary to prevent glitches.



(a) Circuit (b) Timing Diagram

Figure 5.3: 1-latch structure [13].

Now if there are quadrature clock phases available, these phases can create the necessary delay between each Multiplexer by the proper choice of clock edges, L_3 can effectively be replaced. This has already been used in [29] to establish a longer hold time of the Multiplexer input, and in [13] to remove latches completely, see Figure 5.4.

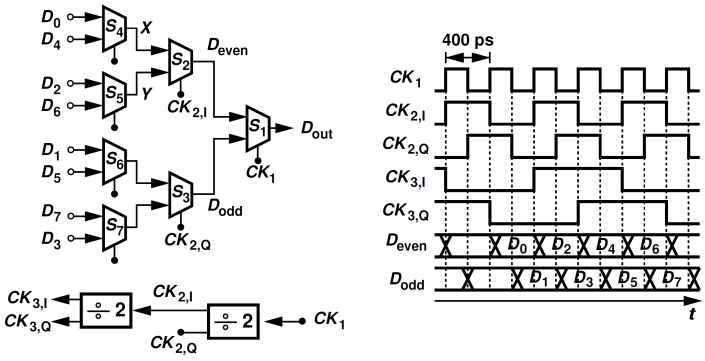


Figure 5.4: IQ clock timing scheme [13].

The example structure in Figure 5.4 shows how D_{even} and D_{odd} will never transition at the same time due to the 90 degrees phase shift between $CK_{2,I}$ and $CK_{2,Q}$. This does require the clock $CK_{2,Q}$ to always lag with an exact phase delay of 90 degrees or half a CK_1 clock cycle. Then at each transition, there is half a CK_1 clock cycle available, so D_{even} and D_{odd} will have sufficient time to settle before S_1 selects D_{out}. This architecture can be extended continuing the odd and even structure, ensuring there are enough clock phases available. The first Multiplexers $S_4 - S_7$ make use of clock phases $CK_{3,II,IQ,QI,QQ}$, each lagging with a phase delay of 90 degrees or half a CK_2 clock cycle. Implementing this extended structure will save a total of 1200 latches compared to the conventional solution. Only $10 \times 5 \times 8 = 400$ latches are still necessary for retiming the SRAM data at the input to the clock phases. An extra row of retimers is added to be sure the data from the SRAM is retimed with the same clock. The complete 16-to-1 Multiplexer is shown in Figure 5.6.

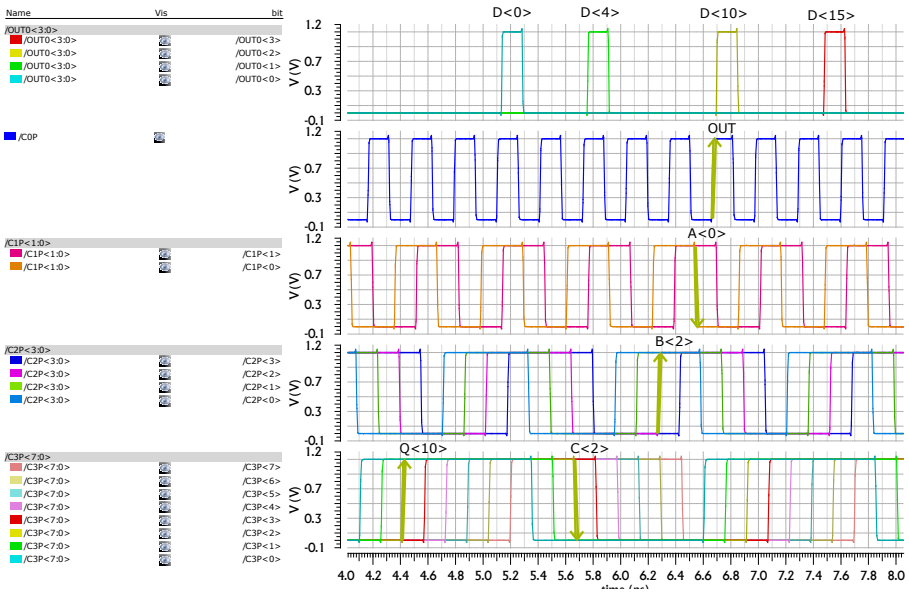


Figure 5.5: 16-to-1 Multiplexer timing diagram, the top row shows the output of the 16-to-1 Multiplexers when D<0>, D<4>, D<10>, D<15> transition from 0 to 1, the lower rows illustrate the in-between transitions on the clock phases from nodes Q<10>,C<2>,B<2>,A<0> to OUT. The nodes are shown in Figure 5.6

Figure 5.5 shows a simulation of the transitions D<0>, D<4>, D<10>, D<15> from 0 to 1, while the rest of the inputs stay 0. In order to understand the structure, let us follow signal D<10>, from left to right in Figure 5.6. This signal will come from the SRAM and thus needs to be retimed with the first DFF, which makes sure it is synchronized with CK3<0>. Then the data is retimed again to the negative edge of CK3<2> and then retimed again by the positive edge CK3<2> creating Q<10>. Hence Q<10> lags behind Q<2> for half a clock cycle.

After being retimed, the first Multiplexer S_5 will be selecting C<2> to become Q<10> on the positive edge of CK3<2>. Then the second Multiplexer S_6 will be selecting B<2> to become C<2> on the positive edge of CK2<2>, followed by S_4 making A<0> equal to B<2> on the positive edge of CK1<0>. Lastly Multiplexer S_8 will select A<0> on the positive edge of CK0, resulting in OUT to transition from $0 \rightarrow 1$.

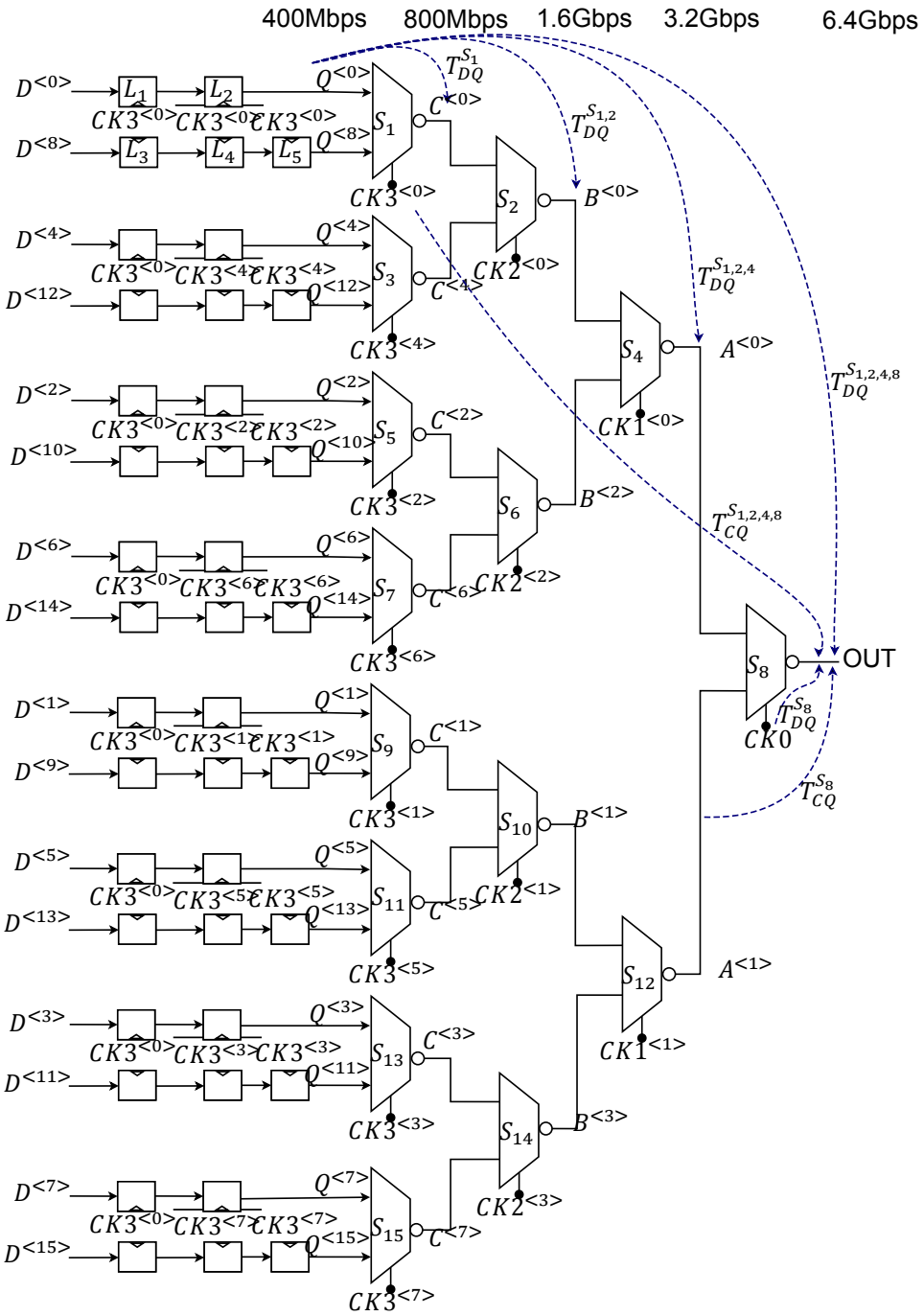


Figure 5.6: 16-to-1 Multiplexer architecture.

5.2. CMOS 2-to-1 Multiplexer

There are two common logic implementations of a 2-to-1 Multiplexer. The first implementation makes use of pass gates and is shown in Figure 5.7. When $SEL = 1$, the top NMOS and PMOS transistor will turn on and allow the signal of A to be passed to the inverter at the output, while the bottom two transistors are turned off and block incoming signal B. When $SEL = 0$, the bottom PMOS and NMOS transistor will turn on and pass signal B, while the top two transistors are turned off to stop A. This implementation is popular in complex logic circuits as it does not require high power and saves area. For high-speed Cryo-CMOS purposes, the pass gates will form an issue. At cryogenic temperatures, the threshold voltage V_T increases, also the subthreshold slope increases. This means that when having a limited swing, it could occur the Gate-Source voltage V_{GS} on the NMOS pass gate of A is not large

enough $V_{GS} > V_T$ to make the pass gate conduct. The subthreshold slope will mean the pass gate is not fully open and hence the increased resistance will limit the speed of the system.

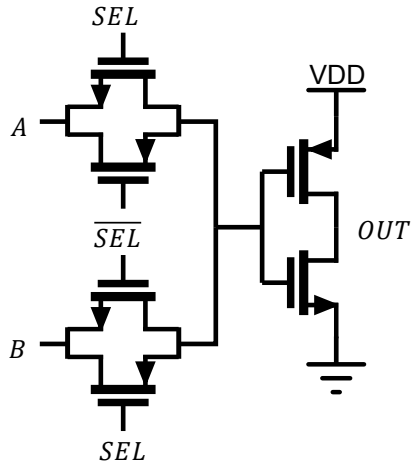


Figure 5.7: Pass-gate 2-to-1 Multiplexer circuit.

The second option to implement a 2-to-1 Multiplexer is a complementary CMOS circuit, as shown in Figure 5.8. It consists of two branches and an inverter. The middle 4 PMOS and NMOS transistors select either branch to be connected to the inverter. For example, if $A = 1$ and $SEL = 0$ then the right bottom two NMOS transistors will conduct and pull the input of the inverter down, thus the output will be $OUT = 1$. The biggest advantage of this implementation is the output has a rail-to-rail swing, thus not very susceptible to the cryogenic threshold problem. Since both the regular and inverted clocks are already available, power consumption is also relatively low.

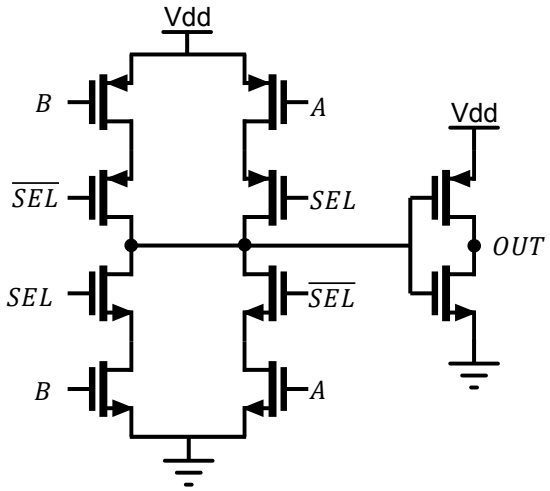


Figure 5.8: CMOS 2-to-1 Multiplexer circuit.

Sizing the CMOS logic was done by taking three factors into account. (1) the power consumption of each Multiplexer should be kept as low as possible; (2) the delay of the Multiplexer should be within boundaries so it does not cause a bit error at the high frequency even when driving the load; (3) the ratio between pmos and nmos should be optimized such so the load resistance is equal for both and rise and fall times are similar.

The ratio between nmos and pmos can be calculated by equalizing the current equations for both n and p channel devices, both are approximately equal if PMOS is sized twice as large. After simulating the rise and fall times, the bottom transistors were optimized to be $\frac{W}{L}_{NMOS} = \frac{200\text{nm}}{40\text{nm}}$ and the top transistors $\frac{W}{L}_{PMOS} = \frac{410\text{nm}}{40\text{nm}}$. These values are for both the complementary stage and the inverter stage,

since the former one is stacked, the fan-out factor for the latter inverter is approximately 2 which is acceptable.

5.3. Layout

To keep the area of the complete chip as small as possible and prevent long clock lines or too much skew between the different Multiplexers, the complete 16-to-1 Multiplexer has been cramped into a small but long structure of 4um by 90 um. This was done by making arrays of 8 DFF retimers and putting them side by side. Similarly, the Multiplexers have been moved side to side thereby minimizing the delay between each and aligning them directly on top of the DFF array.

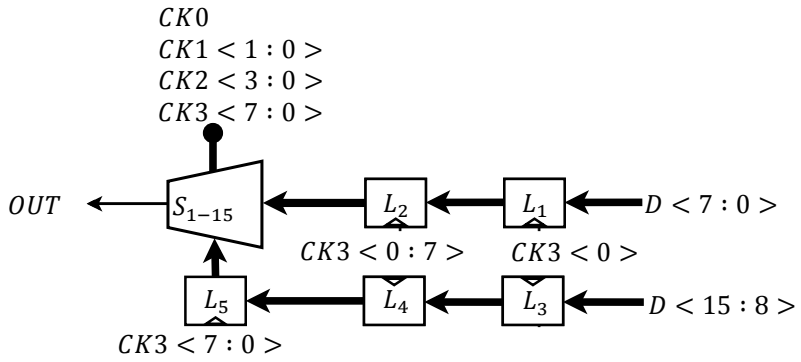


Figure 5.9: Layout sketch 16-to-1 Multiplexer and retimers.

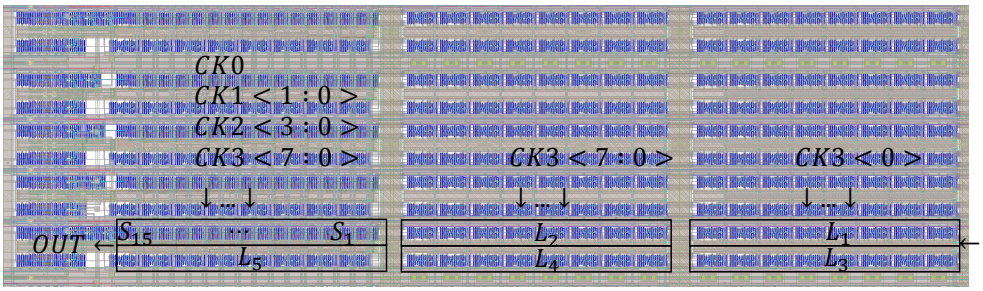


Figure 5.10: Virtuoso Layout 16-to-1 Multiplexer and retimers.

5.4. Issues

The latch-less Multiplexer structure could fail in a few scenarios, due to signals or clock delays leading or lagging for longer than the available selection time of a Multiplexer. It is assumed that the selectors will retime the incoming data signal by the clock edges and thereby resetting the error σ at each stage. However, Monte Carlo simulations are needed to prove this. If this is known, a few requirements can be set for each stage, looking at the blue arrows in Figure 5.6. The total data delay should be smaller than the duration of the incoming data signal minus the duration of the output signal $T_{DQ}^{S_{1,2,4,8}} < \frac{1}{400Mbps} - \frac{1}{6.4Gb/s} = 2.5ns - 165ps = 2.34ns$. This is to make sure the incoming data information is still available when selected and not delayed too much so a new data signal is selected. This requirement will continue for the rest of the structure so $T_{DQ}^{S_1} < 1.25ns$, $T_{DQ}^{S_{1,2}} < 1.875ns$, $T_{DQ}^{S_{1,2,4}} < 2.1875ns$. Another requirement is the maximum allowed error $\sigma_{T_{CQS_8}}$ on the output, based on the bit error requirements and the available setup time of the following retimer stage this should be determined.

Clock skew could also cause a big issue in this structure since everything is phase-related. If a clock phase comes in too late at one of the Multiplexers in Figure 5.6 this will cause the wrong data to be sent to the next stage. Therefore to validate the functionality of the system, the clock skew needs to be added to the sigma of the T_{CQ} and T_{DQ} delays. Clock skew can only be simulated after extraction of the layout since it is dependent on the length of traces and the characteristics of the transistors.

5.5. Results

The first simulations were Monte Carlo runs of the complete 16-to-1 Multiplexer to find the delay times and errors σ of the blue arrows in Figure 5.6. The measured delay results of the Monte Carlo runs are shown in Table 5.1. Note that these simulations are just an estimation since only 100 Monte Carlo runs have been executed while more are needed to accurately predict the $\sigma_{mismatch}$, due to mismatch in devices. It can be concluded from the results that the total T_{DQ} delays have more than sufficient margin, so no bit error will occur due to an incoming signal coming in late. Moreover, it can be seen that the different stages will reset the error at each multiplexing stage since the clock delay error from only the last stage is $\sigma_{DQ}^{S_8} = 2\,ps$, while the clock delay from the first to the last stage is larger $\sigma_{DQ}^{S_{1,2,4,8}} = 12.2\,ps$. Hence, the timing error of the last multiplexing stage is likely only related to itself and unrelated to the total delay error of the system.

Delay path	T_{tot}	σ
$T_{DQ}^{S_1}$	29 ps	2 ps
$T_{DQ}^{S_{1,2}}$	79 ps	6 ps
$T_{DQ}^{S_{1,2,4}}$	124 ps	9.4 ps
$T_{DQ}^{S_{1,2,4,8}}$	172 ps	12.9 ps
$T_{CQ}^{S_{1,2,4,8}}$	174 ps	12.2 ps
$T_{DQ}^{S_8}$	24 ps	2 ps

Table 5.1: Monte Carlo time delay crossing simulation results.

The second simulation that has been done was a fully extracted layout simulation of all 16-to-1 Multiplexers and retimers. This way the effect of clock skew can be best analyzed since the long clock lines are taken into account. Each Multiplexer was given the same input data stream, so each Multiplexer would have the same output. Therefore, any clock skew will appear as a delay between each Multiplexer output. The result of all outputs is plotted as an eye diagram in Figure 5.11.

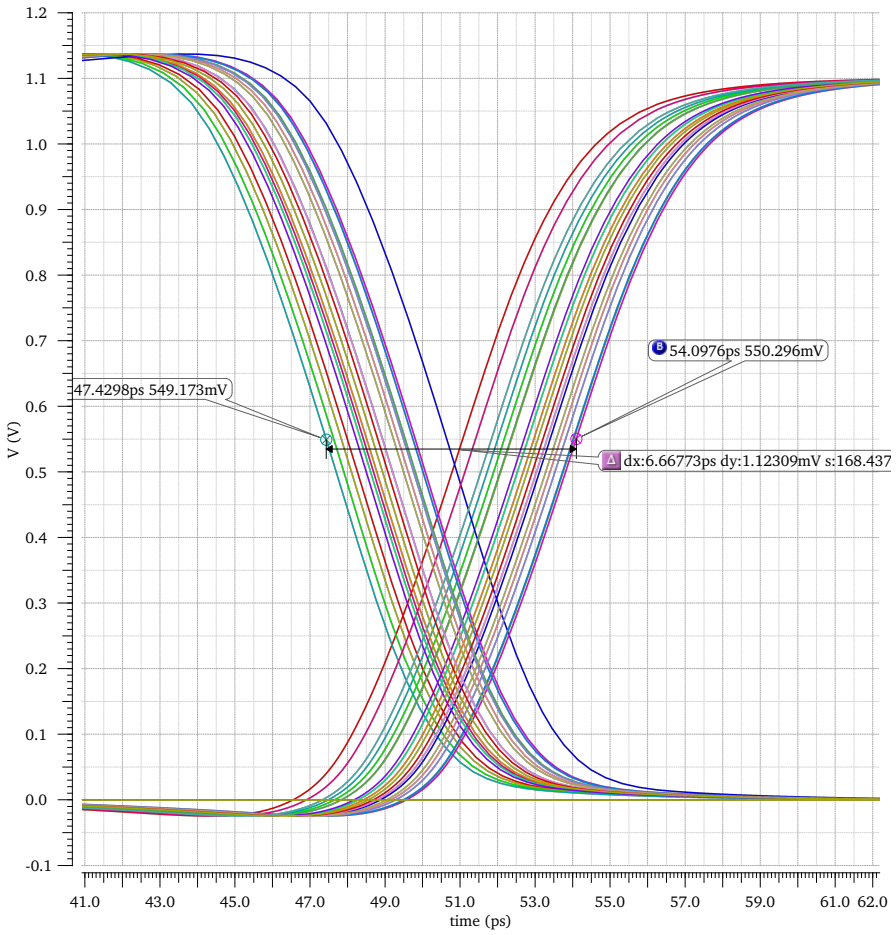


Figure 5.11: Extracted simulation 16-to-1 Multiplexer, clock skew crossings.

From Figure 5.11, it can be concluded that the total peak-to-peak clock skew in the system accounting for both rising and falling edges is about 6 ps. Looking only at rising edges the exact clock skew delay between the first and last output is just 3 ps. Combining both the jitter from the mismatch of the devices, the clock skew, and the random clock jitter of the 3.2GHz clock, the total jitter can be approximated as follows:

$$\sigma_{tot} = \sqrt{\sigma_{mismatch}^2 + \sigma_{skew}^2 + \sigma_{random}^2} = 6.3ps \tag{5.1}$$

To determine if the combined jitter and clock skew is a problem, the crest factor needs to be calculated. This can be done using the following equation:

$$P\left(|\text{ Jitter }| \geq \frac{N\sigma}{2}\right) = DTD \times \text{erfc}\left(\frac{N}{\sqrt{8}}\right) \tag{5.2}$$

Where P is the chance of a bit error, data-transition Density (DTD) (= 0.5 for data signal), σ the RMS jitter, and lastly N is the crest factor.

For the required BER of 10^{-6} the crest factor will be 9.5. This means that the total peak-to-peak jitter will be:

$$J_{peak-peak} = N\sigma = 9.5 \cdot 6.3 = 60ps \tag{5.3}$$

Since at a data speed of 6.4 Gb/s there is 156 ps available to sample the data, the setup time of the next high-speed retimer should be below 95 ps.

Power In total all the 10x low-speed multiplexing slices including retimers consume an average power of 2.85 mA at 1.1-V nominal supply.

5.6. Conclusion

A power-efficient low-speed 16-to-1 Multiplexer structure has been designed to convert the 400 Mb/s input signals into a 6.4 Gb/s stream. A C2MOS 2-to-1 Multiplexer is chosen for its full swing capabilities and likely will work well at cryogenic temperatures. A layout strategy was introduced to minimize the area and clock skew, which turned out to be sufficient for preventing any issues as could be seen in the simulation results.

6

High-speed retimer and Direct 4:1 Multiplexer

For high-speed operations above 5 Gb/s, the latch-less strategy will be difficult in CMOS logic. Therefore, the data must be retimed using a high-speed retimer and 25% duty cycle pulses need to be generated for a 4:1 CML Multiplexer to finally create the 25.6 Gb/s data stream.

This chapter is structured as follows, Section 6.1 introduces the architecture of the Direct 4:1 Multiplexer, pulse generator, and retimer. The retimer DFF circuit is chosen in Section 6.2. The pulse generator circuit is analysed in Section 6.3 and the 4:1 Multiplexer in Section 6.4. The clock tree layout approach is discussed in Section 6.5. The resulting pulses and MUX output are shown in Section 6.6.

6.1. Architecture

The incoming 6.4 Gb/s data from the previous Multiplexer needs to be converted to a 25.6 Gb/s signal, making use of 4x 50% duty cycle 12.8GHz clock phases. A Direct 4:1 Multiplexer will convert 4 incoming data streams into a single 25.6 Gb/s output. However, before this can be done, the data needs to be retimed, made complementary, and adapted to 25% duty cycle pulses. The structure to do this is shown in Figure 6.1, where the High-speed retimer, pulse generator, and 4:1 Multiplexer are drawn for the second input A<1>. The first DFF in this structure will retime all 6.4 Gb/s incoming data streams to the second clock phase $\Phi 2$. Then a complementary copy should be made and retimed. Then the data should be reduced to a 25% duty cycle pulse at each phase. This will generate interleaving pulses for the 4:1 Multiplexer to generate one high speed 25.6 Gb/s output.

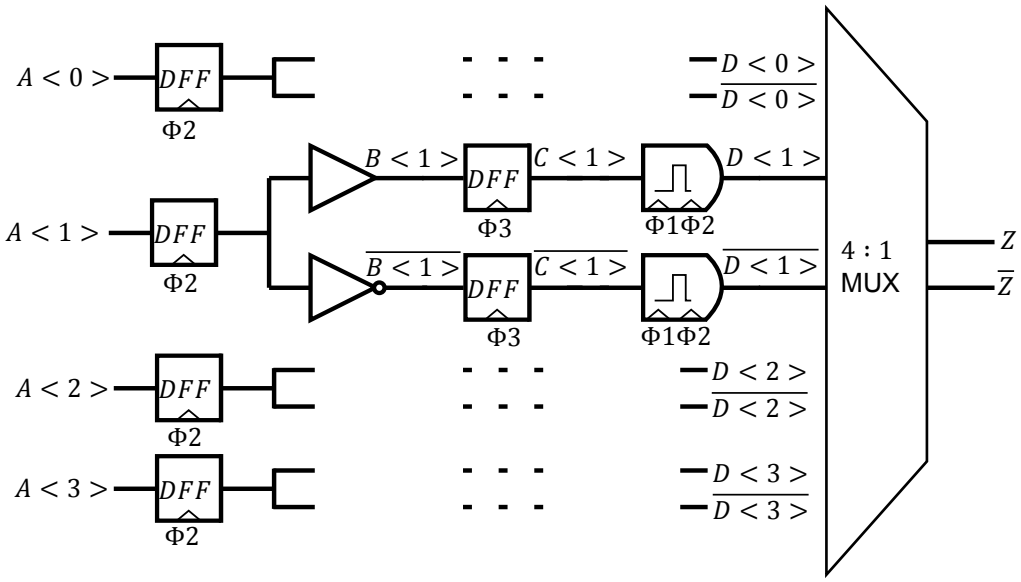


Figure 6.1: Architecture High-speed retimer, pulse generator and 4:1 Multiplexer.

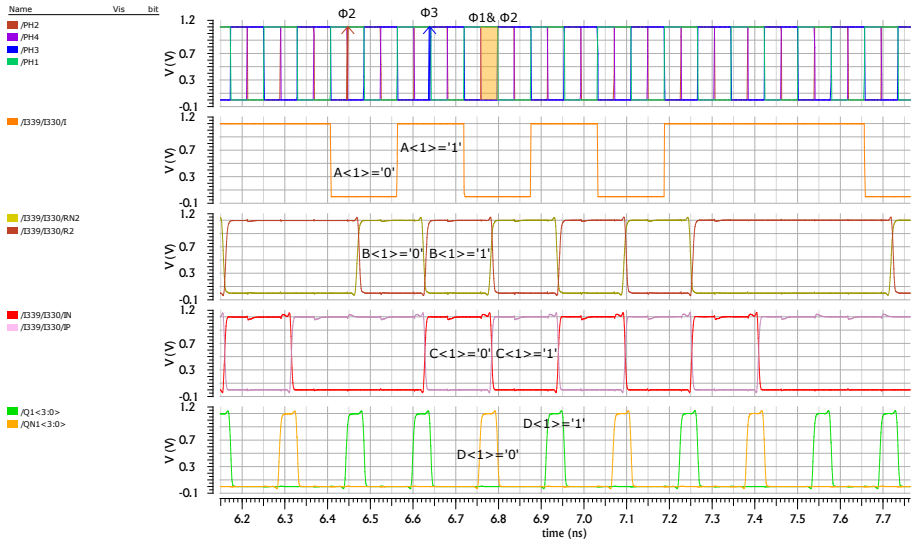


Figure 6.2: Timing diagram of high-speed retimer, showing the transitions from A<1> to D<1>.

6.2. DFF

A commonly used Dynamic Flip-Flop (DFF) implementation is a transmission gate based master-slave DFF, shown in Figure 6.3. In basic operation, when the clock is high, the incoming signal D latches on the master, setting up point $\overline{QM} = \overline{D}$. Then when the clock is low, the slave will latch point $\overline{QM}$ making the output equal to $Q = D$ on the next positive edge. The setup time of this structure is very limited by the performance of the transmission gates, and as was already discussed in the previous section, these pass gates do not perform well at cryogenic temperatures. Therefore, other options should be considered.

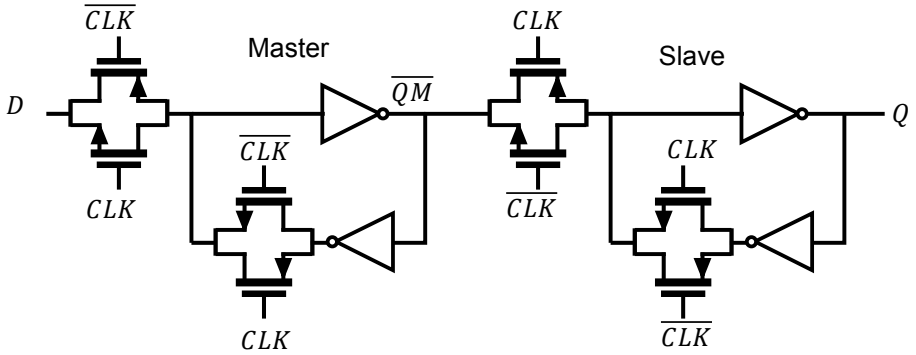


Figure 6.3: Transmission gate based DFF.

Another well-known way to implement a DFF is utilizing the True Single-Phase Clocking (TSPC) topology, shown in Figure 6.4. Using a technique with ”split” outputs [30], the amount of transistors in the clock path is reduced to only 2. Degraded voltage levels at points X and Y, will slow down the transition speed, therefore the clock switch transistors are implemented complementary[31]. This way, the setup time is predicted to be around 20 ps, which is well below the available 95 ps.

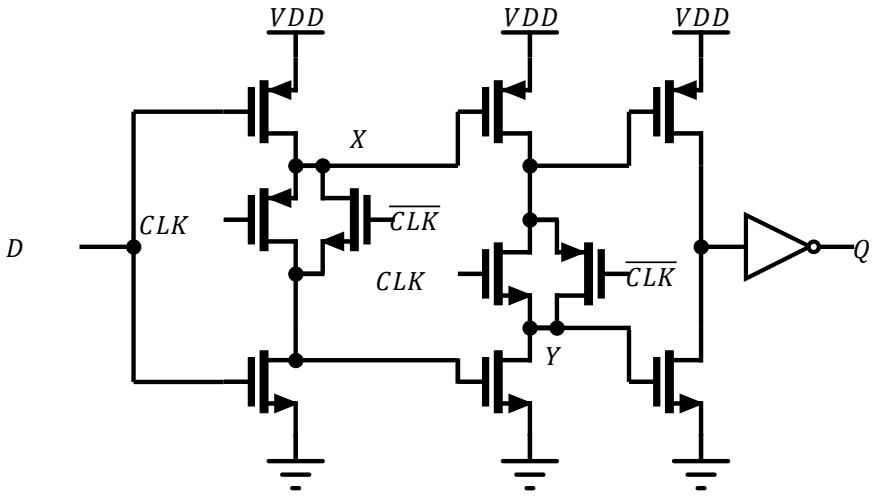


Figure 6.4: TSPC DFF [31].

6.3. Pulse generator

The pulse generator will drive the switches of the 4:1 Multiplexer with a rail-to-rail swing. Therefore it should be able to drive the capacitive load of both the long line and the switching transistor of the 4:1 Multiplexer. It should switch fast enough so the 25% duty cycle can be generated with sufficient steep edges. The basic goal of the pulse generator is a simple 3-input AND operation, $D^{<out>} = \Phi_A \& \Phi_B \& C^{<in>}$. Combining two incoming 50% phases to one 25% pulse only if the input $C_{<in>} = 1$. To implement this some designs use custom circuits [2][25]. However, these circuits are really hard to tweak to make sure the delay paths between Φ_A and Φ_B are equal. A much simpler approach would be to use logic 2-AND ports and configure them such so the delay of $\Phi_A \& \Phi_B$ and $\Phi_B \& C_{<in>}$ are equal, like what has been done in [32]. The pulse generation circuit is shown in Figure 6.5. It has been designed with minimum size LVT transistor-based AND ports for optimal speed followed by a buffer to provide enough power to drive the load line and the Direct 4:1 Multiplexer.

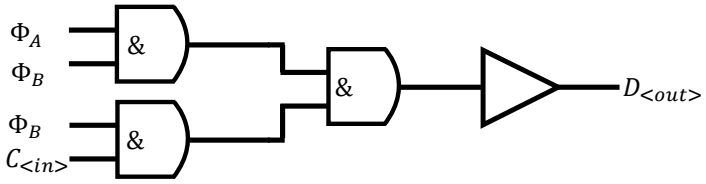


Figure 6.5: Pulse generator circuit.

6.4. Direct 4:1 Multiplexer

Making a high-speed Multiplexer at 25.6 Gb/s calls for a CML implementation. There are several reasons for this choice. The most important is CML Multiplexers can reach much higher speeds than conventional logic Multiplexers in 40-nm technology due to the fact a current steered implementation can switch faster than a voltage swing circuit. A second advantage here is that the output of the CML Multiplexer can be directly used as input of the CML output driver since it converts the rail-to-rail input inherently to the required differential common-mode signal. The last reason is that only a single NMOS transistor is needed to switch the current, thereby minimizing the parasitic capacitance and thus having a higher bandwidth than any other stacked implementation. Figure 6.6 shows the implemented Direct 4:1 Multiplexer circuit.

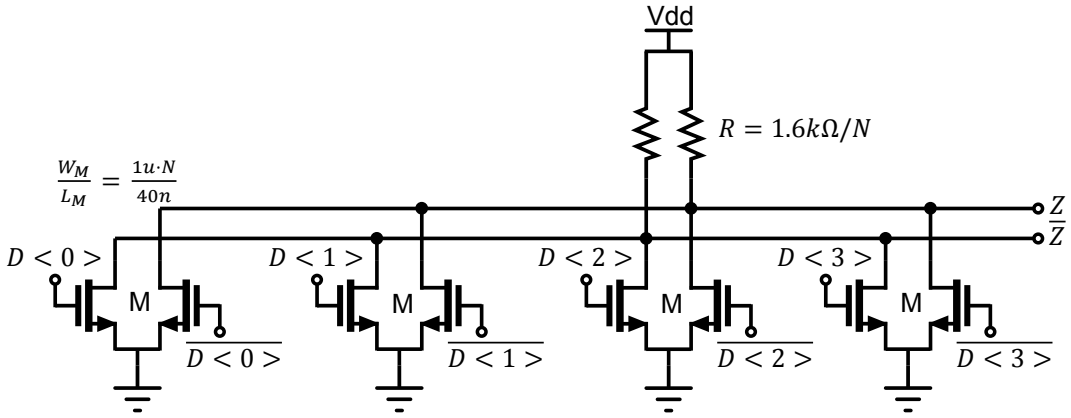


Figure 6.6: Direct 4:1 Multiplexer circuit.

The sizing of the circuit is done with bandwidth and sufficient swing in mind. The required swing should be large enough to completely switch the next output driving DAC, as the NMOS transistors of the differential pair must operate only in saturation to guarantee high-speed operation. To satisfy this requirement, the differential voltage swing must exceed the $\Delta V_{in,max}$ of the next DAC driver [33].

$$R_D I_{SS} \geqslant \sqrt{\frac{2I_{tail_{DAC}}}{\mu_n C_{ox} \frac{W}{L}_{DAC}}} \tag{6.1}$$

However, the driver DAC bits will differ in size, meaning the input capacitance for the LSB is about 8× smaller than the MSB bit. To keep a constant delay, the resistance of the Direct 4:1 Multiplexer should be scaled with the load capacitance of the following DAC bit. To keep the bandwidth sufficient, the following equation must hold.

$$BW = \frac{1}{2\pi R_M C_M} \leq 0.7 \cdot \text{ Bitrate } = 0.7 \cdot 25.6\text{GHz} \tag{6.2}$$

Since the input capacitance for the LSB driver bit plus the line is approximately $5fF$ and increases each unary bit up to $C_M = N \cdot C_{M_{LSB}} = 8 \cdot 5fF = 40fF$. This means the resistance should be scaled accordingly to keep a constant delay $\mathrm{T} = \frac{1}{BW}$, the resistance should be $R_M = R_{M_{LSB}}/N = 1.6k\Omega/N$. However, as discussed before the output swing should stay constant for each DAC bit to guarantee complete switching, meaning for a lower R_M the current should be increased to compensate. Using

all the same NMOS switching transistors and the required swing being over 400 mV, this leads to the width of the current switches being scaled to $\frac{W_M}{L_M} = \frac{1\mu m \cdot N}{40nm}$.

6.5. Layout

The layout of the High-speed retimer and Direct 4-to-1 Multiplexer is shown in Figure 6.7. The data pulses $D_{0-3}, \overline{D_{0-3}}$ comes in from the right where it was serialized to a speed of 6.4 Gb/s and then moves to the left into the high-speed retimers. The retimers work with the 4 clock phases which are distributed to all the high-speed retimers by a higher metal clock tree. The clock tree starts from the left top and gets buffered in the middle, it then splits into branches of 2 and 3 until it reaches the Multiplexers indicated with green arrows. The retimed data then flows from the right to the left through a lower metal line underneath the clock tree into the Direct 4-to-1 Multiplexer, indicated with pink arrows. In the 4-to-1 Multiplexer, these separate pulses get multiplexed into a single high-speed 25.6 Gb/s signal ready to go into the DAC on the left.

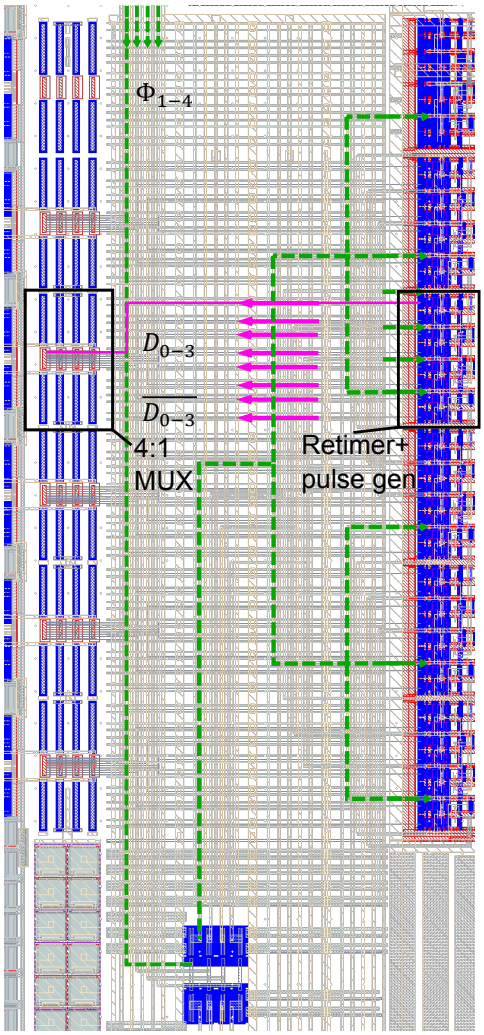


Figure 6.7: Layout High-speed retimer and Direct 4-to-1 Multiplexer.

6.6. Results

The layout has been extracted and a functional time-domain simulation was run with one retiming, pulse generation and 4:1 multiplexing slice. The result is shown in Figure 6.8, where the top shows all the 25 % duty cycle pulses generated and the bottom illustrates the differential output of the Multiplexer.

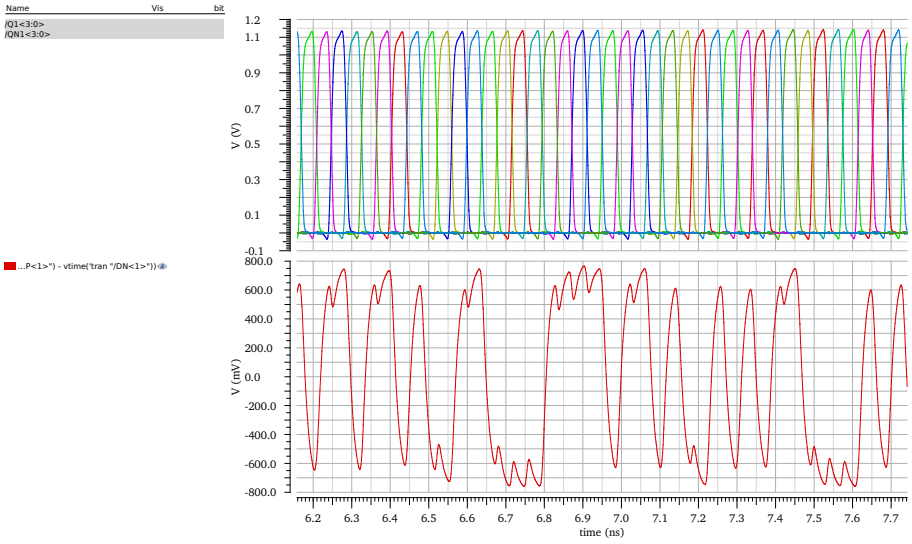


Figure 6.8: Extracted simulation of high-speed retimer pulse output and Direct 4:1 Multiplexer differential output.

Power The average power consumption breakdown is shown in table 6.1. The total power consumption of the 4:1 Multiplexer, retimer and pulse generator for all 10 slices is 16.3 mA at 1.1 V nominal.

Part	Current @1.1V
Retimers and pulse generation	5.119 mA
Direct 4:1 Multiplexer	11.19 mA
Total	16.3 mA

Table 6.1: Power consumption retimers, pulse generators, and 4:1 Multiplexers.

6.7. Conclusion

This chapter showed the design of the high-speed retimer, pulse generator, and Direct 4:1 Multiplexer. It is critical for the system to retime the incoming data correctly so no bit errors appear at the output, a TSPC DFF was designed to do this at the required 6.4 Gb/s data rate. Then a pulse generator was designed to prepare the data for the ”Direct” 4:1 Multiplexer to generate a reliable 25.6 Gb/s output for the DAC. The results show a successful generation of the pulses and output signal.

7

DAC

The Digital-to-Analog Converter (DAC) has to be designed to drive the wireline with sufficient swing, linearity, and bandwidth.

This chapter is ordered as follows: Section 7.1 will introduce the two main driver architectures and justify the choice for CML. Section 7.2 will go through the design of the DAC in detail based on the swing, linearity, and bandwidth requirements. Section 7.3 shows the design of the biasing circuitry for the LSB element. Section 7.4 addresses the design of the output termination and peaking inductor. The layout of the DAC is shown in Section 7.5. The resulting eye diagram is discussed in Section 7.6.

7.1. Architecture

The DAC will be the final stage of the wireline transmitter, and therefore, the driver of the load. This means for a given voltage swing, the driver needs to deliver the voltage/current to achieve the required swing over the output resistor. There are two categories of output drivers, (1) the Current Mode Logic driver (CML) driver steering current to generate logic levels and (2) voltage-mode drivers or Source Series Terminated driver (SST) directly putting voltage on a matched termination load. There are also hybrid drivers combining both current and voltage.

7.1.1. CML driver

The Current Mode Logic driver (CML) driver generates a voltage swing over the load resistor by driving a current through the line. An example implementation of a CML driver is shown in Figure 7.1. In this case, the structure is simplified to a 2-bit DAC to generate the 4 levels required for PAM4. Each bit consists of a switching pair, and a tail source, together these are connected to one termination resistor. The tail source generates the current for the required level swing on the load and termination resistor, while the switching pair turns on either of the branches to steer the current in a positive or negative direction. In this way, a differential swing can be generated on the load resistors.

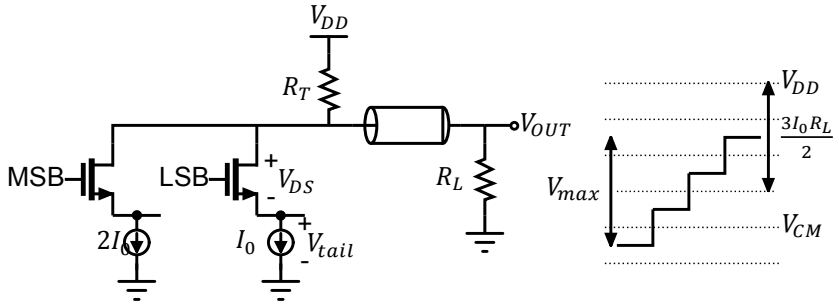


Figure 7.1: Single-ended 2-bit CML driver [13].

Figure 7.1 shows only one side of the driver to illustrate the single-ended swing over the load resistor R_L. On the driver side also a termination resistor is chosen equal to the load $R_T = R_L = 50\Omega$. The

common-mode voltage level at the output node will be equal to $V_{CM} = V_{DD} - 3I_0R_T/2$, and the single-ended swing $V_{max} = 3I_0(R_T||R_L) = 3I_0R_L/2$. With this, the minimum required supply voltage can be calculated.

$$V_{DD\,\min} = \frac{3IR_L}{4} + V_{\max} + V_{DS} + V_{tail} \tag{7.1}$$

Here, V_{DS} is the minimum allowable drain-source voltage for the switching transistor, which will be approximately $V_{DS} = 0.1V$. V_{tail} is the minimum allowable voltage for the current source and approximately $V_{tail} = 0.3V$. For a peak-peak swing of 800 mV, $V_{max} = 0.4V$ and the total power consumption for the driver can be calculated as follows:

$$\begin{aligned} P_{driver} &= (V_{DDmin} \cdot 3I = 1.5V_{\max} + V_{DS} + V_{tail}) \cdot \frac{2V_{\max}}{R_L} \\ P_{driver} &= (V_{DDmin} \cdot 3I = 1.5 \cdot 0.4 + 0.3 + 0.1) \cdot \frac{2 \cdot 0.4}{50} = 16mW \end{aligned} \tag{7.2}$$

This structure has a few disadvantages. Firstly, the current is always flowing through the tail source, meaning the power consumption is high. Secondly, the tail source limits the maximum achievable swing due to the required V_{DS} on the tail source. The main advantage of a CML DAC is the high speed it can achieve. Due to the use of only NMOS transistors and relatively small required input swing, the parasitic input capacitance of the DAC is small. This allows for a large bandwidth compared to other driver techniques and hence can achieve very high speeds.

7.1.2. SST driver

A Source Series Terminated driver (SST) driver can generate a voltage level on the load by switching the supply voltage 'on' or 'off' of the chosen series resistors. An examples structure of a 2-bit DAC implementation is shown in Figure 7.2. The SST DAC consists of two inverters and two resistors. The MSB resistor is scaled to $R_{MSB} = 1.5R_L$, and LSB resistor to $R_{LSB} = 3R_L$. This will yield the maximum single-ended swing of $V_{max} = V_{DD}/2$, retaining the correct termination of $R_{LSB}||R_{MSB} = R_L$.

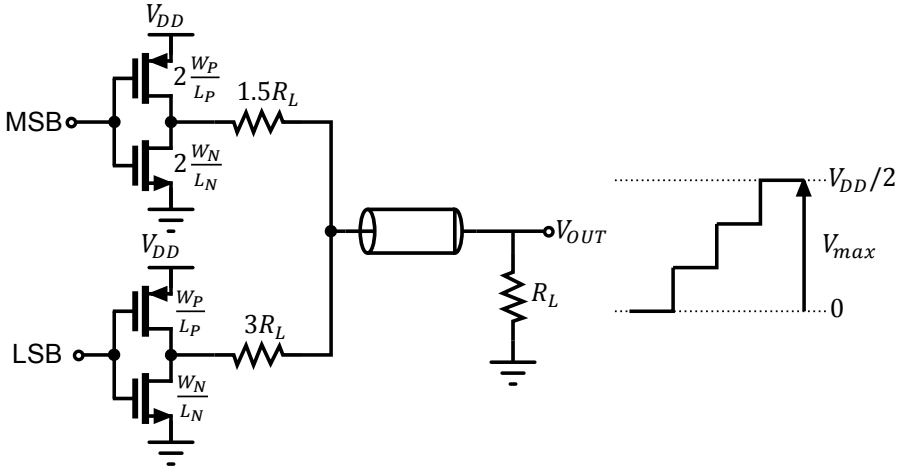


Figure 7.2: 2-bit SST driver [13].

The output voltage swing of the SST driver is determined by the supply voltage. Assuming the same voltage swing $V_{max} = 0.4V$ the required supply voltage will need to be $V_{DD} = 0.8V$. Since the switches either turn on or off, the power consumption is different for MSB and LSB levels. However, for PAM4 these will happen with equal probability. The power consumption can be approximated as follows [13]:

$$P_{\text{driver}} = \frac{13V_{\max}^2}{9R_L} = \frac{13 \cdot 0.4^2}{9R_L} = 4.6\ \text{mW} \tag{7.3}$$

This power consumption is less than half compared to the similar swing CML type driver. However, looking only at the power consumption of the last stage is not very representative. The biggest issue

of this structure is that the on-resistance of the transistors is part of the termination resistance. This means the r_o of the output transistors for both PMOS and NMOS need to be sized to reach a total output resistance of R_L. To be exact $r_{oP} + R_{LSB} = r_{oN} + R_{LSB} = 3R_L$, e.g. choosing $R_{LSB} = 100\Omega$, would require $r_{oP} = r_{oN} = 50$, which requires transistor sizes of many μm. This means the inverters present a large input capacitance requiring much more powerful pre-drivers than for a CML. Also a larger output capacitance limits the bandwidth and therefore rapidly degrades the eye diagram at high data rates.
Another problem arises with PVT variations of the transistors directly influencing the output impedance and thus directly degrading the performance. This could be a serious issue cooling down to cryogenic temperatures.
Lastly, the current pulled by the driver from V_{DD} will depend on the output voltage and could fluctuate at the data rate, so a bond wire of a few hundred pHs will already cause significant ringing.

7.2. 6b DAC

The issues, limited bandwidth, and complexity of an SST DAC were the main reasons to decide to design a CML driver. The CML DAC is often called the current-steering DAC and has a similar design procedure. The design is ordered in the steps taken to achieve the system requirements in terms of RLM, mismatch, swing, and bandwidth.

7.2.1. Number of bits

For normal current-steering DACs, linearity is important to accurately generate a sinusoidal waveform without spurious tones. For a PAM4 system, the Ratio of Level Mismatch (RLM) determines the performance since any deviation in the levels will degrade the horizontal eye-opening. In 2.6, the definition of the RLM was stated as the difference between the smallest eye height divided by the full swing. Since for an ideal PAM4 signal, the eye height is a third of the total swing A, introducing an error Δ into the equation leads to a RLM of

$$RLM = \frac{3 \cdot \min\left(\Delta + \frac{A}{3}\right)}{A} > 0.95 \tag{7.4}$$

Since the requirement of the RLM was at least 95% the maximum error should be

$$\Delta < \frac{0.05A}{3} = \frac{A}{60}. \tag{7.5}$$

Then assuming the error $\Delta = \frac{1}{2}LSB$ of the DAC, the minimum number of bits can be calculated by

$$n = \log_2 \frac{A}{LSB} = \log_2 \frac{1}{30} = 4.9. \tag{7.6}$$

The required number of bits is 4.9. However, since there are also other errors other than nonlinearity that could degrade the performance like jitter, it is better to take a margin. In this design, a 6-bit DAC is chosen.

7.2.2. Current-source Resistance

The largest contributor to the nonlinearity is the finite output resistance of the DAC current sources. A current-steering DAC with non-ideal current sources is shown in Figure 7.3. Depending on the output code, the resistance at the output varies from $R_L||r_O$ with only one switch turned on to $R_L||(r_o/N)$ with all switches turned on. The output voltage over the load resistance will thus vary from $I_0(R_L||r_O)$ at the lowest to $N \cdot I_0(R_L||(r_o/N))$ at the highest code. The output characteristic will show a compression as the amplitude codeword increases, leading to a maximum INL of $I_0 R_L^2 N^2/(4r_O)$. Normalized to the output voltage, this maximum is approximately equal to $I_0 R_L N/(4r_O)$. Since the required RLM is 95%, the INL should be lower than 5%. So with a $R_L = 50\Omega$ and $N = 2^6 = 64$, the required LSB current source resistance should be at least $r_O = 16k\Omega$.

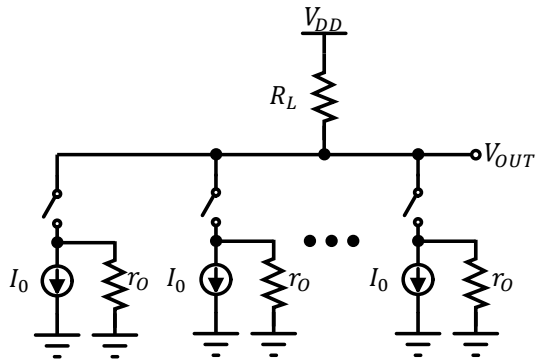


Figure 7.3: Finite current-source resistance.

An output resistance of $r_o = 16k\Omega$ is hard to reach with a single tail transistor, so it is better to use a cascode structure. The cascode tail current source implementation is shown in Figure 7.4. M2 will boost r_{o1} by its intrinsic gain and results in a total resistance of

$$\begin{aligned} R_{out} &= r_{O1} r_{O2} g m_2 \\ &= \frac{L_1}{\lambda_0 I_D} \frac{L_2}{\lambda_0 I_D} \sqrt{\frac{2\mu_n C_{ox} W_2}{I_D L_2}} \end{aligned} \tag{7.7}$$

The drain current through the transistor can be rewritten to fit this equation:

$$I_D = \frac{1}{2}\mu_n C_{ox} \frac{W}{L} V_{GT}^2 \rightarrow \frac{\mu_n C_{ox} W_2}{L_2 I_D} = \frac{2}{V_{GT,2}^2} \tag{7.8}$$

Since the total output resistance should be larger than $R_{OUT} = 16k\Omega$, $V_{GT} \approx 0.2V$, $\lambda_0 \approx 0.1\mu m/V$ the required transistor length can be approximated.

$$R_{out} = \frac{2L_1 L_2}{\lambda_0^2 I_D V_{GT,2}} > 16k\Omega \rightarrow \quad L_1 L_2 > 0.004\mu m^2 \tag{7.9}$$

The parasitic capacitance of M2 should be kept as low as possible since this directly degrades the output bandwidth [34]. Therefore M2 is chosen minimum size $L_2 = 40nm$ and the length of the lower transistor M1 can be $L_1 = 400nm$ to be on the safe side for output resistance.

$$L_1 = 400nm, L_2 = 40nm \rightarrow L_1 L_2 = 0.016\mu m^2 \tag{7.10}$$

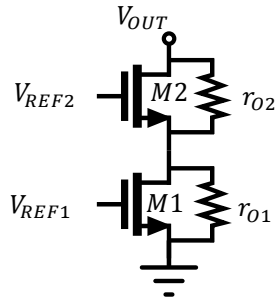


Figure 7.4: Cascode Tail current source.

7.2.3. Swing

SNR should be reached with a maximum FFE of 7 dB for compensating the channel and assuming 0.95 RLM. In the requirements the Signal-to-Noise Ratio should be at least $SNR_{\min} = 32.5dB$. To find the required swing, first the noise power should be approximated.

$$P_n = 2 \cdot 4kT\left(\gamma g_m + \frac{1}{R_D}\right)R_D^2 \approx 8kT\left(1 \cdot 0.02 + \frac{1}{50}\right)50^2$$
$$= 3.3 \cdot 10^{-18}\frac{V^2}{Hz} \approx -144.8\frac{dBm}{Hz} \tag{7.11}$$

The signal power can now be approximated subtracting the noise floor from the minimum SNR requirement $SNR_{\min} = 32.5\,dB$. Additionally the inherent losses in swing due to FFE channel compensation $L_{FFE} = 7\,dB$, plus the effective reduction of eye height due to non-linearity $L_{RLM} = 0.95 = 0.4\,dB$ have to be taken into account. Now, the signal power will be:

$$P_{\text{sig}} = P_n + SNR_{\min} + L_{FFE} + L_{RLM} + 10\log_{10} B$$
$$= -144.8\frac{dBm}{Hz} + 32.5dB + 7dB + 0.4dB + 104dB = -0.9dBm \tag{7.12}$$

$$V_{\text{swing diff}} = \sqrt{10^{-0.09-3} \cdot 8 \cdot 100\Omega} = 0.806V_{pp} \tag{7.13}$$

For a 6 bit current-steering DAC, this means the LSB current I_{LSB} will be:

$$I_{LSB} = \frac{V_{LSB}}{R_L} = \frac{V_{swing_{diff}}/2^6}{50\Omega} = 0.25mA \tag{7.14}$$

7.2.4. Mismatch

INL Designing the sizes of the DAC transistors begins with the unit cell, the tail current-source transistor must be sized and biased to guarantee a maximum static INL lower than 0.5 LSB. The INL specification mainly depends on random mismatches and code-dependent output resistance. The maximum INL can be defined as

$$INL_{MAX} = \frac{1}{2}\sqrt{2^n - 1}\frac{\sigma_i}{\Delta I}, \tag{7.15}$$

where σ_i is a random error in the LSB current ΔI and n is the DAC's number of bits. A cascoded current source will have a threshold mismath equal to

$$\sigma_{\Delta VT} = \frac{A_{VT}}{\sqrt{WL}}, \tag{7.16}$$

where A_{VT} is a process-dependent parameter, which is about $3mV\mu m(\text{NMOS})$ for 40-nm process.

It is known that two relatively large identical current sources will have a current mismatch equal to the product of the threshold mismatch $\sigma_{\Delta VT}$ and the trans-conductance:

$$\sigma_i = \sigma_{\Delta VT} gm = \frac{A_{VT}}{\sqrt{WL}}\sqrt{2\mu_0 C_{ox}\frac{W}{L}I_D} = \frac{A_{VT}}{L}\sqrt{2\mu_0 C_{ox} I_D} \tag{7.17}$$

Substituting equation 7.17 into 7.15 will give a size requirement for the current source.

$$INL_{MAX} = \frac{1}{2}\frac{A_{VT}}{L}\sqrt{\frac{2\mu_0 C_{ox}}{I_M}}\,(2^n - 1) \tag{7.18}$$

where I_M is the total required current to reach full swing. Rewriting this equation to extract the length L results in a minimum length of:

$$L = \frac{1}{2}A_{VT}\sqrt{\frac{2\mu_0 C_{ox}}{I_M}}\frac{2^{ntot} - 1}{INL_{MAX}}$$
$$L = \frac{1}{2}3mV\mu m\sqrt{\frac{2 \cdot 220\mu A/V^2}{16mA}}\frac{2^6 - 1}{0.5} = 31\,\text{nm} \tag{7.19}$$

31 nm is smaller than the minimum 40-nm length available. Thus, for this 6-bit design, the individual mismatch of the current sources will not be a problem. The requirement from the current-source resistance dominates the choice for a longer length tail transistor. The transistors shown in Figure 7.4 are chosen $L_2 = 40\,nm$, $L_1 = 400\,nm$. A Monte Carlo simulation has been run to check if the LSB current source would indeed be sufficient in size to stay below the calculated mismatch. The resulting histogram for the current source is shown in Figure 7.5. The Mean LSB current is 0.25 mA with a standard deviation of $\sigma_i = 17\,\mu m$. These results can be filled into equation 7.15, $INL_{MAX} = \frac{1}{2}\sqrt{2^6 - 1}\frac{17\,\mu A}{0.25\,mA} = 0.27$, this is well below the required 0.5 LSB.

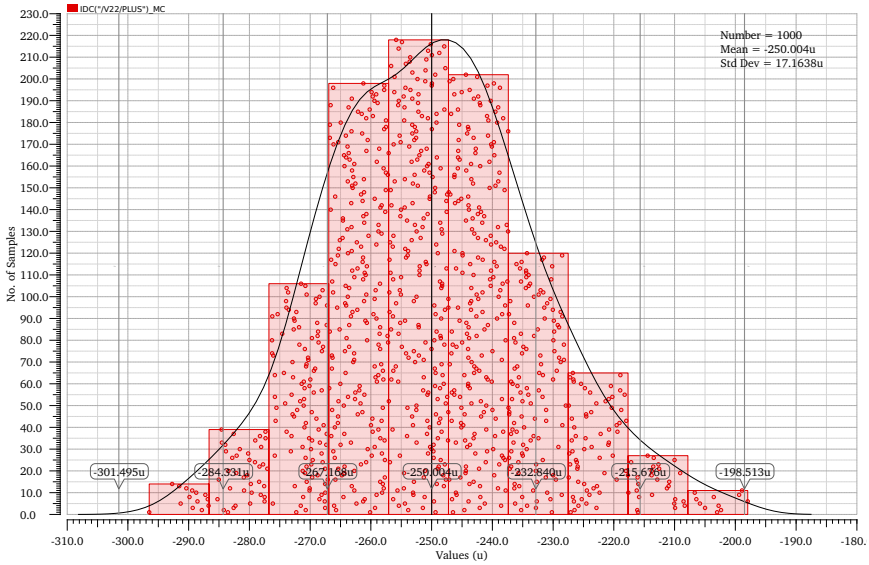


Figure 7.5: Monte Carlo simulation of a single post-layout extracted LSB current source, the histogram shows the LSB current distribution of a 1000 samples.

DNL For a complete unary coded Current-steering DAC, the DNL errors are similar to the above calculated INL, since any major transitions are prevented. However, a complete unary implementation would require $2^6 - 1 = 63$ separately switchable current sources, which for high-data rates cost a lot of digital power. Therefore a mix of both binary current sources for the small and unary current sources for the major transitions is considered. Hereby the number of unary bits should be minimized as long as the DNL performance is within 0.5 LSB. The maximum DNL error will occur at the transition from the binary section to the next unary section, as indicated in Figure 7.6. As stated in [35] the variance can be estimated as follows:

$$\sigma^2_{\text{one DNL}} = \frac{2A_{VT}^2 2^{2N_{\text{binary}}}}{(WL)_{\text{unary}}\,(V_{GS} - V_T)^2} + \frac{2A_{VT}^2\left(2^{N_{\text{binary}}} - 1\right)^2}{(WL)_{\text{all binary}}\,(V_{GS} - V_T)^2} \tag{7.20}$$

For this application, the gate area of the binary section is similar to the gate area of one unary section. The DNL caused by the major transition is then approximately:

$$\sigma_{\text{one DNL}} \approx 2^{N_{\text{binary}}+0.5}\frac{\sigma_{\text{Iunary}}}{I_{\text{unary}}} \approx \frac{2A_{VT}2^{N_{\text{binary}}}}{\sqrt{(WL)_{\text{unary}}}\,(V_{GS} - V_T)} \tag{7.21}$$

Filling the current source size and process-dependent parameters into this equation and trying different versions of a 6-bit DAC gives the following DNL results

$$\begin{aligned} \sigma_{\text{one } DNL} &= 2 \cdot \frac{3mV\mu m \cdot 2^{N_b}}{\sqrt{400n \cdot 30\mu} \cdot (110m)} \\ \sigma_{1DNL}\,(N_b = 6) &= 1.0LSB \\ \sigma_{1DNL}\,(N_b = 5) &= 0.5LSB \\ \sigma_{1DNL}\,(N_b = 4) &= 0.25LSB \end{aligned} \tag{7.22}$$

It is preferred to have a DNL smaller than $\sigma_{1DNL} < 0.5LSB$, which is the case for a $N_b = 4$ bit binary and $N_u = 2$ bit unary DAC.

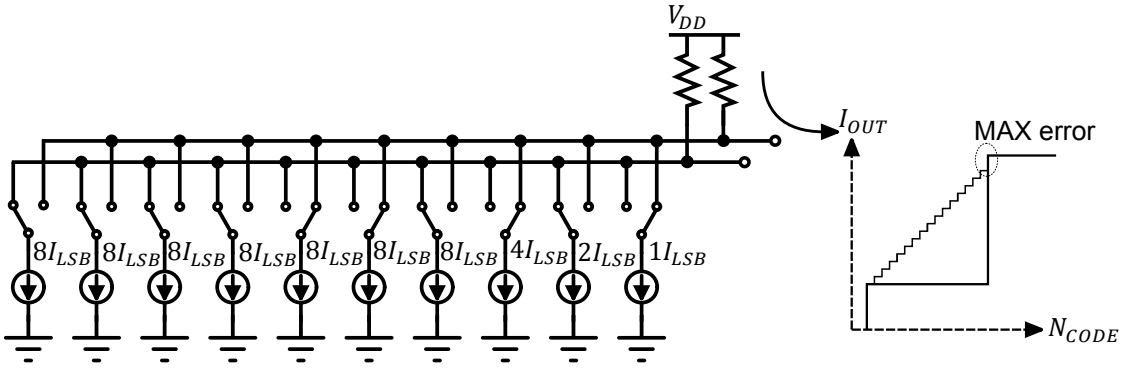


Figure 7.6: Ideal 4b binary, 2b unary current-steering DAC, maximum DNL transition [36].

7.3. Bias

The biasing circuit is designed to provide the reference voltages for all cascode current sources of the DAC. The length of the transistors M1 and M2 was already determined by the linearity requirements $L_{M1} = 400\,\mathrm{nm}, L_{M2} = 40\,\mathrm{nm}$. The width of the transistors should be sized such that with a limited overdrive voltage they are still able to provide the $I_{LSB} = 0.25\,\mathrm{mA}$ current. To be able to still reach the 800 mVpp swing at the output node, the total available drain-source voltage for the cascode is $V_{tail} = V_{DD} - V_{DS_{switch}} - V_{swing} = 1.1 - 0.1 - 0.4 = 0.3\,V$. Both devices need to be in saturation to drive the current so the W/L of both transistors M1 and M2 have a saturation voltage of $V_{Dsat} \approx 150\,\mathrm{mV}$, leading to $W_{M1} = 30\,\mu\mathrm{m}, W_{M2} = 200\,\mathrm{nm}$. Two PMOS current mirror branches will mirror the reference current, which is approximately 5x smaller than the LSB current, so $I_{ref} = I_{LSB}/5 = 50\,\mu\mathrm{A}$. The reference voltages V_{ref1} and V_{ref2} are then generated by NMOS transistors in the diode configuration and sized equal to the cascode tail source. Figure 7.7 shows the complete circuit.

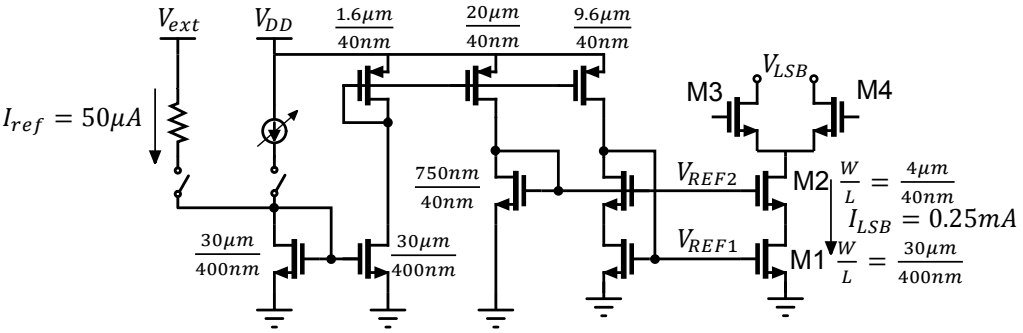


Figure 7.7: Biasing circuit for LSB current source.

The reference current I_{ref} can be selected to be generated by an external reference voltage or using an onboard 6 bit current DAC. The reason for using a variable source is that at cryogenic temperatures, the threshold of the tail devices could increase. This means that the swing is limited if not compensated for by a higher reference current. The reference current DAC has a range from approximately $10\mu A$ to $100\mu A$.

7.4. Termination

The termination of the output driver is not an ideal 50Ω termination. Multiple parasitics will limit the output bandwidth. First, the current DAC driver has a parasitic capacitance, mainly due to the large switching and tail transistors. Secondly, the pad has a parasitic capacitance due to its large size and incorporation of ESD protection. Lastly, the bond wires that bring the signal from the chip to the board introduce a series inductance of approximately $L_{bond} = 100pH$.

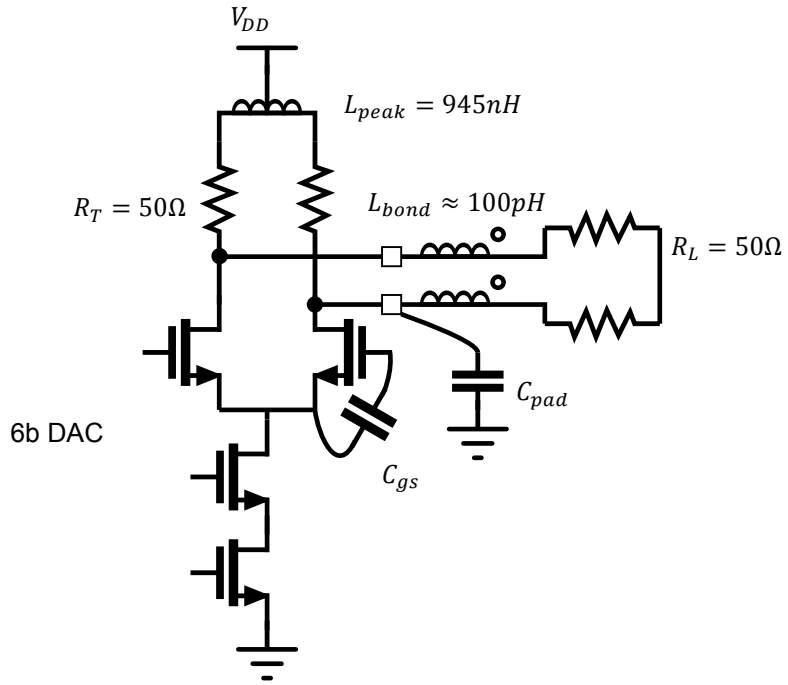


Figure 7.8: Termination circuit.

The termination circuit is shown in Figure 7.8. Due to the ”switching” of the two branches the effective capacitance observable from the output node becomes the C_{gs} of the switching transistors M3/M4 [34]. The total capacitance seen at the output can be calculated by adding up the total C_{gs} for all DAC bits and the post-Layout RC extracted pad capacitance.

$$\begin{aligned} C_{gs(LSB)} &= 15.6\text{fF} \\ C_{PAD} &= 198\text{fF} \\ C_{\text{tot}} &= 63 \cdot C_{gs(LSB)} + C_{\text{PAD}} = 1.18pF \end{aligned} \tag{7.23}$$

The total capacitance will limit the bandwidth. To improve the bandwidth of the system, the inductive peaking technique is employed by adding an inductor L_{peak} in series with the termination transistors with a center tap to V_{DD}. As discussed in section 2.8.2 for a series peaking inductor, the size can be calculated with the following equation.

$$L = mR^2C = 0.33 \cdot 50^2 \cdot 1.18\,\text{pF} = 975\,\text{nH}, \tag{7.24}$$

where $m = 0.33$ is chosen to get a Maximally flat envelope delay which is ideal for this optimally square wave application. The complete extracted output termination has then been simulated to see the bandwidth improvement for the complete system. The result is shown in Figure 7.9, the 3 dB bandwidth at half of the output impedance is increased by more than 6 GHz after adding the peaking inductor.

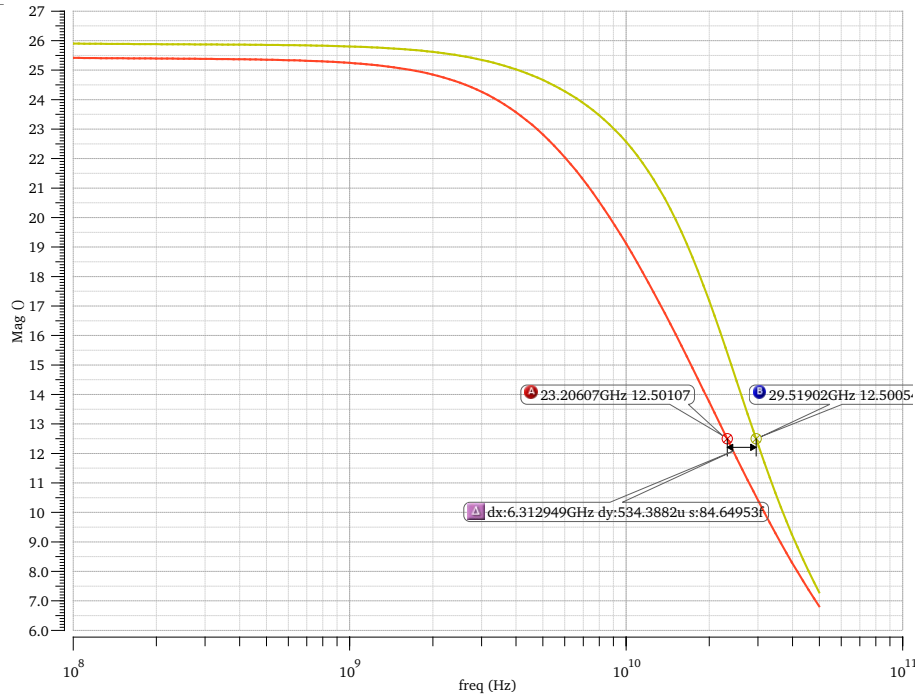


Figure 7.9: Single-ended Z11 at output node, without L_{peak} inductor (red), with L_{peak} inductor (yellow).

7.5. Layout

The layout design of the DAC has been made with symmetry and mismatch in mind. The DAC consists of 10 elements with each $8 \times I_{LSB}$ current sources. For the unary bits all $8 \times I_{LSB}$ sources are connected, while for the binary the unused sources are replaced with dummies. Moreover, two extra elements of $8 \times I_{LSB}$ dummy sources are located on the top and bottom of the DAC to improve mismatch. Each LSB current cell/source is designed symmetrically, where the switching transistors(M3, M4) and signal lines are located in the middle and the larger size current sources(M1, M2) are split into two and located on the top and below the switches. In this way, the horizontal size is minimized to prevent long lines or large capacitance. The effective current source size will stay the same as splitting the current source and attaching it in parallel will effectively have the same $W \cdot L$. The biasing circuit is located on the bottom with the transistors aligned in the same direction, to minimize any mismatch from the POLY gates being oriented in a different direction. Figure 7.10 shows the complete DAC structure with termination. The incoming multiplexed 25.6 Gb/s data signals come in from the right and are amplified by the DAC drivers to drive the termination. The termination lines are designed as a clock tree to minimize any time skew between the outputs. The termination resistors R_T and peaking inductor L_{peak} are located on one side of the line and connected to V_{DD} with a large amount of decoupling capacitors on the bottom. The output lines then go directly to the pads on the left which are placed in a GSGSG form, so these can be probed with a high-speed probe station.

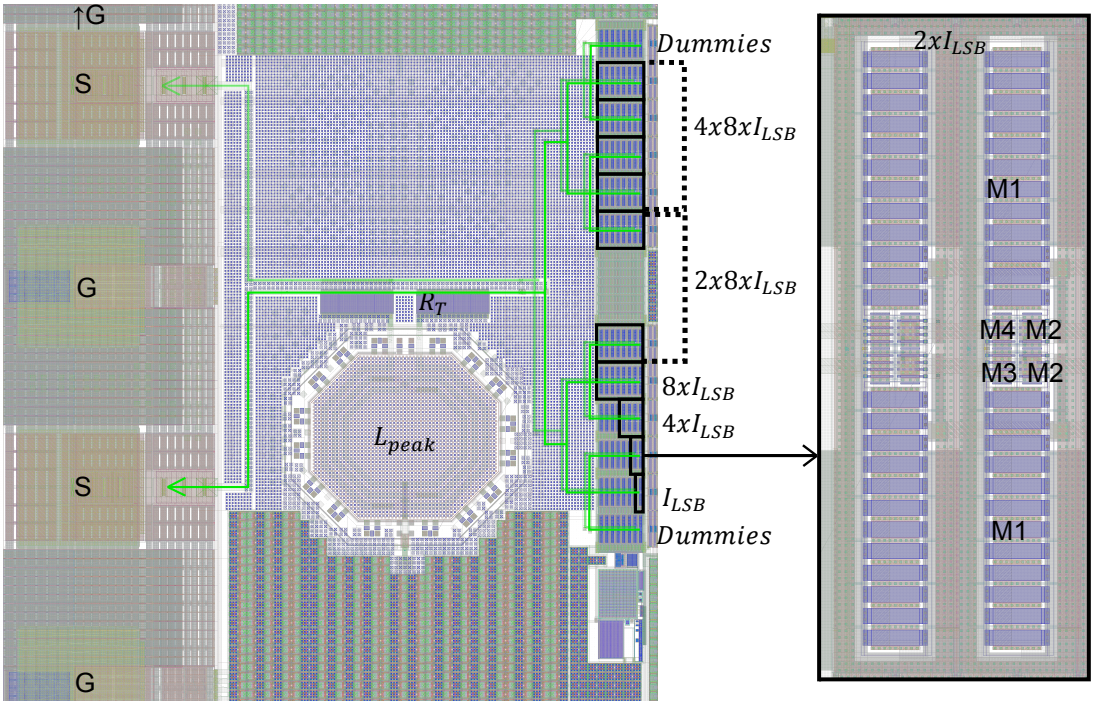


Figure 7.10: DAC and Termination layout.

7.6. Results

All the system blocks of the termination, DAC, and retimer structures have been designed in layout and extracted. Then a complete simulation was run with random input data for 10 ns to plot an eye diagram shown in Figure 7.11. In this simulation, there is no FFE or other digital pre-distortion used, the PAM4 signal is purely generated by randomly changing the even and odd bits to make the 4 differential levels.

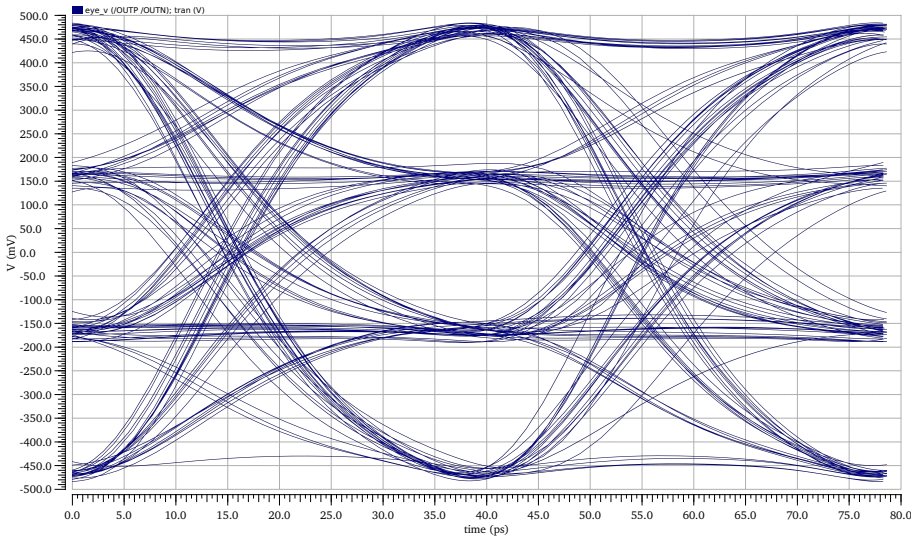


Figure 7.11: Extracted simulation DAC and Termination output, eye diagram.

The total maximum swing is $471.4\,mV + 468.2\,mV = 939.5\,mVpp$, but this is assuming an ideal 50Ω load. So it is good that there is some margin over the required 800 mVpp, as in reality, the load might be different resulting in a smaller swing at the output.

The RLM can be calculated using equation 2.6 and concluding from the eye diagram $V_0 = -468.2\,mV, V_1 = 167.3\,mV, V_2 = -167.5\,mV, V_3 = 471.4\,mV$:

$$RLM = \frac{3 \cdot min(V_3 - V_2, V_2 - V_1, V_1 - V_0)}{V_3 - V_0} = 96\%$$

This is sufficiently over the RLM requirement of 95% and could be still improved by adjusting the settings of the 6-bit DAC.

However, in the time domain measuring the vertical eye opening is not representative for determining jitter. This jitter consists of the random jitter of the clock, the introduced phase noise and mismatch of the pulse generators, the clock skew between the phases, the clock skew of the clock tree, and lastly, any additional jitter from the DAC and termination circuitry. To determine what this would approximately be, a maximum swing clock signal of 25.6 GHz was generated at the output of the DAC. The resulting eye diagram is shown in Figure 7.12, the total measured peak-to-peak jitter is about $J_{pp} = 400\,\text{fs} \approx 0.01\,\text{UI}_{\text{pp}}$. This is meeting the specification of Uncorrelated Bounded High Probability Jitter $T_{UBHPJ} = 0.05\,\text{UI}_{\text{pp}}$.

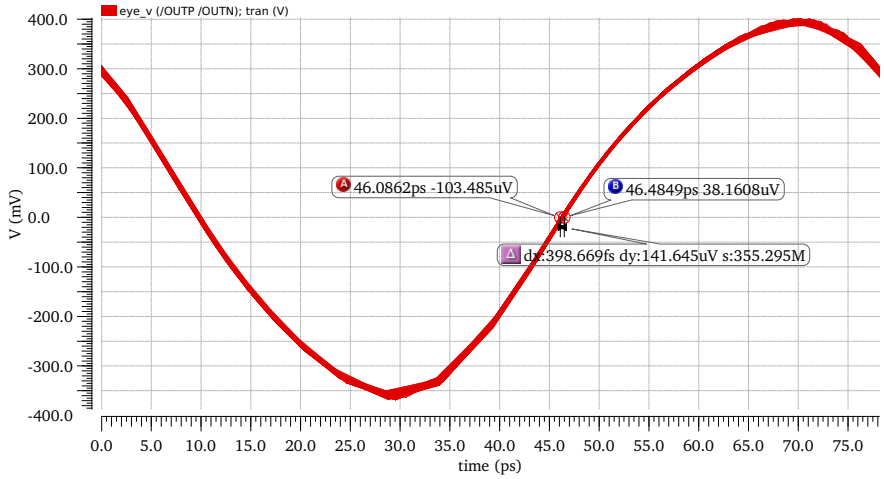


Figure 7.12: Extracted simulation of clock signal DAC output, to measure total peak-to-peak jitter.

Power The total power consumption breakdown of the DAC is shown in Table 7.1.

Part	Current @1.1V
Bias	655.7 uA
Termination	28.88 mA
Total	29.54 mA

Table 7.1: Power consumption DAC.

7.7. Conclusion

The goal of this chapter was to design a DAC capable of reaching the swing, linearity, and bandwidth requirements. First, a CML driver has been chosen for the high speed and relatively simple design. Then the current steering DAC and its biasing were designed based on fundamental calculations of linearity, swing, and mismatch. A series peaking inductor was chosen to boost the bandwidth of the output termination and the layout of the complete DAC was simulated to prove its eye diagram could reach the requirements.

8

Conclusion

This thesis set out to analyze the requirements of a cryogenic wireline transmission system and provide a high-speed transmitter design solution in 40-nm technology. The design of A DAC-based Cryo-CMOS 51.2 Gb/s PAM4 Wireline Transmitter has been presented and simulation results show that this design is suitable to meet the requirements.
The thesis outcomes are summarized in Section 8.1, suggestions for future research are given in Section 8.2.

8.1. Thesis Outcome

The chapters within this thesis worked towards the goal of analyzing the requirements of a cryogenic wireline transmission system and come up with a high-speed transmitter design. Multiple challenges were addressed such as the limitations of the 40-nm technology, the limited channel bandwidth, and the influence of a cryogenic environment. All these challenges have been coped with to finally present a design. In this section, the outcomes of each chapter are discussed.
In chapter 2, some wireline communication data formats have been analyzed. The channel has been measured and different equalization techniques have been introduced to compensate for the low pass behavior. From these definitions, the requirements for the design have been set.
Chapter 3 compared multiple state-of-the-art architectures based on their performance. The most striking difference in architecture was the FFE implementation, the most flexible DAC approach was chosen. This 4b unary 2b binary DAC implementation requires the following elements: an SRAM memory block, 10 serializer slices containing a low speed 16:1, and a high speed 4:1 Multiplexer plus the DAC drivers. The serializers all make use of quadrature clock phases which are generated by the clock generation architecture from chapter 4. The clock generation makes use of a CML divide-by-2 for the high-frequency phases and a dynamic C2MOS divide-by-2 for the generation of all other quadrature phases.
The low-speed retimers and Multiplexers were introduced in chapter 5. The retimers were necessary to retime the unknown data coming in from the SRAM memory block. Then a latchless 16:1 Multiplexer design was proposed to serialize the data making use of the phase delay from the available quadrature clocks.
Next in chapter 6 a direct 4:1 multiplexing structure was introduced. This required the data to be retimed and made complementary at a high speed using multiple TSPC DFFs. A pulse generator is required to generate the non-overlapping pulses from the quadrature clock, so the direct 4:1 Multiplexer can serialize the data into a single stream for the DAC drivers.
The designed CML DAC in chapter 7 is in essence a high-speed current steering DAC. The linearity requirements of the RLM mainly define the sizing of the DAC current source elements. To make sure the current sources have a sufficiently large resistance not to cause any distortion on the output eye these should also be biased correctly. The increased mismatch at cryogenic temperatures led to the choice of a 4b binary 2b unary DAC, although requiring more serializer slices and thus higher power consumption. Finally, the termination has been designed with a single tap peaking inductor to be able to reach the bandwidth requirement and simulations show a sufficient eye diagram.

Finally the results from the simulations have been summarized in the requirements table 8.1. To show which specifications have been met. The empty places will have to be determined by measurement.

Characteristic	Symbol	Min.	Goal	Max.	Simulated	Unit
Feed Forward Equalisation Taps	FFE_{taps}	2	3	-	-	-
Data Rate	R_b	36	51.2	58	51.2	Gb/s
Baud Rate	$BAUD$	18	25.6	29	25.6	Gsym/s
Bit Error Rate(pre-FEC)	BER	-	-	1e-6	<1e-6	-
Signal-to-Noise-and-Distortion-Ratio (SNDR)	$SNDR$	31	-	-	>31	dB
Maximum Output Differential Voltage	$V_{PPd_{MAX}}$	-	800	1200	939.5	mVppd
Relative Level Mismatch	RLM	0.95	-	-	0.96	%
Uncorrelated Bounded High Probability Jitter	T_{UBHPJ}	-	-	.05	.01	UIpp
Uncorrelated Unbounded Gaussian Jitter	T_{UUGJ}	-	-	.01	-	UIrms
Power	P_{tot}	-	300	500	85.9*	mW
Insertion Loss	IL	-	$15_{@14GHz}$	$30_{@14GHz}$	-	dB
Operating Temperature	T	4	-	300	-	Kelvin

* Without SRAM and PLL power

Table 8.1: Wireline transmitter requirements and results.

8.2. Future work

This thesis presented many aspects and components of wireline transmitter design, from setting up requirements to breaking down circuit level design. This work leads to some suggestions for future research which are listed in this section.

Algorithms like FEC and FFE and their implementation is still open to be programmed into the digital SRAM memory. Simulation of a wireline link is shortly evaluated in Section 2.9. However, this was limited to predictions of the behavior of a standard available IBIS-AMI transmitter model and algorithm. To accurately predict the optimal amount of FFE taps and pre-emphasis values, a far more accurate link measurement should be done including all interconnects, such as PCB traces and connectors. Then an IBIS-AMI model of the designed output driver should be extracted, so an actual link analysis simulation can be done. After this, the optimal channel compensation can be predicted. If the BER is insufficient, FEC techniques could be added. The optimal algorithms can then be programmed into the SRAM memory of the designed chip to be tested on their performance.

Measurements are still necessary to prove the working of the design. Simulations have been done to predict random jitter and deterministic jitter. However, random jitter simulations were done using ideal transistor models which is not representative of the actual frequency noise that might occur. The effect of clock skew could also not have been predicted since simulating the complete extracted layout was too labor-intensive for the server, so only parts have been simulated. Moreover, the output eye diagram is not completely representative. The post-layout extracted RC have been simulated including pad, but effects from the package or the inductance from the bond wire are still unknown. When the chip is fabricated, measurements should be done using a high-speed (real-time or sampling) oscilloscope with a real-time bandwidth of at least the Nyquist frequency of 25.6 GHz. The SRAM memory should be programmed with a PRBS sequence, possibly including pre-emphasis algorithms. Then the results should be analyzed to determine if the requirements can be met. The measurements should include:

- Output swing
- SNDR
- Horizontal eye opening
- Vertical eye opening
- BER curves

- Deterministic jitter
- Random jitter
- RLM
- Power
- Temperature

Output Termination could be improved in future designs by making the termination resistors adjustable. Due to design time limitations, only two $50\,\Omega$ unsilicided poly resistors have been used as termination. However, the accuracy of this resistor could be far off, definitely due to the large operating temperature range. This could degrade the SNDR performance, both due to the output power not being delivered efficiently, decreasing total swing, and reflections not being terminated, increasing distortion. Making the termination resistors adjustable, they could be optimized during measurements to match the $100\,\Omega$ differential load, thereby maximizing the transmitter performance.

Acronyms

BER Bit Error Rate. 5, 11–13, 20, 24, 46, 68

CDR Clock and Data Recovery. 1, 10, 11

CML Current Mode Logic driver. 28, 31, 49, 52, 55–57, 65, 67

CTLE Continuous time linear equalizer. 23

DAC Digital-to-Analog Converter. 3, 20, 29, 55, 65, 67

DFE Decision feedback equalizer. 23

DFF Dynamic Flip-Flop. 31, 39, 49–51, 67

DMUX Demultiplexer. 1

DTD data-transition Density. 46

FEC Forward Error Correction. 5, 13, 24, 68

FFE Feed Forward Equalisation. 20, 27–31, 59, 64, 67, 68

IBIS-AMI I/O Buffer Information Specification - Algorithmic Modeling Interface. 25, 68

IIR infinite impulse response. 23

ISI Inter-symbol Interference. 5, 7, 23

LSB Least Significant Bit. 28, 59

MFED maximally flat envelope delay response. 21

MSB Most Significant Bit. 28

MUX Multiplexer. 1–3, 21, 28, 30, 31, 33, 35, 39–47, 49–54, 67

NRZ non-return-to-zero. 5, 7, 10, 11, 13, 14, 25

PAM4 4-level pulse-amplitude-modulation. 1, 2, 5, 7, 10–14, 22, 24, 25, 27, 28, 56, 57, 64

PDF probability density function. 9, 11–13

PLL Phase-Locked Loop. 1

PSD Power Spectral Density. 5–7

RLM Ratio of Level Mismatch. 5, 14, 15, 57, 64, 67, 69

RX Receiver. 1

SNDR Signal-to-Noise-and-Distortion-Ratio. 25, 68, 69

SNR Signal-to-Noise Ratio. 10, 20, 58, 59

SST Source Series Terminated driver. 28, 29, 55–57

TSPC True Single-Phase Clocking. 51, 67

TX Transmitter. 1, 2, 20

UI Unit Interval. 8, 10, 13, 20

A

Floor plan

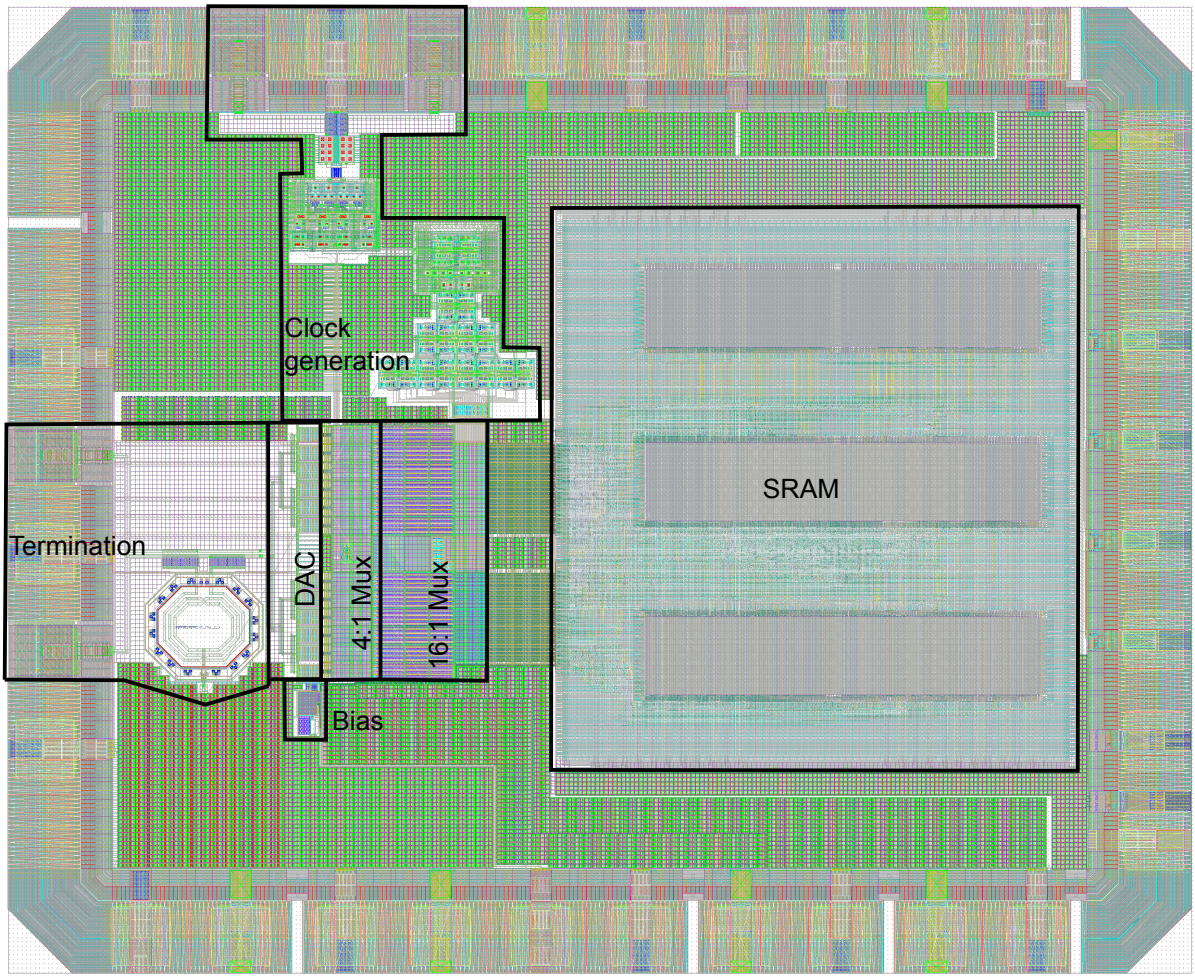


Figure A.1: Layout floor plan of complete chip.

Bibliography

[1] B. Razavi, “Historical Trends in Wireline Communications,” *Imid 2009*, pp. 1069–1072, 2009, ISSN: 19430582. DOI: 10.1109/MSSC.2015.2477016. [Online]. Available: papers://b601f53a-84e1-4f0c-abd7-620bc101dbfb/Paper/p7135.

[2] J. Kim, S. Kundu, A. Balankutty, M. Beach, B. C. Kim, S. Kim, Y. Liu, S. K. Murthy, P. Wali, K. Yu, H. S. Kim, C.-c. Liu, D. Shin, A. Cohen, Y. Fan, and F. O’Mahony, “8.1 A 224Gb/s DAC-Based PAM-4 Transmitter with 8-Tap FFE in 10nm CMOS,” in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, Feb. 2021, ISBN: 978-1-7281-9549-0. DOI: 10.1109/ISSCC42613.2021.9365840.

[3] B. Patra, R. M. Incandela, J. P. Van Dijk, H. A. Homulle, L. Song, M. Shahmohammadi, R. B. Staszewski, A. Vladimirescu, M. Babaie, F. Sebastiano, and E. Charbon, “Cryo-CMOS Circuits and Systems for Quantum Computing Applications,” *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 309–321, Jan. 2018. DOI: 10.1109/JSSC.2017.2737549.

[4] S. Borkar, “Design challenges of technology scaling,” *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul. 1999. DOI: 10.1109/40.782564.

[5] E. Groen, C. Boecker, M. Hossain, R. Vu, S. Vamvakos, H. Lin, S. Li, M. Van Ierssel, P. Choudhary, N. Wang, M. Shibata, M. H. Taghavi, N. Nguyen, and S. Desai, “A 10-to-112Gb/s DSP-DAC-Based Transmitter with 1.2Vppd Output Swing in 7nm FinFET,” in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 2020-February, Institute of Electrical and Electronics Engineers Inc., Feb. 2020, pp. 120–122, ISBN: 9781728132044. DOI: 10.1109/ISSCC19947.2020.9063130.

[6] E. Charbon, F. Sebastiano, A. Vladimirescu, H. Homulle, S. Visser, L. Song, and R. M. Incandela, “Cryo-CMOS for quantum computing,” *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 1–13, Jan. 2017. DOI: 10.1109/IEDM.2016.7838410.

[7] R. M. Incandela, L. Song, H. Homulle, E. Charbon, A. Vladimirescu, and F. Sebastiano, “Characterization and Compact Modeling of Nanometer CMOS Transistors at Deep-Cryogenic Temperatures,” *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 996–1006, Mar. 2018. DOI: 10.1109/JEDS.2018.2821763.

[8] P. A. T’Hart, J. P. Van Dijk, M. Babaie, E. Charbon, A. Vladimircscu, and F. Sebastiano, “Characterization and model validation of mismatch in nanometer CMOS at cryogenic temperatures,” *European Solid-State Device Research Conference*, vol. 2018-September, pp. 246–249, Oct. 2018. DOI: 10.1109/ESSDERC.2018.8486859.

[9] T. Duobinary, N. R. Z. Data, J. Lee, M.-s. Chen, and H.-d. Wang, “Design and Comparison of Three 20-Gb / s Backplane,” *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 43, no. 9, pp. 2120–2133, 2008.

[10] J. H. Sinsky, A. Adamiecki, and M. Duelk, “10-Gb/s electrical backplane transmission using duobinary signaling,” in *IEEE MTT-S International Microwave Symposium Digest*, vol. 1, 2004, pp. 109–112. DOI: 10.1109/mwsym.2004.1335814.

[11] J. Van Kerrebrouck, T. De Keulenaer, R. Pierco, J. De Geest, J. H. Sinsky, B. Kozicki, X. Yin, G. Torfs, and J. Bauwelinck, “NRZ, Duobinary, or PAM4?: Choosing Among High-Speed Electrical Interconnects,” *IEEE Microwave Magazine*, vol. 20, no. 7, Jul. 2019, ISSN: 1527-3342. DOI: 10.1109/MMM.2019.2909517.

[12] J. L. Zerbe, C. W. Werner, V. Stojanovic, F. Chen, J. Wei, G. Tsang, D. Kim, W. F. Stonecypher, A. Ho, T. P. Thrush, R. T. Kollipara, M. A. Horowitz, and K. S. Donnelly, “Equalization and Clock Recovery for a 2.5-10-Gb/s 2-PAM/4-PAM Backplane Transceiver Cell,” *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, vol. 38, no. 12, 2003. DOI: 10.1109/JSSC.2003.818572.

[13] Y. Chang, “Low-Power Wireline Transmitter Design,” UCLA, Tech. Rep., 2018, p. 63. [Online]. Available: https://escholarship.org/uc/item/60b6812k%0Ahttps://escholarship.org/uc/item/7842d9b4%0Ahttps://escholarship.org/uc/item/9jv565v7.

[14] “Implementation Agreement OIF-CEI-04.0 Common Electrical I/O (CEI) Optical Internetworking Forum-Clause 0 : Document Structure and Contents IA Title: Common Electrical I/O (CEI)-Electrical and Jitter Interoperability agreements for 6G+ bps, 11G+ bps, 25G+ bps I/O and 56G+ bps IA # OIF-CEI-04.0,” 2017. [Online]. Available: www.oiforum.com.

[15] H. Johnson and M. Graham, *High-speed digital design: a handbook of black magic*, 1993. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=793170%5Cnhttp://www.lavoisier.fr/livre/notice.asp?id=OR2W26AAL3KOWI.

[16] P. Duthil, “Material Properties at Low Temperature,” DOI: 10.5170/CERN-2014-005.77. [Online]. Available: http://dx.doi.org/10.5170/CERN-2014-005.77.

[17] C. Menolfi, M. Braendli, P. A. Francese, T. Morf, A. Cevrero, M. Kossel, L. Kull, D. Luu, I. Ozkaya, and T. Toifl, “A 112Gb/S 2.6pJ/b 8-Tap FFE PAM-4 SST TX in 14nm CMOS,” in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 61, Institute of Electrical and Electronics Engineers Inc., Mar. 2018, pp. 104–106, ISBN: 9781509049394. DOI: 10.1109/ISSCC.2018.8310205.

[18] J. S. Walling, S. Shekhar, and D. J. Allstot, “Wideband CMOS amplifier design: Time-domain considerations,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 7, pp. 1781–1793, 2008, ISSN: 10577122. DOI: 10.1109/TCSI.2008.926977.

[19] S. Gondi and B. Razavi, “Equalization and Clock and Data Recovery Techniques for 10-Gb/s CMOS Serial-Link Receivers,” *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, vol. 42, no. 9, 2007. DOI: 10.1109/JSSC.2007.903076.

[20] H. Wu, M. Shimanouchi, and M. Pengli, “Effective Link Equalizations for Serial Links at 112 Gbps and beyond,” in *EPEPS 2018 - IEEE 27th Conference on Electrical Performance of Electronic Packaging and Systems*, Institute of Electrical and Electronics Engineers Inc., Nov. 2018, pp. 25–27, ISBN: 9781538693032. DOI: 10.1109/EPEPS.2018.8534219.

[21] G. C. Clark and J. B. Cain, *Error-Correction Coding for Digital Communications*. Springer US, 1981. DOI: 10.1007/978-1-4899-2174-1.

[22] P.-J. Peng, Y.-T. Chen, S.-T. Lai, C.-H. Chen, H.-E. Huang, and T. Shih, “6.7 A 112Gb/s PAM-4 Voltage-Mode Transmitter with 4-Tap Two-Step FFE and Automatic Phase Alignment Techniques in 40nm CMOS,” in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, vol. 2019-Febru, IEEE, Feb. 2019, pp. 124–126, ISBN: 978-1-5386-8531-0. DOI: 10.1109/ISSCC.2019.8662361. [Online]. Available: https://ieeexplore.ieee.org/document/8662361/.

[23] J. Kim, A. Balankutty, R. K. Dokania, A. Elshazly, H. S. Kim, S. Kundu, D. Shi, S. Weaver, K. Yu, and F. O’Mahony, “A 112 Gb/s PAM-4 56 Gb/s NRZ Reconfigurable Transmitter With Three-Tap FFE in 10-nm FinFET,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 29–42, Jan. 2019, ISSN: 0018-9200. DOI: 10.1109/JSSC.2018.2874040. [Online]. Available: https://ieeexplore.ieee.org/document/8500752/.

[24] X. Zheng, H. Ding, H. Ding, F. Zhao, D. Wu, L. Zhou, J. Wu, F. Lv, J. Wang, and X. Liu, “A 50-112-Gb/s PAM-4 transmitter with a fractional-spaced FFE in 65-nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1864–1876, Jul. 2020, ISSN: 1558173X. DOI: 10.1109/JSSC.2020.2987712.

[25] Z. Toprak-Deniz, J. E. Proesel, J. F. Bulzacchelli, H. A. Ainspan, T. O. Dickson, M. P. Beakes, and M. Meghelli, “A 128-Gb/s 1.3-pJ/b PAM-4 Transmitter With Reconfigurable 3-Tap FFE in 14-nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 19–26, Jan. 2020, ISSN: 0018-9200. DOI: 10.1109/JSSC.2019.2939081. [Online]. Available: https://ieeexplore.ieee.org/document/8848421/.

[26] B. Razavi, *Design of CMOS Phase-Locked Loops*. Cambridge University Press, Jan. 2020, ISBN: 9781108626200. DOI: 10.1017/9781108626200.

[27] J. Gong, F. Sebastiano, E. Charbon, and M. Babaie, “A 10-to-12 GHz 5 mW Charge-Sampling PLL Achieving 50 fsec RMS Jitter,-258.9 dB FOM and-65 dBc Reference Spur,” *Digest of Papers - IEEE Radio Frequency Integrated Circuits Symposium*, vol. 2020-August, pp. 15–18, Aug. 2020. DOI: 10.1109/RFIC49505.2020.9218380.

[28] J. Gong, E. Charbon, F. Sebastiano, and M. Babaie, “A 2.7mW 45fsrms-Jitter Cryogenic Dynamic-Amplifier-Based PLL for Quantum Computing Applications,” *Proceedings of the Custom Integrated Circuits Conference*, vol. 2021-April, Apr. 2021. DOI: 10.1109/CICC51472.2021.9431541.

[29] P. J. Peng, J. F. Li, L. Y. Chen, and J. Lee, “A 56Gb/s PAM-4/NRZ transceiver in 40nm CMOS,” *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 60, no. 28, pp. 110–111, 2017, ISSN: 01936530. DOI: 10.1109/ISSCC.2017.7870285.

[30] J. Yuan and C. Svensson, “High-Speed CMOS Circuit Technique,” *IEEE Journal of Solid-State Circuits*, vol. 24, no. 1, pp. 62–70, 1989. DOI: 10.1109/4.16303.

[31] L. Kong, Y. Chang, and B. Razavi, “An Inductorless 20-Gb/s CDR with High Jitter Tolerance,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 10, pp. 2857–2866, Oct. 2019. DOI: 10.1109/JSSC.2019.2930899.

[32] Y. Frans, S. McLeod, H. Hedayati, M. Elzeftawi, J. Namkoong, W. Lin, J. Im, P. Upadhyaya, and K. Chang, “A 40-to-64Gb/s NRZ transmitter with supply-regulated front-end in 16nm FinFET,” *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 59, pp. 68–70, Feb. 2016. DOI: 10.1109/ISSCC.2016.7417910.

[33] P. Heydari and R. Mohanavelu, “Design of ultrahigh-speed low-voltage CMOS CML buffers and latches,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 10, pp. 1081–1093, Oct. 2004. DOI: 10.1109/TVLSI.2004.833663.

[34] C. H. Lin, F. M. Van Der Goes, J. R. Westra, J. Mulder, Y. Lin, E. Arslan, E. Ayranci, X. Liu, and K. Bult, “A 12 bit 2.9 GS/s DAC with IM3 < 60 dBc beyond 1 GHz in 65 nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 12, pp. 3285–3293, Dec. 2009. DOI: 10.1109/JSSC.2009.2032624.

[35] A. Van den Bosch, M. A. Borremans, M. S. Steyaert, and W. Sansen, “A 10-bit 1-GSample/s Nyquist current-steering CMOS D/A converter,” *IEEE Journal of Solid-State Circuits*, vol. 36, no. 3, pp. 315–324, Mar. 2001. DOI: 10.1109/4.910469.

[36] M. Pelgrom, “Digital-to-Analog Conversion,” in *Analog-to-Digital Conversion*, Cham: Springer International Publishing, 2017. DOI: 10.1007/978-3-319-44971-5{_}7.