Master's Thesis

Design of High-Speed PAM-4 Transmitter for Memory Interface

메모리 인터페이스를 위한 고속 PAM-4 송신기

by

Hoyeon Cho

August, 2023

Department of Electrical and Computer Engineering
College of Engineering
Seoul National University

Design of High-Speed PAM-4 Transmitter for Memory Interface

지도 교수 정 덕 균

이 논문을 공학석사 학위논문으로 제출함
2023 년 8 월

서울대학교 대학원
전기·정보공학부
조 호 연

조호연의 석사 학위논문을 인준함
2023 년 8 월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Design of High-Speed PAM-4 Transmitter for Memory Interface

by

Hoyeon Cho

A Thesis Submitted to the Department of
Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for the Degree of
Master of Science
at
SEOUL NATIONAL UNIVERSITY

August, 2023

Committee in Charge:

Professor Suhwan Kim, Chairman

Professor Deog-Kyoon Jeong, Vice-Chairman

Professor Woo-seok Choi

Abstract

This thesis presents a circuit design for a high-speed PAM-4 transmitter for memory interfaces. The design of the transmitter requires the development of techniques, the ability to transmit high-quality data at high bandwidths and the minimization of power consumption.

First, a circuit that minimizes the power consumption of the serializer and the 4:1 multiplexer are studied. Symbol interference is compensated using feed-forward equalization. The quad-error corrector is designed using only CMOS to achieve error correction for clock phases divided into four phases while occupying minimal area. By maintaining 50Ω impedance without resistors connected in series to the driver, area and power are optimized.

The proposed high-speed PAM-4 transmitter circuit is fabricated using a 40-nm CMOS process and occupied an area of 0.00021mm². During the operation of transmitting 32Gb/s PAM-4 data, it consumes 30.4mW of power and achieves a power efficiency of approximately 0.95pJ/b. Additionally, an RLM value of 0.9 or higher is achieved, ensuring uniform output impedance in the high-frequency range and maintaining high-quality signals.

Keywords : Memory Interface, PAM-4, 4:1 MUX, Serializer, Driver

Student Number : 2021-22528

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In order for a computer's CPU to perform computational tasks, it needs to retrieve data from data storage. One important factor in data storage is obviously data capacity. However, the speed of processing HDD data designed with a focus on storage capacity is too slow, so there is a need for intermediate memory. In 1966, Dr. Robert H. Dennard proposed the structure of DRAM that can store data using only one transistor and one capacitor[1]. The proposed structure showed power efficiency and high integration density due to the advantage of using a small number of transistors, and it has the characteristic of RAM that can be accessed and read and written at the same speed from any location. More complex commands are possible because the bus signals use a synchronized clock. Double-data DRAM using both edges of the clock signal has also been proposed and is widely used in various forms.

The growth of industries requiring large-scale data processing, such as AI intelligence, requires DRAM that can transmit more data quickly. However, there are several factors causing bottlenecks. To address the first factor, the limit of clock frequency, clocking schemes have been proposed. The method of placing data on each phase of the differing clocks and later converting it to 1 bit through MUX is widely used. Additionally, methods such as PAM-4[2], PAM-3[3], and duobinary[4] have been proposed to increase data rate without changing the Nyquist frequency of data. However, transmitters that transmit data using these methods are highly affected by Inter-Symbol Interference, so attention should be paid to the eye diagram. In this thesis, methods for achieving high data rates while creating good eye diagrams are studied in the memory interfaces.

1.2 Thesis Organization

This thesis is organized as follows. In Chapter 2 provides the background of memory interfaces and explains the operation and motivation of each building block, including their advantages and disadvantages.

Chapter 3 describes the proposed PAM-4 transmitter for memory interfaces and discusses the considerations involved in circuit implementation. It includes a detailed explanation of the overall structure, as well as the implementation and simulation results of the serializer, 4:1 mux, feed-forward equalizer, and driver.

Chapter 4 presents the measurement results, along with the chip photomicrograph and measurement setup. The results are analyzed mainly from the perspectives of eye diagram and power consumption.

Chapter 5 summarizes the proposed work and concludes this thesis.

Chapter 2

Backgrounds

2.1 Overview

In recent times, several methods have been proposed to achieve high data rates[5], with one such method being the adoption of PAM-4 signaling, which transmits 2-bit data at once. Traditional data channels carry 1-bit data, represented as 0 or 1. However, PAM-4 signaling sends 2-bit data through one channel, with the most significant bit (MSB) having twice the voltage weight of the least significant bit (LSB), resulting in a full swing voltage range divided into four quartiles that represent values like 00, 01, 10, and 11. For instance, with a voltage swing range between 0V and 1.2V, the data values 01 and 10 are transmitted as 0.4V and 0.8V, respectively.

The primary advantage of PAM-4 signaling is the ability to achieve fast data transmission rates without increasing the clock frequency beyond the Nyquist frequency. There are clear limitations to generating high-frequency clocks with low clock jitter using circuits such as

PLLs that incorporate loops and feedback. Additionally, when the Nyquist frequency is low, the operation frequency and complexity of the serializer circuit decrease, leading to reduced power consumption. Finally, there are certain ranges where PAM-4 signaling outperforms NRZ in terms of SNR, which can be mathematically calculated. The intrinsic SNR loss of PAM-4 signaling is determined by the vertical eye opening for PAM-4 signals, which is one-third of the range[6].

$$\mathbf{20} \cdot \boldsymbol{log}\left(\frac{\mathbf{1}}{\mathbf{3}}\right) \simeq \mathbf{-9.54}\ (\boldsymbol{dB}) \tag{2.1}$$

However, PAM-4 signaling is affected more significantly by nonlinearity characteristics in data transmission. Therefore, when SNR degradation exceeds 10 dB as the Nyquist frequency is doubled in NRZ signaling, adopting PAM-4 signaling is advantageous from the SNR aspect. On the other hand, the disadvantage of PAM-4 is that it requires greater effort to obtain an eye diagram due to the reduced voltage swing at high frequencies caused by ISI, which is divided into three equal parts. For instance, the circuit becomes more complex when using equalizers to improve ISI, and the number of equalizers increases twice to control MSB and LSB separately. At the same time, clock distribution layout becomes difficult to eliminate clock skew.

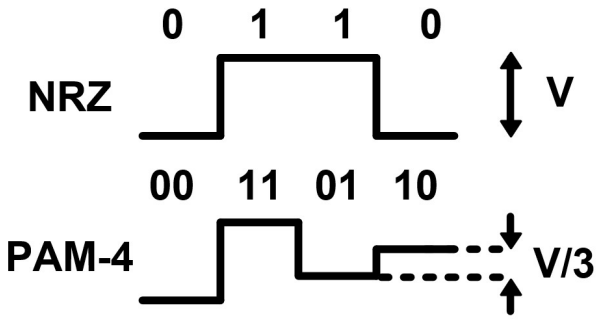
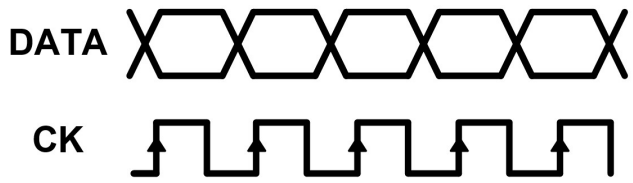
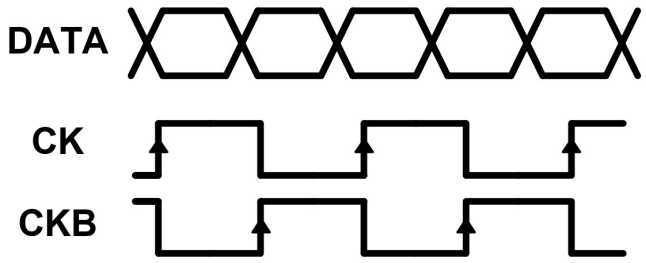


Fig. 2.1 Binary (NRZ) and pulse-amplitude modulation 4-level (PAM-4).

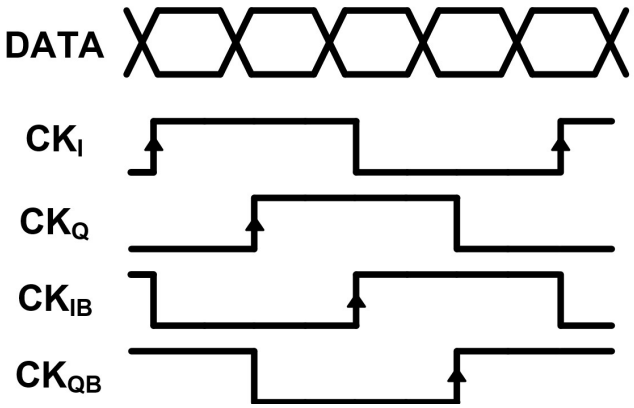
With the limitations of clock frequency, three clocking schemes have been proposed. The basic full-rate clock synchronizes all data to the rising edge of the clock and uses only one phase clock, making it easy to use without any other constraints. However, the symbol frequency and clock frequency are the same, providing little assistance in increasing the data rate. On the other hand, the fast clock, which is designed to increase the data rate, is difficult to generate using circuits like PLL and has the disadvantage of challenging clock distribution. In contrast, the quarter-rate clock can transmit data at a symbol rate that is four times the clock frequency[7]. Compared to using a full-rate clock with twice the frequency, it has the advantage of easier clocking distribution and lower power consumption. However, the disadvantage is that additional circuits are needed to maintain the exact 90-degree difference between the clocks of each phase.



(a)



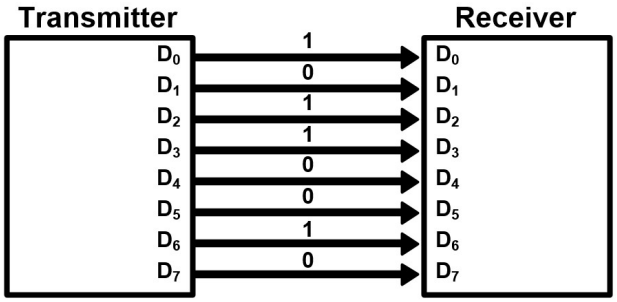
(b)



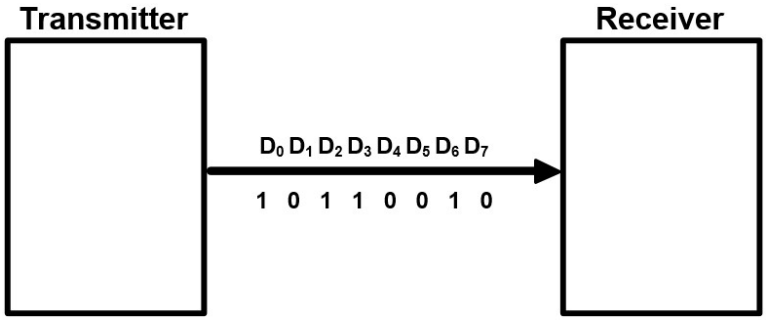
(c)

Fig. 2.2 Data and timing diagram of (a) full-rate, (b) half-rate, and (c) quarter-rate.

In chip-to-chip communication, there are two commonly used methods: parallel communication and serial communication. Each method has its own advantages and disadvantages, so they are used in different domains. Parallel communication has the advantage of requiring less power and having lower latency up until the driver because there is no need to serialize multiple data into a 1-bit channel. However, there are drawbacks such as crosstalk and clock skew that can affect the quality of data due to multiple channels running in parallel[8]. Additionally, there is an area problem that arises from having to lay multiple channels. Therefore, serial communication is widely used in communication methods such as HDMI and Ethernet. In this thesis, a serial link transmitter using PAM-4 with a quarter-rate clock has been implemented.



(a)



(b)

Fig. 2.3 Transceiver with (a) parallel communication and (b) serial communication.

2.2 Building Blocks

2.2.1 Serializer

A serializer is a circuit that converts parallel data at a low speed into a single bit data at a high speed. Typically, serializers are constructed by stacking multiple 2:1 serializers. For example, an 8:1 serializer is made by connecting three stages of 2:1 serializers, requiring a total of seven 2:1 serializers[9]. Since data from 2 bits must be combined into 1 bit data at each stage, a clock that is twice as fast as the previous stage is required in the next stage. Therefore, the 8:1 serializer is represented as shown in the following figure.

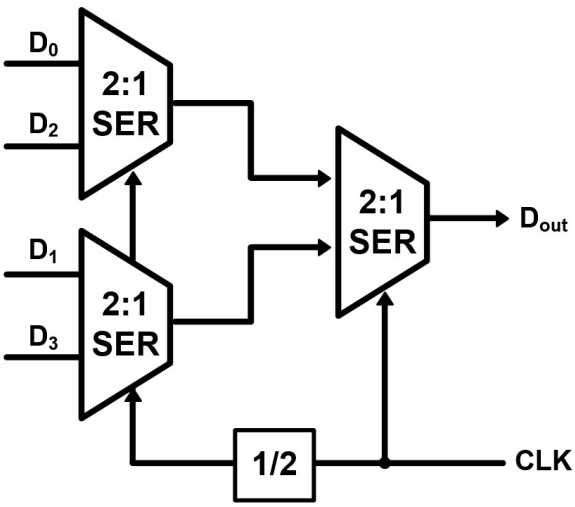
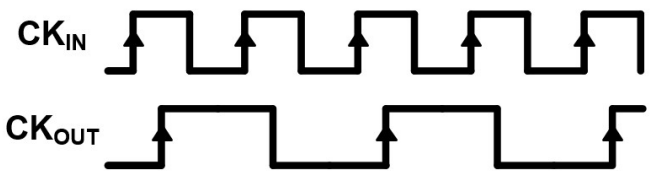
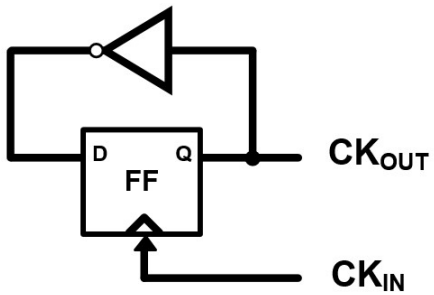


Fig. 2.4 4:1 Serializer architecture.

The 2:1 serializer is usually implemented using 2:1 MUXs and latches. The clock is used as the control signal for the 2:1 mux, while the latched data is received as the input signal. To meet time constraints such as setup time, five latches are used, including one flip-flop and a latch that uses an inverted clock. First, the Flip-Flop synchronizes data D0 and D1 with the clock. Then, when the clock rises, the previously waited data D1 is sent as the output. Similarly, when the clock falls, the previously waited data D0 is sent as the output. This provides sufficient timing margin as input changes before the clock's half period. However, one 2:1 serializer requires five latches, which results in high power consumption and area usage.



(a)



(b)

Fig. 2.5 Frequency divider (a) timing diagram and (b) architecture.

As mentioned earlier, the clock used in the serializer needs to have half the frequency of the clock used in the following stage. To achieve this, a frequency divider circuit consisting of Flip-Flops and inverters is created based on the clock used in the final serializer. The frequency of the clock is doubled and input into the Flip-Flop's clock. The output signal of the Flip-Flop is then inverted and fed back into the input. As a result, a clock with a frequency that transitions every other rising edge of the original clock is generated.

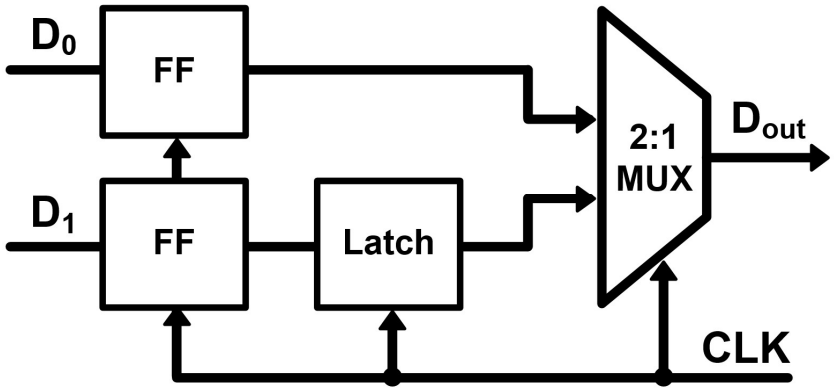


Fig. 2.6 2:1 serializer architecture.

2.2.2 4:1 MUX

If the data rate reaches a high frequency, the serializer presented earlier may not work, and a different circuit is required. Therefore, we studied the process of creating full-rate data by applying quadrature clocks to 4-bit data to create 1-UI pulses and combining them into 1 bit through a 4:1 MUX.

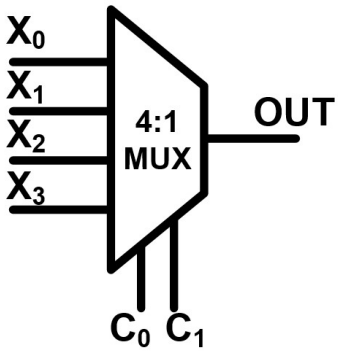


Fig. 2.7 4:1 MUX architecture.

C0	C1	M
0	0	X0
0	1	X1
1	0	X2
1	1	X3

Table 2.1 4:1 MUX truth table

The simplest structure to multiplex the synthesized data of four phases of clock using a 4:1 MUX, as shown in the figure below, is by using an AND gate[10]. However, a CML structure is adopted because this structure needs to operate at high clock frequencies, even though it

induces a relatively large amount of power consumption. There are three main methods to combine two phases of clock with input data and transmit the signal to the MUX. One method is to use 1-stack MUX, which uses a 1-UI pulse generator to combine two clocks and input data and transmit the signal to the MUX. Another method is to use 2-stack MUX, which combines one clock with input data and transmits the other clock to the MUX. The other is to use 3-stack MUX, which transmits all clocks and input data separately to the MUX. The first 1-stack MUX has the advantage of a small stack, resulting in a large swing and a clean eye diagram. However, it has the disadvantage of requiring a complex pulse generator, which consumes a lot of power. Conversely, the 3-stack MUX has several stacks, resulting in a small swing and a poor eye diagram. However, it does not consume much power because it does not require a separate pulse generator. The 2-stack MUX has the intermediate characteristics of both 1-stack MUX and 3-stack MUX.

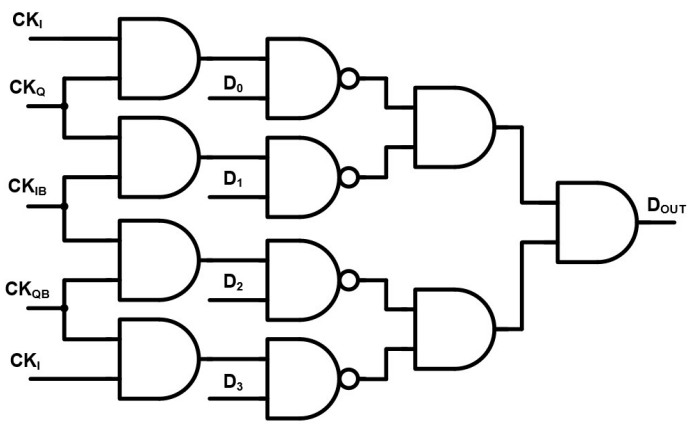


Fig. 2.8 Schematic of 4:1 MUX.

In general, a transmission gate or CMOS gate is used to generate a 1-UI pulse, as shown in the figure below. Although this approach has the advantage of easy design, it can exhibit unstable operation at high frequencies due to gate delay. Therefore, a current trend is to study the essential characteristics of the pulse output and remove some MOSFETs from the CMOS gate-based circuit. This approach aims to minimize parasitic capacitances attached to the data transmission path and is widely used to increase data rate.

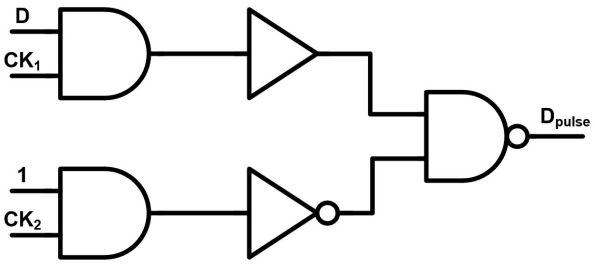


Fig. 2.9 Schematic of 1-UI pulse generator.

2.2.3 Feed-Forward Equalizer

The output at high data rates often exhibits waveform distortions influenced by both preceding and subsequent data, leading to improper waveform shapes. Additionally, as the required data bandwidth increases in memory interfaces, the need for data equalization extends beyond the receiver to include the transmitter. Specifically, the receiver lacks information about the upcoming data, making it unable to remove pre-cursors caused by the subsequent data. On the other hand, the transmitter, being aware of the data it intends to transmit, can effectively eliminate post-cursors and pre-cursors observed in the single-bit response by applying suitable weights. However, designs often prioritize the removal of pre-cursors when designing the equalizer, as compared to the receiver equalizer. The discrete-time equalizer is commonly employed to compensate for such inter-symbol interference (ISI). An example of the operation of a discrete-time equalizer is depicted in the following figure.

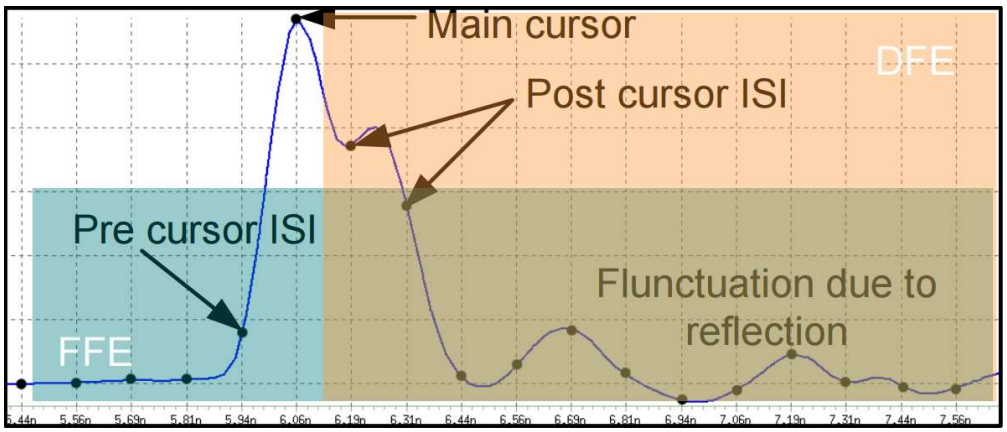


Fig. 2.10 Single bit response (SBR) and inter-symbol interference (ISI).

In essence, multiple instances of the main driver circuit with the same structure are designed, and the input of each driver is multiplied by an appropriate weight and combined to obtain the resulting output. The widely used approach to adjust the weights is by controlling the MOSFET sizing, which allows for weight adjustments. The cursors observed in the single bit response (SBR) typically exhibit the same sign as the main cursor. However, unlike the input to the main driver, the data entering the PMOS and NMOS inputs have opposite signs. This equalization technique, which reduces the signal amplitude, is referred to as de-emphasis feed-forward equalization (FFE). Compared to other equalizers, it offers more flexibility in terms of timing constraints and does not require a time-critical feedback path, thereby ensuring stability. However, since it operates by reducing the signal swing, it leads to a decrease in signal amplitude and higher power consumption, which are considered as disadvantages.

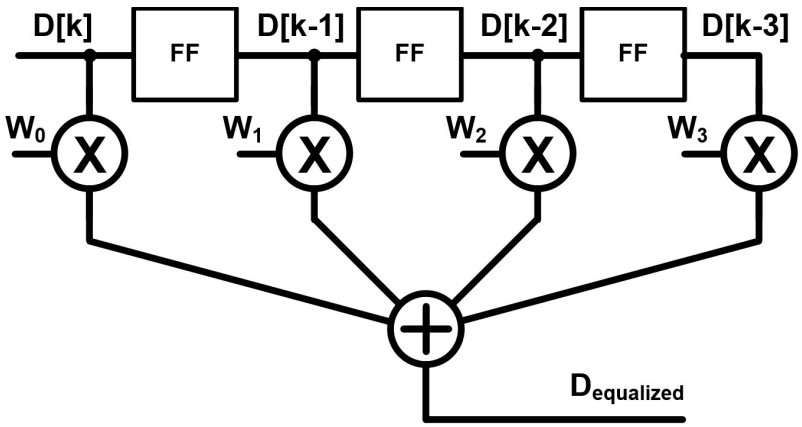


Fig. 2.11 Schematic of feed-forward equalizer (FFE).

2.2.4 Driver

To transmit data through a transmission line, it is necessary to send the data out via the driver in the transmitter. The driver is located in the final stage of the transmitter and is directly connected to the data output circuit. It is a circuit that crucially impacts the eye diagram and requires meticulous design and tuning. When transmitting high-frequency data through a transmission line with an impedance of Z_o, impedance matching becomes necessary. The design of the driver can be broadly classified into two categories, current mode driver and voltage mode driver.

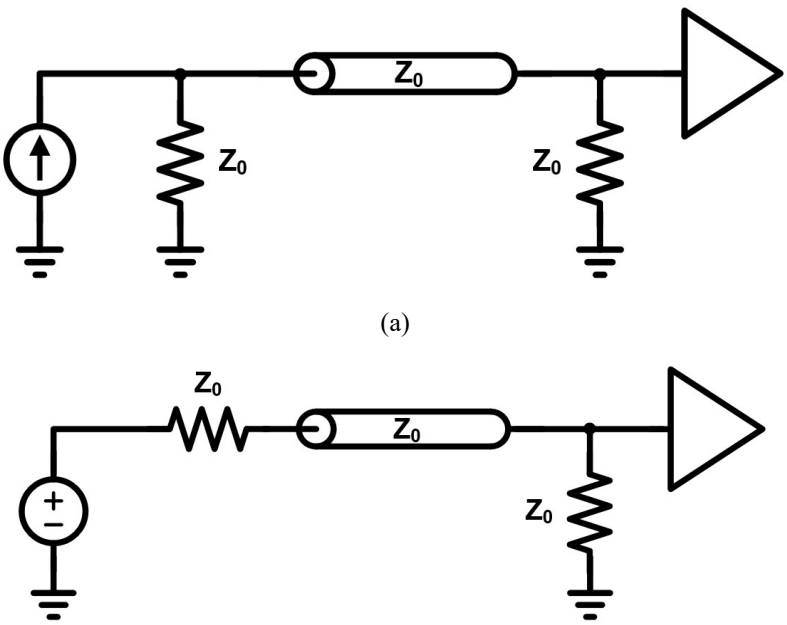


Fig. 2.12 Schematic of (a) current mode driver and (b) voltage mode driver.

The current mode driver, as the first type, is a driver that connects a current source to the transmission line to transmit output data. The output impedance of the current source is characterized as infinite and is connected in parallel with the load impedance Z_o. Although the actual current source in practical implementations may not have an infinite impedance, designing it to be significantly larger than the typical values of the transmission line impedance is sufficient to consider it as a current source. Additionally, the implemented current source operates in the saturation region with a relatively constant current, so the output impedance remains relatively unaffected. The use of a passive resistor as the load impedance ensures that impedance does not vary significantly due to PVT variations or output voltage. However, since the current source is always active, it consumes significant power. Moreover, the use of current mirrors or differential outputs, which are common in current mode drivers, can result in a larger circuit area requirement.

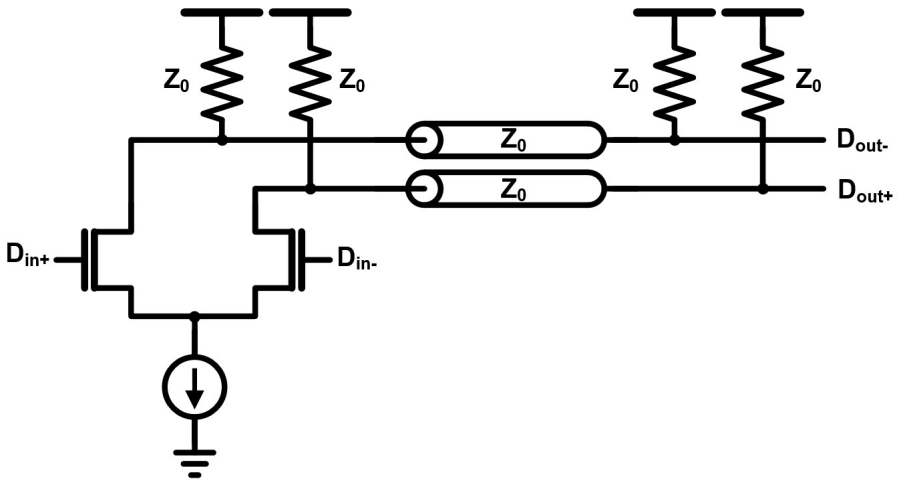


Fig. 2.13 Schematic of current mode logic driver.

The second type, voltage mode driver, transmits output data by connecting a voltage source to the transmission line. Unlike the current source, the output impedance of the voltage mode driver is close to zero, and it is connected in series with the load impedance Z_o. A non-ideal voltage source has a non-zero output impedance, and a commonly used P-over-N driver can be designed as shown in the figure. The resistor and source are connected in series to reduce the variation range of the transistor's transconductance (g_m) with respect to the output voltage. This configuration is known as source-series termination (SST). When assuming that the PMOS and NMOS transistors of the driver have a turn-on resistance (R_{ON}) in the linear region, the series resistance (R_S) and R_{ON} must be designed to sum up to Z_o. However, despite this design, additional circuitry is needed for impedance calibration due to variations in R_{ON}. Additionally, the output swing is reduced to half of VDD, which is a disadvantage. Furthermore, the load resistance being connected in series with the source makes the driver susceptible to supply noise. However, the reduced output swing leads to a clear advantage in terms of power consumption.

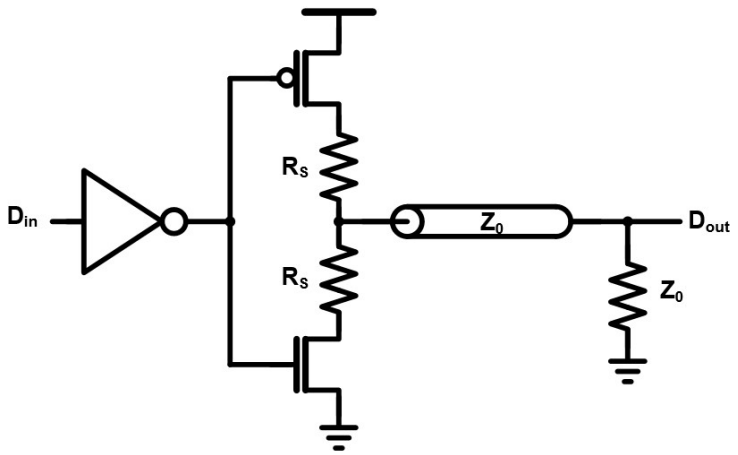


Fig. 2.14 Schematic of source-series terminated driver.

On the other hand, there are two commonly used methods for impedance calibration. The first method involves adjusting the number of driver slices for calibration. This method takes advantage of the inverse relationship between the number of driver slices and the output impedance. By increasing the number of driver slices, the output impedance decreases. The impedance of the transmission line is generally predictable, allowing the calibration range to be predetermined through simulations. This approach offers a wide range of calibration and stability advantages. However, it is not an ideal method for fine control of impedance, as it leads to increased parasitic capacitance of the output node and requires a larger area. The second method involves digitally controlling the impedance. In this approach, the calibration range is solely determined by transistor sizing, limiting the ability to achieve a wide range of calibration. However, it provides the advantage of fine-tuning with significantly less area compared to the driver slice method. Therefore, it is often beneficial to combine both methods appropriately in the design to achieve optimal results.

Chapter 3

Design of PAM-4 Transmitter for Memory Interface

3.1 Design Consideration

The demand for high-frequency interfaces is increasing, emphasizing the importance of considering key factors to achieve high-quality output data with a high data rate and good eye diagram. Bandwidth considerations play a crucial role in circuit design. For instance, simple CMOS-based circuits like a 4:1 MUX face challenges when operating at high frequencies due to inherent gate delays caused by parasitic capacitances. As a result, widely-known circuit designs may not be readily applicable in such cases. Accurate impedance matching is another critical aspect for designing drivers that generate desired output. Inaccurate impedance

matching can lead to reflections and impedance mismatches at high frequencies, adversely affecting the quality of the output data. Additionally, high-frequency data not only impacts the main cursor aligned with the clock but also influences the data of the subsequent clock cycles. Therefore, designing equalizers that adjust the weights of post-cursors and main cursors to minimize the effects of Inter-Symbol Interference (ISI) is of utmost importance. Lastly, careful attention must be given to clock distribution in both the 4:1 MUX and the equalizers. Clocks at high frequencies have short periods, making even minor clock skew introduced during layout design potentially detrimental.

Considering the memory interface, there are specific considerations to be considered during circuit design. Memory interfaces involve the simultaneous transmission of data by thousands of transmitters, making power consumption a critical concern. High power consumption can lead to undesirable heat generation, which can negatively impact the overall memory system's performance and potentially require additional cooling hardware, leading to increased costs. Therefore, circuit design should focus on minimizing power consumption while meeting the desired specifications. In the case of DDR/GDDR DRAM, single-ended output is commonly adopted as the standard. Although the commonly used differential output provides advantages such as increased swing resulting in better eye diagrams and effective noise reduction, the single-ended output adopted due to limitations in memory fabrication processes lacks these benefits. Thus, greater attention needs to be paid to ensuring the quality of the output data in terms of the single-ended output's limitations.

3.2 Overall Architecture

The operation of the proposed transmitter circuit is initiated by the incoming 16GHz clock signals, CKP and CKN, as shown in Fig [3.1]. To describe the overall flow, let's start with the clock signals.

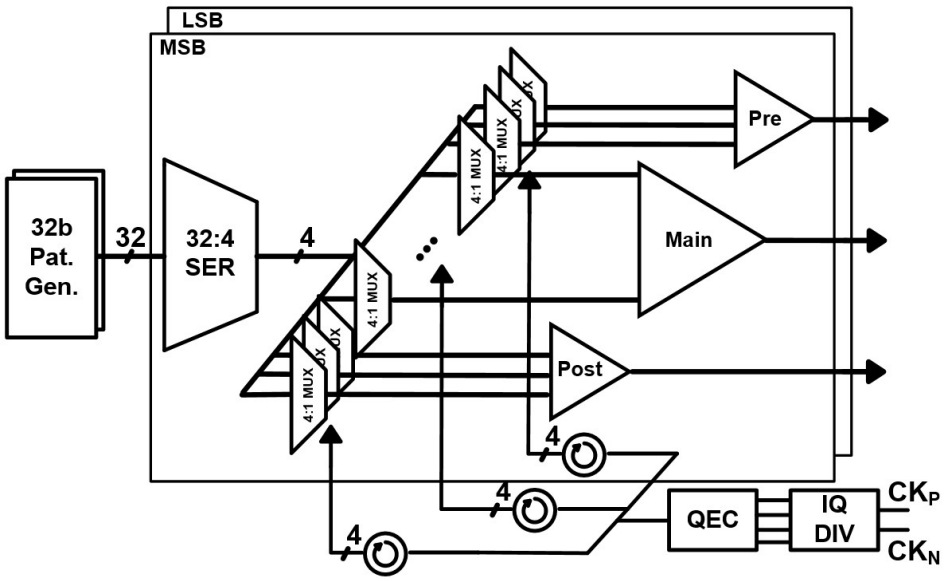


Fig. 3.1 Overall architecture of the proposed transmitter.

The buffered differential clock signals, CKP and CKN, are divided by the IQ divider. The IQ divider generates a 4-phase 8-GHz clock, ensuring a 90-degree phase difference through the quadrature error corrector. These four clock phases enter the 1-UI pulse generator within the 4:1 MUX, where they are used to modulate the data with clock information, creating pulses. Among the multiple clock signals, one clock selected by the I2C interface undergoes a

frequency reduction to half using a clock divider and operates the 32:4 serializer. As the data progresses through the serializer's stages, the frequency decreases by half, facilitating the serialization process. The clock passing through the clock divider after the last stage is used as a synchronized clock for two digitally synthesized 32-bit pattern generators. These generators produce 64-bit data, generated in prbs-7 format, with MSB and LSB each generating 32 bits. The two 32-bit data are serialized into 4 bits each using a 32:4 serializer. Each set of four 1-bit data, along with a 2-phase clock, generates a 1-UI pulse. These four 1-UI pulses are multiplexed into 1 bit using a 4:1 MUX and then passed to the feed-forward equalizer (FFE). The FFE, controlled by the I2C interface, processes the signals and passes them to the driver. The MSB and LSB, each representing 1 bit, are multiplied by different weights and transmitted from the driver.

3.3 Proposed Circuit Implementation

3.3.1 Serializer

The digital output, consisting of 32 bits each for MSB and LSB in PRBS-7 pattern, is combined into 4-bit serial data through a 32:4 serializer for each section. The schematic, divided into three stages with a clock divider and a 2:1 serializer, is shown in Figure [3.2]. The IQ divider enables the selection of the appropriate timing clock among the quadrature clocks composed of four phases. To serialize the data in the correct order, the inputs to the serializer are arranged according to a predefined sequence. The MSB 4 bits and LSB 4 bits produced by the serializer are then fed as inputs to the 1-UI pulse generator within the 4:1 MUX.

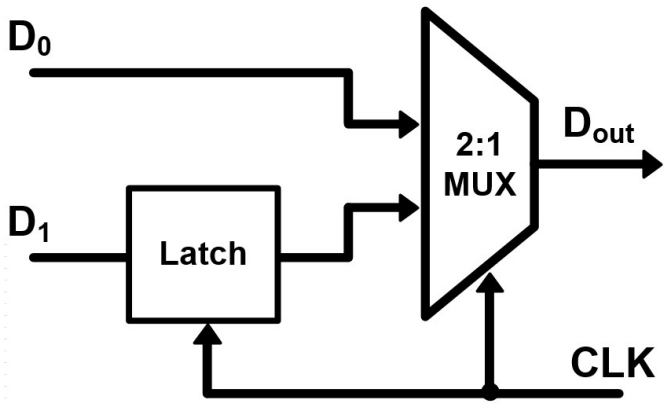


Fig. 3.2 Schematic of the proposed 2:1 serializer.

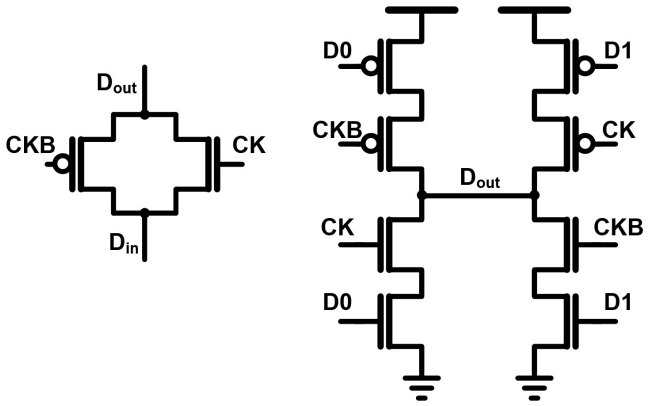


Fig. 3.3 Schematic of the latch and MUX.

To minimize area and power, the 2:1 serializer, which reduces the number of latches and eliminates the MUX, is implemented as shown in the following figure. Each 2:1 serializer consists of one pass gate and two latches, and the outputs of the two latches are combined to form the output of the 2:1 serializer. The transparency issue that could occur between the output of the previous 2:1 serializer and the input of the next stage is addressed by the characteristics of the clock divider.

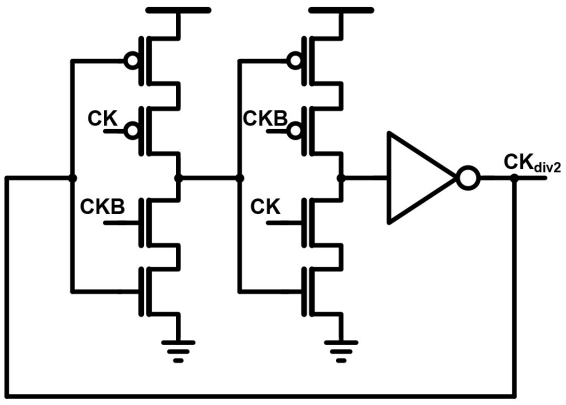


Fig. 3.4 Schematic of the clock divider.

The clock divider, as depicted in the figure, divides the frequency of the fast clock by 2 to generate the slower clock. The delay between the input and output clocks in the flip-flops helps mitigate the potential transparency issue between stages of the 2:1 serializer. In summary, by reducing the number of latches and incorporating a structure widely used with 2 latches instead of a MUX, we can expect a power reduction of around 70%, as the number of components decreases from 7 to 2.

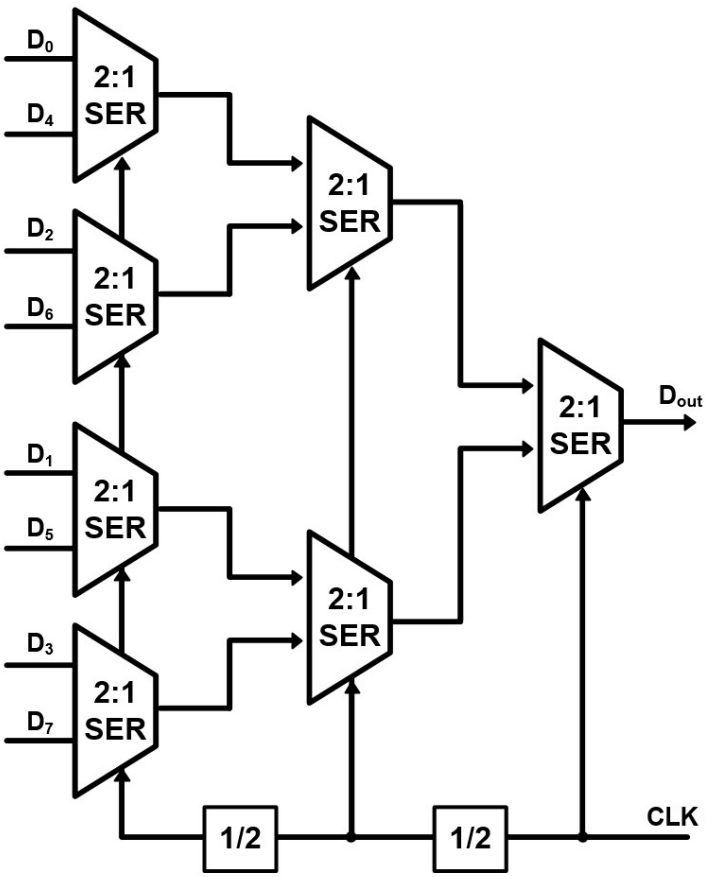


Fig. 3.5 Architecture of the 8:1 serializer.

3.3.2 4:1 MUX

When designing a 4:1 multiplexer (MUX) operating at high frequencies, a different structure is required compared to the serializer. For this purpose, a current mode logic (CML) structure is selected, which combines data with two clock pulses consisting of two phases. There are three possible structures to consider for combining data and the two clocks: 1-stack, 2-stack, and 3-stack MUX. The comparison of these structures should take into account power consumption, eye diagram quality, and time constraints.

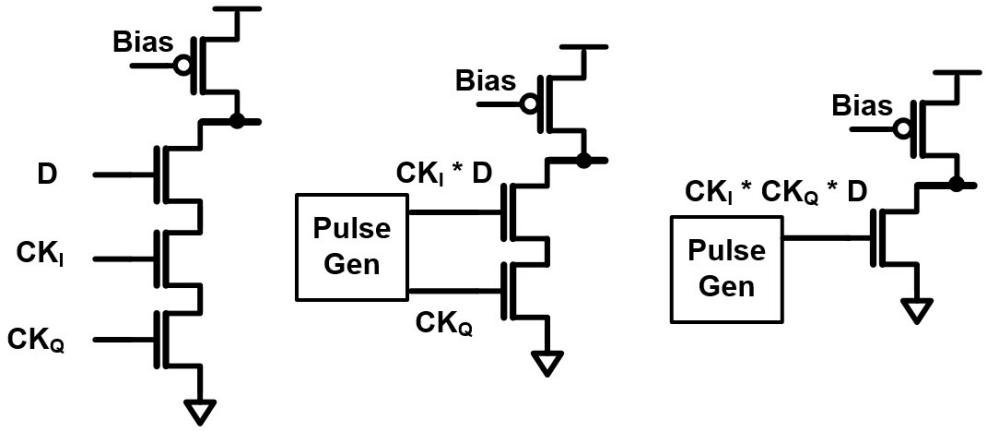
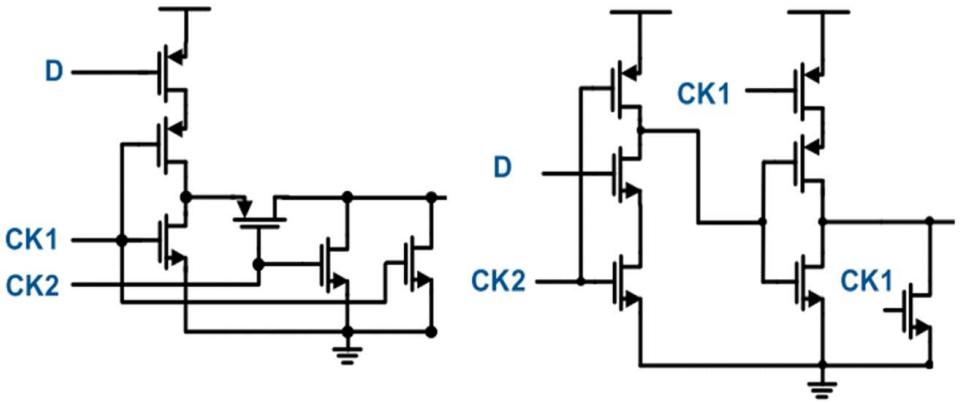


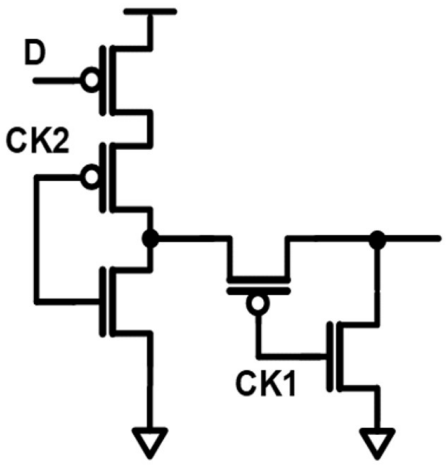
Fig. 3.6 Three possible architectures of the CML 4:1 MUX.

In the case of 1-stack and 2-stack MUX, a separate pulse generator needs to be placed before the MUX stage, resulting in higher power consumption compared to the 3-stack MUX. Specifically, the 1-stack MUX requires two clocks to be combined with the data, leading to larger power consumption in the pulse generator that generates the pulses in two stages. As for the eye diagram, structures with fewer stacks exhibit better voltage swing. In terms of time

constraints, the 3-stack MUX has more flexibility as it does not require a separate pulse generator.



(a)



(b)

Fig. 3.7 (a) Two preceding schematics and (b) presented schematic of the 1-UI pulse generator.[11-12]

Considering the focus on achieving a good quality eye diagram rather than power consumption, the 1-stack MUX structure is designed. To generate such pulses, a separate 1-UI pulse generator needs to be designed. Using CMOS gates to construct the 1-UI pulse generator increases power consumption and may not handle the voltage swing at high data frequencies. In previous pulse generator models, pulses were generated in two stages or by carrying the leading-phase clock first in the data and following with the lagging-phase clock. In this design, an approach that minimizes the use of MOSFETs in CMOS gate operations is employed. This approach is the most efficient in terms of power consumption and provides an advantage in terms of time constraints, as the lagging-phase clock is carried in the data first. The results of post-simulation are presented in the following figure.

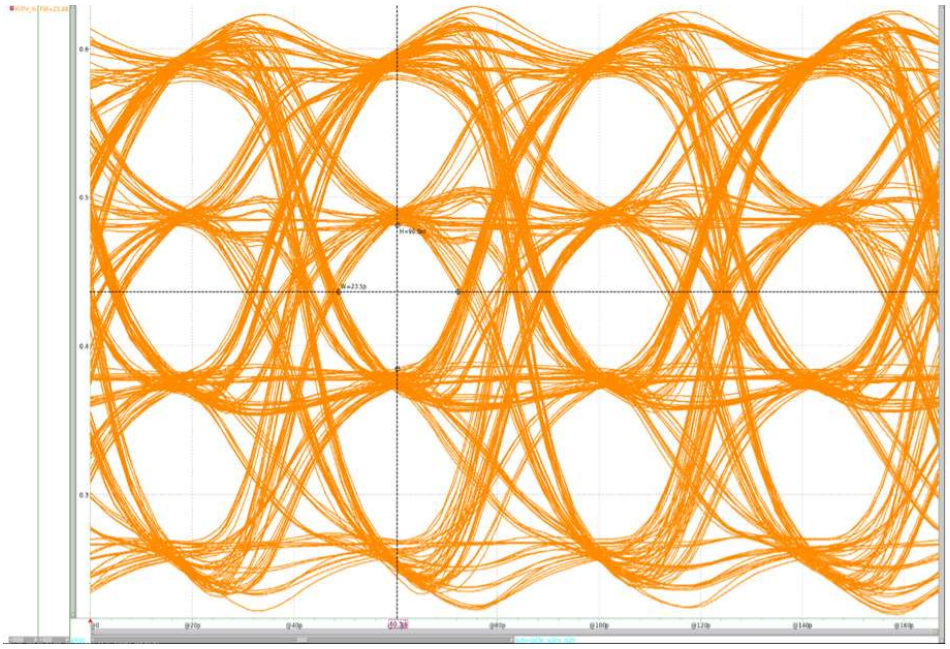


Fig. 3.8 Post simulation waveform of the 4:1 MUX.

3.3.3 Driver and Feed-Forward Equalizer

A 3-tap filter is employed, allowing the selection of inputs for each driver to control the pre-cursor, main-cursor, or post-cursor. Additionally, the number of drivers controlling each cursor is designed to be adjustable while keeping the total number of operating drivers fixed, facilitating easy control of weights. To reduce power consumption, only a few slides of the feed-forward equalizer (FFE) are utilized, and unused slides are gated by clock signals to save power while maintaining parasitic capacitance. This allowed doubling the weight of the MSB without using the corresponding slide.

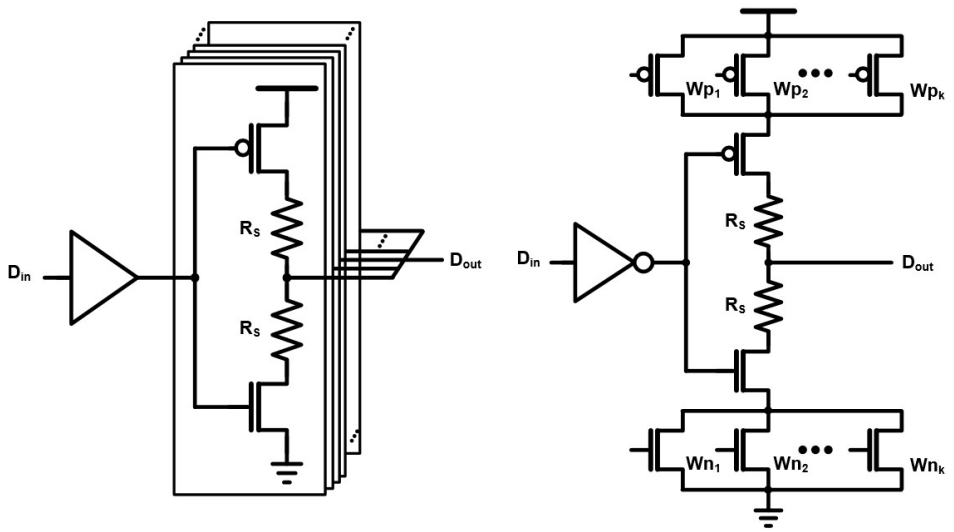


Fig. 3.9 Examples of the feed-forward equalizer(FFE).

Incorporating resistors as passive components in the driver design presents two challenges. Increasing the resistor size is typically necessary for linearity. However, as the resistor size

increases, the CMOS size in the driver and even the pre-driver must also be increased, resulting in increased power consumption. Moreover, since resistors are typically implemented using CMOS, they occupy significant area by themselves. When designing multiple slices for the FFE, the increase in driver area has a greater impact.

To mitigate the nonlinearity caused by the quadratic term in the current equation of CMOS, NP CMOS is employed as an auxiliary to improve the nonlinearity of the driver. Setting VDDQ to a value equal to VDD-Vth allows for the implementation of auxiliary CMOS that replaces the pull-up resistor.

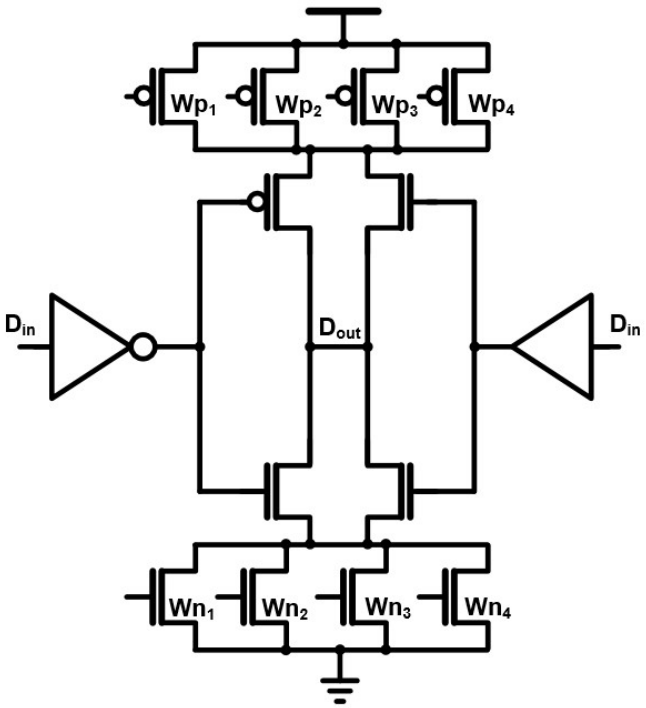


Fig. 3.10 Overall architecture of the proposed transmitter.

This structure offers two advantages. First, it effectively reduces power consumption by decreasing the CMOS size of the driver and pre-driver. Furthermore the DC impedance and AC impedance do not have similar values by using transistors instead of passive resistors. Therefore, an increase in output swing can be expected.

Chapter 4

Measurement Results

4.1 Die Photomicrograph

The prototype chip is fabricated using a 40nm CMOS process. The corresponding die photomicrograph is depicted in Fig. 4.1. The total area, including the pads, measures 1.1mm x 0.8mm, resulting in a total of 0.88mm2. The I2C block occupies an area of 0.024mm2, while the Digital block occupies 0.01mm2. By significantly reducing the number of latches used, the newly proposed 32:4 Serializer occupies a remarkably small area of 35um x 75um. From the MUX to the Driver, the components are arranged in an interconnected structure without separating them to minimize area.

The area of the components, including the 4:1 MUX employing a 1-UI pulse generator for power reduction and the FFE-implemented Driver through multiple slicers, measures 60um x 100um. The Quadrature Error Corrector (QEC), responsible for correcting clock errors in the

four-phase clock, occupies an area of 13um x 16um. The overall area of the transmitter, comprising all components, is a compact 110um x 100um, demonstrating the utilization of a very small area.

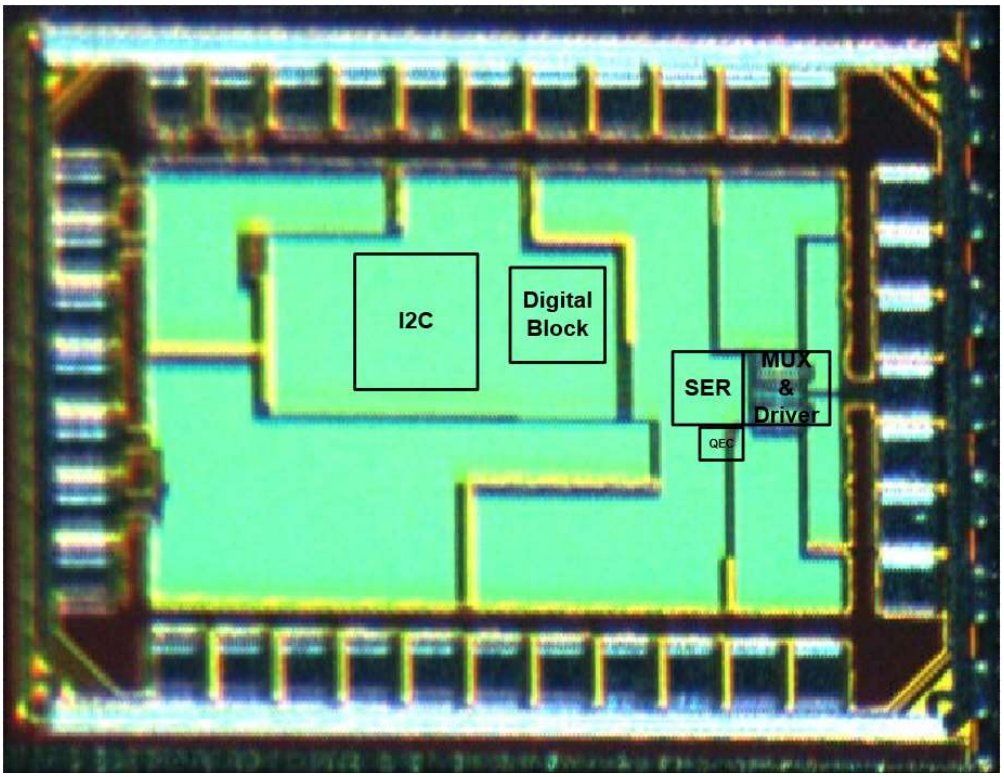


Fig. 4.1 Die Photomicrograph

4.2 Measurement Setup

The experimental setup for characterizing the prototype chip is illustrated in Fig. 4.2. To measure the power consumption of each module in the transmitter, three power supplies (Agilent E36313A, E3631A, E3649A) are employed. The voltage domains include the 4:1 MUX, 32:4 Serializer, Driver & FFE, Digital, and I2C, allowing for convenient power analysis using simultaneous measurements. A voltage of 1V is supplied to all modules, except for the 0.6V VDDQ in the driver and the 3.3V used for I2C's high voltage. A signal quality analyzer (MP1800A) is utilized to generate a 16GHz differential clock from an external source. Additionally, an oscilloscope (MSO 73304DX) generates a 4GHz auxiliary clock, which serves as a reference for the clock in the differential clock's quarter frequency. Communication with the phase quadrature error corrector (QEC), driver's output impedance control, and clock domain crossing between modules is facilitated through the I2C module, enabling interaction with a personal computer (PC). The device under test (DUT) generates a 32Gb/s PAM-4 output from the designed transmitter, which is then sent to the oscilloscope for measurement purposes.

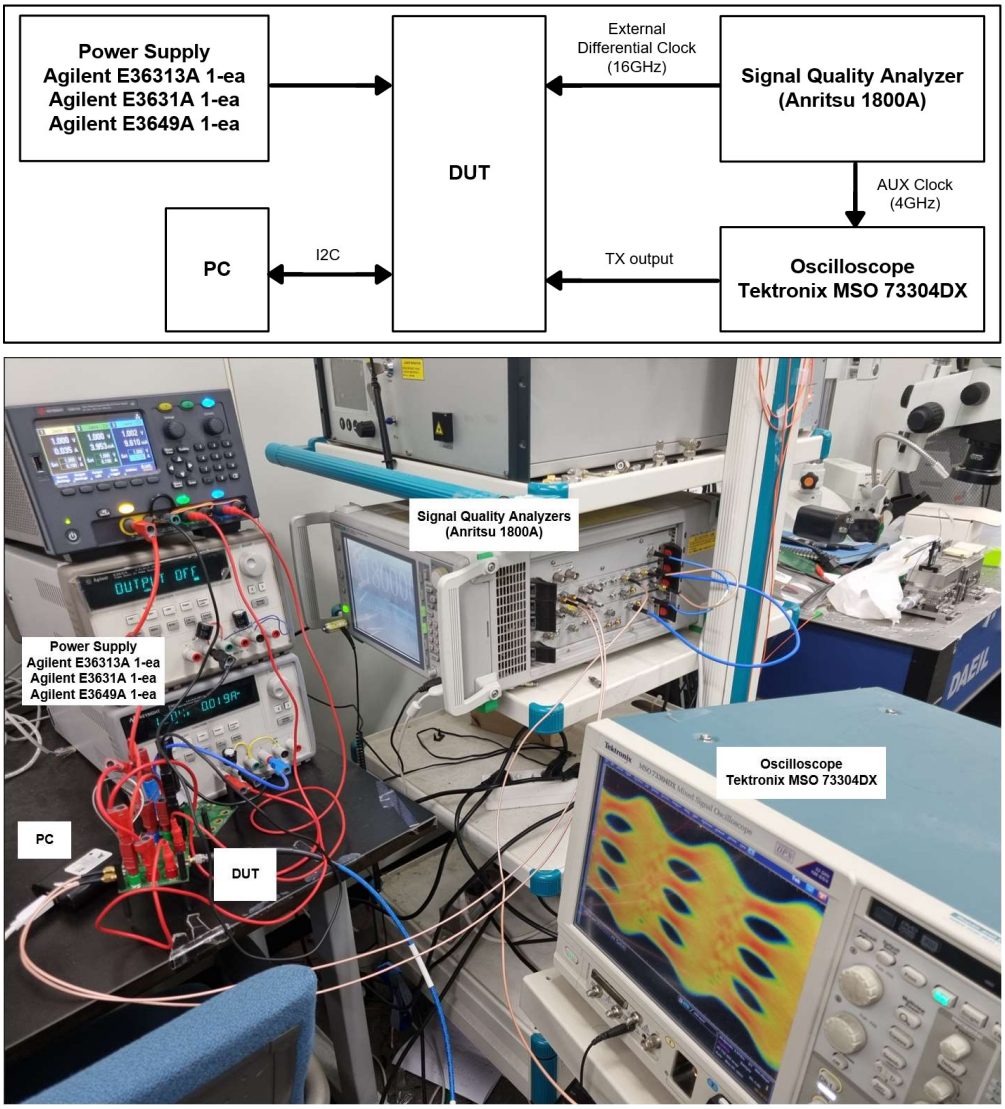
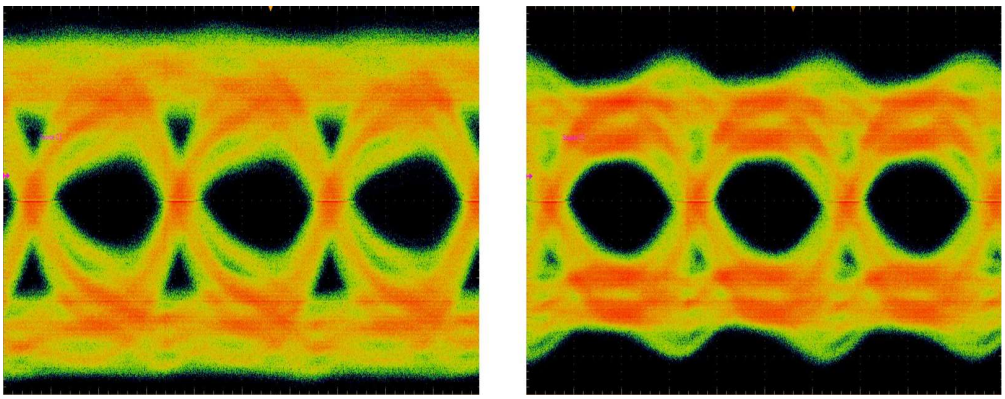


Fig. 4.2 Measurement environment setup.

4.3 Measurement Results

The output of the designed transmitter was measured using an oscilloscope. A termination resistance was 50Ω, and the input consisted of a PRBS-7 pattern, with a pair of 32-bit data assigned to the MSB and LSB, generated by the pattern generator in the digital block. Figure 4.3 illustrates the eye diagram of the 16Gb/s NRZ output. Without the use of an equalizer, a vertical eye opening of 110mV was achieved, but it exhibited some asymmetry due to the influence of inter-symbol interference (ISI). On the other hand, when an equalizer was employed, the eye diagram became perfectly symmetrical, and the vertical eye opening improved slightly to 125mV.



(a) (b)

Fig. 4.3 Eye diagrams measured at 16Gb/s (a) without FFE and (b) with FFE

Figure 4.4 illustrates the eye diagram of the 28Gb/s PAM-4 output. In the absence of an equalizer, the eye diagram exhibited vulnerability due to the influence of inter-symbol interference (ISI). However, when the equalizer was employed, the eye diagram was effectively preserved. The vertical eye openings were measured to be 61mV, 59mV, and 56mV, respectively. The Ratio Level Mismatch (RLM), a crucial indicator of transmitter linearity, was calculated to be approximately 0.95, indicating distinguished performance. Moreover, the horizontal timing margin exceeded 0.5UI, suggesting that the receiver would have no difficulty in accurately discerning the received data.

Moving on to below, it showcases the eye diagram of the 32Gb/s PAM-4 output. In the absence of an equalizer, the eye diagram was susceptible to ISI and did not exhibit secure characteristics. However, with the inclusion of an equalizer, the eye diagram achieved satisfactory stability. The vertical eye openings measured for each level were 36mV, 34mV, and 30mV, indicating acceptable performance at different signal levels. The RLM was calculated to be approximately 0.900, further affirming the positive vertical eye characteristics. The horizontal timing margin was around 0.4UI, slightly smaller than that observed in the 28Gb/s output, suggesting a slight reduction in timing margin but still within acceptable limits.

Fig. 4.4 Eye diagrams measured at 28Gb/s and 32Gb/s

4.4 Performance Summary

The proposed chip has a total active area of 0.0088mm^2, excluding the I2C and digital blocks. The power used for transmitting 32Gb/s PAM-4 data is 30.4mW, resulting in an energy efficiency of 0.95pJ/b. Table shows the supply voltage and power breakdown for each block. The 4:1 MUX consumes a large amount of power because it adopts the CML structure. Since the VDD used in the driver is shared with the 4:1 MUX block, the VDDQ that uses 0.6V has relatively lower power consumption than the 4:1 MUX. The exact values for the power breakdown are calculated from the results of post-layout simulations.

Table 4.2 Area of each block.

		AREA
I2C		200um X 120um
Digital Block		100um X 100um
Transmitter	**32:4 SER**	35um X 75um
	4:1 MUX & DRV	60um X 100um
	QEC	13um X 16um

Table 4.3 Supply voltage and power breakdown of each block.

	Supply Voltage	**Power breakdown**
32:4 SER & QEC	1 V	4.4 mW
4:1 MUX & PRE-DRV	1 V	20.2 mW
DRV with VDDQ	0.6 V	5.8 mW

Table summarizes the performance of the proposed transmitter compared to other single-ended transmitters for memory interfaces. To avoid clock skew issues, the layout is designed symmetrically from the 4:1 MUX to the driver, resulting in a very small transmitter area. The power consumption is reduced by using a 4:1 MUX that is designed targeting less power consumption, while the PN-NP structure of the driver ensured that the output impedance was uniform, improving the linearity and eye diagram of the PAM-4 output.

Table 4.4 Performance summary and comparison.

	ISSCC'20 [2]	JSSC'21 [13]	JSSC'22 [14]	ISSCC'22 [15]	This work
Technology	65nm CMOS	65nm CMOS	28nm CMOS	28nm LPP	40nm CMOS
Data rate [Gb/s]	32	28	21	20	32
Signaling	PAM-4	PAM-4	Duobinary	NRZ	PAM-4
Energy efficiency [pJ/b]	0.97	0.58	0.67	1.18	0.95
Area [mm^2]	0.009	0.033	0.0072	0.0012	0.00021

Chapter 5

Conclusion

In this research study, a high-speed transmitter for memory interfaces is proposed, which aims to achieve low power consumption even at high speeds. The proposed transmitter structure includes a 32:4 Serializer which reduces latch counts, and a 4:1 MUX resulting in significant reductions in power consumption and area. The advantages of the CML structure of the 4:1 MUX proposed in prior research are combined to reduce power consumption while maintaining data quality.

The proposed chip is fabricated using a 40nm CMOS process and has a total area of 0.88mm2, with a pure transmitter design area of 0.00021mm2, excluding I2C and digital blocks. The power consumption, excluding I2C and digital blocks, is 30.4mW, with a power efficiency of 0.95pJ/b. The 32Gb/s PAM-4 output, which is the target data rate, exhibits good eye opening at the RLM 0.900 level. To ensure clock skew is not an issue, the layout is symmetrically designed from the 4:1 MUX to the Driver, resulting in a small transmitter area. By incorporating a 4:1 MUX with low power consumption and using a PN-NP structure for

the Driver, the output impedance is made uniform, improving the linearity and eye diagram of the PAM-4 output. The performance of the proposed transmitter is summarized in Table, and it is compared with other single-ended transmitters for memory interfaces.

Bibliography

[1] Robert H. Dennard, “Field-effect transistor memory”, U.S. Patent No. 3,387,286A, June, 1967.

[2] P. -W. Chiu and C. Kim, "22.4 A 32Gb/s Digital-Intensive Single-Ended PAM-4 Transceiver for High-Speed Memory Interfaces Featuring a 2-Tap Time-Based Decision Feedback Equalizer and an In-Situ Channel-Loss Monitor," 2020 IEEE International Solid- State Circuits Conference - (ISSCC), 2020, pp. 336-337.

[3] H. Park, J. Song, Y. Lee, J. Sim, J. Choi and C. Kim, "23.3 A 3-bit/2UI 27Gb/s PAM-3 Single-Ended Transceiver Using One-Tap DFE for Next-Generation Memory Interface," 2019 IEEE International Solid- State Circuits Conference - (ISSCC), 2019, pp. 383-384.

[4] S. -M. Lee et al., "An 80 mV-Swing Single-Ended Duobinary Transceiver With a TIA RX Termination for the Point-to-Point DRAM Interface," in IEEE Journal of Solid-State Circuits, vol. 49, no. 11, pp. 2619-2630, Nov. 2014.

[5] Wulf, Wm A., and Sally A. McKee. "Hitting the memory wall: Implications of the obvious," ACM SIGARCH computer architecture news 23.1, pp. 20-24, 1995.

[6] Haram Ju. (2019). “A Study on Low-Power, High-Speed PAM-4 Transmitter with Current-Driven Feedback Driver.”, (Ph.D. dissertation). Seoul National University, Korea.

[7] X. Zheng et al., “A 40-Gb/s Quarter-Rate SerDes Transmitter and Receiver Chipset in 65-nm CMOS,” in IEEE Journal of Solid-state Circuits, vol. 52, no. 11, pp. 2963-2978, Nov. 2017.

[8] Wikipedia, serial communication. [Online] [Accessed on 2nd Nov. 2021] http://en.wikipedia.org/wiki/serial_communication.

[9] D. Gong et al., “Development of A 16:1 serializer for data transmission at 5 Gbps,” in Topical workshop on electronics in particle physics, pp. 21-25, Sep. 2009.

[10] J. Kim et al., "A 112Gb/s PAM-4 transmitter with 3-Tap FFE in 10nm CMOS," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), 2018, pp. 102-104.

[11] Z. Toprak-Deniz et al., "A 128-Gb/s 1.3-pJ/b PAM-4 Transmitter With Reconfigurable 3-Tap FFE in 14-nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 55, no. 1, pp. 19-26, Jan. 2020.

[12] J. Kim et al., "8.1 A 224Gb/s DAC-Based PAM-4 Transmitter with 8-Tap FFE in 10nm CMOS," 2021 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 2021, pp. 126-128.

[13] Y. -U. Jeong, H. Park, C. Hyun and S. Kim, "A 28-Gb/s/pin PAM-4 Single-Ended Transmitter with High-Linearity and Impedance-Matched Driver and 3-Point ZQ Calibration for Memory Interfaces," 2020 IEEE Symposium on VLSI Circuits, 2020, pp. 1-2.

[14] D. Kang et al., "A 21-Gb/s Duobinary Transceiver for GDDR Interfaces With an Adaptive Equalizer," in IEEE Journal of Solid-State Circuits, vol. 57, no. 10, pp. 3084-3093, Oct. 2022.

[15] C. Moon, J. Seo, M. Lee, I. Jang and B. Kim, "A 20 Gb/s/pin 1.18pJ/b 1149μm2Single-Ended Inverter-based 4-tap Addition-Only Feed-Forward Equalization Transmitter with Improved Robustness to Coefficient Errors in 28nm CMOS," 2022 IEEE International Solid-State Circuits Conference (ISSCC), 2022, pp. 450-452.

초 록

본 논문은 메모리 인터페이스를 위한 고속 송신기를 설계하기 위한 회로를 제안하였다. 송신기를 설계하는데 있어서 일반적으로 발생하는 두 가지 요구사항들을 개선하는 기술들이 요구된다. 높은 대역폭에서 고품질의 데이터를 송신할 수 있는 기술과 동시에 전력소비를 최소화하는 기술이 연구되었다.

먼저 직렬변환기와 4:1 멀티플렉서의 전력 소비를 최소화하는 회로가 연구되었다. 피드-포워드 보상기를 통해 심볼간 간섭을 보상하였다. 쿼드러쳐 오류 정정기는 CMOS 만을 사용해 적은 면적으로 설계하여 4 개의 위상으로 구분된 클럭들의 오류를 정정하였다. 드라이버에 직렬로 연결된 저항 없이도 50Ω 임피던스 값을 유지할 수 있게 하여 면적과 전력을 최소화하였다.

제안된 고속 4 펄스진폭변조(PAM-4) 송신기 회로는 40-nm CMOS 공정으로 제작되었다. 0.00021mm^2 의 면적을 차지하였다. 32Gb/s 데이터를 송신하는 동작에서 30.4mW 의 전력을 소모하였고 약 0.95pJ/b 의 전력효율을 달성하였다. 동시에 0.9 이상의 불일치비율(RLM) 값을 달성하여 고주파수 대역에서도 출력 임피던스를 균일하게 유지하여 고품질의 신호를 유지하였다.

주요어 : 메모리 인터페이스, 펄스진폭변조, 4:1 멀티플렉서, 직렬변환기, 드라이버

학 번 : 2021-22528