Article

A 0.17 pJ/bit 28 Gb/s/pin Single-Ended PAM-4 Transmitter for On-Chip Short-Reach Unterminated Channels

Soyeon Park and Jintae Kim *

Department of Electrical and Electronics Engineering, Konkuk University, Seoul 05029, Korea
* Correspondence: jintkim@konkuk.ac.kr

Abstract: This paper presents the design of a single-ended four-level pulse-amplitude modulation (PAM-4) transmitter for an on-chip short-reach unterminated channel. To achieve multi-output generation, a local voltage buffer consisting of a diode-connected device and a leaker transistor is introduced. By charge-sharing between a local reservoir capacitor and an unterminated channel, the proposed transmitter generates mid-level output voltages without using the DC current, thereby realizing multi-level signaling without significantly increasing the static current. A prototype chip was fabricated by 28 nm CMOS process, and the transmitter exhibits an energy efficiency of 0.17 pJ/bit at 28 Gb/s/pin, which is state-of-the-art energy efficiency as a multi-level transmitter having a data rate beyond 20 Gb/s.

Keywords: single-ended signaling; transmitter; four-level pulse amplitude modulation (PAM-4); unterminated on-chip channel; short-reach channel

1. Introduction

The advent of data-intensive applications such as artificial intelligence (AI) and cloud services requires huge amount of data processing between processors and memories. In such massively parallel memory–processor interfaces, a single-ended link is a preferred electrical interface because reducing the pin count is the most compelling system constraint. A few such examples are the dynamic random access memory (DRAM) interface and high bandwidth memory (HBM) interface. In addition, a short-reach chip-to-chip “chiplet” interface such as bunch-of-wires (BoW) also adopts a single-ended link as a physical layer.

In high-speed single-ended links, a primary design objective is to maximize the data transfer rate per pin while minimizing the energy cost to transfer the bit. Traditionally, increasing the data rate means reducing the bit time, or equivalently increasing the symbol frequency, in combination with mild channel equalization. However, recent developments [1,2] demonstrated the possibility of using multi-level signaling for single-ended links, which allows us to increase the data rate without reducing the bit time. For instance, the authors in [1] demonstrated 22 Gb/s/pin for a single-ended GDDR6X interface using four-level pulse-amplitude modulation (PAM-4) signaling scheme, where a source series termination (SST) voltage-mode structure is used as a transmitter. In [3,4], a pseudo-open-drain logic (PODL) transmitter generates a PAM-4 signal by adjusting the resistance value in the driver.

One challenge of using multi-level signaling is minimizing the energy cost. Specifically, generating mid-level outputs in both SST and PODL structures relies on resistive voltage dividers. Considering the impedance matching with channel impedance, this approach inevitably consumes a large amount of static current when generating mid-level outputs. Accordingly, the energy cost of a previously published single-ended PAM-4 transmitter is generally higher than a binary transmitter, e.g., approximately 1 pJ/bit for 12 Gb/s [5] and 3.1 pJ/bit for 18 Gb/s [3].

This paper presents a low-power 28 Gb/s/pin PAM-4 transmitter optimized for an on-chip short-reach unterminated channel, achieving an energy efficiency of 0.17 pJ/bit.
The proposed PAM-4 transmitter generates mid-level outputs using capacitive charge sharing rather than resistive division, leading to substantial power saving. While the use of a capacitive-coupled nonreturn-to-zero (NRZ) driver without termination resistance has been subject of a previous publication [6], such structures are not compatible with the generation of multi-level outputs. In contrast, our proposed transmitter structure overcomes such limitations and achieves both a low energy cost and multi-level generations for unterminated channels.

This paper is organized as follows. Section 2 describes the architecture of the transmitter. Section 3 presents the concept and transistor-level design of the proposed transmitter circuit. The measured performance is shown in Section 4. Section 5 concludes the paper with summary.

2. Transmitter Architecture

Figure 1a shows the block diagram of the proposed transmitter along with an embedded eye monitor for measuring the on-chip eye diagram. The transmitter consists of a 2^7 − 1 pseudo random binary sequence (PRBS) generator, a PAM-4 encoder, two PAM-4 drivers and a 2-to-1 analog multiplexer (MUX). The PRBS generator drives the PAM-4 encoder with 4-bit-wide random digital bits, producing a pair of 4-bit-wide bitstream D_E<3:0> and D_O<3:0>.

![Diagram of TX and Eye Monitor](image)

**Figure 1.** (a) Overall architecture of the proposed PAM-4 TX for a short-reach channel and an eye monitor; (b) PAM-4 encoder input–output table; and (c) the timing diagram of the TX and eye monitor.

The PAM-4 encoder, whose encoding table is shown in Figure 1b, is constructed to control the switches in the transmitter in such a way that four distinct levels are generated. The two bitstreams from the encoders are synchronized at both the rising and the falling
3. Circuit Implementation

Figure 2 shows a transistor-level circuit diagram of the PAM-4 driver that generates four-level outputs, i.e., \( V_{\text{DDQ}} \), \( V_{\text{L2}} \), \( V_{\text{L1}} \) and \( V_{\text{SS}} \). The highest and lowest levels are generated by turning on \( M_8 \) and \( M_6 \), respectively, which is essentially same as SST drivers. The key difference is generating mid-level outputs, \( V_{\text{L1}} \) and \( V_{\text{L2}} \). Unlike the SST driver that uses resistive voltage division [1], the proposed transmitter utilizes capacitors and diode-connected devices having two different flavors of threshold voltage to define the mid-level outputs.

![Figure 2. Circuit implementation of the proposed PAM-4 driver.](image-url)

More specifically, the mid-level voltage levels are defined by the local voltage buffer consisting of diode-connected devices \( (M_1 \) and \( M_2) \) and the leakier transistors \( (M_4 \) and \( M_6) \). The diode-connected \( M_1 \) and \( M_2 \) operate in a saturation region and therefore the gate-source voltage increases with threshold voltage and bias current. To generate two different mid-levels, we use super low-\( V_{\text{TH}} \) (SLVT) device for \( M_1 \) and high-\( V_{\text{TH}} \) (HVT) device for \( M_2 \) so that \( V_{\text{L1}} \) is higher than \( V_{\text{L2}} \). The leakier transistors provide a static current path to the diode-connected devices when the corresponding voltage level is not transmitted, and hence slightly degrade the overall power efficiency. However, they are required to finely adjust the \( V_{\text{L1}} \) and \( V_{\text{L2}} \) to the desired voltage levels. In our implementation in 28 nm CMOS, \( V_{\text{L1}} \) and \( V_{\text{L2}} \) are tuned at 720 mV and 330 mV, respectively, by choosing the DC current in the leakier as 90 uA.

The generated mid-levels are transmitted by charge sharing between the local reservoir capacitor \( C_{\text{big1}} \) or \( C_{\text{big2}} \) and the total capacitance of the unterminated channel. Note that in a short-reach unterminated interface whose trace length is less than 1 mm, it is common to model the channel as purely capacitive with channel capacitance ranging from 200 fF/mm to 500 fF/mm depending on the channel structure and process [7,8]. Therefore, the on-chip TX can be designed as a high-impedance capacitive driver.

Figure 3 illustrates the details of the operation of the PAM-4 driver. For the convenience of notation, we refer to the signal levels as +3, +2, +1 and +0 from the highest to the lowest, respectively. When transmitting at the +3 level, the PFET \( M_0 \) turns on to connect \( V_{\text{OUT}} \) to...
$V_{\text{DDQ}}$. Similarly, the NFET $M_8$ connects $V_{\text{OUT}}$ with ground when transmitting $+0$ level. For +2 or +1 levels, the pre-charged $C_{\text{big2}}$ or $C_{\text{big1}}$ is connected to the channel through $M_3$ or $M_4$ while the leaker transistors are disconnected. Note that $M_3$ is PFET because +2 level is close to $V_{\text{DDQ}}$, while $M_4$ is NFET given that +1 level is close to ground. Since the output is not truly “driven” when transmitting +2 or +1 level, the output of the PAM-4 encoder ensures that $C_{\text{big1}}$ and $C_{\text{big2}}$ are fully pre-charged when they are not being used. For instance, when transmitting at the +3, +1 or +0 level, the leaker transistor $M_3$ turns on so that $C_{\text{big2}}$ is quickly pre-charged to the desired voltage level. Similarly, the PAM-4 encoder ensures that $C_{\text{big1}}$ is pre-charged when the transmitter it not transmitting at the +1 level.

<table>
<thead>
<tr>
<th>+3 Level</th>
<th>+2 Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2 = 0$, $S_3 = 0$, $S_4 = 1$, $S_5 = 0$</td>
<td>$S_2 = 0$, $S_3 = 0$, $S_4 = 0$, $S_5 = 1$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>+1 Level</th>
<th>+0 Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2 = 0$, $S_3 = 1$, $S_4 = 1$, $S_5 = 1$</td>
<td>$S_2 = 1$, $S_3 = 0$, $S_4 = 1$, $S_5 = 1$</td>
</tr>
</tbody>
</table>

Figure 3. Operation of the PAM-4 driver.

Figure 4 shows a conceptual circuit diagram along with a simulated eye diagram of the driver where $C_{\text{ch}}$ is the channel capacitor. When mid-levels are transmitted, charge sharing occurs between the $C_{\text{big1}}$ or $C_{\text{big2}}$ and the channel to form the TX voltage. Specifically, the voltage formed by the charge sharing can be expressed as

$$V_{\text{L1,tx}} = \frac{C_{\text{big1}} \cdot V_{\text{L1}} + C_{\text{ch}} \cdot V_{\text{ch}}}{C_{\text{big1}} + C_{\text{ch}}} \quad \text{and} \quad V_{\text{L2,tx}} = \frac{C_{\text{big2}} \cdot V_{\text{L2}} + C_{\text{ch}} \cdot V_{\text{ch}}}{C_{\text{big2}} + C_{\text{ch}}}$$

(1)

Figure 4. Simplified circuit diagram of the PAM-4 driver and the simulated eye diagram.
Therefore, to keep the transmitted mid-level voltage as close as possible to the $V_{L2}$ or $V_{L1}$, Equation (1) indicates that $C_{big1}$ or $C_{big2}$ needs to be much greater than $C_{ch}$. Note that a subthreshold conduction of $M_1$ or $M_2$ may impact the $V_{L1,TX}$ or $V_{L2,TX}$ if long repeated +1 or +2 levels are transmitted because the subthreshold current slowly charges $C_{big1}$ or $C_{big2}$. In this design, up to approximately 40 repeated +1 or +2 levels can be sent without causing a noticeable voltage level change. In practical situations, most memory interface systems utilize some encoding schemes such as 8 b/10 b in peripheral component interconnect express (PCIe) or cyclic-redundancy-check (CRC) in a double data rate (DDR) system, and these extra encodings prevent the transmitter from sending a long and identical repeated bit pattern.

In this work, we target the total channel capacitance less than 100 fF, which corresponds to approximately 0.1 mm of the on-chip interconnect. Therefore, our design uses 9 pF of $C_{big1}$ and $C_{big2}$, respectively, which is approximately 190 times greater than $C_{ch}$. For area efficiency, we use a low-$V_{TH}$ NFET to build the MOS capacitor to implement $C_{big1}$ and $C_{big2}$ instead of the metal-finger capacitor because the linearity of the capacitor is not critical.

The micrograph of the prototype and the cross-section structure of a short-reach channel used in this work are shown in Figure 5. The channel structure, which is similar to that used in [7], uses metal 6 as a signal layer with metal 7 and metal 5 as the top and bottom ground shielding layers, respectively. The signal line is 0.6 µm-wide and the adjacent shielding wires are spaced 0.4 µm apart. Our extracted simulation reveals that this structure exhibits a channel capacitance of 0.47 fF/µm. In this work, the length of on-chip interconnect is 100 µm, which corresponds to total channel capacitance of roughly 47 fF.

![Figure 5. (a) Micrograph of the prototype and (b) channel cross-section.](image)

**4. Measurement Results**

Figure 6 shows the measurement setup and the test board for the prototype chip fabricated by the 28 nm CMOS process. The proposed TX and eye-monitor occupy 8670 µm² and 2620 µm², respectively. The external signal source provides 7 GHz differential clock to the transmitter chip. The transmitter output waveform is captured by the embedded on-chip eye monitor by sweeping DTC control bits and DC reference voltages using two on-chip comparators. From the obtained outputs of the two comparators, we determine whether the voltage output of the channel lies between the two provided references at a
specific time, which allows us to calculate the histogram of the voltage distribution and consequently construct an eye diagram.

![Figure 6](image1.png)

Figure 6. (a) Measurement setup and (b) the test board.

Figure 7 shows the obtained 28 Gb/s eye diagram by applying the described method as well as the vertical histogram of the signal at the end of the channel. Although our eye-measurement is limited by the timing resolution of the DTC as well as the accuracy of the externally provided references, the eye diagram shows that the worst-case horizontal eye opening is approximately 0.2 UI and the vertical eye opening is approximately 40 mV. Figure 7b shows the vertical histogram of the transmitter at specific DTC setting, which clearly shows four distinct levels.

![Figure 7](image2.png)

Figure 7. (a) Measured eye diagram from the on-chip eye monitor and (b) vertical histogram at the sampling clock phase.

The measured power breakdown is shown in Figure 8. With the V_{DDQ} of 1.1 V, the total power consumption of the chip is 6.47 mW and the transmitter excluding the PRBS generator consumes roughly 75% of the total power. Table 1 compares the performance of the prototype transmitter with recently published single-ended transmitters. The proposed transmission achieves both the best energy efficiency of 0.17 pJ/b and the highest data rate among all high-speed single-ended transmitters even though the length of the channel used in the measurement is relatively shorter.
**Total Power Consumption** = 6.4702mW @ 28Gb/s/pin

Figure 8. The power breakdown of the PAM-4 transmitter at 28 Gb/s/pin.

Table 1. Performance comparison with other recent transmitters.

<table>
<thead>
<tr>
<th>Low Speed On-Chip Transmitter</th>
<th>High Speed On-Chip Transmitter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td><strong>Technology</strong></td>
</tr>
<tr>
<td>90 nm CMOS</td>
<td>130 nm CMOS</td>
</tr>
<tr>
<td><strong>Signaling</strong></td>
<td><strong>Signaling</strong></td>
</tr>
<tr>
<td>NRZ</td>
<td>NRZ</td>
</tr>
<tr>
<td><strong>Line Type</strong></td>
<td><strong>Line Type</strong></td>
</tr>
<tr>
<td>On-chip metal</td>
<td>On-chip metal</td>
</tr>
<tr>
<td><strong>Supply Voltage (V)</strong></td>
<td><strong>Supply Voltage (V)</strong></td>
</tr>
<tr>
<td>1.2</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Data Rate (Gb/s)</strong></td>
<td><strong>Data Rate (Gb/s)</strong></td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td><strong>Energy Efficiency</strong></td>
<td><strong>Energy Efficiency</strong></td>
</tr>
<tr>
<td><strong>0.28 pl/b</strong></td>
<td><strong>0.06 pl/b</strong></td>
</tr>
<tr>
<td><strong>Channel Length</strong></td>
<td><strong>Channel Length</strong></td>
</tr>
<tr>
<td>10 mm</td>
<td>10 mm</td>
</tr>
<tr>
<td><strong>Area (mm²)</strong></td>
<td><strong>Area (mm²)</strong></td>
</tr>
<tr>
<td>N/A</td>
<td>0.0034</td>
</tr>
<tr>
<td><strong>0.008673</strong></td>
<td><strong>0.001773</strong></td>
</tr>
</tbody>
</table>

*Data embedded clock signaling ** Include RX.

5. Conclusions

We presented a high-speed single-ended PAM-4 TX for a short-reach channel. Implemented by 28 nm CMOS process, the prototype chip achieved a data rate of 28 Gb/s/pin for the 100 μm on-chip unterminated channel with a state-of-the-art energy efficiency of 0.17 pl/b when using 1.1 V supply voltage. A key to achieving a high energy efficiency is using the capacitor and leakier to generate mid-level outputs instead of voltage division. With demonstrated performance, we believe that the proposed structure can be a promising transmitter topology for next-generation massively parallel on-chip interconnect systems.

**Author Contributions:** S.P. and J.K. proposed the architecture. S.P. designed the circuit and performed all measurements. S.P. wrote the initial manuscript and J.K. supervised the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by Samsung Chip Interconnect Solutions and by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-01171, A Development of Intelligent PHY Interface for High-Speed PIM Data Transfer).

**Acknowledgments:** The authors thank IDEC, KAIST, for the CAD tool support.

**Conflicts of Interest:** The authors declare no conflict of interest.
References


