A Cryo-CMOS DAC-Based 40-Gb/s PAM4 Wireline Transmitter for Quantum Computing

Niels Fakkel, *Graduate Student Member, IEEE*, Mohsen Mortazavi, Ramon W. J. Overwater, Fabio Sebastiano, *Senior Member, IEEE*, and Masoud Babaie, *Senior Member, IEEE*

***Abstract*— Addressing the advancement toward large-scale quantum computers, this article presents the first four-level pulse amplitude modulation (PAM4) wireline transmitter (TX) operating at cryogenic temperatures (CTs). With quantum computers scaling up toward thousands of quantum bits (qubits), but having too limited fidelity for robust operation, continuous rounds of quantum error correction (QEC) are necessary. However, QEC requires a large amount of data to be transferred from a cryogenic controller at 4 K to a classical processor at room temperature (RT). To bridge the gap, a high-speed data link between the quantum processor at CT and the classical counterpart at RT is needed. The proposed PAM4 TX architecture integrates a low-power 64:4 serializer structure, a high-speed 4:1 current-mode logic (CML) multiplexer, and a linear 6-bit digital-to-analog converter (DAC). Considering the challenges and benefits of CMOS operating at CTs, the TX architecture and circuitry are designed to exploit the maximum speed, while maintaining sufficient linearity. The fabricated 40-nm CMOS chip achieves a data rate of 40-Gb/s (36-Gb/s), an energy efficiency of 2.46 pJ/b (2.47 pJ/b), and 97.8% (96.6%) ratio of level mismatch (RLM) at CT (RT). While demonstrating an energy efficiency comparable to prior-art TXs in more advanced CMOS nodes at RT, the broad operating temperature of the proposed TX enables the required high-speed wireline link for large-scale quantum computers.**

***Index Terms*— Cryo-CMOS, high-speed digital-to-analog converter (DAC), quantum computing, quantum error correction (QEC), serializer, wireline transmitter (TX).**

I. Introduction

QUANTUM computers have the potential to solve certain computing tasks significantly faster than classical computers. In the current stage of quantum computing, with small-scale processors comprising up to a 100 quantum bits (qubits), simple algorithms can already outperform a classical computer [1], [2]. In the next intermediate stage, with processors achieving up to 10k qubits, specific chemistry or physics algorithms, such as computational catalysis, can be solved within weeks [3]. In a far future with $>10^6$ qubits, large-scale quantum computer systems can practically revolutionize computing, cracking modern-day encryption within a day [4].

Looking at the intermediate-scale stage for the coming years, qubits are hard to scale up and too noisy for robust computations [5]. Hence, different quantum error correction (QEC) strategies [6], [7], [8], such as surface code (SC), were proposed to realize a large-scale quantum computer with error rates low enough for useful calculations. In general, QEC encodes information across multiple *physical* qubits to form a *logical* qubit, thereby suppressing the errors. On the other hand, to reach good inherent physical qubit fidelity, thermal noise should also be minimized, and consequently, the qubits must be located at sufficiently low temperatures, typically at the mK stage of a dilution refrigerator.

As the classical controller and readout are currently located at room temperature (RT), a large number of cables are needed to transfer the extensive amount of *analog* information from/to the qubits. Fitting all these cables in a dilution refrigerator not only leads to an integration challenge, but also imposes a 1-mW/cable thermal load on the fridge [9]. To tackle this imminent bottleneck, CMOS circuits operating at cryogenic temperatures (cryo-CMOS) have been developed to move the control/readout equipment from RT to the cryogenic environment [10]. Efforts to date in characterization have resulted in multiple CMOS technologies being functional at cryogenic temperatures (CTs), accompanied by some challenges, including increased threshold voltage and device mismatch [11], [12], [13].

So far, cryo-CMOS circuits have been placed at the 4-K stage to control and read out the qubits [14], [15], [16], [17], [18], [19], while (de)multiplexers are designed for operating at the mK stage to reduce the amount of interconnect between qubit and controller [20], [21]. Moreover, *hot* qubits operating with high gate fidelities at temperatures above 1 K are being developed to completely close the temperature gap between the electronics and qubits, thus improving the scalability of future quantum computers [22], [23]. However, in any scenario, QEC is still needed to protect the quantum information from errors

Manuscript received 29 August 2023; revised 27 November 2023 and 19 January 2024; accepted 5 February 2024. Date of publication 21 February 2024; date of current version 25 April 2024. This work was supported in part by Intel Corporation and in part by the Netherlands Organization for Scientific Research under Project 17303. This article was approved by Associate Editor Kenichi Okada. *(Corresponding author: Niels Fakkel.)*

Niels Fakkel, Ramon W. J. Overwater, and Fabio Sebastiano are with the Department of Quantum and Computer Engineering, Delft University of Technology, 2628 CD Delft, The Netherlands, and also with QuTech, 2628 CJ Delft, The Netherlands (e-mail: nef@live.nl).

Mohsen Mortazavi is with the Department of Microelectronics, Delft University of Technology, 2628 CD Delft, The Netherlands, and also with Bosch, 5624 CL Eindhoven, The Netherlands.

Masoud Babaie is with the Department of Microelectronics, Delft University of Technology, 2628 CD Delft, The Netherlands, and also with QuTech, 2628 CJ Delft, The Netherlands.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2024.3364968.

Data is available on-line at https://doi.org/10.24433/CO.7837097.v1

Digital Object Identifier 10.1109/JSSC.2024.3364968

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

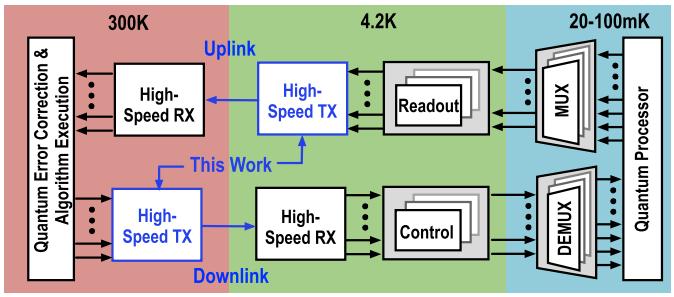


Fig. 1. Simplified block diagram of a scalable quantum computing system incorporating a cryo-CMOS high-speed wireline link.

[6], [7], [8]. After each periodic readout cycle, the combined state of the physical qubits must be decoded to extract any eventual error. Considering the computational complexity of QEC decoding algorithms and their required short execution time (well below 1 μs for typical solid-state qubits [24], [25]), a low-latency, high-performance classical processor is needed for the decoder implementation. Hence, it is challenging to integrate the required processor inside the refrigerator due to the limited available cooling power [26].

Assuming that the readout data from the qubits are digitized at the 4-K plate of the dilution refrigerator, these data then need to be serialized from different readout blocks at CT and transferred to the classical processor at RT, as illustrated in Fig. 1. Furthermore, for the cryo-CMOS control electronics to then apply the correction to the qubits, a fast downlink is necessary to send the gate instructions down. Hence, to accomplish those tasks and to significantly reduce the number of cables between the dilution refrigerator and RT equipment, there is a need for a high-speed wireline transmitter (TX) operating at CT/RT, as presented in [27]. Even if, in the future, QEC can be implemented within the cryogenic environment, a high-speed TX is still necessary for chip-to-chip communication.

Compared with [27], this article aims to quantify the system's requirements, extend the design procedure, and give a more detailed evaluation of the results. This article is organized as follows. Section II defines the system specifications. Section III discusses the architecture and gives a detailed analysis of the circuit design. Section IV evaluates the measurement results, and this article is concluded in Section V.

II. Wireline TX System Specifications

Based on QEC requirements and channel loss, this section gives a guideline for determining the TX specifications, such as data rate, power consumption, modulation format, output swing, jitter, and linearity. First, we estimate the required uplink and downlink data rates based on QEC requirements. Fig. 2(a) shows a simplified system diagram of QEC using a repeated SC with a distance of d. In this scheme, each logical qubit contains d^2 data qubits (D) and d^2-1 ancilla measurement qubits (X/Z). In each SC cycle, the measurement data of the X/Z ancilla qubits are serialized by the uplink wireline TX at 4.2 K and transmitted to a set of decoders located at RT. The decoder then finds the estimated error on the data qubits. Based on this outcome, the Pauli frame unit (PFU)

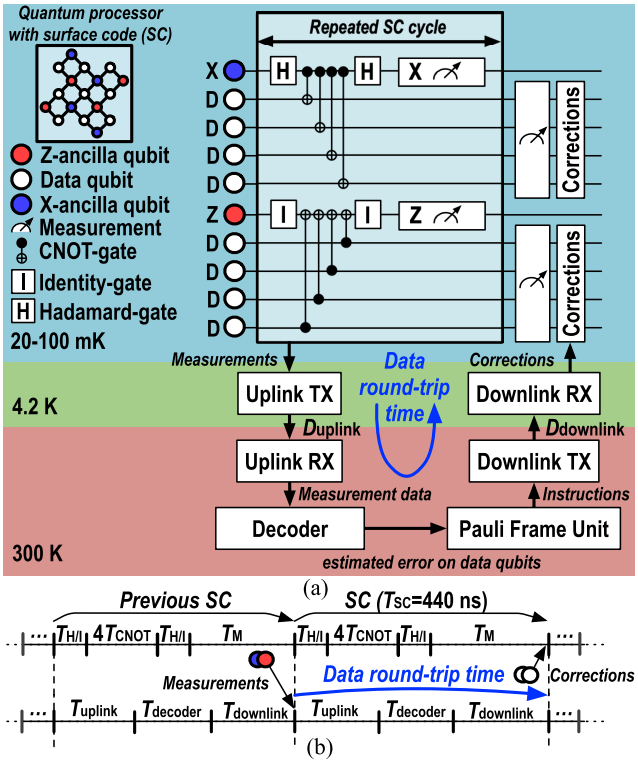


Fig. 2. (a) System diagram of quantum processor with repeated SC. The X/Z ancilla qubit measurement data are transferred at D_{uplink} to a real-time decoder that updates the PFU. The PFU stores the estimated errors of the physical data qubits and transmits the correction instructions at D_{downlink}. (b) Timing diagram illustrating the SC cycle, including single-/two-qubit gate time and measurement time. (c) Timing diagram of the data round trip, including uplink, decoder, and downlink. To prevent backlog, the total data round-trip time should be lower than the SC cycle time.

tracks the error and passes the correction instruction to the downlink TX. The downlink receiver (RX) then de-serializes the instruction data for the cryogenic controller to apply the corresponding correction gates to the data qubits. Each SC cycle includes two Hadamard/Identity gates, four controlled-NOT (CNOT) gates, and the measurement of the X/Z ancilla qubits. The corrections can be applied after measuring the data qubits. The frequency of data qubit measurements depends on the type of quantum algorithm. In the worst case, a correction should be applied between each cycle, but in practice, this measurement rate is always lower [28], [29], [30]. By considering Fig. 2(b) and assuming a Hadamard/Identity gate time ($T_{\text{H/I}}$) of 20 ns, a CNOT gate time (T_{CNOT}) of 40 ns, and a measurement time (T_{M}) of 200 ns, the total SC cycle time (T_{SC}) is 440 ns for state-of-the-art Transmons [31]. As can be gathered from Fig. 2(b) and (c), to prevent a backlog or increase in the execution time of quantum algorithms, the total data round-trip time, including the transfer time for the uplink (T_{uplink}), decoder (T_{decoder}), and downlink (T_{downlink}), should be lower than the rate at which X/Z measurement data are produced [32]. Since the measurement of each ancilla qubit produces a 1-bit data stream ("0" or "1"),[1] the uplink data rate

[1]Some decoding algorithms need the entire measured data stream of in-phase and quadrature analog-to-digital converters (ADCs) of the readout chain [33], thus requiring a substantially higher uplink data rate.

can be estimated by

$$D_{\text{uplink}} = \frac{N \cdot \left(d^2 - 1\right)}{T_{\text{uplink}}} \tag{1}$$

where N is the number of logical qubits. Considering the QEC decoder takes $T_{\text{decoder}} \approx 90\,\text{ns}$ [31], 175 ns of transfer time will be available for either T_{uplink} or T_{downlink}.[2] On the other hand, with the current inherent infidelity of $\sim 10^{-3}$ for state-of-the-art physical qubit technologies [24], [25], at least a distance of 23 is needed to reach a logical error below 10^{-12} for a practical fault-tolerant quantum computer [34]. Considering $T_{\text{uplink}} = 175\,\text{ns}$ and one logical qubit with $d = 23$, an uplink data rate of 3-Gb/s is required.

On the other hand, for the worst-case scenario where a physical data qubit is measured in each SC cycle, the PFU should decide which of the m possible gate instructions needs to be applied. This instruction code should be transmitted down to CT at the same rate. Consequently, the speed required in the downlink TX may be estimated by

$$D_{\text{downlink}} \leq \log_2(m) \cdot \frac{N \cdot d^2}{T_{\text{downlink}}}. \tag{2}$$

Considering the same number of qubits and code distance and assuming that a universal instruction set of eight quantum gates is sufficient to control the qubit system [35], [36], the maximum downlink data rate for one logical qubit is 9-Gb/s. In this work, we target a data rate of 40-Gb/s, thus providing the required throughput for four logical qubits. With current qubit technologies and QEC architectures, moving toward larger numbers of logical qubits in the far future would require Tb/s throughput and extremely challenging energy efficiency for the data link [37]. Consequently, both the relaxation time (T_2^*) and fidelity of the qubits need to improve to relax the requirements on the cycle time and distance of the SC. Another approach to reducing the throughput bottleneck would be to implement a power-efficient decoder within the cryogenic environment to detect local errors, thus reducing the data rate going to the more complex decoder at RT [38].

To provide both uplink and downlink, the wireline TX must operate at CT and RT. The cooling power available in a typical dilution refrigerator is limited to roughly 1 W at the 4-K plate [9], of which the largest portion is available for the electronics, while the other part is reserved for thermal loading. Considering that the control and readout circuits take a significant part of the power budget for electronics, we target an active power consumption of 100 mW for the wireline TX, including data retiming, serializing, and driving of the wireline. Note that, based on the timing diagram in Fig. 2(c), the uplink TX at the 4-K plate is turned on for less than half the SC cycle, so the expected thermal loading will be even lower.

In the next step, the data modulation format between non-return-to-zero (NRZ) signaling and four-level pulse amplitude modulation (PAM4) should be chosen. Essentially, at the same bit rate and TX maximum output swing, PAM4 increases the number of voltage levels from *two* to *four*. Thus, the required circuit speed and bandwidth are reduced by half, while the noise tolerance is degraded, as its eye height is reduced by 3×. Fig. 3 shows the estimated loss of the system between the cryogenic TX and room-temperature RX, considering the cable insertion loss and TX's output bandwidth limitations due to the parasitic capacitance of the output pads and electrostatic discharge (ESD) diodes. At half of the baud rate (i.e., 20 GHz for NRZ and 10 GHz for PAM4), the total loss is ∼6 dB higher for NRZ signaling. Consequently, the maximum eye height normalized to the TX output swing will be ∼3 dB larger for NRZ signaling at the RX input. However, since the delay and rise/fall time of simple gates is ∼12 ps in the used technology (i.e., 40 nm), the timing skews can easily be comparable with a baud rate duration of 25 ps for an NRZ TX, thus degrading the eye width significantly. Therefore, PAM4 signaling is chosen in this design.

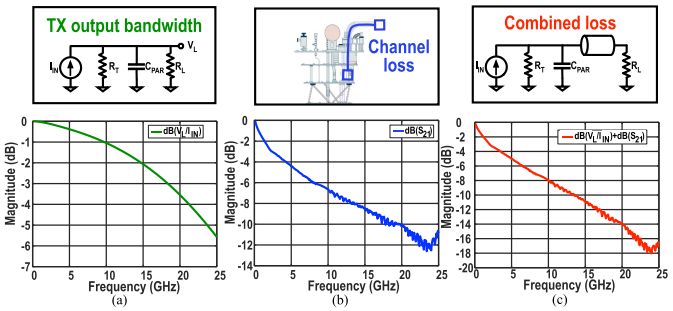


Fig. 3. Estimated loss versus frequency (a) extracted simulation of signal attenuation due to limited chip bandwidth caused by the parasitic capacitance of the ESD and pads (∼200 fF) and (b) measured insertion loss of a typical coaxial cable connecting the fridge 4-K stage to its output connector at RT. (c) Total expected loss.

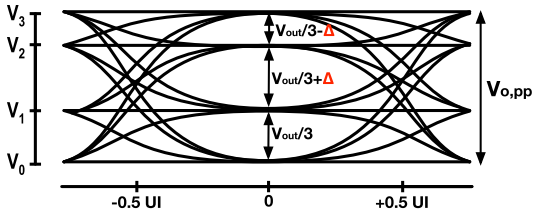


Fig. 4. Effect of the amplitude error Δ on PAM4 eye-diagram linearity.

Considering the targeted baud rate (i.e., 20 Gbaud) and estimated channel loss of 8-to-14 dB at 20 GHz, the long-range wireline standard is used as a guideline to determine the required differential output swing ($V_{o,\text{pp}}$) and jitter of the TX [39]. The standard demands an 800-mV_{pp} steady-state swing without pre-emphasis. Furthermore, the time interval that includes 99.99% of the jitter distribution (J4u) should be below 0.118 unit interval (UI). Hence, the total jitter should not exceed 5.9 ps in our 40-Gb/s PAM4 TX. Note that both deterministic jitter caused by clock skew and random jitter due to noise should be considered in J4u calculations.

Finally, the PAM4 TX should be sufficiently linear to realize the same height for all three eyes and achieve similar bit error rate (BER) for all symbols. For PAM4 signals, the ratio of level mismatch (RLM) determines the linearity performance, as illustrated in Fig. 4. The RLM is defined as the difference

[2]When a linear slowdown of the quantum algorithm is allowed, T_{uplink} and T_{downlink} can be as large as T_{SC}, and consequently, the required data rate per logical qubit reduces.

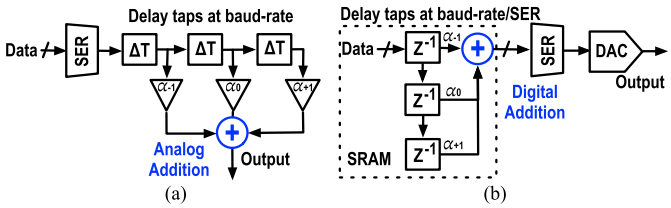


Fig. 5. FFE implementation in (a) analog domain and (b) digital domain.

between the smallest vertical eye opening divided by the full swing

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

where V_{0-3} are the mean values of PAM4 signal levels. Based on the long-range wireline standard [39], an RLM of at least 95% is required. Consequently, the maximum allowed quantization error must be less than 1.6%, which results in at least a 5-bit amplitude resolution for the TX output stage.

III. Wireline TX Architecture and Circuit Design

To overcome bandwidth limitations and compensate for the low-pass channel response, the wireline TX architecture should implement a feed-forward-equalization (FFE) filter in the analog or digital domain. The first option is to realize the pre-emphasis filter in the final stage by weighted combining of the delayed taps of the full-rate signal, as shown in Fig. 5(a). This requires the delay taps to run at the baud rate, increasing design complexity and limiting the flexibility to the fixed amount of retimers installed in the system. Alternatively, the pre-emphasis can be calculated in the digital domain and then transmitted by a high-speed digital-to-analog converter (DAC), as can be gathered from Fig. 5(b). This allows the delay taps to run at the baud rate divided by the serializer steps. However, this requires a multi-bit DAC to run at the baud rate. Using on-chip memory for FFE implementation, the DAC-based architecture is chosen, as it adds flexibility to generate multiple equalization options in the digital domain, as well as the programmability to configure different modulation sequences (i.e., NRZ and PAM4).

The block diagram of the DAC-based wireline TX architecture is shown in Fig. 6. In total, 6×, 512-word, 64-bit programmable SRAM modules, with a total size of ∼197 kb, are synthesized to allow for exploring different bit patterns, data formats, and equalization techniques in measurement. The 6× 64-UI parallel SRAM modules are decoded to feed the 10× DAC serializer slices (3-b binary, 3-b unary coded). In each serializer slice, the data are multiplexed to drive a single DAC bit. First, the SRAM data are retimed by a 64-UI retimer with a selectable clock phase (i.e., $\psi_i \in \psi_{0-7}$) to align all input data. Next, a 64:4 multiplexer (MUX) serializes this data stream to a 4-UI output stream, utilizing the available clock phases in an efficient manner. This 5-Gb/s signal is then retimed by a quarter-rate retimer and converted to 2 × 4 complementary streams, each having 25% duty-cycle pulsewidth. A current-mode logic (CML)-based 4:1 MUX then interleaves the full-swing data pulses to a 20-Gb/s differential output, driving a DAC element. Finally, the DAC element drives the output network, consisting of two 50-Ω termination resistors, a peaking coil, ESD, and differential output pads with sufficient swing. The termination resistors are implemented by the *unsilicided* polysilicon resistor, as its sheet resistance is fairly constant over temperature [10], [40]. The peaking inductor is designed for maximally flat envelope delay, increasing the bandwidth by a few GHz. The values of the output capacitance and peaking inductance stay constant down to CT, while their quality factor increases by 2×–3× [41]. Hence, due to lower passive loss, the output bandwidth is expected to slightly improve at 4.2 K. A clock generation circuit provides the necessary clock phases for all serializers and retimers in the architecture, from an external 10-GHz clock input.

A. 6-bit Current-Steering DAC

The DAC should drive the wireline with sufficient swing, linearity, and bandwidth. The final stage can be implemented in two ways: utilizing a current-mode driver or employing a source-series-terminated (SST) driver, both of which are commonly seen in wireline TXs [42], [43]. The SST or voltage-mode DAC creates the output levels based on a voltage division between the effective output resistance of the DAC cells and the load. Due to its class-D operation and not drawing constant current from supply, an SSL driver typically consumes ∼4× lower power compared with the current-mode counterpart, when generating the same output swing. Despite this advantage, a current-steering DAC is adopted in this design for a number of reasons.

First, the switches in the SST cells must be wide enough to exhibit an ON resistance (r_{ON}) well below their corresponding series resistors. These wide transistors come with a high input capacitance and need a rail-to-rail input swing to switch fast, thus demanding power-hungry pre-drivers. However, current-mode drivers only need a moderate input voltage swing and small size switching pair to steer current, thus requiring less power consumption in the pre-driver stage and achieving higher speeds due to their lower input and output parasitic capacitance. Second, r_{ON} significantly varies with process, voltage, and temperature (PVT) variations. Since the output resistance and voltage of the SST drivers are dependent on r_{ON}, additional calibration or trimming circuitry is needed to satisfy output matching and linearity of the SST driver both at RT and CT. Third, the SST structure draws a significant transient current during data transitions, thus requiring large decoupling capacitors and experiencing some data-dependent supply voltage (V_{DD}) variations. This can heavily affect the TX linearity, as the output voltage is proportional to V_{DD}. Yet, current-steering drivers draw a relatively constant current from V_{DD} for different data transitions and voltage levels. Moreover, the linearity and output swing of current-steering DACs are mainly determined by the current source accuracy, which can be stabilized more conveniently over PVT variations. Consequently, considering the possible supply voltage drop due to the long wires between the cryogenic wireline TX and the room-temperature voltage source, and potential supply/ground interferences, the current-steering driver is a more promising candidate for this design.

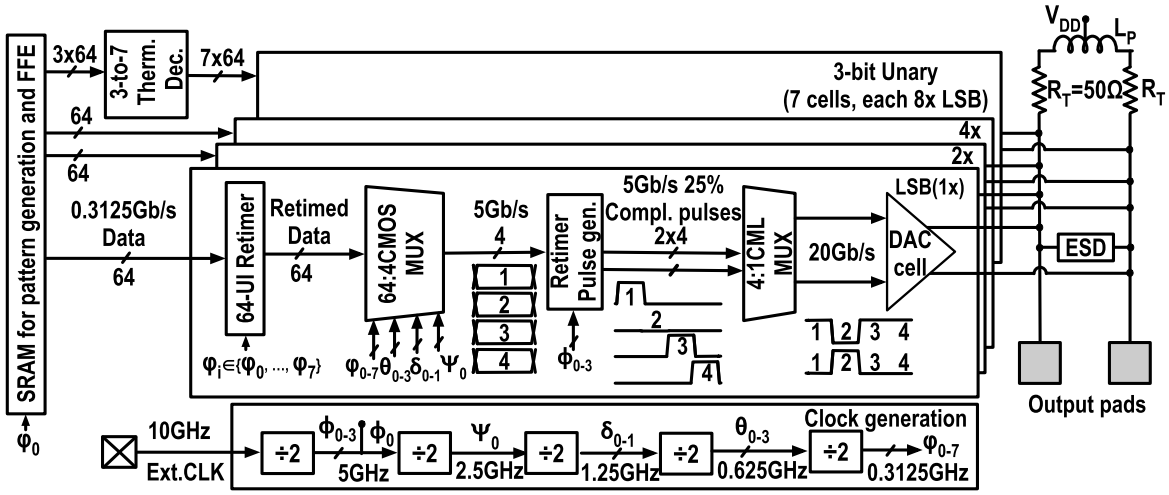


Fig. 6. Block diagram of the proposed DAC-based PAM4 wireline TX architecture, operating at 20-GHz baud rate.

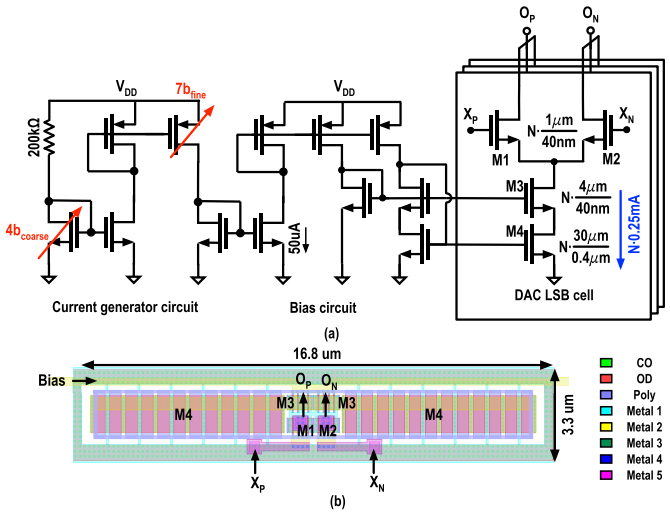


Fig. 7. (a) Schematic of the bias circuit and DAC LSB cell. (b) Layout of the DAC LSB cell.

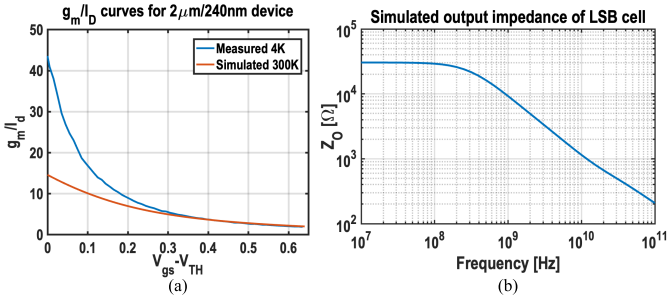


Fig. 8. (a) RT simulation and CT measurement of g_m/I_D curves for a 2-μm/240-nm transistor. (b) Simulated output impedance of the DAC LSB cell versus frequency.

As discussed at the end of Section II, at least a 5-bit DAC is needed to satisfy the RLM requirement. Taking a margin for the other errors that could be introduced, a 6-bit DAC is chosen in this design. Fig. 7 shows the schematic and layout of the DACs' least significant bit (LSB) cell. To achieve the linearity performance, the integral nonlinearity (INL) caused by current source mismatches should be safely below 0.5 LSB at both CT and RT. According to [44], the maximum INL of an n-bit DAC with $N = 2^n$ voltage levels may be approximated by

$$\mathrm{INL_{MAX}} = \frac{1}{2}\sqrt{N-1}\cdot\left(\frac{\sigma_i}{I_0}\right) \tag{4}$$

where I_0 is the LSB current and σ_i is its standard deviation. On the other hand, by using the Croon model [45], the drain-current mismatch can be predicted by

$$\frac{\sigma_i}{I_0} = \frac{A_{V_{\mathrm{TH}}}^2(g_m/I_0)^2 + A_\beta^2}{W_4 \times L_4} \tag{5}$$

where W_4, L_4, and g_m represent the width, length, and transconductance of the current source transistor (M_4), respectively. Moreover, $A_{V_{\mathrm{TH}}}$ and A_{V_β} are, respectively, the threshold-voltage and current-factor proportionality parameters [46]. According to [11], for long-channel NMOS transistors in 40-nm CMOS, the measured $A_{V_{\mathrm{TH}}}$ and A_{V_β}, respectively, increase 1.5× and 3× from RT to 4.2 K, indicating a more severe device mismatch at CT. As can be gathered from (5), to achieve a lower drain-current mismatch, g_m/I_0 should be minimized. Fig. 8(a) shows g_m/I_0 versus overdrive voltages for a 2-μm/240-nm transistor at RT and CT. It can be gathered that, in weak inversion, g_m/I_0 increases by 3× at CT, while in strong inversion, it stays constant [47], [48]. Hence, the tail sources should be biased in strong inversion with sufficient margin to avoid the increase in current mismatch.

The active area of the M_4 transistor is then calculated by substituting (5) into (4)

$$W_4 \times L_4 \geq \frac{N-1}{4\cdot \mathrm{INL_{MAX}}}\cdot\left(A_{V_{\mathrm{TH}}}^2(g_m/I_0)^2 + A_\beta^2\right). \tag{6}$$

Now, assuming that the maximum allowed INL error should be within LSB/8, and considering $A_\beta \approx 3\% \cdot \mu$m, $A_{V_{\mathrm{TH}}} \approx 10\,\mathrm{mV}\cdot\mu$m, and $(g_m/I_0) \approx 10$ S/A, the transistor core area should be at least 11 μm^2. Increasing the size further would create a large parasitic capacitance, reducing the output impedance of the current source at high frequencies.

Another contributor to nonlinearity is the finite output impedance of the current sources [49]. Depending on the DAC input code, the impedance seen at the output node varies, thus distorting the output voltage, and degrading the uniformity of eye heights. Assuming m current switches ($M_{1,2}$) are active on the left and $N-m$ on the right branch [50], the DAC output voltage may be estimated by

$$V_{\text{out}} = -R_L I_0 \times \frac{(N-2\,m)Z_o^2}{(Z_o + mR_L)(Z_o + (N-m)R_L)} \tag{7}$$

where R_L is the 50 Ω, and Z_o is the output impedance of the LSB current source. Due to the finite output impedance of the current sources, the voltage difference between the maximum and minimum levels is reduced to

$$V_3 - V_0 = 2R_L N I_0 \times \frac{Z_o}{Z_o + R_L N}. \tag{8}$$

On the other hand, the minimum eye height can be approximated by

$$V_{\text{eye,min}} = 2R_L N I_0 \times \frac{Z_o}{(3Z_o + 2R_L N)(3Z_o + R_L N)}. \tag{9}$$

By replacing (8) and (9) in (3), we have

$$\text{RLM} = \frac{9Z_o(Z_o + R_L N)}{(3Z_o + 2R_L N)(3Z_o + R_L N)}. \tag{10}$$

To obtain RLM $\geq 95\%$, $|Z_o|$ should be at least 5 kΩ. Moreover, by using (8), I_0 should be larger than 220 μA to achieve an 800-mV_{pp} differential output swing. To achieve the required output resistance, a cascode transistor M_3 is added to the current source. M_3 uses a minimum length transistor, since its drain parasitic capacitance limits the output impedance of the current source and, thus, DAC linearity at high frequencies. Consequently, by considering the required current, output impedance, and device matching, $L_4 = 400$ nm and $W_4 = 30\,\mu$m are chosen. Note that the transistor's early voltage and output resistance drop at CT [47]. Hence, the output resistance is overdesigned at RT. Fig. 8(b) shows the simulated output impedance of the DAC LSB cell at RT, offering an impedance over 5 kΩ for frequencies below ~4.7 GHz.

The DAC unary and binary segmentation is decided based on the trade-off between the DAC differential nonlinearity (DNL) and the TX area and power consumption. The total number of TX slices is determined by

$$\text{TX}_{\text{slices}} = n_b + 2^{n_u} - 1 \tag{11}$$

where n_b and n_u are the number of binary and unary bits, respectively. Note that each slice requires a complete serializer path. Therefore, it is preferred to employ a fully binary DAC to optimize chip area and power consumption. However, this approach increases DNL significantly [51], as the DNL error due to the major transition can be approximated by

$$\text{DNL} = \sqrt{2^{n_b}} \times \left(\frac{\sigma_i}{I_0}\right). \tag{12}$$

By considering this trade-off, a 3-bit binary, 3-bit unary DAC is chosen. In this way, the number of slices reduces from 63 for a fully unary structure, to 10, and the DNL improves from 0.5 LSB, for a fully binary structure, to 0.18 LSB. Further increasing the number of unary bits, e.g., a 4-bit unary DAC, could achieve 0.125-LSB DNL but would require 17 TX slices, thus increasing the digital power consumption of the serializing paths by 70%.

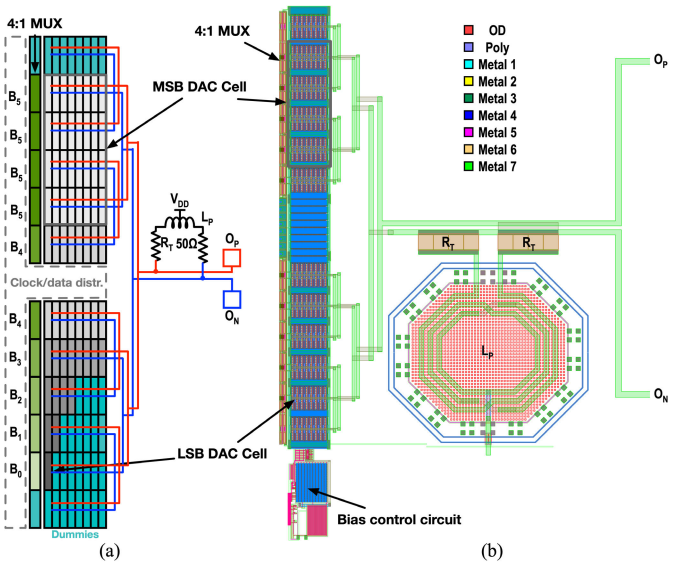


Fig. 9. (a) DAC structure, representing the binary bits B_{0-2} and unary bits B_{3-5} shown in the gray-coded unit cells; the light blue cells symbolize dummies; the red and blue lines represent the output H-tree. (b) DAC layout, including the 4:1 MUX (left) and output matching network (right).

The biasing circuit shown in Fig. 7(a) mirrors a reference current by a factor of 5 to generate the LSB current of 0.25 mA. The reference current is generated internally by an unsilicided resistor, which is fairly constant over temperature [10], [40]. To allow for wide temperature operation with constant output swing, the bias current can be adjusted digitally by trimming the coarse and fine current mirror DAC. In layout, the big current source transistors, $M_{3,4}$, are split symmetrically in two, such that the height of DAC unit cells can be reduced to 3.3 μm, as illustrated in Fig. 7(b). This way, the length of the wires connecting 4:1 MUX outputs to DAC cells is minimized, thus reducing parasitics and delay mismatches of the input signals.

During the input data transition, the voltage across the current source is affected. This data-dependent disturbance couples to the bias voltage of the current source through $M_{3,4}$ parasitic capacitances, degrading the DAC settling and linearity [52]. To reduce this coupling, large and distributed decoupling capacitors are placed at $M_{3,4}$ gate bias voltages, and special attention is given to reducing M_3 drain–source and M_4 gate–drain parasitic capacitances during layout. Besides, the distribution of the bias voltages is done perpendicular to the data lines. Moreover, since the substrate sheet resistance increases by five orders of magnitude at CT [41], substrate contacts are placed around and close to DAC unit cells to avoid possible distortions due to the floating body of transistors. The resulting DAC structure and layout are shown in Fig. 9. The unit-cell slices have been laid out in symmetric rows and connected by H-trees to minimize delay mismatch between the different branches.

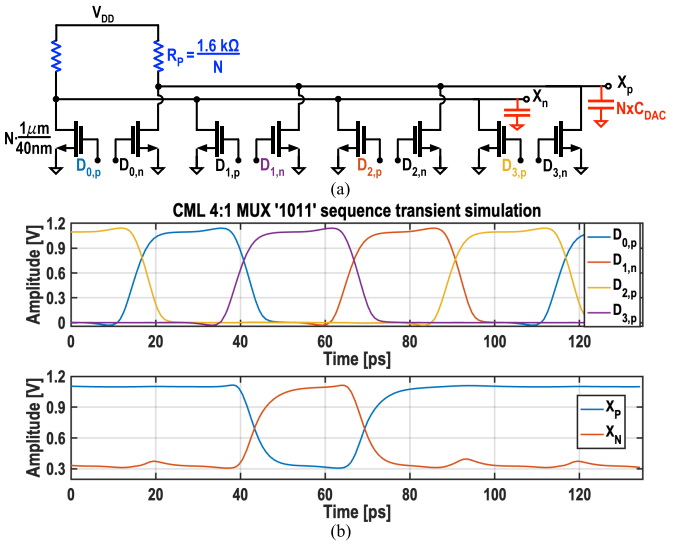


Fig. 10. (a) Schematic of the 4:1 CML MUX. The pull-up resistors (R_P) are inversely proportional to the size of their corresponding DAC slice with an input parasitic capacitance of $N \times C_{\text{DAC}}$. (b) Simulated input and output waveforms of the 4:1 CML MUX for a "1011" sequence.

B. 4:1 Multiplexer

As shown in Fig. 10(a), a CML-based 4:1 MUX is chosen, because it can reach higher speeds than conventional logic, as it only needs to steer a small current rather than switch a full-swing voltage. The circuit combines the 25% rail-to-rail interleaved input pulses in the current domain to generate a differential output data stream, directly driving the associated DAC element. The MUX is designed without a tail transistor to counteract the reduced voltage headroom due to the higher threshold voltage at CT [12]. Fig. 10(b) displays the functioning of the 4:1 CML MUX with a transient simulation of a "1011" sequence. All interleaved data pulses pull down the left branch (i.e., X_n) except for the second data pulse (i.e., $D_{1,n}$), which pulls down the right branch (i.e., X_p). This results in a differential "1011" data stream at the output nodes.

Note that the MSB cells have a larger input capacitance than the LSB cells in the DAC structure. Consequently, if the same components' values are utilized in the 4:1 MUX of all slices, the data reach the DAC input ports with different delays, thus increasing the system's deterministic jitter. To mitigate this issue, the load conductance and pseudo-differential pairs are proportionally scaled with the size of their corresponding DAC slice to achieve a similar bandwidth and voltage amplitude at the output of all 4:1 multiplexers, thus preventing any systematic delay mismatch.

C. Quarter-Rate Retimer

The quarter-rate retimer, complementary data generator, and pulse generator prepare the data for the 4:1 MUX. The circuit schematic and its timing diagram are shown in Fig. 11. The last stage of the 64:4 MUX uses a different clock tree (i.e., ψ_0) than the pulse generator running at the 4-UI domain (i.e., ϕ_{0-3}). This results in an unknown time offset, which can hardly be predicted by simulating the extracted layout. Therefore, to prevent any data violation, a flip-flop retimer

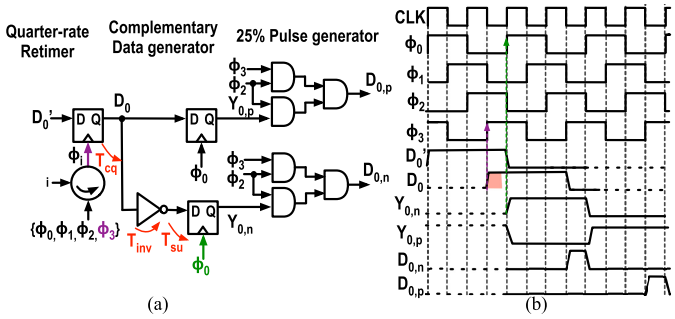


Fig. 11. Quarter-rate retimer, complementary data generator, and 25% pulse generator. (a) Schematic illustrating the worst case critical path in red. (b) Timing diagram showing the total delay.

is added whose clock phase (i.e., $\phi_i \in \{\phi_0, \ldots, \phi_3\}$) can be selected with a multiplexer-based phase rotator. Next, the data are converted to complementary form and again retimed to ensure any delay mismatch caused by the additional inverter in one of the complementary paths is removed.

The pulse generator then generates the required 25% duty-cycle pulses (i.e., 1-UI) by combining two 50% overlapping 4-UI clock phases and the complementary data using two AND gates in cascade. Cascaded AND gates have been chosen, as opposed to a single three-input AND, as this would have multiple stacked devices, resulting in a much slower rise/fall time at CT due to the limited voltage overdrive. In this design, the pulse generation is done locally, since alternatively distributing complementary 25% non-overlapping clocks would require ~2× more power-hungry buffers in the clock path and set tighter constraints to the clock distribution. Similar to the output network, the differential phases of the quadrature clocks are distributed through H-trees and upper level metals to minimize power consumption and clock delay mismatches.

Considering the worst-case scenario and the timing constraint diagram shown in Fig. 11, the delay between the triggering edge of the second retimer (i.e., ϕ_0) and the first sampling edge (i.e., $\phi_i = \phi_3$) must be long enough to accommodate the inverter propagation delay (T_{inv}), the edge-to-edge jitter between the ϕ_0 and ϕ_3 clock phases, the flip-flop clock-to-q delay (T_{cq}), and the setup time (T_{su}). The total edge-to-edge jitter is a combination of peak-to-peak deterministic jitter DJ_ϕ caused by clock skew and device mismatches, as well as random jitter RJ_ϕ due to noise. Targeting a BER of 10^{-12}, we aim for a 14 sigma (i.e., $\pm 7\sigma$) design for the random jitter. Consequently, the maximum baud rate of the system may be estimated by

$$f_{\text{baud}} \leq \frac{1}{\left(T_{\text{inv}} + T_{\text{cq}} + T_{\text{su}} + \text{DJ}_\phi + 14 \cdot \text{RJ}_\phi\right)}. \tag{13}$$

Considering $T_{\text{inv}} = 12$ ps, a simulated $T_{\text{cq}} = T_{\text{su}} = 22$ ps for a high-speed flip-flop at RT, a simulated $\text{DJ}_\phi = 934$ fs from the extracted layout, and a combined random jitter of 48 fs_{rms} from the divider and potential PLL clock source [53], a maximum speed of ~17.4 Gbaud can be achieved.

To ensure that the flip-flop retimers do not slow down at CT, the use of transmission gates or multiple stacked devices should be prevented due to the increased threshold voltage. Therefore, in this design, a true single-phase clock (TSPC)

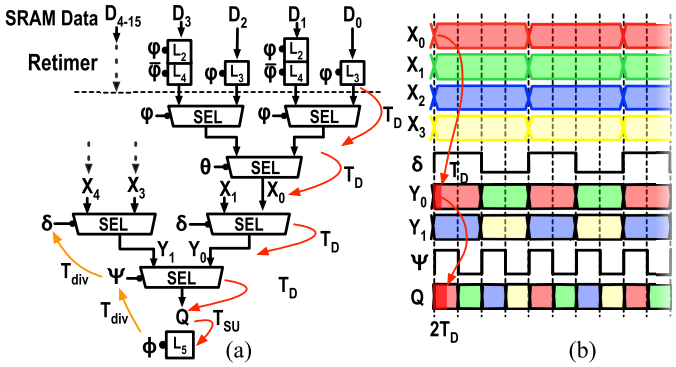


Fig. 12. Simplified schematic of a binary-tree CMOS MUX using a single clock phase in each rank—(a) circuit and (b) timing diagram.

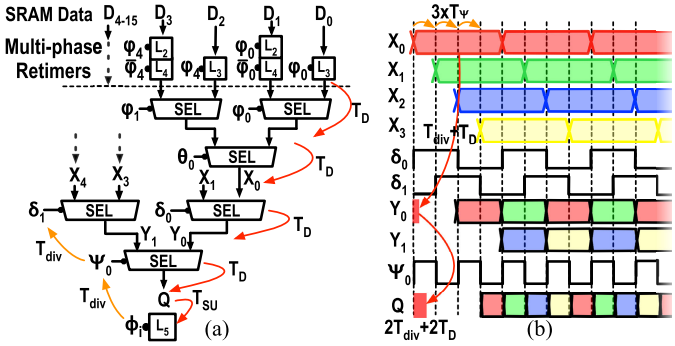


Fig. 13. Simplified schematic of a binary-tree CMOS MUX using quadrature clock phases in each rank—(a) circuit and (b) timing diagram.

dynamic flip-flop is adopted, limiting the number of stacked devices and maximizing the system's speed. The drawback of TSPC logic is its susceptibility to transistor leakages originating from sub-threshold conduction, gate oxide, and source and drain junctions [54]. Since the states are stored on the parasitic capacitance of small-sized transistors, if the clock period is excessively long (i.e., $\gg$10 ns), the leakage current could discharge this capacitor, thereby corrupting the logic state. This sets a constraint on the minimum operating frequency of TSPC flip-flops. Since the TSPC flip-flop is only used in the quarter-rate retimer, it is intended to only work at frequencies higher than a few GHz. Moreover, at CT, due to the 3$\times$ larger sub-threshold slope and increased threshold voltage, the sub-threshold conduction and leakage of source/drain junctions become significantly lower [55], [56]. The gate leakage is also expected to decrease by $\sim$2 [57]. Hence, the minimum operating frequency of the TSPC flip-flop becomes significantly lower at CT.

Due to the transistors' higher mobility at CT [13], T_{cq} and T_{inv} are expected to reduce by $\sim$10%, thus increasing the maximum baud rate to $\sim$20 Gbaud. Besides, the TX speed can be further improved in future implementation by removing the inverter from the critical timing path and realizing complementary data streams earlier in the serializer chain. However, this comes with the cost of power consumption and complexity due to the extra serializing paths needed for the additional inverted signals.

D. 64:4 Multiplexer

For serializing at lower data rates, a conventional 2:1 MUX cell can be used. Each conventional 2:1 MUX cell has one selector and three latches, two of which block glitches from previous stages [58]. A D:Q (e.g., 64:4) binary-tree CMOS MUX requires D–Q (e.g., 60) 2:1 MUX cells and 3$\times$ (D–Q) (e.g., 180) latches operating at frequencies ranging from f_{baud}/D (e.g., 312.5 MHz) to $f_{baud}/2Q$ (e.g., 2.5 GHz), thus increasing the TX power consumption. As shown in Fig. 12, to reduce power, the input and output latches of the MUX tree are maintained, but the intermediate latches are removed. If all selectors in each MUX rank are clocked with the same clock phase, the delay of the selectors (T_D) and frequency dividers (T_{div}) will eventually limit the baud rate at the output of D:Q MUX to

$$f_{Bd,MUX1} \leq \frac{1}{\log_2(D/Q)\cdot(T_D+T_{div})+T_{su}}. \quad (14)$$

Considering $T_D \approx 25$ ps, $T_{div} \approx 25$ ps, and $T_{su} \approx 22$ ps, the maximum baud rate of 64:4 MUX cannot be higher than 4.5 Gbaud, thus limiting the system speed. Alternatively, the selectors can be clocked using the available quadrature phases of the clocks, as proposed in [59]. As shown in Fig. 13, the lower rank selectors use the quadrature clocks generated by the divided clock of their corresponding higher rank selector. In this way, the required delay between each selector's inputs is guaranteed by the appropriate selection of available clock edges in successive stages. Therefore, this structure is scalable as long as the sum of the selector and divider delays does not exceed half the period of the highest rank clock. Since the intermediate-rank selectors mitigate the effect of the data and divider delays of their previous stages, the maximum baud rate of this structure is mainly determined by its output retimer. Hence

$$f_{Bd,MUX2} \leq \frac{1}{T_{su}+T_{div,\phi\rightarrow\Psi}+DJ_{\phi,\Psi}+14\cdot RJ_{\phi}} \quad (15)$$

where $T_{div,\phi\rightarrow\Psi}$ is the divider delay between ϕ and Ψ clock domains, and $DJ_{\phi,\Psi}$ accounts for the additional delay between these two clocks due to the routing. Considering $T_{div,\phi\rightarrow\Psi} =$ 22 ps, $T_{su} = 22$ ps, $DJ_{\phi,\Psi} = 1.38$ ps, and $RJ_{\Psi} = 51$ fs$_{rms}$, the maximum baud rate of 64:4 MUX is 21.7 Gbaud, thus not limiting the performance.

E. Clock Generation

As described in the architecture in Fig. 6, both the quarter-rate retimer and 64:4 multiplexer architecture make extensive use of quadrature clock phases. An external differential clock input is divided using a CML-based quadrature divide-by-2 circuit to generate the four overlapping 5-GHz clock phases (ϕ_{0-3}), as shown in Fig. 14(a) and (b). The subsequent divide-by-2 circuits for generating the lower frequency clock phases (Ψ, δ, θ, and φ) utilize cascaded C^{2}MOS latches shown in Fig. 14(c) and (d). To ensure the correct clock phase relation, the bottom two C^{2}MOS latches in Fig. 14 are reset at the start; thereby, the subsequent divided clock

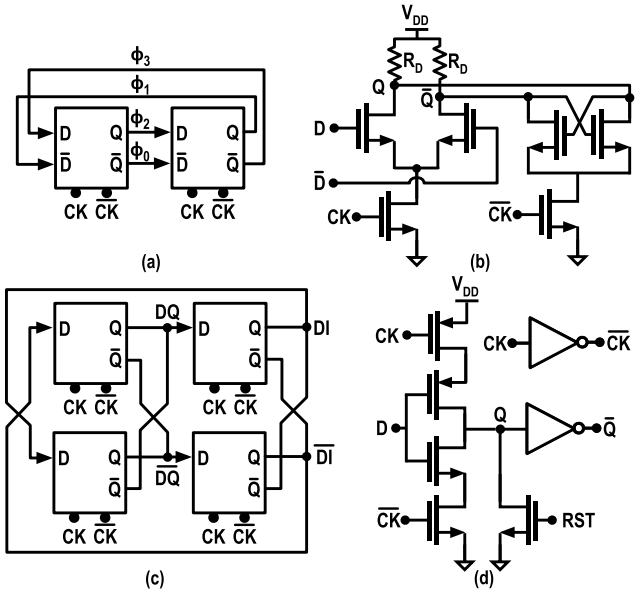


Fig. 14. Block diagrams and schematics of divide-by-2 circuits—(a) CML divider, (b) CML latch, (c) C^{2}MOS divider, and (d) C^{2}MOS latch.

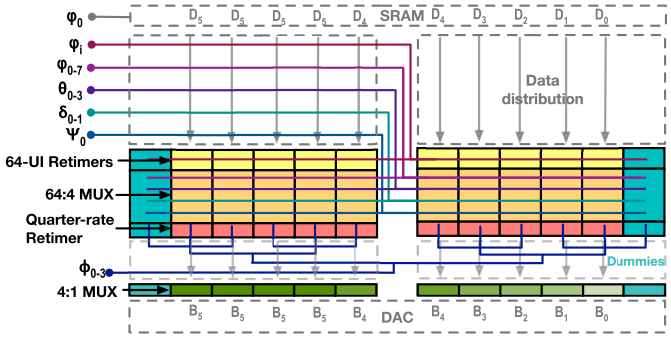


Fig. 15. Illustration of the clock distribution layout. The data distribution is done vertically, indicated in gray; the clock lines are distributed horizontally indicated with the different colors.

phases will come in the correct order. The clock distribution is illustrated in Fig. 15. The highest frequency quadrature clock (i.e., ϕ_{0-3}) is distributed directly to the quarter-rate retimers with a similar H-based clock tree as the output combiner to minimize delay mismatches. The lower frequency clocks (i.e., φ, θ, δ, and Ψ) are routed to the center of 64:4 MUX and then distributed over the digital circuitry to keep the clock skew low. The synthesized SRAM is clocked with φ_0, while the 64-UI retimers are clocked with a selectable $\varphi_i (\in \{\varphi_0, \ldots, \varphi_7\})$ to compensate for the skew in the data path. The data distribution is routed perpendicular to the clock lines to minimize crosstalk. Similar to the DAC, dummy retimers and multiplexers rows are added on left and right, since it is crucial for the timing to keep device mismatches low.

IV. Measurement Results

The wireline TX has been fabricated in a standard 40-nm bulk CMOS process. Fig. 16(a) shows the die micrograph. The total active area, excluding the SRAM, comprises 0.146 mm^2. Due to the self-heating effect at CT [60], as a precaution, the digital 64:4 serializer, CML 4:1 MUX, and DAC are

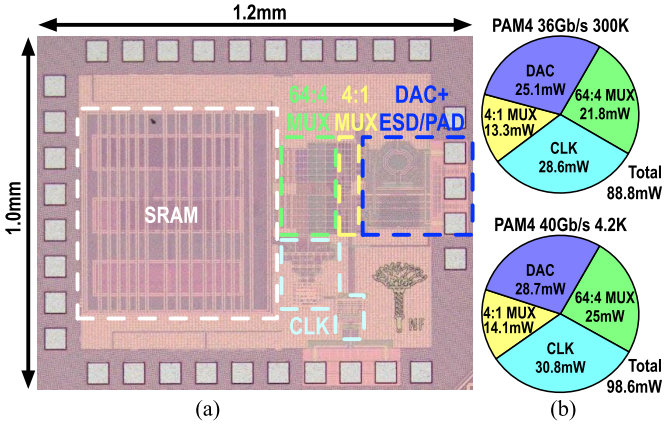


Fig. 16. (a) Die micrograph and (b) power consumption breakdown measured at 4.2 and 300 K.

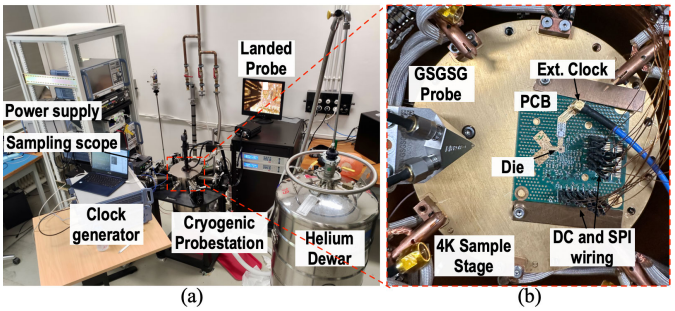


Fig. 17. (a) Complete overview of measurement setup. (b) Test board inside the probe station.

separated from SRAM by at least 100 μm. In addition, since the heat dissipation through the silicon substrate with a small contact area (i.e., $\sim$1 mm^2) is very low [61], and there is no convection possible in the vacuum chamber, most heat needs to be dissipated by the metal connections. Therefore, it is of extra importance that the on-chip power lines are designed with thick upper metals to dissipate the heat.

The chip has been characterized both at 4.2 and 300 K to test the functionality at the required operating temperature range. The chip was measured inside a Lakeshore CPX cryogenic probe station. The measurement setup is shown in Fig. 17. The die is glued onto a sample PCB, which is clamped on top of the 4.2-K sample stage of the probe station. No solder mask is used on the bottom of the PCB, and extra vias are added to the ground plane to thermally anchor the chip to the cold plate. A 10-GHz single-ended external clock with 18.5-fs rms jitter is generated using the R&S SMA-100B and connected to the chip with an on-PCB balun. Then, the differential clock signals are terminated by two series on-chip 50-Ω resistors and amplified by an ac-coupled buffer to drive the frequency dividers. The center tap of the termination resistors is connected to ground to provide a return path for common-mode signals. The output pads are probed using a Cascade 100-μm pitch GSGSG Z-probe rated for CTs. The SRAM test sequences are loaded with a serial peripheral interface (SPI) module, connected through the dc wire connections of the probe station.

The channel loss (S_{dd21}) of the test setup, including the probe, a 30-cm cable inside the probe arm, and a 10-cm cable

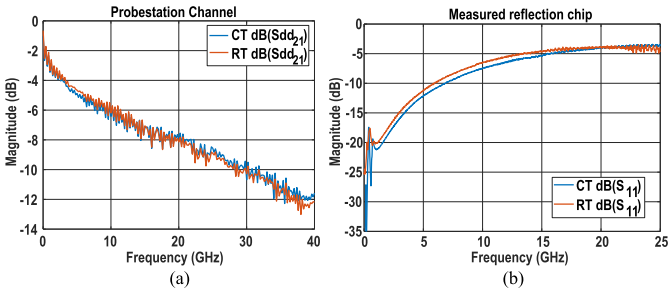


Fig. 18. (a) Measured loss of the differential probe and the cable connecting the chip to the measurement instrument at RT. (b) Measured reflection coefficient (S_{11}) at the output of the chip at 4.2 and 300 K.

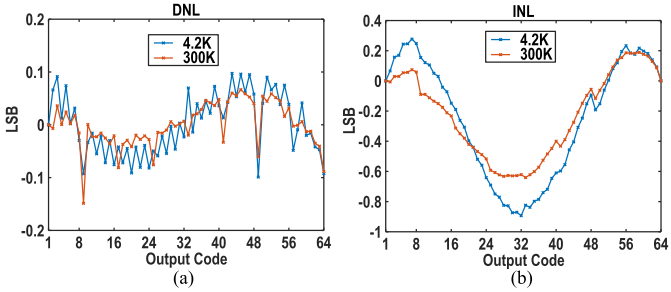


Fig. 19. Measured (a) DNL and (b) INL at 4.2 and 300 K.

outside the probe station, was measured with a vector network analyzer (VNA). First, the error terms up to the output ports of the VNA were measured using the Anritsu 3652K calibration kit. Second, the complete channel, including cables and probe, was calibrated using the short–open–load–reciprocal (SOLR) standard of a CSR-30 calibration substrate at both CT and RT. The total channel loss of the setup could then be extracted by taking the difference in the error terms of the VNA calibration and complete channel calibration. Fig. 18(a) shows that the measured channel loss is about ∼8 dB at 20 GHz for both temperatures.

The output reflection coefficient (S_{11}) of the chip was measured and de-embedded using the same CSR-30 calibration substrate. As can be gathered from Fig. 18(b), without any trimming, the chip is completely matched at low frequencies at 4.2 and 300 K, indicating that the sheet resistance of the unsilicided polysilicon termination resistors is fairly constant over temperature [10], [40]. Moreover, the bandwidth in which S_{11} remains below −10 dB is ∼10% larger at CT, mainly due to the reduction of the parasitic capacitance to ground as the silicon substrate becomes highly resistive [62].

The INL and DNL are extracted from a ramp signal generated with the SRAM, and the results are shown in Fig. 19. The largest jumps in DNL happen in every eighth code and are attributed to the 3-bit unary-to-binary transitions. The peak-to-peak INL is 0.8 LSB at RT and increases to 1.2 LSB at CT. This is expected as the device matching degrades, and the DAC shows more third-order nonlinearity, since the output impedance of its current sources reduces, as discussed in detail in Section III-A. To evaluate the DAC's performance without being limited by the clocks' deterministic and random jitter, a relatively low-frequency (i.e., ∼20 MHz) sine wave was loaded into the on-chip SRAM, and the output of the 20-GS/s

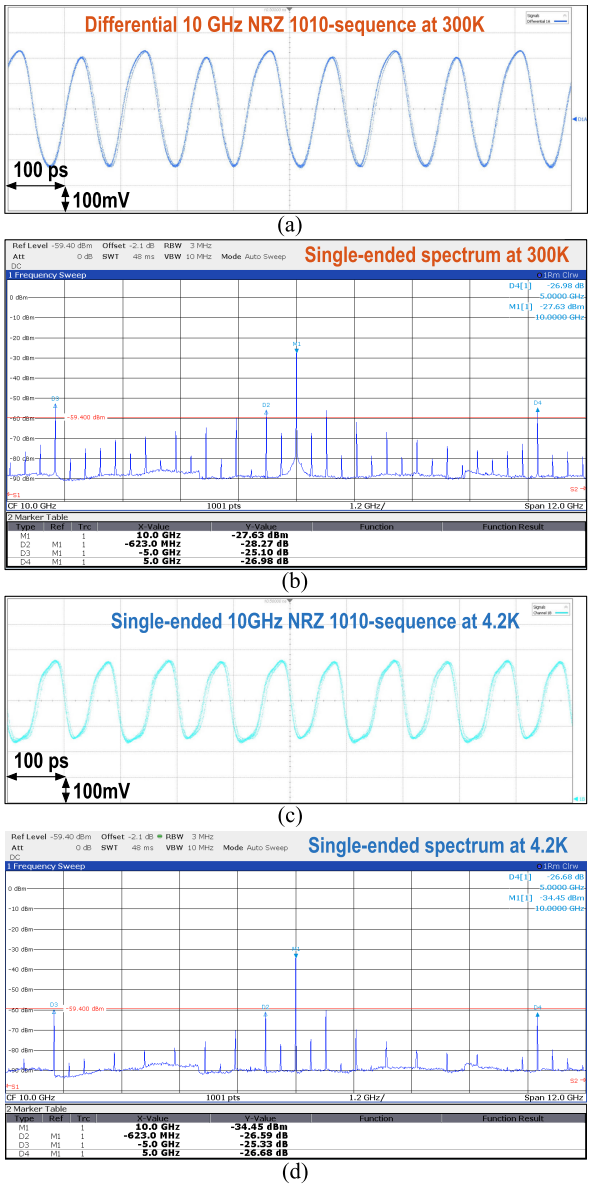


Fig. 20. Measured output of the TX delivering a 10-GHz periodic 1010 NRZ sequence (a) differential output waveform and (b) its spectrum at 300 K, and (c) single-ended waveform and (d) its spectrum at 4.2 K. Note that the spectrum was measured with relative power.

wireline TX was captured by a sampling scope. At RT and CT, a signal to noise and distortion ratio (SINAD) of 35.8 and 36.2 dB is achieved, thus satisfying the linearity requirement and showing the reduction of the DAC's output resistance at CT.

The clock's deterministic and random jitter limits the DAC performance at high frequencies. To evaluate this, a 20-Gb/s 1010 NRZ sequence (i.e., a 10-GHz period signal) is generated, and the resulting waveform and spectrum are measured at the TX output at 4.2 and 300 K, as shown in Fig. 20. Note that the output of the CML 4:1 MUX is not retimed while traveling to the TX output. Consequently, any systematic duty-cycle mismatch among the *four* non-overlapping clocks results in a deterministic jitter at the output. Due to the duty-cycle mismatches in 4:1 MUX, spurious tones with a level of approximately −26 dBc at ±5 GHz from 10-GHz carrier are observed. This results in a deterministic rms (peak-to-peak)

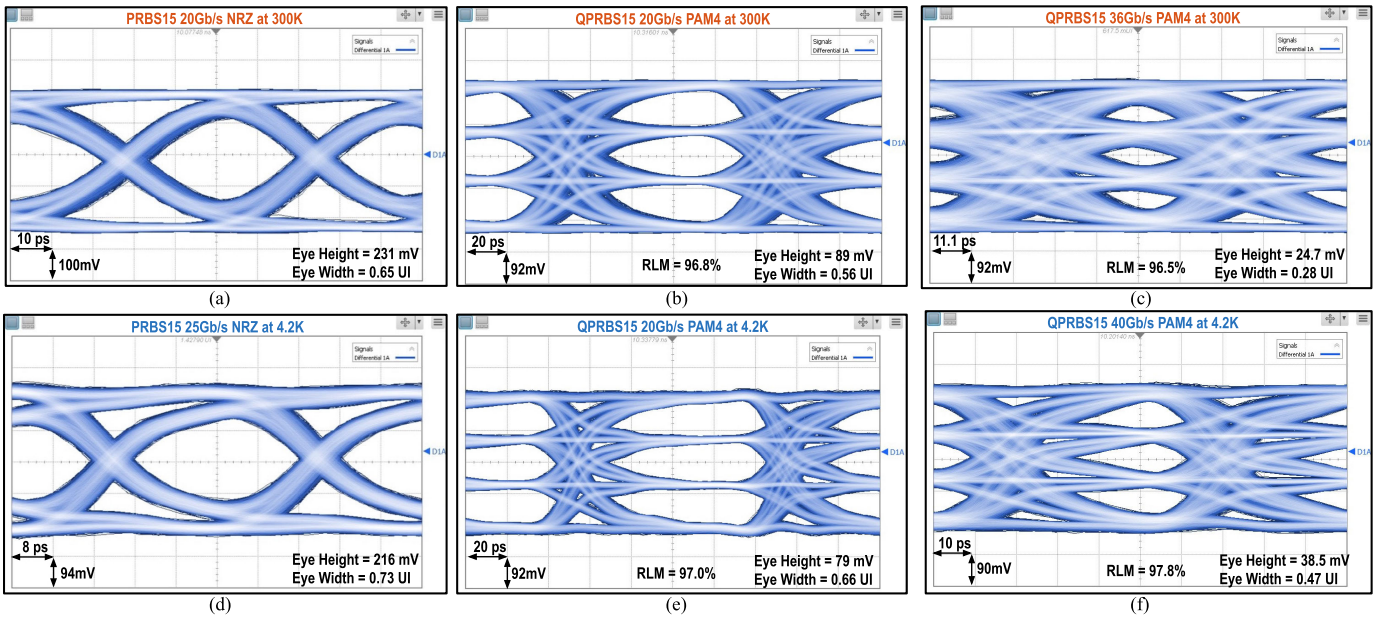


Fig. 21. Measured eye diagrams with 10-cm channel—(a) 20-Gb/s NRZ at 300 K, (b) 20-Gb/s PAM4 at 300 K, (c) 36-Gb/s PAM4 at 300 K, (d) 25-Gb/s NRZ at 4.2 K, (e) 20-Gb/s PAM4 at 4.2 K, and (f) 40-Gb/s PAM4 at 4.2 K.

jitter of 1.12 ps (3.17 ps) at CT and 1.14 ps (3.22 ps) at RT [59], [63]. This amount of deterministic jitter is somewhat higher than the simulated value of 1 ps but still low enough to satisfy the J4u requirement for our targeted baud rate. To alleviate this issue in a future version, a duty-cycle error correction circuit similar to [64] can be implemented.

The eye diagrams shown in Fig. 21 were measured using a Keysight N1094B sampling scope without using additional equalization or de-embedding. The SRAM is programed with a 2^{15}-length pseudorandom binary sequence (PRBS)-15 for NRZ and a quarternary QPRBS-15 for PAM4. The maximum speeds with sufficient eye opening at a BER of $<10^{-15}$ are explored for different modulation schemes and operating temperatures. At RT, a maximum speed of 20-Gb/s NRZ and 36-Gb/s PAM4 is achieved. For the highest baud rates at RT, the measured eye heights (widths) of NRZ and PAM4 are 231 mV (0.65 UI) and >24.7 mV (0.28 UI), respectively, with 96.5% RLM. At CT, due to the mobility improvement, the baud rate could be increased, and therefore, 25-Gb/s NRZ and 20- and 40-Gb/s PAM4 are measured. At CT, the measured eye heights (widths) of NRZ and PAM4 are 216 mV (0.73 UI) and >38.5 mV (0.47 UI), with 97.8% RLM. Consequently, at the maximum data rate, the TX achieves 2.46-pJ/b (2.47-pJ/b) energy efficiency at CT (RT), while consuming 98.6 mW (88.8 mW) from a 1.1-V power supply. The power breakdown charts, excluding the SPI controller and SRAM, are shown in Fig. 16(b). The dynamic power remains almost constant at RT and CT, since the power consumptions of the digital part of the serializer and DAC are mainly determined by $CV_{\mathrm{DD}}^2 f$ and the required voltage swing, respectively. The static power consumption of the digital circuits was measured when the chip was idle. Due to the larger threshold voltage, higher subthreshold slope, and lower gate leakage, the static power consumption of the multiplexers and SRAM reduces from 1.55 mW at RT to 0.17 mW at CT.

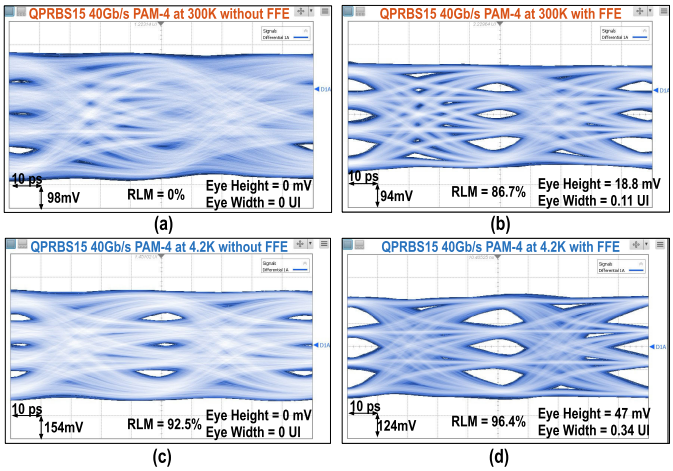


Fig. 22. Measured 40-Gb/s PAM4 eye diagrams with 1.2-m channel—(a) without FFE at 300 K, (b) with FFE at 300 K, (c) without FFE at 4.2 K, and (d) with FFE at 4.2 K.

To illustrate the effectiveness of FFE, another set of measurements was done by adding an extra 1.2-m cable to the channel. The required FFE taps were calculated based on the measured impulse response of the channel and quantized to one pre-cursor and five post-cursors, with values $[-0.0156\ 1.000\ -0.1562\ -0.0156\ 0.0156\ -0.0156\ -0.0156]$. A QPRBS15 PAM4 40-Gb/s waveform was generated with and without FFE, and the results are shown in Fig. 22. At RT, by applying FFE, the minimum eye width and height extend from a closed eye to 18.8 mV and 0.11 UI, respectively. At CT, the measured eye width and height extend from eye closure to 47 mV and 0.34 UI, respectively. Hence, FFE is very effective at extending the eye opening for long wireline channels.

Table I summarizes the performance of the proposed TX and compares it with relevant prior DAC-based and multi-tap

TABLE I
COMPARISON TABLE WITH PRIOR DAC-BASED AND MULTI-TAP PAM4 WIRELINE TXS

	This work		[42] ISSCC'17	[43] ISSCC'18	[65] ISSCC'19
Temperature [K]	300	4.2	300	300	300
Data-rate [Gb/s]	36	40	36	45	112
Power [mW]	88.8	98.6	84	120	175*
Efficiency [pJ/b]	2.47	2.46	2.33	2.67	1.56*
RLM [%]	96.5	97.8	-	92	94
Max. Vpp	0.8		0.8	1.3	1
Signalling	PAM4		PAM4	PAM4	PAM4
Output driver	CML		CML	SST	H-bridge
FFE technique	DAC		4-taps	DAC	DAC
Supply [V]	1.1		1/1.5	1	0.9/1.2
Technology [nm]	40		28	28	7
Active area [mm^2]	0.146		0.05	0.28	0.193

*including PLL

TABLE II
SUMMARY OF THE CRYO-CMOS CHALLENGES, DIVIDED INTO DIFFERENT BEHAVIORS, THEIR CONSEQUENCE, AND THE IMPLEMENTED TECHNIQUE

Behavior	Consequence	Implemented technique
Increased threshold voltage	Drawback for digital speed	Stacked devices are minimized in 4:1 CML MUX and TSPC flip-flops
g_m/I_d increase in weak inversion	Drawback for linearity	DAC cells are biased with sufficient overdrive
Increased mobility	Advantage for digital	Exploited to increase speed in quarter-rate retimer
Increased substrate resistance	Advantage for bandwidth	Bandwidth of interconnect is exploited to increase speed
Lower thermal noise	Advantage for SNR	Relaxes swing requirements for same eye opening
Larger device mismatch	Drawback for linearity	Sizing current-sources in DAC
Self-heating	More severe	Increased spacing for sensitive circuits and thicker metal routing
Long wiring	Worse supply variation	Current-steering DAC is employed

PAM4 wireline TXs. This work achieves a similar data rate and energy efficiency as prior art, while it stands out by maintaining its linearity and RLM down to CT. The drawbacks associated with cryo-CMOS devices are mitigated, while their speed enhancement is harnessed to achieve full performance. Moreover, by demonstrating, for the first time, both full functionality and high efficiency over the wide temperature range, this work addresses the required high-speed wireline link for quantum computing applications.

V. CONCLUSION

This article presented the first PAM4 cryo-CMOS wireline TX for quantum computing applications. Based on QEC requirements and the channel loss, the specifications for the data link between the control electronics at CT, and a classical processor at RT have been quantified. Guidelines were also developed to efficiently design different building blocks of the proposed DAC-based TX at CT. As summarized in Table II, by circumventing the drawbacks of cryo-CMOS devices (i.e., higher threshold, larger mismatch) and exploiting their higher speed, the TX maintains high power efficiency, linearity, and data rate down to CT. At CT (RT), the prototype achieves 40-Gb/s (36-Gb/s) PAM4 transmission with 2.46-pJ/b (2.47-pJ/b) efficiency and 97.8% (96.5%) RLM. Therefore, this work satisfies the requirements of a high-speed data link between classical and quantum processors, paving the way toward realizing large-scale quantum computers.

ACKNOWLEDGMENT

The authors would like to thank Bishnu Patra, Mohammad Ali Montazerolghaem, Bagas Prabowo, Ehsan Shokrolahzade, Atef Akhnoukh, and Zu-Yao Chang for the technical discussions and measurement support.

REFERENCES

[1] A. Morvan et al., "Phase transition in random circuit sampling," 2023, *arXiv:2304.11119*.
[2] Y. Kim et al., "Evidence for the utility of quantum computing before fault tolerance," *Nature*, vol. 618, no. 7965, pp. 500–505, Jun. 2023, doi: 10.1038/s41586-023-06096-3.
[3] V. von Burg et al., "Quantum computing enhanced computational catalysis," *Phys. Rev. Res.*, vol. 3, no. 3, Jul. 2021, Art. no. 033055, doi: 10.1103/PhysRevResearch.3.033055.
[4] C. Gidney and M. Ekerå, "How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits," *Quantum*, vol. 5, p. 433, Apr. 2021, doi: 10.22331/q-2021-04-15-433.
[5] J. Preskill, "Quantum computing in the NISQ era and beyond," 2018, *arXiv:1801.00862*.
[6] P. W. Shor, "Scheme for reducing decoherence in quantum computer memory," *Phys. Rev. A, Gen. Phys.*, vol. 52, no. 4, pp. R2493–R2496, Oct. 1995. [Online]. Available: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.52.R2493
[7] E. Knill and R. Laflamme, "Theory of quantum error-correcting codes," *Phys. Rev. A, Gen. Phys.*, vol. 55, no. 2, p. 900, 1997. [Online]. Available: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.55.900
[8] D. Bacon, "Operator quantum error-correcting subsystems for self-correcting quantum memories," *Phys. Rev. A, Gen. Phys.*, vol. 73, no. 1, Jan. 2006, Art. no. 012340, doi: 10.1103/PhysRevA.73.012340.
[9] S. Krinner et al., "Engineering cryogenic setups for 100-qubit scale superconducting circuit systems," *EPJ Quantum Technol.*, vol. 6, no. 1, p. 2, Dec. 2019, doi: 10.1140/epjqt/s40507-019-0072-0.
[10] B. Patra et al., "Cryo-CMOS circuits and systems for quantum computing applications," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 309–321, Jan. 2018.
[11] P. A. T. Hart, M. Babaie, E. Charbon, A. Vladimirescu, and F. Sebastiano, "Subthreshold mismatch in nanometer CMOS at cryogenic temperatures," in *Proc. 49th Eur. Solid-State Device Res. Conf. (ESSDERC)*, Sep. 2019, pp. 98–101.
[12] A. Beckers, F. Jazaeri, A. Grill, S. Narasimhamoorthy, B. Parvais, and C. Enz, "Physical model of low-temperature to cryogenic threshold voltage in MOSFETs," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 780–788, 2020.
[13] 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 J. Electron Devices Soc.*, vol. 6, pp. 996–1006, 2018.
[14] J. P. G. Van Dijk et al., "A scalable cryo-CMOS controller for the wideband frequency-multiplexed control of spin qubits and transmons," *IEEE Sensors J. Solid-State Circuits*, vol. 55, no. 11, pp. 2930–2946, Nov. 2020.
[15] K. Kang et al., "A 40-nm cryo-CMOS quantum controller IC for superconducting qubit," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3274–3287, Nov. 2022.
[16] S. Chakraborty et al., "A cryo-CMOS low-power semi-autonomous transmon qubit state controller in 14-nm FinFET technology," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3258–3273, Nov. 2022.
[17] G. Kiene et al., "13.4 A 1 GS/s 6-to-8 b 0.5 mW/qubit cryo-CMOS SAR ADC for quantum computing in 40 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 214–216.
[18] B. Prabowo et al., "A 6-to-8 GHz 0.17 mW/qubit cryo-CMOS receiver for multiple spin qubit readout in 40 nm CMOS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 212–214.

[19] M. Prathapan et al., "A cryogenic SRAM based arbitrary waveform generator in 14 nm for spin qubit control," in *Proc. IEEE 48th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2022, pp. 57–60.

[20] L. Enthoven, J. van Staveren, J. Gong, M. Babaie, and F. Sebastiano, "A 3 V 15 b 157 μW cryo-CMOS DAC for multiplexed spin-qubit biasing," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 228–229.

[21] R. Acharya et al., "Multiplexed superconducting qubit control at millikelvin temperatures with a low-power cryo-CMOS multiplexer," *Nature Electron.*, vol. 6, no. 11, pp. 900–909, Sep. 2023. [Online]. Available: https://www.nature.com/articles/s41928-023-01033-8

[22] S. Pezzagna and J. Meijer, "Quantum computer based on color centers in diamond," *Appl. Phys. Rev.*, vol. 8, no. 1, Mar. 2021, Art. no. 011308. [Online]. Available: https://pubs.aip.org/aip/apr/article-abstract/8/1/011308/238677/Quantum-computer-based-on-color-centers-in-diamond?redirectedFrom=fulltext

[23] L. Petit et al., "Design and integration of single-qubit rotations and two-qubit gates in silicon above one Kelvin," *Commun. Mater.*, vol. 3, no. 1, pp. 1–7, Nov. 2022, doi: 10.1038/s43246-022-00304-9.

[24] R. Barends et al., "Coherent Josephson qubit suitable for scalable quantum integrated circuits," *Phys. Rev. Lett.*, vol. 111, no. 8, Aug. 2013, Art. no. 080502, doi: 10.1103/PhysRevLett.111.080502.

[25] R. Barends et al., "Superconducting quantum circuits at the surface code threshold for fault tolerance," *Nature*, vol. 508, no. 7497, pp. 500–503, Apr. 7497. [Online]. Available: https://www.nature.com/articles/nature13171

[26] F. Battistel et al., "Real-time decoding for fault-tolerant quantum computing: Progress, challenges and outlook," *Nano Futures*, vol. 7, no. 3, Sep. 2023, Art. no. 032003.

[27] N. Fakkel, M. Mortazavi, R. Overwater, F. Sebastiano, and M. Babaie, "A cryo-CMOS DAC-based 40 Gb/s PAM4 wireline transmitter for quantum computing applications," in *Proc. IEEE Radio Freq. Integr. Circuits Symp. (RFIC)*, Jun. 2023, pp. 257–260.

[28] S. Krinner et al., "Realizing repeated quantum error correction in a distance-three surface code," *Nature*, vol. 605, no. 7911, pp. 669–674, May 2022.

[29] L. Skoric, D. E. Browne, K. M. Barnes, N. I. Gillespie, and E. T. Campbell, "Parallel window decoding enables scalable fault tolerant quantum computation," *Nature Commun.*, vol. 14, no. 1, pp. 1–8, Nov. 2023.

[30] T. Haner, M. Roetteler, and K. M. Svore, "Factoring using $2n+2$ qubits with Toffoli based modular multiplication," *Quantum Inf. Comput.*, vol. 17, nos. 7–8, pp. 673–684, May 2017.

[31] R. W. J. Overwater, M. Babaie, and F. Sebastiano, "Neural-network decoders for quantum error correction using surface codes: A space exploration of the hardware cost-performance tradeoffs," *IEEE Trans. Quantum Eng.*, vol. 3, pp. 1–19, 2022.

[32] B. M. Terhal, "Quantum error correction for quantum memories," *Rev. Mod. Phys.*, vol. 87, no. 2, pp. 307–346, Apr. 2015.

[33] J. Bausch et al., "Learning to decode the surface code with a recurrent, transformer-based neural network," Google Deepmind, London, U.K., Tech. Rep., 2023. [Online]. Available: https://arxiv.org/abs/2310.05900v1

[34] Z. Chen et al., "Exponential suppression of bit or phase errors with cyclic error correction," *Nature*, vol. 595, pp. 383–387, Jul. 2021. [Online]. Available: https://www.nature.com/articles/s41586-021-03588-y

[35] C. M. Dawson and M. A. Nielsen, "The Solovay–Kitaev algorithm," 2005, *arXiv:quant-ph/0505030*.

[36] L. de Jong, J. I. Bas, J. Gong, F. Sebastiano, and M. Babaie, "A 10-Gb/s 275-fsec jitter cryo-CMOS charge-sampling CDR for quantum computing application," *IEEE Microw. Wireless Technol. Lett.*, vol. 33, no. 6, pp. 875–878, May 2023.

[37] A. Ganguly et al., "Interconnects for DNA, quantum, in-memory, and optical computing: Insights from a panel discussion," *IEEE Micro*, vol. 42, no. 3, pp. 40–49, May 2022.

[38] S. Gicev, L. C. L. Hollenberg, and M. Usman, "A scalable and fast artificial neural network syndrome decoder for surface codes," *Quantum*, vol. 7, p. 1058, Jul. 2023, doi: 10.22331/q-2023-07-12-1058.

[39] K.-H. Otto, *Clause 21*, document OIF-CEI-05.1, 2022. [Online]. Available: https://www.oiforum.com/wp-content/uploads/OIF-CEI-5.1.pdf

[40] D. S. Russell, "Technology advances for radio astronomy," Ph.D. dissertation, Dept. Elect. Eng., California Inst. Technol., Pasadena, CA, USA.

[41] B. Patra, M. Mehrpoo, A. Ruffino, F. Sebastiano, E. Charbon, and M. Babaie, "Characterization and analysis of on-chip microwave passive components at cryogenic temperatures," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 448–456, 2020.

[42] A. Nazemi et al., "A 36 Gb/s PAM4 transmitter using an 8 b 18 GS/S DAC in 28 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2015, pp. 58–59.

[43] M. Bassi, F. Radice, M. Bruccoleri, S. Erba, and A. Mazzanti, "A 45 Gb/s PAM-4 transmitter delivering 1.3 V_{ppd} output swing with 1 V supply in 28 nm CMOS FDSOI," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 66–67.

[44] C.-H. Lin and K. Bult, "A 10-b, 500-MSample/s CMOS DAC in 0.6 mm^2," *IEEE J. Solid-State Circuits*, vol. 33, no. 12, pp. 1948–1958, Dec. 1998.

[45] J. A. Croon, M. Rosmeulen, S. Decoutere, W. Sansen, and H. E. Maes, "An easy-to-use mismatch model for the MOS transistor," *IEEE J. Solid-State Circuits*, vol. 37, no. 8, pp. 1056–1064, Aug. 2002.

[46] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[47] J. van Dijk et al., "Cryo-CMOS for analog/mixed-signal circuits and systems," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Mar. 2020, pp. 1–8.

[48] M. Mehrpoo et al., "Benefits and challenges of designing cryogenic CMOS RF circuits for quantum computers," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.

[49] C.-H. Lin et al., "A 12 bit 2.9 GS/s DAC with IM3 $\ll -60$ dBc beyond 1 GHz in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 12, pp. 3285–3293, Dec. 2009.

[50] S.-M. Babamir and B. Razavi, "Relation between INL and ACPR of RF DACs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 10, pp. 3877–3887, Oct. 2022.

[51] A. van den Bosch, M. A. F. Borremans, M. S. J. Steyaert, and W. Sansen, "A 10-bit 1-GSample/s Nyquist current-steering CMOS D/A converter," *IEEE J. Solid-State Circuits*, vol. 36, no. 3, pp. 315–324, Mar. 2001.

[52] Y. Chen, J. Gong, R. B. Staszewski, and M. Babaie, "A fractional-N digitally intensive PLL achieving 428-fs jitter and <-54-dBc spurs under 50-mV_{pp} supply ripple," *IEEE J. Solid-State Circuits*, vol. 57, no. 6, pp. 1749–1764, Jun. 2022.

[53] J. Gong, E. Charbon, F. Sebastiano, and M. Babaie, "A cryo-CMOS PLL for quantum computing applications," *IEEE J. Solid-State Circuits*, vol. 58, no. 5, pp. 1362–1375, May 2023.

[54] B. Razavi, "TSPC logic [a circuit for all seasons]," *IEEE Solid State-Circuits Mag.*, vol. 8, no. 4, pp. 10–13, Fall 2016.

[55] R. Saligram, S. Datta, and A. Raychowdhury, "CryoMem: A 4K–300K 1.3 GHz eDRAM macro with hybrid 2T-gain-cell in a 28 nm logic process for cryogenic applications," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2021, pp. 1–2.

[56] R. A. Damsteegt, R. W. J. Overwater, M. Babaie, and F. Sebastiano, "A benchmark of cryo-CMOS 40-nm embedded SRAM/DRAMs for quantum computing," in *Proc. IEEE 49th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2023, pp. 165–168.

[57] R. Asanovski et al., "Understanding the excess 1/f noise in MOSFETs at cryogenic temperatures," *IEEE Trans. Electron Devices*, vol. 70, no. 4, pp. 2135–2141, Apr. 2023.

[58] B. Razavi, "Multiplexers and laser drivers," in *Design of Integrated Circuits for Optical Communications*. Hoboken, NJ, USA: Wiley, 2012, ch. 10, pp. 356–392.

[59] Y. Chang, A. Manian, L. Kong, and B. Razavi, "An 80-Gb/s 44-mW wireline PAM4 transmitter," *IEEE J. Solid-State Circuits*, vol. 53, no. 8, pp. 2214–2226, Aug. 2018.

[60] P. A. T. Hart, M. Babaie, A. Vladimirescu, and F. Sebastiano, "Characterization and modeling of self-heating in nanometer bulk-CMOS at cryogenic temperatures," *IEEE J. Electron Devices Soc.*, vol. 9, pp. 891–901, 2021.

[61] J. C. Thompson and B. A. Younglove, "Thermal conductivity of silicon at low temperatures," *J. Phys. Chem. Solids*, vol. 20, nos. 1–2, pp. 146–149, Jun. 1961.

[62] J. Gong, Y. Chen, E. Charbon, F. Sebastiano, and M. Babaie, "A cryo-CMOS oscillator with an automatic common-mode resonance calibration for quantum computing applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 12, pp. 4810–4822, Dec. 2022.

[63] I. Galton and C. Weltin-Wu, "Understanding phase error and jitter: Definitions, implications, simulations, and measurement," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 1, pp. 1–19, Jan. 2019.

[64] 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," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 19–26, Jan. 2020.

[65] E. Groen et al., "A 10-to-112 Gb/s DSP-DAC-based transmitter with 1.2 V_{ppd} output swing in 7 nm FinFET," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2020, pp. 120–121.

Niels Fakkel (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in microelectronics from the Delft University of Technology, Delft, The Netherlands, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree.

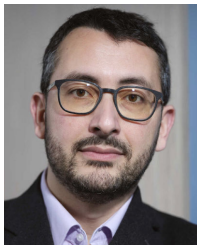
His current research interests include cryogenic electronics for quantum computing, RF transmitters for qubit control, and high-speed wireline systems.

Mohsen Mortazavi received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from the University of Tehran, Tehran, Iran, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the ELCA Research Group, Delft University of Technology, Delft, The Netherlands.

His research interests include energy-efficient transmitters at mm-wave frequencies for cellular network applications.

Ramon W. J. Overwater received the B.S. degree in electrical engineering, the M.S. degree (cum laude) in microelectronics, and the M.S. degree in computer engineering from the Delft University of Technology, Delft, The Netherlands, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree in cryogenic electrical engineering.

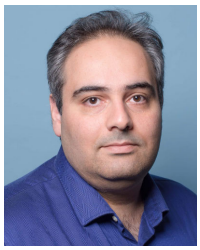
His research interests include cryogenic electronic characterization, mixed-signal design, and high-performance computing.



Fabio Sebastiano (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (cum laude) in electrical engineering from the University of Pisa, Pisa, Italy, in 2003 and 2005, respectively, the M.Sc. degree (cum laude) from the Sant'Anna School of Advanced Studies, Pisa, in 2006, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011.

From 2006 to 2013, he was with NXP Semiconductors Research, Eindhoven, The Netherlands, where he conducted research on fully integrated CMOS frequency references, nanometer temperature sensors, and area-efficient interfaces for magnetic sensors. In 2013, he joined the Delft University of Technology, where he is currently an Associate Professor. He has authored or coauthored one book, 11 patents, and over 100 technical publications. His main research interests are cryogenic electronics, quantum computing, sensor readouts, and fully integrated frequency references.

Dr. Sebastiano is on the Technical Program Committee of the ISSCC and the IEEE RFIC Symposium, and has been on the Program Committee of IMS. He was a co-recipient of several awards, including the 2008 ISCAS Best Student Paper Award, the 2017 DATE Best IP Award, the ISSCC 2020 Jan Van Vessem Award for Outstanding European Paper, and the 2022 IEEE CICC Best Paper Award. He has served as a Distinguished Lecturer for the IEEE Solid-State Circuit Society. He has served as a Guest Editor for JSSC. He is also serving as an Associate Editor for IEEE TRANSACTIONS ON VLSI and JSSC.



Masoud Babaie (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from the Amirkabir University of Technology, Tehran, Iran, in 2004, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2006, and the Ph.D. degree (cum laude) in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2016.

From 2006 to 2011, he was with the Kavoshcom Research and Development Group, Tehran, where he was involved in designing wireless communication systems. From 2014 to 2015, he was a Visiting Scholar Researcher with the Berkeley Wireless Research Center, Berkeley, CA, USA. In 2016, he joined the Delft University of Technology, where he is currently an Associate Professor. He has authored or coauthored one book, three book chapters, 11 patents, and more than 100 peer-reviewed technical articles. His research interests include RF/millimeter-wave integrated circuits and systems for wireless communications and cryogenic electronics for quantum computation.

Dr. Babaie was a co-recipient of the 2015–2016 IEEE Solid-State Circuits Society Pre-Doctoral Achievement Award, the 2019 IEEE ISSCC Demonstration Session Certificate of Recognition, the 2020 IEEE ISSCC Jan Van Vessem Award for Outstanding European Paper, the 2022 IEEE CICC Best Paper Award, and the 2023 IEEE IMS Best Student Paper Award (second place). He received the Veni Award from the Netherlands Organization for Scientific Research (NWO) in 2019. He is the Co-Chair of the Emerging Computing Devices and Circuits Subcommittee of the IEEE European Solid-State Circuits Conference (ESSCIRC) and the Technical Program Committee of the IEEE International Solid-State Circuits Conference (ISSCC). He is serving as an Associate Editor for the IEEE SOLID-STATE CIRCUITS LETTERS (SSC-L).