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

Fakkel, Niels; Mortazavi, Mohsen; Overwater, Ramon; Sebastiano, Fabio; Babaie, Masoud

DOI
10.1109/RFIC54547.2023.10186114

Publication date
2023

Document Version
Final published version

Published in
2023 IEEE Radio Frequency Integrated Circuits Symposium, RFIC 2023

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

Niels Fakkel[1], Mohsen Mortazavi, Ramon Overwater, Fabio Sebastiano, Masoud Babaie
Delft University of Technology, the Netherlands
[1]n.e.fakkel@tudelft.nl

***Abstract*— State-of-the-art quantum computers already comprise hundreds of cryogenic quantum bits (qubits), and prototypes with over 10k qubits are currently being developed. Such large-scale systems require local cryogenic electronics for qubit control and readout, leaving the digital controllers for algorithm execution and quantum error correction (QEC) at room temperature due to the limited cryogenic cooling budget. The entire process, including qubit readout, data transmission, QEC, and algorithm execution, should be completed well within the qubit decoherence time, thus requiring a low-power high-speed communication link between the cryogenic quantum processor and classical processor located at room temperature. To this end, this paper presents the first cryo-CMOS high-speed 4-level pulse amplitude modulation (PAM4) wireline transmitter. Thanks to a power-efficient serializing architecture driving a 6-bit digital-to-analog converter (DAC), the 40-nm CMOS chip achieves a data rate of 40 Gb/s PAM4 with an efficiency of 2.46 pJ/b and a ratio of level mismatch (RLM) of 97.8% at 4.2 K. While demonstrating an energy efficiency comparable to state-of-the-art transmitters in more advanced CMOS nodes, the extremely wide temperature operating range (4.2 K - 300 K) will enable future large-scale quantum computers.**

***Keywords*— Cryo-CMOS, quantum computing ICs, high-speed DAC, wireline transmitter.**

I. Introduction

Quantum computers can potentially solve problems intractable by classical computers, with applications ranging from cryptography and pharmaceuticals to artificial intelligence. Yet, the physical implementations of quantum bits (qubits) are too noisy to be used for robust computations. The robustness can be improved by realizing a logical qubit from multiple physical qubits and applying continuous rounds of quantum error correction (QEC). With current technologies, QEC requires reading out each qubit at about 1 Mb/s rate [1], and applying the corresponding corrections in real time. With roadmaps predicting the availability of systems with more than 10k qubits in only four years [2], the amount of data required for QEC grows accordingly. Electronics for qubit control and readout can be integrated within the cryogenic chamber of a quantum computer to tackle the scalability issues. However, the digital controllers for algorithm execution and QEC need to be placed at room temperature (RT) due to the restricted cryogenic cooling budget, as shown in Fig. 1. To prevent any backlog between the QEC and measurement cycles, a large amount of qubit readout data should be serialized and transmitted at >10 Gb/s rate. Hence, a need emerges for a high-speed wireline transmitter (TX) to transfer the data generated by the cryogenic readout circuits to a classical processor located at RT. Such a cryogenic TX must dissipate low power to comply with the cooling budget of a practical cryogenic refrigerator, while achieving the highest possible speed and linearity despite non-idealities in cryogenic CMOS device behavior, such as higher threshold voltage [3] and worse matching [4].

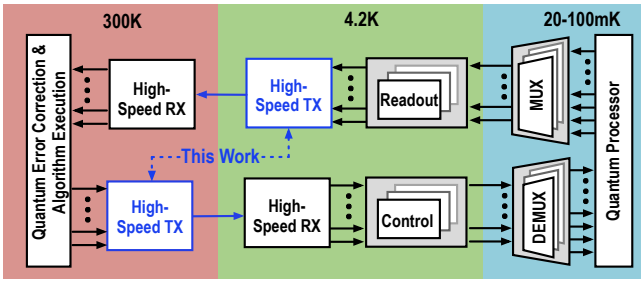


Fig. 1. Simplified block diagram of a scalable quantum computing system incorporating a cryo-CMOS high-speed wireline link.

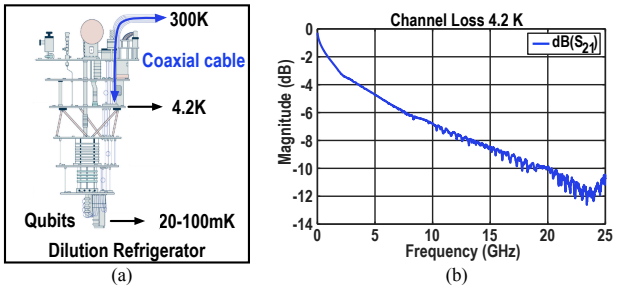


Fig. 2. (a) Dilution refrigerator illustration, and (b) the measured insertion loss (S_{21}) of a typical coaxial cable connecting the fridge 4.2 K stage to its output connector at room temperature.

This paper addresses the imminent need for a cryogenic wireline link in quantum computer applications and tackles the challenges associated with the extremely wide operating temperature range (4.2 K – 300 K). To the authors' best knowledge, this paper presents the first cryogenic CMOS (cryo-CMOS) wireline transmitter. By employing a 4-level pulse amplitude modulation (PAM4) protocol, the TX achieves a 40 Gb/s data rate and 2.46 pJ/b energy efficiency.

II. Wireline Transmitter Architecture

In a typical dilution refrigerator setup, shown in Fig. 2, a copper coaxial cable connects the 4.2 K stage to its output connector at RT. Considering a channel loss (i.e., S_{21}) of ~10 dB at 20 GHz and the limited TX output bandwidth due to the ESD and pad parasitic capacitance, the PAM4 modulation scheme is chosen over non-return-to-zero (NRZ)

2023 IEEE Radio Frequency Integrated Circuits Symposium (RFIC) | 979-8-3503-2122-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/RFIC54547.2023.10186114

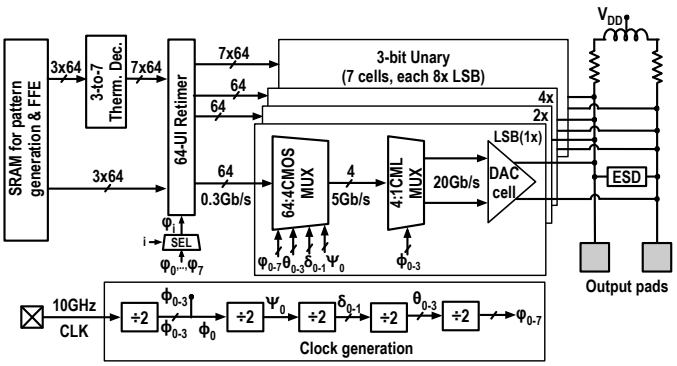


Fig. 3. Simplified block diagram of the proposed DAC-based wireline TX.

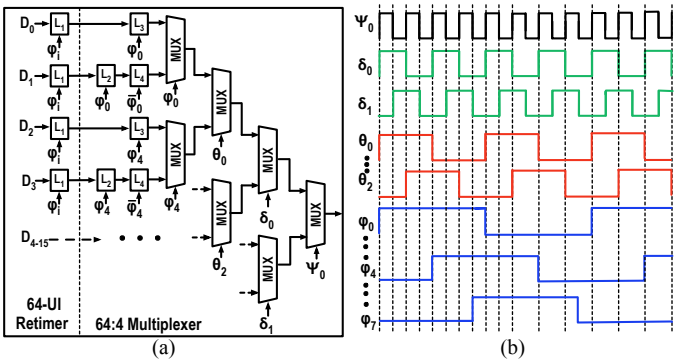


Fig. 4. (a) Schematic of 64:4 MUX, and (b) a timing diagram of its required clock phases to remove latches between the selectors.

to avoid extensive loss at higher frequencies and enhance the throughput. A digital-to-analog converter (DAC)-based transmitter architecture is employed since it can support different modulation formats (e.g., NRZ, PAM4), overcome channel imperfections, and transfer data without losing signal integrity. Furthermore, a current-mode logic (CML)-based DAC is adopted instead of a voltage-based output stage, due to its higher bandwidth, better supply rejection, and lower sensitivity to variations of the transistor characteristic at 4.2 K.

The block diagram of the TX architecture is shown in Fig. 3. A programmable 512×64×6 bit SRAM is implemented to allow for the exploration of different test sequences, data formats, and equalization techniques in measurement. The 64-Unit-Interval (UI)×6 bit parallel SRAM data is decoded and fed into the 10 DAC slices (3b thermometer + 3b binary coded). In each slice, a 64:4 multiplexer (MUX) structure serializes the data to a high-speed 4UI×10 slice signal. This signal is then retimed by a selectable clock phase to complementary 25% pulses and fed into a 4:1 CML-based MUX. A dedicated CML output driver after the 4:1 MUX is employed to improve TX linearity and bandwidth compared to the case where the 4:1 MUX also acts as the output stage. The output network is designed with two differential 50 Ω termination resistors, and a center-tapped peaking inductor to compensate for the bandwidth reduction due to the parasitic capacitance of the ESD protection and output pads. A clock generation circuit converts an external clock input into all the divided clock phases necessary for the retimers and multiplexers.

A. 64:4 Multiplexer

Fig. 4 (a) shows the schematic of the 64:4 MUX. All incoming bits from the SRAM are retimed in L_1 latches by one selectable clock phase (i.e., $\varphi_i \in \{\varphi_0, \ldots, \varphi_7\}$) that is digitally calibrated to provide optimal setup- and hold-time margin. To avoid narrow pulses or glitches at the outputs of the first selectors, the time difference between the data transitions at the inputs of each first-rank selector is adjusted to 32UI, using different phases (i.e., φ_{0-7}) of the 64UI clock in L_{2-4} latches. However, this approach is not power-efficient at higher-rank selectors, as it would need 140 more latches operating at higher frequencies. As depicted in Fig. 4, those latches can be removed if lower-rank selectors use the quadrature clocks generated by dividing the clock frequency of their corresponding higher-rank selector [5].

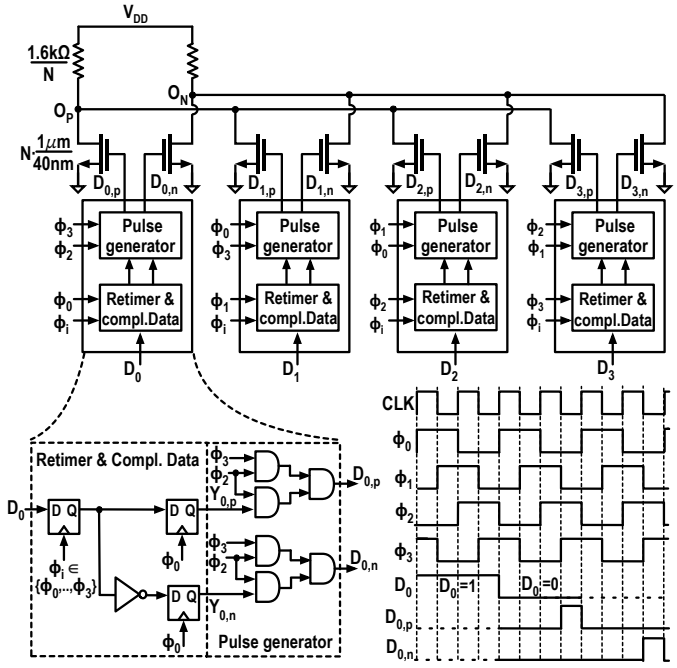
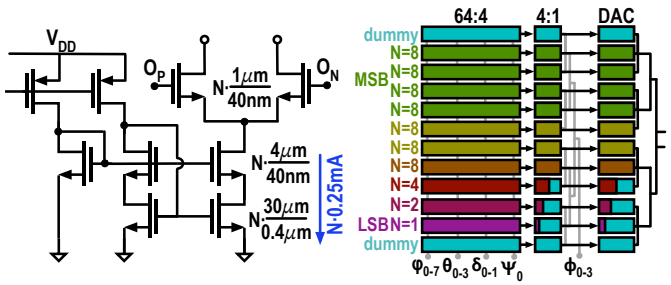


Fig. 5. Implementation of proposed quarter-rate retimer, pulse generator, and 4:1 CML MUX with the corresponding timing diagram.

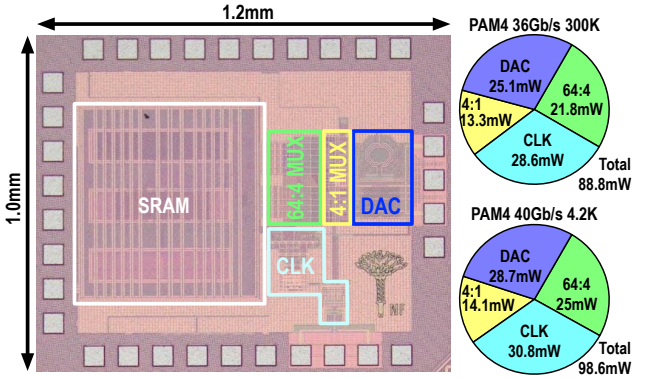
B. 4:1 Multiplexer

In the quarter-rate retimer and pulse generator, the data is prepared for the 4:1 CML MUX, as shown in Fig. 5. Due to the different clock routing schemes, there is an unknown skew between the 8UI clock domain (i.e., ϕ_0) used in the last stage of 64:4 MUX and 4UI clock domain (i.e., ϕ_{0-3}) employed in the pulse generator. Hence, a retimer is added between these blocks whose clock phase (i.e., $\phi_i \in \{\phi_0, \ldots, \phi_3\}$) is selectable using a multiplexer-based phase rotator. The data is converted to complementary form and again retimed to ensure any delay mismatch from the inverter is compensated. The required 25% duty-cycle pulse (i.e., 1UI) is generated by combining the corresponding 50% overlapping 4UI clock phases and complementary data using two cascaded AND gates. Note that using a single 3-input AND gate would result



(a) (b)

Fig. 6. (a) Schematic of LSB DAC cell, and (b) DAC layout floor plan.



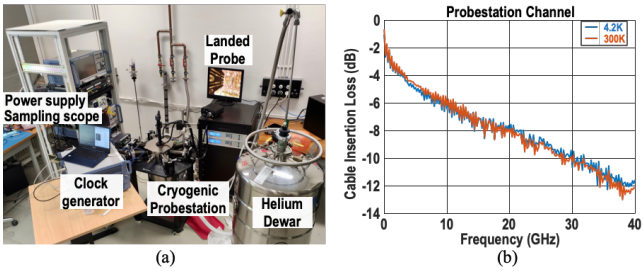
(a) (b)

Fig. 7. (a) Chip micrograph, and (b) measured power breakdown at 300 K and 4.2 K.

in a much slower rise/fall time at cryogenic temperatures (CT) due to the threshold voltage increase and, therefore, limited voltage overdrive. The pulse generation is done locally, because distributing 25% non-overlapping clock pulses as an alternative would require more power-hungry buffers in the clock path and set tighter skew constraints to the clock distribution. The differential phases of the quadrature clocks use an upper-level metal and are distributed through H-trees to reduce the power consumption of the clock buffer and clock delay mismatches.

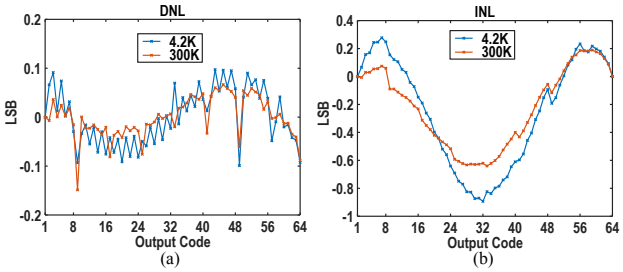
The TX maximum speed is determined by the clock-to-q delay and setup time of the high-speed flip-flops as they are in the critical path. Among different flip-flop architectures, a True Single Phase Clock (TSPC) dynamic flip-flop that employs a lower number of stacked devices is adopted, thus maximizing speed by fully exploiting transistors' higher mobility at CT.

The CML-based 4:1 MUX is designed without a tail transistor to counteract the reduced voltage headroom due to the higher threshold at CT and allow for smaller size switches. It combines the interleaved pulses from the quarter-rate retimer to drive the final output stage. Note that the MSB cells have a larger input capacitance than the LSB cells in the DAC structure. Consequently, the load resistance and the transistor size of the corresponding 4:1 MUX as the pre-driver are proportionally scaled to maintain the bandwidth and input voltage amplitude of all DAC slices, thus preventing any systematic delay mismatch.



(a) (b)

Fig. 8. (a) Cryogenic probe station measurement setup, and (b) the measured insertion loss of the probe and cable, realizing the channel between the chip and measurement instrument.



(a) (b)

Fig. 9. (a) DNL, and (b) INL, measured at 4.2 K and 300 K.

C. DAC

This design employs a 6-bit CML DAC to satisfy the ratio of level mismatch (RLM) of at least 95% defined in wireline standards with sufficient margin. The tail current sources of each DAC cell (Fig. 6 (a)) are sized with a margin to reach the noise and differential non-linearity (DNL) requirements, accounting for the 20% increased device mismatch at CT [4]. To optimize the tradeoff between the integral non-linearity (INL) requirement and the total power consumption due to the number of serializer paths, a 3b binary + 3b unary coded structure is employed. As shown in Fig. 6 (b), the DAC slices, including serializers and retimers, are laid out symmetrically, and dummy rows are added at the top and bottom to minimize mismatch.

III. Results

A TX prototype was fabricated in 40-nm CMOS. As shown in the chip micrograph in Fig. 7 (a), the active area of the transmitter, excluding SRAM, is 0.146 mm^2. The TX performance was measured with a cryogenic probe station to lower the prototype's ambient temperature to 4.2 K. The measurement setup shown in Fig. 8 introduces a total loss of ~8 dB at 20 GHz due to probes and cables. Fig. 9 shows that the maximum INL is below 0.6 LSB at RT and, as expected, increases to 0.9 LSB at CT due to increased device mismatch. Fig. 10 displays the measured eye diagrams at RT and CT without using additional feedforward equalization, de-embedding, or scope equalization. To perform the measurement, the on-chip SRAM is loaded with a 2^{15}-length pseudorandom binary sequence (PRBS)-15 for NRZ and a quaternary QPRBS-15 for PAM4. At RT, 20 Gb/s NRZ and 20&36 Gb/s PAM4 signals are measured. At the highest rate, the measured eye heights (widths) of NRZ and PAM4 are 231 mV (0.65UI) and >24.7 mV (0.28UI), respectively, with 96.5% RLM. Due

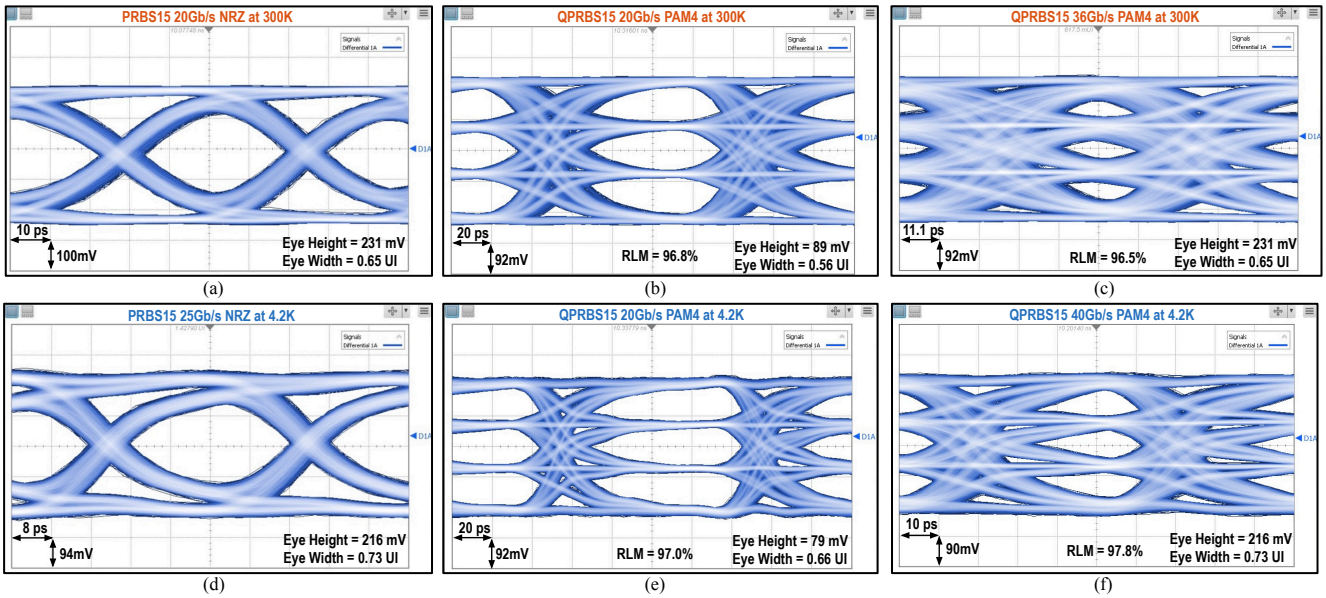


Fig. 10. Measured eye diagrams: (a) 20 Gb/s NRZ at 300 K, (b) 20 Gb/s PAM4 at 300 K, (c) 36Gb/s PAM4 at 300 K, (d) 25 Gb/s NRZ at 4.2 K, (e) 20 Gb/s PAM4 at 4.2 K, and (f) 40 Gb/s PAM4 at 4.2 K.

Table 1. Comparison table with prior DAC-based and multi-tap wireline transmitters.

	This work		[6] ISSCC'17	[7] ISSCC'18	[8] 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	98.8	-	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

to the digital speed improvement at CT, the baud rate could be increased, and consequently, 25 Gb/s NRZ and 20&40 Gb/s PAM4 are measured. At the highest rate, the measured eye heights (widths) of NRZ and PAM4 are 216 mV (0.73UI) and >38.5 mV (0.47UI), respectively, with 97.8% RLM. The power breakdown chart is shown in Fig. 7 (b). At RT (CT), the TX can achieve a maximum configurable swing of 0.8 Vpp and consumes 88.8 mW (98.6 mW) at 36 Gb/s (40 Gb/s), thus achieving 2.47 pJ/b (2.46 pJ/b) energy efficiency. Measurements above those baud rates violate the timing constraints in the retimers before the 4:1 MUX, and therefore result in incorrect behavior.

Table 1 benchmarks the performance of this work with prior art. At RT, even by using a less advanced technology node, the proposed transmitter can reach similar energy efficiency as state-of-the-art TXs while achieving the highest RLM. Moreover, by demonstrating, for the first time, both full functionality and high efficiency over the wide temperature range down to CT, the proposed TX enables the required high-speed wireline link for quantum computing applications.

IV. Conclusion

This paper demonstrates the first cryogenic wireline transmitter. At CT (RT), the prototype achieves 40 Gb/s (36 Gb/s) PAM-4 transmission with 2.46 pJ/b (2.47 pJ/b) efficiency and 97.8%(96.5%) RLM. By circumventing the disadvantages of cryo-CMOS devices (higher threshold, larger mismatch) and exploiting their higher speed in the design of the serializer and DAC, the transmitter maintains high power efficiency, linearity, and data rate down to CT. This result enables high-speed data communication between classical and quantum processors, which is essential in the scale-up of future quantum computers.

References

[1] R. W. J. Overwater *et al.*, "Neural-network decoders for quantum error correction using surface codes: A space exploration of the hardware cost-performance tradeoffs," *IEEE Transactions on Quantum Engineering*, vol. 3, pp. 1–19, 2022.

[2] J. Gambetta. (2022, May) Expanding the IBM Quantum roadmap to anticipate the future of quantum-centric supercomputing. [Online]. Available: https://research.ibm.com/blog/ibm-quantum-roadmap-2025

[3] A. Beckers *et al.*, "Physical model of low-temperature to cryogenic threshold voltage in MOSFETs," *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 780–788, 2020.

[4] P. A. T. Hart *et al.*, "Subthreshold mismatch in nanometer CMOS at cryogenic temperatures," *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 797–806, 2020.

[5] Y. Chang *et al.*, "An 80-Gb/s 44-mW wireline PAM4 transmitter," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 8, pp. 2214–2226, 2018.

[6] A. Nazemi *et al.*, "A 36Gb/s PAM4 transmitter using an 8b 18GS/S DAC in 28nm CMOS," in *IEEE International Solid-State Circuits Conference*, San Francisco, USA, Feb. 2015, pp. 58–59.

[7] M. Bassi *et al.*, "A 45Gb/s PAM-4 transmitter delivering 1.3Vppd output swing with 1V supply in 28nm CMOS FDSOI," in *IEEE International Solid-State Circuits Conference*, San Francisco, USA, Feb. 2016, pp. 66–67.

[8] E. Groen *et al.*, "A 10-to-112Gb/s DSP-DAC-Based transmitter with 1.2Vppd output swing in 7nm FinFET," in *IEEE International Solid-State Circuits Conference*, San Francisco, USA, Feb. 2020, pp. 120–121.