A Benchmark of Cryo-CMOS Embedded SRAM/DRAMs in 40-nm CMOS

Damsteegt, Rob A.; Overwater, Ramon W.J.; Babaie, Masoud; Sebastiano, Fabio

DOI
10.1109/JSSC.2024.3385696

Publication date
2024

Document Version
Accepted author manuscript

Published in
IEEE Journal of Solid-State Circuits

Citation (APA)
Damsteegt, R. A., Overwater, R. W. J., Babaie, M., & Sebastiano, F. (2024). A Benchmark of Cryo-CMOS Embedded SRAM/DRAMs in 40-nm CMOS. *IEEE Journal of Solid-State Circuits*, *59*(7), 2042-2054. https://doi.org/10.1109/JSSC.2024.3385696

A Benchmark of Cryo-CMOS Embedded SRAM/DRAMs in 40-nm CMOS

Rob A. Damsteegt, Ramon W. J. Overwater, Masoud Babaie, *Senior Member, IEEE*, and Fabio Sebastiano, *Senior Member, IEEE*

***Abstract*— The interface electronics needed for quantum processors require cryogenic CMOS (cryo-CMOS) embedded digital memories covering a wide range of specifications. To identify the optimum architecture for each specific application, this article presents a benchmark from room temperature (RT) down to 4.2 K of custom SRAMs/DRAMs in the same 40-nm CMOS process. To deal with the significant variations in device parameters at cryogenic temperatures, such as the increased threshold voltage, lower subthreshold leakage, and increased variability, the feasibility of different memories at cryogenic temperature is assessed and specific guidelines for cryogenic memory design are drafted. Unlike at RT, the 2T low-threshold-voltage (LVT) DRAM at 4.2 K is up to 2× more power efficient than both SRAMs for any access rate above 75 kHz since the lower leakage increases the retention time by 40 000×, thus sharply cutting on the refresh power and showing the potential of cryo-CMOS DRAMs in cryogenic applications.**

***Index Terms*— Cryogenic CMOS (cryo-CMOS), DRAM, eDRAM, memory, quantum computing, SRAM.**

I. Introduction

QUANTUM computers (QCs) can deliver an exponential speedup for several computational problems [1], [2], [3], [4], [5], [6]. However, scaling up the number of quantum bits (qubits) to the thousands or millions necessary for useful computations requires an impractical amount of wires connecting the cryogenic qubits to the room-temperature (RT) control electronics. To overcome such an interconnect bottleneck, electronics integrated in commercial CMOS technology but operating at cryogenic temperature, i.e., cryogenic CMOS (cryo-CMOS), has been proposed [7], [8]. As the power consumption of the cryo-CMOS control electronics must be kept below the cooling power of the cryogenic refrigerators adopted in QC applications, designing power-efficient cryo-CMOS circuits is crucial.

The control electronics consist of analog/RF circuits directly interfacing with the qubits to perform operations and measurements, in combination with the digital system-on-chip (SoC) for scheduling the quantum-algorithm execution [9] and processing a large amount of measurement results, e.g., as required for quantum error correction [10], [11], [12], [13], [14]. In modern digital systems, significant fractions of the area and power are consumed by the memory, thus making the optimization of cryo-CMOS embedded memories essential. However, accurately estimating the power consumption of a memory at cryogenic temperatures is challenging due to the lack of reliable cryogenic device models.

Furthermore, the cryo-CMOS controllers will require memories for several distinct functions covering a wide range of access rates (read and write operations per second) and write/read (W/R) ratios, ranging from high-speed lookup tables for generating the waveforms for qubit control (multi-GHz, $W/R = 0$) [15], [16], [17] to low-speed buffer queues for the quantum-algorithm instructions (sub-MHz, $W/R = 1$) [9]. Static memories (SRAMs) are well-suited for high access-rate applications but they suffer from excessive operation energy and limited density. The density issue can be alleviated by dynamic memories (DRAMs), which store data as the charge on a (parasitic) capacitor and require fewer transistors per cell. Unfortunately, frequent refreshes are required to counteract charge leakage, resulting in a large power consumption independent of the access rate. While the charge leakage is strongly mitigated by the significant decrease in subthreshold leakage at cryogenic temperatures [18], [19], it is unclear whether a cryo-CMOS DRAM can outperform a cryo-CMOS SRAM, due to both the shortcomings of existing device models and the absence of comprehensive studies in the literature.

Cryogenic memories have been actively explored for (superconducting) high-performance computing [20], [21], [22], [23], [24], [25] and, more recently, for QC applications. From both perspectives, commercial DRAM memories have been investigated down to 77 K [26], [27], [28], [29], [30], [31]. However, for deep-cryogenic (4.2 K) applications, such as superconducting computing [32], [33], [34], [35], [36], [37], [38], [39] and QC, custom embedded memories have been investigated, including static-cell designs [10], [35], [36], [39], [40], [41], [42] and dynamic-cell designs [32], [43], [44], [45], [46], [47], [48], [49], [50]. Additionally, less well-known cell designs in specific technologies have also been investigated around both 77 K [51], [52], [53] and 4.2 K [54]. Analyses based only on simulations have been attempted but do not capture the full range of cryogenic effects, e.g., not properly

Manuscript received 8 December 2023; revised 19 March 2024; accepted 25 March 2024. This article was approved by Associate Editor Filip Tavernier. This work was supported by Intel Corporation. *(Corresponding author: Rob A. Damsteegt.)*

Rob A. Damsteegt, 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: r.a.damsteegt@tudelft.nl).

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.2024.3385696.

Digital Object Identifier 10.1109/JSSC.2024.3385696

modeling leakage and dynamic effects [10], [41], [42], [49], [50]. Still, the large variety in memory architectures, adopted CMOS processes, and temperature ranges in those prior works hinders the compilation of a fair comparison. Thus, identifying the best memory design in terms of area and power for each memory application is still a challenging open question.

To overcome this issue, this work compares eight different dynamic and static memory cell designs, embedded in identical memory architectures in a nanometer CMOS process (TSMC 40-nm) typically adopted for QC cryo-CMOS interfaces, by comparing the experimental characterization at both RT and 4.2 K. Due to the limited cooling power available in dilution refrigerators, the main focus is on minimizing the memory power consumption. Since the power consumption of the dynamic memories is limited by their refresh power for medium-to-high frequency applications, a detailed characterization of the data-retention time is required for these cells. This article, an extension of our work in [55], is structured as follows. Section II offers a brief overview of the cryogenic effects in CMOS devices. Section III describes the circuit designs of the adopted memories, for which the experimental characterization is presented in Section IV. Section V discusses the results and Section VI concludes this article. The data shown in this article are also available here [56].

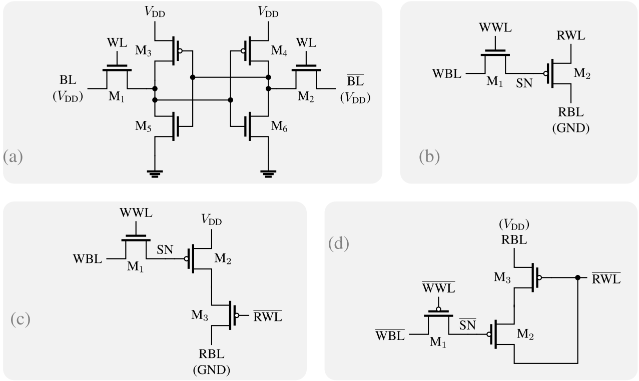
II. Cryo-CMOS Device Behavior

Cooling down to cryogenic temperatures affects the characteristics of short-channel NMOS and PMOS transistors by increasing their threshold voltage V_{th} (100–200 mV), subthreshold slope (∼3× steeper), and carrier mobility (∼2× for low-field mobility) [18], [19], [32], [57], [58], [59], [60], [61]. Additionally, the mismatch between devices increases, as shown in [62] and [63] for 40-nm bulk CMOS and 28-nm bulk CMOS, respectively, interconnect resistance drops (∼30%) [64], and the capacitance of source/drain junctions decreases due to wider depletion regions due to freeze-out [19]. For analog circuits, this results in an increased bandwidth and reduced power consumption.

For full-swing digital circuits, the mobility increase compensates the effects of the larger V_{th} and, together with the reduced resistance and capacitance, results in a speed-up for digital circuits from 10% to 20% for 40-nm bulk CMOS [65], [66], [67]. For more advanced technology nodes, the speed-up from RT to 4.2 K is reduced due to the increased relative importance of interconnect capacitance and lower supplies, enhancing the relative V_{th} increase [65]. However, the speed-up could be recovered for FinFET technologies by scaling V_{th} [40]. The increased V_{th} and the steeper subthreshold slope lead to severely reduced subthreshold leakage, while gate leakage stays approximately constant (<2× smaller) [68]. For these digital circuits, this will result in greatly reduced leakage power, while keeping the dynamic power consumption similar.

III. Circuit Design

The memory cells in this work have been mainly optimized for maximum density, and, where possible, for



(R)BL followed by (precharge voltage)
(R/W)WL: active high, $\overline{\text{(R/W)WL}}$: active low
(W)BL/SN: high when data = 1, $\overline{\text{(W)BL/SN}}$: low when data = 1
GC RBLs: low after reading data = 1, high after reading data = 0

Fig. 1. Schematics of the four cell designs: (a) 6T static cell; (b) 2T NW-PR dynamic gain cell; (c) 3T NW-PR dynamic gain cell; and (d) 3T PW-PR preferentially boosted dynamic gain cell. The readout current of the dynamic cells always flows from top to bottom.

optimum (expected) performance at cryogenic temperature. All memory cells are implemented in two versions, using either standard-threshold-voltage (SVT) or low-threshold-voltage (LVT) devices. LVT cells are expected to perform worse at RT since their higher subthreshold leakage reduces the retention time of dynamic memories and increases the static power consumption. At cryogenic temperatures, however, the V_{th} increase may cause SVT designs to fail due to the insufficient overdrive voltage limiting the readout currents. Although forward-biasing the bulk–source voltage [66] could help circumvent the cryogenic V_{th} increase, no individual bulk contacts have been employed to avoid an excessive increase in the design effort and the area of the memory cells. The memory peripherals always use LVT devices, unless otherwise noted, to ensure functionality at cryogenic temperatures and minimize their effect on memory performance, while the synthesized digital circuits, e.g., the controllers, adopt SVT devices with extra hold margin to anticipate the cryogenic logic speed-up.

A. 6T Static Cell

As the most commonly used embedded-memory cell, the conventional six-transistor static cell [6T, Fig. 1(a)] represents a good reference for comparison with alternative designs. It consists of a latch formed by two inverters ($M_{3\text{–}6}$) and two access transistors ($M_{1,2}$) that connect the latch nodes to the differential bitlines (BLs) (BL and $\overline{\text{BL}}$). The latch state is written by differentially driving the BLs and pulling the wordline (WL) high. To read the state, both BLs are first precharged to V_{DD} before enabling the WL. Then, the BL connected to the low side of the latch will be discharged by one of the pull-down transistors ($M_{5,6}$).

To minimize the cell area, most transistors have minimum size ($W/L = 120$ nm/40 nm). Since the cell design is ratioed, the pull-down transistors ($M_{5,6}$) are sized 1.5× larger ($W/L =$ 180 nm/40 nm) to ensure writing and reading under device mismatch. For a fair comparison with the other cells, the static cell is manually implemented using the logic design rule check (DRC) rule set and occupies 0.435 μm^2 using a lithographically symmetrical layout [69]. This is 80% larger than the foundry-offered cells (0.242 μm^2 [70]) that violate several logic DRC rules.

B. 2T NW-PR Dynamic Cell

A higher density can be reached by dynamic memory cells, as they require fewer transistors. Since the popular one-transistor-one-capacitor (1T1C) dynamic cell [71] is only advantageous with a high-density-capacitor technology option [72], gain-cell dynamic memories are preferred here to achieve low area in standard CMOS. In the simplest gain cell with two transistors [2T, Fig. 1(b)] [46], the data, stored as charge on the storage node (SN), are written from the write bitline (WBL) through a write pass-transistor (M_1) when the write wordline (WWL) is enabled. For reading, the read bitline (RBL) is precharged to ground and charged by the readout current of M_2 when the read wordline (RWL) is enabled, depending on the voltage of the SN. The output data are obtained by comparing the RBL voltage to a reference.

We could use common device types to implement both transistors, allowing for a high cell density due to the lack of N-well transitions. However, different device types are preferred for the following reasons. To keep the design simple and reliable, all voltages are kept within the supply rails. This means that WL boosting, i.e., pulling the WWL beyond the supply rails to counter the V_{th} drop across M_1, cannot be used. The resulting V_{th} drop will limit the voltage range on SN, reduce M_2's overdrive, and, therefore, limit the readout current. This will be worse at cryogenic temperatures due to the V_{th} increase. The charge on SN leaks away through M_1's subthreshold leakage and M_2's gate leakage. Since the gate leakage is expected to dominate at cryogenic temperatures and the PMOS gate leakage is smaller in the target technology (according to the RT model), an NMOS is used for writing (NW) and a PMOS for reading (PR).

Although a wider M_1 would speed up the writing, its width is minimized to reduce the area and the subthreshold leakage since the minimum-size write speed is still very high (10–100 ps). For M_2, a larger width asks for more area but also increases the SN capacitance and the readout current, and therefore the retention time. At −40 °C, i.e., the lowest valid temperature for the standard models, $W = 300$ nm results in a good tradeoff between area and retention time by minimizing the area-refresh-power product for a fixed read duration of 1 ns and a fixed margin (>300 mV) between the RBL voltage levels for the different stored bits. The resulting cell area is 0.184 μm^2 (58% smaller than the custom 6T cell, and 24% smaller than the foundry 6T cell).

Unfortunately, the retention time and readout speed of the 2T cell are limited due to capacitive coupling between the RWL and the SN. Due to the M_2 gate–source coupling, the SN voltage is pulled up at the start of a read operation. While this ensures M_2 to be off for high SN voltages, it limits M_2's overdrive for low SN voltages, thus limiting the readout current and increasing the read time. Since such a gate–source coupling is stronger when M_2 is in inversion, the increase in SN voltage will be larger for lower SN voltages. As a result, the SN voltage levels for the two states move closer during readout, making them harder to distinguish. Additionally, the RBL voltage is limited by the other cells connected to the same RBL and with low SN voltages, as they will also conduct when the RBL voltage approaches V_{th} of the readout transistors. Although this effect is mitigated by the cryogenic V_{th} increase, the RBL voltage swing is usually kept well below this limit by limiting the duration of the RWL pulse, so as to minimize the read energy and stay within the functional range of the sense amplifiers.

C. 3T NW-PR Dynamic Cell

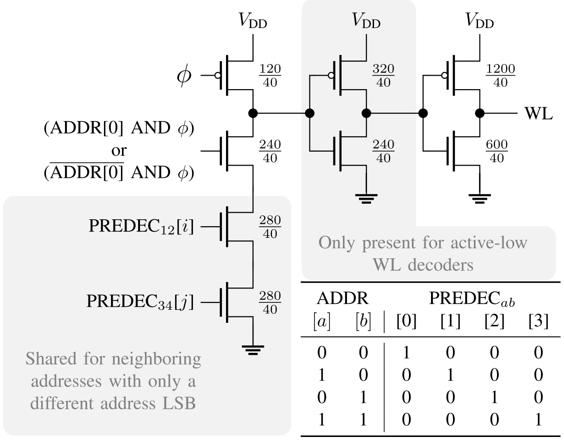
A three-transistor cell [73] [3T, Fig. 1(c)] circumvents the readout limitations of the 2T cell. The source of the readout transistor (M_2) is connected to a fixed voltage (V_{DD}) and a read pass-transistor (M_3) is added to select the row. This results in a faster readout due to larger M_2 overdrive, no shrinking of the SN voltage margin during readout, and no leakage through the readout transistors of other cells when RBL gets charged higher.

The 3T-cell sizing follows the principles adopted for the 2T cell for $M_{1,2}$, resulting in the same sizes for these transistors. Within the layout, with the sizes of $M_{1,2}$ now fixed, M_3 is sized as wide as possible ($W = 190$ nm) to minimize its ON-resistance without significantly increasing the area, which is 0.242 μm^2 (only 32% larger than the 2T cell and equal to the foundry 6T cell).

The largest SN voltage that can be written is $V_{\text{DD}} - V_{\text{th,n}}$, while M_2's gate voltage must be larger than $V_{\text{DD}} - |V_{\text{th,p}}|$ to turn M_2 off. To ensure M_2 to be off, $M_{2,3}$ are implemented as transistors with higher threshold voltages (SVT for the LVT cell version, and high threshold voltage (HVT) for the SVT cell version). Since the V_{th} increase is larger for PMOS than for NMOS at 4.2 K, the margin $|V_{\text{th,p}}| - V_{\text{th,n}}$ will be larger, making it easier to turn M_2 off. However, this will also lead to a reduced overdrive and slightly lower readout currents.

D. 3T PW-PR Dynamic Cell

Instead of avoiding the SN-RWL coupling, it can be exploited to increase the SN voltage margin during readout by using preferential boosting [74]. In such a cell [Fig. 1(d)], the RWL is connected to both the gate of the readout pass-transistor (M_3) and the drain of the readout transistor (M_2). Since the RWL is now pulled down, the RBL will be discharged from V_{DD} through the PMOS stack. As the RWL pull-down coupling to the SN is larger for low SN voltages, the SN voltage margin between the two logic levels now improves due to the coupling. Since the SN voltage is now pulled down by the preferential boosting, the write transistor (M_1) has to be a PMOS (PW) to ensure that a high enough SN voltage can be written to turn off M_2. Consequently, the SN voltage cannot be set lower than the $|V_{\text{th}}|$ of M_1. The overdrive of M_2 is then significantly reduced due to the cryogenic V_{th} increase for both M_1 and M_2, pointing to a high chance of failure that



ADDR [a]	ADDR [b]	PREDEC$_{ab}$ [0]	[1]	[2]	[3]
0	0	1	0	0	0
1	0	0	1	0	0
0	1	0	0	1	0
1	1	0	0	0	1

Fig. 2. Row decoder schematic, including transistor (W/L) in nm and the 2-bit predecoder truth table. The optional inverter in gray is only included for the row decoders where the active WL is low, i.e., overlined WWL and RWLs in Fig. 1.

should be experimentally studied to assess the feasibility of the cell design at 4.2 K.

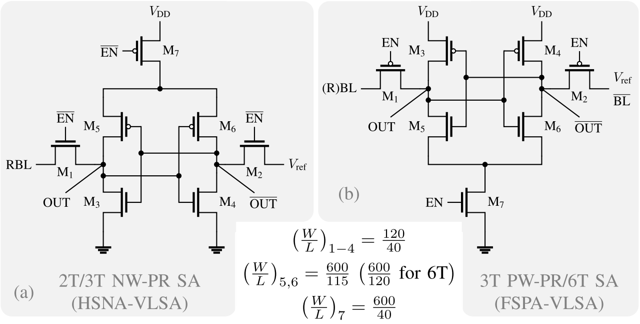
With the same sizing as for the 3T NW-PR cell, the area is 0.254 μm^2, 38% larger than the 2T cell and slightly larger than the 3T NW-PR cell since the RWL connection of M_2 cannot be shared with neighboring cells as effectively as for the 3T NW-PR.

E. Memory Peripherals

To focus on the differences in performance due to different cell designs, the simplest memory architecture is adopted with a single bank with 1024 cells (32 rows and 32 columns) without peripheral sharing. The peripherals are nearly identical among different memories, with only small adaptations for different cell pitch and signal polarities, to minimize their effect on performance.

1) Row Decoders: Row decoders decode the 5-bit address (0–31) into a one-hot signal on one of the 32 WLs. The dynamic memories have two decoders, one for the RWLs and one for the WWLs. For low-latency and regular-layout design, the dynamic decoder in Fig. 2 is adopted with two 2-bit predecoders for the address' four most-significant bits (MSBs). The lower two NMOS transistors in the pull-down stack (left gray block) are shared between neighboring addresses, as they only have a different LSB. A large output inverter is used to minimize the WL rise/fall time and, only for 2T NW-PR and 3T PW-PR, to supply the readout current without excessive voltage drop.

2) Sense Amplifiers: The voltage-latched sense amplifiers (VLSAs) shown in Fig. 3 [75] determine whether the RBL voltage at the end of the readout phase is above or below an external reference voltage. When $M_{1,2}$ sample the reference voltage and the RBL voltage, the power-gated latch formed by two inverters is turned off. The latch is then disconnected from the inputs and turned on to amplify the input difference. The VLSA is sized to fit within the pitch of a single memory cell, and for offset and noise not to limit the cell performance.



$$\left(\frac{W}{L}\right)_{1-4} = \frac{120}{40}$$
$$\left(\frac{W}{L}\right)_{5,6} = \frac{600}{115}\ \left(\frac{600}{120}\text{ for 6T}\right)$$
$$\left(\frac{W}{L}\right)_{7} = \frac{600}{40}$$

Fig. 3. Schematics of the two SA designs with sizing in nm: (a) HSNA-VLSA for low RBL voltages and (b) FSPA-VLSA for high RBL voltages.

Two variations of the VLSA are used. Due to $M_{3,4}$ starting to conduct during sampling, the headswitch NMOS access (HSNA) VLSA [Fig. 3(a)] is used for the 2T and 3T NW-PR cells as it works well for inputs below $V_{\text{th,n}}$, while the footswitch PMOS access (FSPA) VLSA is used for the 3T PW-PR and static cells as it works well for inputs above $V_{\text{DD}} - |V_{\text{th,p}}|$ [75]. In both designs, $M_{5,6}$ perform the comparison and dominate the input-referred offset. Thus, they are sized larger to lower the offset and laid out in a regular grid with surrounding dummies to improve the matching. Transistor M_7 is also wider to supply sufficient current to the latches and not to limit the decision speed. All other transistors ($M_{1–4}$) are minimum-sized and $M_{3,4}$ are implemented using HVT devices to increase the functional range of the SAs. In this case, the high V_{th} is not a problem since $M_{3,4}$ must only ensure the (dis)charge to the supply rails.

The input-referred offset standard deviation of the HSNA-VLSA is expected to be around 12.6 mV based on RT Monte Carlo simulations. This is significantly less than the expected RBL voltage variation due to cell mismatch (in the order of $\sigma = 50$ mV). The input-referred rms noise is expected to be around 3.5 mV with a decision time of around 200–250 ps at RT. At 4.2 K, the mismatch is expected to increase roughly 10%–15% [62] while the rms noise is expected to decrease by at least 50% [76].

For the dynamic memories, the reference input of the SA is always connected to an external reference voltage pad. A minimum-sized NMOS/PMOS pass-transistor (the same type as the access transistor) is added to the BLs so they can be connected to a second external pad. This allows for the characterization of the offset and noise of the SAs by controlling both input voltages. The SAs are followed by transmission-gate-based latches implemented with minimum-sized devices. During a read operation, these prevent glitching at the output and isolate the SAs to prevent interference.

3) BL Driver: In the BL driver for the dynamic memories [Fig. 4(a)], a multiplexer selects the external data input (D_{IN}) or the data from the last read operation for a refresh (D_{REF}). The BL driver for the static memories [Fig. 4(b)] implements a different functionality: when idle (W and R low), the BL is pulled up and precharged to V_{DD}; when reading (R high), the BL is left floating to be discharged by the cell being read; when writing (W high), D_{IN} is written to the BL. For each differential BL pair, two of these drivers are used with an extra

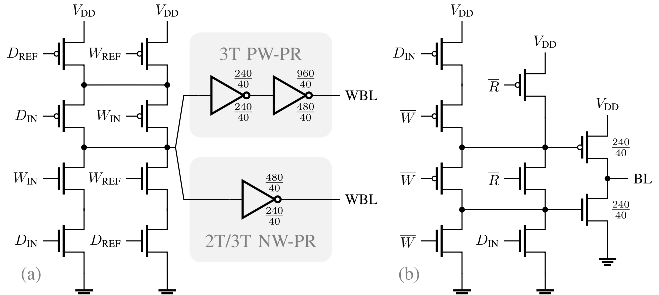


Fig. 4. BL driver for: (a) dynamic memories and (b) static memories with transistor (W/L) sizes in nm. In (a), unannotated transistors are minimum-sized ($W/L = 120/40$) and inverter PMOS/NMOS sizes are shown above/below the inverters, respectively; the driving inverters in the gray boxes are alternatively used for the respective memory. In (b), unannotated NMOS transistors are minimum-sized ($W/L = 120/40$) while unannotated PMOS transistors are double-width, minimum-length ($W/L = 240/40$).

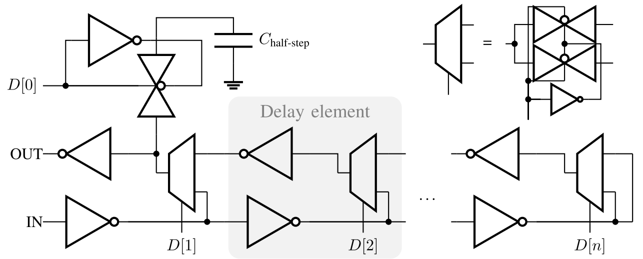


Fig. 5. Delay chain used to generate the timing for the memory control signals. The (W/L) of all transistors is (300 nm/40 nm) and (500 nm/40 nm) for NMOS and PMOS, respectively, except for the transistors in the inverters driving the transmission gates, which are minimum-sized (120 nm/40 nm).

inverter (minimum-size NMOS and double-width PMOS) to generate $\overline{D_{\text{IN}}}$ for the $\overline{\text{BL}}$ driver.

4) Timing Control: Since the exact cell behavior at 4.2 K was unknown at design time, designing a fixed timing circuit was not possible. To allow also detailed cell characterization or debugging, the timing of the control signals is derived asynchronously using programmable delay chains (Fig. 5). The lengths of two inverter chains running in opposite directions are set by transmission-gate-based multiplexers. The delay is determined by the first non-zero element in $D[1{:}n]$. A 3.6-fF metal-oxide-metal (MOM) capacitor $C_{\text{half-step}}$ can be added to the final stage to increase its delay by approximately 50%, resulting in a delay resolution of about 20 ps.

For reading, a 192-step delay chain with a maximum delay of 3.84 ns determines the total duration of the SA's sampling phase. A 16-step (320-ps) delay chain determines how much of the sampling time is spent on precharging the internal SA node on the RBL side to fully reset the SA. The write duration is derived from a single 64-step (1.28-ns) delay chain.

The control-signal generation circuits for both write and read operations are identical for all memories, such that the same settings on different memories result in similar delays. Since the delay chains consume a constant but large amount of energy, their supplies are separated and not included in the reported power budget. For an actual memory, such fine programmability is not needed, allowing for a low-power design. To estimate the delay, especially at cryogenic temperature, a 256-step delay chain is configured as a ring oscillator (RO) by selectively shorting the output to the input through a NAND gate. The frequency of its buffered output can be measured through a pad for several delay settings to estimate the stage delay.

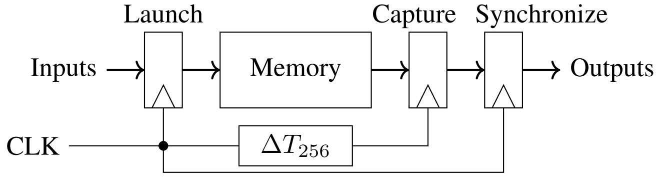


Fig. 6. Latency-measurement setup. If the memory delay is larger than ΔT_{256}, the outputs will be delayed by a clock cycle.

F. Testing Infrastructure

To measure the total read-access latency, the setup shown in Fig. 6 is used. A 256-step (5.12-ns) delay chain generates a programmable delay between the launch and capture registers. The total latency is estimated as the lowest delay setting for which the outputs of the synchronize register are correct. This will include the clock-to-Q and setup time of the launch and capture registers, respectively, which are not removed since they are small and these registers would be needed in any real application to synchronize to the clock to prevent race conditions.

A local controller is connected to each individual memory to execute read, write, and refresh operations. Additionally, it stores and decodes all memory settings, such as the delay chain settings and special test mode flags. A programmable, global controller is connected to local controllers through a shared bus. The global controller is a custom 32-bit, single-cycle microprocessor with 16 different instructions, 32 registers, and a 32-word instruction memory. Additional hardware compares memory read instruction results with the expected values and accumulates the error count for the various tests. The registers, instruction memory, and read-error accumulators are written and read through a shift register (SR). All (automatically synthesized) controllers are clocked at 100 MHz. To account for the cryogenic logic speed-up, 50-ps margin is added to the hold time in the synthesis flow. Due to the bus communication overhead, the maximum memory operation frequency is lower (six cycles per write and eight cycles per read).

G. Additional Test Structures

An often-used metric in static cell design is the static-noise margin (SNM) [69], which indicates the read and write stability of the cell design. The SNM can be estimated by plotting the butterfly curves, which are created by overlaying the dc voltage transfers of both SRAM cell sides. The distance between the curves gives an indication of the noise amplitude needed to flip the state of the cell. A larger SNM, therefore, indicates a more stable cell.

To experimentally characterize the SNM, an array with 256 SVT and 256 LVT half-cells (organized in 32 rows by 16 columns) is included, which matches the actual 6T-cell array layout as accurately as possible up to 1 μm around

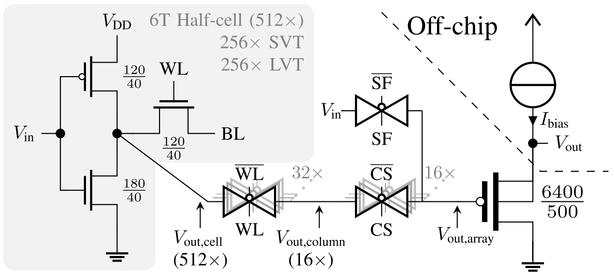


Fig. 7. Structure for static-cell SNM characterization, including a single half-cell, the cell selection hierarchy, and the output buffer with transistor (W/L) in nm.

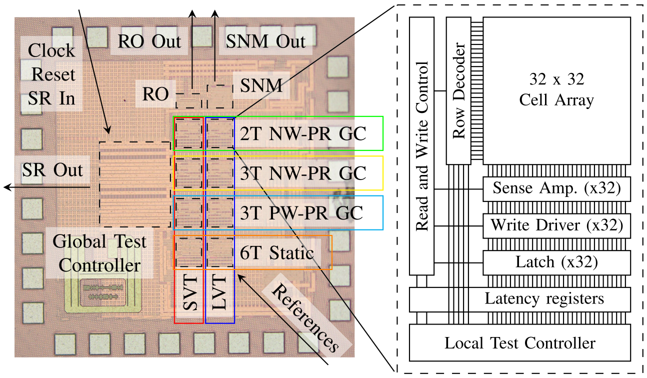


Fig. 8. Annotated micrograph of the chip (left) and a block diagram of the architecture for all memories (right).

the half-cells (Fig. 7). All V_{in} are shorted and driven by a pad. The BLs are driven by tri-state buffers, allowing them to be floating (hold SNM curve), pulled up (read SNM curve), or pulled down (write SNM curve), as shown in Section IV-C. The WLs select one cell from each column to be connected to each $V_{out,column}$ through a transmission gate (all $W/L = 120$ nm/40 nm). A column select (CS) signal selects one $V_{out,column}$ to be connected to $V_{out,array}$ through a transmission gate (all $W/L = 300$ nm/40 nm). A thick-oxide source follower buffers V_{out} to a pad. To characterize the buffer's dc shift, V_{in} can also be connected directly to $V_{out,array}$ using the SF signal.

IV. Measurement Results

A. Measurement Setup

Fabricated in TSMC 40-nm bulk CMOS process (Fig. 8), the test chip has been bonded to a dual in-line (DIL) package and mounted on a printed circuit board (PCB) at the end of a dipstick for testing at RT and 4.2 K by submerging into liquid helium.

The SA reference voltages and SNM input voltage are set by a programmable R&S HMC8043 dc power supply, while the supply for the digital controllers (1.1 V) and pad ring supply (2.5 V) are set by manually tuned low-noise adjustable RT low-dropout (LDO) regulators which operate far from their rated limits to ensure a stable output voltage. The memory macro supplies (1.1 V for all reported measurements) are divided across three pins: one for the dynamic memories, one for the SVT static memory, and one for the LVT static memory. These pins are connected to relays on the PCB to select between the 1.1-V LDO supply or a Keithley 2636B source measure unit (SMU) channel for current measurements. A second SMU channel is used to drive a Lakeshore DT-670 cryogenic temperature sensor, located slightly above the test chip to monitor the approximate environmental temperature. The test-chip digital interface is connected to an RT field-programmable gate array (FPGA), through an optocoupler board for noise isolation, for reprogramming the global test controller and manually sending messages on the shared controller bus.

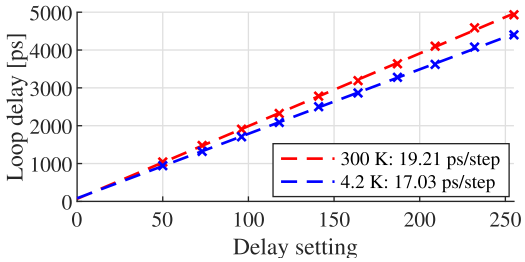


Fig. 9. RO loop delay as a function of the delay chain settings with a least-squares linear fit ($V_{DD} = 1.1$ V).

The average delay of a single delay-chain setting step is determined by measuring the RO frequency with an oscilloscope and fitting the resulting oscillation period for various settings to a linear equation, as shown in Fig. 9. The resulting step delay, later used for latency estimation, shows a cryogenic speed-up of 11%.

B. Dynamic Memories

The retention time of all dynamic memory cells shown in Fig. 10 is measured by writing data to the cell, waiting for varying hold time with the opposite voltage on the WBLs for worst-case leakage, and reading back the data. The retention time is defined as the maximum hold time for which the read data match the written data for both data polarities. A data mismatch for the shortest possible hold time (80 ns) is considered a failure of the cell. The longest measurable retention time is limited to 20 ms to contain the total characterization time (100 ms for the 2T LVT cell at 4.2 K). The SA V_{ref} is optimized for each memory type to give the best retention time performance. The read duration is chosen to be as short as possible without increasing the fail rate. A log-normal cumulative distribution function (cdf) is least-squares fit to the cumulative histogram of the non-failing cells for which a retention time can be determined within the measurement limit. As shown in the following, the good fit is compatible with an exponential distribution of the leakage currents at both temperatures, which is expected for both subthreshold leakage and gate leakage.

At RT, the SVT implementation of the 2T NW-PR cell outperforms the LVT version due to the lower subthreshold leakage through the write transistor. While all SVT cells are functional, nine LVT cells always fail [Fig. 10(b-2)]. Both versions show a clear improvement in retention time from RT to 4.2 K, thanks to the reduced subthreshold leakage. At 4.2 K, both types show a similar retention time since they are both limited by the gate leakage of the readout transistor. At this

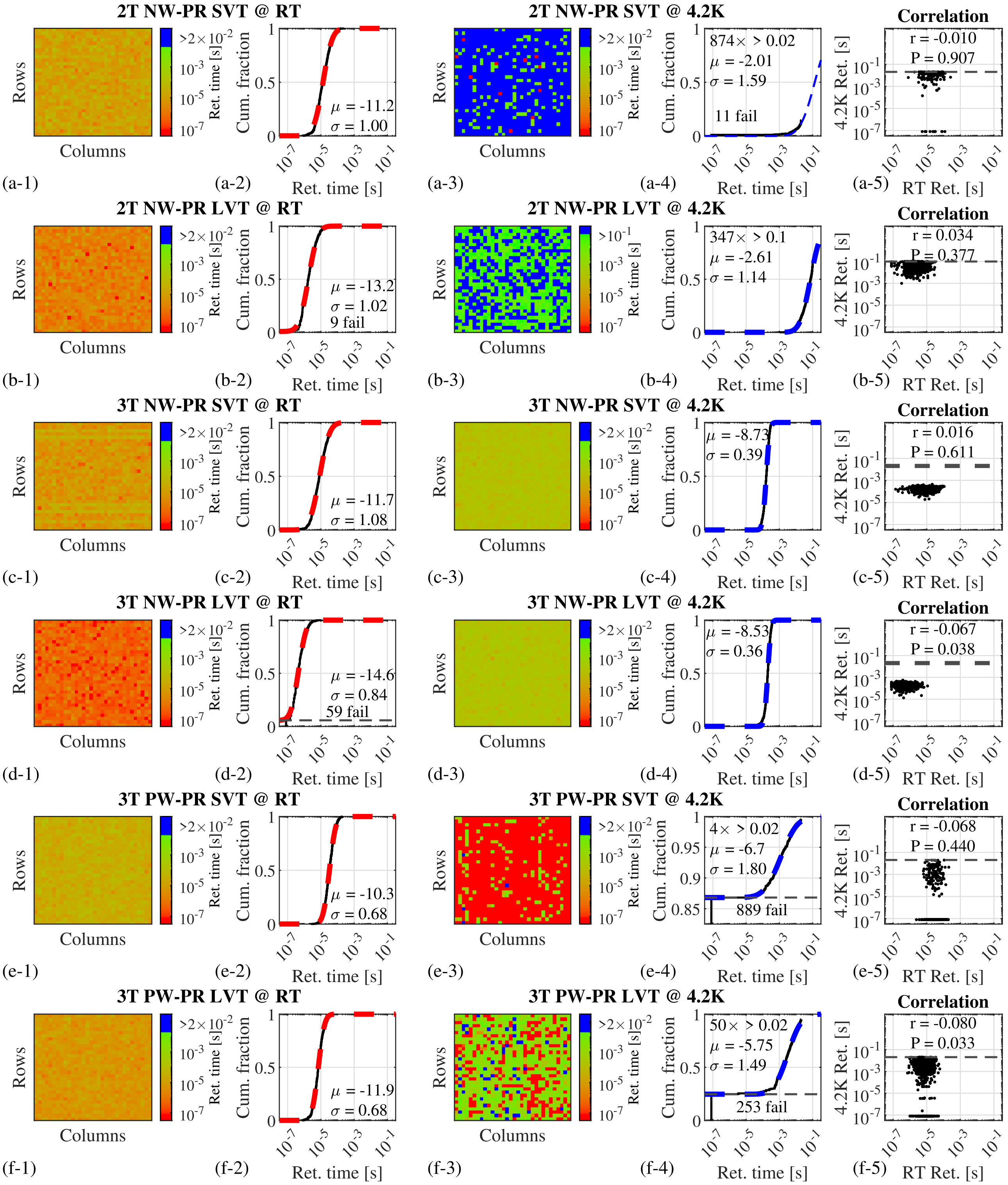


Fig. 10. Retention time measurements for all dynamic cell designs ($V_{\mathrm{DD}} = 1.1$ V) shown in subplots (y-x). The first part of the label (y) indicates the cell design: a = 2T NW-PR SVT; b = 2T NW-PR LVT; c = 3T NW-PR SVT; d = 3T NW-PR LVT; e = 3T PW-PR SVT; f = 3T PW-PR LVT. The second part of the label (x) indicates the type of plot: 1 = RT retention time heatmap per cell; 2 = cumulative distribution of the RT retention time per cell with a log-normal cdf least-squares fit to the non-failing cells with its μ and σ (base-e) and the total number of failing cells (retention time less than 80 ns); 3 = same as 1, but at 4.2 K, where blue cells indicate a retention time longer than the limit imposed by the measurement setup; 4 = same as 2, but at 4.2 K, including the total number of cells with a retention time larger than the measurement setup limit; 5 = scatterplot where each point corresponds to the retention time of a single cell at both temperatures, indicating little to no significant ($P < 0.05$) log–log Pearson correlation (r) between RT and 4.2-K retention times.

temperature, all LVT cells are functional, while 11 SVT cells always fail [Fig. 10(a-4)]. Some cells have a high M_2 $|V_{\mathrm{th}}|$, resulting in low readout currents for both SN voltages. Due to the V_{th} increase at 4.2 K, these currents drops below the readout currents of some other cells with high M_1 V_{th} and low M_2 $|V_{\mathrm{th}}|$. One of these cells must be deemed failing since there exists no reference voltage for which the states of both cells can be distinguished. While LVT failures at RT are related to

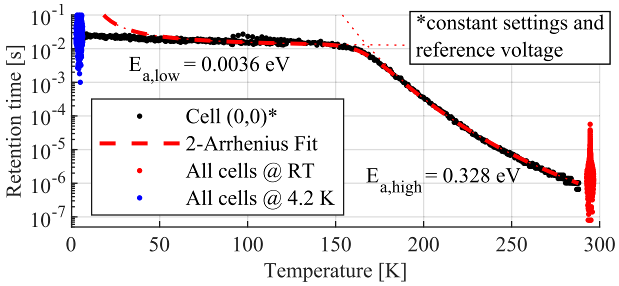


Fig. 11. Retention time over temperature for a single 2T NW-PR LVT cell with a two-term Arrhenius fit, and retention-time distribution of all cells at RT and 4.2 K ($V_{\mathrm{DD}} = 1.1$ V). The RT and 4.2-K retention-time distributions use different optimized settings and reference voltages. The single-cell temperature sweep uses the 4.2-K settings.

large leakage currents, the SVT failures at 4.2 K are related to insufficient SN margins.

For both the SVT and LVT designs, several cells exceed the retention time limit, resulting in an unreliable log-normal fit for the SVT memory (Fig. 10(a-4), fit to only 139 out of 1013 functional cells). For the LVT cells, the fit is more reliable and shows an increase in both the average retention time ($4 \times 10^4 \times$) and spread of the retention time. Note that the increased spread cannot be directly attributed to the cryogenic increase in transistor mismatch, as the retention time is limited by different physical effects at the two temperatures, as explained in the following. Both cell designs show no significant correlation between the retention times at RT and 4.2 K [Fig. 10(a-5) and (b-5)].

The 3T NW-PR designs show a smaller improvement in retention time from RT to 4.2 K. Their RT retention time is similar to the 2T NW-PR cells, but their 4.2 K retention time is much lower and with lower spread. The lower spread may be due to the lower relative impact of the mismatch on a larger leakage. The LVT implementation shows a significant (weak) negative correlation of the retention time [Fig. 10(d-5)], which can be explained by the fact that, while a low V_{th} of the write transistor causes a large RT leakage, it also provides better SN voltage margins that improve the retention time at 4.2 K.

The retention time of the 3T PW-PR designs is superior to the other cell flavors at RT, thanks to the preferential boosting technique. However, 889 SVT cells [Fig. 10(e-4)] and 253 LVT cells [Fig. 10(f-4)] always fail at 4.2 K (dashed lines). This is attributed to the V_{th} increase of all transistors, which limits both the SN voltages that can be written and the readout current. This is explained in more detail in the following. The LVT implementation also shows a significant weak negative retention-time correlation [Fig. 10(f-5)], similar to what happens for the 3T NW-PR cells and attributed to the same effects.

Overall, at RT, the LVT cells show higher error rates and shorter retention time than the SVT cells due to their larger subthreshold leakage. At 4.2 K, however, the LVT cells show fewer failures than the SVT cells due to the compensation for the cryogenic V_{th} increase and better SN voltage margins. The differences in retention times between LVT and SVT for functional cells are also smaller, indicating that their leakage is much more similar.

The transition between subthreshold leakage and gate leakage can be observed by continuously sweeping the ambient temperatures. This can be accomplished by slowly raising/lowering the chip's vertical position in the helium vapors above the liquid helium surface. Two regions clearly appear, as shown for the 2T NW-PR LVT cell in Fig. 11. For high temperatures (>160 K), the subthreshold leakage dominates while for low temperatures (<160 K), the gate leakage dominates. The temperature dependence of each leakage process can be determined by fitting to a sum of two Arrhenius equations

$$\frac{1}{t_{\mathrm{ret}}} \propto I_{\mathrm{leak}} = A_{\mathrm{high}} e^{-\frac{E_{a,\mathrm{high}}}{k_b T}} + A_{\mathrm{low}} e^{-\frac{E_{a,\mathrm{low}}}{k_b T}} \qquad (1)$$

where k_b is Boltzmann's constant, T is the absolute temperature, A_{high} and A_{low} are the proportionality constants, and $E_{a,\mathrm{high}}$ and $E_{a,\mathrm{low}}$ are the activation energies, as is usually done in the literature. The high-temperature activation energy $E_{a,\mathrm{high}} = 0.328$ eV matches the expected value for subthreshold leakage $E_{a,\mathrm{subth}} = \ln(10) k_b (V_{\mathrm{th}}(0)/s_0)$ with $V_{\mathrm{th}}(0)$ the extrapolated V_{th} at 0 K and s_0 the linearized subthreshold slope temperature dependence [30]. The low-temperature activation energy $E_{a,\mathrm{low}}$ indicates very little temperature dependence, which is in-line with the expected very small temperature dependence of the gate leakage. However, since gate leakage does not actually follow an Arrhenius equation, the fit fails below 50 K. This shows that simply fitting an Arrhenius equation for temperatures above 50 K is not sufficient to predict the cell's retention time at 4.2 K.

The retention time over temperature for the 3T cell designs is shown in Fig. 12 for both data polarities separately. These also show the subthreshold leakage limitation at high temperatures, mainly for the high SN voltages. Since the readout transistor gate leakage pulls up the SN, the high-SN retention time becomes infinite when the subthreshold leakage becomes smaller than the gate leakage.

For temperatures below 200 K, the retention time becomes limited by the state with a low SN voltage due to the gate leakage. Especially for the 3T NW-PR cells, the readout transistor is then in strong inversion ($|V_{\mathrm{gs}}| = V_{\mathrm{DD}}$), resulting in a much larger gate leakage than for the 2T NW-PR cells where the readout transistor is never in inversion during the hold time. Although the gate leakage is assumed to be roughly constant over temperature, the retention time decreases over temperature due to the V_{th} shift of the readout transistors. This results in smaller readout currents and readout margins at lower temperatures and thus earlier cell failures.

For the 3T PW-PR cells, the V_{th} shift has a double effect, also increasing the lowest voltage that can be written through the PMOS. This means that the readout transistor overdrive $V_{\mathrm{gs}} - |V_{\mathrm{th}}|$ during readout decreases due to the decrease in V_{gs} and the increase in $|V_{\mathrm{th}}|$, resulting in a reduced retention time for lower temperatures and even in failing cells. The limited write SN voltage issue could be overcome using WL-boosting, which is not used here since it would increase the design complexity and could impact the reliability of the devices.

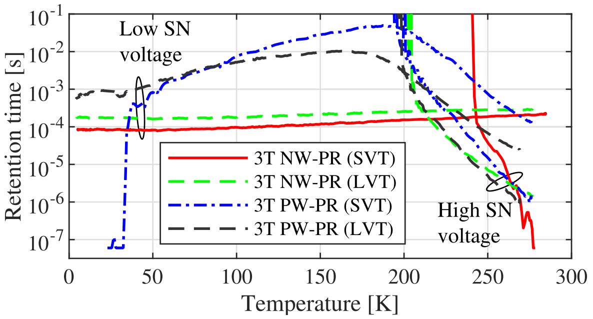


Fig. 12. Retention time over temperature for a single cell of each 3T cell design for the low-SN-voltage and high-SN-voltage states separately ($V_{\mathrm{DD}} = 1.1$ V).

TABLE I

INPUT-REFERRED OFFSET AND NOISE OF 64 SAS ($V_{\mathrm{DD}} = 1.1$ V)

	Offset $\mu \pm \sigma$ [mV]		Mean Noise [mV$_{\mathrm{RMS}}$]	
Cell type	RT	4.2 K	RT	4.2 K
2T NW-PR	$33(1) \pm 11(1)$	$41(1) \pm 12(1)$	2.44(8)	0.68(3)
3T NW-PR	$35(2) \pm 18(2)$	$43(2) \pm 19(2)$	2.11(7)	0.64(4)
3T PW-PR	$-39(2) \pm 15(1)$	$-47(2) \pm 16(1)$	1.89(4)	0.58(3)

Uncertainty indicated in concise uncertainty notation.

Table I lists the input-referred offset and noise of the dynamic-memory SAs. These are determined by accumulating the average SA output over 1023 comparisons for a fixed V_{ref} and variable RBL voltage by connecting the RBLs to a pad. A binary search is performed to find the RBL voltage for which the average SA output equals 0.5. The input-referred offset and rms noise are found by fitting a normal cdf to the average output for various differential input voltages, similar to the method used in [76]. The measurement is performed for all 6 × 32 dynamic-memory SAs. Since the SAs in the SVT and LVT cell memories are identical, there are three unique designs with 64 samples each.

The 2T NW-PR and 3T NW-PR memories use identical SA designs, resulting in no significant difference in systematic offset (μ). The offset spread (σ) and RT noise are significantly different and attributed to differences in the layout needed to fit different column pitches. Their 4.2-K noise is similar.

All designs show a significant mean offset due to the unequal loading of the two output nodes, which is positive for the HSNA-VLSAs and negative for the FSPA-VLSAs. Although the SA design difference for the 3T PW-PR memories results in different mean offset, the offset spread and noise performance for all SAs are similar.

The absolute systematic offset of all SA designs increases by 22%–25% when cooling down from RT to 4.2 K. This is attributed to a reduction in parasitic capacitance due to the reduction of the source/drain junction capacitance, which could increase the effects of, e.g., charge injection. The offset spread increases by 5%–9% (although with very limited statistical confidence) due to the increase in mismatch, while the rms noise decreases by ∼70% due to the reduced thermal noise. While the SA designs are different, the changes in input-referred offset and noise of the NMOS and PMOS versions do not show significant differences.

The operation energy, leakage power, and full-memory latency of the dynamic memories are shown in the first six rows of Table II. These are given in ranges since various timing settings are possible, resulting in different tradeoffs. In general, shorter timing settings result in lower latency, lower operation energy due to reduced BL swing, and lower retention time. The lower retention time will, however, give a higher static power consumption, resulting in a tradeoff between static and dynamic power.

The leakage power (P_{leakage}) is determined by measuring the average power consumption without any memory operations. Note that the reported DRAM P_{leakage} is the average leakage per memory bank, as all the dynamic memories share the same supply. For the SRAM, the leakage per individual bank (SVT or LVT) is measured and reported. The average write power is measured by writing random data to random addresses of the selected memory, generated using a linear feedback shift register pseudo random number generator. By subtracting P_{leakage} from the average write power and dividing the result by the write operation frequency f_{write}, the write energy per operation E_{write} is obtained. Next, a combination of random writes and reads are performed to obtain the average combined write and read power from which P_{leakage} and write power ($E_{\mathrm{write}} \times f_{\mathrm{write}}$) are subtracted and divided by the read operation frequency f_{read} to obtain the read energy per operation E_{read}. The full memory refresh energy E_{refresh} is determined by dividing the average power when refreshing the entire memory by the refresh frequency f_{refresh}. Note that E_{read} (E_{write}) is the energy required to read (write) a single 32-bit word, while E_{refresh} is the energy required to fresh a full memory bank (32 words). The latency is determined by reading alternating data polarities while reducing the latency-measurement delay chain setting until the read is unsuccessful.

In general, there is a decrease in operation energy from RT to 4.2 K, which is mainly attributed to the decrease in source/drain junction capacitance. Furthermore, the leakage reduces to inappreciable levels and the latency decreases due to the improved digital speed.

C. Static Memories

The SNMs of the 6T cells are measured using the special test structures in Fig. 7 by sweeping the input voltage while looping over different cells. The cell's output voltage is determined from the source follower's output voltage after compensating for the source-follower transfer. Fig. 13 shows the measured SNM curves of the SVT and LVT 6T half-cells at RT (red) and 4.2 K (blue) overlaid on mirrored versions of the curves to show the SNM gaps. At RT, all 256 SVT and 256 LVT half-cells are measured. At 4.2 K, only eight SVT and eight LVT half-cells have been measured because much longer measurement times are required at 4.2 K due to the lower currents and larger transmission-gate impedance around the digital-level transitions [66].

At 4.2 K, the hold curves [Fig. 13(a-1) and (b-1)] show sharper corners due to the steeper subthreshold slope. Furthermore, the V_{out} versus V_{in} curves slightly shift to the right, thus moving toward the middle of the voltage range and marginally increasing the hold SNM. Such a shift is due to the

TABLE II
CELL AREA, OPERATION ENERGY, LEAKAGE POWER, AND LATENCY OF ALL MEMORY DESIGNS AT RT AND 4.2 K ($V_{\mathrm{DD}} = 1.1$ V)

Cell type	Cell area [µm^2]	E_{read} [fJ/op]		E_{write} [fJ/op]		E_{refresh} [pJ]		P_{leakage} [µW]		Latency [ns]	
		RT	4.2 K	RT	4.2 K	RT	4.2 K	RT	4.2 K	RT	4.2 K
2T NW-PR (SVT)	0.184	380-396	308-336	163-165	153-154	23.5-23.8	21.7-22.1	0.13	0*	2.52	2.03-2.17
2T NW-PR (LVT)		303-454	334-358	134-190	153-154	20.8-23.2	21.8-22.1	0.13	0*	1.82-2.56	2.05-2.36
3T NW-PR (SVT)	0.242	455-511	355-438	153-184	156-157	26.0-26.2	22.4-23.7	0.13	0*	1.60	N/A
3T NW-PR (LVT)		486-496	407-413	181-183	156-157	26.3-26.5	23.5	0.13	0*	1.38-1.40	1.22
3T PW-PR (SVT)	0.254	243-330	204-223	208-237	211-213	25.2-25.8	16.7-16.9	0.13	0*	1.96-2.34	N/A†
3T PW-PR (LVT)		277-342	258-266	196-241	211-213	24.0	22.2	0.13	0*	1.84	2.37
6T Static (SVT)	0.435	707-712	613-621	489-494	462	-	-	0.68	0*	1.42	1.23
6T Static (LVT)		695-725	609-627	484-502	473	-	-	5.5	0*	1.40	1.18

*: Limited by setup accuracy. †: Too many failing cells.

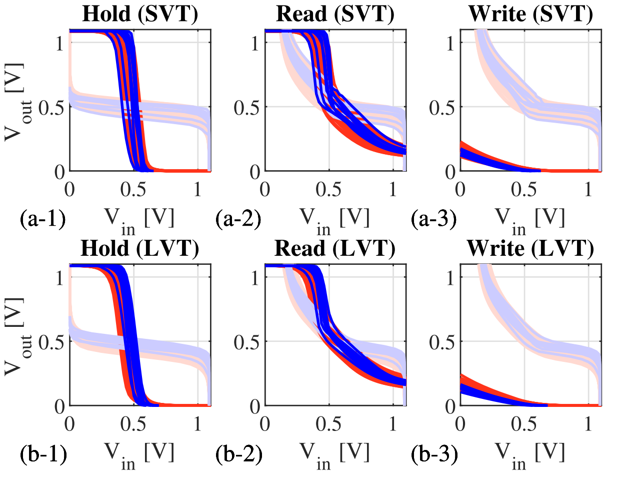


Fig. 13. SNM curves at RT (red) and 4.2 K (blue) of the SVT 6T cells with BL floating (hold curves, a-1), BL pulled up (read curves, a-2), and BL pulled down (write curves, a-3) and the LVT 6T cells with BL floating (hold curves, b-1), BL pulled up (read curves, b-2), and BL pulled down (write curves, b-3).

cryogenic threshold increase of the NMOS dominating over the PMOS threshold increase due to NMOS being stronger, thanks to its larger mobility.

The left half of the read curves [Fig. 13(a-2) and (b-2)] follows the hold curves, including the sharper corners and the shift to the right. In the right half of the plot, the inverter NMOS pulls down while the access NMOS pulls up. Since both transistors are equally affected by the temperature change, their effects partially cancel out. Thanks to the curve shift in the left half, the read SNM increases slightly at 4.2 K.

For the write curves [Fig. 13(a-3) and (b-3)], the right half follows the hold curves. Since in the left part, the inverter PMOS pulls up and the access NMOS pulls down, the 4.2 K curves are pulled down more due to the NMOS cryogenic shift dominating over the PMOS. This also results in an increase of the write SNM.

The operation energy, leakage power, and latency of the static memory designs are shown in the two bottom rows of Table II. These metrics are determined using the same method as for the dynamic memories. For the static memories, a slight decrease by ~13% and ~5% in read and write energy is observed, respectively. Since voltage swings stay approximately constant, this is expected to be caused by the reduced node capacitance due to the reduction in source/drain junction capacitance. While there is a significant static leakage at RT, especially for the LVT cells, it becomes inappreciable at 4.2 K. Furthermore, the latency decreases by about 14% due to the increased readout current.

V. DISCUSSION

For nearly all memories, only a marginal improvement in operation energy and latency from RT to 4.2 K has been observed, in combination with very significant improvements in leakage power and in DRAM retention time, which lowers the refresh power. Since subthreshold leakage becomes negligible, LVT devices become a natural choice to improve performance with their lower V_{th}, resulting in faster operation for the static cells and larger retention times for the dynamic cells, thanks to the larger SN margins.

Furthermore, some relevant guidelines for cryogenic design can be inferred. For the dynamic cells, RT techniques to improve the retention time may fail at 4.2 K, as shown for the 3T PW-PR cell. In that case, the V_{th} shift of the readout and write transistors severely reduces the readout currents and the retention time. For the readout transistor, the overdrive for the current-generating state must be large enough to mitigate the V_{th} shift. For the write transistor, the current-generating state must be written strongly. As a result, the readout and write transistors should be of a different type (PMOS/NMOS). Additionally, since gate leakage is the dominant leakage source at 4.2 K, it should be minimized by avoiding readout devices in strong inversion and selecting the device type with the lowest gate leakage.

For the static memories, a slight increase in SNMs is expected. Despite being small at RT, the read SNM for the LVT cells is apparently larger than that of SVT cells at 4.2 K, allowing the use of LVT cells with similar leakage power and lower latency than the SVT cells, although definitive conclusions cannot be drawn due to the limited sample size. Given the improved write SNM, a different sizing in favor of the read stability (larger pull-down transistors) may allow for even better cells under mismatch, at the cost of a slight increase in area.

Using the values reported in Table II, a quantitative comparison of the expected power consumption for various applications can be drafted. Each application is defined by its access rate (limited by the memory's latency) and the W/R ratio, ranging from 0 to 1 assuming that we do not write

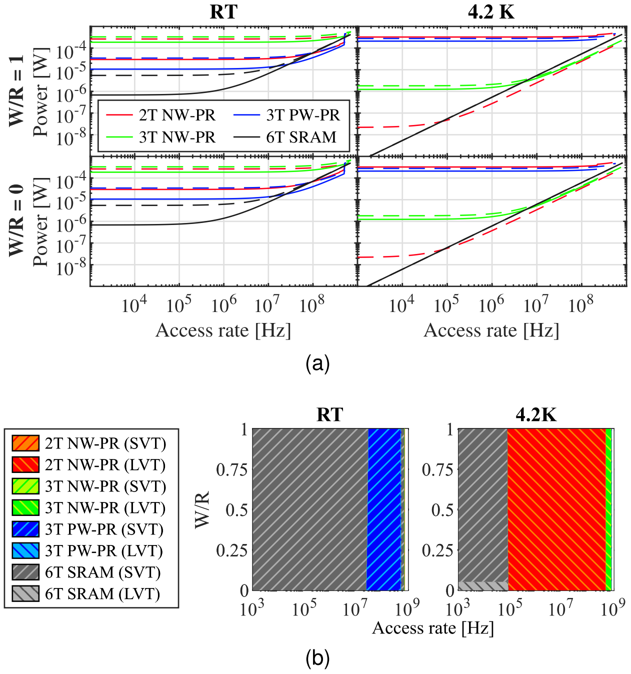


Fig. 14. Expected full memory power consumption ($V_{\mathrm{DD}} = 1.1$ V, excluding control signal generation) over the full application space, showing: (a) power consumption for $W/R = 0$ and $W/R = 1$ (solid SVT and dashed LVT) and (b) memory with the lowest power consumption. The refresh rate is determined by the worst cell retention time measured over multiple runs. Using the setting configurations with the minimum power results in discontinuities for high access rates where some configurations are too slow for the required speed.

data that is never read. For each memory, Fig. 14(a) shows a flat refresh/leakage-dominated region and an operation-power-dominated region, with little dependence on W/R. Since the write energy is lower than the read energy, there will be a minor power consumption decrease for all memories. At RT, all LVT memories consume much more power than the SVT versions. However, at 4.2 K, some LVT memories perform better (2T) or roughly equal (6T). The gap decreases for the 3T memories. Only the 2T NW-PR (LVT) and 3T NW-PR designs improve from RT to 4.2 K since their worst cell retention time improves, resulting in lower refresh rates. In Fig. 14(b) at RT, the 6T SVT memory is most efficient below 25 MHz, thanks to the leakage power being lower than the DRAMs' refresh power. Above 25 MHz, the 3T PW-PR SVT memory consumes the lowest power, thanks to the low operation energy and the highest retention time. At 4.2 K, the LVT 2T NW-PR outperforms the SRAM already beyond 75 kHz, also being the smallest and even 24% smaller than the foundry SRAM cell, since its retention time at 4.2 K is much longer than the SVT 3T PW-PR at RT, resulting in a significantly lower refresh power. Finally, although the higher latency of the 2T NW-PR and 3T PW-PR memories may limit their maximum access rate, multi-bank architectures could be adopted to reach a much higher throughput using slower banks, therefore not constituting a fundamental issue.

Based on the proposed guidelines, more advanced cell designs could be considered beyond this work (see [77]). Additionally, the presented tradeoffs do not capture the full range of considerations for memory selection. For instance, refresh operations will reduce the DRAM availability and noise-limited bit-error rates must be acceptable for the application. Reliability and security aspects may also be relevant, such as retention time limitations due to row-hammer attacks [24].

VI. Conclusion

By comparing single-bank static and dynamic memories at cryogenic temperature, this article shows that well-designed dynamic memories can outperform static memories for middle-to-high frequency applications in terms of area and power. While the subthreshold leakage reduces substantially from RT to 4.2 K, gate leakage stays approximately constant, thus still limiting the retention time. Still, adopting dynamic cells with enhanced resistance to gate leakage and cryogenic V_{th} shifts can significantly increase retention time, thus lowering the refresh power. The increased variability in both cells and peripherals may increase the number of outlier cells, while the lower noise reduces the read error rate. Embracing the design guidelines outlined here for cryogenic embedded memories will facilitate the adoption of dynamic-memory cells for high-density low-power cryogenic memories, thereby enabling the complex cryo-CMOS SoCs needed in future QCs.

Acknowledgment

The authors would like to thank Atef Akhnoukh for technical support.

References

[1] P. W. Shor, "Algorithms for quantum computation: Discrete logarithms and factoring," in *Proc. 35th Annu. Symp. Found. Comput. Sci.*, Nov. 1994, pp. 124–134.

[2] L. K. Grover, "Quantum mechanics helps in searching for a needle in a haystack," *Phys. Rev. Lett.*, vol. 79, no. 2, p. 325, 1997.

[3] A. Montanaro, "Quantum algorithms: An overview," *npj Quant. Inf.*, vol. 2, no. 1, pp. 1–8, 2016.

[4] A. W. Harrow and A. Montanaro, "Quantum computational supremacy," *Nature*, vol. 549, no. 7671, pp. 203–209, Sep. 2017.

[5] D. J. Egger et al., "Quantum computing for finance: State-of-the-art and future prospects," *IEEE Trans. Quantum Eng.*, vol. 1, pp. 1–24, 2020.

[6] F. Bova, A. Goldfarb, and R. G. Melko, "Commercial applications of quantum computing," *EPJ Quantum Technol.*, vol. 8, no. 1, p. 2, Dec. 2021.

[7] F. Sebastiano et al., "Cryo-CMOS electronic control for scalable quantum computing," in *Proc. 54th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2017, pp. 1–6.

[8] 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.

[9] X. Fu, L. Lao, K. Bertels, and C. G. Almudever, "A control microarchitecture for fault-tolerant quantum computing," *Microprocessors Microsyst.*, vol. 70, pp. 21–30, Oct. 2019.

[10] P. Wang, X. Peng, W. Chakraborty, A. I. Khan, S. Datta, and S. Yu, "Cryogenic benchmarks of embedded memory technologies for recurrent neural network based quantum error correction," in *IEDM Tech. Dig.*, Dec. 2020, pp. 38.5.1–38.5.4.

[11] 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.

[12] P. Das, A. Locharla, and C. Jones, "LILLIPUT: A lightweight low-latency lookup-table decoder for near-term quantum error correction," in *Proc. 27th ACM Int. Conf. Architectural Support Program. Lang. Operating Syst.* New York, NY, USA: Association for Computing Machinery, Feb. 2022, pp. 541–553, doi: 10.1145/3503222.3507707.

[13] P. Das et al., "AFS: Accurate, fast, and scalable error-decoding for fault-tolerant quantum computers," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Apr. 2022, pp. 259–273.

[14] F. Battistel et al., "Real-time decoding for fault-tolerant quantum computing: Towards higher decoding speed and lower communication latency," *Bull. Amer. Phys. Soc.*, vol. 7, no. 3, Aug. 2023, Art. no. 032003, doi: 10.1088/2399-1984/aceba6.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Y. Liu, L. Lang, Y. Chang, Y. Shan, X. Chen, and Y. Dong, "Cryogenic characteristics of multinanoscales field-effect transistors," *IEEE Trans. Electron Devices*, vol. 68, no. 2, pp. 456–463, Feb. 2021.

[20] F. Ware et al., "Do superconducting processors really need cryogenic memories? The case for cold DRAM," in *Proc. Int. Symp. Memory Syst.*, Oct. 2017, pp. 183–188.

[21] G.-H. Lee, D. Min, I. Byun, and J. Kim, "Cryogenic computer architecture modeling with memory-side case studies," in *Proc. ACM/IEEE 46th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2019, pp. 774–787.

[22] D. Min, I. Byun, G.-H. Lee, S. Na, and J. Kim, "CryoCache: A fast, large, and cost-effective cache architecture for cryogenic computing," in *Proc. 25th Int. Conf. Architectural Support Program. Lang. Operating Syst.*, Mar. 2020, pp. 449–464.

[23] I. Byun, D. Min, G.-H. Lee, S. Na, and J. Kim, "CryoCore: A fast and dense processor architecture for cryogenic computing," in *Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA)*, May 2020, pp. 335–348.

[24] G.-H. Lee, S. Na, I. Byun, D. Min, and J. Kim, "CryoGuard: A near refresh-free robust DRAM design for cryogenic computing," in *Proc. ACM/IEEE 48th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2021, pp. 637–650.

[25] D. Prasad et al., "Cryo-computing for infrastructure applications: A technology-to-microarchitecture co-optimization study," in *IEDM Tech. Dig.*, Dec. 2022, p. 23.

[26] P. Wyns, R. Anderson, and W. D. Jardins, "Temperature dependence of required refresh time in dynamic random access memories," in *Proc. Symp. Low Temp. Electron. High Temp. Superconductors*, vol. 88, 1988, p. 123.

[27] P. Wyns and R. L. Anderson, "Low-temperature operation of silicon dynamic random-access memories," *IEEE Trans. Electron Devices*, vol. 36, no. 8, pp. 1423–1428, Aug. 1989.

[28] J. A. Halderman et al., "Lest we remember: Cold-boot attacks on encryption keys," *Commun. ACM*, vol. 52, no. 5, pp. 91–98, May 2009.

[29] S. S. Tannu, D. M. Carmean, and M. K. Qureshi, "Cryogenic-DRAM based memory system for scalable quantum computers: A feasibility study," in *Proc. Int. Symp. Memory Syst.*, Oct. 2017, pp. 189–195.

[30] F. Wang, T. Vogelsang, B. Haukness, and S. C. Magee, "DRAM retention at cryogenic temperatures," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2018, pp. 1–4.

[31] T. Kelly et al., "Some like it cold: Initial testing results for cryogenic computing components," *J. Phys., Conf. Ser.*, vol. 1182, Feb. 2019, Art. no. 012004.

[32] N. Yoshikawa et al., "Characterization of 4 K CMOS devices and circuits for hybrid Josephson-CMOS systems," *IEEE Trans. Appiled Supercond.*, vol. 15, no. 2, pp. 267–271, Jun. 2005.

[33] Q. Liu et al., "Simulation and measurements on a 64-kbit hybrid josephson-CMOS memory," *IEEE Trans. Appl. Supercond.*, vol. 15, no. 2, pp. 415–418, Jun. 2005.

[34] Q. Liu et al., "Latency and power measurements on a 64-kb hybrid josephson-CMOS memory," *IEEE Trans. Appl. Supercond.*, vol. 17, no. 2, pp. 526–529, Jun. 2007.

[35] T. Van Duzer et al., "64-kb hybrid josephson-CMOS 4 Kelvin RAM with 400 ps access time and 12 mW read power," *IEEE Trans. Appl. Supercond.*, vol. 23, no. 3, Jun. 2013, Art. no. 1700504.

[36] K. Kuwabara, H. Jin, Y. Yamanashi, and N. Yoshikawa, "Design and implementation of 64-kb CMOS static RAMs for josephson-CMOS hybrid memories," *IEEE Trans. Appl. Supercond.*, vol. 23, no. 3, Jun. 2013, Art. no. 1700704.

[37] H. Jin, K. Kuwabara, Y. Yamanashi, and N. Yoshikaw, "Investigation of robust CMOS amplifiers for Josephson-CMOS hybrid memories," *Phys. Proc.*, vol. 36, pp. 229–234, Jan. 2012.

[38] M. Tanaka, M. Suzuki, G. Konno, Y. Ito, A. Fujimaki, and N. Yoshikawa, "Josephson-CMOS hybrid memory with nanocryotrons," *IEEE Trans. Appl. Supercond.*, vol. 27, no. 4, pp. 1–4, Jun. 2017.

[39] G. Konno, Y. Yamanashi, and N. Yoshikawa, "Fully functional operation of low-power 64-kb josephson-CMOS hybrid memories," *IEEE Trans. Appl. Supercond.*, vol. 27, no. 4, pp. 1–7, Jun. 2017.

[40] H. L. Chiang et al., "Cold CMOS as a power-performance-reliability booster for advanced FinFETs," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2.

[41] Z. Wang et al., "Designing EDA-compatible cryogenic CMOS platform for quantum computing applications," in *Proc. 5th IEEE Electron Devices Technol. Manuf. Conf. (EDTM)*, Apr. 2021, pp. 1–3.

[42] V. P. Hu and C.-J. Liu, "Static noise margin analysis for cryo-CMOS SRAM cell," in *Proc. IEEE Int. Symp. Radio-Frequency Integr. Technol. (RFIT)*, Aug. 2021, pp. 1–2.

[43] W. Link and H. May, "Low temperature characteristics of MOS single-transistor memory cells," *Archiv Elektronik Uebertragungstechnik*, vol. 33, pp. 229–235, Jun. 1979.

[44] T. I. Chappell et al., "A 3.5 ns/77 K and 6.2 ns/300 K 64 k CMOS RAM with ECL interfaces," *IEEE J. Solid-State Circuits*, vol. 24, no. 4, pp. 859–868, Aug. 1989.

[45] W. H. Henkels et al., "A 12-ns low-temperature DRAM," *IEEE Trans. Electron Devices*, vol. 36, no. 8, pp. 1414–1422, Aug. 1989.

[46] R. C. Jaeger and T. N. Blalock, "Quasi-static RAM design for high performance operation at liquid nitrogen temperature," *Cryogenics*, vol. 30, no. 12, pp. 1030–1035, Dec. 1990.

[47] W. H. Henkels et al., "A 4-Mb low-temperature DRAM," *IEEE J. Solid-State Circuits*, vol. 26, no. 11, pp. 1519–1529, Nov. 1991.

[48] 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.

[49] E. Garzón, Y. Greenblatt, O. Harel, M. Lanuzza, and A. Teman, "Gain-cell embedded DRAM under cryogenic operation—A first study," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 7, pp. 1319–1324, Jul. 2021.

[50] E. Garzón, A. Teman, and M. Lanuzza, "Embedded memories for cryogenic applications," *Electronics*, vol. 11, no. 1, p. 61, Dec. 2021.

[51] J.-H. Bae et al., "Characterization of a capacitorless DRAM cell for cryogenic memory applications," *IEEE Electron Device Lett.*, vol. 40, no. 10, pp. 1614–1617, Oct. 2019.

[52] S. Chakraborty and J. P. Kulkarni, "Cryo-TRAM: Gated thyristor based capacitor-less DRAM for cryogenic computing," in *Proc. Device Res. Conf. (DRC)*, Jun. 2022, pp. 1–2.

[53] W. Chakraborty et al., "Multi-bit per-cell 1T SiGe floating body RAM for cache memory in cryogenic computing," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Jun. 2022, pp. 302–303.

[54] W. Chakraborty et al., "Pseudo-static 1T capacitorless DRAM using 22nm FDSOI for cryogenic cache memory," in *IEDM Tech. Dig.*, Dec. 2021, p. 40.

[55] 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.

[56] R. A. Damsteegt, "A benchmark of cryo-CMOS 40-nm embedded SRAM/DRAMs for quantum computing," 4TU.ResearchData, Dataset, Delft, 2024, doi: 10.4121/3886ef40-1b1a-479f-a5c4-d3584b90e8a4.

[57] M. Mouis, J. W. Lee, D. Jeon, M. Shi, M. Shin, and G. Ghibaudo, "Source/drain induced defects in advanced MOSFETs: What device electrical characterization tells," *Phys. Status Solidi C*, vol. 11, no. 1, pp. 138–145, Jan. 2014.

[58] F. Balestra and G. Ghibaudo, "Physics and performance of nanoscale semiconductor devices at cryogenic temperatures," *Semicond. Sci. Technol.*, vol. 32, no. 2, Feb. 2017, Art. no. 023002.

[59] A. Beckers, F. Jazaeri, and C. Enz, "Characterization and modeling of 28-nm bulk CMOS technology down to 4.2 K," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 1007–1018, 2018.

[60] 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.

[61] S. S. Parihar, V. M. van Santen, S. Thomann, G. Pahwa, Y. S. Chauhan, and H. Amrouch, "Cryogenic CMOS for quantum processing: 5-nm FinFET-based SRAM arrays at 10 K," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 8, pp. 3089–3102, Aug. 2023.

[62] P. A. t. Hart, M. Babaie, E. Charbon, A. Vladimirescu, and F. Sebastiano, "Characterization and modeling of mismatch in cryo-CMOS," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 263–273, 2020.

[63] A. Grill et al., "Temperature dependent mismatch and variability in a cryo-CMOS array with 30k transistors," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2022, pp. 1–6.

[64] R. Saligram, S. Datta, and A. Raychowdhury, "Scaled back end of line interconnects at cryogenic temperatures," *IEEE Electron Device Lett.*, vol. 42, no. 11, pp. 1674–1677, Nov. 2021.

[65] H. Homulle, "Cryogenic electronics for the read-out of quantum processors," Ph.D. dissertation, Delft Univ. Technol., Delft, The Netherlands, May 2019.

[66] R. W. J. Overwater, M. Babaie, and F. Sebastiano, "Cryogenic-aware forward body biasing in bulk CMOS," *IEEE Electron Device Lett.*, vol. 45, no. 2, pp. 152–155, Feb. 2024.

[67] N. Fakkel, M. Mortazavi, R. W. J. Overwater, F. Sebastiano, and M. Babaie, "A cryo-CMOS DAC-based 40-Gb/s PAM4 wireline transmitter for quantum computing," *IEEE J. Solid-State Circuits*, early access, 2024, doi: 10.1109/JSSC.2024.3364968.

[68] 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.

[69] K. Ishibashi and K. Osada, *Low Power and Reliable SRAM Memory Cell and Array Design*, vol. 31. London, U.K.: Springer, 2011.

[70] TSMC. *40 nm Technology*. Accessed: Feb. 5, 2023. [Online]. Available: https://www.tsmc.com/english/dedicatedFoundry/technology/logic/l_40nm

[71] R. H. Dennard, "Field-effect transistor memory," U.S. Patent 3 286 387, Jun. 4, 1968.

[72] B. Keeth, R. J. Baker, B. Johnson, and F. Lin, *DRAM Circuit Design: Fundamental and High-Speed Topics*, vol. 13. Hoboken, NJ, USA: Wiley, 2007.

[73] W. Regitz and J. Karp, "A three transistor-cell, 1024-bit, 500 ns MOS RAM," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 1970, pp. 42–43.

[74] K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim, "A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1495–1505, Jun. 2011.

[75] T. Na, S.-H. Woo, J. Kim, H. Jeong, and S.-O. Jung, "Comparative study of various latch-type sense amplifiers," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 2, pp. 425–429, Feb. 2014.

[76] G. Kiene, A. G. Sreenivasulu, R. W. J. Overwater, M. Babaie, and F. Sebastiano, "Cryogenic comparator characterization and modeling for a cryo-CMOS 7b 1-GSa/s SAR ADC," in *Proc. IEEE 48th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2022, pp. 53–56.

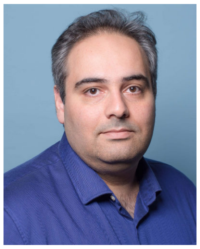
[77] Y. Shu, H. Zhang, H. Sun, Q. Deng, and Y. Ha, "CSDB-eDRAM: A 16 Kb energy-efficient 4T CSDB gain cell eDRAM with over 16.6s retention time and 49.23uW/kb at 4.2K for cryogenic computing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2023, pp. 1–5.

Rob A. Damsteegt received the B.Sc. and M.Sc. degrees (cum laude) in electrical engineering and the M.Sc. degree (cum laude) in computer engineering from Delft University of Technology, Delft, The Netherlands, in 2019 and 2022, respectively, where he is currently pursuing the Ph.D. degree in cryogenic electrical engineering.

His research interests include mixed-signal design, high-performance computing, novel computing paradigms, and accelerator design.

Ramon W. J. Overwater received the B.S. degree in electrical engineering, the M.S. degree (cum laude) in micro-electronics, and the M.S. degree in computer engineering from 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.

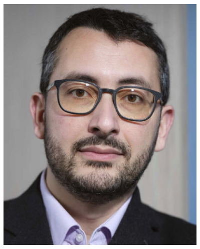
His research interests include cryogenic electronic characterization, mixed-signal design, and high-performance computing.



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 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 Berkeley Wireless Research Center, Berkeley, CA, USA. In 2016, he joined Delft University of Technology, where he is currently an Associate Professor. He has authored or coauthored one book, three book chapters, and more than 100 peer-reviewed technical articles, and holds 11 patents. 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 International Solid-State Circuits Conference (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 on the Technical Program Committee of the IEEE ISSCC. He is currently serving as an Associate Editor for IEEE SOLID-STATE CIRCUITS LETTERS (SSC-L).



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 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 Delft University of Technology, where he is currently an Associate Professor. He has authored or coauthored one book and over 100 technical publications, and holds 11 patents. His main research interests are cryogenic electronics, quantum computing, sensor read-outs, and fully integrated frequency references.

Dr. Sebastiano 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 is on the Technical Program Committee of the ISSCC and the IEEE RFIC Symposium, and has been on the Program Committee of IMS. He is currently serving as an Associate Editor for IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS and IEEE JOURNAL OF SOLID-STATE CIRCUITS (JSSC) and has also served as a Guest Editor for IEEE JSSC. He has served as a Distinguished Lecturer for the IEEE Solid-State Circuit Society.