A Benchmark of Cryo-CMOS 40-nm Embedded SRAM/DRAMs for Quantum Computing

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

DOI
10.1109/ESSCIRC59616.2023.10268788

Publication date
2023

Document Version
Final published version

Published in
ESSCIRC 2023 - IEEE 49th European Solid State Circuits Conference

A Benchmark of Cryo-CMOS 40-nm Embedded SRAM/DRAMs for Quantum Computing

Rob A. Damsteegt*, Ramon W.J. Overwater*, Masoud Babaie†, and Fabio Sebastiano*
*Department of Quantum & Computing Engineering & QuTech, TU Delft, The Netherlands
†Department of Microelectronics & QuTech, TU Delft, The Netherlands
Email: r.a.damsteegt@tudelft.nl

ESSCIRC 2023- IEEE 49th European Solid State Circuits Conference (ESSCIRC) | 979-8-3503-0420-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ESSCIRC59616.2023.10268788

***Abstract*—The cryogenic electronic interface for quantum processors requires cryo-CMOS embedded memories that cover a wide range of specifications. The temperature dependence of device parameters, such as the threshold voltage, the gate/subthreshold leakage, and the variability, severely alters the memories' performance between room temperature (RT) and cryogenic temperatures (4.2 K). To assess the best memory design for a given application, this paper benchmarks three custom DRAMs and a custom SRAM in 40-nm CMOS at 4.2 K and RT, e.g., identifying that, while the SRAM is more power efficient for moderate-to-high speeds at RT, the 2T DRAM performs better than SRAM and 3T DRAMs at 4.2 K.**

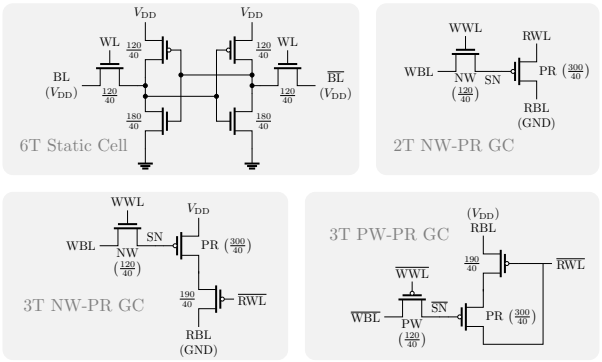
***Index Terms*—Cryo-CMOS, eDRAM, SRAM, DRAM.**

I. Introduction

Quantum computers (QCs) promise to solve some computational problems intractable by classical computers. However, scaling up QCs to the thousands (or more) of quantum bits (qubits) needed for such computations requires excessively many wires connecting the cryogenic qubits to their room-temperature control electronics. To mitigate this, cryogenic CMOS (cryo-CMOS) control interfaces have been proposed [1]. Such cryo-CMOS controllers will require memories for various functions, from high-speed lookup tables (multi-GHz, W/R = 0) to low-speed I/O buffer queues (sub-MHz, W/R = 1), covering a wide range of access rates (read and write operations per second) and write/read (W/R) ratios.

Static memories (SRAMs) are typically adopted for high-access-rate applications since they offer high speeds and can be embedded in a logic process. However, they suffer from excessive power consumption 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 static power consumption. At cryogenic temperatures, however, this effect is strongly mitigated by the significant decrease in subthreshold leakage [2]–[6]. Still, it is unclear whether a cryo-CMOS DRAM can outperform a cryo-CMOS SRAM, and for which applications since the lack of reliable cryogenic device models and the large variety of memory architectures and CMOS processes adopted in prior works hinder the compilation of a fair comparison.

To fill this gap, this work benchmarks 40-nm CMOS SRAM and DRAM embedded memories (Fig. 1) optimized for cryogenic operation. The whole application space is explored by comparing them over a wide range of access rates (1 kHz to 1 GHz) and W/R ratios (0 to 1). The most efficient memory for each use case is then identified, by assessing their retention time, power/energy consumption, and latency.



(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, $\overline{\text{(W)BL}}/\overline{\text{SN}}$: low when data = 1
GC RBLs: low after reading data = 1, high after reading data = 0

Fig. 1. Schematics of the proposed memory cells, with the signal polarity of wordlines (WLs) and bitlines (BLs) and the $\frac{W}{L}$ of the transistors in nm.

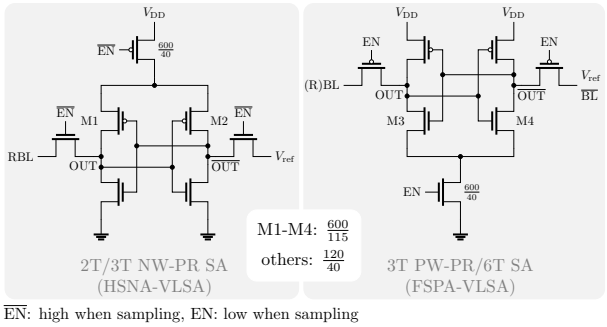
In the following, Section II presents the characteristics of the cryo-CMOS effects technology, and Section III describes the circuit design. The measurements are presented and discussed in Section IV. Conclusions are drawn in Section V.

II. Cryo-CMOS device behavior

At cryogenic temperatures, digital logic speeds up due to the 50 % mobility increase [7], the lower interconnect resistance [8], and the lower source/drain junction capacitance [3], but is limited by the threshold voltage increase (up to 150 mV [7]). Also, the subthreshold leakage is reduced due to the steeper subthreshold slope [9], and the thermal noise decreases to the level that shot noise may dominate in the sense amplifiers [10]. Finally, as matching deteriorates [7], there are more variations between cells and a larger sense amplifier offset.

III. Memory design

The dynamic gain cells (GC) (Fig. 1) store charge on the storage node (SN), which is translated into a current during readout. A two-transistor (2T) hybrid (N Write - P Read, NW-PR) GC cell has been implemented for the highest density [2],



$\overline{\text{EN}}$: high when sampling, EN: low when sampling

Fig. 2. Sense amplifiers for cells with charging RBLs (left) and discharging RBLs (right) with $\frac{W}{L}$ indicated in nm.

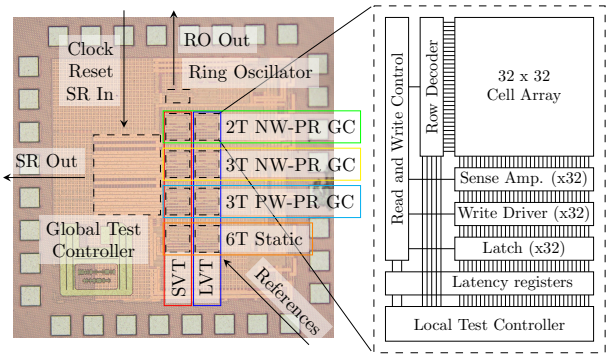


Fig. 3. Die micrograph and architecture of each memory.

[5], [6]. NW is minimum-sized to minimize subthreshold leakage, while PR is slightly larger to increase the SN capacitance and the readout current, only moderately increasing the area. Unfortunately, due to PR gate-source capacitance coupling, SN is pulled up during readout, limiting the achievable overdrive. Since the capacitance is largest when SN is low, the difference between the SN voltage levels is decreased during readout. To solve this issue, a faster three-transistor (3T) hybrid (NW-PR) GC [5] is also implemented, in which the source of the readout transistor is fixed to V_{DD}. NW and PR are sized as in the 2T cell, and the additional transistor width is maximized without increasing the cell area. The gate-source coupling can even be used to improve the SN levels in the PMOS-only (P Write - P Read, PW-PR) 3T cell [11]. Due to the coupling, SN is pulled down during readout. Since the capacitance is again largest when SN is low, the difference between the SN voltage levels is increased, making the levels easier to distinguish. Finally, a custom static 6T cell is implemented complying to standard logic layout rules for an equal comparison. Each cell is implemented using low threshold voltage (LVT) devices to counter the expected increase of the threshold voltage at cryogenic temperatures, and standard threshold voltage (SVT) as a reference.

The peripheral design is re-used as much as possible among the different memories to highlight the impact of the memory cells and uses LVT devices to ensure functionality at 4.2 K.

The adopted sense amplifiers (Fig. 2) consist of a cross-coupled-inverter latch that can be disconnected from the supply by a header/footer when sampling. The read bitline (RBL) and reference voltages are sampled by pass transistors, and their difference is amplified by the latch. The adopted sense amplifiers are small and efficient, but their offset is sensitive to the systematic difference in load capacitance and the mismatch of M1-M2/M3-M4 threshold, thus requiring larger M1-M4.

The control signals are generated using programmable delay chains to individually tune the duration of the sense amplifier's reset and the read/write operations. Since such tunability is only used for the characterization and is not required in a real application, this circuit is not optimized for power consumption and uses a separate supply.

The memories are connected to a local controller to perform any read/write/refresh sequence and store the settings of the programmable delay chains. Such sequences are sent over a shared bus by a global controller running programs to generate the desired address and data patterns. The delays in the control signal generation are determined by externally monitoring the frequency of a ring oscillator made out of a matching programmable delay chain.

The memories are compared based on latency, energy consumption, and retention time. The latency is defined as the time from the read trigger rising edge until the latch data are valid. The read trigger is launched by a clock-triggered register, while the latch outputs are captured by a register connected to a delayed clock. The delay between the two clocks is reduced until the data are incorrect. The energy per read/write operation or full refresh is determined by measuring the power of programs reading and writing pseudo-random data to pseudo-random addresses and normalizing it to energy per operation. The retention time is measured by increasing the delay between successive write and read operations until the read data do not correspond to the written data.

IV. Measurement Results

The 1-mm^2 TSMC 40-nm CMOS test chip (Fig. 3) is mounted on a dipstick. Its temperature is measured by a Lake Shore DT-670 temperature sensor and swept by varying the height of the dipstick in either liquid helium or helium vapors in a dewar. The global test controller is programmed at 200 kHz through a shift register (SR) driven by an opto-coupled FPGA at RT. The reference voltage for the sense amplifiers and the memory supply rails are supplied by a R&S HMC8043 supply and a Keithley 2636B SMU, respectively.

Table I shows that the sense amplifier's offset mean changes slightly over temperature due to the temperature dependence of device capacitances. To avoid a deterioration of the retention time, the reference voltage is then independently optimized for RT and 4.2 K (except for temperature sweeps in Figs. 4 and 5). The offset spread increases by roughly 7 %, which corresponds to the expected increase in the Pelgrom-law scaling factor for the threshold voltage (A_{VT}) for similar-length transistors [7]. This may also cause a reduced retention time due to a larger

TABLE I
INPUT-REFERRED OFFSET AND NOISE OF 64 SENSE AMPLIFIERS

Cell type	Offset: μ, σ [mV] RT	4.2 K	Mean Noise [mV_{RMS}] RT	4.2 K
2T NW-PR	32.6, 10.8	40.7, 11.7	2.44	0.677
3T NW-PR	35.2, 17.8	43.2, 18.9	2.11	0.644
3T PW-PR	−38.6, 15.3	−47.2, 16.3	1.89	0.584

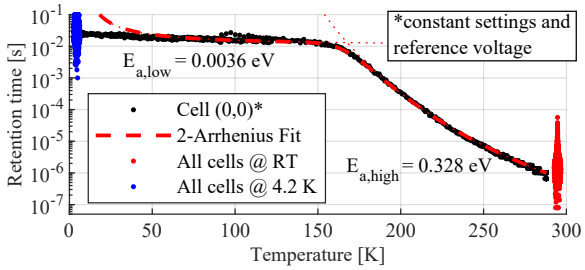


Fig. 4. 2T NW-PR (LVT) retention time (limited by high SN voltage state) for a single cell over temperature and for all cells at RT and 4.2 K. Measurements saturate to 100 ms to limit the total characterization time.

apparent cell variation between columns. The sense amplifier's input-referred noise only decreases by approximately 70 %, likely due to the shot noise dominating the performance at cryogenic temperature [10]. This should slightly improve the bit error rate and retention time since the difference between the RBL and reference voltage can be smaller.

Fig. 4 shows the retention time t_{ret} of all 2T NW-PR (LVT) cells at RT and 4.2 K, and a typical cell's t_{ret} over temperature exhibiting a high-temperature region ($>$ 160 K) and a low-temperature region ($<$ 160 K). By assuming t_{ret} is inversely proportional to the total leakage current I_{leak} and modeling the leakage currents with the Arrhenius equation above 50 K in the two regions, we get:

$$\frac{1}{t_{\text{ret}}} \propto I_{\text{leak}} = Ae^{-\frac{E_{a,\text{high}}}{k_b T}} + Be^{-\frac{E_{a,\text{low}}}{k_b T}} \quad (1)$$

where E_a is the activation energy, k_b is the Boltzmann constant, and T is the absolute temperature. The high-temperature activation energy matches well with that of the subthreshold leakage current through the write transistor $E_{a,subth} = \ln(10) k_b \frac{V_{th}(0)}{s_0}$, with s_0 the linearized subthreshold slope temperature dependence [4]. Below 160 K, the expected temperature-independent gate leakage of the readout transistor likely dominates, which matches with the behavior in Fig. 4.

Fig. 5 shows the retention time of single cells of the 3T memories for storing both a high or a low voltage on SN, which do not follow the pattern from Fig. 4. First, the PR of the 3T NW-PR cell is in strong inversion for low SN voltages. This results in a large gate leakage limiting the retention time over the entire temperature range. Additionally, the threshold voltage increase of PR even reduces the retention time over temperature for all 3T cell types due to a reduction in overdrive that results in a smaller readout current. Finally, the PW-PR cells also suffer from the threshold-voltage increase of PW,

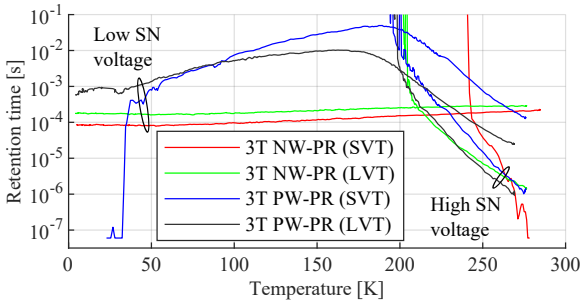


Fig. 5. Retention time over temperature for a single cell of each 3T memory cell design for high- and low-SN voltage states for constant settings and V_{REF}.

TABLE II
LOG-NORMAL μ AND σ AND WORST CELL RETENTION TIME DISTRIBUTION WITH OPTIMIZED SETTINGS AND REFERENCE VOLTAGES

Cell type	Retention time [s] Log-normal μ, σ RT	4.2 K	Worst cell retention time [µs] RT	4.2 K
2T NW-PR (SVT)	-11.2, 1.00	-2.01, 1.58	0 (2×)	0 (11×)
2T NW-PR (LVT)	-13.4, 1.03	-2.89, 1.16	0 (16×)	459
3T NW-PR (SVT)	-11.7, 1.07	-8.77, 0.39	0.14	18
3T NW-PR (LVT)	-14.8, 0.92	-8.92, 0.39	0 (71×)	9.38
3T PW-PR (SVT)	-10.4, 0.70	N/A*	2.2	0 (889×)
3T PW-PR (LVT)	-12.2, 0.77	-6.56, 2.23	0.44	0 (253×)

*: Too many failing cells for meaningful distribution.

which increases the minimum voltage that can be written to SN. This shows that techniques improving the retention time at RT do not necessarily work at 4.2 K, and can be counterproductive.

The retention time of all cells of a memory follows a log-normal distribution with a slight low-side tail, resulting in lower worst-cell retention times than expected (Table II). At RT, the SVT designs are better than the LVT designs thanks to the lower leakage. As intended (Section III), the 3T PW-PR is better than the 3T NW-PR, which, in turn, is better than the 2T NW-PR. However, at 4.2 K, most LVT designs outperform the SVT designs, since the subthreshold leakage is no longer significant. The μ and σ of the 2T design increase due to reduced leakage and increased mismatch. As expected from Fig. 5, the 3T NW-PRs gain significantly less retention time due to the large gate leakage and the 3T PW-PRs become worse due to the PW threshold-voltage increase.

Table III shows the energy consumption per operation and full refresh cycle, the leakage power, and the latency at RT and 4.2 K. A range is reported for each metric due to different delay chain settings and optimized reference voltages. Overall, the energy per operation at 4.2 K is lower than at RT. This is attributed to the decrease in source/drain junction capacitance and increase in threshold voltage resulting in less gate charge. The leakage power reduces significantly due to the reduction of the subthreshold leakage. Finally, the latency reduces due to the faster readout and faster logic.

The memory power consumption is compared versus access rate for W/R = 0 and W/R = 1 applications at RT and 4.2 K

TABLE III
CELL AREA, OPERATION ENERGY, LEAKAGE POWER, AND LATENCY OF ALL MEMORY DESIGNS AT 300 K AND 4.2 K

Cell type	Cell area [µm^2]	E_{read} [fJ/op] RT	E_{read} [fJ/op] 4.2 K	E_{write} [fJ/op] RT	E_{write} [fJ/op] 4.2 K	$E_{refresh}$ [pJ] RT	$E_{refresh}$ [pJ] 4.2 K	$P_{leakage}$ [µW] RT	$P_{leakage}$ [µW] 4.2 K	Latency [ns] RT	Latency [ns] 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.

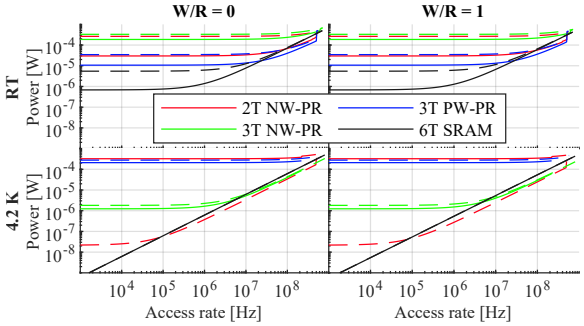


Fig. 6. Power consumption of each memory design (SVT: solid, LVT: dashed) versus access rate for W/R ratios of 0 (left) and 1 (right) at RT (top) and 4.2 K (bottom).

in Fig. 6. Using the energy per operation from Table III and a refresh frequency determined by the worst cell retention time (Table II), the total expected power consumption is computed. 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. For each memory, the graph 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. Looking for the most efficient design, the SVT SRAM is the best for low speeds at RT due to its comparatively low leakage power, while the 3T PW-PR (SVT) becomes better above 25 MHz thanks to its lower operation energy. At 4.2 K, the 2T NW-PR (LVT) outperforms the SRAM already beyond 75 kHz, also being the smallest and even 24 % smaller than the foundry SRAM cell.

V. CONCLUSIONS

The comparison of various embedded memories at RT and 4.2 K shows how traditional RT design paradigms break down at cryogenic temperatures. Different from RT, the best retention time is achieved at 4.2 K by the 2T DRAM thanks to the temperature dependence of subthreshold leakage and threshold voltage and the impact of gate leakage. Additionally, the variability between cells and in sense amplifiers increases at 4.2 K, likely resulting in more outlier cells and a wider retention time distribution, while the lower noise may improve the bit error rate. Thanks to the lower operation energy and the lower refresh energy due to the lower leakage, the 2T DRAM cells are more power efficient than the static cells at 4.2 K for moderate-to-high speeds, thus representing the best alternative for high-density low-power embedded memories in cryo-CMOS processors and interfaces.

ACKNOWLEDGMENT

The authors would like to thank Intel corporation for funding and Atef Akhnoukh for technical support.

REFERENCES

[1] B. Patra *et al.*, "Cryo-CMOS circuits and systems for quantum computing applications," *IEEE JSSC*, vol. 53, no. 1, pp. 309–321, 2017.
[2] 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, 1990.
[3] N. Yoshikawa *et al.*, "Characterization of 4 K CMOS Devices and Circuits for Hybrid Josephson-CMOS Systems," *IEEE Trans. Appl. Supercond.*, vol. 15, no. 2, pp. 267–271, 2005.
[4] F. Wang *et al.*, "DRAM Retention at Cryogenic Temperatures," in *2018 IEEE International Memory Workshop (IMW)*, 2018, pp. 1–4.
[5] E. Garzón *et al.*, "Gain-Cell Embedded DRAM Under Cryogenic Operation—A First Study," *IEEE Trans. VLSI Syst.*, vol. 29, no. 7, pp. 1319–1324, 2021.
[6] R. Saligram *et al.*, "CryoMem: A 4K-300K 1.3 GHz eDRAM macro with hybrid 2T-gain-cell in a 28nm logic process for cryogenic applications," in *CICC*, 2021, pp. 1–2.
[7] P. A. 't Hart *et al.*, "Characterization and Modeling of Mismatch in Cryo-CMOS," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 263–273, 2020.
[8] R. Saligram *et al.*, "Scaled Back End of Line Interconnects at Cryogenic Temperatures," *IEEE EDL*, vol. 42, no. 11, pp. 1674–1677, 2021.
[9] R. M. Incandela *et al.*, "Characterization and Compact Modeling of Nanometer CMOS Transistors at Deep-Cryogenic Temperatures," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 996–1006, 2018.
[10] G. Kiene *et al.*, "Cryogenic Comparator Characterization and Modeling for a Cryo-CMOS 7b 1-GSa/s SAR ADC," in *ESSCIRC 2022*, pp. 53–56.
[11] K. C. Chun *et al.*, "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, 2011.