Performance Benchmarking of Cryo-CMOS Embedded SRAM and DRAM in 40nm CMOS Technology

Shaik Abdul Khayum[1], G Mahendra[2]
[1]*PG Student, QUBA College of engineering and technology*
[2]*Assistant Professor, QUBA College of engineering and technology*

Abstract: Cryogenic Complementary Metal-Oxide-Semiconductor (Cryo-CMOS) technology has recently gained significant attention as an enabling platform for quantum computing and ultra-low temperature electronic systems. Embedded memory, particularly Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM), forms a critical component in such systems, where performance, energy efficiency, and reliability at cryogenic temperatures are of paramount importance. This work presents a comprehensive performance benchmarking of embedded SRAM and DRAM designed in a 40nm CMOS process for cryogenic operation. Key performance metrics such as access latency, read/write stability, leakage power, retention time, and energy-delay product are systematically evaluated across temperature scaling from room temperature down to cryogenic levels. The benchmarking framework highlights the trade-offs between SRAM and DRAM under extreme conditions, emphasizing the design challenges posed by device variability, leakage suppression, and refresh mechanisms. The results demonstrate that SRAM exhibits superior read/write speed at cryogenic temperatures, while DRAM offers improved density and scalability but suffers from reduced retention without specialized refresh strategies. This analysis provides critical insights into the feasibility of integrating embedded memories in Cryo-CMOS platforms, paving the way for future quantum-classical co-processing architectures.

***Index Terms*— Cryo-CMOS, 40nm CMOS technology, embedded memory, SRAM, DRAM, cryogenic operation, performance benchmarking, memory stability, leakage power, retention time, quantum computing, energy-delay product**

I. INTRODUCTION

QUANTUM computers (QCs) can deliver an exponential speedup for several computational problems [1],[2]. 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 [3], [4]. 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 [5] and processing a large amount of measurement results, e.g., as required for quantum error correction [6]. 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) [9] to low-speed buffer queues for the quantum-algorithm instructions (sub-MHz, W /R = 1) [5]. 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 [12], [13], 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.

1.1 CRYO-CMOS DEVICE BEHAVIOUR Cooling down to cryogenic temperatures affects the characteristics of short-channel NMOS and PMOS transistors by increasing their threshold voltage Vth (100–200 mV), subthreshold slope (~3× steeper), and carrier mobility (~2× for low-field mobility) [34],[35], Additionally, the mismatch between devices increases, as shown in [66] and [67] for 40-nm bulk CMOS and 28-nm bulk CMOS, respectively, interconnect resistance drops (~30%) [38], and the capacitance of source/drain junctions decreases due to wider depletion regions due to freeze-out [13]. For analog circuits, this results in an increased bandwidth and reduced power consumption.

Applications and uses

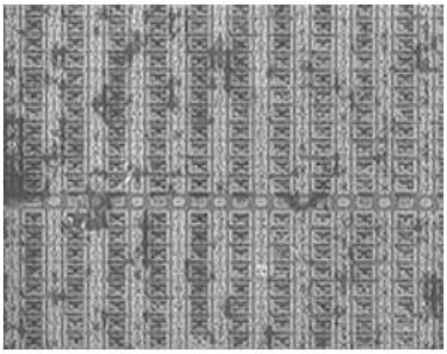


Fig 1.1: SRAM Seen by SEM using 180 nm

Fig 1.2: SRAM Seen by Optical Microscope using 180 nm

Operational Read Failure

Since a read is typically the slowest memory operation, its timing is the most vulnerable to failure [35]. During a read operation, the amount of differential voltage generated on the bit lines is directly proportional to two parameters: the width of the word line signal.

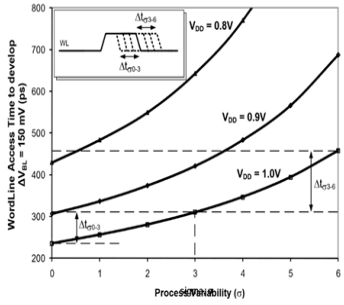


Figure 1.17: Effect of process and voltage variations on required cell access time and the strength of the SRAM cell.

This would cover 99.73% of the variability cases. A flexible timing scheme would have three benefits. It could increase the yield by providing extra time for the read operation to complete in the cases of variability beyond 3 σ. For the majority of dies whose variability is less than 3 σ, a flexible timing scheme would create more optimal timing signals, allowing those dies to be operated at with a higher DNM and reduced power dissipation because the cell is being accessed for a short period of time.

Moreover, the supply voltage can be reduced, while still maintaining a guard band of a given number of σ. Additionally, a fabricated array will have an unknown amount of variability. By using flexible timing, the edges of the control signals can be moved to not only correct failures, but also to characterize the array's variability. By starting with the most aggressive timing setting, and relaxing that timing until the SRAM performs correctly, or vice versa, with the most relaxed timing, and pushing the timing until failure, the residual difference between nominal timing setting and those of the chip-under-test can be characterized. This can lead to "binning" of chips based on their amount of variability.

Cell Stability Failure

In an SRAM array containing multiple words per row, a cell is said to be half-selected when it is accessed via the word line, but its bit lines are not routed to the sense amplifier. In the case of a half-selected cell, the

dynamic noise margin is determined by the width of the wordline access time window. Cells weakened due to process variation and aging experience a lower DNM. To illustrate this response, simulations were performed on a 6T SRAM cell in a 65 nm standard CMOS process. Resistors are used to symmetrically weaken the cell, as shown in Figure 3.2. If the resistance is relatively low, it models the effects of process variability,

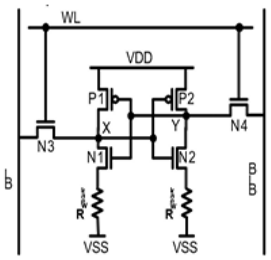


Figure 1.18: Schematic of a Weak 6T SRAM Cell

When the value of Rweak is low, or when the access time is low, the cell is stable; however, if the resistance is large enough, and the access time is sufficiently long, the cell can become unstable. This behavior shows a strong dependence on the supply voltage. For example, a weakened SRAM cell with Rweak = 10 kΩ is stable with a supply voltage of 1 V. If the supply voltage is reduced to 0.7 V however, the width of the word line signal must be kept to less than 100 ps or else the cell will become unstable. These results are similar to those of Sharif Kani and Sachdev [39]. In their work, they show measured results that illustrate the relationship between cell stability and access time, as can be seen in Figure 3.4. Care must be taken when designing the timing for the SRAM array so that enough time is available for the selected cells to develop the required differential voltage on the bit lines for the sense amplifier to resolve the data; however, not so much time as to upset the half-selected cells.

II SYSTEM ANALYSIS

2.1 LITERATURE SURVEY

B. Patra et.al. described about the design and experimental validation of several major circuit blocks critical for the implementation of a CMOS classical electronic controller to operate at cryogenic temperatures (i.e., cryo-CMOS) in order to interface with a practical quantum processor, a 4-K 160-nm LNA capable of reading out 150 1-MHz qubit channels with a power efficiency better than 700 μ W/qubit and a 4-K 40-nm 6-GHz class-F2,3 oscillator with an integrated FN of 3.4 kHzrms over ~10 MHz bandwidth, which is low enough to drive the state-of-the-art qubits without limiting their performance. Such performance is achieved by carefully employing standard circuit design techniques while exploiting specific characteristics of the adopted cryo-CMOS devices, such as increased speed and low thermal noise. In this paper further effort is required to develop a full cryogenic controller demonstrating the required performance in the very tight power budget set by existing dilution refrigerators, the proposed circuits show that cryo-CMOS is a viable technology for the implementation of such classical electronic controllers, thus establishing cryo-CMOS circuits and systems as an enabling technology for the fabrication of practical quantum computers with thousands or even millions of qubits. [11] S. Chakraborty et al. described about a scalable, non-multiplexed cryogenic 14 nm FinFET quantum bit (qubit) state controller (QSC) for use in the semi-autonomous control of superconducting transmon qubit.The QSC includes an augmented general-purpose digital processor that supports waveform generation and phase rotation operations combined with a low-power current-mode single sideband upconversion I / Q mixer-based RF arbitrary waveform generator (AWG).

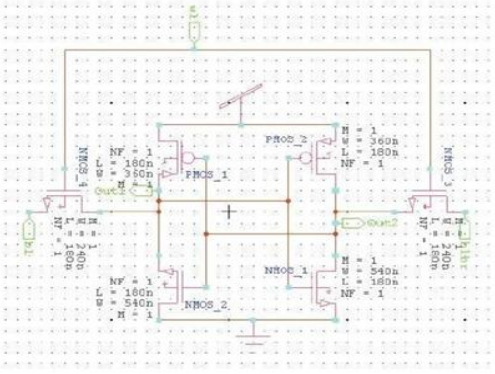


Figure 2.1: Circuit diagram of Existing System

From the literature survey the following are the drawbacks

1. It is proved that conventional 6 T fails to maintain its stability in scaled technology, particularly in deep-subthreshold regime.
2. The size of SRAM cell is about 90nm and 180nm technologies which is large

3. High Power Consumption
4. Complex to Design & Higher Manufacturing Cost
5. Low Storage Density

To overcome the above drawbacks A 4 Transistor based SRAM was proposed.

III PROPOSED SYSTEM

3.1. Introduction:

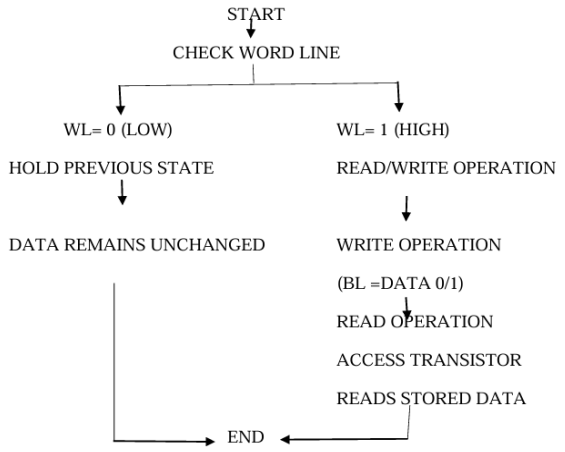
In the proposed system Memory Circuits Volatile and non-volatile memories are two major categories of CMOS-based memory. Volatile memories, such as static random-access memory (SRAM) and dynamic RAM (DRAM), exhibit improved performance metrics at cryogenic temperatures due to reduced leakage currents and enhanced carrier mobility. Additionally, the high compatibility of volatile memory with silicon-based CMOS processes facilitates achieving high levels of integration.

3.2. 4T SRAM Circuit operation:

Proposed 4T SRAM structure optimized for operation at 77 K by eliminating two PMOS transistors from the pull-up network. A 4T (Four-Transistor) SRAM cell is a type of SRAM that uses four transistors to store a single bit of data. Out of four transistors two transistors are used for storage and two were used to control read and write operations (access transistors). The 4T SRAM achieved a reduction of cell area by 20.3%, compared to the standard 6T SRAM structure. Moreover, it also provided faster read and write operations.

3.3. Flowchart:

Here's a simple flowchart for the 4T SRAM operation:



IV. RESULT ANALYSIS

Existing System results:

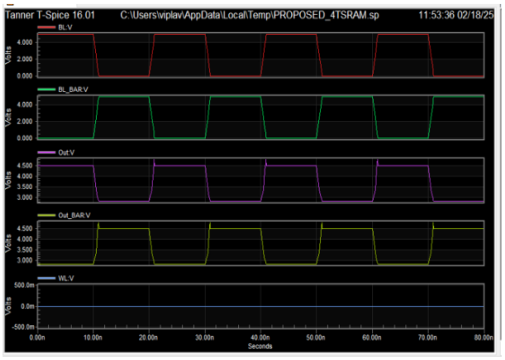


Figure 4.1: 6T SRAM wave forms

```
Power Results

VV2 from time 0 to 1e-008
Average power consumed -> 1.475947e+001 watts
Max power 1.475947e+001 at time 0
Min power 1.475947e+001 at time 0

VV3 from time 0 to 1e-008
Average power consumed -> 1.812614e-003 watts
Max power 1.812614e-003 at time 0
Min power 1.812614e-003 at time 0

VV4 from time 0 to 1e-008
Average power consumed -> 0.000000e+000 watts
Max power 0.000000e+000 at time 0
Min power 0.000000e+000 at time 0
```

Figure 4.2: 6T SRAM Power Results

S.NO	PARAMETER	EXISTING METHOD	PROPOSED METHOD
1	No. of Transistors	6	4
2	Area per cell	Larger due to usage of 6 transistors	Smaller due to usage of 4 transistors
3	CMOS Technology used	90nm and 180nm technologies	40nm technology
4	Data stored	Using cross-coupled transistors	Transistor latch
6	complexity	More complex	Moderate

Table 4.1: comparison between the existing and proposed method.

5. CONCLUSION

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 Vth 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. FUTURE SCOPE: In future, this project can be built by comparing the performance of 40nm cryo-CMOS using newer technology nodes such as 28nm, 22nm and 14nm which helps to develop more energy-efficient, reliable cryogenic memory architectures.

REFERENCE

[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] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.