CryoMem: A 4–300-K 1.3-GHz Hybrid 2T-Gain-Cell-Based eDRAM Macro in 28-nm Logic Process for Cryogenic Applications

Rakshith Saligram, Suman Datta, *Fellow, IEEE*, and Arijit Raychowdhury, *Senior Member, IEEE*

***Abstract*—This letter demonstrates the first CMOS logic compatible cryogenic memory solution operating from 4 to 300 K designed in the 28-nm high-K metal gate (HKMG) CMOS. With the growing applications of cryogenic systems from quantum computing to space electronics, there is a need for memory capable of reliable functionality. While the prevailing low-temperature memories suffer from temperature scalability and integrability, the proposed test chip of 1-kb 2T hybrid gain cell-based embedded DRAM macro overcomes these issues while providing $10^6\times$ better retention time, 1.3-GHz peak frequency at 4 K, sub nW/kb array refresh power, and 1.7× energy efficiency at 4 K compared to 300 K. This is due to the near absence of leakage, improved ON current, and subthreshold slope which leads to enhanced performance of critical path circuits at lower temperatures.**

***Index Terms*—Charge injection, cryogenic memory, data retention time, gain cell embedded DRAM (GC-eDRAM), quantum computing.**

I. Introduction

Cryogenic CMOS has gained traction for its ability to achieve higher performance and/or better power with process retargeted for cryo-high performance computing (cryo-HPC) [1], [2]. Besides, cryo-CMOS also presents as one of the most feasible solutions for interface and peripheral circuitry required to build scalable quantum and superconducting computers and is an integral part of space electronics and digital control in fuel cell electric vehicles [3]. This wide range of temperature from 10 mK to 300 K demands for a memory system that can reliably operate, easily integrate, and scale in capacity. Also, in order to bridge the processor–memory performance gap (memory wall) in cryo-HPC systems, the memory needs to have higher bandwidth (BW) and lower latency. This emphasizes the need to increase on-die memory which provides several orders of magnitude improvement in power and performance. In superconducting computers, though many memory technologies have been proposed [4]–[9], *viz.*, Josephson junctions, pi junctions, single quantum flux, etc., they suffer from low density, poor scalability, low reliability, higher design complexity, single operating temperature, integrability, and other issues. In this backdrop, we present the 2T gain-cell (GC)-based embedded DRAM (eDRAM) macro test chip in 28-nm high-K metal gate CMOS targeted for a range of aforementioned cryogenic applications [10]. While doing so, we discuss the CMOS transistor characteristics that aid the memory design in Section II, analyze the different GC topologies and obtain an overview of the eDRAM array architecture in Section III, and present various characterizations performed on the array in Section IV.

Manuscript received June 30, 2021; revised September 25, 2021; accepted October 25, 2021. Date of publication October 28, 2021; date of current version November 10, 2021. This work was supported in part by the Lab for Physical Sciences; in part by the Samsung GRO Program; in part by ORNL; and in part by SRC through ASCENT Center. This article was approved by Associate Editor Shreyas Sen. *(Corresponding author: Rakshith Saligram.)*

Rakshith Saligram and Arijit Raychowdhury are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: rakshith.saligram@gatech.edu).

Suman Datta is with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA.

Digital Object Identifier 10.1109/LSSC.2021.3123866

II. Transistor Characteristics and Models at Cryogenic Temperatures

The measured transistor output characteristics (for iso overdrive of 0.5 V) and transfer characteristics (linear and saturation region) for 28-nm bulk CMOS are shown in Fig. 1(a)–(c). The figures indicate that the current increases with a decrease in temperature caused by higher carrier mobility (μ) due to lower lattice vibrations and resulting in reduced phonon scattering. The threshold voltage (V_{th}) increases with decrease in temperature due to bandgap widening, the shift in Fermi potential, and incomplete ionization [11]. We see around 133 mV increase in the extracted V_{th} from 300 to 4 K for both nMOS and pMOS. The measured subthreshold leakage current decreases by a factor of more than 10^6 from 300 to 4 K shown in Fig. 1(d). The increase in the ON current along with an exponential decrease in the OFF current leads to linear improvement in subthreshold swing (SS) with a decrease in temperature shown in Fig. 1(d) (extracted) for both nMOS and pMOS in linear and saturation regimes. With these properties, bulk CMOS renders itself as a promising technology for low-temperature digital operation.

BSIM4 models are calibrated to the measured device data using the BSIMProPlus tool by varying key transistor parameters like V_{th}, μ, interface trap capacitance, drain–source resistances, etc., based on the MIT-Virtual Source Model [12] to obtain a closely fit model. The models here are temperature dependent meaning the parameters are tuned for each temperature point and fit for five key temperatures of 6, 30, 70, 150, and 300 K.

III. Gain Cell Topologies and eDRAM Array Architecture

A. 2T Gain Cells

GC eDRAM is a fully logic compatible memory technology that is used to circumvent the area and power scaling problem of conventional 6-T SRAM. 2T GCs comprise a write and a read transistor and the gate capacitance of the read transistor acts as the storage node. GCs typically use stored voltage to turn ON the transistors in the read-out path making it a nondestructive read operation and since the charge flow in the sense line is higher than that stored in the storage capacitor, the name GC [13]. Depending on the type of read (R) and write (W) transistors (nMOS/pMOS), we have four configurations: 1) nMOS only (NW-NR); 2) pMOS only (PW-PR); and 3,4) Hybrid (NW-PR/PW-NR).

Analysis of the pMOS-Only cell reveals that it suffers from charge injection from write word line (WWL) to the storage node during a $0 \rightarrow 1$ transition worsened by a subsequent charge injection during a $0 \rightarrow 1$ transition on read word line (RWL) elevating the stored "0" voltage level. Similarly in the nMOS-Only cells, the stored "1" voltage level is affected by the $1 \rightarrow 0$ transition on WWL followed by $1 \rightarrow 0$ transition on the RWL. In case of Hybrid PW-NR, the $1 \rightarrow 0$ transition on RWL affects stored 1 and $1 \rightarrow 0$ transition on WWL affects the stored 0. Note in the case of Hybrid GC, the two charge injections assist the stored node voltage readout. The amount of charge injection is a function of the device current and

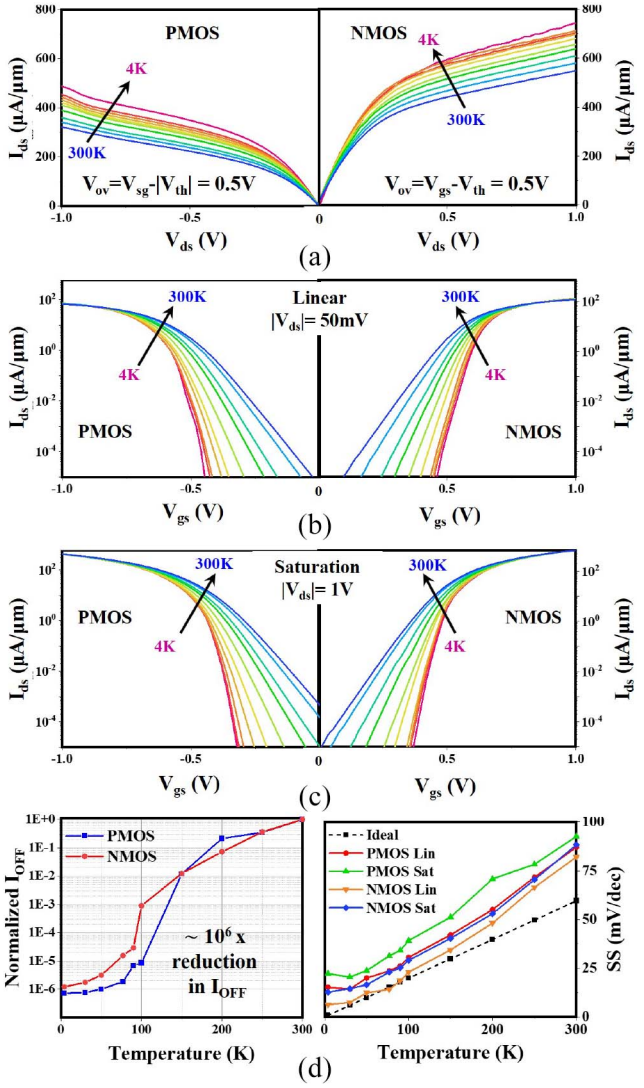
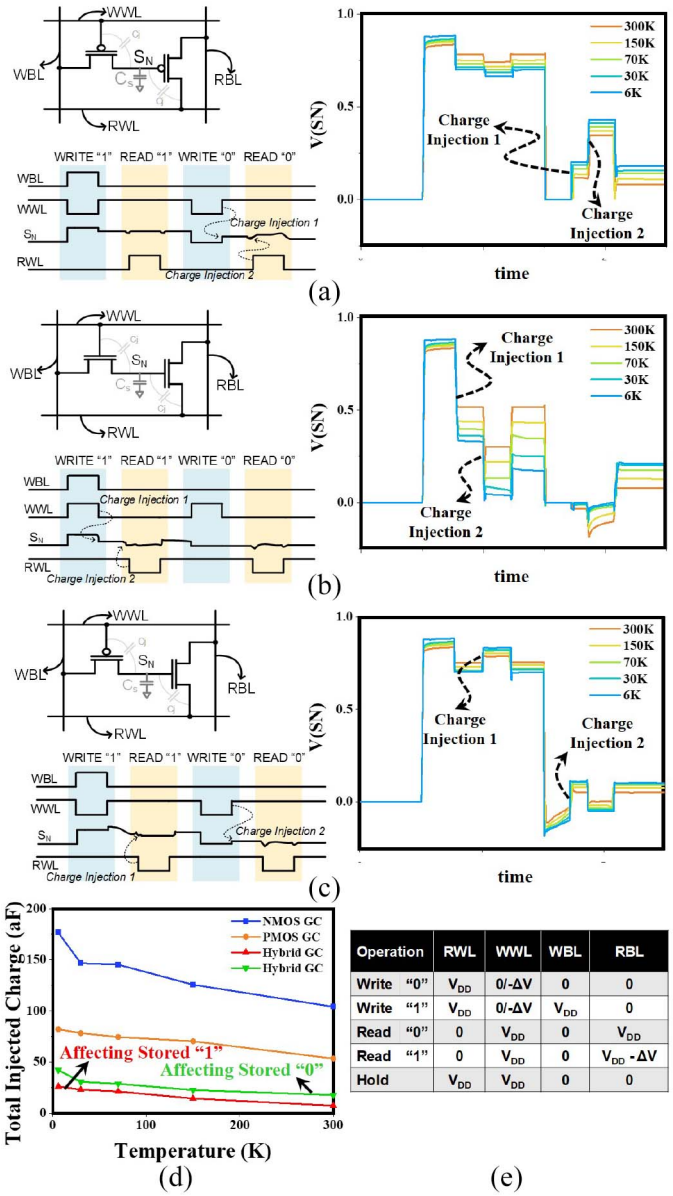


Fig. 1. Measured device characteristics showing (a) I_{DS}-V_{DS} for iso overdrive voltage of 0.5 V, (b) I_{DS}-V_{GS} for linear region, and (c) I_{DS}-V_{GS} for saturation region. (d) Left: Normalized OFF current (I_{OFF}); right: subthreshold slope (SS) variation with temperature.



Operation	RWL	WWL	WBL	RBL
Write "0"	V_{DD}	0/-ΔV	0	0
Write "1"	V_{DD}	0/-ΔV	V_{DD}	0
Read "0"	0	V_{DD}	0	V_{DD}
Read "1"	0	V_{DD}	0	V_{DD} - ΔV
Hold	V_{DD}	V_{DD}	0	0

Fig. 2. 2T GC topologies showing circuit, model timing diagram, and transients of storage node voltage and charge injections simulated with calibrated BSIM4 models across temperature for (a) pMOS-Only GC, (b) nMOS-Only GC, (c) Hybrid GC, (d) variation of total injected charge versus temperature, and (e) bit/word line voltages for different operations.

the parasitic capacitance and exacerbates at a lower temperature in scaled nodes due to higher current and sharper slews. This can be seen in the transients of the stored node voltage for the three structures simulated through the calibrated BSIM4 models (Fig. 2). The total injected charge (Q_{inj}) in case of hybrid GC indicates two curves one for each transition as they affect different stored voltages while that for nMOS only GC is for stored 1 and pMOS only GC is for stored 0. The Q_{inj} increases by 70% (29%) for nMOS-GC (pMOS-GC) going from 300 to 6 K. Thus, the macro in this work features the hybrid GC which despite larger in the area [13% (15%) higher than pMOS-GC (nMOS-GC)], provides higher read voltage margin due to lower charge injection, balanced P-N density, error-free operation, and relaxed physical design constraints.

B. GC eDRAM Array Architecture

The 2T hybrid GC-based eDRAM array architecture is shown in Fig. 3. It is a 1024-bit (1-kb) subarray arranged in 1 bank with 32-bit word length in an open bit line architecture. Individual words can be selected by an address decoder which decodes 5 bits of the address to 1 of 32 hot decoded word line outputs. Predecoding stages in the address decoder reduce the area. The array uses per column cross-coupled strong ARM-based latched comparators as sense amplifiers, with an external tunable analog reference voltage to decode the read bits into the read-out flip-flops. The sense amplifier is enabled by the SAE_B strobe signal applied to the gate of pMOS which will resolve the sense nodes to the rails. The bit lines must be precharged before the read operation and done using the precharge circuit. The strobe signal generator and write circuitry generate all the necessary control signals internally through a series of delay buffers and XOR gates starting from external clocks.

The key signals and their timing diagram are depicted in Fig. 3(f). First, the write address is provided followed by the write clock signal. This will internally generate the WBL and WWL signals to store the data into the chosen cell. In order to read the data, the read address needs to be issued followed by read clock which will generate the

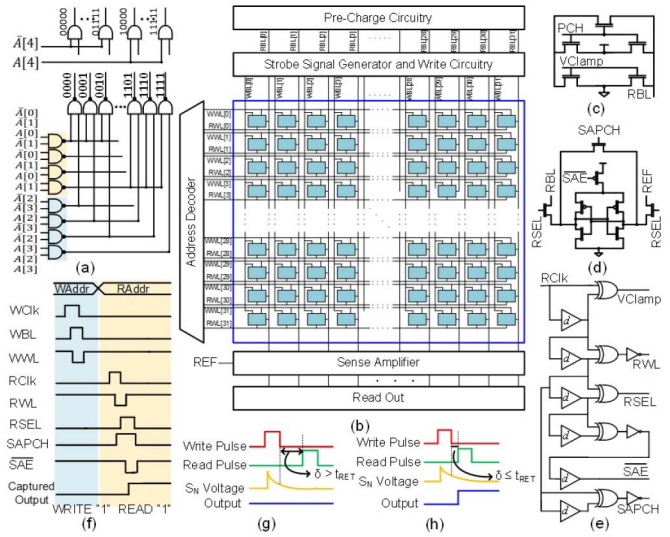


Fig. 3. 2T GC eDRAM array architecture with constituent components. (a) Address decoder. (b) Memory organization. (c) Precharge circuit. (d) Sense amplifier. (e) Strobe signal generator. (f) Sample timing diagram. (g) Read failure. (h) Read success.

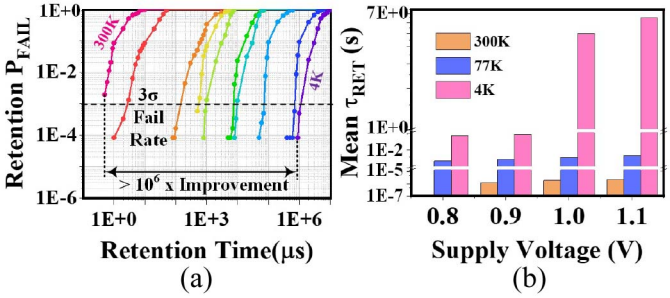


Fig. 4. Retention characterization showing (a) retention probability of failure versus retention time across temperature indicating $>10^6\times$ improvement from 300 to 4 K and (b) mean retention time (τ_{RET}) versus supply voltage for three key temperatures.

RWL, read select (RSEL), sense amplifier precharge (SAPCH), sense amplifier enable (SAE_B), and other signals to capture the output.

IV. Memory Array Characterization

A. Data Retention Time

The retention time is characterized by the capacitance of the storage node and the leakage currents of the cell. Based on the worst case retention time and the target yield, the data need to be refreshed periodically to prevent incorrect reads. The retention time of the eDRAM cell can be determined by the relative phase of the write and the read clocks both of which are controlled externally. If the read pulse is issued after the retention time of the cell, the output is not captured correctly as the cell does not carry sufficient charge to be interpreted as logic 1 by the sense amplifier (read failure) [Fig. 3(g)]. When the relative phase between the pulses is less than or equal to the retention time of the eDRAM GC, a correct value is captured at the output [Fig. 3(h)]. The retention time can thus be determined by measuring the maximum phase difference between the two external clocks.

The linear scaling of subthreshold slope with decreasing temperature results in an exponential decrease in the leakage current and therefore provides higher retention time. The retention time is characterized over 12 eDRAM subarrays (12 kb) at $V_{DD} = 1.0$ V across temperature and the retention probability of failure is plotted

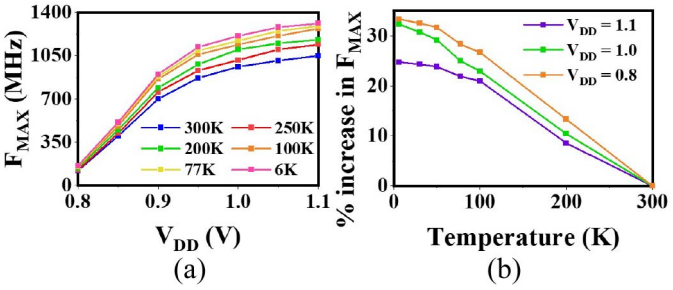


Fig. 5. Performance characterization showing (a) maximum frequency (F_{MAX}) versus supply voltage across temperature and (b) percentage improvement in F_{MAX} versus temperature for three supply voltages.

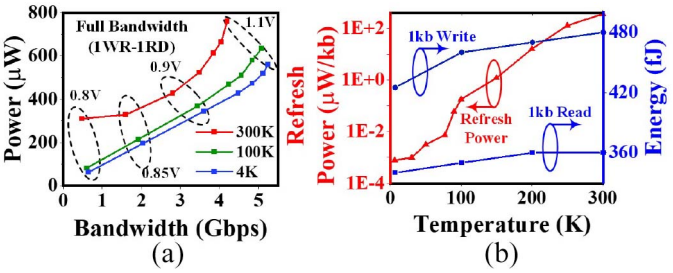


Fig. 6. Power characterization showing (a) power versus BW across three temperatures for full BW (1WR-1RD) and (b) refresh power and read/write energies for 1-kb memory across temperature.

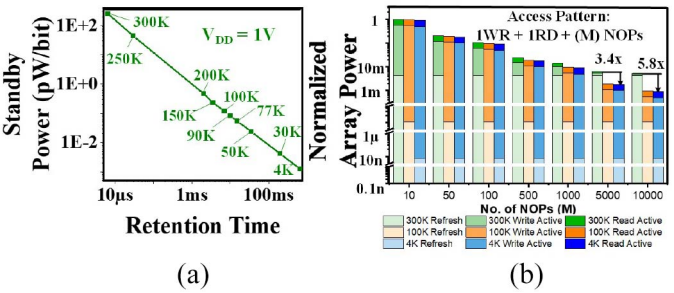


Fig. 7. (a) Standby power versus retention time across temperature. (b) Normalized array power for different access patterns.

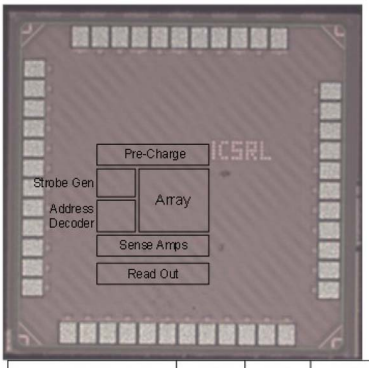
[Fig. 4(a)]. The median retention time improves from 2.4 μs (300 K) to 6.5 s (4 K) demonstrating a 2.7×10^6 improvement while the 3σ fail rate improves by an equivalent amount. The retention time statistics calculated across V_{DD} [Fig. 4(b)] shows super-linear scaling with increasing V_{DD} across temperature.

B. Performance

The eDRAM array is characterized at multiple temperature points. The maximum frequency of operation (F_{MAX}) at different supply voltages is shown in Fig. 5(a). A peak F_{MAX} of 1.3 GHz is measured at 4 K. The percentage increase in F_{MAX} from 300 to 4 K varies between 24.7% ($V_{DD} = 1.1$ V) and 33.3% ($V_{DD} = 0.8$ V) [Fig. 5(b)].

C. Power

The array is tested under full BW conditions with back-to-back read/write (RD/RW) cycles across temperature [Fig. 6(a)]. A peak array BW of 4.2 Gb/s at 761 μW/kB is measured at 300 K and it improves to 5.24 Gb/s at 560 μW/kB at 4 K yielding a net energy efficiency improvement (in terms of Gb/s/W) of 1.7$\times$. The array refresh power [Fig. 6(b)] reduces to <1 nW/kb at 4 K while the



Technology	28nm HKMG CMOS
Flavor	HPC+
Metal Stack	1P-9M
Die Area	1mm²
Core Voltage	0.9V
IO Voltage	1.8V
Word Size	32 bits
Cell Area	0.5704μm²
Banks	1
Memory Size	1kb

***VT: Vortex Transition Memory Device**
JJ: Josephson Junction
MRAM: Magnetic RAM
nTron: Nanocryotronics

	[4]	[5]	[6]	[7]	[8]	[9]	This Work
Technology	VT Memory Cells	JJ & Toggle MRAM	VT Memory Cells	JJ & CMOS	JJ, nTron & CMOS	JJ & CMOS	CMOS
CMOS Node	--	--	--	65nm	65nm	180nm	28nm
Memory Cell Type				SRAM	DRAM	SRAM	2T Gain Cell eDRAM
Cell Area (μm²)	99	2	3025	3.36	--	--	0.57
Read/Write Access Times	--	300ps/ 2000ps	800ps/--	430ps/ 300ps	660ps/--	1340ps/--	763ps
Dynamic Power		<100μW	9.5mW	12mW Read/ 21mW Write	0.78mW Read/ 2.2mW Write	28mW Read/ 41mW Write	0.76mW (300K) 0.56mW (6K)
Operating Temperature	4K						4K-300K
Integration with Si Logic	Requires Heterogeneous Integration						Monolithic

Fig. 8. Test chip die shot, chip characteristics, and comparison table.

RD and WR energies are measured at 360 fJ/kB (340 fJ/kB) and 480 fJ/kB (425 fJ/kB) at 300 K(4 K), respectively.

The characterization of standby power (including refresh power and leakage from peripherals) is performed across mean retention time across multiple temperatures [Fig. 7(a)]. Memory arrays particularly used as scratch pad or cache are seldom used at full-BW. We measure the macro array power as a function of the activity factor [Fig. 7(b)]. Every WR and RD operation is followed by M No-ops and the corresponding array power is measured. At higher values of M, most of the array power is consumed in refresh operations at 300 K and a 5.8× improvement in array power is noted at 4 K.

V. Conclusion

The GC macro is compared against existing prototypes for low-temperature memory (Fig. 8), although no temperature scalable and monolithically integrable CMOS solution is noted. Existing technologies [2]–[7] (typically used for superconducting quantum computers) have been shown to operate at 4 K and consume a large array power at low density/capacity using non-CMOS and hybrid processes. By comparison, we demonstrate a 4–300-K 2T-GC-based macro with high energy efficiency, 5.24 Gb/s of BW operating at 1.3 GHz (at 4 K) on a 28-nm CMOS logic technology process.

References

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

[2] R. Saligram, D. Prasad, D. Pietromonaco, A. Raychowdhury, and B. Cline, "A 64-bit arm CPU at cryogenic temperatures: Design technology co-optimization for power and performance," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, 2021, pp. 1–2.

[3] J. X. Jin, X. Y. Chen, L. Wen, S. C. Wang, and Y. Xin, "Cryogenic power conversion for SMES application in a liquid hydrogen powered fuel cell electric vehicle," *IEEE Trans. Appl. Supercond.*, vol. 25, no. 1, Feb. 2015, Art. no. 5700111.

[4] V. K. Semenov, Y. A. Polyakov, and S. K. Tolpygo, "Very large scale integration of josephson-junction-based superconductor random access memories," *IEEE Trans. Appl. Supercond.*, vol. 29, no. 5, Aug. 2019, Art. no. 1302809.

[5] J.-B. Yau, Y.-K.-K. Fung, and G. W. Gibson, "Hybrid cryogenic memory cells for superconducting computing applications," in *Proc. IEEE ICRC*, 2017, pp. 1–3.

[6] 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.

[7] S. Nagasawa, Y. Hashimoto, H. Numata, and S. Tahara, "A 380 ps, 9.5 mW Josephson 4-Kbit RAM operated at a high bit yield," *IEEE Trans. Appl. Supercond.*, vol. 5, no. 2, pp. 2447–2452, Jun. 1995.

[8] 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, Jun. 2017, Art. no. 1800904.

[9] 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.

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

[11] A. Beckers, F. Jazaeri, and C. Enz, "Cryogenic MOSFET threshold voltage model," in *Proc. 49th Eur. Solid-State Dev. Res. Conf. (ESSDERC)*, 2019, pp. 94–97.

[12] W. Chakraborty, K. Ni, J. Smith, A. Raychowdhury, and S. Datta, "An empirically validated virtual source FET model for deeply scaled cool CMOS," in *Proc. IEEE IEDM*, 2019, pp. 39.4.1–39.4.4.

[13] W. K. Luk and R. H. Dennard, "A novel dynamic memory cell with internal voltage gain," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 884–894, Apr. 2005.