2021 IEEE Custom Integrated Circuits Conference (CICC) | 978-1-7281-7581-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/CICC51472.2021.9431527

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

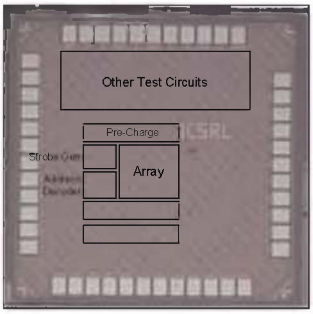
Rakshith Saligram[1], Suman Datta[2], Arijit Raychowdhury[1]

[1] Georgia Institute of Technology, [2] University of Notre Dame

In the pursuit of higher digital performance, as well as for finding new applications in emerging computing models and operating environments, low-temperature (300K to 150K) and cryogenic (<150K) computing is gaining momentum. In particular, space electronics (4K-200K), digital-control in fuel-cell electric vehicles (20K-80K) and digital-assist/ peripherals of quantum/ superconducting computers (10mK – 100K) require temperature-scalable process technologies and digital logic/memory [1-7]. Further, cryo-CMOS computing (100K – 150K) has been recently shown to be a significant booster for high-performance computing (HPC), with process retargeted for cryo-HPC [1]. For all these applications, there is a demand for high-density, large-capacity, high-bandwidth (BW) memory, which cannot be addressed by non-CMOS technologies (e.g., Vortex Transitional devices, Josephson Junction based Memory arrays etc. [2-7]) that suffer from single-temperature operation, low-density, poor-scalability, poor-reliability and high design-complexity. On the other hand, scaled CMOS with improved characteristics at low-temperature, such as steep sub-threshold slope (SS), improved channel transport, low-I_{OFF} and reduced thermal noise, provides a promising, yet largely unexplored pathway for integrating large on-die memory with both CMOS and non-CMOS cryo-computing, across a wide range of temperatures and applications. In this research test-chip, we present a 2T-gain-cell (GC) based embedded-DRAM (eDRAM) macro in 28nm HKMG CMOS targeted for a range of cryo-applications and enabled by superior transistor characteristics at low-temperature. It features: (1) a hybrid P/N gain-cell for stable storage and low coupling noise, (2) an open bit-line architecture taking advantage of the low noise, (3) reliable operation from 300K to 4K and (3) 10^6x improvement of retention time from 300K to 4K. The cell architecture, operating voltages and the design-space for eDRAM (1T-1C and gain-cell) are summarized in Fig. 1. While low storage capacitor in eDRAM on logic leads to low retention time and high refresh power, the ultra-low leakage at cryo-temperature makes it a promising technology with a measured retention time of >1s at 4K.

Previous demonstrations of room-temperature gain-cell arrays feature P-only or N-only cells (Fig. 1). Analysis of the P-cell reveal charge injection from the write WL (WWL) to the storage node during a 0→1 transition, worsened by a subsequent charge injection during 0→1 transition on the read WL (RWL), elevates the voltage level during "0" storage. Similarly, in the N-cell 1→0 transitions on the WWL add to the 1→0 transition on the RWL affecting the "1" storage. The problem of the charge-injection is exacerbated at scaled nodes due to higher overlap capacitance and worsened at low temperature due to sharper signal transitions arising from steeper Ids-Vgs characteristics. To address this challenge, the macro features a hybrid P-N gain-cell where charge is injected from only one WL for each of the storage configurations (Fig. 2); and the 1→0 (0→1) RWL (WWL) transition further assists the "0" ("1") storage. With an optimized sense-amplifier threshold, the hybrid cell provides high noise-margin, balanced P/N density, error-free operation and relaxes physical design constraints to improve array density. Measurement of key transistor properties from 300K to 4K reveal linear SS scaling for both NMOS and PMOS in linear and saturation. Further, the transistor threshold increases at low temperature; and both these effects result in >10^6 decrease in the measured transistor leakage. The higher carrier mobility at low temperature also improves saturation current by 1.9x (1.35x) for the PMOS (NMOS). Thus, CMOS itself provides a promising technology for low-temperature digital operation. The gain-cell macro is arranged in a 1Kb subarray with per-column, cross-coupled strong-arm based latched comparators in an open-BL architecture. Pre-charge circuits enable externally controlled BL voltages for test and debug. Peripherals, including timers, strobe generators and decoders are synthesizable. Sample timing diagram illustrating signal transitions are shown in Fig. 3. The retention time is characterized by the capacitance of the storage node and the net leakage of the cell, and the data needs refreshed periodically to prevent incorrect reads.

The gain-cell macro is measured and characterized across multiple temperature points and the V_{DD}-Frequency shmoo plots at three critical temperature points (T=300K, 100K and 4K) demonstrate (1) wide operating range down to 0.8V, (2) <1ns RD/WR cycles and (3) temperature scalability from room temperature to deep cryo. The maximum array frequency (F_{MAX}) is measured across V_{DD} and temperature. Improved transistor characteristics at low temperature is reflected in a 24.7% (33.3%) improvement in F_{MAX} at V_{DD}=1.1V (0.8V). A peak F_{MAX} of 1.3 GHz at T=4K is measured.



Technology	28nm HKMG CMOS
Flavor	HPC+
Metal Stack	1P-9M
Die Area	1mm^2
Core Voltage	0.9V
IO Voltage	1.8V
Word Size	32bits
Cell Area	0.5704µm^2
Banks	1
Memory Size	1kb

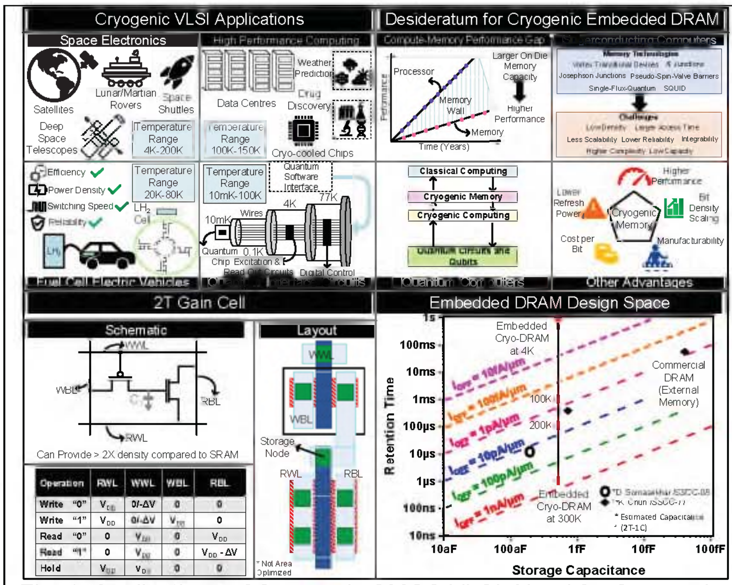
The linear scaling of SS results in exponentially lower leakage and higher retention time as the temperature is lowered. Retention time is characterized over 12 eDRAM subarrays (12Kb) at VDD=1.0V across temperature and the retention-PFAIL is illustrated in Fig. 5. The median retention time improves from 2.4us (300K) to 6.5s (4K) demonstrating a 2.7x10^6 improvement while the 3σ fail-rate improves by an equivalent amount. The retention time statistics are calculated across V_{DD} points, showing super-linear scaling with increasing V_{DD} across temperature (Fig. 5). The array is tested under full BW condition with back-to-back WR/RD cycles and across temperature. A peak array BW of 4.2 Gbps at 761 µW/kB is measured at 300K, and it improves to 5.24 Gbps at 560 µW at 4K. Consequently, a net energy-efficiency improvement (in terms of Gbps/W) of 1.7x is obtained because of (1) enhanced performance of critical path circuits, (2) 10^6 decrease in refresh rate and (2) near-absence of leakage at lower temperatures. The array refresh power decreases to <1 nW/Kb at 4K while the RD and WR energies are measured at 360 fJ/kB (340 fJ/kB) and 480 fJ/kB (425 fJ/kB) at 300K (4K) respectively.

The characterization of standby power (including refresh power and leakage from peripherals) is performed across mean retention time by varying the operating temperature and the results are presented in Fig. 6. 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. Every WR and RD operation is followed by M No-ops and the corresponding array power is measured. At higher values of M, we observe that most of the array power is consumed in refresh operations at 300K and a 5.8x improvement in array power is noted at 4K. The gain-cell macro is compared against existing prototypes for low temperature memory, 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 4K and consume a large array power at low density/capacity using non-CMOS and hybrid processes. By comparison, we demonstrate a 4K-300K 2T-gain-cell based macro with high energy-efficiency, 5.24 Gbps of BW operating at 1.3GHz (at 4K) on a 28nm CMOS logic technology process. The die micrograph and the chip characteristics are shown above.

References: [1] H. L. Chiang et al.,VLSI-T. Symp., 2020. [2] V. K. Semenov et al., IEEE Trans. Appl. Superconductivity 2019. [3] J. Yau, et al., ICRC, 2017 [4] T. Van Duzer et al., IEEE Trans. Appl. Superconductivity 2013 [5] S. Nagasawa et al., IEEE Trans Appl. Superconductivity, 1995 [6] M. Tanaka et al., IEEE Trans. Appl. Superconductivity, 2017. [7] K. Kuwabara et al., IEEE Trans. Appl. Superconductivity, 2013.

Acknowledgements: This work was supported by the Lab of Physical Sciences, Samsung GRO Program, ORNL and SRC via the ASCENT center.

978-1-7281-7581-2/21/$31.00 ©2021 IEEE



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. 1. Applications of Cryogenic CMOS VLSI, Need for Cryogenic eDRAM, 2T Gain Cell and Design Space of eDRAM.

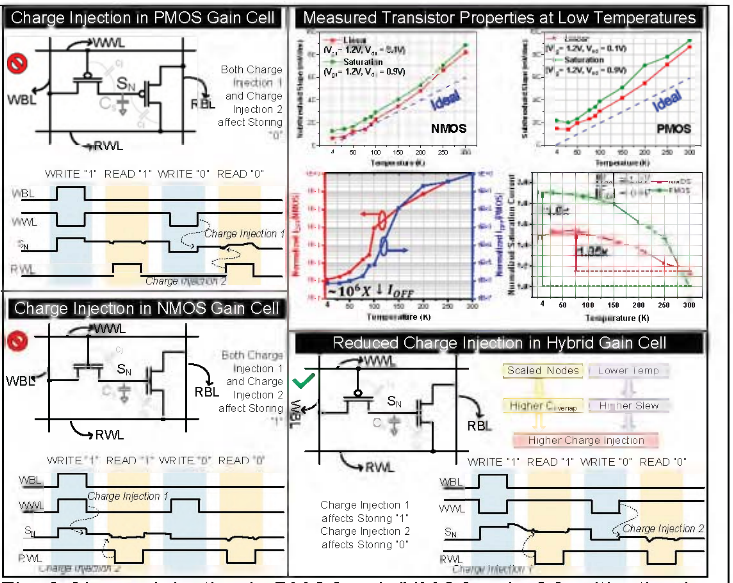

Charge Injection in PMOS Gain Cell
WWL
WBL
SN
RBL
RWL
Both Charge Injection 1 and Charge Injection 2 affect Storing "0"
WRITE "1" READ "1" WRITE "0" READ "0"
WBL
WWL
SN
RWL
Charge Injection 1
Charge Injection 2
Charge Injection in NMOS Gain Cell
Both Charge Injection 1 and Charge Injection 2 affect Storing "1"
Measured Transistor Properties at Low Temperatures
Linear (V_{ds} = -1.2V, V_{gs} = -0.1V)
Saturation (V_{ds} = 1.2V, V_{gs} = 0.9V)
Ideal
NMOS
PMOS
Temperature (K)
~$10^6X \downarrow I_{OFF}$
1.6X
1.35X
Reduced Charge Injection in Hybrid Gain Cell
Scaled Nodes
Lower Temp
Higher $C_{overlap}$
Higher Slew
Higher Charge Injection
Charge Injection 1 affects Storing "1"
Charge Injection 2 affects Storing "0"

Fig. 2.Charge Injection in PMOS only/NMOS only GC mitigation in hybrid GC, measured transistor properties at Low Temperatures.

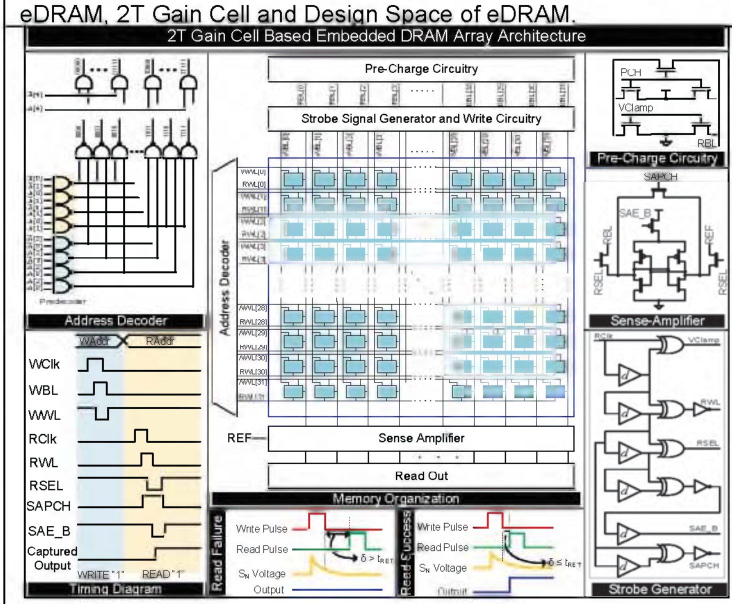


Fig. 3. 2T Gain Cell based Embedded DRAM Array Architecture showing Memory Organization, components and Timing Diagrams.

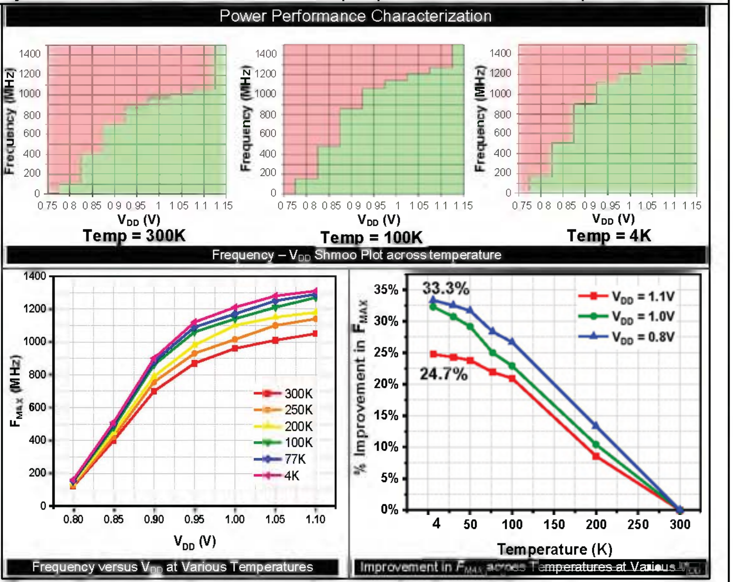


Fig. 4. Power Performance Characterization of Embedded DRAM with Frequency-V_{DD} Shmoo plots and F_{MAX}-V_{DD} plots.

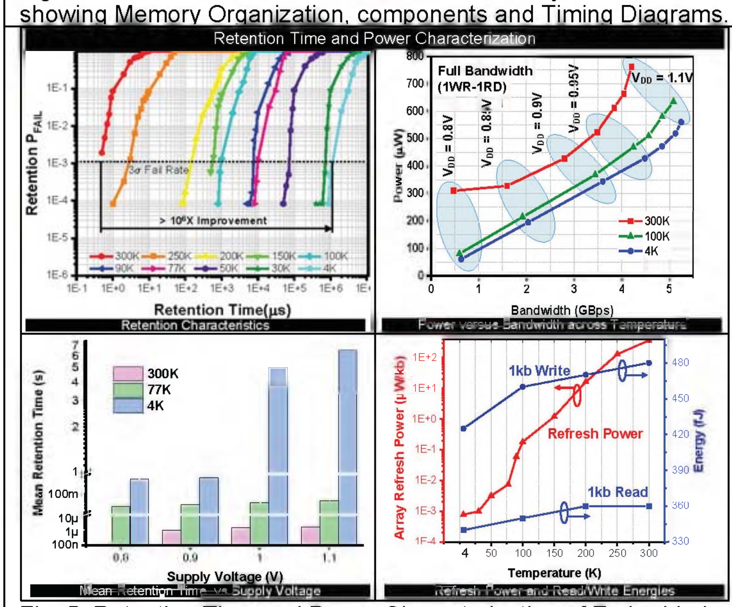
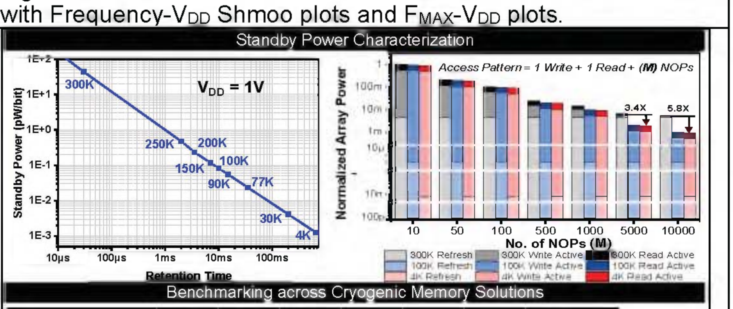


Fig. 5. Retention Time and Power Characterization of Embedded DRAM at Cryogenic Temperatures.



Benchmarking across Cryogenic Memory Solutions

	[2]	[3]	[4]	[5]	[6]	[7]	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	99µm²	2µm²	3025µm²	3.36µm²	--	--	0.59µm²
Read/Write Access Times	–	300ps/ 2000ps	800ps/--	430ps/ 300ps	660ps/--	1340ps/--	763ps
Dynamic Power		<100µW	9.5mW	12mW Read/21 mW Write	0.78mW Read/2.2 mW Write	28mW Read/41 mW Write	0.76mW (300K) 0.56mW (6K)
Operating Temperature	4K						4K-300K
Integration with Si Logic	Requires Heterogeneous Integration						Monolithic

VT Vortex Transitional Memory Devices
MRAM Magnetic RAM
JJ Josephson Junctions
nTron Nanocryotrons

Fig. 6. Standby Power Characterization and benchmarking of current work with other cryogenic memory solutions.