Pseudo-Static 1T Capacitorless DRAM using 22nm FDSOI for Cryogenic Cache Memory

Wriddhi Chakraborty[1], Rakshith Saligram[2], Aniket Gupta[1], Matthew San Jose[1], Khandker Akif Aabrar[1], Sourav Dutta[1], Abhishek Khanna[1], Arijit Raychowdhury[2] and Suman Datta[1]
University of Notre Dame, Notre Dame, IN 46556, USA, [2]Georgia Institute of Technology, Atlanta, GA 30332, USA
Email: wchakrab@nd.edu

Abstract—Cryogenic CMOS processors need low latency, high bandwidth access to high-density on-die cache memory to maximize performance. In this work, we experimentally demonstrate, for the first time, pseudo-static random access memory operation of a 1T Capacitorless Floating Body DRAM using 22nm FDSOI transistor, down to 4.8K, suitable for cryogenic cache memory. We demonstrate a 1T Cryo-DRAM (W/L_G=120nm/20nm) that exhibit : (a) record high sensing current and sense margin ($\Delta I_{Read}=I_{Read,1}-I_{Read,0}$), (b) pseudo-static retention characteristics (>10^5 sec); (c) high write endurance > 10^{10} cycles, and (d) non-destructive read cycles > 10^{16}, suitable for cache application. Benchmarking reveals that 1T Cryo-DRAM outperforms Cryo-SRAM and Cryo-STT-MRAM in memory density by 10x and 50x; in read/write energy by 2.7x/2.4x and 1.3x/1.5x and in read latency by 1.46x and 1.80x respectively for a cache size of 2MB. Hence, 1T Cryo-DRAM is a viable option for L2/L3 cache in high-performance cryogenic computing.

I. Introduction

Low temperature (~77K) logic technology is a promising approach for high performance cloud computing. Temperature is a viable knob to enable steep subthreshold switching characteristics, accompanied by other performance boosters like improved mobility, improved reliability, lower wiring resistance and reduced self-heating [1], [2]. Improvement in cryogenic logic core performance must be balanced by expanding the on-die cache memory capacity, in order to avoid the memory-wall (Fig. 1(a)). On-chip memory solutions like embedded-DRAM (eDRAM) and STT-MRAM have been proposed as higher-density alternative to 6T-SRAM at cryogenic temperature [3][4][5]. Here, we propose and demonstrate a pseudo-static, deeply-scaled capacitor-less 1T DRAM for cryogenic cache, taking advantage of the (a) temperature invariant property of Gate-Induced-Drain-Leakage (GIDL) for hole injection in the transistor body, (b) boost in transconductance resulting from reduced phonon scattering at cryogenic temperature for a high current sense window and (c) suppressed Shockley-Read-Hall (SRH) generation-recombination rate which results ultra-high retention time (~10^5 sec) at 77K. 1T Capacitorless DRAM utilizes the floating body effect of single MOS transistor for memory operation. Holes can be injected into the body through GIDL Program/ Write '1' method (Fig. (2(a)). On the other hand, these injected holes can be expunged by forward biasing the drain to body junction during Erase i.e. Write '0' operation (Fig. (2(a)). The excess and absence of holes in the floating body modulate the MOSFET source barrier and modulates the drain current (I_{DS}) during read operation. In this work, we experimentally demonstrate, for the first time, memory operation of 1T DRAM on 22nm FDSOI platform [6], from 300K to 4.8K, with a projected cell layout of $6F^2$. Finally, we perform Energy-Delay comparison between 1T Cryo-DRAM, Cryo-STT-MRAM and Cryo-SRAM based on experimentally obtained memory parameters to determine its potential for high-speed, high-density, pseudo-static, low power cache level memory for cryogenic processor.

II. Device Characterization

Fig 3. (a) shows well-tempered transfer characteristics of 20nm gate length FDSOI nFET at 300K, 77K and 4.8K. Threshold voltage undergoes a positive shift, while the subthreshold slope steepens at low temperature [7]. Reduced phonon scattering results in 35% boost in peak g_m at 77K (Fig. 3(b)). Hence, 1T Cryo-DRAM will have a higher sense current at 77K compared to 300K, resulting in faster read latency. Interestingly, GIDL is temperature invariant (Fig. 3(c)) as it totally dominated by Band-to-band-tunneling component. Thus, GIDL programming allows efficient injection of holes in the body, resulting in lower write energy at 77K. Fig. 4(a) shows the projected layout of the 1T DRAM cell using 22nm FDSOI transistor with an ultra-compact layout area of $6F^2$ or 0.013 μm^2. Fig 4(b) shows the cross-sectional TEM of 22nm FDSOI Si nFET n [6]. The device dimension of the memory cell is 120nm/20nm (W/L_G). Fig. 5(a) shows the timing diagram of the voltage waveforms, with corresponding voltages listed in TABLE I. Write pulses of +/- 1.7V and 20ns duration were asserted on BL and WL, respectively. in the corresponding read current, I_{READ} after Write '1' and Write '0' are shown for 4.8K, 77K, 150K and 300K operation (Fig. 5(b), (c), (d), (e)) indicating the presence and absence of holes in the body. Interestingly, the sense current window i.e. ΔI_{Read} increases monotonically at low temperature, which is attributed to higher g_m at low temperature. The dependence of ΔI_{Read} on write voltage and write pulse width is characterized to define the operational design space of 1T Cryo-DRAM (Fig. 6(a)). ΔI_{Read}>5μA was achieved at +/-1.4V, 20ns write pulse, the lowest program time being limited by measurement setup. ΔI_{Read} variation of 1T DRAM cell over 100 cycles show wide sense margin of ΔI_{Read}=8.5μA for 3σ devices (Fig. 6(b)). Retention characteristics of State'1' and State '0' for 4.8K, 77K and 300K are shown in Figs. 7(a), (b) and (c) respectively, with no hold voltage applied. At 4.8K

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720578

and 77K, State '1' remains unaffected for more than 100s. Under hold condition (V_G=V_D=0V), the MOSFET is in accumulation at cryogenic temperature, where as it enters depletion at 300K. Thus, State '1' does not undergo any change at cryogenic temperature. However, State '0' approaches State '1' during hold, but at a slower rate than 300K due to suppressed SRH generation rate. Retention time of $2x10^5$ s is obtained at 77K by extrapolating the sense margin to 3μA. Fig. 8(a) shows the timing diagram while Table II shows the applied voltages for Write, Hold, and Read operations during Hold. We confirm the non-destructive nature of the read-scheme, for both states '1' and '0'. ΔI_{Read} margin is maintained after 10^9 cycles and ΔI_{Read}=7μA is projected after 10^{16} read cycles, indicating that the drain voltage applied during Read does not cause loss of '0' state.

Fig. 9(a) shows the timing diagram with voltage waveforms for Write, Hold, and Read operations for investigating Write Endurance of the states under consequent (Bipolar) Program-Erase cycles. ΔI_{Read} margin is maintained above 50% of initial ΔI_{Read} for $2x10^9$ and 10^8 cycles under +/-1.6V and +/-1.7V pulse respectively. There are two mechanisms responsible for the endurance degradation in 1T Cryo DRAM. As shown in Fig. 10(a) GIDL current measured after each endurance cycles decreases with cycling. The reduction is more prominent for higher pulse amplitude. This is caused by hot-holes generated during GIDL programming. Hot holes trapped in high-K in the gate-drain overlap region screens the drain-to-gate electric field. This reduces the GIDL and causes less holes to be injected in the channel with cycling [8]. The transconductance (g_m) was found to degrade after cycling with a positive shift in V_{TH}. This is due to acceptor states generated in the source-drain to gate overlap region during GIDL, and can cause a positive shift of V_{TH} [8]. Hence optimizing the write voltage space for optimum sense margin, endurance and retention is the key for 1T Cryo DRAM operation.

To further evaluate 1T DRAM feasibility for cryogenic cache application, we investigate the disturb challenges. Fig. 11. (a) shows the timing diagram for Write, Hold, and Read operations for measuring Write Disturb on unselected rows during read operation in a 1T DRAM array. TCAD simulation of 22nm FDSOI device, calibrated to experimental data, is used to understand the origin of write disturbs in 1T Cryo-DRAM cell. TCAD simulation suggests that SRH recombination of stored holes during '0' disturb and BTBT generation of holes during '1' disturb at drain-body junction depletion region are the two major disturb mechanisms. Fig. 11(c) shows that increasing the magnitude of disturb voltages after Program and Erase, results in the closure of read margin, ΔI_{Read}, for both state '0'/'1', however, the sense margin of 5.7μA is still maintained after 10^7 cycles, indicating the immunity of the states to external access disturbs. For 1T Cryo DRAM to be a viable cache memory, it is important to control the disturbs by applying opposite polarity voltage of optimized value during the hold. A summary of the measured and projected value of memory cell parameters are summarized in Table V.

III. Array Simulation of Cryogenic 1T DRAM at 77K

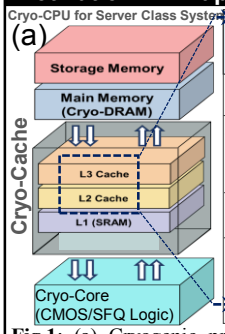
The device level demonstration of 1T Cryo-DRAM cell is extended to memory array using a simulation framework. The typical organization of cryogenic cache memory is shown in Fig. 12(a) and comprises of multiple banks, MATs and subarrays. Experimentally calibrated temperature dependent energy-delay models for decoders, sense amplifiers, pre-charge circuits, multiplexers and read out circuits, interconnect data from industry standard foundry 22nm FDSOI PDK, along with the device level measurements for 1T Cryo-DRAM are used to accurately model the cache memory performance. Fig. 12 (b) illustrates the percentage break down of different components of read access latency for a cache size of 1MB at 77K to be dominated by cell read time (74%). The read latency increases with increase in cache size up to 1MB due to increase in the cell read time with increasing interconnect capacitance and is exasperated by increased logic delay for higher cache size while still providing 45% improvement compared to corresponding 300K SRAM (Fig. 12(d)). Similarly, the read energy (Fig. 12 (c)) increases with cache size due to increase in energy expended in charging the interconnect capacitances and later worsened by logic, but still yielding 75% (70%) improvement for 256kB (2MB) respectively. The write energy limited by the minimum write pulse width provides a constant improvement of approximately 65% across varying cache sizes. The write latency is 50% higher than 300K SRAM for lower cache sizes due to the higher cell write time but gets amortized by the logic delay for larger cache sizes of above 1MB (20% at 2MB). Thus cryogenic floating body cell based 1T-DRAM at 77K shows betterment in performance and energies compared to SRAM and STT-MRAM making it suitable candidate for last level of cache (L2/L3).

IV. Conclusion

1T Capacitorless DRAM is demonstrated down to 4.8K for high-density cryogenic cache memory application. The 1T DRAM operates at 77K with 2x higher, 4 orders of magnitude longer the retention than its, counterpart at 300K. Record high write endurance of 10^{10} cycles and non-destructive reads > 10^{16} cycles further prove the feasibility of 1T DRAM for cache memory in cryogenic CPUs. Cache architecture level simulation study shows that 1T Cryo-DRAM has 10x and 50x higher memory density than Cryo-SRAM and Cryo-STT-MRAM respectively. It provides 2.7x/2.4x and 1.3x/1.5x benefit in read/write energy compared to 77K SRAM and 77K STT-MRAM. In terms of read latency it also gains 1.46x and 1.80x improvement respectively over 77K SRAM and 77K STT-MRAM as cache capacity approaches 2MB. Hence, 1T Cryo-DRAM is a viable option for L2/L3 cache in high-performance cryogenic computing.

Reference: [1] W.Chakraborty et. al. VLSI Sym. 2021, [2] W.Chakraborty et. al. IRPS. 2020 [3] E. Garzón et. al. IEEE Tran. Nano 2021, [4] E. Garzón et. al. *IEEE Tran. VLSI Sys.* 2021 [5] Saligram et. al. IEEE CICC 2021, [6] Carter et. al IEEE IEDM 2016, [7] W.Chakraborty et. al. IEEE IEDM 2019. [8] B. Doyle et. al. TED 1990. **Acknowledgement: This work was supported by ASCENT center, one of the six centers in JUMP.**

Motivation: 1T Capacitorless DRAM for Cryogenic Cache



(b)

	Cryo-SRAM [3]	Cryo-STT-MRAM [3]	Cryo-eDRAM 3T Gain Cell [4]	Cryo-eDRAM 2T Gain Cell [5]	1T [This Work]
Demonstration	Simulation	Simulation	Simulation	Experimental	Experimental
Cell Size (μm²)	0.11	0.505	0.015	0.59	0.013
Access Time	-	-	-	0.8-0.9ns	1-10ns
RetentionTime	Static	Static	23ms	40ms	Pseudo Static
Destructive Read	No	No	No	No	No
Leakage Power/Bit	3.36nW	-	-	-	0.45nW

Fig.1: (a) Cryogenic processor requires compact high-speed random access memory (RAM) that expand the on-chip memory capacity at various levels of cache hierarchy (b) One Transistor (1T) Capacitorless Cryo-DRAM on 22nm FDSOI platform, first experimentally demonstrated in this work, shows pseudo static data retention behavior at 77K, making it an excellent choice for high density Cryo-cache application

seudo-Static 1T Capacitorless Cryogenic DRAM

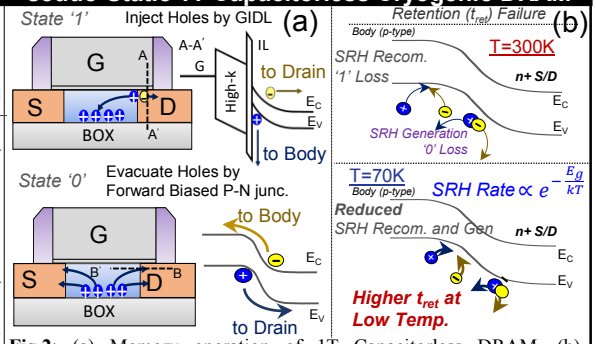


Fig.2: (a) Memory operation of 1T Capacitorless DRAM, (b) Retention failure in 1T DRAM involves Shockley-Read-Hall (SRH) recombination/generation of holes in the body which is strongly suppressed at cryogenic temperature, resulting pseudo-static retention behavior at low temperature

Cryogenic Characterization of 22nm FDSOI FET

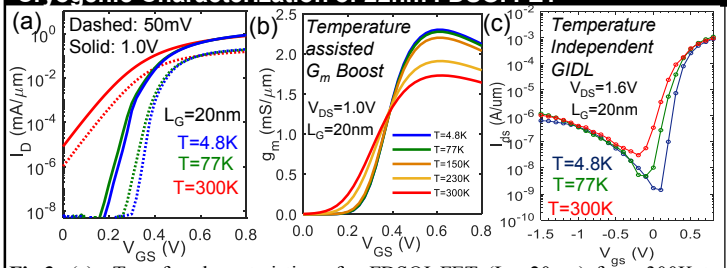


Fig.3: (a) Transfer characteristics of n-FDSOI FET (L_G=20nm) from 300K to 4.8K, (b) Reduced phonon scattering at low temperature results in 35% boost in g_m at 77K, (c) Gate-induced-Drain-Leakage (GIDL) programming of 1T DRAM cell still feasible at cryogenic temperature as GIDL remains unaffected by temperature

22nm FDSOI Cryogenic 1T Capacitorless DRAM Cell

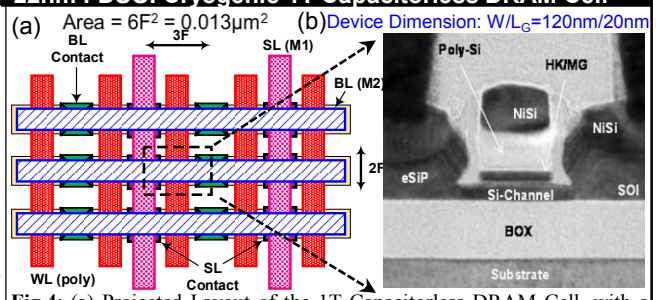


Fig.4: (a) Projected Layout of the 1T Capacitorless DRAM Cell, with a ultra compact layout area of 6F², (b) xTEM image of 22nm FDSOI nFET, with device dimension W=120nm and L_G=20nm [6]

Memory Operation of Cryogenic 1T Capacitorless DRAM Cell

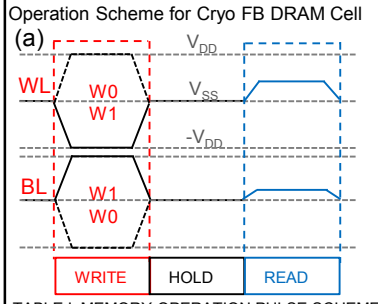


TABLE I: MEMORY OPERATION PULSE SCHEME

	V_D	V_G	V_S
W1	+1.7V	-1.7V	0V
W0	-1.7V	+1.3V	0V
HOLD	0V	0V	0V
READ	0.3V	0.6V	0V

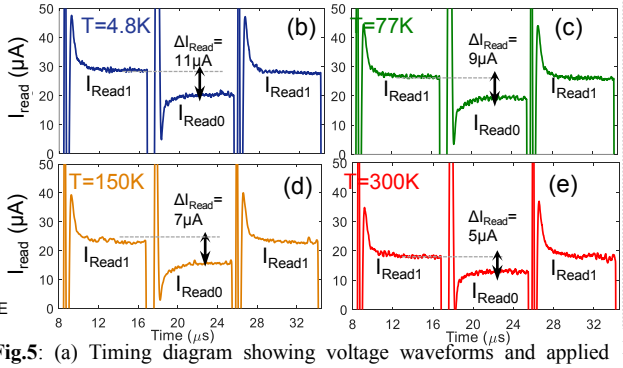


Fig.5: (a) Timing diagram showing voltage waveforms and applied voltages for Write, Hold, and Read operations. Transient modulation in FET read current after Program (injecting holes) and erase (evacuating holes) at (b) 4.8K, (c) 77K, (d) 150K and (e) 300K. Monotonic increase in sense margin (ΔI_{Read}) at low T can be attributed to improved g_m

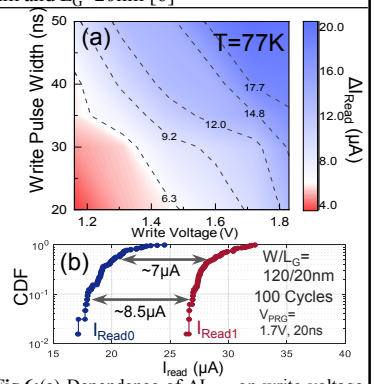


Fig.6:(a) Dependence of ΔI_{Read} on write voltage and write pulse width at 77K. shows ΔI_{Read}>5μA achievable at 1.4V. Lowest program time (20ns) limited by measurement setup, (b) ΔI_{Read} variation of 1T DRAM cell over 100 cycles show ΔI_{Read}=8.5μA for 3σ devices

Retention and Non-Destructive Read Characteristics of Cryogenic 1T Capacitorless DRAM Cell

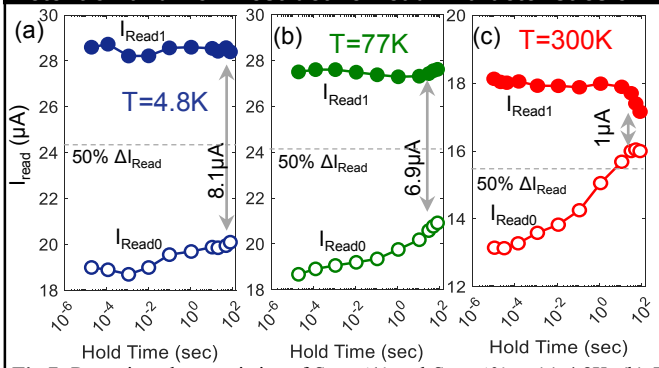


Fig.7: Retention characteristics of State '1' and State '0' at (a) 4.8K, (b) 77K and (c) 300K with no Hold voltage applied, shows State '1' remains unaffected more than 100s, attributed to suppressed SRH recombination rate of the stored holes at cryogenic temperature. However, State '0' approaches State '1' during hold, but at a damped rate than 300K. Retention time of $2x10^5$s obtained at 77K by extrapolating the sense margin to close at 3μA

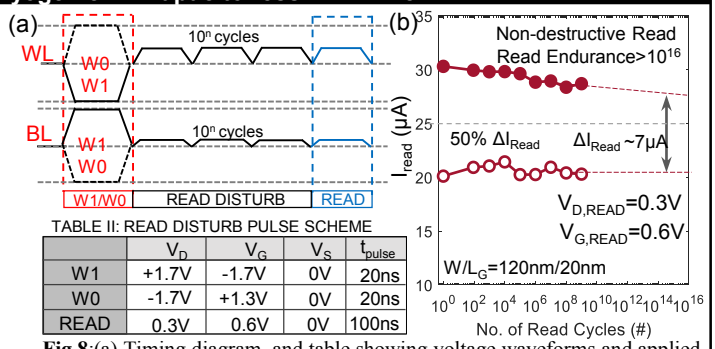


TABLE II: READ DISTURB PULSE SCHEME

	V_D	V_G	V_S	t_{pulse}
W1	+1.7V	-1.7V	0V	20ns
W0	-1.7V	+1.3V	0V	20ns
READ	0.3V	0.6V	0V	100ns

Fig.8:(a) Timing diagram and table showing voltage waveforms and applied voltages for Write, Hold, and Read operations during Hold, for investigating Non-destructive nature of the read-scheme, during hold of state '1'/'0'. (b) ΔI_{Read} margin was maintained even after 10^9 cycles and 7μA ΔI_{Read} was projected after 10^{16} read cycles, indicating drain voltage during Read does not cause loss of '0' state by generating excess holes through SRH generation or Band-to-Band-Tunneling

Endurance of Cryogenic 1T Capacitorless DRAM at 77K

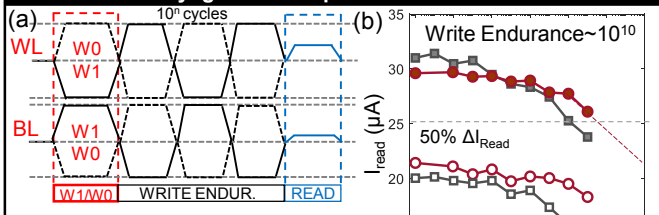


TABLE III: WRITE ENDURANCE PULSE SCHEME

	V_D	V_G	V_S	t_{pulse}
W1	+1.7V	-1.7V	0V	20ns
W0	-1.7V	+1.3V	0V	20ns
READ	0.3V	0.6V	0V	-

Fig.9: (a) Timing diagram and table showing voltage waveforms and applied voltages for Write, Hold, and Read operations for investigating Write Endurance of the states under consequent Program-Erase cycles (b) ΔI_{Read} margin was maintained above 50% of initial ΔI_{Read} for $2x10^9$ and 10^8 cycles under +/-1.6V and +/-1.7V pulse respectively.

Origin of Endurance Degradation of 1T DRAM at 77K

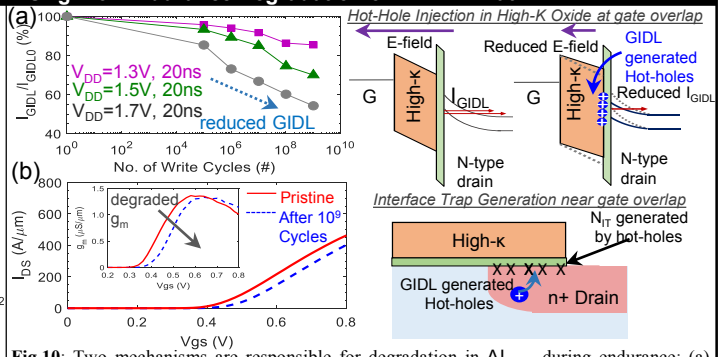


Fig.10: Two mechanisms are responsible for degradation in ΔI_{Read} during endurance: (a) Hot-holes generated during GIDL, when trapped in high-K oxide above the gate-drain overlap region, screens the drain-to-gate electric field. This reduces GIDL and hence no. of injected holes during consecutive programing. (b) Interface Traps generated at the gate-drain overlap regio by the hot-holes, degrades the transconductance of the FET, degrading ΔI_{Read}

Array Disturb Conditions in Cryogenic 1T Capacitorless DRAM Cell at 77K

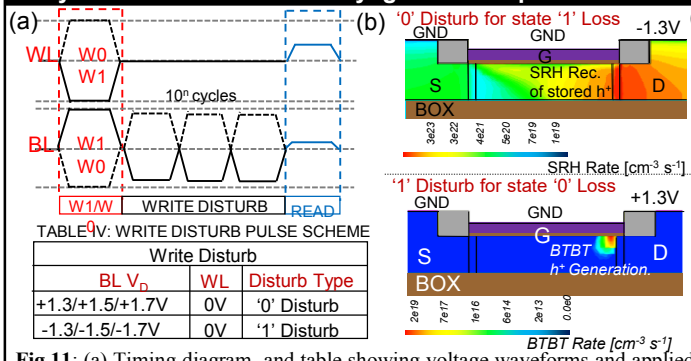


TABLE IV: WRITE DISTURB PULSE SCHEME

Write Disturb		
BL V_D	WL	Disturb Type
+1.3/+1.5/+1.7V	0V	'0' Disturb
-1.3/-1.5/-1.7V	0V	'1' Disturb

Fig.11: (a) Timing diagram and table showing voltage waveforms and applied voltages for Write, Hold, and Read operations for measuring Write Disturb on unselected rows during read operation in a 1T DRAM array (b) TCAD simulation suggested SRH recombination of stored holes and BTBT generation of holes at Drain-body junction depletion region are the two major mechanisms for '0' and '1' disturb respectively. (c) Increasing the magnitude of disturb voltages after Program and Erase results in collapsing of ΔI_{Read} for both state '0'/'1', however sense margin of 5.7μA was still maintained after 10^7 cycles, indicating the immunity of the states to external disturbs.

Cell Parameters for Array Simulation of Cryogenic 1T Capacitorless-DRAM at 77K

Cell Size	0.013μm²
Access Time*	1.5ns
Retention Time* (@ΔI_{READ}=3μA)	$2x10^5$ s
Write Energy	23.77pJ
Read Energy	17.92pJ
Non-destructive Reads	$>10^{16}$
Write Endurance	$\sim10^{10}$

(*Projected)

Cell Array Simulation of Cryogenic 1T Capacitorless DRAM at 77K

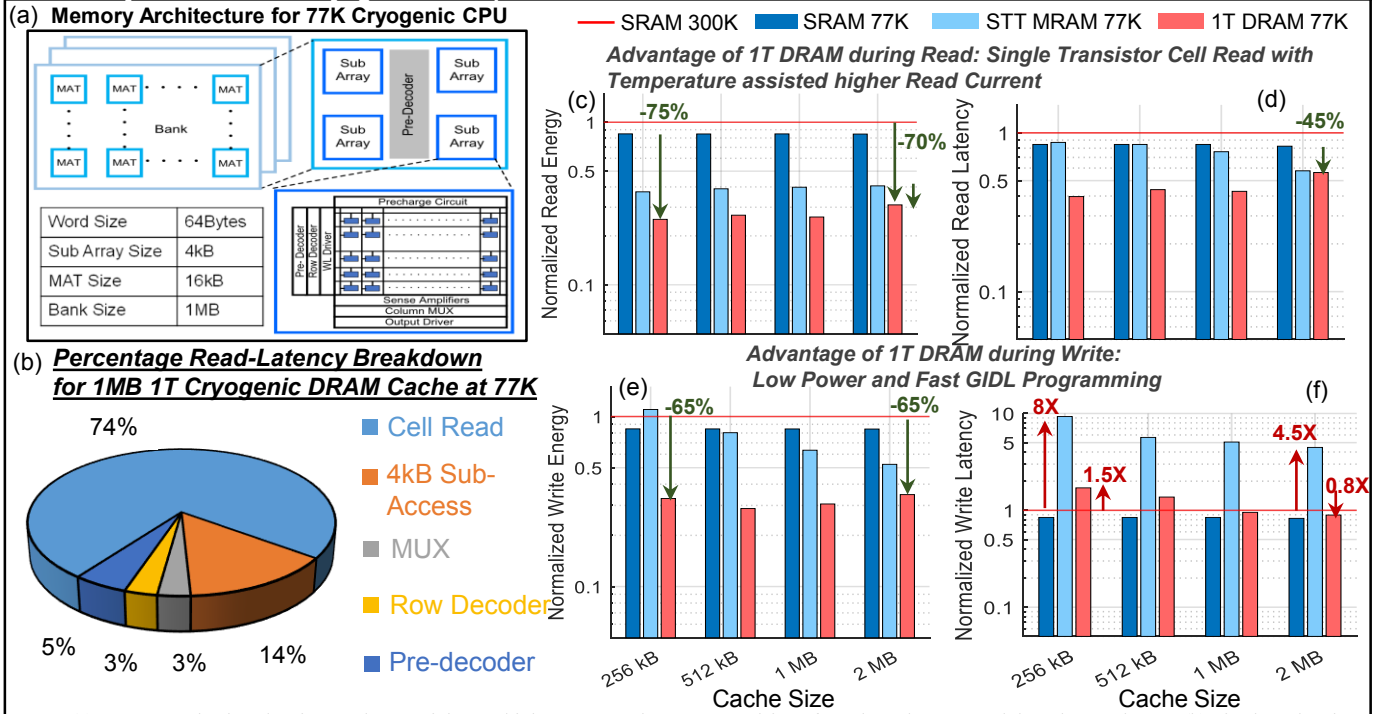


Fig.12: (a) Array organization showing Banks containing multiple MATs, each MAT comprising of 4 16kB sub arrays and the sub array layout. (b) Pie chart showing cell read time dominated read latency at 77K for 1MB cache size. (c,e) Read/Write Energy and (d ,f)Latency for 77K SRAM, 77K STT-MRAM, 77K 1T-DRAM all normalized to 300K SRAM for multiple cache sizes. 77K 1T-DRAM Read/Write Energy and Read Latency show significant gain over 77K SRAM and STT-MRAM at all cache size due to (1) single transistor cell read with temperature assisted higher read current and (2) Low power GIDL programming scheme. Write latency also improves over STT-MRAM of same memory capacity, with comparable values with 77K SRAM, specially for higher capacities, which indicates deeply scaled 1T Capacitorless DRAM cells ($1.5F^2$) are a potential candidate for L2/L3 cache level in the memory hierarchy of future Cryogenic-memory.