Multi-bit per-cell 1T SiGe Floating Body RAM for Cache Memory in Cryogenic Computing

W.Chakraborty[1*], P. Shrestha[2,3], A.Gupta[1], R.Saligram[4], S.Spetalnick[4], J. Campbell[2,3], A. Raychowdhury[4] and S. Datta[1]

[1]University of Notre Dame, Notre Dame, USA, [2]Theiss Research, La Jolla, USA, [3]National Institute of Standards and Technology, Gaithersburg, USA,

[3]Georgia Institute of Technology, Atlanta, USA *Email: wchakrab@nd.edu

2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) | 978-1-6654-9772-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/VLSITechnologyandCir46769.2022.9830483

Abstract: Cryogenic computing requires high-density on-die cache memory with low latency, high bandwidth and energy-efficient access to increase cache hit and maximize processor performance. Here, we experimentally demonstrate, high-speed multi-bit memory operation in 1T SiGe Floating-body RAM (FBRAM) using 22nm FDSOI transistor at 77K, for cryogenic cache memory application. The 1T SiGe FBRAM cell (W/L_G=170nm/20nm) at 77K exhibits : (a) record write time of <5ns with write voltage (V_{Write}) 1.5V; (b) high sense current ($I_{Read,1}$~75μA) with read margin ($\Delta I_{Read}=I_{Read,1}-I_{Read,0}$) ~14 μA; (c) 2-bit/cell operation; (d) pseudo-static retention (~$8x10^3$ s) for single-bit and worst case retention of 100 s for 2-bit per cell, and (e) high write endurance >10^{12}. Array-level benchmarking shows that compared to 6T SRAM, 1T SiGe FBRAM shows 8.3x higher memory density with 2.3x/1.8x gain in read/write energy, 3.3x/1.7x in read/write latency and 4.6x in energy-delay product (EDP) for a cache size of 16MB at 77K. Considering the cooling energy cost, FBRAM exhibit 60% EDP reduction compared to 300K 6T SRAM. Hence, SiGe FBRAM is a promising option for L2/L3 cache in high-performance cryo-computing.

Introduction: Performance boosters like steep subthreshold switching, enhanced mobility, improved reliability and lower wire resistance makes low temperature (~77K) logic technology a promising option for High Performance Computing (HPC) systems [1]. To further maximize system performance, high-density on-die cache memory like embedded-DRAM, STT-MRAM and Si FBRAM have been proposed as alternative to 6T-SRAM at 77K [2][3][4]. In this work, we experimentally demonstrate memory operation of 1T SiGe channel FBRAM on 22nm FDSOI[5] platform at 77K, showing significant advantages over other candidates, such as faster write time, lower V_{Write} and energy, pseudo-static retention characteristics and high write endurance (Fig. 1(b)). P-channel SiGe FBRAM utilizes the floating body effect of MOS transistor for memory operation. Majority carrier electrons are injected into the body through Gate-Induced-Drain-Leakage (GIDL) during Write '1' (Fig. (2(a)), while injected electrons can be evacuated by forward biasing the drain to body junction during Write '0' operation (Fig. (2(a)). The excess presence and absence of electrons in the floating body modulate the FET source barrier which, in turn, modulates the drain current (I_{DS}) during cell read. Higher GIDL in p-SiGe compared to n-Si FET (Fig. 3), due to lower band-gap of SiGe [6], enables lower V_{Write} and higher ΔI_{Read}. In this work, we demonstrate multi-bit memory operation of 1T SiGe FBRAM to further enhance cache size. Finally, we benchmark write/read energy-delay metrics of 1T FBRAM against 6T Cryo-SRAM, to determine its potential for high-speed, high-density, pseudo-static cache level memory for cryogenic processor.

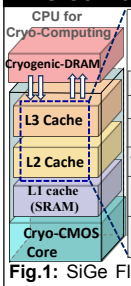
Results and Discussions: Fig. 4(a) shows the ultrafast cryogenic measurement setup. Fast V_G, V_D pulses were applied to 50Ω terminated probes with continuous grounds to efficiently suppress signal reflection. Read current from the cell is sensed using low noise amplifier with gain of 10^3 V/A. Fig. 4(b) shows the timing diagram for applied pulse scheme during Write '1', Write '0' and Read operation. Fast programming pulses with Full-width at Half-Max (FWHM) of 4.2ns were probed at the 1MΩ oscilloscope termination with transmission line delay (t_D)<20% of the pulse rise time to subside the reflections due to 1MΩ. Write pulses of +/-1.6V were asserted on WL and BL, respectively, of SiGe FBRAM cell (W/L_G=170/20nm). Fig. 5(a) shows the corresponding transient modulation in read current (I_{READ}), with ΔI_{READ}=10μA at 77K, after Write '1' and Write '0', indicating the presence and absence of injected electrons in the body. ΔI_{READ} for V_{DD}=1.6V as a function of pulse width (t_{PW}) shows >2.5x increase in ΔI_{READ} at 77K compared to 300K (Fig. 5(b)), due to improved transconductance at cryogenic temperature [7]. Moreover, p-SiGe FBRAM provides 2.7x higher ΔI_{READ} than n-Si FBRAM at 77K, attributed to increased concentration of injected carriers due to higher GIDL, (Fig. 5(c)). This also enables reduction of V_{Write} and consequently write energy in SiGe FBRAM for iso-ΔI_{READ}. We demonstrate the multi-bit program capability of SiGe FBRAM varying the injected electron concentration with varied V_{Write}. Fig. 6(a) shows the pulse scheme for multi-bit cell operation, with corresponding transient modulation in I_{Read} (Fig. 6(b)), demonstrating four distinct I_{Read} levels (corresponding to 2bit/cell) achieved by tuning V_{Write}. For writing bits '00', '01', '10' and '11', V_{Write} of -/+1.7V, +/-1.2V, +/-1.4V and +/-1.7V were asserted to WL/BL with pulse widths of 30ns. Cycle-to-cycle variation over 120 cycles show distribution of the well-separated four I_{Read} states (Fig. 6(c)). Retention characteristics of the states for 77K are shown in Fig. 6(d). I_{Read} '11' and '10', remains unaffected up-to 300s, owing to suppression of thermally activated Shockley-Read-Hall (SRH) recombination at 77K. The I_{Read} '00' and '01' approach each other during hold time. Still, Δ_{IRead} =9μA (50% of initial ΔI_{READ}) was observed till about 100s. Thus, pseudo-static retention of $8x10^3$s and worst case retention of 100s was estimated at 77K for 1-bit/cell and 2bit/cell respectively, extrapolating I_{READ} to 50% initial ΔI_{READ}. Fig. 7(a) shows the timing diagram of pulse scheme during write-endurance under bipolar program-erase cycles. 80% of initial ΔI_{Read} is preserved till 10^{10} cycles of +/-1.4V, 30ns pulse, as shown in Fig. 7(b). Endurance vs V_{Write} for 77K FBRAM (Fig. 8(a)), shows improved cycle-to-failure (failure condition: ΔI_{Read} < 50% initial ΔI_{Read}) in SiGe compared to Si. Power law model fitting suggests 10^{12} write endurance can be achieved in SiGe with 150mV higher V_{Write} compared to Si cell, enabling broader design space for co-optimizing ΔI_{Read}, retention and endurance. Improved write endurance in SiGe can be attributed to lower hot-carrier injection during erase operation due to the presence of valence band offset (ΔE_V ~0.23eV) barrier between SiGe channel and thin-Si capping layer (Fig. 8(c)) [8]. This was further validated with ΔV_{TH} time kinetics of SiGe and Si channel FETs under hot-carrier stress ($|V_G|=|V_D|$=1.7V), showing 60mV lower V_{TH} shift in SiGe FETs after 10^3s of DC stress (Fig. 8(b)).

Cell Layout & Array Simulation of 77K SiGe FBRAM: Cell layout of $6F^2$ is projected for 1T FBRAM, which can increase the cache size by 8.3x compared to 6T SRAM, resulting in reduced cache-miss/10^3 instructions by 2.86x and 2.3x for 1-bit/cell and 2-bit/cell respectively (Fig. 9(a)). Moreover, 75%/62% lower WL/BL capacitance/cell and 80% lower WL resistance/cell in 1T FBRAM (Fig. 9(b, c)), results in 2.3x/1.8x gain in read/write energy, 3.3x/1.7x in read/write latency compared to 6T SRAM for a 16MB cache at 77K (Fig 9(d)). Thus SiGe FBRAM shows better latency and energy compared to 6T SRAM at 77K, making it a potential candidate for last level of cache (L2/L3).

Conclusion: 1T SiGe FBRAM operation is demonstrated at 77K for high-density ($6F^2$), high-speed (<5ns) cryogenic cache memory. In comparison with Si FBRAM cells SiGe FBRAM provides 2.7x higher ΔI_{Read}, 0.15V lower V_{Write}, 20% lower write power and 100x higher endurance, making it an excellent choice for L2/L3 cache. Multi-bit program capability further increases the on-die cache capacity, resulting in 20% lower cache miss per 1K instructions. Array level benchmarking reveals for a large cache size (~16MB), SiGe FBRAM can outperform 77K 6T SRAM in terms of EDP gain by 4.6x. Even considering the cooling energy cost at 77K, FBRAM exhibit 60% EDP reduction compared to 6T SRAM at 300K (Fig. 9(e)). Hence, 1T SiGe FBRAM is a viable option for L2/L3 cache in high-performance cryo-computing.

Reference: [1] H.L Chiang et al., VLSI'20, [2] H.L Chiang et al., VLSI'21, [3] R. Saligram et al., IEEE CICC 2021, [4] W.Chakraborty et al., IEEE IEDM 2021, [5] R.Carter et al., IEEE IEDM 2016, [6] J.Franco et al., IRPS 2010, [7] W.Chakraborty et al., IEEE VLSI 2021, [8] W.Chakraborty et al., IEEE IRPS 2020. **Acknowledgement:** This work was supported through DARPA sponsored LTLT program. We thank William Taylor and Nigel Cave of GLOBALFOUNDRIES USA for 22nm FDX® test wafer.

1T SiGe Floating Body (FB) RAM for Cryogenic Cache



Memory Candidates	6T SRAM [1]	STT-MRAM [2]	2T Gain Cell [3]	1T FBRAM (Si) [4]	1T FBRAM (SiGe) [This Work]
Cell Size	~50F²	~10F²	20F²	6F²	6F²
Write Time	-	20ns	0.76ns	*10ns	>5ns
Write Voltage	0.75V	-	1.0V	1.8V	1.5V
Write Energy	-	-	-	1.08fJ	0.82fJ
Retention @77K	Static	Non-Volatile	23ms	*2x10⁵ s	*8x10³ s
Write Endurance	-	10¹²	-	10¹⁰	10¹²

*Projected

Fig.1: SiGe Floating Body RAM (FBRAM) on 22nm FDSOI platform with 6F² cell size, 5ns write time and pseudo static data retention-a promising candidate for high density cache in Cryogenic Processor

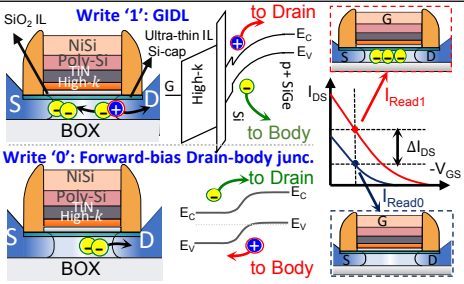


Fig.2 Memory operation of SiGe FBRAM with corresponding band diagrams during Write '1', Write '0' and difference in cell current during read

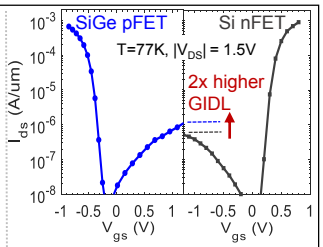
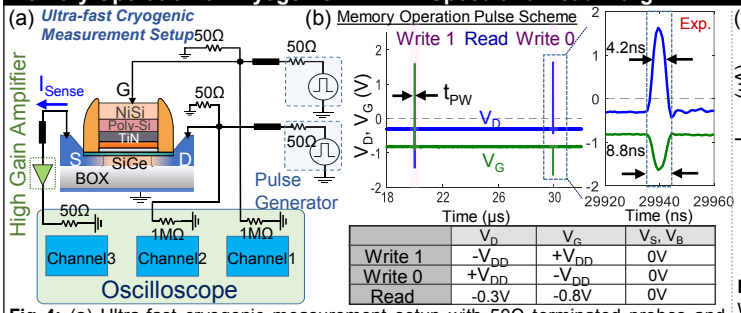


Fig.3: 77K Gate-induced-Drain-Leakage (GIDL) characteristics of n-Si and p-SiGe channel FDSOI FETs with L_G=20nm; SiGe pFET show 2x higher GIDL than Si nFET

Memory Operation of Cryogenic FBRAM: Speed and Read Margin



	V_D	V_G	V_S, V_B
Write 1	$-V_{DD}$	$+V_{DD}$	0V
Write 0	$+V_{DD}$	$-V_{DD}$	0V
Read	-0.3V	-0.8V	0V

Fig 4: (a) Ultra-fast cryogenic measurement setup with 50Ω terminated probes and Low noise amplifier with gain of 10^3 V/A; (b) transient pulse scheme for Write 1, Read and Write 0 operation. Fast programming pulses with Full-Width at Half-Maximum (FWHM) of 4.2ns were probed at the 1MΩ oscilloscope termination.

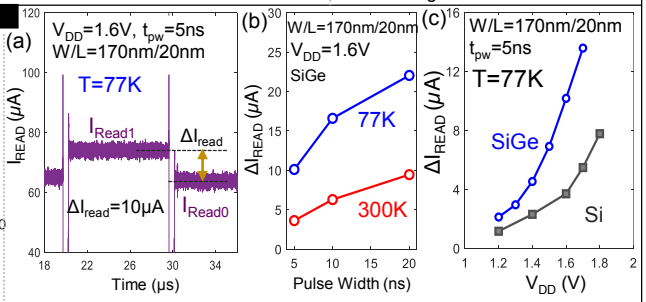


Fig 5: (a) Transient modulation in I_{READ} with ΔI_{read}=10μA obtained, after Write '1' and '0' with 1.6V, 5ns pulse at 77K, indicating the presence and absence of injected electrons in the body (b) Increased ΔI_{read} in SiGe at 77K due to higher transconductance at 77K, (c) Higher ΔI_{read} in SiGe compared to Si due to higher injected electron concentration during write via GIDL.

Multi-Bit Memory Operation in Cryogenic FBRAM at 77K

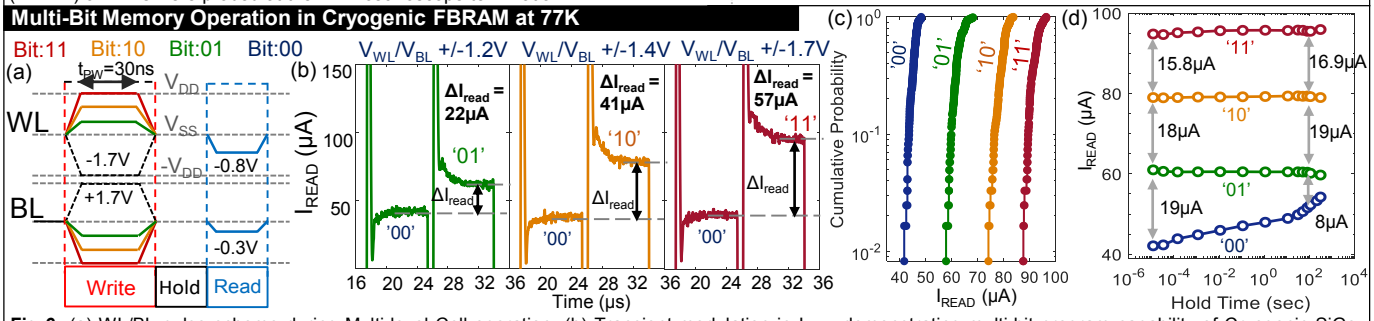
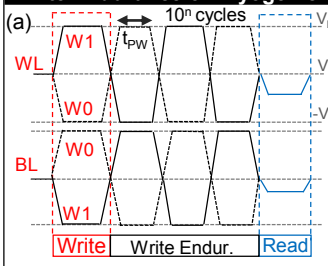


Fig 6: (a) WL/BL pulse scheme during Multi-level Cell operation, (b) Transient modulation in I_{READ} demonstrating multi-bit program capability of Cryogenic SiGe FBRAM, (c) Distribution of well-separated 4 I_{READ} states ('00', '01', '10' and '11') over 120 cycles, (d) Retention characteristics of 4 distinct I_{READ} states show I_{Read} '11' and '10' unaffected upto 300s, owing to suppression of thermally activated SRH recombination at 77K. Δ_{IRead}=9μA (50% of initial ΔI_{READ}) between '00' and '01' observed till 100s. Pseudo-static retention of 8x10³s and worst case retention of 100s was estimated at 77K for 1-bit/cell and 2bit/cell.

Write Endurance of Cryogenic FBRAM



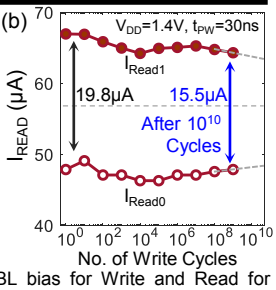


Fig.7: (a) Timing diagram of WL and BL bias for Write and Read for Write-Endurance under bipolar Program-Erase cycles, (b) 80% of ΔI_{read} margin is preserved till 10^{10} cycles of +/- 1.4V endurance pulse at 77K

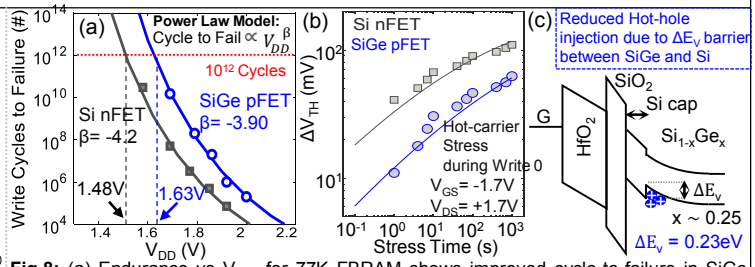
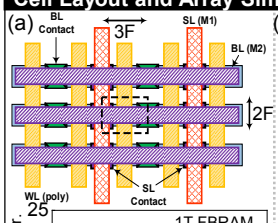
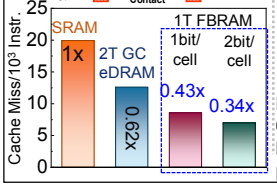
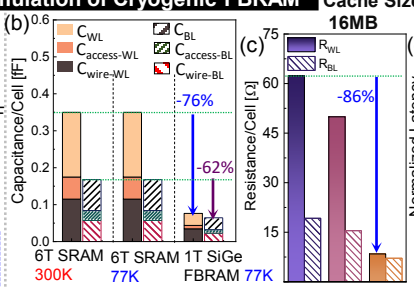


Fig.8: (a) Endurance vs V_{DD} for 77K FBRAM shows improved cycle-to-failure in SiGe compared to Si. >10^{12} write endurance achieved in SiGe with 150mV higher write V_{DD}, (b) Hot-carrier injection during erase operation cause higher degradation in ΔV_{TH} and I_{READ} in Si nFET compared to SiGe pFET, due to (c) presence of valence band offset (ΔE_V~0.23eV) barrier between SiGe channel and thin-Si capping layer.

Cell Layout and Array Simulation of Cryogenic FBRAM







Cache Size 16MB ■ 6T SRAM 300K ■ 6T SRAM 77K ■ 1T SiGe FBRAM 77K

Cache Size 16MB	6T SRAM	6T SRAM	1T SiGe FBRAM [This Work]
Temperature	300K	77K	77K
WL Cap./Cell (fF)	1x	1x	-76%
RWL Res./Cell (Ω)	1x	-20%	-86%
Write Latency	1x	-13%	-49%
Write Energy	1x	-20%	-57%
Read Latency	1x	-13%	-74%
Read Energy	1x	-21%	-66%
EDP w/o cooling cost	1x	-30%	-85%
EDP w cooling cost	1x	2.72x	-40%

Fig 9: (a) Ultra-compact layout area of 6F² enable 57% and 66% reduction in cache miss rate for 1bit and 2bit/cell 1T FBRAM from 6T SRAM, (b) 75%/62% lower WL/BL capacitance/cell and (c) 80% lower WL resistance/cell in 1T FBRAM results in (d) 70%/41% lower read/write latency compared to 6T SRAM for a 16MB cache at 77K, (e) 78% EDP gain in 1T FBRAM compared to SRAM projected at 77K. Including cooling cost at 77K for an ideal Carnot engine, 60% EDP gain over 6T SRAM at 300K can be achieved.