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

W. Chakraborty1, P. Shrestha2, A. Gupta1, R. Saligram1, S. Spetala2, J. Campbell1,2, A. Raychowdhury2, and S. Datta1

1University of Notre Dame, Notre Dame, USA, 2Theiss Research, La Jolla, USA, National Institute of Standards and Technology, Gaithersburg, USA,
3Georgia Institute of Technology, Atlanta, USA  *Email: wchakrab@nd.edu

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 FD-SOI transistor at 77K, for cryogenic cache memory application. The 1T SiGe FBRAM cell (W/L=170/20nm) at 77K exhibits: (a) record write time of ~5ns with write voltage (V\text{Write}) 1.5V; (b) high sense current (I_{\text{Read}} > 75\mu A) with read margin (\Delta I_{\text{Read}} > 14 \mu A); (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 exhibits 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 FD-SOI[5] platform at 77K, showing significant advantages over other candidates, such as faster write time, lower V\text{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 body barrier which, in turn, modulates the drain current (I_{\text{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\text{Write} and higher I_{\text{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 Vg, Vd 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 1M\text{Ω} oscilloscope termination with transmission line delay (td)<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=170/20nm). Fig. 5(a) shows the corresponding transient modulation in read current (I_{\text{Read}}), with \Delta I_{\text{Read}} =10\mu A at 77K, after Write ‘1’ and Write ‘0’, indicating the presence and absence of injected electrons in the body. I_{\text{Read}} for V_{\text{DD}} =1.6V as a function of pulse width (tpw) shows ~2.5x increase in I_{\text{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 I_{\text{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\text{Write} and consequently write energy in SiGe FBRAM for iso-\Delta I_{\text{Read}}. We demonstrate the multi-bit program capability of SiGe FBRAM varying the injected electron concentration with varied V\text{Write}. Fig. 6(a) shows the pulse scheme for multi-bit cell operation, with corresponding transient modulation in I_{\text{Read}} (Fig. 6(b)), demonstrating four distinct I_{\text{Read}} levels (corresponding to 2bit/cell) achieved by tuning V\text{Write}. For writing bits ‘00’, ‘01’, ‘10’ and ‘11’, V_{\text{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_{\text{Read}} states (Fig. 6(c)). Retention characteristics of the states for 77K are shown in Fig. 6(d), I_{\text{Read}} ‘11’ and ‘10’, remains unaffected up-to 300s, owing to suppression of thermally activated Shockley-Read-Hall (SRH) recombination at 77K. The I_{\text{Read}} ‘00’ and ‘01’ approach each other during hold time. Still, \Delta I_{\text{Read}} =9\mu A (50% of initial \Delta I_{\text{Read}}) was observed till about 100s. Thus, pseudo-static retention of 8x10^5 and worst case retention of 100s was estimated at 77K for 1-bit/cell and 2bit/cell respectively, extrapolating I_{\text{Read}} to 50% initial \Delta I_{\text{Read}}. Fig. 7(a) shows the timing diagram of pulse scheme during write-endurance under bipolar program-erase cycles. 80% of initial \Delta I_{\text{Read}} is preserved till 10^10 cycles of +/-1.4V, 300s pulse, as shown in Fig. 7(b). Endurance vs V_{\text{Write}} for 77K FBRAM (Fig. 8(a)), shows improved cycle-to-failure (failure condition: \Delta I_{\text{Read}} < 50% initial \Delta I_{\text{Read}}) in SiGe compared to Si. Power law model fitting suggests 10^12 write endurance can be achieved in SiGe with 150mV higher V_{\text{Write}} compared to Si cell, enabling broader design space for co-optimizing \Delta I_{\text{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 (E_{V\text{B}} ~ 0.23eV) barrier between SiGe channel and thin-Si capping layer (Fig. 8(c)) [8]. This was further validated with \Delta V_{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 10^5 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/10^9 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 \Delta I_{\text{Read}}, 0.15V lower V_{\text{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.

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

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

- **Memory Candidates**: 1T SRAM [1], 2T MRAM [2], 3T GFRAM [3], 3T GFRAM [Si][4]
- **Cell Size**: -50F, -10F, 20F
- **Write Time**: 20ns
- **Write Voltage**: 0.7V, 1.0V, 1.8V
- **Write Energy**: -1.1kW
- **Retention**: 7K Static Non-Volatile 23mS
- **Endurance**: -10^10

**Fig. 2** Memory operation of SiGe FBRAM 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- and p-SiGe channel FDSOI FETs with Lp=20nm. SiGe pFET show 2x higher GIDL than Si nFET

**Memory Operation of FFRAM: Speed and Read Margin**

- **Ultra-fast Cryogenic Measurement Setup**: 500µA, 5ns current probe and low noise amplifier with gain of 10^2 V/A
- **Transistor and transient scheme for Write 1, Read, and Write 0 operation.** Fast programming pulses with Full-Width at Half-Maximum (FWHM) of 2ns and energy of 100nJ were used.

**Fig. 4** (a) Ultra-fast cryogenic measurement setup with 500µA terminated probes and low noise amplifier with gain of 10^2 V/A. (b) Transient pulse scheme for Write 1, Read, and Write 0 operation. Fast programming pulses with FWHM of 2ns and energy of 100nJ were used.

**Multi-bit Memory Operation in Cryogenic SiGe FBRAM at 77K**

- **Bit:11 Bit:01**
- **Write Endurance of Cryogenic SiGe FBRAM**
- **Cell Layout and Array Simulation of Cryogenic SiGe FBRAM**

**Fig. 5** The effect of pulsed conditions on 1T SiGe FBRAM compared to Si nFET. SiGe pFET shows lower capacitance and higher GIDL compared to Si nFET.

**Fig. 6** (a) W/L=170nm/20nm SiGe FET with W/L=170nm/20nm 1T SiGe FBRAM at 77K. (b) Time evolution of VDD and VSS during WRITE (WRITE) and READ cycles.

**Fig. 7** (a) Timing diagram of WL and BL bias for Write and Read for W/L=170nm/20nm 1T SiGe FBRAM. (b) Endurance under regular program/erase cycles. (c) Write endurance under regular program/erase cycles.

**Fig. 8** (a) Endurance vs. VDD for 77K SiGe FBRAM compared to Si nFET. (b) Hot-cotier injection during erase operation causes higher degradation in ΔVDD and ΔVSS in Si nFET compared to SiGe pFET, due to (c) presence of valence band offset (ΔE = 0.3eV) and (d) SiGe channel and thin-Si capping layer.

**Fig. 9** (a) Ultra-compact layout area of 6F2 enable 57% and 66% reduction in cache miss rate for 1bit and 2bit/2cell 1T SiGe FBRAM from 6T SRAM. (b) 75%/82% lower WL/BL capacitance/cell and (c) 90%/92% lower WL resistance/cell in 1T SiGe 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 SiGe FBRAM compared to SRAM projected at 77K. Including cooling cost at 77K for an ideal Carnot cycle, 60% EDP gain over 6T SRAM at 300K can be achieved.