Ternary Volatile Random Access Memory based on Heterogeneous Graphene-CMOS Fabric

Santosh Khasanvis*, K. M. Masum Habib, Mostafizur Rahman, Pritish Narayanan,
Roger K. Lake and Csaba Andras Moritz
*khasanvis@ecs.umass.edu

***Abstract*— Graphene is an emerging nano-material that has garnered immense research interest due to its exotic electrical properties. It is believed to be a potential candidate for post-Si nanoelectronics due to high carrier mobility and extreme scalability. Recently, a new graphene nanoribbon crossbar (xGNR) device was proposed which exhibits negative differential resistance (NDR). In this paper, we present an approach to realize multistate memories, enabled by these graphene crossbar devices. We propose a ternary graphene nanoribbon tunneling volatile random access memory (GNTRAM) and implement it using a heterogeneous integration with CMOS transistors and routing. Benchmarking is presented with respect to state-of-the-art CMOS SRAM and 3T DRAM designs. Ternary GNTRAM shows up to 1.77x density-per-bit benefit over CMOS SRAMs and 1.42x benefit over 3T DRAM in 16nm technology node. Ternary GNTRAM is also up to 9x more power-efficient per bit against low-power CMOS SRAMs during stand-by, while maintaining comparable performance to high-performance designs. Thus GNTRAM has the potential to realize ultra-dense nanoscale memories exceeding those achievable by mere physical scaling. Further improvements may be possible by using graphene more extensively, as graphene transistors become available in future.**

Keywords- Graphene Nanoribbons; NDR; Multistate Memory; Heterogeneous Integration.

I. INTRODUCTION

SRAM has been widely used to implement on-chip embedded memory due to its high performance. Over the years, on-chip SRAM caches have been steadily increasing in density to meet the computing needs of high performance processors. In order to maintain this historical growth in memory density, SRAM bit cells have been aggressively scaled down for every generation along the semiconductor technology roadmap. However, there has been a slowdown in SRAM area scaling from 50% to 30% reduction per generation [1] due to several challenges such as increased variability at nanoscale [2][3]. This calls for new concepts and technological improvements to meet growing performance demands.

One such concept is to use memory cells which have more than two stable states. Such a multi-state memory provides a new dimension for scaling as an alternative to physical scaling, by compressively storing multiple bits in a single cell. This is enabled by emerging nanoscale materials, like graphene and unique material interactions between novel device structures.

Graphene is an atomically-thin allotrope of carbon and is considered to be a potential candidate for post-Si nanoscale computing systems [4]. It exhibits extra-ordinary electrical and thermal properties featuring Dirac fermion [5] with very high conductivity [6] and extreme scalability. Its planar structure also makes it compatible with current CMOS fabrication processes [7]. Several graphene based transistors have been proposed [8]-[12], however challenges still exist which preclude their use in digital systems [13]. A novel bi-layer graphene nanoribbon crossbar tunneling device (xGNR) was reported recently [14][15], which exhibits negative differential resistance (NDR). This xGNR NDR device has potential applications in multi-state logic and memory circuits.

Multi-state circuits using NDR based resonant tunneling diodes (RTDs) have been extensively researched in the past [16]-[19]. However, RTDs were implemented using non-lithographic processes and III-V technology. Such processes were expensive and incompatible with those for Si, which prohibited its integration with conventional Si technology [20]. Due to technological and economical barriers, RTDs using III-V materials could only be used in niche applications. On the other hand, Graphene based NDR devices like xGNR overcome such integration challenges and have the potential to be used in mainstream applications.

In this paper, we propose a ternary volatile memory using the xGNR device for on-chip multi-state memories, called ternary graphene nanoribbon tunneling random access memory (GNTRAM). The contributions include (i) introducing and validating the ternary GNTRAM concept, (ii) heterogeneous graphene-CMOS implementation and (iii) benchmarking against state-of-the-art CMOS SRAM and 3T DRAM memory cells. Our evaluations show that the proposed ternary GNTRAM has up to 1.77x density-per-bit benefit against CMOS SRAMs and 1.42x benefit against 3T DRAM in 16nm technology node. Ternary GNTRAM is also up to 9x more power efficient per-bit when compared against the low-power CMOS designs in idle periods, while still having comparable performance to high-performance designs. This work is the first step towards high-density multi-state volatile memories using graphene. Further work on device and circuit level techniques to increase the number of memory states per cell could potentially lead to ultra-dense multi-state nanoscale memories. Even further improvements may be possible by using graphene more extensively instead of silicon MOSFETs, as advances are made in graphene technology.

We acknowledge support from the Focus Center Research Program (FCRP) - Center on Functional Engineered Nano Architectonics (FENA) and the Center for Hierarchical Manufacturing (CHM), UMass Amherst.

K. M. Masum Habib and Roger K. Lake are with University of California Riverside. The rest of the authors are with University of Massachusetts at Amherst.

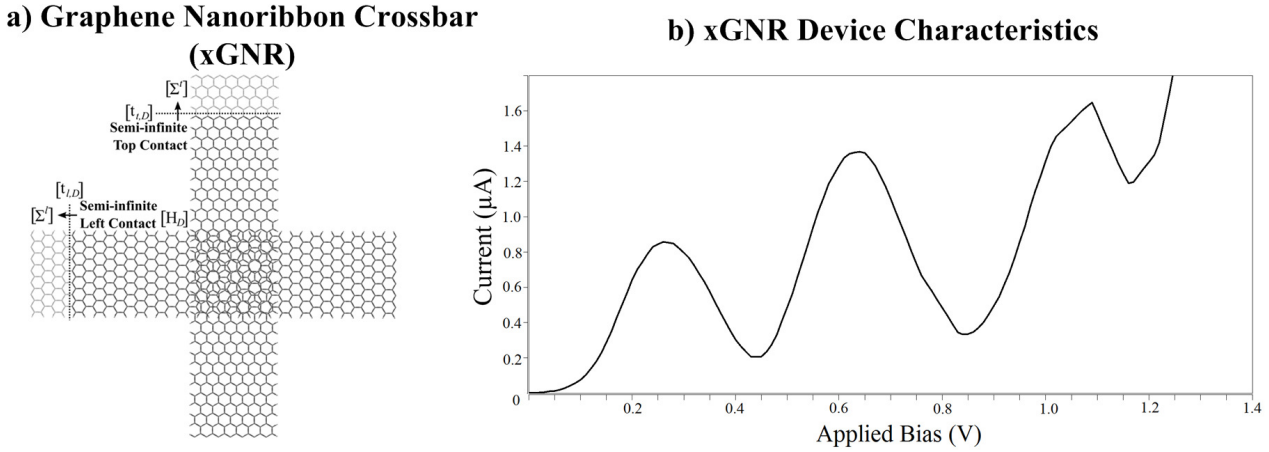

Figure 1 (a) Atomistic geometry of the GNR crossbar. Two hydrogen passivated relaxed armchair type GNRs are placed on top of each other at a right angle with a vertical separation of 3.35 Å. The relaxation was done using Fireball. The extended parts of the GNRs are used as contacts. A bias is applied by independently contacting each GNR such that one is held at ground while the other has a potential applied to it.
(b) Simulated I-V characteristics of the crossbar structure exhibiting NDR with multiple current peaks and valleys.

The rest of the paper is organized as follows. Section II provides a background on the xGNR device and previous work based on this device. Section III proposes a new ternary memory cell and Section IV describes a physical implementation with heterogeneous integration between CMOS and Graphene. Methodology and benchmarking is presented in Section V followed by conclusion in Section VI.

II. Background and Previous Work

A. Graphene Nanoribbon Crossbar (xGNR) Device

The graphene nanoribbon crossbar shown in Fig. 1a consists of two semi-infinite, H-passivated, armchair type GNRs (AGNRs) with one placed on top of the other at right angles and a vertical separation of 3.35 Å in between [14][15]. The GNRs are chosen to be 14-C atomic layers [(3n + 2) ~1.8 nm] wide to minimize the bandgap resulting from the finite width. The bandgap of the 14-AGNR calculated from density functional theory (DFT) code Fireball [21][22] is 130 meV which is in good agreement with Son *et al.* [23]. The contacts are single layer GNRs modeled by the self-energies of semi-infinite leads. A bias is applied to the top GNR with respect to the bottom one. Assuming the majority of the potential drop occurs in between the two nanoribbons, the potential difference between the GNRs is the applied bias.

The current voltage (I-V) characteristic of the xGNR is calculated using the first principle DFT coupled with the non-equilibrium Green's functions formalism (NEGF). The Hamiltonian matrix element used in the NEGF calculations are generated from the quantum molecular dynamics, DFT code, Fireball, using separable, nonlocal Troullier-Martins pseudopotentials [24], the BLYP exchange correlation functional [25][26], a self-consistent generalization of the Harris-Foulkes energy functional [27]-[30], and a minimal sp^3 Fireball basis set. The radial cutoffs of the localized pseudoatomic orbitals forming the basis are r_c^{1s} = 4.10 Å for hydrogen and r_c^{2s} = 4.4 Å and r_c^{2p} = 4.8 Å for carbon [31]. These matrix elements are used in the recursive Green's function (RGF) algorithm to calculate the transmission and the current as described in [32].

The simulated I-V characteristic of the xGNR is shown in Fig. 1b exhibiting negative differential resistance (NDR) with multiple peak and valley currents, which makes it suitable for RTD-based applications [33]. The NDR is attributed to the localization of the electronic states near the cut-ends of the GNRs [15]. The electronic waves are reflected back from these cut-ends. The interference between the incident and the reflected electronic waves give rise to these localized states which, in turn, results in resonances and anti-resonances in the transmission. The strengths of the resonant peaks in the transmission are strongly modulated by the applied bias leading to NDR. This phenomenon is analogous to the stub effect in microwave theory. In this case the GNR cut-ends act as open ended stubs for the electrons.

B. Application of xGNR Device as a Latch

We explore one application where xGNR devices can be used in a latch configuration for volatile memory [34]. A latch can be built to leverage on NDR behavior by connecting two xGNRs in series, as shown in Fig. 2a. One of the devices (xGNR1) is connected to supply voltage (Vref) and acts as a pull-up device. The other device (xGNR2) is connected to ground terminal acting as the pull-down device. The circuit schematic of this configuration is shown in Fig. 2b. The common terminal between these devices acts as the state node (SN) where information is latched. DC load line analysis of this configuration exhibits three stable states A, B & C under applied voltage, as shown in Fig. 2c. Consider state C as an example. When the state node is at voltage corresponding to state C, a constant static current flows through the devices. Any external perturbation may cause the state node voltage to either increase or decrease. This is countered by restoring currents which pull-up (or pull-down) the state node when the external noise brings its voltage down (or up), as shown in the figure. The magnitude of the restoring current is given by the

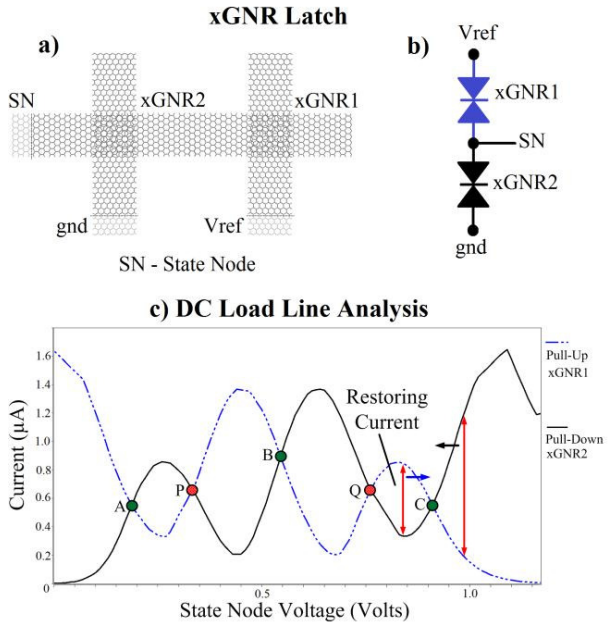

Figure 2 (a) xGNR latch configuration; (b) Circuit schematic; and (c) DC load line analysis showing multiple stable states.

difference between the pull-up and pull-down currents. As long as the noise current is smaller than this restoring current, the state information is retained.

States denoted by P and Q in Fig. 2c are unstable and hence the corresponding voltages are the transition voltages. Consider state Q where any external noise would cause the state node voltage to transition to one of the surrounding states depending on the direction of the perturbation. The details of latch operation are explained in our previous work [34]. This xGNR series configuration can be used as a binary latch or multi-state latch, where the information is stored on the common terminal (the state node) of the xGNR devices. Previously, we had explored the validity of this concept by building a binary memory cell [34]. We now build on this concept to propose a ternary latch based memory cell, where all three states are used to store information.

III. Proposed Ternary Memory Cell

The xGNR latch with three stable states can be used to compressively store more than one bit electrically in a single cell. To build a volatile random access memory cell using this memory core, access to the state node is required. This is achieved with the help of FETs for cell selection, write and read operations [35]. A static implementation using this scheme would however lead to large static currents and thus large stand-by power dissipation.

We propose a dynamic memory cell to enable a low-leakage volatile ternary graphene-tunneling random access memory (GNTRAM) as shown in Fig. 3. This cell uses all three stable states (A, B & C in Fig. 2c) to store information. The xGNR devices are arranged in a latch configuration and a write FET is used to access the state node. To mitigate static power, we switch OFF the xGNR latch and use a capacitor (C_{SN}) at the state node to store the voltage value written into the cell. The state node capacitance is isolated from the power/ground lines during stand-by with the help of a Schottky Diode and a sleep FET. The Schottky diode provides current rectification during stand-by and helps preserve the state node voltage. Two read FETs are used to read the stored information. The cell operation is explained next.

A. Write Operation

During a write operation, the required cell is selected by activating the corresponding write-line and applying the required voltage onto the data-line. Here, the value of the applied voltage on the data line denotes the state to be written and is in ternary representation (0V – logic 0, 0.6V – logic 1 and 1.0V – logic 2). The voltage values are chosen based on the voltages at which stable states occur in the xGNR latch. Fig. 4 shows the DC load line analysis for the xGNR latch in conjunction with the Schottky diode and the sleep FET. The stable states are marked with their respective logic states in the graph.

Consider that the state node is initially at logic 0. To write logic 1, the appropriate voltage (0.6V) is applied on the data line. The write signal is applied which starts charging the state capacitance. Once the capacitance is charged to a voltage close to the required value, the restore signal is applied. This supplements the write operation by providing restoring currents to pull-up the state node. After the voltage value is written onto the state capacitance, the word-line is switched-off followed by the data-line. The restore signal is still maintained to latch the information and ensure that the switching transients do not

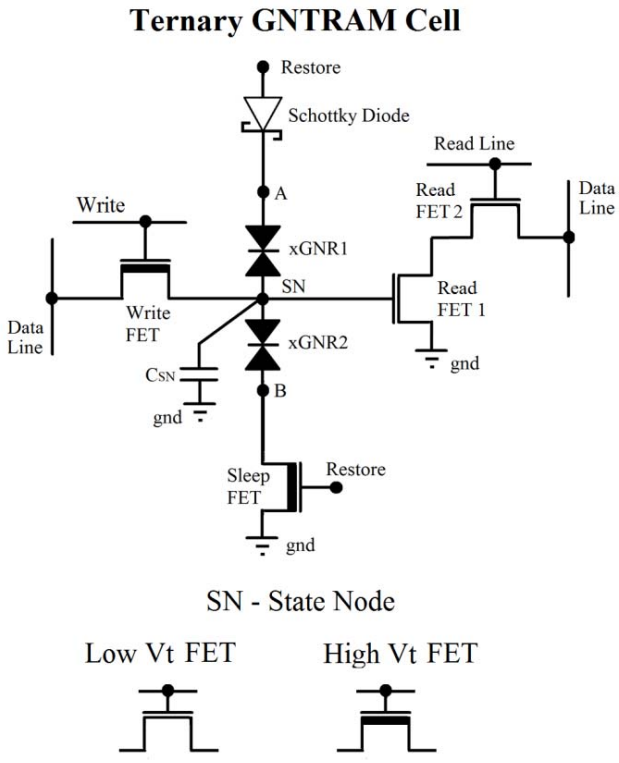

Figure 3. Proposed Ternary GNTRAM Circuit Schematic

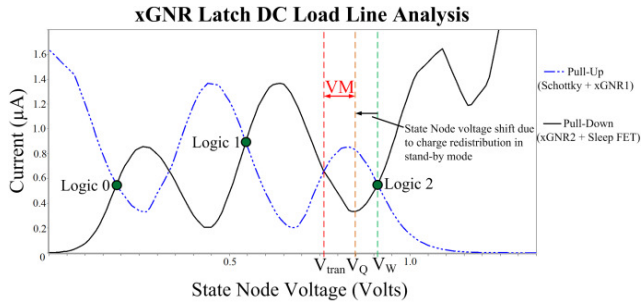

Figure 4. DC Load Line Analysis for xGNR latch including Schottky Diode and Sleep FET showing multiple stable states – Logic 0, 1 and 2.

affect the state node voltage. After the stored voltage is stabilized, the restore signal is switched OFF and the information is stored dynamically on the state capacitance. Similarly, logic 2 is written by applying the corresponding data voltage and charging up the state capacitance.

To write a logic 1 or logic 0 when the state node is initially at logic 2, the appropriate voltage is applied on data line. This results in a discharge operation of the state capacitance when the write signal is activated and proceeds along the same lines as discussed above.

B. *Read Operation*

The state node is used to gate read FET1 and hence is isolated from the output data line. This scheme ensures that the read operation is non-destructive. The ON-current through the read path is determined by the value of the state node voltage which gates read-FET1. Since the voltage level stored is different for each of the logic states, the read current varies in each case. This enables the detection of multiple voltage levels at the data output.

To initiate a read operation, the data line is pre-charged to full VDD and then the read signal pulse is applied for a pre-determined time. This read time is chosen such that when logic 2 is stored at the state node, the data line is completely discharged and can be deciphered as logic 2. A stored logic 1 would cause read-FET1 to have a higher ON resistance compared to that of logic 2. Thus applying the read pulse would lead to the data line being only partially discharged to an intermediate value, which can be deciphered as logic 1. When logic 0 is stored, the read-FET1 is completely switched OFF and the data line remains at VDD. Hence this scheme results in an inverting read-out mechanism. Such a pull-down scheme is used because nMOS transistors are suited for pull-down operation, obviating the need for gate voltage boosting to overcome the threshold voltage drop in a pull-up scheme.

C. *Restore Operation*

In an on-chip cache, data access is typically centered on a fixed number of words due to the principle of locality. Thus a major part of the cache cells are in a stand-by mode most of the time. A static scheme would have lead to a tremendous amount of static power dissipation when the memory is idle. In GNTRAM, the data is stored on a capacitor during stand-by, thus mitigating static power dissipation. However, the stored charge starts to leak and has to be restored. This is done by asserting the restore signal, which switches-ON the sleep FET and the Schottky diode. The restoring currents flowing through the state node charge-up the capacitor and restore its value, as long as the noise/leakage currents are small enough to be countered. Unlike a DRAM, the GNTRAM restore operation does not require a read followed by write to be able to restore the charge and is a low-power operation.

GNTRAM offers a separate channel for charge restoration enabled by the unique properties of the xGNR latch. The restore operation is independent of read and write-operations. This considerably eases the restoration without the need for complex restore control schemes.

D. *Circuit Implementation*

In order to maximize the retention time, the circuit is implemented as an asymmetric cell [36] as shown in Fig. 3, i.e. performance-critical paths use low threshold voltage (V_t) devices while others use high-V_t devices to minimize leakage. Since the write and sleep FETs are directly connected to the state node, they are implemented using high-V_t transistors to minimize charge leakage during stand-by. The read FET1 has to be necessarily a low-V_t device to distinguish between the three stored states. Read FET2 can have a high-V_t for a low-power design or low-V_t for a high-performance design.

The value of the state capacitance is determined by two factors – (i) the value of the parasitic capacitances of the diode and the sleep FET and (ii) the worst case voltage margin. Due to the parasitic capacitances, the charge written onto the state node is immediately redistributed as soon as the cell goes into stand-by. This is denoted by the voltage level V_Q in Fig.4, for the case of storing logic 2. This is the final quiescent voltage at the state node as soon as the write and restore signals are deactivated and the cell goes into stand-by mode. If V_Q falls below transition voltage (V_{tran} in Fig. 4), the restore operation causes a state transition to logic 1 instead of restoring logic 2 at the state node. Thus the total state capacitance (C_{SN}) should be large enough to ensure that the state information is not lost.

The quiescent voltage (V_Q) should ensure that enough voltage-margin (VM) is maintained for dynamic data retention. This is shown in Fig. 4. This voltage margin determines the maximum time available for the information to be stored dynamically, before a restore operation needs to occur. By choosing an appropriate V_Q, the retention time can be optimized. The minimum value of the total capacitance at the state node can be derived using the following relation:

$$C_{SN}.V_w = (C_{SN} + C_{PT}).V_Q \quad (1)$$

In (1), C_{SN} is the total capacitance at the state node, which includes the explicit capacitance to be formed at the state node, parasitic diffusion capacitance of the write FET, gate capacitance of read FET1 and the capacitance due to routing lines. C_{PT} is the total parasitic capacitance, which includes the diffusion capacitance of the sleep FET and the capacitance of the Schottky diode. V_W is the voltage to which the state node is charged during a write operation. The available voltage margin for retention is given by the difference between V_Q and V_{tran}.

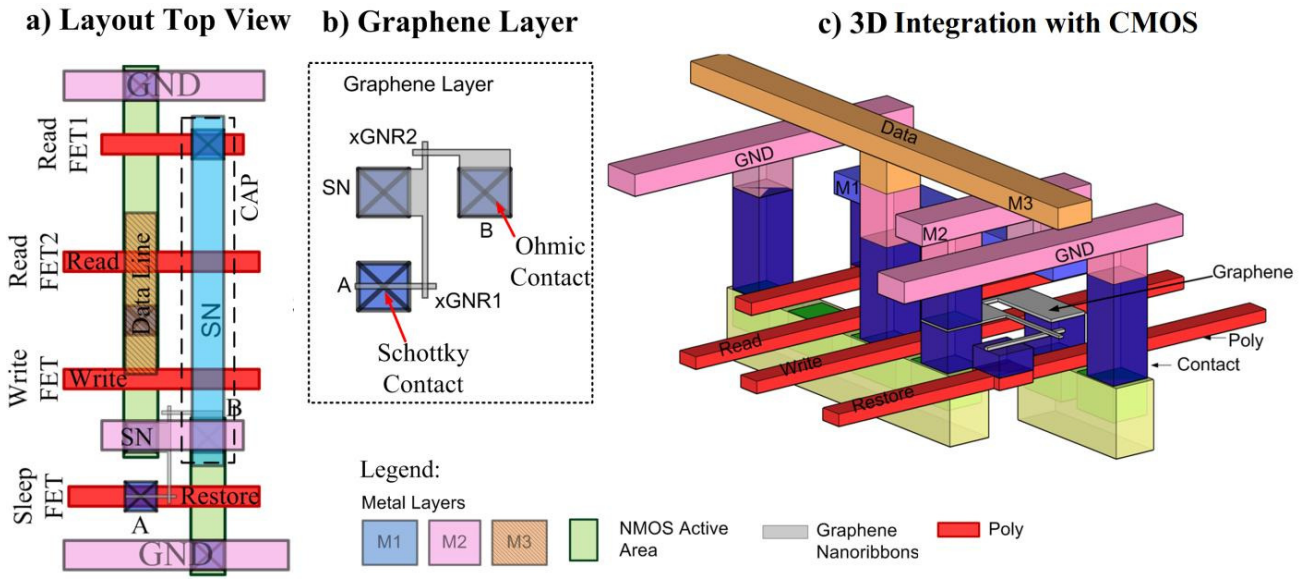

Figure 5. Proposed Ternary GNTRAM Layout - (a) Layout Top View, (b) Graphene Layer, (c) 3D Integration between CMOS and Graphene.

Alternate implementations are possible where pMOS is used instead of NMOS as a write FET. This could be beneficial from a low-power perspective since a pMOS can easily pull-up the state node without the need for gate overdrive, as in the case of an nMOSFET. Since the stored logic 0 is at voltage of about 0.15V, a complete discharge is not even required when writing logic 0. The trade-offs with using PMOS would be (i) lower performance and (ii) area overhead due to the separation needed between n-well and p-well. The cell design can be specifically tuned to the required application.

IV. Physical Layout

A cross-technology heterogeneous implementation is used between CMOS and graphene [34], as shown in Fig. 5. The MOS transistors are formed at the bottom layer on the substrate. The xGNR devices are implemented in a graphene layer on top of the MOS layer. Interfacing between these layers is done with the help of metal vias. GNRs can form either Ohmic contacts or Schottky contacts with metals, depending on whether they are metallic or semiconducting [37][38]. This feature is leveraged to realize the Schottky diode with the help of a Schottky contact between a narrow semiconducting armchair GNR and metal, as shown in Fig. 5b. The rest of the graphene-metal contacts are Ohmic to ensure proper operation and this is achieved by using wide GNRs [39]. Both Schottky diode and Sleep FET receive the same restore signal. Hence the layout is arranged so that the restore signal reaches both devices almost simultaneously. The data line is multiplexed between read and write-operations since only one of these operations is performed on a memory cell at a given time.

A lithography-friendly grid-based layout is used with minimum sized nMOS transistors for high density and ease of fabrication. Some of these can be replaced with pMOS depending on the application. Routing is achieved with the help of a conventional metal stack. The state capacitor can be implemented either as a trench or as a stacked capacitor over the state node routing area shown in Fig. 5a.

TABLE I. Design Rules

1D Gridded Design [41]	*M1, M2 Interconnect*	*Poly*
Pitch (16nm technology node)	40~60 nm	60~80nm

V. Methodology and Benchmarking

HSpice circuit simulator was used to simulate and verify the operation and for benchmarking against the state-of-the-art. The xGNR devices were modeled as piece-wise linear voltage controlled current sources, based on current-voltage data points. A generic integrated circuit Schottky diode model was used for a first order analysis and 16nm CMOS PTM models [40] were used to simulate the read, write and sleep FETs. The value of the state capacitance was chosen to be 200aF for proper circuit behavior, based on the discussion in Section III. This ensures that when a restore signal is applied at a period of 0.7µs, the state node is brought up to the required stable point. A higher capacitance value would lead to a longer retention time.

The simulation waveforms for write and read operations are shown in Fig. 6a. The state node is initialized to 0 and logic 1 is first written and then read. After this, all possible transitions between the three states are simulated and verified for both read and write operations. Fig.6b and Fig. 6c show the data output signals in detail. Restore operation is performed at a period of 0.7µs, as shown in Fig. 6d for the case of restoring logic 2. The circuit operated as outlined in Section III.

For physical layout design and evaluation, 1-D Gridded design rules [41] were used to compare the area of GNTRAM cell with Gridded 8T SRAM cell [42] in 16nm technology

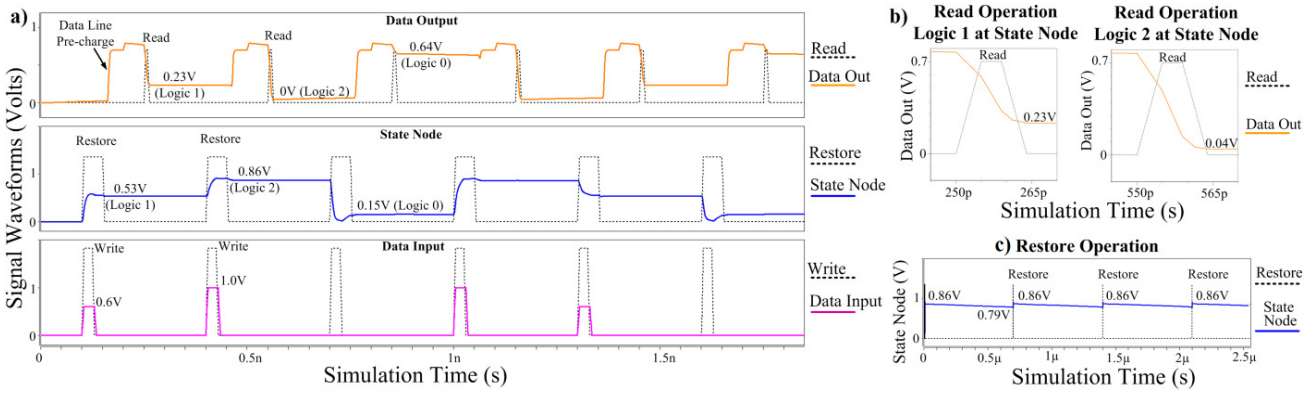

Figure 6. (a) Simulation Waveforms showing GNTRAM Operation, (b) Read Operation for Logic 1, (c) Read Operation for Logic 2 and (d) Restore Operation .

node. Regular 6T CMOS SRAM scaled to 16nm technology node was also used for benchmarking. Area scaling was done based on a wide range of design rules published by the industry. For each parameter (such as metal pitch spacing, etc.), scaling factors across technology nodes were determined. The method is outlined in [43]. This methodology resulted in a range of values for 6T SRAM cell area for a range of design rules. PTM RC models [40] based on scaled interconnect dimensions and 16nm PTM transistor models [40] were used for simulation with HSpice for power and performance evaluation of 16nm CMOS 6T SRAM and Gridded 8T SRAM. Both low power and high performance 6T and 8T SRAM cell designs are considered for comparison since, ternary GNTRAM uses an asymmetric cell design with both low-power and high-performance transistors.

3T DRAM was also investigated for benchmarking since it is a potential candidate for on-chip caches in advanced technology nodes [44][45]. The 3T DRAM cell was designed using 16nm PTM transistor models and the physical layout was done on the same lines as the GNTRAM. The 3T DRAM circuit and layout are shown in Fig. 7. It was simulated using HSpice for power and performance evaluations. Area evaluation was done using the same grid-based design rules as GNTRAM.

Table I shows the design rules used and Table II shows the comparison results.

A. *Area Evaluation*

Ternary GNTRAM showed significant density advantage compared to the other 16nm CMOS RAMs. Although the physical cell size is comparable to that of the SRAMs and the 3T DRAM, ternary GNTRAM's density benefit comes from the fact that it stores more than one bit per cell (*log3/log2* bits per cell). In particular, ternary GNTRAM showed a density-per-bit benefit of up to 1.68x vs. scaled 6T CMOS SRAM, 1.77x vs. gridded 8T CMOS SRAM and 1.42x vs. the 3T DRAM in 16nm technology node.

Considering the current SRAM scaling trend, CMOS SRAM when advanced by one or two technology generations after 16nm node, would have about the same area as ternary GNTRAM in 16nm node. This benefit can further be improved if more states are available per cell, thus providing an alternative to physical scaling. As graphene technology matures, the availability of graphene transistors would enable a monolithic graphene fabric with potentially ultra-dense nanoscale multi-state memories.

B. *Power Evaluation*

In terms of active power, the ternary GNTRAM cell power was comparable to that of high-performance CMOS SRAMs. However when power-per-bit is considered, GNTRAM showed up to 1.84x benefit against CMOS high-power SRAM designs, while being comparable to that of the low power designs. Ternary GNTRAM also showed up to 1.75x active power-per-bit benefit against the 3T DRAM in 16nm node.

Ternary GNTRAM was 9x more power-efficient during idle period against the low-power scaled 6T CMOS SRAM, and 5.63x more power-efficient against low-power 8T gridded SRAM in 16nm node. These benefits are because of two reasons – (i) GNTRAM is dynamic and hence no static paths exist to contribute to idle power, and (ii) GNTRAM stores more than one bit per cell thus amortizing leakage costs. The 3T DRAM exhibits lower stand-by power than GNTRAM

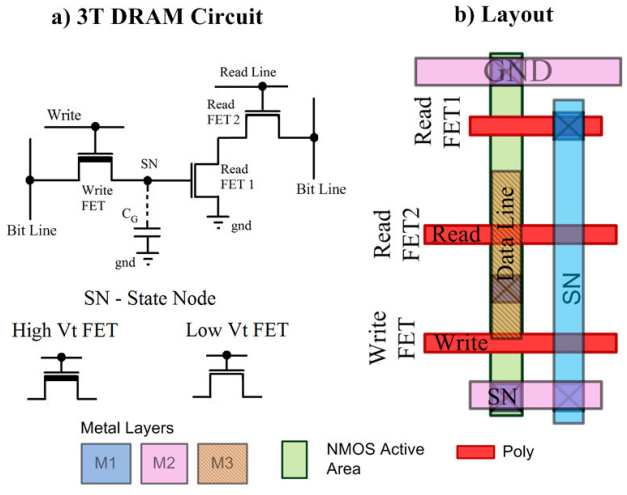

Figure 7. 3T DRAM – (a) Circuit Schematic, and (b) Physical Layout

TABLE II. GNTRAM BENCHMARKING

		Ternary GNTRAM (Per Cell, 1.585 bits)	Ternary GNTRAM (Per Bit)	16nm CMOS 6T SRAM Cell		16nm CMOS Gridded 8T SRAM Cell		16nm 3-T DRAM Cell
				Low Power	High Performance	Low Power	High Performance	
Area Comparison (μm^2)		0.03 – 0.0608	0.019 – 0.038	0.026 – 0.064		0.0336 – 0.0672		0.0264 – 0.054
Power Comparison	Active Power (μW)	2.05 – 2.15	1.29 – 1.35	1.21 – 1.16	2.1 – 2.2	1.45 – 1.47	2.38 – 2.44	2.12 – 2.15
	Stand-by Power (pW)	22.04 – 22.07	13.9 – 13.92	124.18 – 125.12	6152 – 6157	78.38 – 78.44	15552 – 15556	6.49 – 7.01
Performance	Read Operation (ps)	8.98 – 9.8		17.39 – 21.03	8.35 – 9.25	14.82 – 16.08	7.68 – 7.96	9.18 – 9.68
	Write Operation (ps)	16.26 – 16.39		67.27 – 67.54	18.44 – 18.46	58.37 – 63.18	16.62 – 19.16	10.45 – 10.97

since it has lesser number of leakage paths.

C. Performance Evaluation

Ternary GNTRAM was comparable in read performance to high-performance CMOS SRAMs since it uses high-performance devices in its read path. The asymmetric cell design (multi-Vt transistors) thus enables high-performance while reaping the benefits due to low power. An asymmetric (multi-Vt) approach was necessary in ternary GNTRAM because the read FET1 needs to have a low-Vt to successfully differentiate between three stored states. The write performance of GNTRAM is better than the SRAM designs because of the boosted gate voltage to overcome the threshold voltage drop, when storing logic 1 and logic 2 at the state node. The 3T DRAM performs better than GNTRAM during write operation because the state node capacitance to be charged is lower in 3T DRAM.

VI. CONCLUSION

Multi-state memory concept was presented in this paper as an alternative to physical scaling for continued density improvement of on-chip caches in nanoscale computing systems. A ternary graphene nanoribbon crossbar tunneling random access memory (GNTRAM) cell was presented, which was enabled by new nanomaterials like graphene and unique graphene nanoribbon structures. Ternary GNTRAM memory was implemented with a heterogeneous integration between CMOS and graphene technologies. Benchmarking against state-of-the-art CMOS RAM designs showed that ternary GNTRAM exhibited significant benefits, stemming from compressively storing more than 1 bit per cell.

Such a ternary memory would require additional circuits to interface with binary logic. Since such interfacing circuits would be shared across multiple cells or banks, it is expected that the area overhead would be small. Investigation is required to evaluate the architectural performance overhead due to conversion between binary-ternary logic systems.

This work takes the first step towards multi-state volatile memory based on graphene-nanoribbon devices. Future work would explore device and circuit techniques to increase the number of memory states per cell. This could potentially lead to significant density and power benefits per-bit over binary memories to exceed those achievable by mere physical scaling. As progress is made in graphene technology, further benefits may be expected by replacing Si MOSFETs with graphene transistors.

REFERENCES

[1] Smith, K.C.; Wang, A.; Fujino, L.C.; , "Through the Looking Glass: Trend Tracking for ISSCC 2012," Solid-State Circuits Magazine, IEEE , vol.4, no.1, pp.4-20, March 2012

[2] Itoh, K.; , "Embedded Memories: Progress and a Look into the Future," Design & Test of Computers, IEEE , vol.28, no.1, pp.10-13, Jan.-Feb. 2011.

[3] Qazi, M.; Sinangil, M.E.; Chandrakasan, A.P.; "Challenges and Directions for Low-Voltage SRAM," Design & Test of Computers, IEEE , vol.28, no.1, pp.32-43, Jan.-Feb. 2011.

[4] The International Technology Roadmap for Semiconductors http://www.itrs.net/.

[5] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, and A. A. Firsov, "Two-dimensional gas of massless dirac fermions in graphene," Nature, vol. 438, no. 7065, pp. 197–200, November 2005.

[6] T. Ando, "Exotic electronic and transport properties of graphene," Physica E: Low-dimensional Systems and Nanostructures, vol. 40, no. 2, pp. 213 – 227, 2007.

[7] de Heer, W.A.; Berger, C.; Conrad, E.; First, P.; Murali, R.; Meindl, J.; , "Pionics: the Emerging Science and Technology of Graphene-based Nanoelectronics," Electron Devices Meeting, 2007. IEDM 2007. IEEE International , vol., no., pp.199-202, 10-12 Dec. 2007.

[8] G. Fiori and G. Iannaccone, "On the Possibility of Tunable-Gap Bilayer Graphene FET," IEEE Electron Device Letters, vol. 30, no. 3, pp. 261–264, March 2009.

[9] Fiori, G.; Iannaccone, G.; "Ultralow-Voltage Bilayer Graphene Tunnel FET," IEEE Electron Device Letters, vol. 30, no. 10, pp. 1096–1098, Oct 2009.

[10] K.-T. Lam and G. Liang, "A computational evaluation of the designs of a novel nanoelec-tromechanical switch based on bilayer graphene nanoribbon," in IEEE Int. Electron Devices Meeting Tech. Dig. New York: IEEE, 2009, pp. 37.3.1 – 37.3.4.

[11] K.-T. Lam, C. Lee, and G. Liang, "Bilayer graphene nanoribbon nanoelectromechanical system device: A computational study," Applied Physics Letters, vol. 95, no. 14, p. 143107, 2009.

[12] S. K. Banerjee, L. F. Register, E. Tutuc, D. Reddy, and A. H. MacDonald, “Bilayer pseudospin field-effect transistor (bisfet): A proposed new logic device,” IEEE Elect. Dev. Lett., vol. 30, no. 2, pp. 158 – 160, 2009.

[13] Schwierz, Frank. "Graphene transistors." Nature Nanotechnology 5.7 (2010): 487-96.

[14] K. M. Masum Habib and Roger K. Lake, “Numerical Study of Electronic Transport Through Bilayer Graphene Nanoribbons,” Proc. of the 69th Annual Device Res. Conf. (DRC), pp. 109 - 110 (2011).

[15] K. M. M. Habib and R. K. Lake, “Graphene nanoribbon crossbar resonant tunneling diode,” (in preparation).

[16] Wei, S.-J.; Lin, H.C.; , "A multi-state memory using resonant tunneling diode pair," Circuits and Systems, 1991., IEEE International Sympoisum on , vol., no., pp.2924-2927 vol.5, 11-14 Jun 1991.

[17] van der Wagt, J.P.A.; Tang, H.; Broekaert, T.P.E.; Seabaugh, A.C.; Kao, Y.-C.; , "Multibit resonant tunneling diode SRAM cell based on slew-rate addressing," Electron Devices, IEEE Transactions on , vol.46, no.1, pp.55-62, Jan 1999.

[18] van der Wagt, J.P.A.; , "Tunneling-based SRAM," Proceedings of the IEEE , vol.87, no.4, pp.571-595, Apr 1999.

[19] Lin, H.C.; , "Resonant tunneling diodes for multi-valued digital applications," Multiple-Valued Logic, 1994. Proceedings., Twenty-Fourth International Symposium on , vol., no., pp.188-195, 25-27 May 1994.

[20] N. K. Jha, D. Chen Eds., “Nanoelectronic Circuit Design”, Springer, 2011.

[21] O. F. Sankey and D. J. Niklewski, “Ab initio multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems,” Phys. Rev. B, vol. 40, no. 6, pp. 3979 – 3995, 1989.

[22] J. P. Lewis, K. R. Glaesemann, G. A. Voth, J. Fritsch, A. A. Demkov, J. Ortega, and O. F. Sankey, “Further developments in the local-orbital density-functional-theory tight-binding method,” Phys. Rev. B, vol. 64, no. 19, p. 195103, 2001.

[23] Y.-W. Son, M. L. Cohen, and S. G. Louie, “Energy gaps in graphene nanoribbons,” Phys. Rev. Lett., vol. 97, no. 21, p. 216803, 2006.

[24] J. L. Martins, N. Troullier, and S. H. Wei, “Pseudopotential plane-wave calculations for ZnS,” Phys. Rev. B, vol. 43, no. 3, pp. 2213 – 2217, 1991.

[25] A. D. Becke, “Density-functional exchange-energy approximation with correct asymptotic behavior,” Phys. Rev. A, vol. 38, no. 6, pp. 3098 – 3100, 1988.

[26] C. Lee, W. Yang, and R. G. Parr, “Development of the colle-salvetti correlation energy formula into a functional of the electron density,” Phys. Rev. B, vol. 37, no. 2, pp. 785 – 789, 1988.

[27] J. Harris, “Simplified method for calculating the energy levels of weakly interacting fragments,” Phys. Rev. B, vol. 31, pp. 1770–1779, 1985.

[28] W. M. C. Foulkes and R. Haydock, “Tight-binding models and density-functional theory,” Phys. Rev. B, vol. 39, pp. 12 520–2536, 1989.

[29] A. A. Demkov, J. Ortega, O. F. Sankey, and M. P. Grumbach, “Electronic structure approach for complex silicas,” Phys. Rev. B, vol. 52, no. 3, pp. 1618 – 1630, 1995.

[30] P. Jelinek, H. Wang, J. P. Lewis, O. F. Sankey, and J. Ortega, “Multicenter approach to the exchange-correlation interactions in ab initio tight-binding methods,” Phys. Rev. B, vol. 71, no. 23, p. 235101, 2005.

[31] N. A. Bruque, M. K. Ashraf, T. R. Helander, G. J. O. Beran, and R. K. Lake, “Conductance of a conjugated molecule with carbon nanotube contacts,” Phys. Rev. B, vol. 80, no. 15, p. 155455, 2009.

[32] N. A. Bruque, R. R. Pandey, and R. K. Lake, “Electron transport through a conjugated molecule with carbon nanotube leads,” Phys. Rev. B, vol. 76, no. 20, p. 205322, 2007.

[33] Mazumder, P.; Kulkarni, S.; Bhattacharya, M.; Jian Ping Sun; Haddad, G.I.; "Digital circuit applications of resonant tunneling devices," Proceedings of the IEEE, vol.86, no.4, pp.664-686, Apr 1998.

[34] Khasanvis, S.; Habib, K.M.M.; Rahman, M.; Narayanan, P.; Lake, R.K.; Moritz, C.A.; , "Hybrid Graphene Nanoribbon-CMOS tunneling volatile memory fabric," Nanoscale Architectures (NANOARCH), 2011 IEEE/ACM International Symposium on , vol., no., pp.189-195, 8-9 June 2011.

[35] van der Wagt, J.P.A.; Seabaugh, A.C.; Beam, E.A., III.; , "RTD/HFET low standby power SRAM gain cell," Electron Device Letters, IEEE , vol.19, no.1, pp.7-9, Jan 1998.

[36] Azizi, N.; Najm, F.N.; Moshovos, A.; , "Low-leakage asymmetric-cell SRAM," Very Large Scale Integration (VLSI) Systems, IEEE Transactions on , vol.11, no.4, pp.701-715, Aug. 2003.

[37] Ling-Feng Mao; Li, X.J.; Zhu, C.Y.; Wang, Z.O.; Lu, Z.H.; Yang, J.F.; Zhu, H.W.; Liu, Y.S.; Wang, J.Y.; , "Finite-Size Effects on Thermionic Emission in Metal–Graphene-Nanoribbon Contacts," Electron Device Letters, IEEE , vol.31, no.5, pp.491-493, May 2010.

[38] Ximeng Guan; Qiushi Ran; Ming Zhang; Zhiping Yu; Wong, H.-S.P.; , "Modeling of schottky and ohmic contacts between metal and graphene nanoribbons using extended hückel theory (EHT)-based NEGF method," Electron Devices Meeting, 2008. IEDM 2008. IEEE International , vol., no., pp.1-4, 15-17 Dec. 2008.

[39] Unluer, D.; Tseng, F.; Ghosh, A.; Stan, M.; , "Monolithically patterned wide-narrow-wide all-graphene devices," Nanotechnology, IEEE Transactions on , vol.PP, no.99, pp.1, 0.

[40] Predictive Technology Model, http://ptm.asu.edu/.

[41] C. Bencher, H. Dai, and Y. Chen. “Gridded design rule scaling: Taking the CPU toward the 16nm node”, Proc. SPIE 7274, 2009.

[42] R. T Greenway, K. Jeong and A. B. Kahng, C.-H. Park and J. S. Petersen, “32nm 1-D regular pitch SRAM bitcell design for interference-assisted lithography”, Proc. SPIE BACUS, 2008.

[43] Rahman, M.; Narayanan, P.; Moritz, C.A.; , "N3asic-based nanowire volatile RAM," Nanotechnology (IEEE-NANO), 2011 11th IEEE Conference on , vol., no., pp.1097-1101, 15-18 Aug. 2011.

[44] K. Itoh, "Ultra-Low Voltage Nano-Scale Memories", Springer, 2007.

[45] Ki Chul Chun; Jain, P.; Jung Hwa Lee; Kim, C.H.; , "A 3T Gain Cell Embedded DRAM Utilizing Preferential Boosting for High Density and Low Power On-Die Caches," Solid-State Circuits, IEEE Journal of , vol.46, no.6, pp.1495-1505, June 2011.