Low-Power Heterogeneous Graphene Nanoribbon-CMOS Multistate Volatile Memory Circuit

SANTOSH KHASANVIS, University of Massachusetts Amherst
K. M. MASUM HABIB, University of California Riverside
MOSTAFIZUR RAHMAN, University of Massachusetts Amherst
ROGER LAKE, University of California Riverside
CSABA ANDRAS MORITZ, University of Massachusetts Amherst

Graphene is an emerging nanomaterial believed to be a potential candidate for post-Si nanoelectronics, due to its exotic properties. Recently, a new graphene nanoribbon crossbar (xGNR) device was proposed which exhibits negative differential resistance (NDR). In this paper, a multi-state memory design is presented that can store multiple bits in a single cell enabled by this xGNR device, called Graphene Nanoribbon Tunneling Random Access Memory (GNTRAM). An approach to increase the number of bits per cell is explored alternative to physical scaling to overcome CMOS SRAM limitations. A comprehensive design for quaternary GNTRAM is presented as a baseline, implemented with a heterogeneous integration between graphene and CMOS. Sources of leakage and approaches to mitigate them are investigated. This design is extensively benchmarked against 16nm CMOS SRAMs and 3T DRAM. The proposed quaternary cell shows up to 2.27x density benefit vs. 16nm CMOS SRAMs and 1.8x vs. 3T DRAM. It has comparable read performance and is power-efficient, up to 1.32x during active period and 818x during stand-by against high performance SRAMs. Multi-state GNTRAM has the potential to realize high-density low-power nanoscale embedded memories. Further improvements may be possible by using graphene more extensively, as graphene transistors become available in future.

Categories and Subject Descriptors: **B.7.1 [Integrated Circuits]:** Types and Design Styles—Advanced Technologies; **B.3.2 [Memory Structures]:** Design Styles— Primary memory

General Terms: Low-power design, Memory circuit, Benchmarking

Additional Key Words and Phrases: GNTRAM, graphene nanoribbons, hybrid integrated circuits, multistate memory, negative differential resistance

1. INTRODUCTION

SRAM has been the industry workhorse for on-chip embedded memory due to its high performance. In the past, on-chip caches have been steadily increasing in density to accommodate the growing demands for high-performance computing. In order to maintain this historical growth in memory density, SRAM bit cells have been aggressively scaled down physically 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 [Smith et al. 2012] due to several challenges such as increased leakage and variability at nanoscale [Itoh 2011; Qazi et al. 2011]. 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, as shown in Fig. 1. This provides a new dimension for scaling and can potentially overcome the challenges associated with physical downscaling at nanoscale. In addition, it may provide power benefits per-bit, since the power cost associated with each physical cell is amortized over multiple bits. Emerging nanoscale materials like graphene, and unique material interactions between novel device structures can

This work was supported in part by the Focus Center Research Program (FCRP) Center on Functionally Engineering Nano Architectonics (FENA), and the Center for Hierarchical Manufacturing (CHM) at UMass Amherst.

Author's addresses: S. Khasanvis, M. Rahman and C. A. Moritz are with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003 USA. (e-mail: {khasanvis, rahman, andras}@ecs.umass.edu). K. M. M. Habib and R. Lake are with University of California Riverside, CA 92521 USA. (e-mail: khabib@ee.ucr.edu, rlake@ee.ucr.edu).

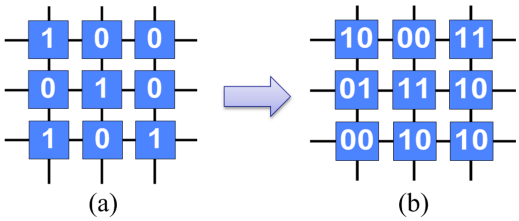


Fig. 1. (a) Current technology uses binary memory storing a single bit per cell; (b) Proposed concept: Multiple bits per cell with novel graphene structures.

enable the implementation of such unconventional circuits. They can potentially lead to low-power ultra-dense nanoscale memories, which cannot be achieved by relying on physical scaling alone.

Graphene is a two-dimensional layer of carbon atoms and is considered to be a potential candidate for post-Si nanoscale computing systems [ITRS]. It exhibits extra-ordinary electrical and thermal properties featuring Dirac fermion [Novoselov et al. 2005] with very high conductivity [Ando 2007] and extreme scalability. Its planar structure also potentially makes it compatible with current CMOS fabrication processes [de Heer 2007]. While graphene based transistors have been proposed [Fiori and Iannaccone 2009a; Fiori and Iannaccone 2009b; Lam and Liang 2009; Lam et al. 2009; Banerjee et al. 2009], challenges still exist which preclude their use in digital systems [Schwierz 2010]. Novel device structures with unique characteristics have been recently explored, such as the bi-layer graphene nanoribbon crossbar tunneling device (xGNR) [Habib and Lake 2011; Habib and Lake 2012; Habib et al. 2013] which exhibits negative differential resistance (NDR). This xGNR 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 [Wei and Lin 1991; van der Wagt 1999; Lin 1994]. However, RTDs were implemented using non-lithographic processes, which were expensive and incompatible with those for Si, which prohibited their integration with conventional technology [Jha and Chen 2001]. On the other hand, graphene based devices like xGNR potentially overcome such integration challenges and may be used in mainstream applications.

Our previous work has explored a binary memory circuit using this xGNR device [Khasanvis et al. 2011], which could also function as ternary memory [Khasanvis et al. 2012] but did not scale further. In this paper, we present a scaling approach that is different from physical scaling, where the number of bits stored in a single cell can be increased. We show a new quaternary memory circuit as baseline using the xGNR device, called quaternary graphene nanoribbon tunneling random access memory (GNTRAM). A heterogeneous graphene-CMOS circuit implementation is used for access and control. Extensive benchmarking against state-of-the-art 16nm CMOS SRAM and 3T DRAM memory cells is also presented. Our evaluations show that the quaternary GNTRAM has up to 2.27x density-per-bit benefit against CMOS SRAMs and 1.8x benefit against 3T DRAM in 16nm technology node. It is also up to 818x more power efficient per-bit when compared against the high-performance CMOS designs in idle periods, while having comparable performance. Even further improvements may be possible by using graphene more extensively instead of silicon MOSFETs, as advances are made in graphene technology.

The rest of the paper is organized as follows. Section 2 presents an overview of the xGNR device and latch configuration. The proposed scaling approach is presented in Section 3. Section 4 details the quaternary GNTRAM design and its operation.

Section 5 discusses the leakage analysis and mitigation in GNTRAM, followed by physical implementation description in Section 6. Methodology and benchmarking are presented in Section 7 and conclusion in Section 8.

2. BACKGROUND AND PREVIOUS WORK

2.1 Graphene Nanoribbon Crossbar (xGNR)

The graphene nanoribbon crossbar (Fig. 2a) is a two-terminal device. It consists of two semi-infinite, H-passivated armchair type GNRs (AGNRs) stacked orthogonally to each other with a vertical separation of 3.35 Å in between [Habib and Lake 2011; Habib and Lake 2012; Habib et al. 2013]. Each of these AGNRs has a truncated end with a zigzag edge. The overlap region of the xGNR is a misoriented or twisted bilayer graphene. Since we are interested in current switching in absence of a bandgap, 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. A voltage 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 equal to the applied bias.

The current-voltage (I-V) response of the xGNR is calculated using first principles atomistic calculations. The simulated I-V characteristics (Fig. 2b) exhibit negative differential resistance (NDR) with multiple peak and valley currents. This makes it suitable for RTD-based applications [Mazumder et al. 1998]. The oscillatory current-voltage response is attributed to the quantum interference between the standing electronic waves inside the twisted bilayer region of the xGNR as explained below.

An electron in a semi-infinite AGNR behaves as standing wave due to the reflection occurring at the truncated end. The wavelength of such a standing wave is a function of the total energy of the electron. Thus, by creating a potential energy difference between the top and bottom layers of xGNR, one can control the phase difference between the standing waves of individual nanoribbons. These standing waves interfere inside the overlap region of the xGNR. Depending on the phase difference (and hence the potential energy difference), the interference can be either constructive or destructive. Constructive interference occurs when the potential difference is $V = (2m + 1)\frac{\pi\hbar v}{2qL}$; where, $m \in \{0, 1, 2, \ldots\}$, v is the speed of the electron in graphene, L is the length of the truncated end of AGNR, $\hbar$ is the reduced Plank's constant and q is the charge of an electron. Similarly, the standing waves interfere destructively when $V = n\frac{\pi\hbar v}{qL}$ where, $n \in \{0, 1, 2, \ldots\}$ [Habib et al. 2013].

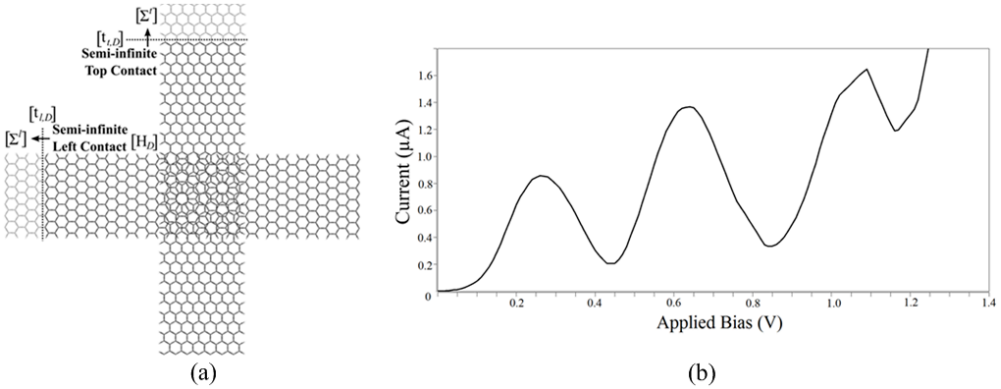


Fig. 2. (a) Atomistic geometry of the GNR crossbar device (xGNR); (b) Simulated I-V characteristics of the crossbar structure exhibiting NDR with multiple current peaks and valleys.

The interlayer tunneling current becomes maximum (or minimum) when the interference is constructive (or destructive). Thus, an external voltage bias applied across the layers of xGNR results in multiple constructive and destructive interferences, which leads to oscillatory current-voltage response with multiple NDR regions.

2.2 Application of xGNR Device in a Multistate Memory Element

A memory element can be built leveraging the NDR characteristics by using two xGNRs in a series configuration (Fig. 3a), similar to a Goto pair [Goto et al. 1960]. The circuit schematic of this configuration is shown in Fig. 3b. The xGNR latch consists of a pull-up leg and a pull-down leg. One of the devices (xGNR1) is connected to supply voltage (V_{dd}) and acts as the pull-up device. The other device (xGNR2) is connected to ground terminal acting as the pull-down device. The common terminal between these devices is the state-node (SN). Data is encoded in the voltage at this state node. DC load line analysis of this configuration exhibits three stable states A, B and C under applied voltage bias, as shown in Fig. 3c. Thus it can be used as a binary latch or ternary latch depending on choice of data representation (see Table I).

The latching mechanism is illustrated in Fig. 4 [Khasanvis et al. 2011]. The following terms will be used in the discussion:

I_{p1}, V_{p1} – First peak current and corresponding voltage
I_{p2}, V_{p2} – Second peak current and corresponding voltage
I_{v1}, V_{v1} – First valley current and corresponding voltage
I_{v2}, V_{v2} – Second valley current and corresponding voltage.

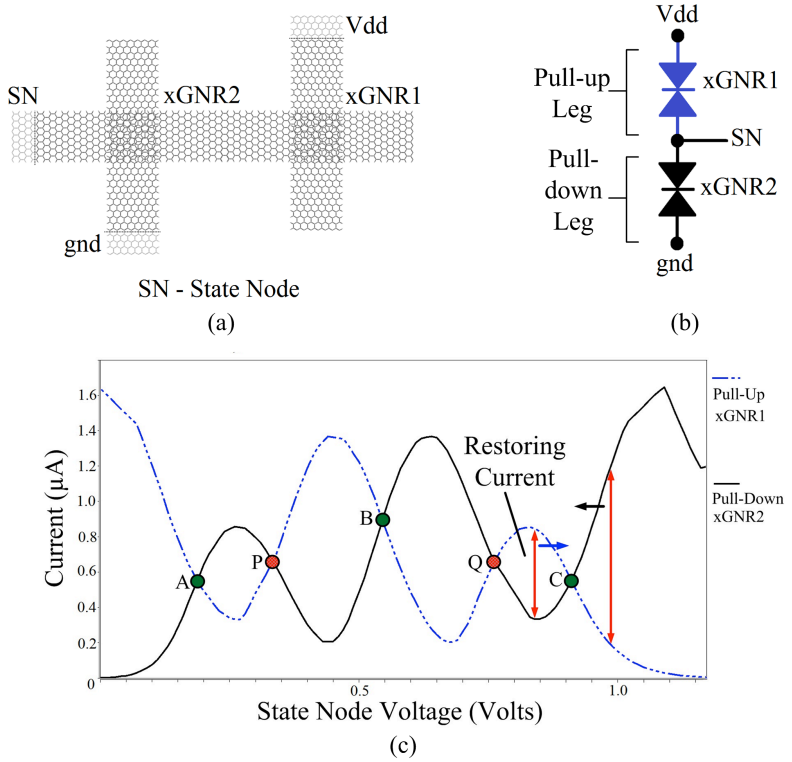


Fig. 3. (a) xGNR latch configuration; (b) Circuit schematic; and (c) DC load line analysis showing three stable states A, B, and C. States P and Q are unstable.

TABLE I. DATA REPRESENTATION USING XGNR LATCH

Data Representation	xGNR Latch State	Logic Value
Binary	State A	0
	State C	1
Ternary	State A	0
	State B	1
	State C	2

Fig. 4a-c illustrates the operation of latching logic HIGH (state C in Fig. 3c) onto the state node by injecting an input pull-up current into the latch (I_{in}). This could represent logic state 1 when used as binary memory, or logic state 2 when used as ternary memory. Y-axis represents currents and X-axis represents voltage at the state node (V_{SN}). The solid line shows pull-down current and dashed line represents pull-up current. Assuming the state node is initially at 0V, when the voltage V_{dd} is gradually increased, the operating point (shown by the dot X in Fig. 4a) is given by the intersection between pull-up and pull-down currents (satisfying Kirchoff's Current Law). Fig. 4a shows the situation when the first pull-down current peak is encountered, which is a decision point. As long as the pull-up current (I_{in} + I_{xGNR1}) is greater than pull-down current (I_{xGNR2}), the state node continues to shift from operating point X (Fig. 4a) to point Y (Fig. 4b). Finally it shifts to point C (Fig. 4c) when V_{dd} reaches its maximum value. When the input current (I_{in}) is switched off, the state node is latched to state C. Hence to be able to latch state C, the following condition should be met–

$$I_{in} + (I_{p1})_{xGNR1} > (I_{p2})_{xGNR2}. \quad (1)$$

Fig. 4d-f shows the process of latching logic LOW onto the state node (state A in Fig. 3c). This represents logic 0 in both binary and ternary representation. Consider the state-node is initially at 0V and the input is logic low. In this case, input pull-down current (I_{ex}) is applied at the state node. The analysis proceeds on the same

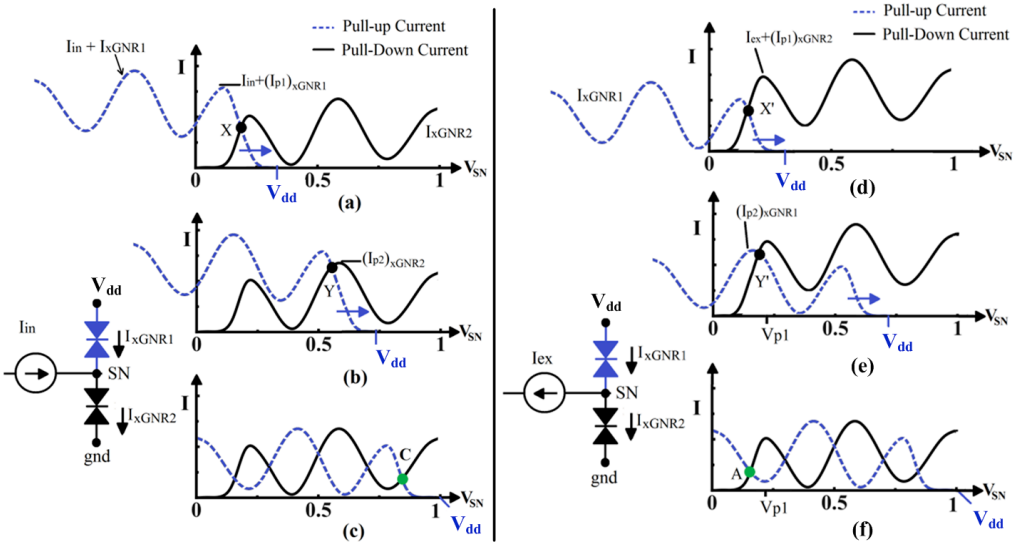


Fig. 4. Load Line Analysis of xGNR latch when latching data. (a)-(b) Input logic HIGH and V_{SN} at decision points, (c) Input switched OFF and logic HIGH latched. (d)-(e) Input logic LOW and V_{SN} at decision points, and (f) Input switched OFF and logic LOW latched.

lines as before. As long as the net pull-up current (I_{xGNR1}) is lower than pull-down currents ($I_{ex} + I_{xGNR2}$), the state node voltage (V_{SN}) will not rise beyond V_{p1} (Fig. 4e-f). After input I_{ex} is switched-off, the state node remains at state A. Thus, to be able to latch logic 0, the following condition has to be satisfied –

$$(I_{p2})_{xGNR1} < I_{ex} + (I_{p1})_{xGNR2}. \quad (2)$$

Similarly, when used as a ternary latch, the state node can be latched to the stable point B (in Fig. 3) if the following condition is satisfied –

$$(I_{p2})_{xGNR2} > I_{in} + (I_{p1})_{xGNR1} > (I_{p1})_{xGNR2}. \quad (3)$$

The retention of data in the latch is discussed next. As mentioned earlier, the states A, B and C (in Fig. 3c) are stable. When the state node is at one of these stable points, any external noise that causes the state voltage to increase or decrease would be countered by strong restoring currents (see Fig. 3c). For example, when the state node is at voltage corresponding to state C, a constant static current flows through the devices. Any external perturbation (noise) results in a noise current (I_{noise}) that may cause the state node voltage to decrease (or increase). This is countered by a net current that pulls-up (or pulls-down) the state node. The magnitude of the restoring current is given by the difference between the pull-up and pull-down currents ($|I_{xGNR1} - I_{xGNR2}|$). As long as the noise current is smaller than this restoring current, the operating point does not move beyond the decision points and data is retained.

$$|I_{noise}| < I_{p1} - I_{v2} \text{ (worst case)} \quad (4)$$

States denoted by P and Q (in Fig. 3c) are unstable and hence the corresponding voltages are transition voltages. Consider state Q; due to lack of a restoring current, external noise would cause the state node voltage to transition to one of the surrounding states depending on the direction of the perturbation. Thus for correct latch operation, the noise currents should be less than the restoring currents to ensure that states P and Q are not reached during latch retention.

Our previous work explored binary random access memory circuit using xGNR latch as the memory core, and access transistors for writing and reading data [Khasanvis et al. 2011]. This could also be used as a ternary memory cell [Khasanvis et al. 2012]. However, these circuits still required physical down-sizing of transistors to scale further. In the following section, we present an approach for scaling that is alternative to physical scaling, where the number of bits in a single cell can be further increased. This can potentially overcome the limitations of down-sizing CMOS transistors, providing an alternative pathway for scaling.

3. PROPOSED SCALING APPROACH

The key requirement for scaling is to increase the number of stable states of the xGNR latch, which would allow storing more bits in a single cell. This can be achieved by increasing the number of current peaks in the pull-up and pull-down legs of xGNR latch. When multiple xGNR devices are used in each leg, the I-V characteristics of such a configuration will exhibit more current peaks than if a single device is used in each leg [Kao et al. 1992].

As shown in Fig. 5a-b, a series combination of 2 xGNRs leads to 4 current peaks. Similarly, 3 xGNRs in series lead to 6 current peaks (Fig. 5c-d). In general, a series configuration of 'N' xGNR devices exhibits '2N' current peaks, since each xGNR

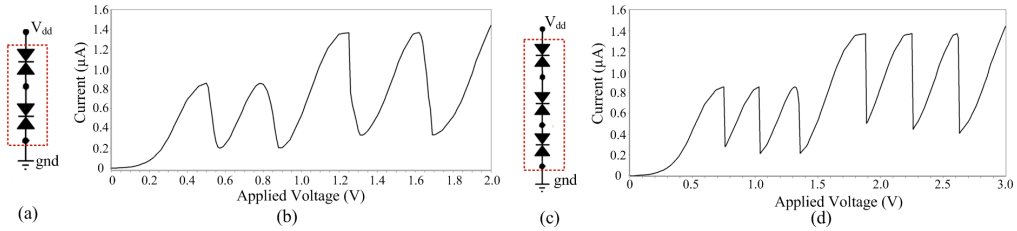


Fig. 5. Circuit technique to increase number of current peaks: (a) 2 xGNRs in series; (b) DC load line analysis showing 4 current peaks for configuration in (a); (c) 3 xGNRs in series; and (d) DC load line analysis showing 6 current peaks for configuration in (c).

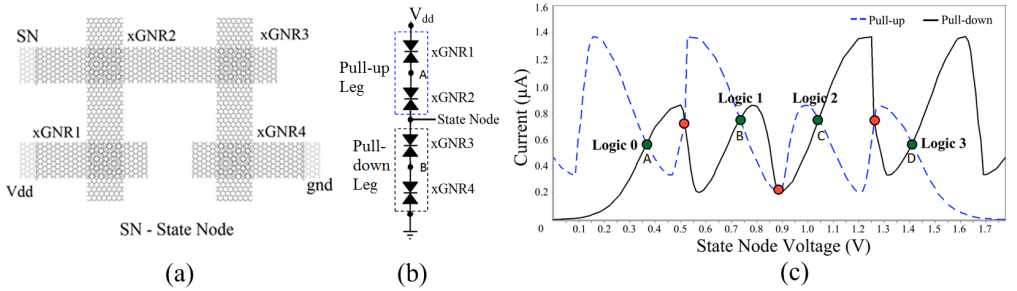


Fig. 6. (a) Quaternary xGNR tunneling latch structure; (b) Circuit schematic; and (c) DC load line analysis showing 4 stable states.

device has 2 current peaks. However, every additional xGNR in the stack would require a higher operating voltage in order to reach all the current peaks. Thus, the operating voltage limitation determines the maximum number of current peaks (and hence the number of stable states) that can be achieved with such a multi-peak xGNR circuit.

For the xGNR latch shown in the preceding section, each device in pull-up and pull-down legs exhibited 2 current peaks in their I-V characteristics, which led to 3 stable states. In general, a latch configuration with devices having 'P' current peaks in each leg would exhibit 'P + 1' stable states. Thus, a configuration of 2 series xGNRs in each leg of the xGNR latch (Fig. 6a-b) would lead to 5 stable states at the state node, since both pull-up and pull-down legs have 4 current peaks. We use 4 of these states (as shown in Fig. 6c) to build a quaternary memory cell, as discussed next.

4. QUATERNARY GRAPHENE NANORIBBON TUNNELING RANDOM ACCESS MEMORY

An xGNR latch configuration with two series xGNR devices in each leg can realize a quaternary latch, and this is used to build quaternary graphene nanoribbon tunneling random access memory (GNTRAM). Such a design will enable storing 2 bits in a single memory cell, resulting in a higher memory density than CMOS SRAMs that store 1 bit per cell.

A dynamic memory cell implementation is adopted for low-leakage, low-power quaternary GNTRAM, as shown in Fig. 7a. This design uses the quaternary xGNR latch as the state holding element, thus exhibiting 4 stable states (see Fig. 7b). Access to the state node is achieved with *write*-FET and *read*-FETs. To mitigate static power dissipation, the xGNR latch is switched OFF during idle periods using a *sleep*-FET and a Schottky diode. A capacitor (C_{SN}) is then required at the state node to retain the state written into the cell. The Schottky diode provides current rectification, mitigating charge leakage through reverse current paths when voltage is switched OFF during idle periods. This implementation is a multi-threshold circuit design, where transistors with high threshold voltage (V_t) are used in leakage critical

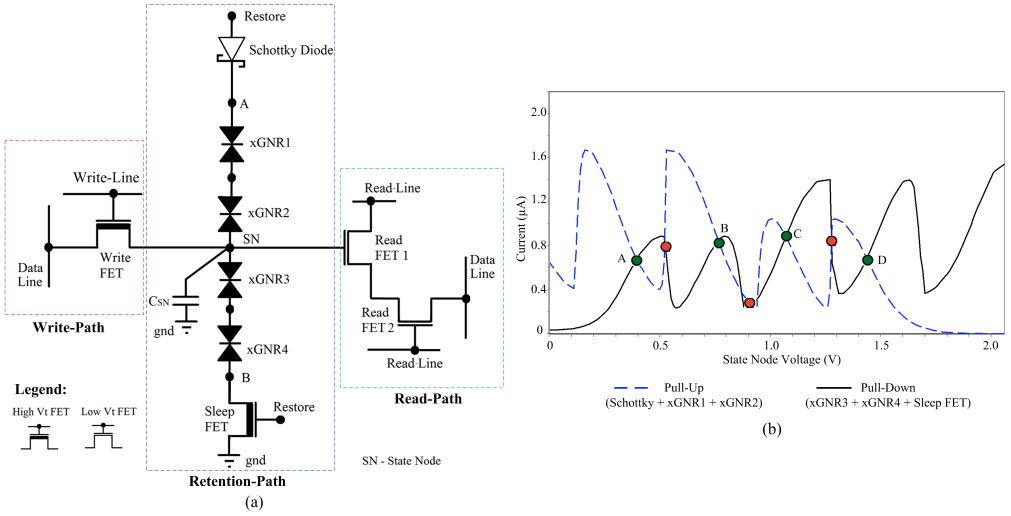


Fig. 7. (a) Proposed quaternary GNTRAM cell; and (b) DC Load Line Analysis for xGNR memory circuit.

paths, and low-V_t transistors are used in other paths. The operation of quaternary GNTRAM is described next.

4.1 Write Operation

The write operation involves charging-up/discharging the state capacitance to the required voltage through the *write*-FET. The gate terminal of the *write*-FET is connected to the *write-line* and the drain terminal is connected to the input *data-line*, with the source terminal acting as the state-node. During a write operation, the memory cell is selected by activating the corresponding *write-line*, and then the *restore* signal is switched ON. Data is written by applying the required input voltage onto the *data-line*, which either charges or discharges the state capacitance depending on the previous state. Here, the input voltages used are in quaternary representation (0V – logic 0, 0.75V – logic 1, 1.1V – logic 2 and 1.5V – logic 3). These voltage values are chosen based on voltages at which stable states A, B, C and D occur in the xGNR latch characteristics respectively (see Fig. 7b).

After the data has been written, the input and write signals are switched OFF while *restore* signal is still ON. This results in restoring currents through xGNR latch that prevent FET switching noise transients from affecting the state-node. Once the *write*-FET is completely switched OFF, the *restore* signal can be de-asserted and the data is held on the state capacitor. Fig. 8a shows the simulated write operation (using HSPICE) for all possible state transitions in the quaternary GNTRAM cell.

4.2 Read Operation

A pre-discharge and evaluate scheme is used to read the stored information in the memory cell. The series stack of read-FETs acts as the evaluation path during read operation (see Fig. 7a). The output *data-line* is connected to the source of *read*-FET2. 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.

To initiate the read operation, the *data-line* is discharged first to 0V and then the *read-line* signal is switched ON (see Fig. 8b-d). This starts to pull up the voltage on the output *data-line*. The voltage to which the output can be pulled-up is limited by

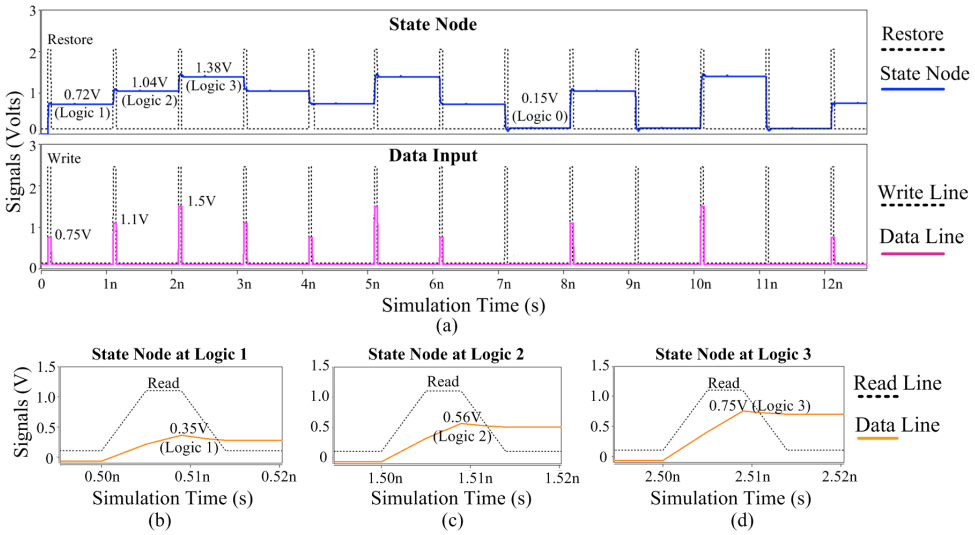


Fig. 8. Quaternary GNTRAM circuit operation simulated with HSPICE: (a) Write operation; and Read operation when storing (b) logic 1; (c) logic 2; and (d) logic 3 at state node.

the state-node voltage at the gate of *read*-FET1. This is due to the intrinsic threshold voltage drop in the nMOS transistor. Thus the final output voltage at the end of read operation is specific to the stored state, which enables the detection of multiple voltage levels at the data output. When logic 0 is stored, *read*-FET1 is completely switched OFF and the data line remains at low voltage. For all other stored logic states, *read*-FET1 is switched ON and the output is pulled up to the corresponding voltage level (see Fig. 8b-d). To successfully distinguish between different stored states low-V_t transistors are used in the read path.

4.3 Restore Operation

During idle periods, the xGNR latch is switched OFF by turning OFF the *restore* signal. Data is then stored on the state capacitance. However, the stored charge on the state capacitance starts to leak and needs to be replenished. This is done by simply switching ON the *restore* signal periodically within a stipulated time interval. For example, consider logic 3 (state D in Fig. 7b) being stored in the memory. During idle periods, the stored voltage gradually reduces due to leakage. When the *restore* signal is switched ON, a net pull-up current in the xGNR latch charges the capacitor back to logic 3. As long as the voltage has not dropped below the transistion state between logic 2 and logic 3 (see Fig. 7b), it can be restored.

The time for which a written state can be maintained before it has to be restored is called retention time, and it is desirable to maximize the retention time. Two factors contribute to this: (i) total capacitance at the state node, and (ii) total leakage current.

The value of the state node capacitance (C_{SN}) is determined by (i) the total 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 after the write operation (when the *write* and *restore* signals are deactivated), and the cell goes into idle mode. This degrades the written voltage value (V_W) to a quiescent voltage level (V_Q). For example, consider the case when logic 3 was written into the memory cell, shown in Fig. 9. If V_Q falls below transition voltage (V_{tran} in Fig. 9) after the write operation, the subsequent restore operation will cause a state transition to logic 2 instead of restoring logic 3 at the state node. Thus the value of V_Q should be high enough to ensure that the state

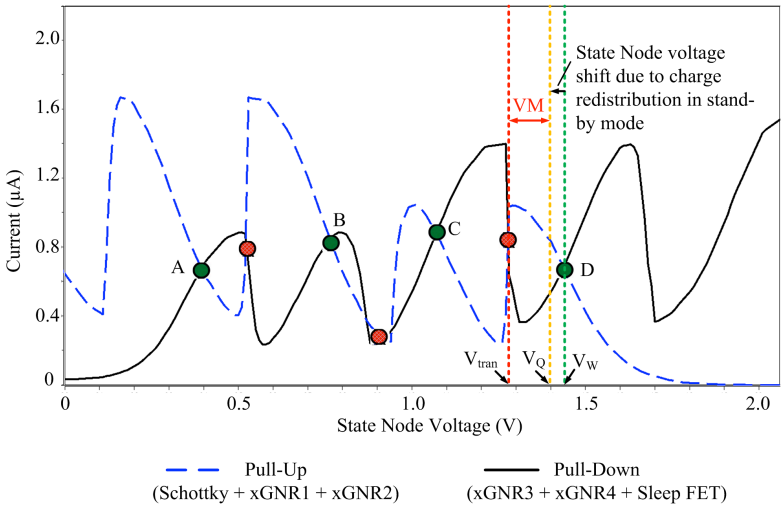


Fig. 9. DC Load Line Analysis for xGNR latch including showing a shift in state node voltage after logic 3 is written, and the available voltage margin during idle mode.

information is not lost immediately after write operation. In addition, the quiescent voltage level (V_Q) should also ensure that sufficient voltage-margin (VM in Fig. 9) is maintained for dynamic data retention. By choosing an appropriate V_Q, the retention time can be optimized. Based on these requirements, the minimum value of the total capacitance at the state node for a particular V_Q can be derived using the following relation:

$$C_{SN}.V_w = (C_{SN} + C_{PT}).V_Q \quad (5)$$

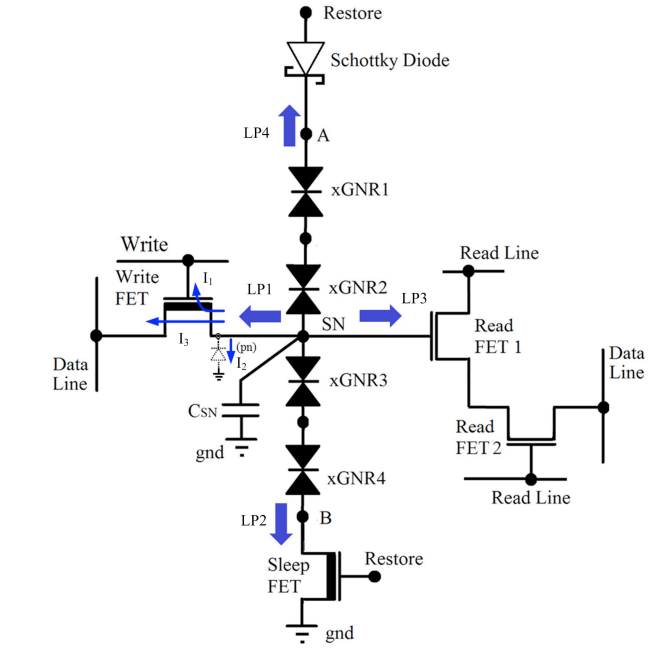
In (5), C_{SN} is the total capacitance at the state node. This includes the explicit capacitance to be formed at the state node, diffusion capacitance of the *write*-FET, gate capacitance of *read*-FET1, and the capacitance due to routing lines. C_{PT} is the total parasitic capacitance, which includes 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 write operation. The available voltage margin for retention is given by the difference between V_Q and V_{tran}.

A higher state capacitance leads to a higher voltage margin, and thus lengthens the retention time. However a large state capacitance is not desirable as it slows down the write operation. The other option is to reduce the magnitude of leakage currents. To minimize charge leakage, the critical leakage paths need to be identified. In this design, the *write*-FET and *sleep*-FET form leakage-critical paths since the transistors are directly connected to the state node. Hence they are implemented with high-threshold voltage (V_t) transistors, which are typically optimized to have very low OFF-state current and minimize leakage during stand-by.

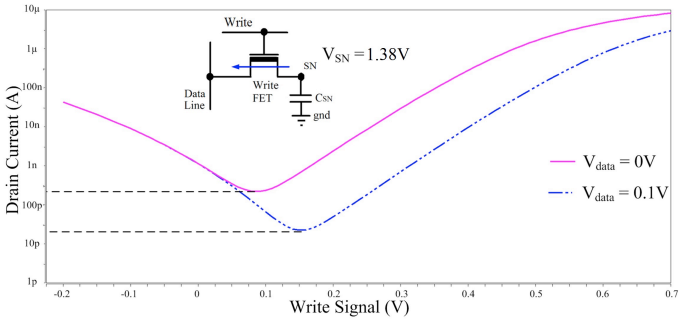
However, even with the use of high-V_t transistors the retention time for the quaternary GNTRAM was found to be low (in the order of a few nano-seconds). This is due to exacerbated leakage at the relatively higher operating voltage when storing logic 3. This necessitates leakage mitigation techniques to improve the retention time, which is discussed next.

5. LEAKAGE ANALYSIS AND MITIGATION

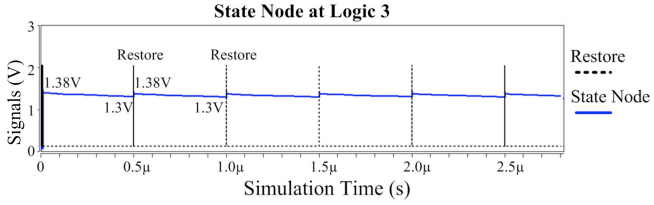
Leakage current in MOS transistors is exacerbated at high voltages. Hence during idle periods, leakage currents are the highest when the memory cell stores logic state 3 (1.38V). Analysis of the leakage paths (denoted by LP1 through LP4 in Fig. 10a) shows that the *write*-FET and *sleep*-FET form critical paths (LP1 and LP2), since they are connected through low impedance paths to the state node. For both devices, the sources of leakage are – gate tunneling current (I_1), reverse-bias junction leakage (I_2) and sub-threshold channel leakage (I_3). It was found that for the 16nm LP PTM devices used, leakage current was dominated by sub-threshold channel leakage (I_3).



(a)



(b)



(c)

Fig. 10. (a) Leakage paths in quaternary GNTRAM; (b) Sub-threshold leakage analysis in write FET when logic 3 is stored at SN; and (c) Simulated restore operation when logic 3 is stored at state node after mitigating leakage.

One of the frequently used circuit techniques in literature to reduce the OFF-state sub-threshold channel leakage is source/gate biasing [Itoh 2007]. This scheme is most effective compared to other techniques such as body-biasing or V_{DS} reduction [Itoh 2007]. The sub-threshold analysis of the devices (see Fig. 10b) shows that when the source is offset by 0.1V during idle periods, the leakage current can be reduced by almost 10x when storing logic state 3. Thus the *data-line* and the source terminal of the *sleep*-FET are maintained at 0.1V during idle mode. This can be achieved either by using a self-biasing scheme with a shared carefully-sized nMOS transistor in series [Itoh 2007] or by selecting a separate voltage source [Elakkumanan et al. 2003].

The remaining leakage sources are the gate-oxide tunneling current through *read*-FET1 (LP3 in Fig. 10a) and the reverse-bias leakage of the Schottky diode (LP4 in Fig. 10a). The gate-oxide leakage can be reduced by increasing the oxide thickness (for 16nm PTM transistor used here, V_{th0} was recalculated using the equation for retro-grade doping CMOS [Morshed et al. 2011]). Thus the *read*-FET1 will need to be engineered to minimize the gate-oxide tunneling current, while still maintaining low-enough V_t to be able to read the stored states. The reverse-bias leakage through the Schottky diode is assumed to be constant at 10pA. These techniques enhanced data retention period to 0.5µs as shown in Fig. 10c.

6. PHYSICAL IMPLEMENTATION

A cross-technology heterogeneous implementation is used between CMOS and graphene [Khasanvis et al. 2011], as shown in Fig. 11. A lithography-friendly grid-based layout is used with minimum sized nMOS transistors for high density and ease of fabrication (Fig. 11a). The MOS transistors are created first on the Si 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 [Mao et al. 2010; Guan et al. 2008]. 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. 11b. The rest of the graphene-metal contacts are Ohmic to ensure proper operation and this is achieved by using wide GNRs [Unluer et al. 2011]. 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

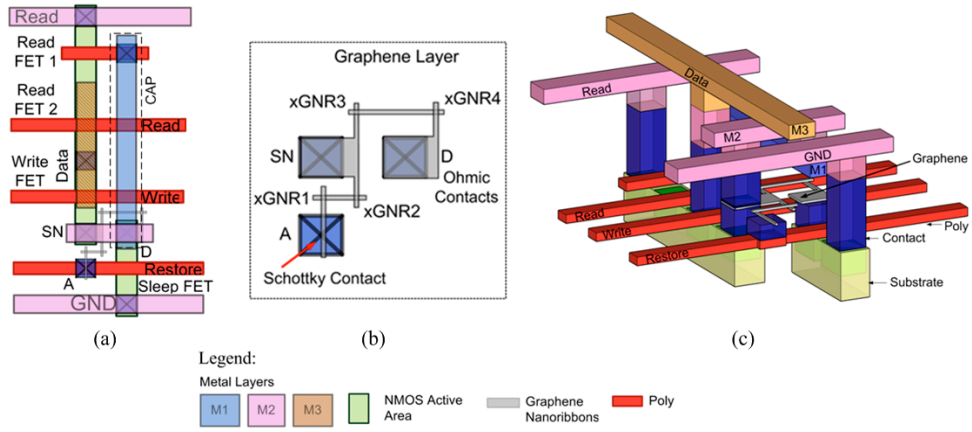


Fig. 11. (a) Quaternary GNTRAM physical layout; (b) Graphene layer showing xGNR devices, Schottky and Ohmic contacts; and (c) Heterogeneous integration with CMOS routing stack.

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. Interconnections are implemented with conventional CMOS routing layers (Fig. 11c). The state capacitor can be implemented either as a trench or as a stacked capacitor over the state node routing area shown in Fig. 11a.

7. METHODOLOGY AND BENCHMARKING

HSPICE circuit simulator was used to verify the GNTRAM operation and for power and performance analysis. The xGNR device was modeled as a HSPICE sub-circuit [van der Wagt 1999] using the structure shown in Fig. 12. The DC I-V characteristics derived from the atomistic simulations (mentioned in Section 2.1) was modeled using a voltage controlled current source (VCCS) with a piece-wise linear approximation between each I-V data point. The geometric capacitance at the GNR crossbar was modeled as a capacitor in parallel to take reactive currents into account in addition to DC response. A generic integrated circuit Schottky diode model was used for a first order analysis and 16nm CMOS PTM models [Predictive Technology Models] were used to simulate the *read*, *write* and *sleep*-FETs. The value of the state capacitance was chosen to be 200aF for required circuit behavior. PTM interconnect RC models based on scaled interconnect dimensions were used in conjunction with the PTM transistor models for power and performance evaluation of GNTRAM using HSPICE. For physical layout design and area evaluation of GNTRAM, 1-D gridded design rules [Bencher et al. 2009] were used as shown in Table II.

For benchmarking against CMOS, 16nm Gridded 8T SRAM cell [Greenway et al. 2008] was used, since this SRAM design utilizes the same grid-based design used in GNTRAM. 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 both high-performance and low-power 6T-SRAM designs at 65nm, 45nm and 32nm technology nodes. This method is detailed in [Rahman et al. 2011]. Using this data, scaling factors were derived based on cell area, Poly, Metal1, Metal2 and Via scaling trends. These were used to calculate 16nm 6T-SRAM design rules and cell area. The aforementioned 16nm Predictive Technology Models (PTM) transistors and RC interconnect models were used for power and performance evaluation of CMOS 6T SRAM and Gridded 8T SRAM using HSPICE. Both low-power and high-performance 6T and 8T SRAM cell designs were considered for comparison since quaternary GNTRAM uses a multi-V_t cell design.

3T DRAM was also investigated for benchmarking since it is a potential candidate for on-chip caches in advanced technology nodes [Itoh 2007; Chun et al. 2011]. The 3T DRAM cell (shown in Fig. 13) was designed using 16nm PTM transistor models, and the physical layout was done on the same lines as the GNTRAM. It was

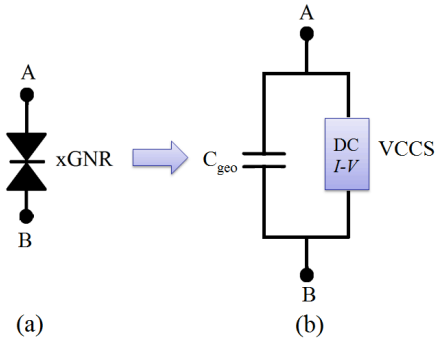


Fig. 12. (a) xGNR device circuit schematic; (b) HSPICE xGNR device model.

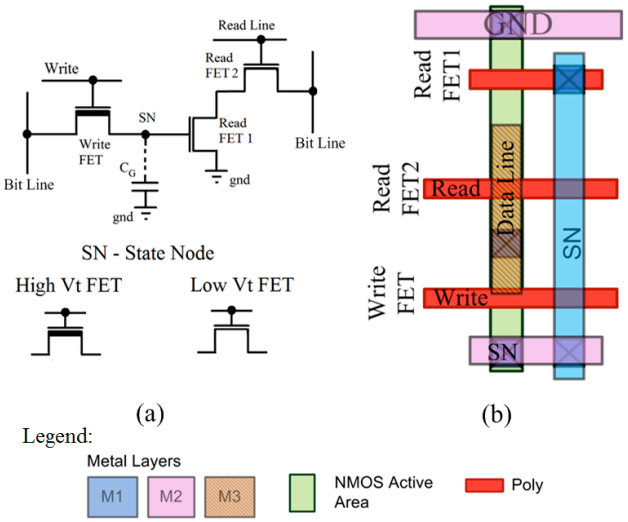


Fig. 13. 3T DRAM: (a) Circuit Schematic; and (b) Physical Layout.

TABLE II. GRIDDED DESIGN RULES

1D Gridded Design [Bencher et al. 2009]	M1, M2 Interconnect	Poly
Pitch (16nm node)	40~60 nm	60~80nm

simulated using HSPICE for power and performance evaluation. Area evaluation was done using grid-based design rules (see Table II). Table III shows the comparison results.

7.1 Area Evaluation

The GNTRAM physical cell area was estimated for the layout in Fig. 11a based on the design rules shown in Table II. Since this is a grid-based design, the area was calculated by counting the nuber of metal and poly pitches along each dimension. This area accounts for spacing required between adjacent GNTRAM cells as well.

Quaternary GNTRAM showed significant density advantage compared to the other 16nm CMOS RAMs. Although the physical cell area is comparable to that of the SRAMs and the 3T DRAM, quaternary GNTRAM's density benefit comes from the fact that it stores 2 bits per cell. In particular, GNTRAM showed a density-per-bit benefit of up to 2.27x vs. CMOS SRAM and 1.8x vs. the 3T DRAM in 16nm technology node.

Considering the current SRAM scaling trend, CMOS SRAM when advanced by two technology generations after 16nm node would have about the same area as 16nm quaternary GNTRAM. Thus GNTRAM provides 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.

7.2 Power Evaluation

For power evaluation, GNTRAM power dissipation was measured using HSPICE simulations during both active (read/write) and idle periods. For active power, the power dissipation was measured for all possible state transistions during write

TABLE III. QUATERNARY GNTRAM BENCHMARKING

		Quaternary GNTRAM (Per Cell, 2 bits)	Quaternary GNTRAM (Per Bit)	16nm CMOS 6T SRAM Cell (High Performance)	16nm CMOS Gridded 8T SRAM Cell (High Performance)	16nm 3-T DRAM Cell
Area Comparison (μm²)		0.03 – 0.06	0.015 – 0.03	0.026 – 0.064	0.0336 – 0.0672	0.0264 – 0.054
Power Comparison	**Active Power (μW)**	3.6 – 4.1	1.8 – 2.05	2.1 – 2.2	2.38 – 2.44	2.12 – 2.15
	Stand-by Power (pW)	38 – 44	19 – 22	6152 – 6157	15552 – 15556	6.49 – 7.01
Performance	**Read Operation (ps)**	7.6 – 8.2		8.35 – 9.25	7.68 – 7.96	9.18 – 9.68
	Write Operation (ps)	31.6 – 32		18.44 – 18.46	16.62 – 19.16	10.45 – 10.97

		Quaternary GNTRAM (Per Cell, 2 bits)	Quaternary GNTRAM (Per Bit)	16nm CMOS 6T SRAM Cell (Low Power)	16nm CMOS Gridded 8T SRAM Cell (Low Power)
Area Comparison (μm²)		0.03 – 0.06	0.015 – 0.03	0.026 – 0.064	0.0336 – 0.0672
Power Comparison	**Active Power (μW)**	3.6 – 4.1	1.8 – 2.05	1.21 – 1.16	1.45 – 1.47
	Stand-by Power (pW)	38 – 44	19 – 22	124.18 – 125.12	78.38 – 78.44
Performance	**Read Operation (ps)**	7.6 – 8.2		17.39 – 21.03	14.82 – 16.08
	Write Operation (ps)	31.6 – 32		67.27 – 67.54	58.37 – 63.18

operation, as well as during the read operation. The worst-case power was then considered and is reported here. The same method was followed for evaluation of CMOS SRAM and 3T-DRAM cells. Quaternary GNTRAM showed up to 1.32x lower active power per bit against CMOS high-power SRAM designs. It also showed up to 1.17x lower active power-per-bit against the 3T DRAM in 16nm node.

Stand-by power dissipation was measured with HSPICE simulations during idle periods (no switching activity), when GNTRAM is storing data on the state capacitance. Quaternary GNTRAM was 818x more power-efficient during idle period against the high-performance CMOS SRAM, and 6.53x more power-efficient against low-power CMOS 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 2 bits per cell thus amortizing leakage costs. The 3T DRAM exhibited lower stand-by power than GNTRAM since it has lesser number of leakage paths.

7.3 Performance Evaluation

GNTRAM performance was evaluated by measuring the time taken to write data onto the state node using HSPICE simulations. All state transitions during write operation were measured and the worst-case write time is reported here. Similarly, time taken to read various stored states was measured and worst-case read time was

considered. CMOS SRAM and 3T-DRAM performance measurements were performed using the same method.

Quaternary GNTRAM was comparable in read performance to high-performance CMOS SRAMs, even though it uses a higher capacitance at state node. This is because GNTRAM uses low-V_t transistors in its read path, which are typically optimized to have high ON current. The asymmetric cell design (multi-V_t transistors) thus enables high-performance, while reaping the benefits due to low power operation. An asymmetric approach was necessary in GNTRAM because the *read*-FETs need to have a low-V_t to successfully differentiate between the stored states. The write performance of GNTRAM was slower because of the increased voltage swing associated with storing logic 3, which requires a longer time to charge the state capacitance. The 3T DRAM performed better than GNTRAM during write operation because the state node capacitance is lower in 3T DRAM (which is the just the gate capacitance of *read*-FET).

8. CONCLUSION

A low-power multi-state memory concept was introduced in this paper enabled by unique graphene nanoribbon crossbar devices (xGNRs). This presented a new direction for scaling where the number of bits stored in a single cell can be increased, as an alternative to physical down-sizing of transistors. This may potentially overcome the challenges associated with transistor scaling. Quaternary graphene nanoribbon crossbar tunneling random access memory (GNTRAM) cell was presented as a baseline, and implemented with a heterogeneous integration between CMOS and graphene. Benchmarking against state-of-the-art 16nm CMOS RAM designs showed that quaternary GNTRAM exhibited significant benefits, which stem from storing 2 bits per cell.

This work takes the initial step towards exploring the potential of multi-state memories for on-chip memory applications enabled by graphene. While operating voltage may limit the maximum number of bits that can be stored, the xGNR device itself can possibly be engineered to have more current peaks within a smaller operating voltage. As progress is made in graphene technology, further benefits may be expected by replacing Si MOSFETs with graphene transistors, thus resulting in ultra-dense nanoscale memories.

REFERENCES

K. C. Smith, A. Wang, and L. C. Fujino. 2012. Through the looking glass: Trend tracking for ISSCC 2012. *IEEE Solid-State Circuits Magazine*, vol.4, no.1, pp.4-20.

K. Itoh. 2011. Embedded memories: Progress and a look into the future. *IEEE Design & Test of Computers*, vol.28, no.1, pp.10-13.

M. Qazi, M. E. Sinangil, and A. P. Chandrakasan. 2011. Challenges and directions for low-voltage SRAM. *IEEE Design & Test of Computers*, vol.28, no.1, pp.32-43.

ITRS – *The International Technology Roadmap for Semiconductors.* http://www.itrs.net/.

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

T. Ando. 2007. Exotic electronic and transport properties of grapheme. *Physica E: Low-dimensional Systems and Nanostructures*, vol. 40, no. 2, pp. 213 – 227.

W. A. de Heer, C. Berger, E. Conrad, P. First, R. Murali, and J. Meindl. 2007. Pionics: The emerging science and technology of graphene-based nanoelectronics. In *Proceedings of IEEE International Electron Devices Meeting, IEDM 2007*, pp.199-202.

G. Fiori and G. Iannaccone. 2009. On the possibility of tunable-gap bilayer graphene FET. *IEEE Electron Device Letters*, vol. 30, no. 3, pp. 261–264.

G. Fiori, and G. Iannaccone. 2009. Ultralow-voltage bilayer graphene tunnel FET. *IEEE Electron Device Letters*, vol. 30, no. 10, pp. 1096–1098.

K.-T. Lam and G. Liang. 2009. A computational evaluation of the designs of a novel nanoelectromechanical switch based on bilayer graphene nanoribbon. In *IEEE Int. Electron Devices Meeting Tech. Dig*., New York: IEEE, pp. 37.3.1 – 37.3.4.

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

S. K. Banerjee, L. F. Register, E. Tutuc, D. Reddy, and A. H. MacDonald. 2009. Bilayer pseudospin field-effect transistor (bisfet): A proposed new logic device. *IEEE Electron Device Letters*, vol. 30, no. 2, pp. 158 – 160.

F. Schwierz. 2010. Graphene transistors. *Nature Nanotechnology*, 5.7, pages 487-96.

K. M. M. Habib and R. K. Lake. 2011. Numerical study of electronic transport through bilayer graphene nanoribbons. In the *Proceedings of the 69th Annual Device Res. Conf. (DRC)*, pp. 109 - 110.

K. M. M. Habib and R. K. Lake. 2012. Current modulation by voltage control of the quantum phase in crossed graphene nanoribbons. *Phys. Rev. B*, 86(4), 045418.

K. M. M. Habib, F. Zahid and R. K. Lake. 2013. Multi-state current switching by voltage controlled coupling of crossed graphene nanoribbons. *J. Appl. Phys.*, vol. 114(15), 153710.

S.-J. Wei and H. C. Lin. 1991. A multi-state memory using resonant tunneling diode pair. In the *Proceedings of IEEE International Symposium on Circuits and Systems*, pp.2924-2927 vol.5.

J. P. A. van der Wagt. 1999. Tunneling-based SRAM. *Proceedings of the IEEE*, vol.87, no.4, pp.571-595.

H. C. Lin. 1994. Resonant tunneling diodes for multi-valued digital applications. In the *Proceedings of Twenty-Fourth International Symposium on Multiple-Valued Logic*, pp.188-195.

N. K. Jha and D. Chen (Eds.). 2011. *Nanoelectronic Circuit Design*. Springer.

S. Khasanvis, K. M. M. Habib, M. Rahman, P. Narayanan, R. K. Lake, and C. A. Moritz. 2011. Hybrid graphene nanoribbon-CMOS tunneling volatile memory fabric. In the *Proceedings of IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp.189-195.

S. Khasanvis, K. M. M. Habib, M. Rahman, P. Narayanan, R. K. Lake, and C. A. Moritz. 2012. Ternary volatile random access memory based on heterogeneous graphene-CMOS fabric. In the *Proceedings of IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp.69-76.

P. Mazumder, S. Kulkarni, M. Bhattacharya, J. P. Sun, and G. I. Haddad. 1998. Digital circuit applications of resonant tunneling devices. *Proceedings of the IEEE*, vol.86, no.4, pp.664-686.

E. Goto, K. Mutara, K. Nakazawa, T. Moto-Oka, Y. Matsuoka, Y. Ishibashi, T. Soma, and E. Wada. 1960. Esaki diode high-speed logical circuits. *IRE Trans. Electron. Comput.*, vol. 9, pp. 25–29.

Y.C. Kao, A.C. Seabaugh, and H.-T. Yuan. 1992. Vertical integration of structured resonant tunneling diodes on InP for multi-valued memory applications. In the *Proceedings of 4th International Conference On Indium Phosphide and Related Materials*, pp.489-492.

K. Itoh. 2007. *Ultra-Low Voltage Nano-Scale Memories*. Springer.

P. Elakkumanan, A. Narasimhan, and R. Sridhar. 2003. NC-SRAM - A low-leakage memory circuit for ultra deep submicron designs. In the *Proceedings of IEEE International Systems-on-Chip (SOC) Conference*, pp. 3- 6.

T. H. Morshed, D. D. Lu, W. Yang, M. V. Dunga, X. Xi, J. He, W. Liu, Kanyu, M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, and C. Hu. 2011. BSIM4v4.7 MOSFET Model – User's Manual. UC Berkeley. Retrieved May 12, 2014 from http://www-device.eecs.berkeley.edu/bsim/Files/BSIM4/BSIM470/BSIM470_Manual.pdf

L.-F. Mao, X. J. Li, C. Y. Zhu, Z. O. Wang, Z. H. Lu, J. F. Yang, H. W. Zhu, Y. S. Liu, and J. Y. Wang. 2010. Finite-size effects on thermionic emission in metal–graphene-nanoribbon contacts. *IEEE Electron Device Letters*, vol.31, no.5, pp.491-493 (May 2010).

X. Guan, Q. Ran, M. Zhang, Z. Yu, and H.-S. P. Wong. 2008. Modeling of Schottky and Ohmic contacts between metal and graphene nanoribbons using extended hückel theory (EHT)-based NEGF method. In the *Proceedings of IEEE International Electron Devices Meeting (IEDM 2008)*, pp.1-4.

D. Unluer, F. Tseng, A. W. Ghosh, and M. R. Stan. 2011. Monolithically patterned wide–narrow–wide all-graphene devices. *IEEE Transactions on Nanotechnology*, vol.10, no.5, pp.931-939 (Sept. 2011).

Predictive Technology Model, http://ptm.asu.edu/.

C. Bencher, H. Dai, and Y. Chen. 2009. Gridded design rule scaling: Taking the CPU toward the 16nm node. *Proc. SPIE 7274, Optical Microlithography XXII*, 72740G (March 16, 2009). DOI:10.1117/12.814435

R. T Greenway, K. Jeong and A. B. Kahng, C.-H. Park and J. S. Petersen. 2008. 32nm 1-D regular pitch SRAM bitcell design for interference-assisted lithography. *Proc. SPIE 7122, Photomask Technology 2008*, 71221L (17 October 2008). DOI: 10.1117/12.801883

M. Rahman, P. Narayanan, and C. A. Moritz. 2011. N^{3}asic-based nanowire volatile RAM. In the *Proceedings of 11th IEEE Conference on Nanotechnology (IEEE-NANO 2011)*, pp.1097-1101 (15-18 Aug. 2011).

K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim. 2011. A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches. *IEEE Journal of Solid-State Circuits*, vol.46, no.6, pp.1495-1505 (June 2011).