Heterogeneous Graphene-CMOS Ternary Content Addressable Memory

Santosh Khasanvis*, Mostafizur Rahman and Csaba Andras Moritz
Department of Electrical and Computer Engineering,
University of Massachusetts Amherst, USA

*Corresponding Author:
201 Knowles Engineering Building, 151 Holdsworth Way, University of Massachusetts Amherst, Amherst, MA – 01003, USA
Email: khasanvis@ecs.umass.edu
Phone: +1-413-695-1531

ABSTRACT

Leveraging nanotechnology for computing opens up exciting new avenues for breakthroughs. For example, graphene is an emerging nanoscale material and 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 propose a novel graphene nanoribbon tunneling ternary content addressable memory (GNTCAM) enabled by xGNR device, featuring heterogeneous integration with CMOS transistors and routing. Benchmarking with respect to 16nm CMOS TCAM (which uses two binary SRAMs to store ternary information) shows that GNTCAM is up to 1.82x denser, up to 9.42x more power-efficient during stand-by, and has up to 1.6x faster performance during match operation. Thus, GNTCAM has the potential to realize low-power high-density nanoscale TCAMs. Further improvements may be possible by using graphene more extensively, as graphene transistors become available in future.

Keywords- *Graphene Nanoribbons; CAM; TCAM; NDR; Ternary Memory; Heterogeneous Integration; GNTCAM.*

1. INTRODUCTION

Content Addressable Memory (CAM) is a widely used hardware search engine for high-speed parallel data search applications. In addition to storing information, CAM provides the capability to search input data across an entire memory array simultaneously. This typically enables CAM to have a single clock cycle throughput, making it faster than other hardware/software based search systems. It is primarily used in networking applications for classification and forwarding of data packets in routers [1], which require high capacity and high speed CAMs. A few other common applications include translation look-aside buffers (TLBs) for memory systems [2][3], tag arrays in associative cache memories [4] and image coding [5].

Two kinds of CAM cells available today are the binary CAM (BCAM) and ternary CAM (TCAM) [1]. BCAM can store and search binary information (*logic 0* and *logic 1*) and is used in applications requiring an exact match with fully specified inputs. TCAM provides the capability of storing and matching against a *don't care* state (X) in addition to *logic 0* and *logic 1*. This feature extends the search capability by enabling partial matching of input data against stored data, which is a requirement in networking applications [1][6]. Conventional CMOS TCAMs achieve this by using two Static Random Access Memory (SRAM) cells for storing ternary information. A typical TCAM cell schematic [1] is shown in Fig. 1. However, TCAMs are more expensive than BCAMs and SRAMs due to increased silicon area and have high power consumption [7]. Scaling trends [8] suggest a growing demand for TCAM capacity, making it critical to reduce the cost and power consumption. While SRAM cells have been physically scaled down over the past to meet TCAM capacity requirements, the area scaling is slowing down recently (50% to 30% reduction per generation [9]) due to several issues at nanoscale, such as increased leakage and variability [10][11]. This calls for new technology and new concepts to meet growing TCAM demands.

One such concept is to use memory cells which have more than two stable states rather than emulating ternary information with binary memory. Such a multi-state memory can compressively store ternary information in a single cell, thus potentially reducing the area and power consumption. 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 believed to be a potential candidate for post-Si nanoscale computing systems [12]. It exhibits extraordinary electrical and thermal properties featuring Dirac fermion [13] with very high conductivity [14] and extreme scalability. Its planar structure also makes it compatible with current CMOS fabrication processes [15]. Several graphene-based transistors have been proposed [16]-[20], however challenges still exist which preclude their use in digital systems [21]. A novel bi-layer graphene nanoribbon crossbar tunneling device (xGNR) was reported recently [22]-[25], which exhibits negative differential resistance (NDR). This xGNR NDR device has potential applications in multi-state logic and memory circuits that will be leveraged in this work.

In this paper, we propose a TCAM cell using the xGNR device, called Graphene Nanoribbon tunneling Ternary Content Addressable Memory (GNTCAM). The contributions include (i) novel GNTCAM cell design using a ternary memory core, (ii) heterogeneous graphene-CMOS implementation and (iii) benchmarking against 16nm CMOS TCAM cell. Our evaluations show that the proposed GNTCAM has up to 1.82x density-per-bit benefit against 16nm CMOS TCAMs, is up to 1.6x faster during match operation, and up to 9.42x more power efficient per-bit when compared against the CMOS TCAM in idle periods. Further improvements may be possible by using graphene more extensively instead of silicon MOSFET transistors, as advances are made in graphene technology. While this work introduces the GNTCAM cell design, the study of effect of line edge roughness in GNRs and noise considerations are out of scope of this paper and will be a part of the future work.

The rest of the paper is organized as follows. Section 2 provides a background on the xGNR device and previous work based on this device. Section 3 proposes a new ternary content addressable memory cell and Section 4 describes a physical implementation with heterogeneous integration between CMOS and graphene. Methodology and benchmarking are presented in Section 5 followed

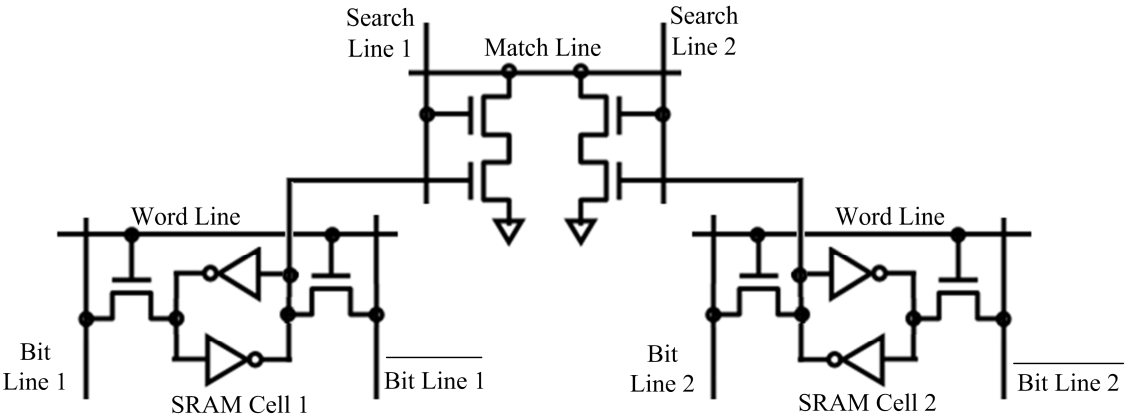

Figure 1. Conventional CMOS Ternary CAM – Uses two SRAM cells to store ternary information.

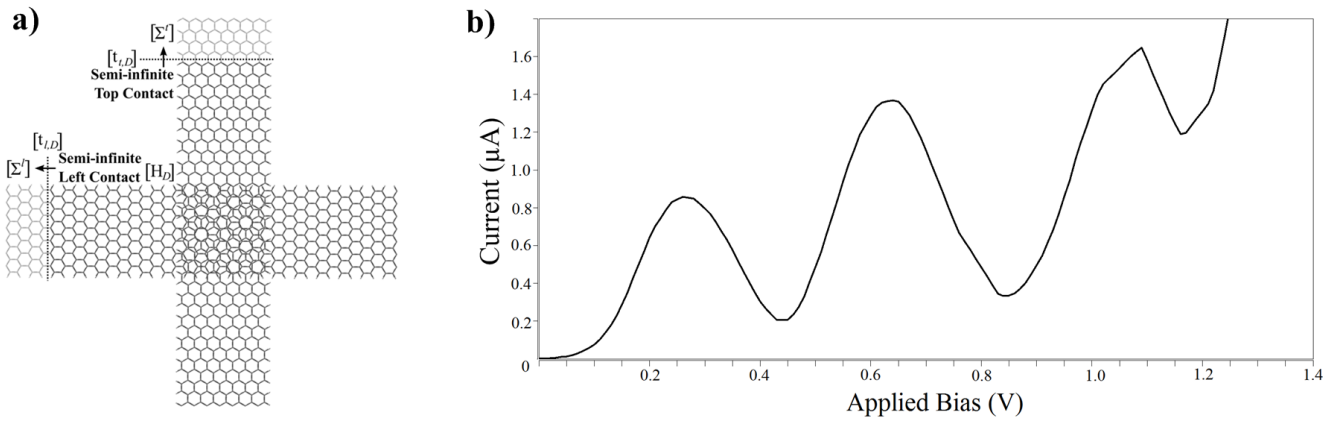

Figure 2. (a) Atomistic geometry of the GNR crossbar device (xGNR). Two hydrogen passivated relaxed armchair type GNRs are placed on top of each other at a right angle with a vertical separation of 3.35 Å [23]-[25]. A bias is applied by independently contacting each GNR such that one is held at ground 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.

by conclusion in Section 6.

2. Background and Previous Work

2.1 Graphene Nanoribbon Crossbar (xGNR) Device

The graphene nanoribbon crossbar shown in Fig. 2a 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 [22]-[25]. 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 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 first principle atomistic calculations [22]-[25]. The simulated I-V characteristic of the xGNR is shown in Fig. 2b exhibiting negative differential resistance (NDR) with multiple peak and valley currents, which makes it suitable for RTD-based applications [27]. The NDR is attributed to the localization of the electronic states near the cut-ends of the GNRs [23][25]. 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.

2.2 Application of xGNR Device as a Ternary Memory Element

A latch can be built to leverage on NDR behavior by connecting two xGNRs in series as shown in Fig. 3a (similar to a Goto pair). 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. 3b. The

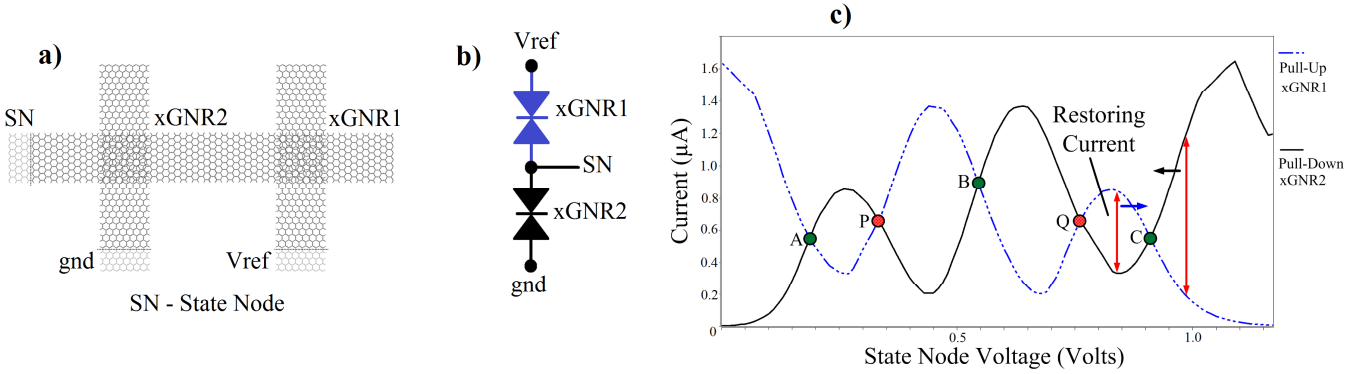

Figure 3. (a) xGNR latch configuration; (b) Circuit schematic; and (c) DC load line analysis showing multiple stable states.

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 and C under applied voltage, as shown in Fig. 3c. 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 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. 3c 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 [24]. 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 binary and ternary random access memory cells using xGNR latch as the memory core [24][25]. We now build on this concept to propose a ternary content addressable memory (CAM) design where all three states are used to store information in a single CAM cell.

3. Proposed Graphene Nanoribbon Ternary Content Addressable Memory

The xGNR latch with three stable states can be used to store ternary information electrically in a single cell. To build a ternary content addressable memory (CAM) cell using this memory core, access to the state node is required. This is achieved with the help of transistors for cell selection, write and match operations. 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-power Graphene Nanoribbon-tunneling Ternary Content Addressable Memory (GNTCAM) as shown in Fig. 4a. This cell uses three stable states to encode ternary information as shown in Fig. 4b. 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. To enable matching the stored information with search data, a match circuit implementing an XNOR is used as shown in the cell design. A match condition occurs at the match line output if it remains at logic 1 after an initial precharge operation to VDD. When a miss is encountered, the match line output is discharged through the match circuit.

The match circuit is a pull-down path, which consists of a pass-transistor implementation of an XNOR gate in series with match-enable transistors M1 and M2. Transistor M1 is gated by the state node and is used to distinguish between *don't care* condition and other stored states. This transistor is in an OFF state when the state node encodes *don't care* condition, thus ensuring that the match-line is cut-off from the pull-down path irrespective of the search data. When state node is at *logic 0/logic 1*, transistor

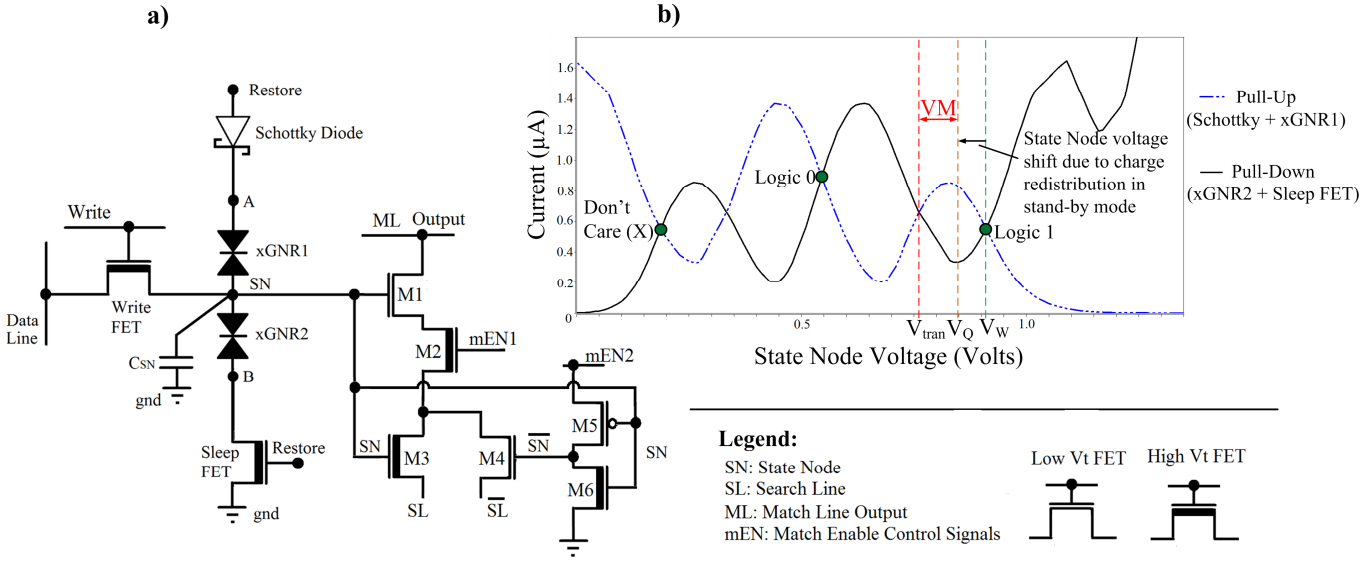

Figure 4. (a) Proposed GNTCAM circuit schematic; (b) DC load line analysis for xGNR latch including Schottky Diode and Sleep FET showing multiple stable states – X, Logic 0 and Logic 1.

Table I. Match Operation

Stored Data at SN	Search Data	Match Output
Don't Care (0.15V)	SL = 1	VDD (Match)
	SL = 0	
Logic 0 (0.53V)	SL = 1	0 (Miss)
	SL = 0	VDD (Match)
Logic 1 (0.86V)	SL = 1	VDD (Match)
	SL = 0	0 (Miss)

M1 is switched ON and the match-line is allowed to discharge through the pull-down path in case of a mismatch. Transistor M2 acts as an evaluate transistor gated by the *Match-Enable* control signal, which initiates the match operation.

Transistors M3 and M4 realize the XNOR gate to compare the stored *logic 0/logic 1* state with the search data. Since a complement of the stored data is required for this operation, the match circuit also includes an inverter (transistors M5 and M6) as shown in Fig. 4a. The inverter is connected to *Match-Enable2* signal which is active during match operation. This signal is switched OFF during stand-by, since the inverter is only used for match operation and not data storage. Transistor M3 is gated by the state node and distinguishes between stored *logic 0* and *logic 1* states. Thus if the search data matches with the stored data, both pull-down paths are deactivated allowing the match-line to remain at VDD. In case of a mismatch, one of the pull-down paths is active leading to the match-line being discharged. The cell operation is described in the following sub-sections.

3.1 Write Operation

During a write operation, the appropriate cell is selected by activating the corresponding write-line and applying the required input voltage onto the data-line. Here, the value of the applied voltage on the data-line encodes the state to be written and is in ternary representation (0V – *don't care (X)*, 0.6V – *logic 0* and 1.0V – *logic 1*). The input voltage values to be applied on the data-line are chosen based on the voltages at which stable states occur in the xGNR latch. Fig. 4b 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 encoding in the graph.

Consider that the state node is initially at a voltage of 0.15V (*don't care* condition). To write *logic 0*, the appropriate input 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 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 1* is written by applying the appropriate input voltage on the data-line and charging up the state capacitance. The state node can be made to store *don't care* condition by applying the appropriate input voltage (0V) 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.

3.2 Match Operation

The state node is used to gate the match transistors and hence is isolated from the output match line. This scheme ensures that this operation is non-destructive. The match-line is initially pre-charged to VDD; the search data (and its complement) is broadcasted on the search lines and the *Match-Enable* control signals are switched ON. If at the end of the operation the match-line remains at VDD, then it implies a match condition between the stored data and search data. A miss condition is indicated by the discharge of the match-line at the end of the match operation.

For example, if the state node encodes *don't care* condition, transistor M1 is switched OFF and hence match-line remains at VDD. When the state node is at *logic 0* (0.55V) and the search data is *logic 0*, M3 is switched OFF while the source terminal of M4 is at VDD. Thus no pull-down path exists in this case. If the search data is *logic 1*, a pull-down path exists through transistor M4. The match operation for a stored *logic 1* state proceeds similarly. Operation conditions are outlined in Table I.

3.3 Restore Operation

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

3.4 Circuit Implementation

This implementation requires a multi-threshold (V_t) circuit design since each of the transistors in the match circuit are activated at different voltages. The transistor M1 needs to have a low threshold voltage between the voltages at *state X* and *logic 0* states in Fig. 4b for the desired operation. Transistor M3 needs to have a relatively high threshold voltage (between the voltages at *logic 0* and *logic 1* states in Fig. 4b) so that it can distinguish between stored *logic 0* and *logic 1* state. The inverter is designed such that *logic 0* (0-0.55V) at its input causes the output to be *logic 1*. This is achieved by using a low-V_t pMOS and a high-V_t nMOS as shown.

The value of the state capacitance (C_{SN}) is determined by (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. 4b, for the case of storing *logic 1*. This is the 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. 4b), the restore operation causes a state transition to *logic 0* instead of restoring *logic 1* 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. 4b. 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:

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

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 match transistors (M1, M3, M5, M6), 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}.

4. PHYSICAL LAYOUT

A cross-technology heterogeneous integration between CMOS and graphene [24][25] is used, as shown in Fig. 5. The Silicon 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 [28][29]. This feature is used to realize a Schottky diode with the help of a Schottky contact between a narrow semiconducting armchair GNR and metal, as shown in Fig. 5c. The rest of the graphene-metal contacts are Ohmic to ensure proper operation and this is achieved by using wide GNRs [30].

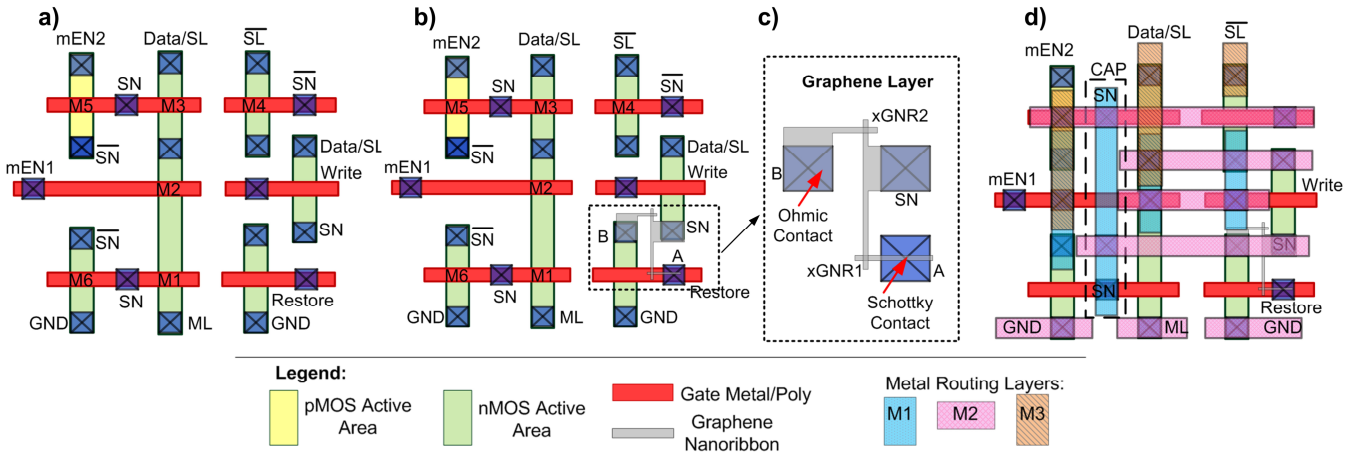

Figure 5. Proposed GNTCAM Layout - (a) Bottom Si layer showing transistors; (b) Graphene layer integrated on top of transistors; (c) Detailed top view of graphene layer showing xGNRs and metal contacts for interfacing with bottom Si layer; and (d) Routing stack integrated on top of graphene and Si layers.

TABLE II. DESIGN RULES

1D Gridded Design [32]	*M1, M2 Interconnect*	*Poly*
Pitch (16nm technology node)	40~60 nm	60~80nm

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 write and match operations (for search input) since only one of these operations is performed on a CAM cell at a given time. A separate routing line is used for the complement of the search input required for match operation.

A lithography-friendly grid-based layout is used with minimum sized transistors for high density and ease of fabrication. Routing is achieved with the help of a conventional CMOS metal stack. The state capacitor may be implemented as a stacked capacitor over the state node routing area shown in Fig. 5d.

5. METHODOLOGY AND BENCHMARKING

HSPICE circuit simulator was used to simulate and verify the GNTCAM cell operation and for benchmarking. 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 [31] were used to simulate the match, 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 3. This ensures that when a restore signal is applied at a period of 0.4µs, the state node is brought up to the required stable point. A higher capacitance value would lead to a longer retention time at the cost of a slower write operation.

The simulation waveforms for write and match operations are shown in Fig. 6a. The state node is initialized to 0V and *logic 0* is first written and then the match operation is performed against input search data. During the period when match enable signal mEN1 is ON, the match output discharges to 0V for a search data of *logic 1* indicating a miss condition. For the search data of *logic*

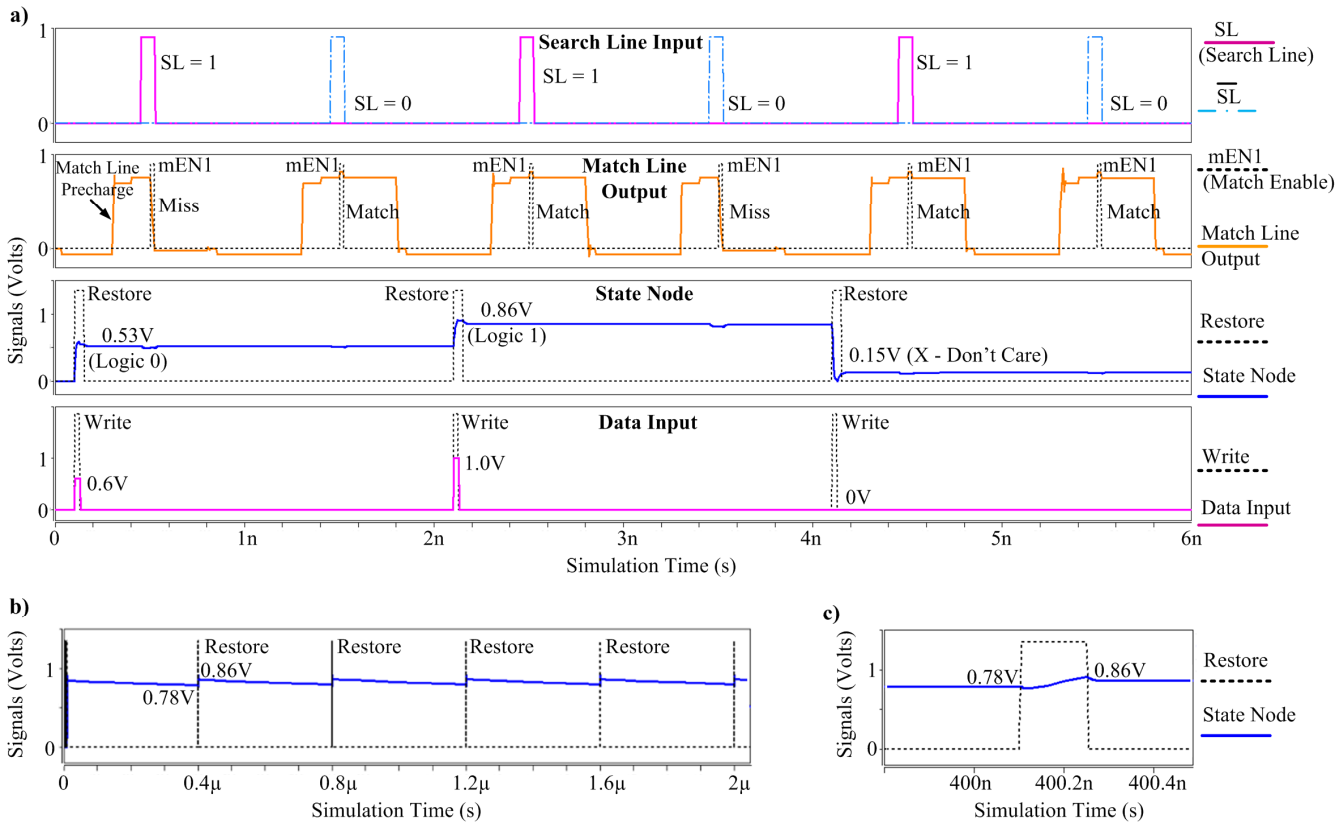

Figure 6. Simulation output for GNTCAM – (a) Write and match operations; (b) Restore operation for logic 1 at a period of 0.4µs; and (c) Detailed restore operation at t = 0.4µs when storing logic 1 at state node

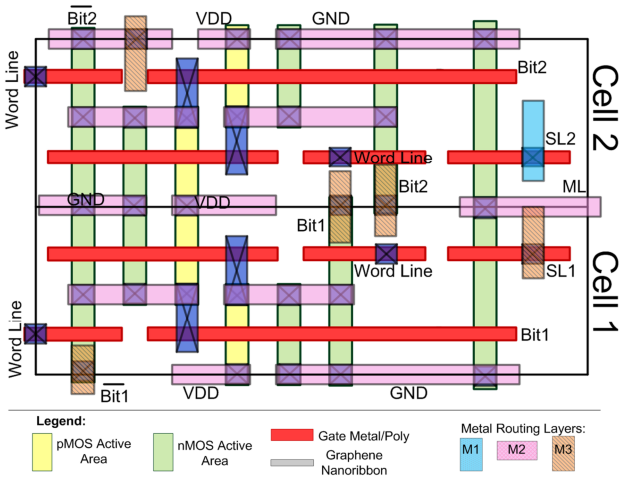

Figure 7. CMOS TCAM layout (ML – Match Line, SL – Search Line) based on circuit schematic in Fig.1. The layout consists of two 8T grid-based SRAM cells [33] with match circuit using 16nm gridded design rules [32].

0, the match output remains at VDD indicating a match condition. This is repeated for a stored data of *logic 1* and also for a stored *don't care* condition. Restore operation is performed at a period of 0.4µs, as shown in Fig. 6b for the case of restoring *logic 1*. The circuit operated as outlined in Section 3. For physical layout design, 1-D Gridded design rules [32] (shown in Table II) were used to evaluate the area of GNTCAM cell in 16nm technology node.

A conventional 16nm CMOS ternary CAM using two 8T SRAM cells was used for benchmarking. The 8T SRAM cell design [33] is based on the same 16nm gridded design rules used for GNTCAM. The circuit schematic [1] is shown in Fig. 1 and the physical layout is shown in Fig. 7. In order to evaluate the power and performance of GNTCAM and 16nm CMOS ternary CAM, PTM RC models [31] and 16nm PTM transistor models [31] were used for simulation with HSPICE. Table III shows the comparison results with 16nm CMOS ternary CAM cell.

5.1 Area Evaluation

GNTCAM showed significant density advantage compared to 16nm CMOS TCAM. Specifically, GNTCAM showed up to 1.82x smaller area per bit compares to CMOS TCAM. This is mainly because GNTCAM uses a single cell for encoding ternary information as opposed to CMOS TCAM which uses two SRAM cells, thus providing an immediate area benefit. The GNTCAM cell area in this design is primarily dominated by silicon transistors and routing. As graphene technology matures, the availability of graphene transistors would enable a monolithic graphene fabric with potentially much higher area benefits.

5.2 Power Evaluation

Since all the cells in a memory array are active during match operation, the power consumption per cell during this period is

TABLE III. GNTCAM BENCHMARKING

		GNTCAM (Per bit – 1 Cell)	***16nm CMOS TCAM (Per bit – 2 Cells)***
Area Comparison (μm^2)		0.0616 – 0.1344	0.1126 – 0.2283
Power Consumption	Active Power *(μW)*	0.85 – 0.94	1.13 – 1.44
	Stand-by Power *(pW)*	49.41 – 79.66	465.35 – 465.75
Performance	Match Operation *(ps)*	14.7 – 15.41	18.82 – 24.55
	Write Operation *(ps)*	14.05 – 15.9	27.07 – 28.25

critical. In terms of active power during match operation, the GNTCAM cell operated at up to 1.53x lower power than 16nm CMOS TCAM. During idle periods, GNTCAM was up to 9.42x more power-efficient compared to 16nm CMOS TCAM. The idle power benefits are because of two reasons – (i) GNTCAM is dynamic and hence no static paths exist to contribute to idle power, and (ii) GNTCAM stores ternary information in a single cell, thus reducing leakage costs.

5.3 Performance Evaluation

GNTCAM was up to 1.6x faster during match operation compared to 16nm CMOS TCAM. This is because – (i) the match line is shorter in GNTCAM leading to lower output load resistance and capacitance; and (ii) GNTCAM uses both high-performance and low-power devices in its match pull-down path, while the CMOS TCAM uses only low-power transistors. An asymmetric (multi-Vt) approach was necessary in GNTCAM to successfully differentiate between three stored states. The write performance of GNTCAM is better than the CMOS TCAM because of the boosted gate voltage necessary to overcome the threshold voltage drop, when writing *logic 0* and *logic 1* at the state node.

6. CONCLUSION

A graphene nanoribbon tunneling ternary content addressable memory (GNTCAM) cell was presented in this paper, enabled by new nanomaterials like graphene and unique graphene nanoribbon structures. This memory technology could be potentially highly beneficial in many computing applications from networks to microprocessor caches. GNTCAM was implemented with a heterogeneous integration between CMOS and graphene technologies. Benchmarking against 16nm CMOS TCAM design showed that GNTCAM exhibited significant benefits, stemming from compressively storing ternary information in a single cell. Specifically, GNTCAM was up to 1.82x denser, up to 1.6x faster and exhibited up to 9.42x lower stand-by power consumption when compared to 16nm CMOS TCAM.

Future work would study the effect of line edge roughness in GNRs and noise analysis. As progress is made in graphene technology, further benefits may be expected by replacing Si MOSFETs with graphene transistors and other graphene xGNR based logic.

REFERENCES

[1] Pagiamtzis, K.; Sheikholeslami, A.; , "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," Solid-State Circuits, IEEE Journal of , vol.41, no.3, pp. 712- 727, March 2006.

[2] Higuchi, H.; Tachibana, S.; Minami, M.; Nagano, T.; , "A 2-ns, 5-mW, synchronous-powered static-circuit fully associative TLB," VLSI Circuits, 1995. Digest of Technical Papers., 1995 Symposium on , vol., no., pp.21-22, 8-10 June 1995.

[3] M. Sumita, "A 800 MHz single cycle access 32 entry fully associative TLB with a 240ps access match circuit," Digest of Technical Papers of the Symposium on VLSI Circuits, pp. 231-232, Jun. 2001.

[4] Perng-Fei Lin; Kuo, J.B.; , "A 1-V 128-kb four-way set-associative CMOS cache memory using wordline-oriented tag-compare (WLOTC) structure with the content-addressable-memory (CAM) 10-transistor tag cell," Solid-State Circuits, IEEE Journal of , vol.36, no.4, pp.666-675, Apr 2001.

[5] Panchanathan, S.; Goldberg, M.; , "A content-addressable memory architecture for image coding using vector quantization," Signal Processing, IEEE Transactions on , vol.39, no.9, pp.2066-2078, Sep 1991.

[6] Gamache, B.; Pfeffer, Z.; Khatri, S.P.; , "A fast ternary CAM design for IP networking applications," Computer Communications and Networks, 2003. ICCCN 2003. Proceedings. The 12th International Conference on , vol., no., pp. 434- 439, 20-22 Oct. 2003.

[7] Ashish Goel; and Pankaj Gupta; "Small subset queries and bloom filters using ternary associative memories, with applications", in Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems (SIGMETRICS '10), ACM, New York, NY, USA, 143-154, 2010.

[8] Rajendran, B.; Cheek, R.W.; Lastras, L.A.; Franceschini, M.M.; Breitwisch, M.J.; Schrott, A.G.; Jing Li; Montoye, R.K.; Chang, L.; Chung Lam; , "Demonstration of CAM and TCAM Using Phase Change Devices," Memory Workshop (IMW), 2011 3rd IEEE International , vol., no., pp.1-4, 22-25 May 2011.

[9] 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

[10] 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.

[11] 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.

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

[13] 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.

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

[15] de Heer, W.A.; Berger, C.; Conrad, E.; First, P.; Murali, R.; MeindI, 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

[22] 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).

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

[24] 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.

[25] Khasanvis, S.; Habib, K.M.M.; Rahman, M.; Narayanan, P.; Lake, R.K.; Moritz, C.A., "Ternary volatile random access memory based on heterogeneous graphene-CMOS fabric," Nanoscale Architectures (NANOARCH), 2012 IEEE/ACM International Symposium on , vol., no., pp.69,76, 4-6 July 2012.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] 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.

FIGURE CAPTIONS

Figure 1. Conventional CMOS Ternary CAM – Uses two SRAM cells to store ternary information.

Figure 2. (a) Atomistic geometry of the GNR crossbar device (xGNR). 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 [23]-[25]. 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.

Figure 3. (a) xGNR latch configuration; (b) Circuit schematic; and (c) DC load line analysis showing multiple stable states.

Figure 4. (a) Proposed GNTCAM circuit schematic; (b) DC load line analysis for xGNR latch including Schottky Diode and Sleep FET showing multiple stable states – X, Logic 0 and Logic 1.

Figure 5. Proposed GNTCAM Layout - (a) Bottom Si layer showing transistors; (b) Graphene layer integrated on top of transistors; (c) Detailed top view of graphene layer showing xGNRs and metal contacts for interfacing with bottom Si layer; and (d) Routing stack integrated on top of graphene and Si layers.

Figure 6. Simulation output for GNTCAM – (a) Write and match operations; (b) Restore operation for logic 1 at a period of 0.4μs; and (c) Detailed restore operation at t = 0.4μs when storing logic 1 at state node.

Figure 7. CMOS TCAM layout (ML – Match Line, SL – Search Line) based on circuit schematic in Fig.1. The layout consists of two 8T grid-based SRAM cells [33] with match circuit using 16nm gridded design rules [32].