2012

Heterogeneous Graphene Nanoribbon-CMOS Multi-State Volatile Random Access Memory Fabric

Santosh Khasanvis
University of Massachusetts Amherst

HETEROGENEOUS GRAPHENE NANORIBBON-CMOS MULTI-STATE VOLATILE RANDOM ACCESS MEMORY FABRIC

A Thesis Presented

by

SANTOSH KHASANVIS

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

September 2012

Department of Electrical and Computer Engineering

HETEROGENEOUS GRAPHENE NANORIBBON-CMOS MULTI-STATE VOLATILE RANDOM ACCESS MEMORY FABRIC

A Thesis Presented

by

SANTOSH KHASANVIS

Approved as to style and content by:

_____
Csaba Andras Moritz, Chair

_____
Israel Koren, Member

_____
C. Mani Krishna, Member

_____
C.V. Hollot, Department Head
Department of Electrical and Computer Engineering

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank all the people without whom this thesis would not be possible.

First and foremost, I would like to extend my sincere gratitude to my advisor, Prof. Csaba Andras Moritz, for his constant support, inspiring advice and encouragement. I would also like to thank my committee members, Prof. Israel Koren and Prof. Mani Krishna for their time, advice and suggestions. I would like to acknowledge the collaboration with Prof. Roger Lake and his graduate student K. Masum Habib at University of California Riverside, who provided the xGNR device used in this work.

I would like to especially acknowledge the guidance and assistance provided by Pritish Narayanan, Mostafizur Rahman and others in the Moritz Research Group. I thank all my friends at Amherst for making my stay so enjoyable.

Last but not least, I would like to thank my family and friends for their support, encouragement and trust in my ability.

ABSTRACT

HETEROGENEOUS GRAPHENE NANORIBBON-CMOS MULTI-STATE VOLATILE RANDOM ACCESS MEMORY FABRIC

SEPTEMBER 2012

SANTOSH KHASANVIS

B.TECH, VELLORE INSTITUTE OF TECHNOLOGY UNIVERSITY

M.S.E.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Csaba Andras Moritz

CMOS SRAM area scaling is slowing down due to several challenges faced by transistors at nanoscale such as increased leakage. This calls for new concepts and technologies to overcome CMOS scaling limitations. In this thesis, we propose a multi-state memory to store multiple bits in a single cell, enabled by graphene and graphene nanoribbon crossbar devices (xGNR). This could provide a new dimension for scaling. We present a new multi-state volatile memory fabric called Graphene Nanoribbon Tunneling Random Access Memory (GNTRAM) featuring a heterogeneous integration between graphene and CMOS. A latch based on the xGNR devices is used as the memory element which exhibits 3 stable states. We propose binary and ternary GNTRAM and compare them with respect to 16nm CMOS SRAM and 3T DRAM. Ternary GNTRAM (1.58 bits/cell) 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 1196x more power-efficient per bit against high-performance CMOS SRAMs during stand-by.

To enable further scaling, we explore two approaches to increase the number of bits per cell. We propose quaternary GNTRAM (2 bits/cell) using these approaches and

extensively benchmark these designs. The first uses additional xGNR devices in the latch to achieve 4 stable states and the quaternary memory shows up to 2.27x density benefit vs. 16nm CMOS SRAMs and 1.8x vs. 3T DRAM. It has comparable read performance in addition to being power-efficient, up to 1.32x during active period and 818x during stand-by against high performance SRAMs. However, the need for relatively high-voltage operation may ultimately limit this scaling approach. An alternative approach is also explored by increasing the stub length in the xGNR devices, which allows for storing 2 bits per cell without requiring an increased operating voltage. This approach for quaternary GNTRAM shows higher benefits in terms of power, specifically up to 4.67x in terms of active power and 3498x during stand-by against high-performance SRAMs.

Multi-bit GNTRAM has the potential to realize high-density low-power nanoscale memories. Further improvements may be possible by using graphene more extensively, as graphene transistors become available in future.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION AND MOTIVATION

The semiconductor IC industry has witnessed a tremendous growth in the functional and processing capabilities of microprocessors over the past few decades. This has primarily been stimulated by physical downscaling of CMOS devices which provides cost, performance, power and density benefits simultaneously. These have been the key drivers of the extra-ordinary progress in electronics leading to the ubiquitous computing with advanced capabilities available today.

Traditionally, the microprocessor development has continued independent of memory which has led to an exponential increase in processor speed, while memory latencies have not shown such a dramatic improvement resulting in a widening processor-to-memory

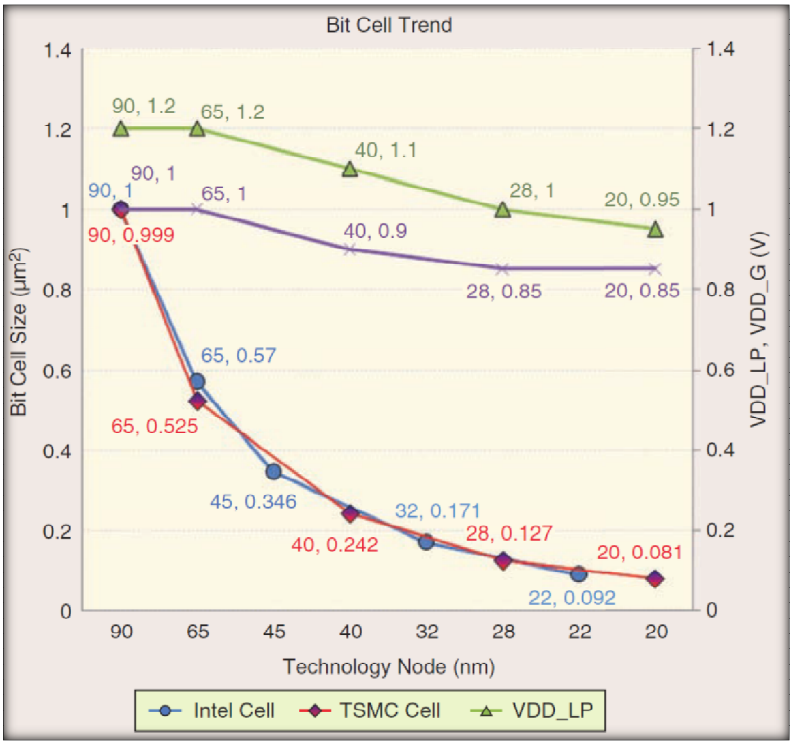


Figure 1. SRAM bit-cell area and VDD trends showing a slowdown in SRAM area scaling from 50% to 30% per generation [2].

gap. This phenomenon is termed as “the Memory Wall” [1] where increasing memory access times severely limit system performance. This problem has been addressed by adding several levels of high-speed caches, in addition to other architectural techniques to hide the memory latency. To implement these on-chip caches, CMOS SRAM has been widely used due to its high access speed.

As microprocessors evolve with increased functionality and higher performance for every new generation, applications get more demanding on computing resources. As a result, on-chip cache memory density has dramatically increased over the past years to accommodate the growing demands for high-performance computing. In order to maintain this historical growth in memory density, SRAM bit cell size has been aggressively scaled down for every generation along the semiconductor technology roadmap. However, there has been a slowdown in SRAM cell area scaling from 50% to 30% reduction per generation [2] (see Figure 1) due to several challenges such as increased leakage and variability at nanoscale [3][4]. 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 Figure 2. We propose a multi-state memory concept which is enabled by

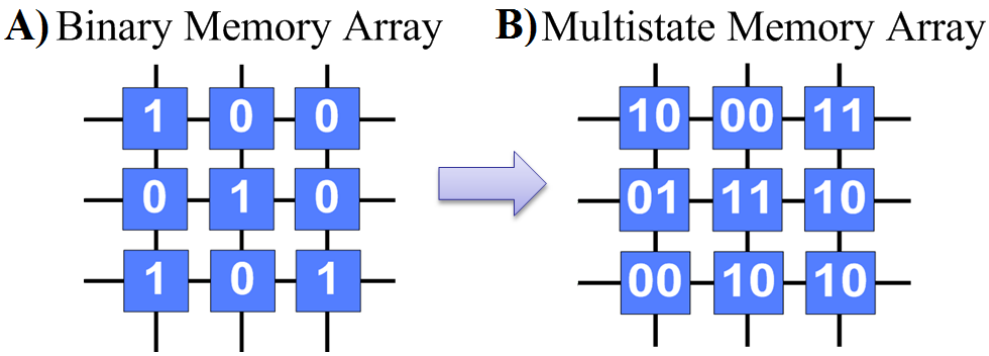


Figure 2. A) Current technology uses binary memory storing a single bit per cell; B) Proposed concept: Multi-bit per cell with novel graphene structures.

emerging nanoscale materials like graphene and unique material interactions between novel device structures. It could potentially provide a new dimension for scaling as an alternative to physical downscaling, by compressively storing multiple bits in a single cell.

Multi-state circuits using negative differential resistance (NDR) based resonant tunneling diodes (RTDs) have been extensively researched in the past [5]-[8]. 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 [9]. Due to technological and economical barriers, RTDs using III-V materials could only be used in niche applications. On the other hand, devices based on emerging materials like graphene overcome such integration challenges and have the potential to be used in mainstream applications.

In this thesis, we explore new multi-state memories enabled by novel graphene nanoribbon devices to replace CMOS SRAM for implementing on-chip caches. We propose a heterogeneous integration between graphene and CMOS technologies to implement a novel Graphene Nanoribbon crossbar (xGNR) based Tunneling volatile Random Access Memory (GNTRAM). We start by introducing the design of a binary memory cell with this approach and proceed to realize the ternary version that stores 1.5 bits per cell. We also explore possible scaling approaches to increase the number of bits that can be stored in a cell, as alternatives to physical scaling. We present quaternary GNTRAM designs based on these scaling approaches which can store 2 bits per cell. We extensively benchmark these designs against 16nm CMOS 6T and 8T SRAMs and 3T DRAM in terms of density, power and performance. Our analysis shows that multistate

GNTRAM designs have significant benefits against state-of-the-art CMOS RAMs in terms of density and power, while having comparable performance. 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.

The rest of the thesis is organized as follow. Chapter 2 provides an overview on graphene and briefly introduces the new graphene nanoribbon crossbar (xGNR) device. Chapter 3 explores the application of this device as a multi-state latch and presents the basics of a multi-state memory cell design using this latch. Chapter 4 discusses specific circuit and physical implementations of binary and ternary GNTRAM. Chapter 5 presents detailed comparison of the binary and ternary GNTRAM with state-of-the-art CMOS SRAM and 3T DRAM. Chapter 6 explores possible scaling approaches with GNTRAM and presents quaternary GNTRAM design using these approaches, with detailed benchmarking of each design vs. CMOS. Finally, chapter 7 concludes the thesis with a brief discussion on possible future work in this direction.

CHAPTER 2

GRAPHENE NANORIBBON CROSSBAR DEVICE: BACKGROUND

We briefly introduce the properties of graphene, an emerging nanoscale material for post-CMOS nanoelectronics. The advantages and challenges of carbon based allotropes are briefly discussed. A new graphene nanoribbon device is also introduced which exhibits negative differential resistance (NDR).

2.1 Graphene

Due to the scaling limitations with CMOS, alternative materials other than Si are a subject of intense research to build integrated circuit logic and memory. Carbon is often seen as a candidate material for post-CMOS electronics; in particular its low-dimensional allotropes like carbon nanotubes and graphene (see Figure 3). Both materials are of great interest for electronics due to their exotic properties such as high electrical and thermal conductivities, extraordinary mechanical strength and ultimate scalability down to the

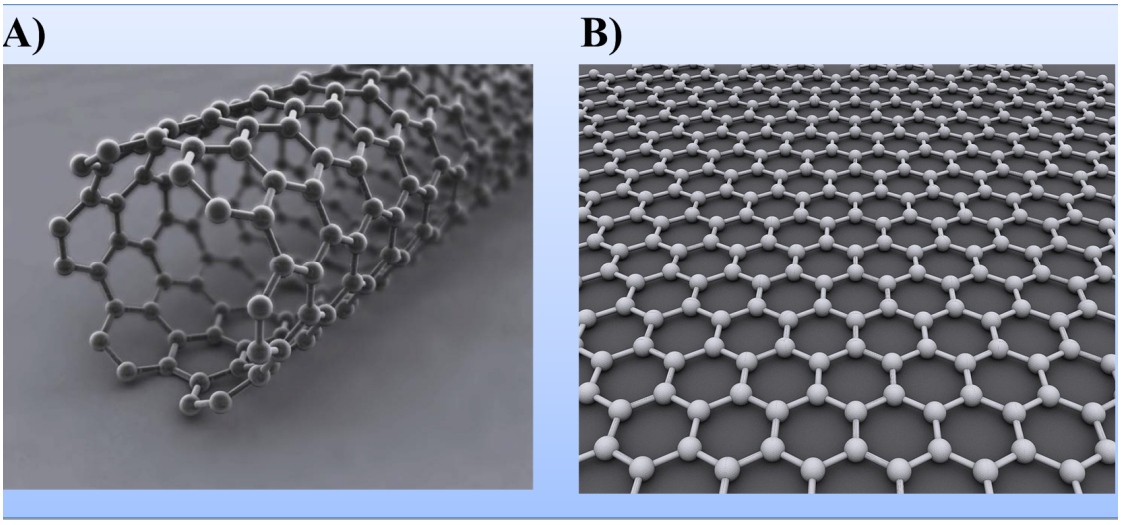


Figure 3. Carbon allotropes – Potential candidates for post-CMOS electronics: A) Carbon nanotube [10]; and B) Graphene (Source: Wikipedia).

atomic level. In the field of electronics, these properties potentially enable very small feature sizes leading to high performance devices and interconnects. Carbon nanotubes, however, face several challenges including requiring processing techniques that provide a tight distribution of semiconductor bandgaps, alignment and placement precision and compatibility with CMOS processing [11].

Graphene is a single atomic layer thick, 2-dimensional allotrope of carbon with a hexagonal lattice structure. It shares all of the extraordinary electronic properties of carbon nanotubes, with the additional benefit of being compatible with CMOS processing techniques due to its planar structure. This fact has led to wide-spread research on graphene electronics, and it is touted to be a potential candidate for "next-generation" post-CMOS nanoelectronics. Low-cost large-scale synthesis of graphene with CVD and epitaxial growth techniques has been shown and other techniques are currently being

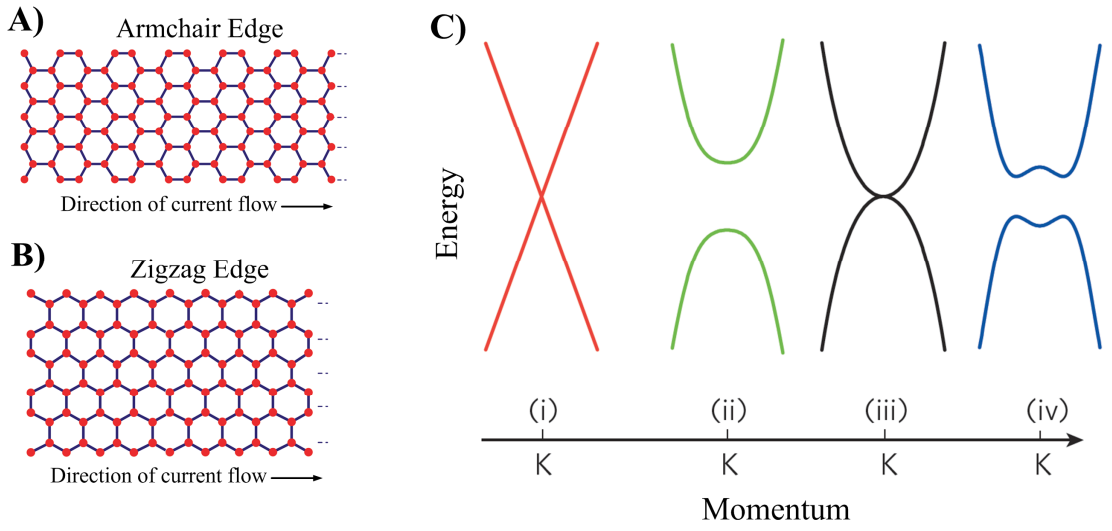


Figure 4. Graphene band-gap manipulation with quantum confinement to create graphene nanoribbons having – A) Armchair geometry and B) Zigzag geometry. C) Energy-Momentum relationships for various graphene configurations: (i) Wide-area graphene; (ii) Graphene nanoribbons; (iii) Unbiased bi-layer graphene; and (iv) Biased bi-layer graphene. Adapted from reference [16].

researched [12]-[14].

Wide area graphene is a semimetal; i.e. it has zero bandgap (see Figure 4C) which limits its application in digital electronics. This is a critical challenge in using graphene as an alternative channel material in switching devices like FETs, since the lack of a bandgap limits any electrostatic control over channel conduction. It is however of great interest as electrical interconnect [17] and in on-chip cooling networks [18][19].

It is possible to modify the band-structure of graphene to open a bandgap (Figure 4C). Some of the techniques used to do this are (i) quantum confinement by patterning monolayer graphene into narrow 1-dimensional nanoribbons, (ii) biasing bi-layer graphene and (iii) applying strain to graphene [16]. Graphene nanoribbons (GNRs) (Figure 4A and B) have been extensively studied for electronic device applications and the bandgap is inversely proportional to ribbon width to a good approximation. Although opening of a bandgap in narrow GNRs (<10nm) has been experimentally verified [20]-[22], they require very precise and well-defined edges to be useful for conventional FET devices. The alternative is to stack two monolayers to form bilayer graphene. This configuration also has a semiconducting band-structure with zero bandgap, but it can be tuned by applying a potential difference between the two layers. FET devices using bilayer graphene as channel have been studied, but they exhibit poor ON-OFF current ratios due to strong band-to-band tunneling. Several other graphene based FETs have been proposed [23]-[27], however several challenges exist which preclude their use in digital applications.

Recently, electronic transport through a bilayer GNR structure has been studied numerically by Prof. Lake's group at UCR [28]. The geometry consists of two GNRs

placed on top of each other in AA or AB sequence with an external bias applied to one with respect to the other. It has been shown that negative differential resistance (NDR) occurs in such configuration. Reference [29] considers a more realistic geometry, consisting of two GNRs placed on top of each other at right angles like a crossbar. Calculations based on ab-initio density functional theory (DFT) coupled with the non-equilibrium Green's function (NEGF) formalism, reveal that NDR also occurs in the model GNR crossbar (xGNR). This configuration is described in the next section.

2.2 Graphene Nanoribbon Crossbar (xGNR) Device

The graphene nanoribbon crossbar shown in Figure 5 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 [28][29][30][31]. 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 [32][33] is 130 meV which is in good agreement with Son et al. [34]. 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 [35], the BLYP exchange correlation functional [36][37], a

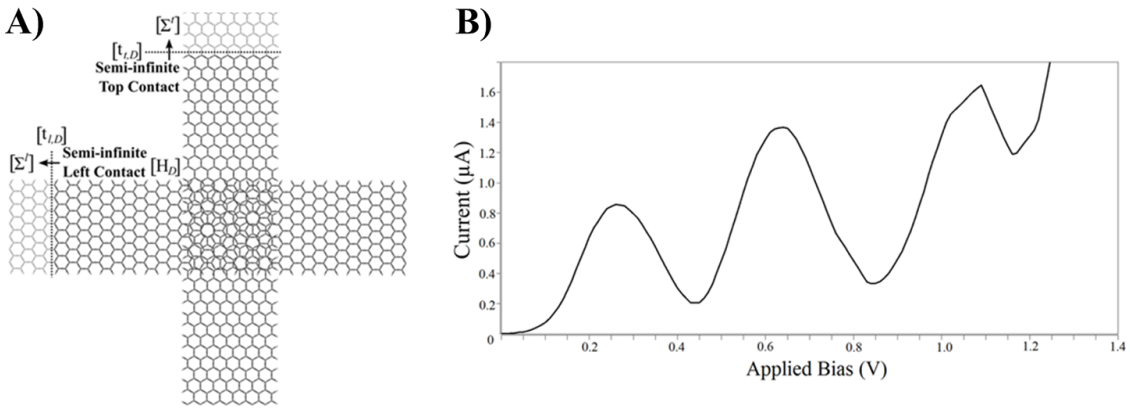


Figure 5. (A) Atomistic geometry of the GNR crossbar. Two hydrogen passivated relaxed armchair type GNRs are placed on top of each other at a right angle with a vertical separation of 3.35 Å. The relaxation was done using Fireball. The extended parts of the GNRs are used as contacts. A bias is applied 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.

self-consistent generalization of the Harris-Foulkes energy functional [38]-[41], 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 [42]. These matrix elements are used in the recursive Green's function (RGF) algorithm to calculate the transmission and the current as described in reference [43].

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 [44]. The NDR is attributed to the localization of the electronic states near the cut-ends of the GNRs [29]. 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 [31]. 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.

2.3 Chapter Summary

In this chapter, a brief background on carbon allotropes was presented. Carbon nanotubes and graphene were discussed as they are highly applicable to electronics due to their exotic electrical properties. The advantages and challenges with each were discussed. A new bilayer graphene crossbar device (xGNR) was introduced, which exhibits negative differential resistance behavior. The next chapter discusses the application of the xGNR device for integrated circuit memory to implement on-chip caches.

CHAPTER 3

GRAPHENE NANORIBBON TUNNELING RAM (GNTRAM) CELL

We now present an application of the novel xGNR device as a memory element. The xGNR device exhibits negative differential resistance (NDR) behaviour similar to resonant tunnelling diodes (RTDs). We explore one possible direction where the xGNR devices can be use in a latch configuration in volatile random access memory. In this chapter, we introduce and explain the xGNR latch configuration and analyse the DC characteristics. We also propose a volatile random access memory cell design and explain the operation.

3.1 Application of xGNR Device as a latch

The xGNR device is a two-terminal device represented using the symbol shown in Figure 6. A series stack of two xGNR devices (Figure 7A) leverages NDR characteristics to exhibit multiple stable states A, B & C as shown in Figure 7C. This xGNR series configuration can be used as a binary latch or multi-state latch, where the information is

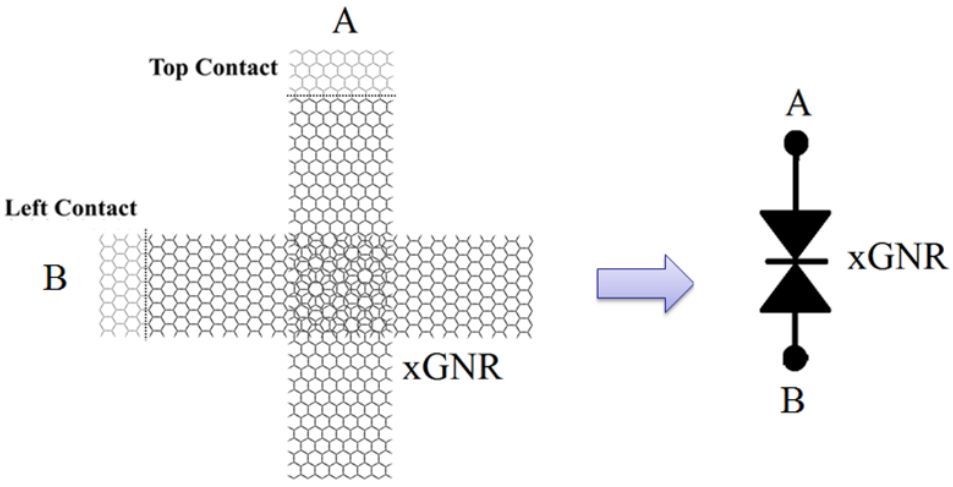


Figure 6. Two terminal xGNR device and its circuit symbol.

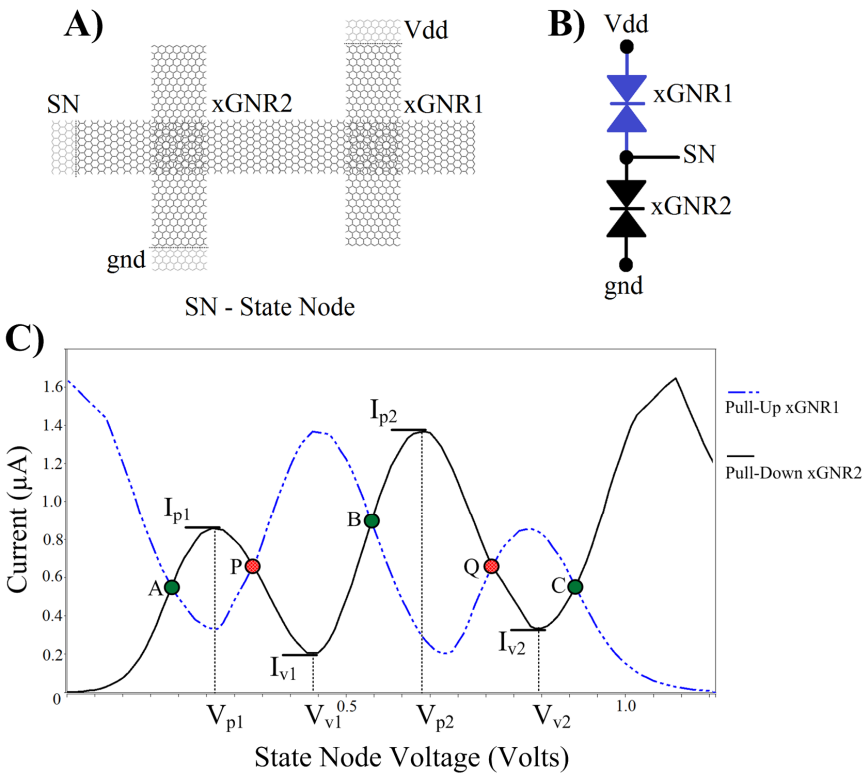


Figure 7. A) xGNR latch configuration; B) Circuit schematic; and C) DC load line analysis showing 3 stable states.

stored in the voltage level at the common terminal (state node).

The latching mechanism, which implements an idea based on early Resonant Tunneling Diodes (RTDs) [45], can be explained using DC load line analysis. In the latch configuration (Figure 7), xGNR1 is connected to the reference voltage (V_{dd}) and acts as a pull-up device. The xGNR2 is connected to ground and acts as the pull-down device. The common terminal of the two devices is the state-node (SN) which stores the bit. The following terms will be used in the analysis –

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

Figure 8 depicts the operation of latching logic 1 onto the state node by injecting currents into the latch (I_{in}). Y-axis represents current flowing through the state-node and X-axis is the voltage of state node (V_{SN}). The solid line represents pull-down current and dashed line represents pull-up current. Assuming the state node is initially at 0, as the reference voltage V_{dd} is increased from 0, the operating point (shown by the dot X in Figure 8) is the intersection between pull-up and pull-down currents (satisfying Kirchoff's Current Law). Figure 8a shows the situation when the first pull-down current peak is encountered, called a decision point. As long as the pull-up current (I_{in} + I_{xGNR1})

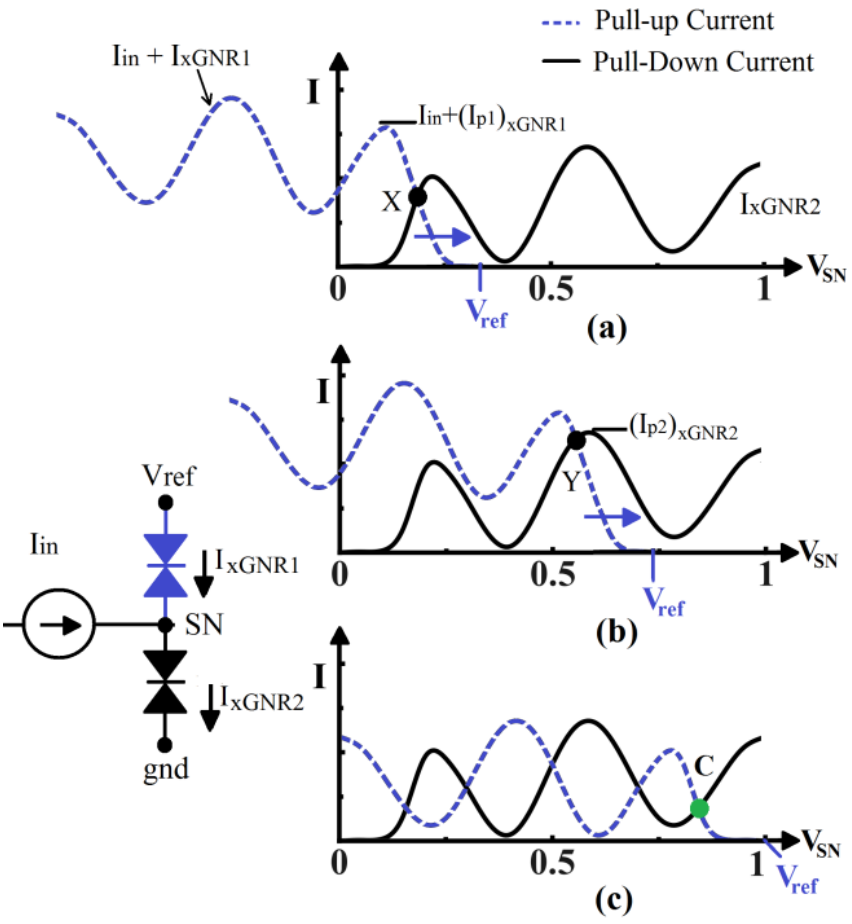


Figure 8. Load Line Analysis of xGNR latch when latching logic high. (a) & (b) Input logic high and V_{SN} at decision points, (c) Input switched off and Logic high latched.

is greater than pull-down current (I_{xGNR2}), the state node continues to shift from operating point X (Figure 8a), to point Y (Figure 8b) and finally to point C (Figure 8c) when V_{dd} reaches its maximum value. When the input current is switched off, the state node is latched to logic high. Hence to be able to latch the state-node to logic 1, the following condition should be met–

$$I_{in} + (I_{p1})_{xGNR1} > (I_{p2})_{xGNR2}$$

Figure 9 shows the process of latching logic 0 onto the state node. Consider the state-node is initially at 0 and the input is logic low. In this case, pull-down current (I_{ex}) exists at the state node. The analysis proceeds on the same lines as before. As long as the pull-

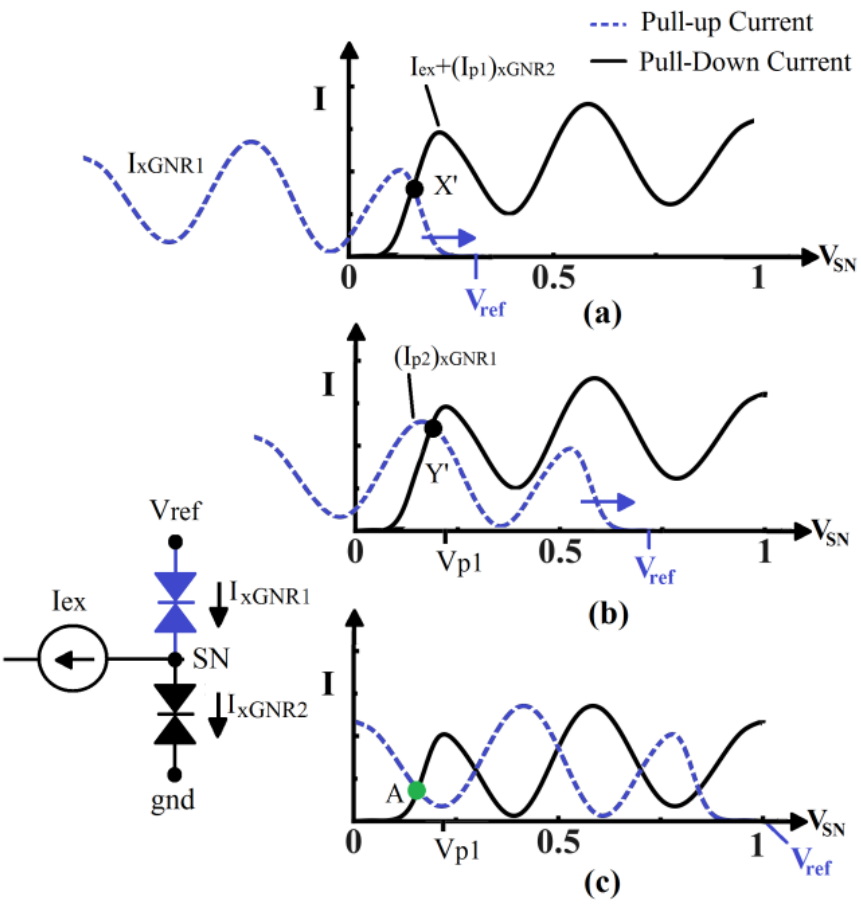


Figure 9. Load Line Analysis of xGNR latch when latching logic low - (a) & (b) Input logic low and V_{SN} at decision points, (c) Input switched off and Logic low latched.

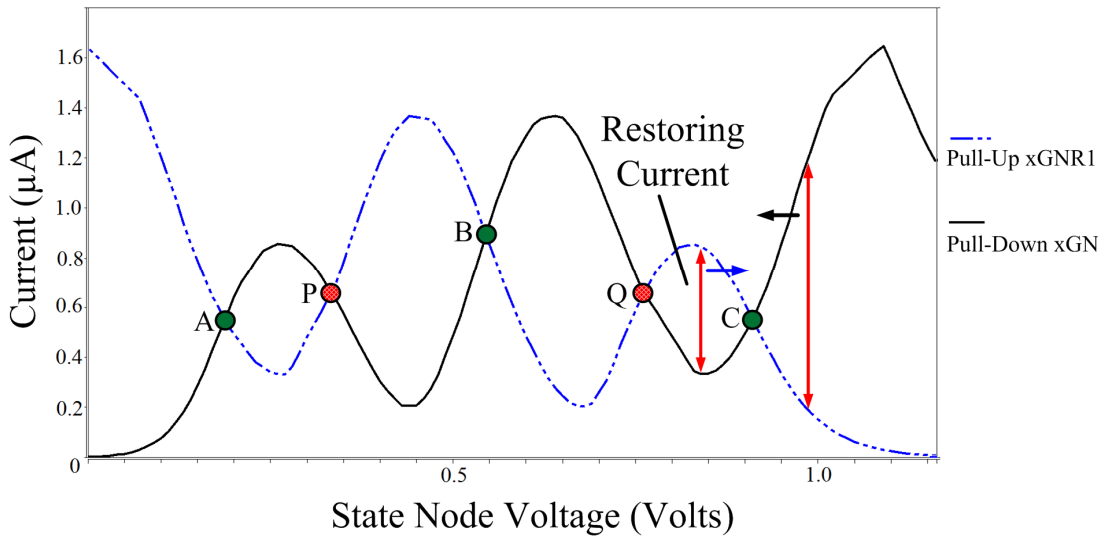


Figure 10. DC load line analysis of xGNR latch configuration showing stable states and restoring currents.

up current (I_{xGNR1}) is lower than pull-down currents (I_{ex} + I_{xGNR2}), the state node voltage (V_{SN}) never rises beyond V_{p1} (Figure 9b-c). After I_{ex} is switched-off, the state node remains at stable point A. Thus, to be able to latch logic 0, the following condition has to be satisfied –

$$(I_{p2})_{xGNR1} < I_{ex} + (I_{p1})_{xGNR2}$$

When used as a multi-state latch, the state node can be latched to the stable point B (in Figure 7) if the following condition is satisfied –

$$(I_{p2})_{xGNR2} > I_{in} + (I_{p1})_{xGNR1} > (I_{p1})_{xGNR2}$$

When the state node is at one of the stable points (A, B or C in Figure 10), any external disturbance that causes the state voltage to increase or decrease would be countered by strong restoring currents [7]. 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. Thus for correct latch operation at stable points, the following condition should be satisfied.

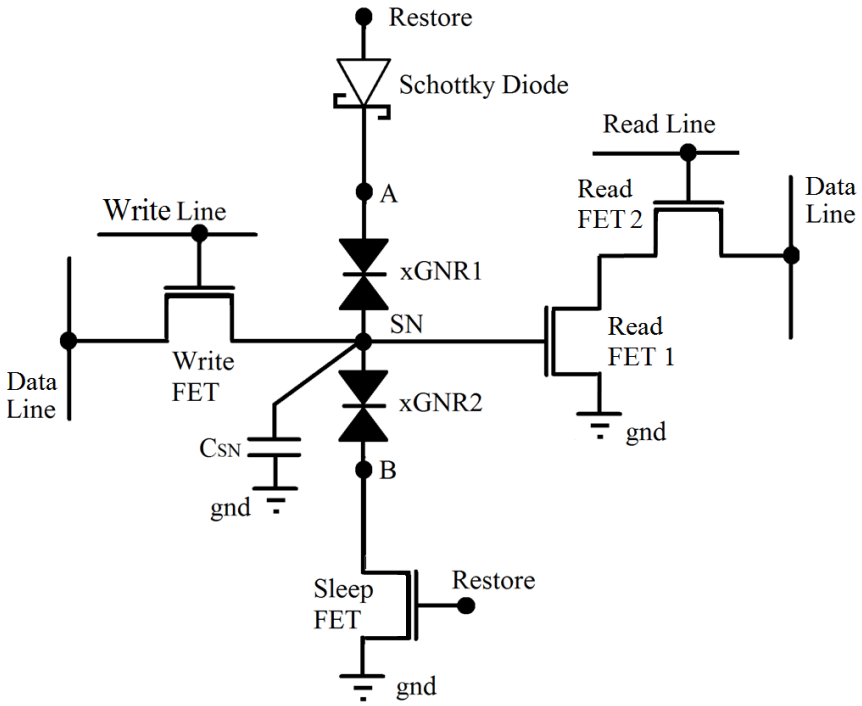
$$I_{noise} < I_{p1} - I_{v2} \text{ (worst case)}$$

States denoted by P and Q in Figure 10 are unstable and hence the corresponding voltages are the transition voltages. Consider state Q, any external noise would cause the state node voltage to transition to one of the surrounding states depending on the direction of the perturbation.

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. We now build on this concept to propose a volatile memory cell.

3.2 GNTRAM Cell Design

The xGNR latch can be used as the state holding element for volatile random access memory. Memory-cell selection, read and write operations can be performed using access transistors similar to the RAM cell proposed in [46]. However, a static



SN – State Node

Figure 11. Proposed GNTRAM Cell

implementation using this scheme would lead to large static currents and thus large stand-by power dissipation (in the order of μW).

We propose a dynamic memory cell to enable a low-leakage, low-power volatile xGNR based Tunneling RAM (GNTRAM). This design (Figure 11) uses two xGNR devices in a latch configuration and a write FET 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 GNTRAM cell can be used to realize a binary volatile memory by using two of the stable states to store information. All three stable states can also be used to compressively store more than 1 bit per-cell, thus realizing a ternary memory cell. The cell operation is explained next.

3.3 GNTRAM Cell Operation

3.3.1 Write Operation

A write operation is basically charging-up/discharging the state capacitance to the required voltage. Access to the state node for this operation is provided by the write FET. The gate terminal of the write FET is connected to the write-line and the drain terminal is connected to the data-line.

During a write operation, the required 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 denotes the state to be written. For binary memory,

logic 0 is represented by 0V and logic 1 is represented by 1V at the input. These input voltages correspond to the stable states A and C after the write operation is completed. When used as a ternary memory cell, the input voltages are in ternary representation (0V – logic 0, 0.6V – logic 1 and 1.0V – logic 2). These voltage values correspond to the voltages at which stable states A, B and C occur in the xGNR latch.

Consider that the state node is initially at logic 0. To write a particular logic value onto the cell, the appropriate input voltage (depending on binary or ternary representation) is applied on the data line (see Figure 12). The write signal is activated, 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 write-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

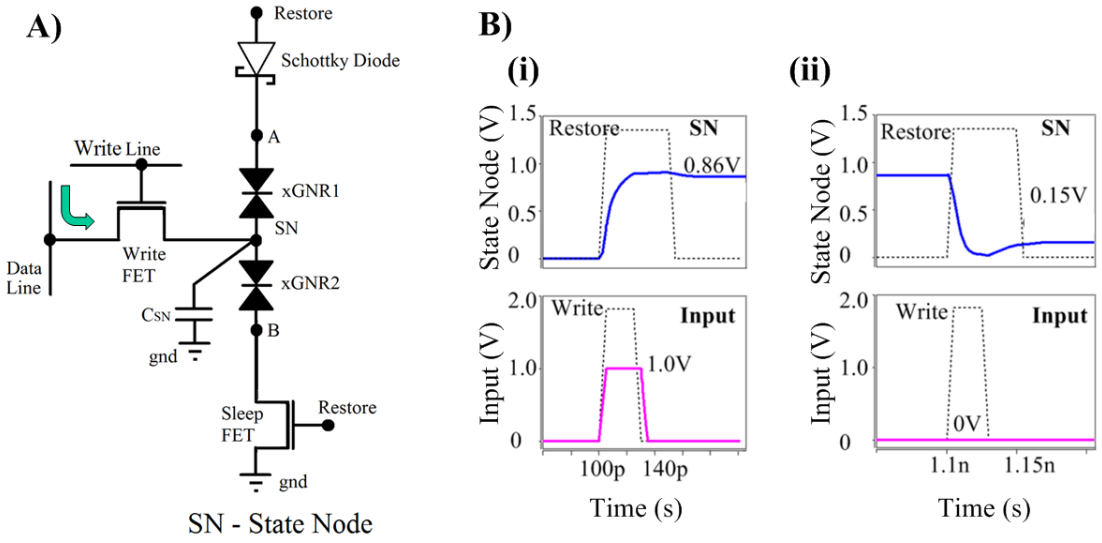


Figure 12. GNTRAM write operation: A) Circuit schematic showing the write path; and B) Voltage signals for writing (i) logic 1 (/logic 2) for binary (/ternary) and (ii) logic 0

restore signal is switched OFF and information is stored dynamically on the state capacitance.

When the state node is initially at a high voltage, a lower logic level can be written by simply applying the appropriate input voltage on the 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.

3.3.2 Read Operation

A pre-charge and evaluate scheme is used to read the stored information in the read path of the memory cell (see Figure 13). The output data line is connected to the drain of read FET1 and this node is pre-charged to VDD prior to a 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 read signal controls the gate of read FET2 and is used to select a particular memory cell for reading. The series stack of read FETs 1 and 2 acts as the evaluation path when the read signal is activated. The ON-

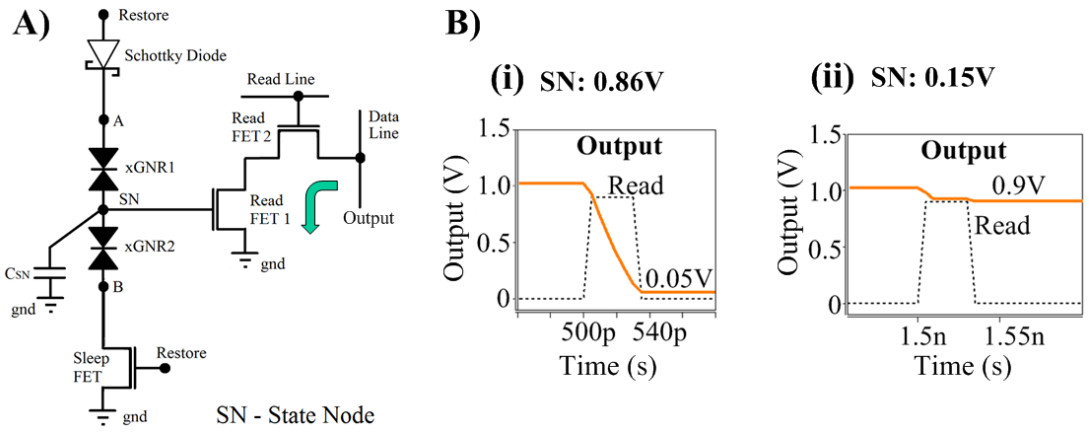


Figure 13. GNTRAM read operation: A) Circuit schematic showing read path; and B) Voltage signals during read operation for (i) logic 1 (/logic 2) for binary (/ternary) and (ii) logic 0

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 VDD and then the read signal pulse is applied for a pre-determined time. This read time is chosen such that when logic 1 is stored at the state node in the case of binary memory (logic 2 for the case of ternary memory), the data line is completely discharged and can be identified by 0V at the output. If a lower logic level is stored at the state node, it would cause read-FET1 to have a higher ON resistance. Thus applying the read signal would lead to the data line being only partially discharged to an intermediate value. 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.

3.3.3 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 dynamically 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 (see Figure 14). The restoring currents flowing through the state node charge-up the capacitor and restore its

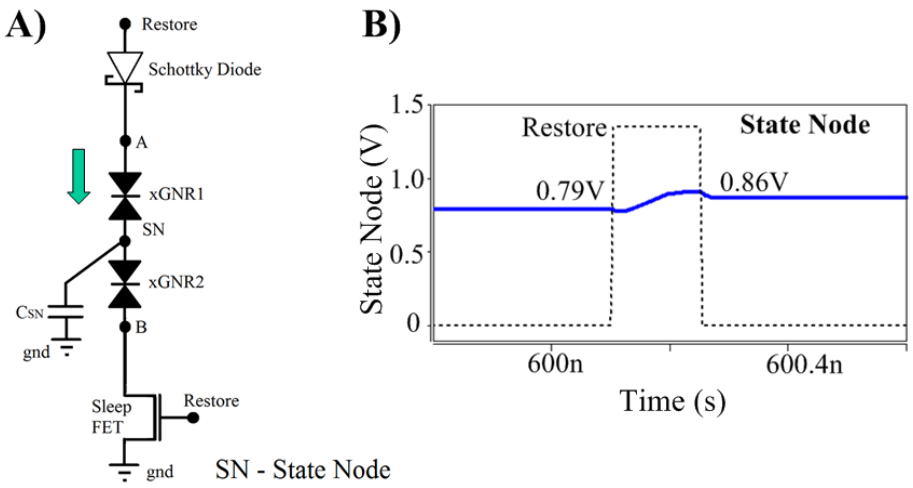


Figure 14. GNTRAM restore operation: A) Circuit schematic showing restore path; and B) Voltage signals during restore operation for logic 1 (/logic 2) for binary (/ternary) GNTRAM

value, as long as the noise/leakage currents are small enough to be countered. Unlike DRAM, the GNTRAM restore operation does not require a read followed by write to be able to restore the charge and is a relatively low-power operation.

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

3.4 Chapter Summary

We presented an application of the novel xGNR device as a memory element in this chapter. We proposed a new volatile memory cell called Graphene Nanoribbon Tunneling Random Access Memory (GNTRAM) which uses the NDR properties to realize multi-state memory. CMOS transistors were used for access and leakage power dissipation was mitigated by using a dynamic memory scheme with the help of a state capacitance. GNTRAM differs from conventional DRAM in two aspects – (i) the read

operation is non-destructive and (ii) the restore operation does not require a read operation and is independent. Binary and ternary GNTRAM implementation details are presented in the next chapter.

CHAPTER 4

BINARY AND TERNARY GNTRAM IMPLEMENTATION

A new volatile GNTRAM cell was proposed in the previous chapter. We present specific implementations of binary and ternary memory based on GNTRAM design in this chapter. We also propose a heterogeneous integration between graphene and CMOS technologies to physically realize GNTRAM. We leverage unique material interactions between graphene nanoribbons and metals to implement the required circuit functionality.

4.1 Binary GNTRAM Circuit Implementation

Binary GNTRAM utilizing two of the stable states can be realized based on the design proposed in the previous chapter, as shown in Figure 15. The access and sleep

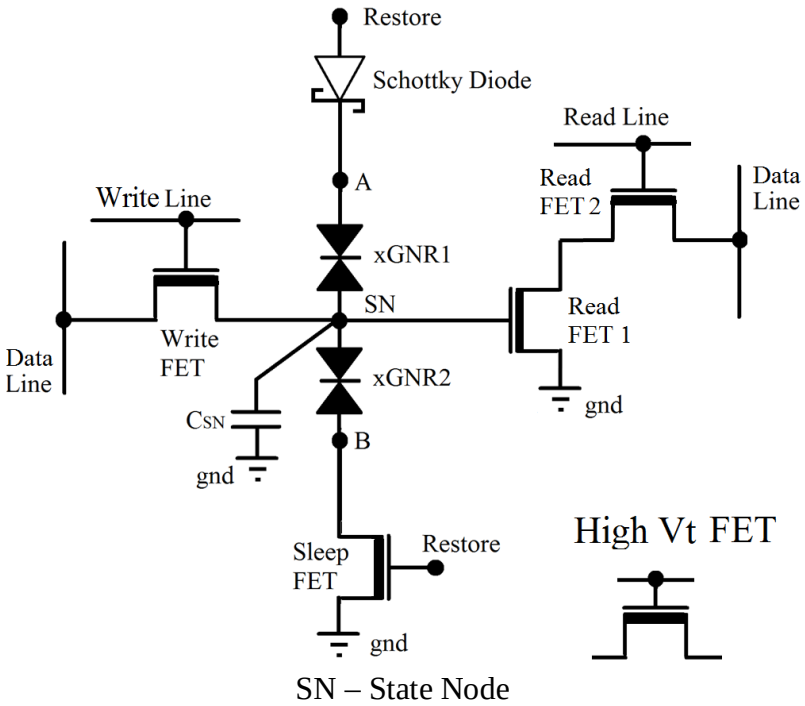


Figure 15. Binary GNTRAM Circuit Implementation

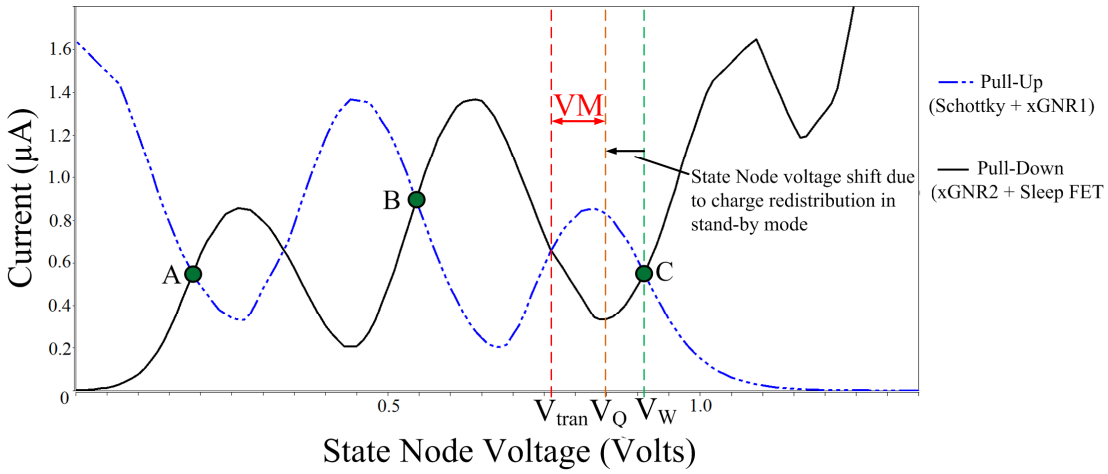


Figure 16. DC Load Line Analysis for xGNR latch including Schottky Diode and Sleep FET showing multiple stable states A, B and C.

FETs are implemented using uniform minimum-sized nMOS transistors. Since the write and sleep FETs are directly connected to the state node, they form leakage-critical paths. Thus to maximize the retention time, high-Vt nMOSFETs are used in this implementation. Alternate implementations with low-Vt devices are possible to improve performance. However, such implementations would suffer from high leakage and short retention time.

Since a dynamic implementation is used, a capacitor is required to retain the state information at the state node during stand-by. 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 Figure 16, for state C. 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 Figure 16), the restore operation causes a transition to the intermediate state B

instead of restoring state C. 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 Figure 16. 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

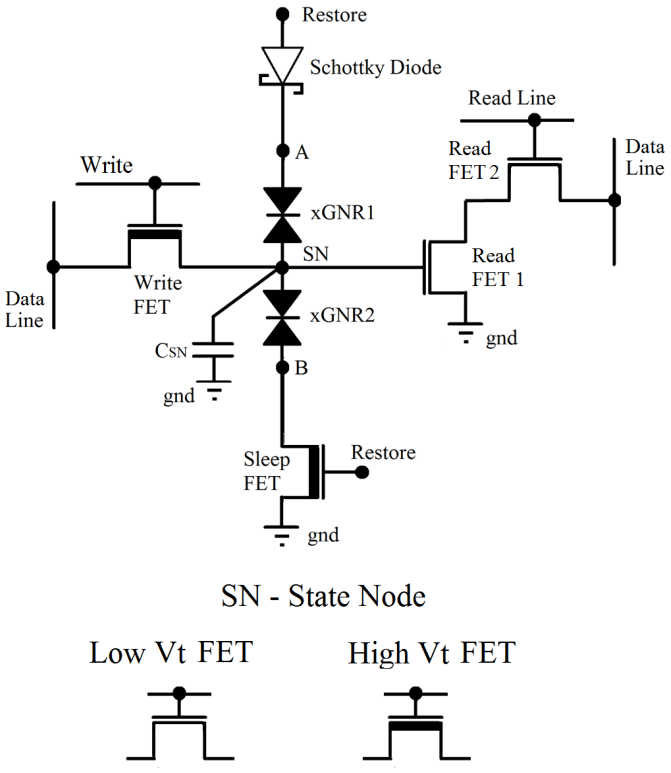


Figure 17. Ternary GNTRAM Circuit Implementation

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}.

The write FET can alternatively be implemented with a pMOSFET. This could be beneficial since a pMOS can easily pull-up the state node without any need of overdriving the gate voltage, 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. However, 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.

4.2 Ternary GNTRAM Circuit Implementation

Ternary GNTRAM can be realized as shown in Figure 17, which utilizes all of the stable states A, B and C. In order to distinguish between three stored voltage values, read FET1 necessarily needs to have a low Vt. Thus an asymmetric cell implementation is

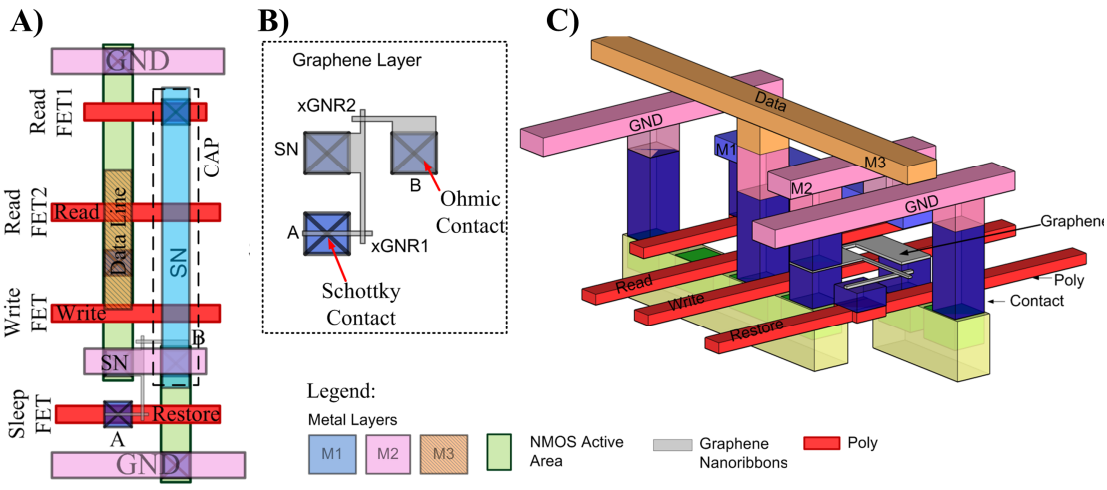


Figure 18. GNTRAM physical Implementation: A) Layout; B) Graphene layer showing schottky and ohmic contacts and the xGNRs; and C) Heterogeneous integration with CMOS.

used here as shown in Figure 17. The state capacitor is designed based on the discussion in the previous section.

4.3 Physical Implementation

We propose a cross-technology heterogeneous implementation between CMOS and graphene as shown in Figure 18. 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 MOSFET 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 [47][48]. 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 Figure 18B. The rest of the graphene-metal contacts are Ohmic to ensure proper operation and this is achieved by using wide GNRs [49]. 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 Figure 18A.

4.4 Chapter Summary

In this chapter, specific circuit implementations of binary and ternary GNTRAM were discussed. Binary GNTRAM was implemented with uniform high Vt transistors to minimize cell leakage. A performance-oriented design could potentially employ low Vt devices or a multi-Vt approach depending on the application. Trade-offs between retention time and performance need to be considered in such designs. An asymmetric cell approach was used for ternary GNTRAM in order to distinguish between the three stable states during read operation.

A novel physical implementation was presented by integrating CMOS transistors with graphene nanoribbon crossbar devices. Material interactions between the graphene nanoribbons and metals were leveraged to realize the Schottky diode. A lithography-friendly layout was used with uniform grid-based design and nMOSFETs. Alternative implementations are possible where some of the nMOSFETs are replaced with pMOSFETs. As graphene technology matures, CMOS transistors can even be replaced with graphene devices. Evaluations in terms of area, power and performance for the proposed designs are presented in the next chapter.

CHAPTER 5

GNTRAM BENCHMARKING

In this chapter, we present the simulation results for circuit validation and benchmarking methodology. Detailed evaluation in terms of area, power and performance is presented and compared to state-of-the-art 16nm CMOS SRAMs and 3T DRAM designs.

HSPICE was used to simulate and verify GNTRAM operation and for benchmarking against the state-of-the-art. 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 reverse bias leakage current through the Schottky diode was assumed to be 10pA [48], which is the same order of leakage currents in the high-Vt 16nm FETs. The value of the state capacitance was chosen to be 200aF for proper circuit behavior, based on the discussion in Chapter 4. A higher capacitance value would lead to a longer retention time.

5.1 xGNR HSPICE Device Model

A HSPICE behavioural model was developed for the xGNR device to conduct circuit simulation. The xGNR was modelled as a HSPICE sub-circuit using the structure shown in Figure 19 [7]. The DC I-V characteristics derived from the atomistic simulations (as explained in Chapter 2) was modelled using a voltage controlled current source (VCCS) with a piece-wise linear approximation between each I-V data point. The VCCS here is a two-terminal element and the current through it depends on the voltage difference across its terminals. The geometric capacitance at the GNR crossbar was modelled as a capacitor in parallel to take reactive currents into account in addition to DC response.

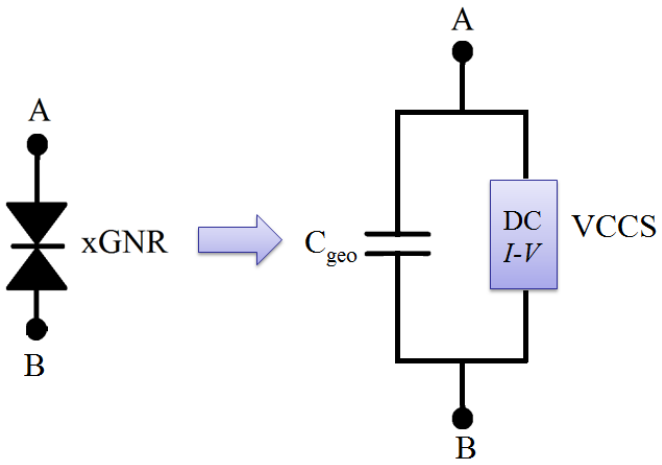


Figure 19. xGNR device modeled as a parallel configuration of its geometric capacitance and a Voltage Controlled Current Source (VCCS).

5.2 Circuit Validation using Simulation

Simulation was carried out using HSPICE for write, read and restore operations for both binary and ternary GNTRAM. Both the xGNR devices were assumed to be identical

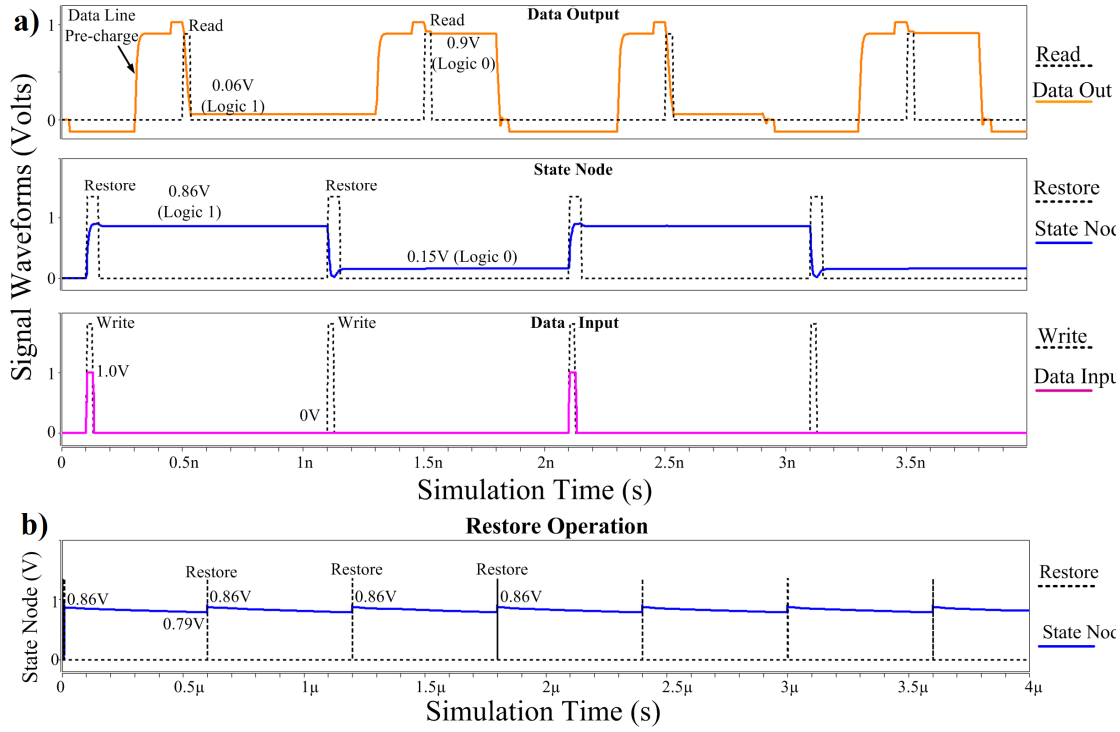


Figure 20. (a) Simulation waveforms showing binary GNTRAM read and write operations; and (b) Restore Operation for Logic 1.

for validation of the concept and circuit design. A more rigorous analysis considering variations between the devices and is beyond the scope of this thesis.

In the case of binary GNTRAM, the state node was initialized to logic 0 and logic 1 was first written and read. After this, logic 0 was written followed by a read operation. Logic 1 was written again and restore signal was applied at a period of 600ns to verify that logic 1 was being restored correctly. The simulation waveforms are shown in Figure 20. For the ternary GNTRAM, the state node was initialized to 0 and logic 1 was first written and then read. After this, all possible transitions between the three states were simulated and verified for both read and write operations (see Figure 21). Figure 21B shows the data output signals in detail. Restore operation is performed at a period of 0.7µs, as shown in Figure 21C for the case of restoring logic 2.

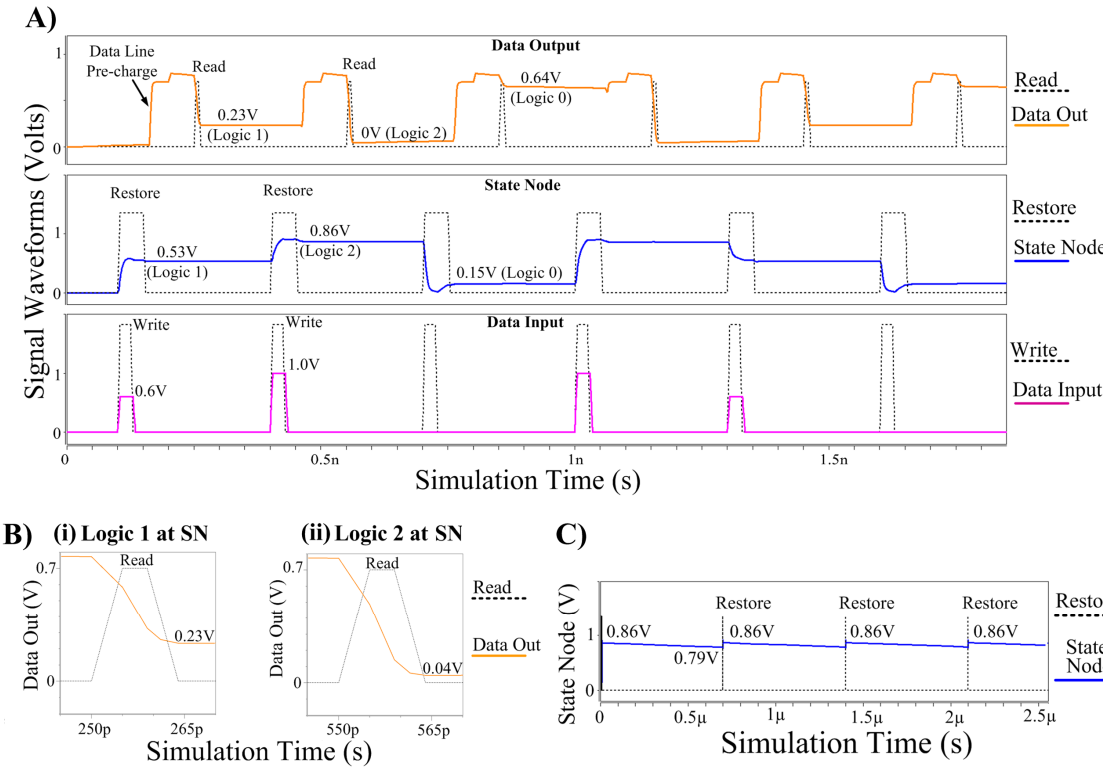


Figure 21. (a) Simulation waveforms showing ternary GNTRAM operation; (b) Read operation for (i) logic 1 and (ii) logic 2 at state node; and (c) Restore Operation for logic 2.

5.3 Benchmarking

Benchmarking was carried out in terms of area, power and performance against state-of-the-art 16nm CMOS SRAMs and 3T DRAM. For physical layout design and evaluation, 1-D Gridded design rules [50] (see Table I) were used to compare the area of GNTRAM cell with Gridded 8T SRAM cell [51] in 16nm technology 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 [52]. This methodology resulted in a range of values for 6T SRAM cell area for a range of design rules. PTM RC models [53] based on scaled interconnect dimensions and 16nm PTM transistor models [53] were used for simulation with HSPICE for power and performance evaluation of 16nm CMOS 6T SRAM and Gridded 8T SRAM.

3T DRAM was also investigated for benchmarking since it is a potential candidate for on-chip caches in advanced technology nodes [54][55]. 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 Figure 22. 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. Design Rules

1D Gridded Design [50]	***M1, M2 Interconnect***	***Poly***
Pitch (16nm technology node)	40~60 nm	60~80nm

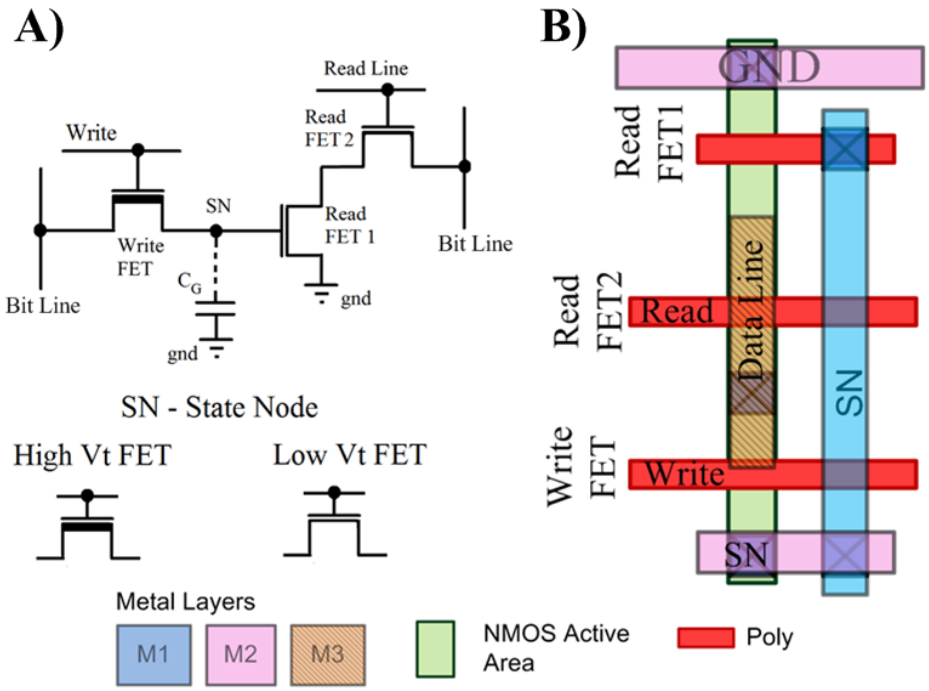


Figure 22. 3T DRAM – (a) Circuit Schematic, and (b) Physical Layout

5.3.1 Binary GNTRAM Evaluation

In this section, we provide our evaluation results for binary GNTRAM in terms of area, power and performance. Table II shows the evaluation results.

A. Area Evaluation

Binary GNTRAM showed a density advantage of up to 1.1x over 16nm CMOS 8T Gridded SRAM, and has comparable area to 16nm regular 6T SRAM cell. The area overhead in GNTRAM is due to routing and state capacitance.

B. Power Evaluation

Active power dissipation of binary GNTRAM was up to 1.23x lower than regular 6T CMOS SRAM and up to 1.48x lower than Gridded 8T CMOS SRAM cell. When compared to CMOS 3T DRAM, binary GNTRAM was up to 2.17x more power efficient during active periods. In terms of stand-by power dissipation, binary GNTRAM was

3.68x lower than CMOS SRAMs. This leakage power benefit is due to dynamic state retention scheme rather than using static currents to retain stored state.

C. Performance Evaluation

In terms of performance, the write operation for binary GNTRAM was faster than that of SRAM mainly because the write transistor operates at a higher than nominal voltage. The read time of binary GNTRAM suffers due to three reasons–

- The read FET operates at lower than nominal voltage since the state node stable point for logic high is 0.86V.
- The bit line capacitance is relatively higher for GNTRAM due to larger cell height.
- GNTRAM uses minimum sized transistors.

D. Trade-off Analysis: State Capacitance vs. Write Time and Retention Time

A study was conducted to investigate the effect of increasing the state capacitance on retention time of the GNTRAM. The trade-off with write performance was also analyzed. It was observed that as the state capacitance was increased, there was orders of

Table II. Binary GNTRAM Benchmarking

		GNTRAM Cell	***16nm CMOS 6T SRAM Cell (LP)***	***16nm CMOS Gridded 8T SRAM Cell (LP)***	***16nm 3-T DRAM Cell***
Area Comparison (μm^2)		0.03 – 0.0608	0.026 – 0.064	0.0336 – 0.0672	0.0264 – 0.054
Power Comparison	Active Power (μW)	0.98 – 0.99	1.16 – 1.21	1.45 – 1.47	2.12 – 2.15
	Stand-by Power (pW)	31.3 – 34	124.18 – 125.12	78.38 – 78.44	6.49 – 7.01
Performance	Read Operation (ps)	24.39 – 27.32	17.39 – 21.03	14.82 – 16.08	9.18 – 9.68
	Write Operation (ps)	16.5 – 16.84	67.27 – 67.54	58.37 – 63.18	10.45 – 10.97

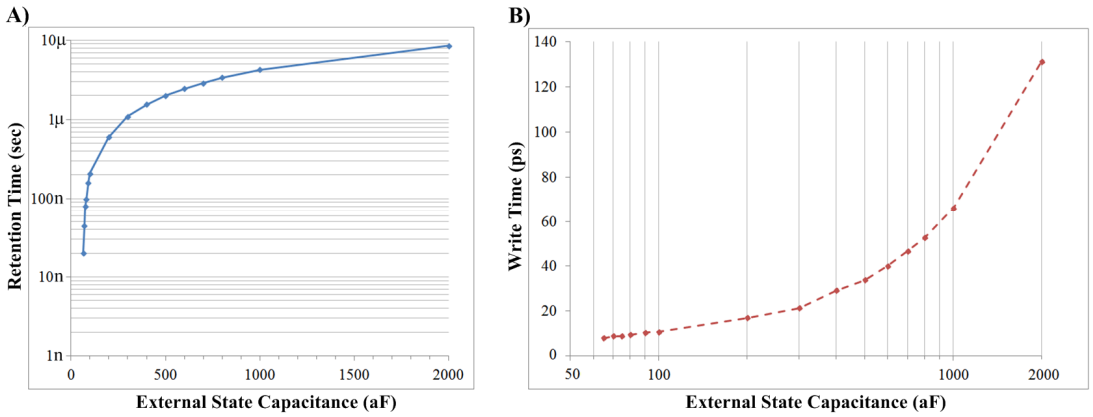


Figure 23. Trade-off Analysis: A) External state capacitance vs. Retention time; B) External state capacitance vs. write time.

magnitude improvement in retention time with only a linear impact on the write time. As the state capacitance is increased further beyond 400aF, the improvement in retention time was only linear while the write time increased steeply as shown in Figure 23A and B.

5.3.2 Ternary GNTRAM Evaluation

In this section, we present our evaluation results for ternary GNTRAM in terms of area, power and performance. Table III shows the evaluation results. 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.

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

Table III. Ternary GNTRAM Benchmarking

		Ternary GNTRAM (Per Cell, 1.585 bits)	***Ternary GNTRAM (Per Bit)***	***16nm CMOS 6T SRAM Cell (High Performance)***	***16nm CMOS Gridded 8T SRAM Cell (High Performance)***
Area Comparison (μm^2)		0.03 – 0.0608	0.019 – 0.038	0.026 – 0.064	0.0336 – 0.0672
Power Comparison	Active Power (μW)	2.05 – 2.15	1.29 – 1.35	2.1 – 2.2	2.38 – 2.44
	Stand-by Power (pW)	22.04 – 22.07	13.9 – 13.92	6152 – 6157	15552 – 15556
Performance	Read Operation (ps)	8.98 – 9.8		8.35 – 9.25	7.68 – 7.96
	Write Operation (ps)	16.26 – 16.39		18.44 – 18.46	16.62 – 19.16

		Ternary GNTRAM (Per Cell, 1.585 bits)	***Ternary GNTRAM (Per Bit)***	***16nm CMOS 6T SRAM Cell (Low Power)***	***16nm CMOS Gridded 8T SRAM Cell (Low Power)***	***16nm 3-T DRAM Cell***
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 Compari-son	Active Power (μW)	2.05 – 2.15	1.29 – 1.35	1.21 – 1.16	1.45 – 1.47	2.12 – 2.15
	Stand-by Power (pW)	22.04 – 22.07	13.9 – 13.92	124.18 – 125.12	78.38 – 78.44	6.49 – 7.01
Perform-ance	Read Operation (ps)	8.98 – 9.8		17.39 – 21.03	14.82 – 16.08	9.18 – 9.68
	Write Operation (ps)	16.26 – 16.39		67.27 – 67.54	58.37 – 63.18	10.45 – 10.97

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.

During stand-by mode, ternary GNTRAM was up to 1196x more power efficient in terms of leakage power when compared to high performance CMOS SRAMs. It was also 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 ternary GNTRAM 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.

5.4 Chapter Summary

In this chapter, GNTRAM evaluation methodology and benchmarking in terms of area, power and performance were presented against state-of-the-art CMOS SRAMs and 3T DRAM. Binary GNTRAM showed up to 10% density benefit over 16nm CMOS SRAMs and was up to 3.68x more power-efficient during stand-by mode. The overhead in the area of binary GNTRAM is attributed to access MOSFETs and routing requirements. For the ternary GNTRAM, as more bits are stored (1.5 bits) in one cell, these costs are amortized. Thus we see higher benefits as expected with up to 1.77x better density compared to 16nm CMOS SRAMs and up to 1196x lower stand-by power compared to high performance CMOS SRAMs, while maintaining comparable performance. Hence, GNTRAM has the potential to overcome the physical scaling limitations of CMOS by storing more than 1 bit in a given cell. The next chapter explores possible approaches to enable scaling of the number of bits that can be stored per cell, to further enhance GNTRAM benefits.

CHAPTER 6

SCALING APPROACHES – QUATERNARY GNTRAM

Previously, binary and ternary GNTRAM cell designs were introduced based on a novel xGNR tunneling device. It was shown that both designs had density and power benefits vs. 16nm CMOS SRAM and 3T DRAM designs. Ternary GNTRAM offered the ability to store multiple data bits in a single cell (1.5 bits per cell), thereby improving the density as compared to CMOS. This also amortized the leakage per cell over multiple bits resulting in reduced stand-by power consumption. In order to further enhance density and leakage benefits, there is a need for an approach to scale further by storing more bits in a single cell. This could provide a new dimension for scaling as an alternative to relying on physical scaling for enhancing benefits.

The key requirement to allow scaling is to increase the number of stable states at the state node of the xGNR latch, by increasing the number of current peaks in the pull-up and pull-down devices. In this chapter, we explore two possible approaches based on circuit techniques and device engineering to meet this requirement. We use these approaches to realize quaternary GNTRAM with 4 stable states, thus allowing for 2 bits being compressively stored in a single cell. We will also investigate the trade-offs with these scaling approaches and benchmark the quaternary GNTRAM designs against 16nm CMOS SRAMs and 3T DRAM.

The first approach is a circuit-level technique based on a concept similar to RTDs [56]. By increasing the number of xGNR NDR devices in each leg of the latch, the I-V characteristics of such a configuration will exhibit more current peaks over an extended

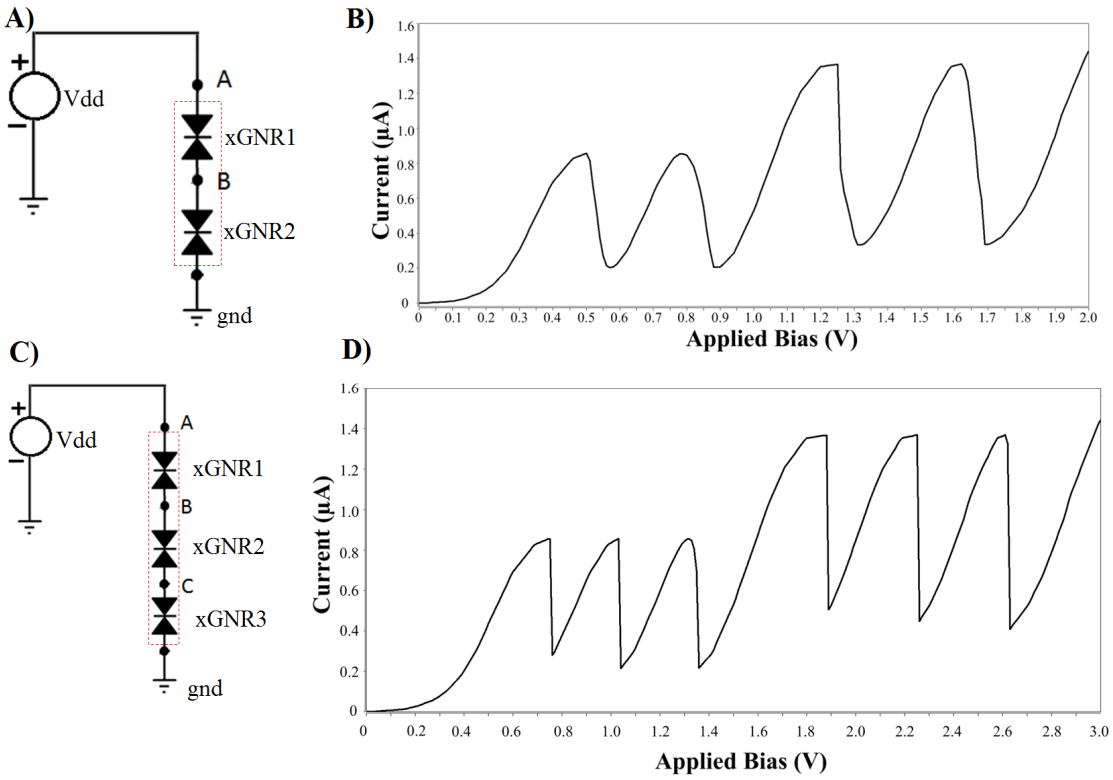


Figure 24. Circuit technique to increase number of current peaks: A) 2 xGNRs in series; B) DC load line analysis showing 4 current peaks for configuration in (A); C) 3 xGNRs in series; and D) DC load line analysis showing 6 current peaks for configuration in (B).

voltage range. The second approach relies on altering the length of the GNR stub of the xGNR device to achieve more current peaks in the I-V curve.

6.1 Approach 1 – Circuit technique to increase number of states

In the case of ternary xGNR latch, both pull-up and pull-down devices exhibited 2 current peaks and valleys in their I-V characteristics which led to 3 stable states. In general, a latch configuration with devices having *'N'* current peaks would exhibit *'N + 1'* stable states. Thus to realize a quaternary xGNR latch, the devices in both legs of the latch would require at least 3 current peaks in their I-V characteristics.

A series configuration of *'N'* xGNR devices exhibits *'2N'* current peaks. As shown in Figure 24, a series combination of 2 xGNRs leads to 4 current peaks when the voltage

across the combination is increased to about 2V. Similarly, 3 xGNRs in series lead to 6 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.

Thus, arranging 2 such series xGNRs in each leg of an xGNR latch 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 to build a quaternary latch, as shown in Figure 25. The latch

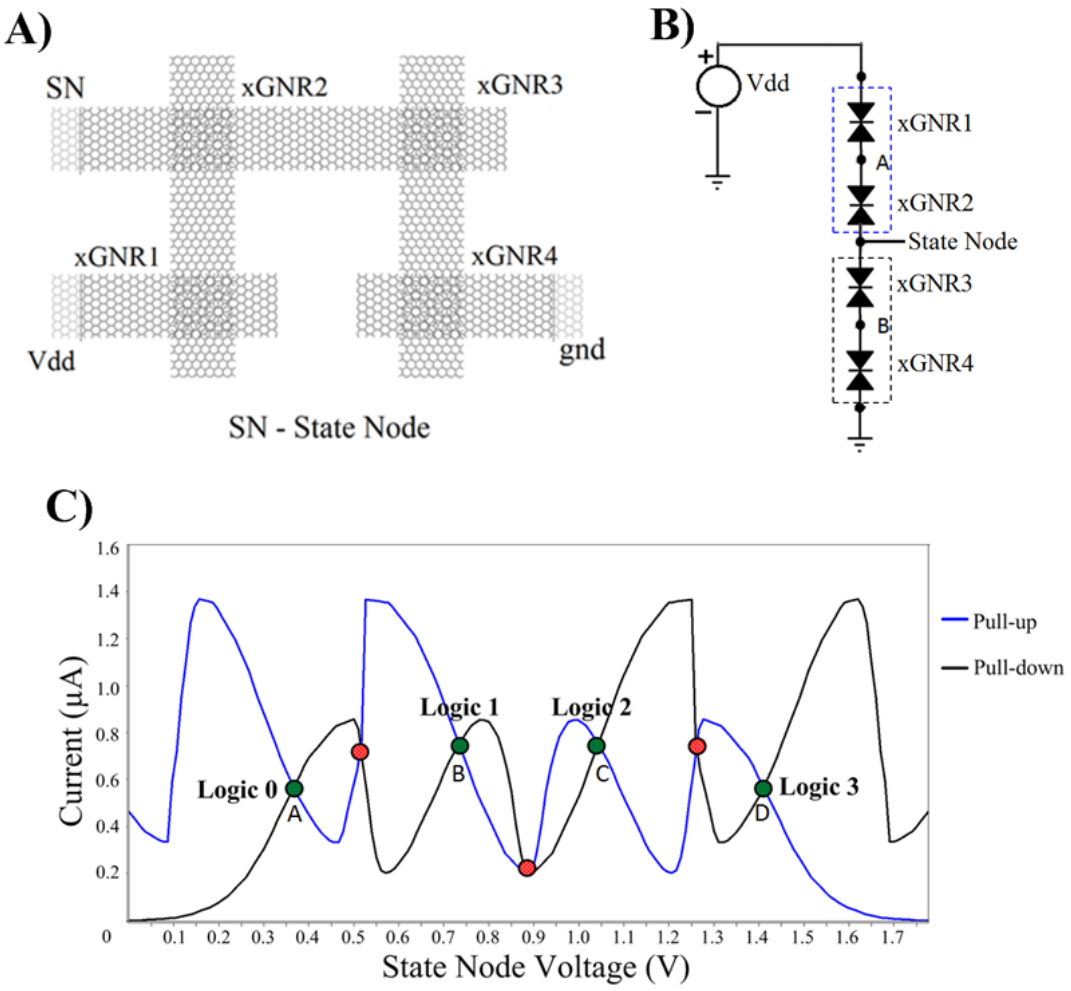


Figure 25. A) Quaternary cross-Graphene Nanoribbon (xGNR) tunneling latch; (B)Circuit schematic; and C) DC Load Line Analysis showing 4 stable states.

operation is the same as outlined in Chapter 3, Section 3.1.

6.1.1 Quaternary GNTRAM

As shown in the previous section, an xGNR latch with 2 series xGNR devices in each leg can realize a quaternary latch, and this is used to build quaternary GNTRAM. Such a design will enable storing 2 bits in a single memory cell and resulting in a higher memory density than CMOS designs that store 1 bit per cell.

Similar to previous GNTRAM designs, a dynamic memory cell implementation is adopted for low-leakage, low-power quaternary GNTRAM. This design (Figure 26) uses the quaternary xGNR latch as the state holding element and a write FET to access the state node. To mitigate static power, the xGNR latch is switched OFF during stand-by and a capacitor (C_{SN}) is used at the state node to store the voltage value written into the

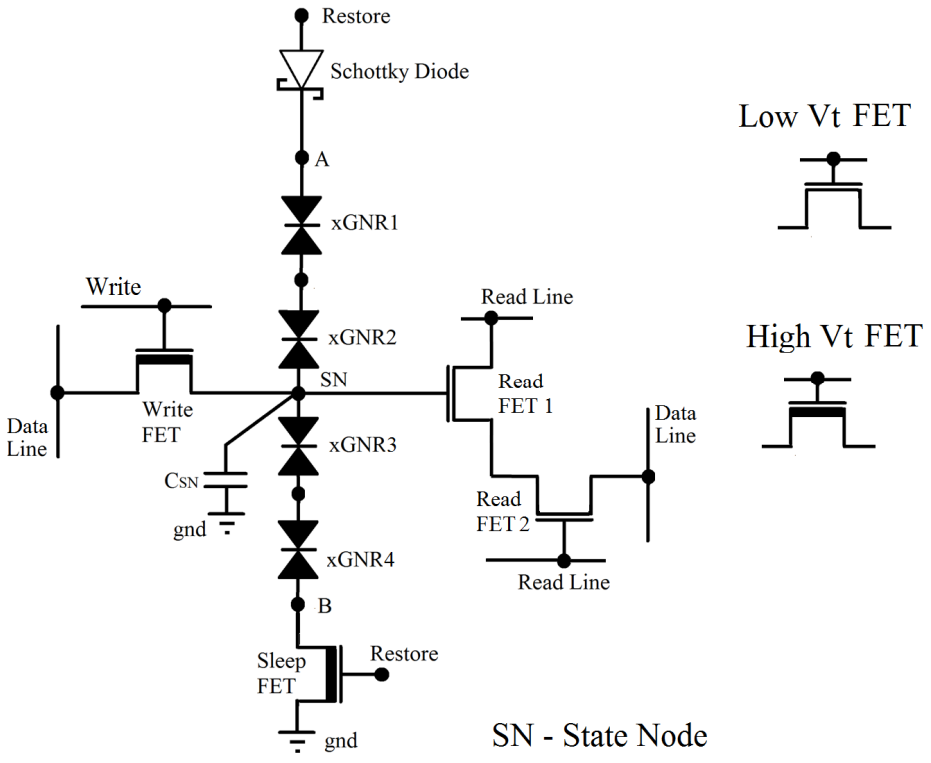


Figure 26. Proposed quaternary GNTRAM cell.

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. The write FET and sleep FET form leakage-critical paths and hence are implemented with high-Vt devices. Two read FETs are used to read the stored information. In order to distinguish between the stored states, low-Vt devices are used in the read path.

6.1.2 Quaternary GNTRAM Operation

A) Write Operation:

Similar to previous GNTRAM designs, the write operation is basically 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 data-line, with the state node at source. During a write operation, the required cell is selected by activating the corresponding write-line and applying the required input voltage onto the data-line. For quaternary memory, the input voltages are in quaternary representation (0V – logic 0, 0.7V – logic 1, 1.1V – logic 2 and 1.5V – logic 3). These voltage values correspond to the voltages at which stable states A, B, C

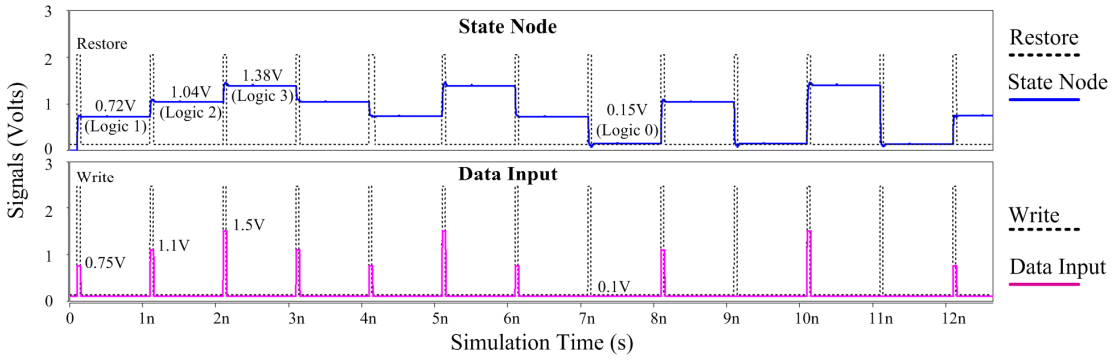


Figure 27. Quaternary GNTRAM write operation.

and D occur in the xGNR latch characteristics (see Figure 25C). Figure 27 shows the write operations for all possible state transitions in the quaternary GNTRAM cell.

B) Read Operation:

A pre-discharge and evaluate scheme is used to read the stored information in the memory cell (see Figure 28). The pull-down scheme (shown before for binary and ternary GNTRAM) was not used here because it did not result in significant margins for distinguishing between logic 2 and logic 3. However, the Vt of the read FET1 may be tuned to achieve better read margin to distinguish the stored states in order to use a pull-down read approach.

In the quaternary GNTRAM design, the drain of read FET1 and gate of read FET2

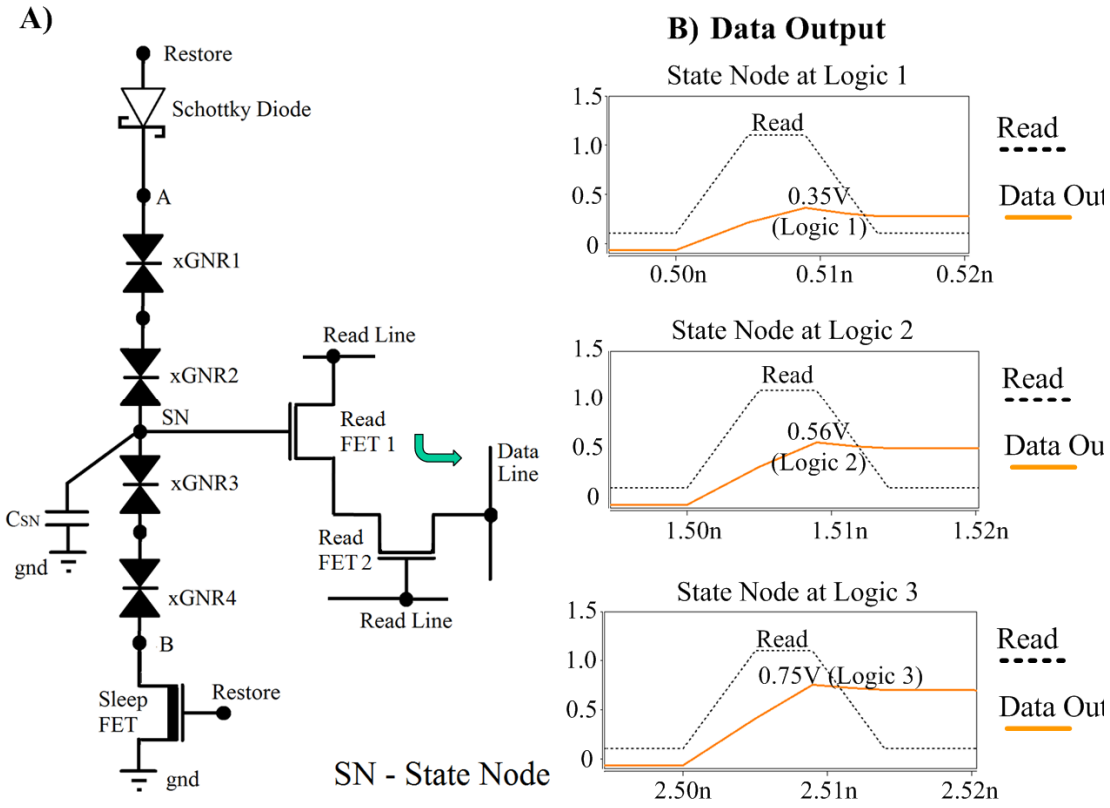


Figure 28. Quaternary GNTRAM read operation: A) Circuit schematic showing read path; B) Data output signals for reading different stored states.

are connected to the READ signal. The output data line is connected to the source of read FET2 and this node is pre-discharged prior to a 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 series stack of read FETs 1 and 2 acts as the evaluation path; when the read signal is activated the output data line is pulled up based on the stored state. The value of the state node voltage at the gate of read FET1 limits the voltage to which the output can be pulled-up due to the intrinsic threshold voltage drop in the nMOS. Thus the final output voltage is specific to a stored state which enables the detection of multiple voltage levels at the data output. This pull-up read scheme is also applicable to the binary and ternary GNTRAM designs.

To initiate a read operation, the data line is discharged and then the read signal is applied. When logic 0 is stored, the read-FET1 is completely switched OFF and the data line remains at low voltage. For all other stored logic states, the output is pulled up to their corresponding voltage levels and hence this scheme results in a non-inverting read-out.

C) Restore Operation:

The stored charge on the state capacitance starts to leak during stand-by mode and needs to be replenished. This is done by simple asserting the restore signal within the stipulated time, similar to the approach in binary and ternary GNTRAM. However, the retention period for the quaternary GNTRAM was found to be low (in the order of a few ns) due to higher leakage due to the relatively higher voltage operation. This calls for leakage mitigation techniques to improve the retention period.

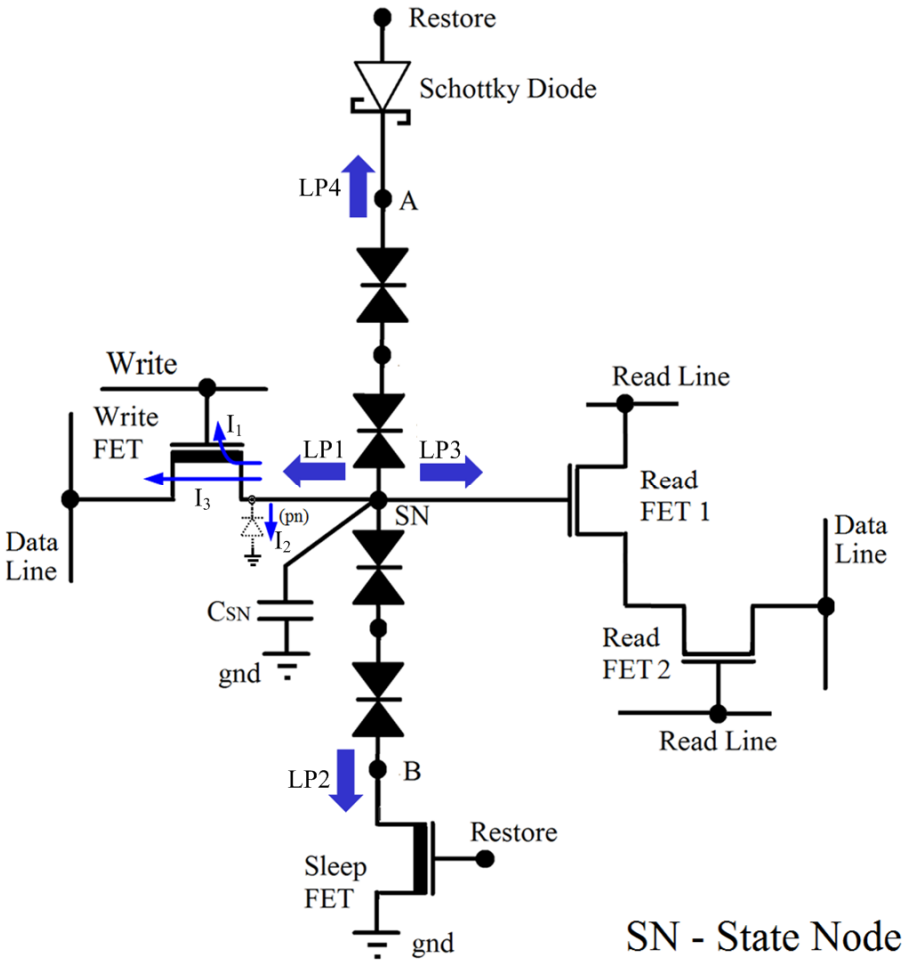


Figure 29. Leakage paths in quaternary GNTRAM.

6.1.3 Leakage Analysis and Mitigation

Due to relatively higher voltage operation, the leakage in the control FETs is exacerbated. During stand-by, the 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 Figure 29) shows that the write FET and Sleep FET form critical paths (LP1 and LP2) since they are directly connected 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). However, it was found that for the 16nm LP PTM devices

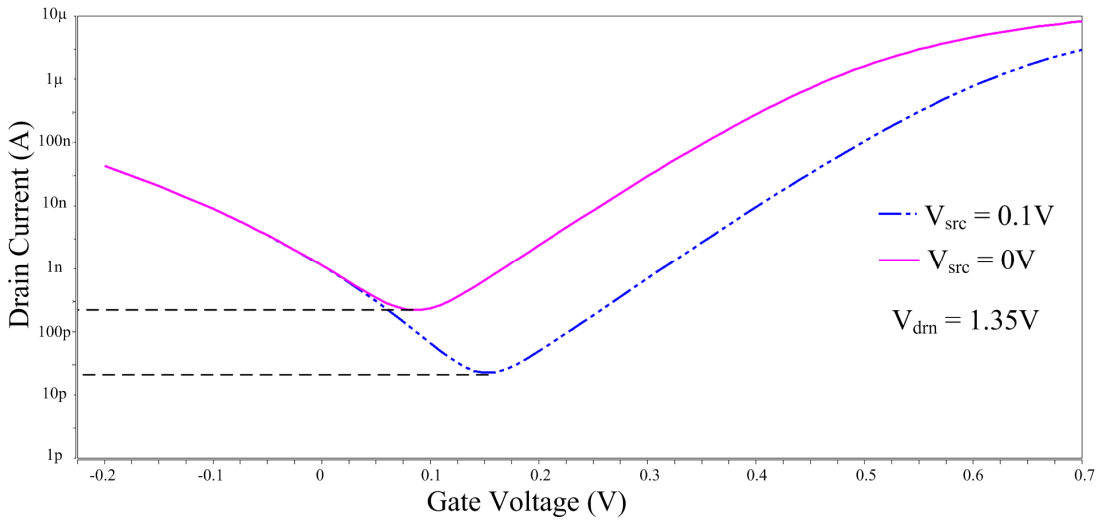


Figure 30. Sub-threshold leakage analysis in write FET when logic 3 is stored at state node.

used, the gate tunneling current and junction leakage were negligible and the leakage current was dominated by sub-threshold channel leakage.

One of the frequently used circuit techniques in literature to reduce the OFF-state sub-threshold channel leakage is source/gate biasing during stand-by [54]. This scheme is most effective in curbing the sub-threshold leakage compared to other techniques such as body biasing or VDS reduction. The sub-threshold current analysis of the devices shows that when the source is offset by 0.1V during stand-by, the leakage current can be reduced by almost 10x when storing logic state 3 (see Figure 30). Thus the data-line and the source terminal of the sleep FET are maintained at 0.1V during stand-by mode. This can be achieved either by using a self-biasing scheme with a shared carefully-sized nMOS transistor in series [54] or by selecting a separate voltage source similar to the approach in reference [57].

The remaining leakage sources are the gate leakage current through read FET1 (LP3 in Figure 29) and the reverse-bias leakage of the Schottky diode (LP4 in Figure 29). The gate leakage can be reduced by increasing the oxide thickness for 16nm HP PTM device

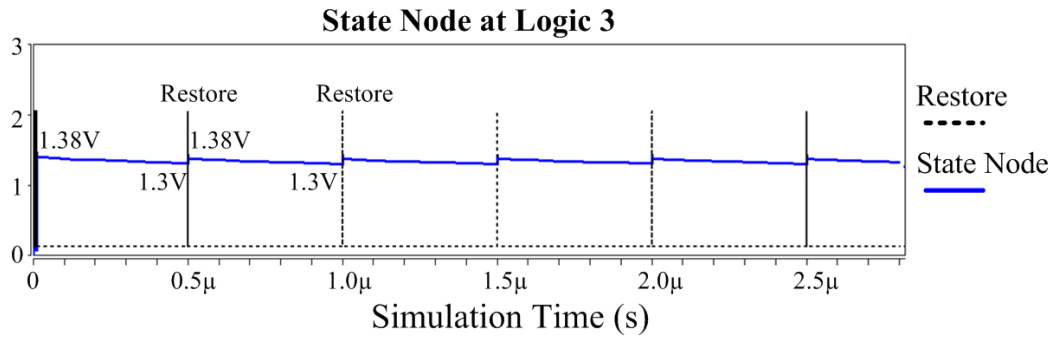


Figure 31. Restore operation when logic 3 is stored at state node.

(Vth0 was recalculated using the equation for retro-grade doping CMOS [58]). The reverse-bias leakage through the Schottky diode is assumed to be constant at 10pA. These leakage reduction techniques enhanced data retention period to 500ns as shown in Figure 31. A larger state capacitor would lead to a longer retention period.

6.1.4 Physical Implementation

A heterogeneous integration between CMOS and graphene was followed to physically realize the quaternary GNTRAM, similar to the binary and ternary versions. This is shown in Figure 32. The graphene layer now contains 4 xGNR devices as shown

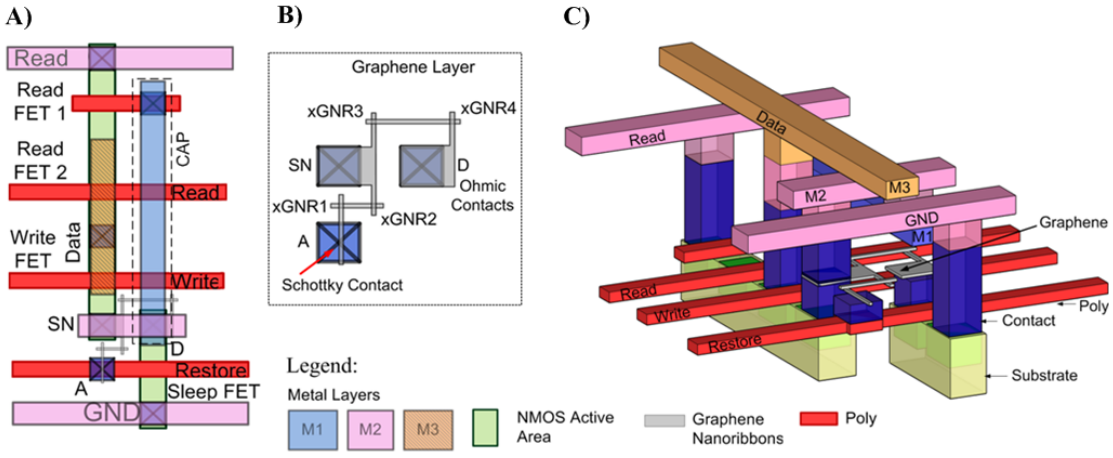


Figure 32. A) Quaternary GNTRAM approach 1— physical layout; B) Graphene layer showing xGNR devices, and Schottky and Ohmic contacts; and C) Heterogeneous integration with CMOS.

in Figure 32B. Since the xGNR devices are quite small relative to the FETs, no additional area requirement is considered in this case.

6.1.5 Benchmarking vs. 16nm CMOS

To understand the benefits of this scaling approach, the quaternary GNTRAM was extensively benchmarked against 16nm CMOS SRAM 6T and 8T gridded designs and 3T DRAM. HSPICE was used to evaluate the performance and power with 16nm PTM MOSFET and RC interconnect models. Similar to binary and ternary GNTRAM

Table IV. Quaternary GNTRAM Approach I 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 (μm^2)		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 (μm^2)		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

evaluation, the piecewise linear VCCS behavioral model was used for the xGNR device. 16nm grid-based design rules were used to evaluate the area (see Table I). The benchmarking results are shown in Table IV.

As expected, the quaternary GNTRAM shows higher density benefits by storing 2 bits per cell, with up to 2.27x benefit against 16nm CMOS SRAM and 1.8x vs. 3T DRAM. The read performance of the quaternary GNTRAM was comparable to high-performance CMOS SRAMs and 3T DRAM. Although the quaternary GNTRAM is power-efficient (up to 1.32x during active period and 818x during stand-by against high performance SRAM), the relative benefit is lower than that for the ternary GNTRAM. This is because of higher operating voltage required for quaternary GNTRAM.

The need for higher operating voltages with further scaling using this approach may eventually limit the number of bits that can be stored in a single cell. High operating voltages may also impact the reliability of the MOSFETs. Hence we explore an alternative scaling approach in the following section where a lower operating voltage can potentially be used.

6.2 Approach 2 – xGNR Device Engineering

In the xNGR device study conducted by Prof. Lake's group at UCR, it was found that the number of current peaks in the xGNR I-V characteristics is a strong function of the xGNR stub length L_S (see Figure 33A) [29]. By increasing the length of the stub, the number of current peaks increases. For an xGNR with stub length L_S = 2.5nm, 2 current peaks were observed for 1V applied bias as in Figure 33B. When the stub length is increased to L_S = 9.3nm, the xGNR device exhibits 6 current peaks for 1V applied bias (Figure 33C). This approach thus enables increasing the number of states per cell without

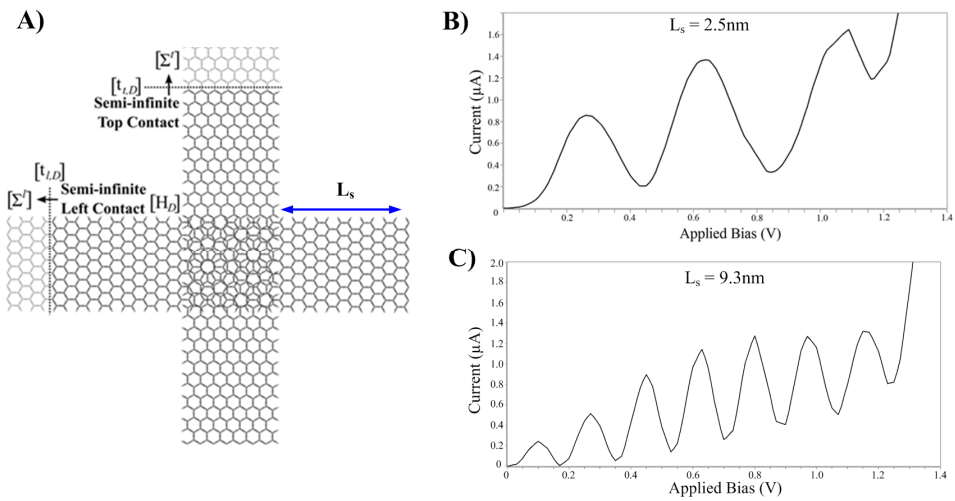


Figure 33. xGNR device engineering to increase the number of current peaks: A) xGNR device structure; B) I-V characteristics showing 2 current peaks between 0-1V for 2.5nm stub length (Ls); and C) I-V characteristics showing 6 current peaks between 0-1V for 9.3nm stub length.

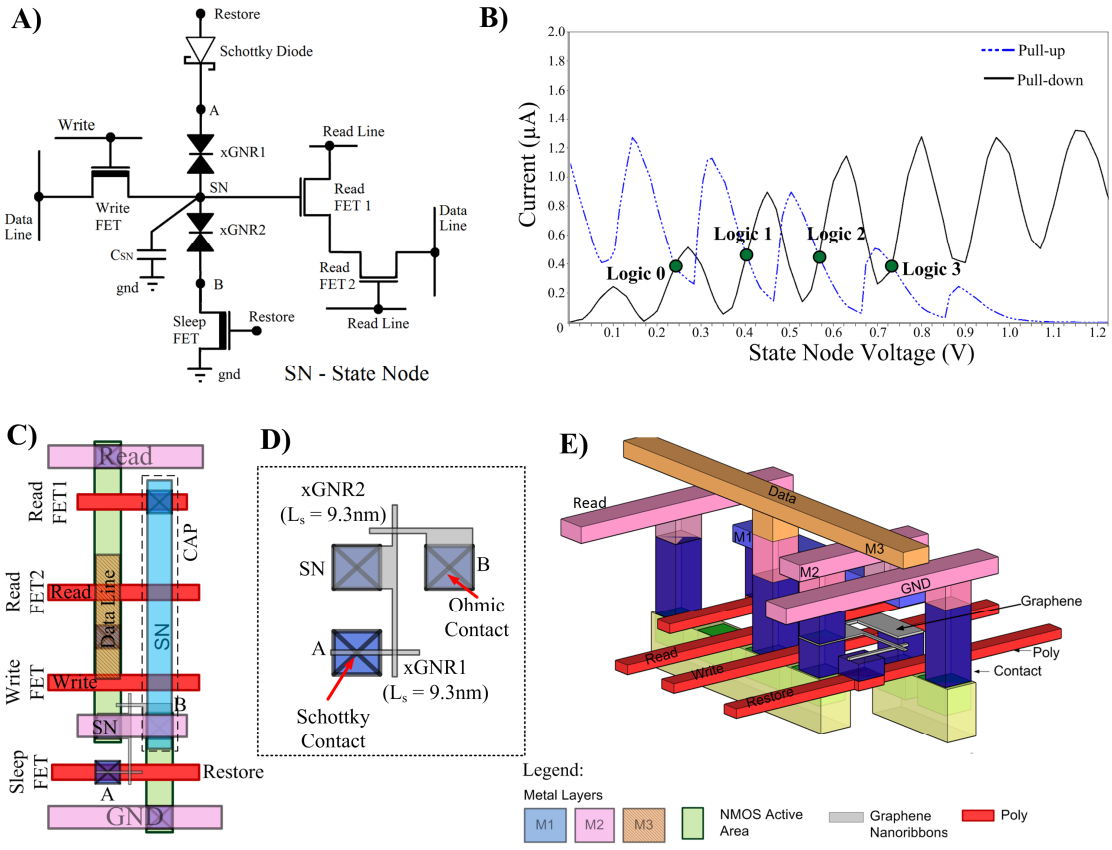


Figure 34. A) Quaternary GNTRAM approach 2; B) DC load line analysis showing 4 stable states; C) Physical layout; D) Graphene layer showing xGNR devices and contacts; and E) Heterogeneous integration with CMOS.

having to increase the operating voltage and hence shows promise for a low-power design.

A quaternary GNTRAM design using xGNR latch with only 2 such xGNR devices is shown in Figure 34. In this design, leakage-critical paths employ high-Vt FETs while read FETs are implemented with low- Vt devices to distinguish between the stored states.

However, due to reduced voltage margins during stand-by, a larger state capacitance is now necessary to dynamically store the state during stand-by. As per the discussion in Chapter 4 Section 4.1, the state capacitance was calculated to be 350aF to ensure dynamic state retention. The memory cell operation is the same as explained in Section 6.1.2. The physical implementation is similar to ternary GNTRAM with 2 xGNR devices and follows a heterogeneous integration between CMOS and graphene as before (Figure 34C-E).

Benchmarking vs. 16nm CMOS:

HSPICE was used to benchmark the quaternary GNTRAM design against 16nm CMOS SRAMs and 3T DRAM. 16nm PTM LP MOSFET models were used for write and sleep FETs. For the read FETs, a low-Vt xnwFET device model [59] was used to meet the Vt requirements of this design. However, the CMOS FETs can be tuned to achieve this Vt or CMOS FINFETs may be used. PTM RC interconnect models were used for parasitic resistances and capacitances. Grid-based design rules in 16nm node were used to evaluate area, as before (Table I).

Detailed benchmarking is shown in Table V. This approach for quaternary GNTRAM exhibits much higher power benefits compared to the previous approach. Up to 4.67x active power and 3498x leakage power benefits were seen compared to high-performance

Table V. Quaternary GNTRAM Approach II 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 (μm^2)		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)	1.03 – 1.12	0.51 – 0.56	2.1 – 2.2	2.38 – 2.44	2.12 – 2.15
	Stand-by Power (pW)	8.89 – 9	4.44 – 4.5	6152 – 6157	15552 – 15556	6.49 – 7.01
Performance	Read Operation (ps)	7 – 9		8.35 – 9.25	7.68 – 7.96	9.18 – 9.68
	Write Operation (ps)	21 – 22.2		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 (μm^2)		0.03 – 0.06	0.015 – 0.03	0.026 – 0.064	0.0336 – 0.0672
Power Comparison	Active Power (μW)	1.03 – 1.12	0.51 – 0.56	1.21 – 1.16	1.45 – 1.47
	Stand-by Power (pW)	8.89 – 9	4.44 – 4.5	124.18 – 125.12	78.38 – 78.44
Performance	Read Operation (ps)	7 – 9		17.39 – 21.03	14.82 – 16.08
	Write Operation (ps)	21 – 22.2		67.27 – 67.54	58.37 – 63.18

CMOS SRAMs with comparable read performance. In terms of density, quaternary GNTRAM showed to 2.27x benefit against 16nm CMOS SRAM and 1.8x vs. 3T DRAM. Although the size of the stub was increased in the xGNR devices, this design requires only 2 xGNR devices which are relatively small compared to MOSFETs, and hence no additional area requirement was considered.

The drawback in this approach is that it requires a larger state capacitance when compared to the previous approach due to smaller voltage margins for state retention. As

the number of states is increased, the decreasing voltage margins will eventually limit the number of bits that can be stored in a cell with this approach.

6.3 Chapter Summary

In this chapter, two possible approaches were presented to allow further scaling of GNTRAM, by increasing the number of bits per cell. The first approach was using circuit techniques where additional xGNR devices in the latch allowed for storing 2 bits per cell (4 stable states). Extensive benchmarking of this quaternary GNTRAM design showed that it was 2.27x denser than 16nm CMOS SRAMs, however the power benefits were only up to 1.32x during active period and 818x during stand-by against high performance SRAMs. This was due to relatively high-voltage operation, which may ultimately limit this scaling approach. An alternative approach was presented to enable increasing the number of states by increasing the stub length in the xGNR devices. This allows for storing 2 bits per cell without requiring an increased operating voltage range. This quaternary GNTRAM evaluation showed that it had much higher benefits in terms of power, specifically up to 4.67x in terms of active power and 3498x during stand-by when compared to high-performance SRAMs. Thus it has the potential to allow further scaling to increase benefits further.

CHAPTER 7

CONCLUSION

In this thesis, we proposed a novel multistate volatile memory called GNTRAM for CMOS SRAM replacement, featuring a heterogeneous integration between graphene nanoribbons and CMOS. Binary and ternary implementations were presented and benchmarked extensively against 16nm CMOS SRAMs and 3T DRAM. Binary GNTRAM showed up to 10% density benefit and was up to 3.68x more power-efficient during stand-by mode when compared to low-power CMOS SRAMs. For the ternary GNTRAM, as more bits are stored (1.5 bits) in one cell, these costs are amortized. Thus we see higher benefits as expected with up to 1.77x better density and up to 1196x lower stand-by power compared to high performance CMOS SRAMs, while maintaining comparable performance. Hence, GNTRAM has the potential to overcome the physical scaling limitations of CMOS by storing more than 1 bit in a given cell.

In order to allow further scaling of GNTRAM, two possible approaches were explored to increase the number of bits per cell. The first approach used additional xGNR devices in the latch allowed for storing 2 bits per cell (4 stable states). Extensive benchmarking of this quaternary GNTRAM design showed that it was 2.27x denser than 16nm CMOS SRAMs, however the power benefits were only up to 1.32x during active period and 818x during stand-by against high performance SRAMs. This was due to relatively high-voltage operation, which may ultimately limit this scaling approach. An alternative approach was presented to enable increasing the number of states by increasing the stub length in the xGNR devices. This allows for storing 2 bits per cell without requiring an increased operating voltage range. Evaluation of this quaternary

GNTRAM design showed that it had much higher benefits in terms of power, specifically up to 4.67x in terms of active power and 3498x during stand-by when compared to high-performance SRAMs. A combination of both approaches may even be used to scale the number of bits per cell.

Thus multi-bit GNTRAM has the potential to realize high-density low-power nanoscale memories by overcoming the challenges associated with physical scaling. As graphene technology matures, the Si components may be replaced with graphene counterparts for even higher benefits. This thesis introduces the concept of using graphene based NDR devices for multi-state memory applications. Future work in this direction would be to take this concept further with an analysis of noise sources and the effect of variability and line-edge roughness in the graphene nanoribbon devices, in addition to exploring the use of graphene more extensively.

BIBLIOGRAPHY

[1] W. Wulf and S. McKee; "Hitting the wall: Implications of the obvious", ACM SIGArch Computer Architecture News, 23(1):20-24, Mar. 1995.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Siegmar Roth, "Carbon nanotubes and graphene: Electronic properties and applications", Lecture notes, Seoul National University (http://ntl.snu.ac.kr/2008summer/index.htm).

[11] The International Technology Roadmap for Semiconductors, 2011 (http://www.itrs.net/).

[12] K. V. Emtsev, et al.; "Towards wafer-size graphene layers by atmospheric pressure graphitization of silicon carbide", Nat. Mater. 8, 203 (2009).

[13] Sakulsuk Unarunotai, et al.; "Transfer of graphene layers grown on SiC wafers to other substrates and their integration into field effect transistors", Appl. Phys. Lett. 95, 202101 (2009).

[14] X. Li, W. Cai, J. An, S. Kim, J. Nah, D. Yang, R. Piner, A. Velamakanni, I. Jung, E. Tutuc, S.K. Banerjee, L. Colombo, and R.S. Ruoff; "Large-area synthesis of high-quality and uniform graphene films on copper foils", Science, 324, 5932, 1312–1314, 2009.

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

[17] Hong Li; Chuan Xu; Banerjee, K.; , "Carbon Nanomaterials: The Ideal Interconnect Technology for Next-Generation ICs," Design & Test of Computers, IEEE , vol.27, no.4, pp.20-31, July-Aug. 2010.

[18] Alexander A. Balandin; "Thermal properties of graphene and nanostructured carbon materials", Nature Materials 10, 569–581 (2011).

[19] Balandin, A.A.; "The Heat Is On: Graphene Applications," Nanotechnology Magazine, IEEE , vol.5, no.4, pp.15-19, Dec. 2011.

[20] Han, M. et al.; "Energy band-gap engineering of graphene nanoribbons", Phys. Rev. Lett. 98, 206805 (2007).

[21] Li, X., Wang, X., Zhang, L., Lee, S. & Dai, H.; "Chemically derived, ultrasmooth graphene nanoribbon semiconductors", Science 319, 1229–1232 (2008).

[22] Chen, Z., Lin, Y-M., Rooks, M. J. & Avouris,Ph.; "Graphene nano-ribbon electronics", Physica E 40, 228–232 (2007).

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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).

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

[30] 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.

[31] Khasanvis, S.; Habib, K.M.M.; Rahman, M.; Narayanan, P.; Lake, R.K.; Moritz, C.A.; , "Ternary Volatile Random Access Memory based on Hetero-geneous Graphene-CMOS Fabric," Nanoscale Architectures (NANOARCH), 2012 IEEE/ACM International Symposium on , in press.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

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

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] Williamson, W., III; Enquist, S.B.; Chow, D.H.; Dunlap, H.L.; Subramaniam, S.; Peiming Lei; Bernstein, G.H.; Gilbert, B.K.; , "12 GHz clocked operation of ultralow power interband resonant tunneling diode pipelined logic gates," Solid-State Circuits, IEEE Journal of , vol.32, no.2, pp.222-231, Feb 1997.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

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

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

[55] 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.

[56] Y.C. Kao, et al., “Vertical Integration of Structured Resonant Tunneling Diodes On InP for Multi-valued Memory Applications”, 4th Intl. Conf. On Indium Phosphide and Related Materials, pp.489-492, 21-24 Apr 1992.

[57] Elakkumanan, P.; Narasimhan, A.; Sridhar, R.; , "NC-SRAM - a low-leakage memory circuit for ultra deep submicron designs," SOC Conference, 2003. Proceedings. IEEE International [Systems-on-Chip] , vol., no., pp. 3- 6, 17-20 Sept. 2003.

[58] Xuemei (Jane) Xi; et al., “BSIM4.3.0 MOSFET Model – User’s Manual”, UC Berkeley, 2003.

[59] Panchapakeshan, P.; Narayanan, P.; Moritz, C.A.; , "N3ASICs: Designing nanofabrics with fine-grained CMOS integration," Nanoscale Architectures (NANOARCH), 2011 IEEE/ACM International Symposium on , vol., no., pp.196-202, 8-9 June 2011.