Chapter 1

INTRODUCTION

Since the evolution of the electronics industry, the memories have always been a fundamental part of every device and circuit. The recent technological advancements have escalated the demand for reliable, cheaper, denser, and faster memories with low power consumption that can provide ample data storage space. Currently, to meet this considerable demand, the conventional metal-oxide-semiconductor field-effect transistor (MOSFET) transistor technology-based Static RAM (SRAM), Dynamic RAM (DRAM), and Flash memories are employed. The 3-D NAND based flash memory device arrays are being used for the last two decades to meet this demand after the outburst of portable electronic devices such as laptops, smartphones, and other gadgets in the market. These devices serve a low cost solution with high density and fast storage. A flash memory cell comprises a structure similar to a MOSFET with an additional charge storing "Floating Gate" (FG) and an oxide layer to isolate FG with the gate terminal. However, FG MOSFETs are prone to shorting defects in gate oxide that may result in drain off the charge. Hence, nowadays, in place of FG, a thin Si_3N_4 charge trapping layer that contains defect sites is commonly used in NAND flash memories, known as silicon-oxide-nitride-silicon (SONOS) technology. The electronics industry is mainly focusing on scaling down the SONOS memories to meet the Moore's law criterion to increase the on-chip density and thereby improving the overall functionality of the processing unit. However, scaling these devices is a challenging task as the charge trapping capability of Si_3N_4 begins to deteriorate when scaled down below 6 nm and becomes even worse at the higher temperature. Moreover, MOSFET scaling also induces short channel effects such as hot carrier injection, velocity saturation, and drain induced barrier lowering.

Hence, after the saturation of scaling limits and performance issues in flash memories, the research was focused on an entirely new class of emerging non-volatile solid-state memory technology (NVM). So after using the magnetic disk commonly known as the hard disk drive (HDD) from the 1950s to easily 2000s and exploring the capabilities of solid-state drives (SSD), the entire memory industry's research momentum got shifted towards the four emerging NVM technologies i.e. resistive random access memories (RRAM), ferroelectric random access memories (FeRAM), phase change random access memories (PCRAM), and spin-transfer torque magnetoresistive random access memories (STT-MRAM). Referring to literature published by research and academic labs, all these technologies carry the potential to substitute not only the flash as storage but also the DRAM as the main memory for their higher storage density and operating speed.

1.1 CLASSIFICATION of MEMORIES

Conventionally, the semiconductor memory is broadly categorized as volatile and NVMs as far as their storage and retention capability is concerned [Szalay and Gray, 2006]. The Memories in which stores the data is stored permanently and does not require the stored

information to be refreshed and termed as NVM [Giovanni Campardo, 2005; Lee *et al.*, 2015]. While the volatile memories retain the data until the power supply is available to them. Currently under the computer memory hierarchy, a computer system consists of both HDD and SSD that are being for different applications. The type of applications assigned to these memories has been governed by their operating speed, storage densities, and volatilities, as shown in Figure 1.1. Conventionally, the data transfer in a traditional computer architecture takes place between the high density NVMs and low latency computational memories [Åkerman, 2005].

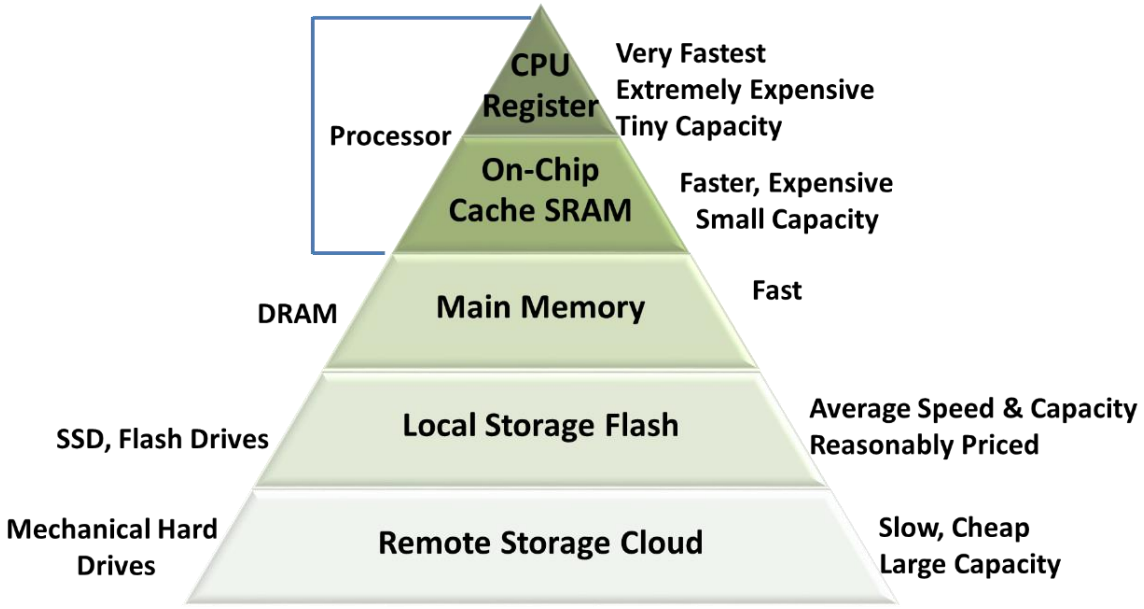

Figure 1.1: Memory hierarchy in computer system. As we move up from mechanical drives to CPU registers, the pricing is too high; however, the speed becomes faster. In contrast, as we move down the storage becomes denser and cheaper.

1.1.1 Volatile Memories

The volatile memories can be further classified into Static Random-Access Memory (SRAM) and Dynamic Random-Access Memory (DRAM), where SRAM is the caches memory that offers ultra-fast speed (0.3 ns) and higher programming cycles [Jeong *et al.*, 2012; Kim *et al.*, 2019c; Smullen *et al.*, 2011]. On the contrary, SRAM comes with complex and expensive fabrication and lower integration density due to larger feature size. The DRAM comprises of a one-transistor one-capacitor (1T1C) architecture and serves as the main memory component of any computer system that follows Von Neumann architecture. The logic "1" and "0" in a DRAM cell are defined by the charged and discharged state of the capacitor, respectively [Slater *et al.*, 1993]. The smaller feature size of a DRAM cell makes integration density higher than that of an SRAM cell; however, it comes with higher latency than the former. The data in a DRAM cell needs to be refreshed regularly during the processing and storing period due to its shorter retention time (a few milliseconds), which makes it a high power-consuming component [Aadithya *et al.*, 2013].

1.1.2 Non-volatile Memories

The typical examples of NVM are HDD and flash memories. The HDDs stores the data in the form of a magnetic field and has been the major power-consuming component with high storage capacity; however, latency, weight, and use of mechanical element were some major issues with them [Takashima *et al.*, 2011]. As time progressed, flash memories and their subsequent variants aggressively replaced the HDDs owing to their high speed read/write operation, less power consuming, compact, and less fragile. The flash memories have been composed of MOSFET devices with an additional floating gate (FG) and blocking oxide, as shown in Figure 1.3 [Burr *et al.*, 2008]. The programming of a flash memory cell has been performed by

channel hot carrier injection and charge accumulation in the floating gate, whereas the data could be erased by Fowler-Nordheim (FN) tunneling that drives the stored charge back into the transistor channel. The charge, once trapped inside FG will remain there for years unless no external bias is applied. The read operation is governed by the amount of charge stored in the FG after the write and erases operations [Pavan *et al.*, 1997]. An adequately low read voltage is applied at CG that doesn't alter amount of charge preserved in FG, but it must be sufficiently higher to discriminate between the charged and uncharged FG. The amount the charge stored

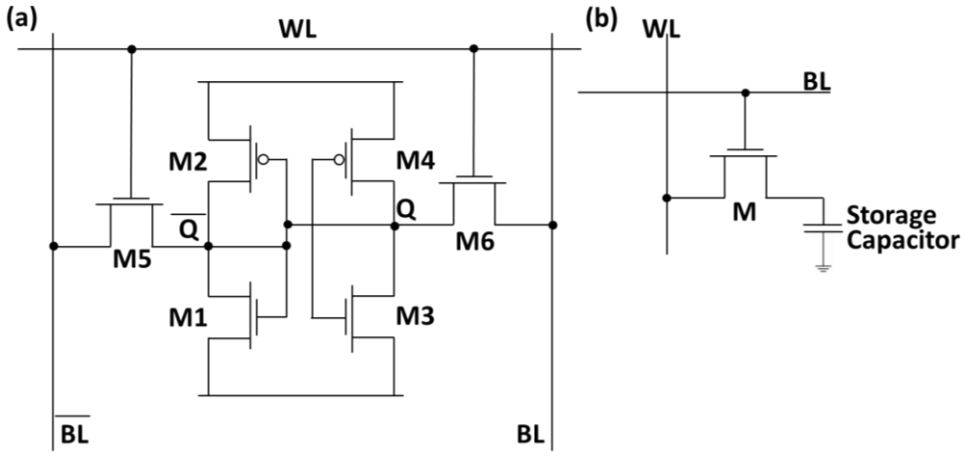

Figure 1.2: A typical conventional 6T SRAM cell architecture **(a)**. Schematic diagram of a DRAM cell **(b)**.

in FG drives the threshold voltage shift of the transistor, and reading the current at a particular voltage complies with "0" and "1" stored in the cell. Currently, flash memories can be further categorized into NOR and NAND based memories, each having its advantages and disadvantages. Higher read performance along with the complete address and data bus access are some of the critical features of NOR-based flash memories; however, larger area consumption and slow write and erase operation make it a weak contender against the NAND-based flash. The NAND-based flash memories demonstrate extremely high-capacity data storage systems that can be utilized in secondary storage devices such as hard disks, memory cards, flash drives, and optical drives [Jiyoung *et al.*, 2009].

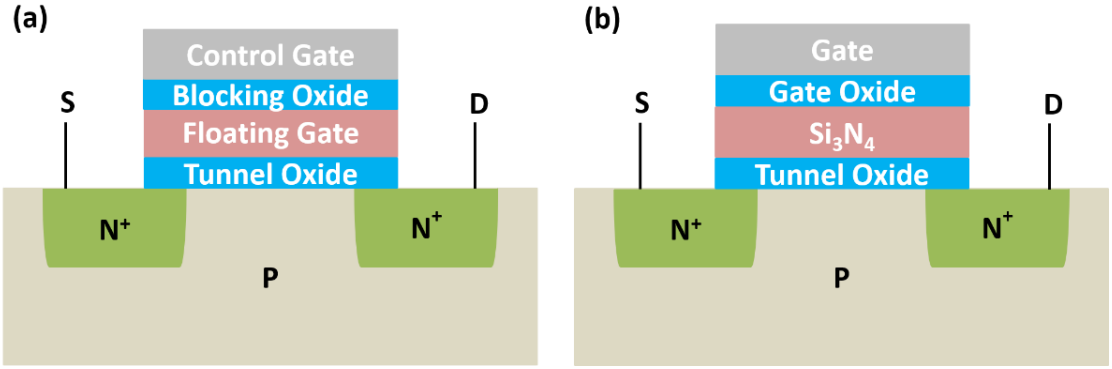

Figure 1.3: Schematic diagrams demonstrating the Flash memory **(a)** and SONOS memory cells **(b)**.

The charge in flash memory cells is stored in FG, which made up of polysilicon that must be reasonably thick for providing an adequate coupling ratio between the FG and the CG. This increases the number of processing steps and the required masks. Moreover, any reduction in the tunnel oxide thickness results in elevated leakage current though FG. Hence, to overcome these issues, an extended version of flash memory has come into limelight i.e., SONOS with a non-conductive silicon nitride layer in place of polysilicon that offers higher deep charge defect levels [Gritsenko *et al.*, 2003]. By using this technique, the gate stack can be scaled down up to

50% and lesser leakage due to the insulating behavior of silicon nitride and hence making the write/erase operation even faster at much lower voltages [Hubert *et al.*, 2009].

1.2 Emerging Non-volatile Memories

The major drawback of flash memories is it architecture that constitutes of MOSFET, which is currently facing scaling limitations [Bohr and Young, 2017]. Hence, industry and researchers are looking for an alternate solution for the last two decades, which can deliver scalable, lower power consumption, high speed, high endurance, CMOS compatible, and cost-effective technology [Yan *et al.*, 2019]. With an increasing pace of research in the emerging NVMs, the memory market is expected to grow at ~45% compound annual growth rate (CAGR) with a valuation of over $8 billion.

The four major technologies that have evolved in recent times are Ferroelectric Memories (FeRAM), Spin-torque Transfer Magnetic Memories (STT-MRAM), Phase Change Memories (PCM), and Resistive Random Access Memories (RRAM) [Chen, 2016; Chen, 2020; Mikolajick *et al.*, 2020; Wong *et al.*, 2010; Zhu, 2008]. The foundries and many electronic equipment manufacturers such as Global Foundries, TSMC, Sony, Hewlett Packard, Toshiba, Intel, Micron, Samsung, Western Digital, IBM, Fujitsu, Crossbar, Hynix and many other have already begun the large-scale production of some of these; however, the research for performance enhancement is still on. The exploration on various materials and process optimization for these technologies provides researchers a pathway to design high performance, cost-effective, and denser NVMs.

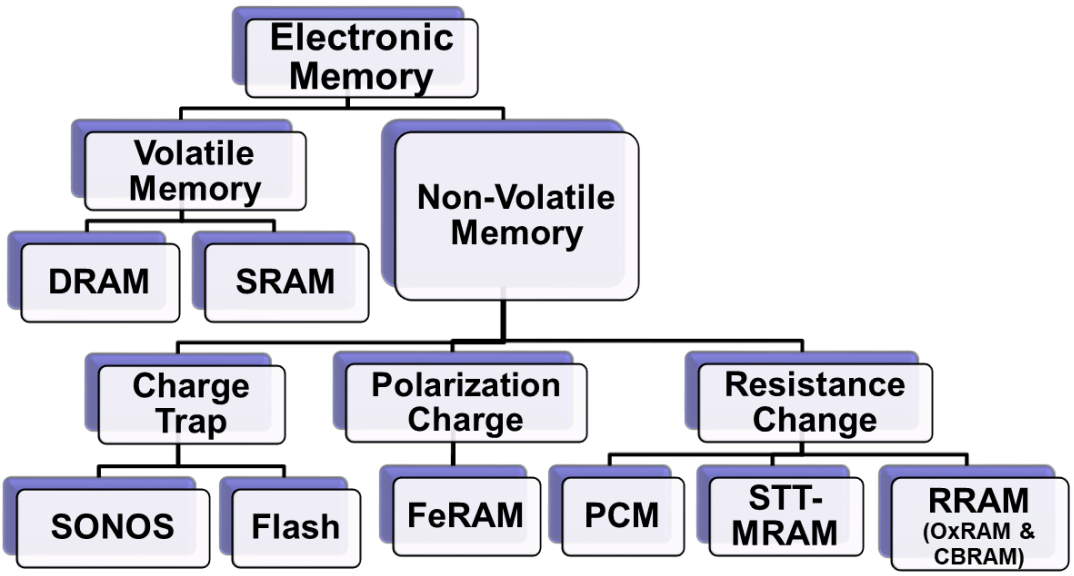

Figure 1.4: Categorization of standard semiconductor memories and emerging NVM technologies.

1.2.2 Ferroelectric Random Access Memories (FeRAM)

Figure 1.5(a) displays the schematic of a FeRAM memory cell with a ferroelectric material's switching layer sandwiched between the two electrodes. The most recent materials explored for ferroelectric switching layer are PZT ($Pb_xZr_{1-x}TiO_3$) and SBT ($SrBi_2Ta_2O_9$), possessing the quality of being highly scalable [Mikolajick *et al.*, 2020]. The application of electric field across the ferroelectric layer resulted in ionic polarization where the displacement of an ion from each unit cell occurs. This displacement and hence the polarization is irreversible even after the biasing potential is removed that in turns induces polarization hysteresis in the device, as shown in Figure 5(b) [Mikolajick *et al.*, 2001]. The positive and negative polarization saturation of FeRAM is considered as the distinct states of a memory system. The read operation is executed by applying a positive write voltage (+V), and the current is measured at saturated electric field levels [Chiu *et al.*, 2017]. The FeRAM read process is destructive in nature, which restricts the device to attain enough operating speed. The main advantage of FeRAM is its fast read and write

speed like DRAM; however, poor scalability remains a vital issue with this technology [Banerjee, 2020; Mikolajick *et al.*, 2020].

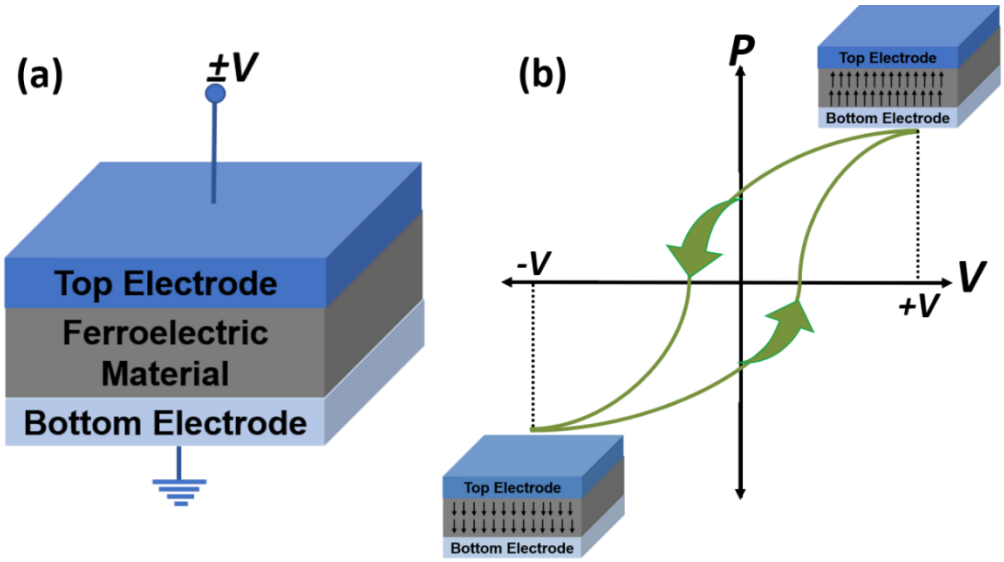

Figure 1.5: (a) Device schematic of a FeRAM device. **(b)** Development of polarization against the voltage applied across the top and bottom electrode (BE). The positive and negative polarization saturation at positive and negative voltages indicates the '1' and '0' states, respectively.

1.2.1 Spin-torque Transfer- Magnetic RAM (STT-MRAM)

The STT-MRAM's structure resembles with the magnetic tunnel junctions (MTJ) and has two ferromagnetic switching layers isolated by thin metal oxide tunnel insulating layer, as shown in Figure 1.5(a) [Gallagher and Parkin, 2006]. When an electric field is applied across the two terminals of the device, the magnitude of tunneling current varies with respect to the direction of magnetization in the two ferromagnetic layers [Hosomi *et al.*, 2005]. The parallel and ani-parallel magnetization will increase and decrease the tunnel current, consecutively modulating the device magnetoresistance (R_m), as shown in Figure 1.6(b). These resistance states corresponding to

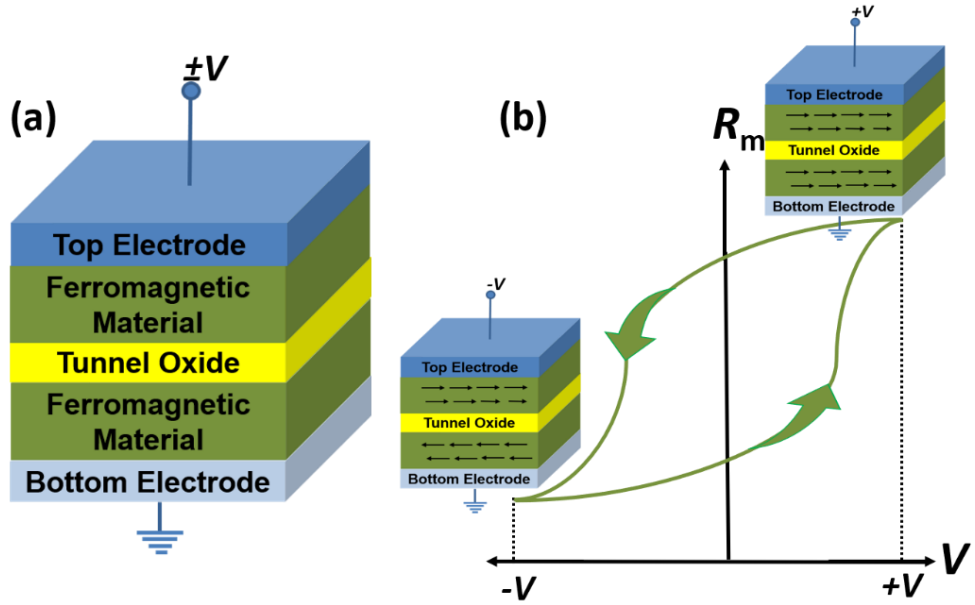

Figure 1.5: (a) Device schematic of a STT-MRAM cell. **(b)** The curve shows the magnetoresistance variation with applied voltage, depending upon the direction of polarization of the bottom ferromagnetic layer with respect to the top.

parallel and anti-parallel magnetizations indicates binary memory states [Khvalkovskiy *et al.*, 2013]. The operating speed, voltages, and cell size are almost similar to FeRAM; however, motion of ions is not needed for attaining the magnetization saturation. Hence, the bias loading-induced degradation issues of FeRAM do not arise in the STT-MRAM. Higher endurance (>10^{16}) and high programming and read speed (<10 ns) are key highlights of STT-MRAM [Ikegawa *et al.*, 2020]. However, STT-MRAMs have their challenges such as negligible memory window and hence smaller noise margin, scalability and complexity issues with a multi-layer structure, and fabrication with CMOS backend process without damaging magnetic properties of materials [Krounbi *et al.*, 2015].

1.2.3 Phase Change Memory (PCM)

As the name indicates, the PCM stores in the information in terms of electrical resistance states while the switching layer material's phase is changing [Wong *et al.*, 2010]. Its structure comprises a phase change chalcogenide alloy material such as $Ge_2Sb_2Te_5$ (GST), a resistive heating element, and two adjacent insulating layers, all sandwiched between the TE and BEs [Liu and Wang, 2020]. GST is already being used in rewritable DVDs. Among the emerging NVMs, PCM has demonstrated its capabilities to enter the mainstream NVM market. This chalcogenide material is temperature sensitive and contains the capability of altering its electrical conductance by changing its structural phase i.e., amorphous and crystalline [Kim and Lee, 2020]. As shown in the device structure in Figure 1.7(a), when an electric field is applied across the device, the heating element switches the structural phase of GST between amorphous and crystalline as per its crystal and melt temperature that changes its electrical conductance to resistive and conductive respectively. These resistance states correspond to the storage of '0' and '1' bit in a memory cell. The set process (writing of '1' bit) is executed by applying an electric pulse to heat some specific part (adjacent to the heater) of the switching layer above the crystallization temperature ($T_{crystal}$) [Salinga *et al.*, 2018]. The reset process (writing of '0' bit) requires a faster pulse of high current amplitude than the set process to melt-quench the GST and program the cell for low conductance state, as displayed in Figure 1.7(b). Higher operating voltage (~3 V) requirement for phase change and higher current requirements for the

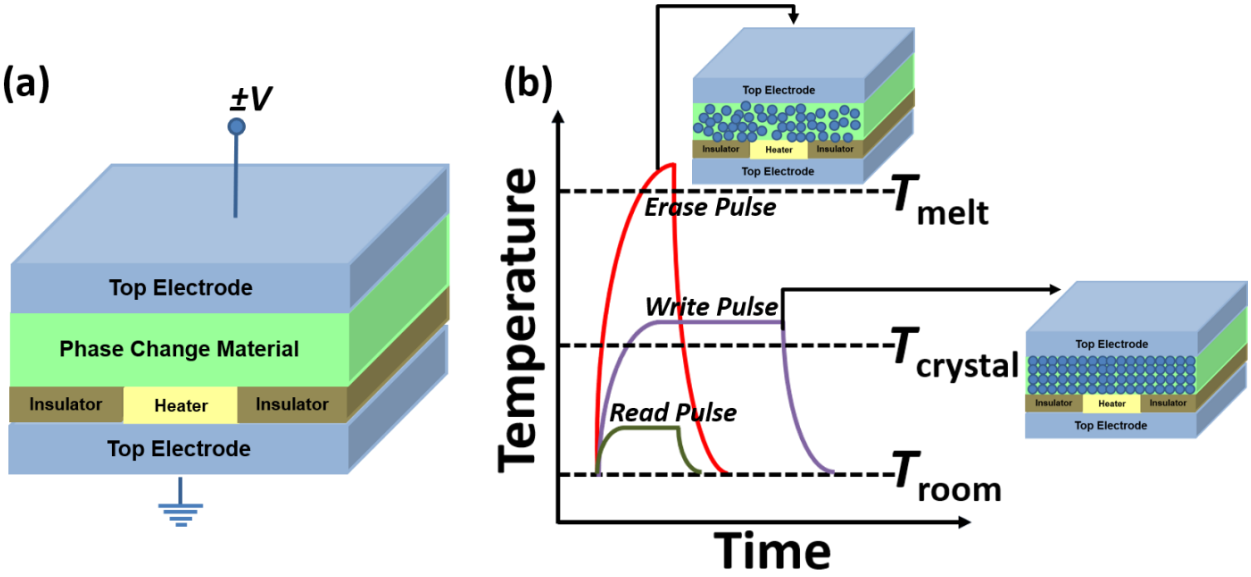

Figure 1.7: (a) Device schematic of a PCM cell. **(b)** Programing scheme of a PCM cell where a transition to amorphous phase happens when a shorter pulse at high temperature is applied that drives the cell to HRS. A longer pulse at low temperature convert the switching layer into crystal phase that switches the cell to LRS.

reset process are the key hindrances that ease its dominance over the other emerging NVMs. PCM has attracted some top companies of electronics industries owing to its advantages over NAND Flash, such as higher storage density, fast, and high switching endurance [Fong *et al.*, 2017].

1.2.4 Resistive Random Access Memory (RRAM)

Resistive random access memories comprise an insulating layer sandwiched between the two conducting electrodes, as shown in Figure. 1.8(a) [Chen, 2020]. RRAM is a typical two-terminal device where the information is stored in terms of multiple electrical resistance states, namely high resistance state and low resistance state (HRS and LRS). An RRAM cell works on the resistive switching phenomenon, which states that the conductance state of a dielectric can be changed when an extremely high electric field is applied across it. The electrical resistance states, i.e. HRS and LRS, correspond to the storage of "0" and "1" bit in an RRAM memory cell. HRS and LRS states are obtained due to the formation and rupture of a conductive path between the two electrodes when an electric field is applied across it [Lee *et al.*, 2015]. A wide variety of materials such as binary and ternary metal oxides, polymer dielectrics, transition metal dichalcogenides, bio dielectrics, and some other dielectrics and semiconductors have been explored so far as the switching layer. Inert metals such as Au, Pt, Ti, TiN, and ITO are generally chosen for the BE and both the electrochemically active or inactive metals can be used for the top electrode (TE). Based on the electrochemical activity of the TE material, the RRAM can be categorized into oxygen vacancy RRAM (VCRAM) and conductive bridge RRAM (CBRAM) [Wang and Yan, 2019]. Furthermore, depending upon the polarity of bias at which the resistive switching is happening; the RRAM can be further classified as "bipolar RRAM" or "unipolar RRAM". For a bipolar RRAM device, if the set process (switching to LRS) is happening one polarity, the reset process (switching to HRS) must occur at the opposite polarity, as shown in Figure 1.8(b). These type of RRAM devices are more popular for their reliable performance and for a better understanding of their switching mechanism. The unipolar RRAM devices switches between HRS and LRS at the same polarity but at different amplitude. The reset process in these devices is generally attributed to thermal effects resulted in dissolution of conductive filament (CF).

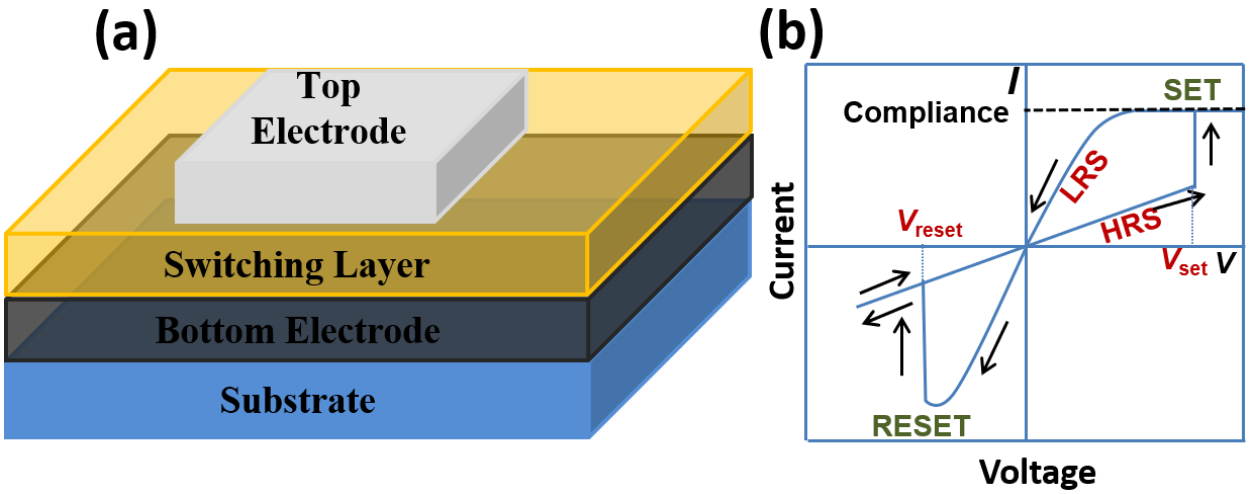

Figure 1.8: Schematic diagram of RRAM metal-insulator-metal structure with single switching layer **(a)** and bi-layer switching layer **(b)**. Schematic diagram representing current-voltage characteristics of unipolar and bipolar resistive switching.

In contrast to the aforementioned emerging and existing NVMs, RRAM carries the advantage of fast speed (<10 ns), low switching voltage (~ 0.5 V), and higher endurance (~10^{10} cycles) that make it a suitable candidate for future high-density memory applications. Moreover, RRAM has proven its capability in flexible electronics applications as well with high electromechanical stability under extreme mechanical stress and strain. Table 1 displays a performance comparison between the existing and emerging NVMs on various parameters such as cell area, operating voltage, energy consumption per bit, read time, write time, switching endurance, and retention time. The detailed discussion about the RRAM technology will be done in Chapter 2 where its working principle, device structures, challenges, and performance parameter will be elaborated.

Table 1.1: A benchmark table for the performance comparison on various parameters of emerging and existing memory technologies [LIU, 2018; Munjal and Khare, 2019; Schenk *et al.*, 2020; Zahoor *et al.*, 2020].

Memory Technology	SRAM	DRAM	NAND Flash	FeRAM	PCM	STT-MRAM	RRAM
Cell area	>100F^2	6F^2	<4F^2 (3D)	22F^2	4-20F^2	6-20F^2	<4F^2
Cell element	6T	1T1C	1T	1T1C	1T/1R	1T/1R	1T/1R
Voltage	<1 V	<1 V	<10 V	~3 V	<3 V	<2 V	<1 V
Read time	~1 ns	~10 ns	~ 10 µs	~5 ns	<10 ns	<10 ns	<10 ns
Write time	~1 ns	~1 ns	10 µs-1 ms	10 ns	~50 ns	<5 ns	<10 ns
Energy/bit	~fJ	~pJ	~1 nJ	50 fJ	3 pJ	~2 pJ	~50 pJ
Retention	N/A	~64 ms	>10 years	>10 years	>10 years	>10 years	>10 years
Endurance	>10^{16}	>10^{16}	>10^4	>10^{14}	>10^9	>10^{15}	>10^9
Multibit capacity	No	No	Yes	No	Yes	Yes	Yes
Non-volatility	No	No	Yes	Yes	Yes	Yes	Yes
Scalability	Yes	Limited	Limited	Challenge	Limited	Good	Excellent

1.3 Research Focus and Thesis Organization

The research interest of the scientific community has grown in RRAMs since early 2000, which is evident from the increasing number of publications and patents being filed every year. The prime focus during these years has been to explore various materials that can deliver highly reliable RRAM devices. The device performance has improved remarkably owing to the tremendous research related to methods of device fabrication, electrical characterization, improved film uniformity, stoichiometry control, and film and material characterization techniques.

This work has been majorly focused around to produce high performance and highly flexible RRAM devices using hybrid bilayer switching layers by taking control over growth and rupture of CF during the switching processes. Nonetheless, the selection of appropriate materials such as switching layer dielectric, TE metal, BE metal, and substrate and then the optimization of device structure is the most crucial step towards achieving a high-performance RRAM device. A proper control over thickness and stoichiometry of the switching layer is essential for improving the reliability and lifetime of the device. In that regards, the atomic layer deposition (ALD) is the most trusted technique for high quality and uniform deposition of high-k metal oxide layers such as AlO_x, TaO_x, TiO_x, and HfO_x and among them, HfO_x was chosen as the primary switching layer for many reasons like its high band gap of 5.3-5.7 eV, low leakage current and compatibility with CMOS fabrication process flow. Moreover, the integration of solution processed low-k polymer dielectrics such as poly(4-vinylphenol) (PVP), polymethyl methacrylate, polyvinylcarbazol, and polyvinylalcohol in the bilayer RRAMs has been explored intensively in recent times and it provides highly uniform resistive switching characteristics that results in lower device-to-device and cycle-to-cycle variability. Hence, with HfO_x as the primary switching layer and PVP based composite thin film as the secondary layer, the aim was to achieve the following goals:

I. Exploring the ALD deposited metal oxides for low voltage switching operation.

II. Improved cyclic endurance and memory window with bilayer structure.
III. Achieve the excellent memory window with high cyclic endurance and low voltage switching on the flexible substrates
IV. To control the switching power consumption by regulating the on current on the flexible substrates.

The targets of the first goal were achieved by using AlO_x as the switching layer which was ALD deposited using multi-temperature deposition scheme to exploit the oxygen vacancies-controlled formation and rupture of CF. For second target, a thin layer HfO_x and solution processed PVP was used where HfO_x improved the memory window and PVP provided the low voltage and decently reliable switching operation. Third and fourth targets were achieved by using PVP composite with GO in the first study and molybdenum disulfide (MoS_2) in the second study that has delivered RRAM devices with high reliability, low voltage switching, low power consumption, and high cyclic endurance along with the excellent flexibility.

This thesis has been organized as following:

Chapter 1 describes in brief background of the motivation of this research. Also, the typical volatile and non-volatile memory technologies are discussed and eventually their comparison with the emerging memory technologies has been made.

Chapter 2 discusses about the various aspects of RRAM working and performance. The resistive switching phenomenon and its type are well explained in detail. Moreover, the various device architectures, their possible application space, and challenges with each of them is elaborated here. Eventually, the chapter is concluded with discussion about major RRAM performance parameters and various challenges associated with them.

Chapter 3 explains various fabrication and analytical tools used for film deposition and material characterization, respectively. In addition, a detailed explanation of electrical characterization of the fabricated devices and data extraction from the obtained results for RRAM is provided.

Chapter 4 presents ALD deposited AlOx based resistive random access memories fabricated with a novel multi-temperature deposition (MTD) scheme. The variation in deposition temperature resulted in the concentration gradient of oxygen concentration. The higher oxygen vacancies in the film created stronger filament and vice-versa. This causes the localization of filament rupture at weak regions and hence the reliable switching operation.

Chapter 5 presents a hybrid bilayer RRAM with HfO_x, and PVP, demonstrating a higher cyclic endurance with a better memory window at lower switching voltage. HfO_x has induced lower leakage current in the device to enhance the memory window. The pin holes present at the PVP surface guides the migration of metal ions through the switching layer. Therefore, a reliable switching operation with improved memory window and low voltage switching is achieved.

Chapter 6 presents the exploration of polymer and 2-dimensional materials composite and ultra-thin HfO_x as the switching layer. After optimizing the graphene oxide (GO) concentration in the composite solution, the devices demonstrated excellent switching behavior with excellent memory window and low voltage switching operation with improved repeatability. Moreover, owing to the strength of GO sheets, devices exhibited decent flexibility without any degradation in the switching performance.

Chapter 7 reports the extended version of the work demonstrated in Chapter 6 where the GO was replaced with molybdenum disulfide (MoS_2) that further enhances the low voltage switching with addition advantage of extremely low ON current. The extremely low ON current is attributed to the formation and rupture of multiple weak conductive filaments. Furthermore, the devices exhibited excellent flexibility under extreme bending conditions.

Chapter 8 summarizes the research work done in this thesis and concludes with scope of the future expansion of this work in the related applications.