ENABLING LOGIC-MEMORY SYNERGY USING INTEGRATED NON-VOLATILE TRANSISTOR TECHNOLOGIES FOR ENERGY-EFFICIENT COMPUTING

by

Sandeep Krishna Thirumala

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Electrical and Computer Engineering
West Lafayette, Indiana
December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Sumeet Kumar Gupta, Chair
School of Electrical and Computer Engineering

Dr. Anand Raghunathan
School of Electrical and Computer Engineering

Dr. Zhihong Chen
School of Electrical and Computer Engineering

Dr. Vijay Raghunathan
School of Electrical and Computer Engineering

Dr. Kaushik Roy
School of Electrical and Computer Engineering

Approved by:
Dr. Dimitrios Peroulis

To my family for their unconditional love and support.

ACKNOWLEDGMENTS

It is a bitter-sweet moment to write this part of my dissertation. Looking back, my PhD chapter of life has been nothing but an incredible voyage. Phenomenal learning curve, overcoming taxing hurdles and amazing collaborations is what has pushed me to achieve my dream and grow me into the person I am today.

First and foremost, I would like to wholeheartedly extend my sincere gratitude to my academic advisor, Professor Sumeet Kumar Gupta, for his excellent guidance and mentorship at every step of my doctoral studies. I am grateful to him for challenging me and providing me with an excellent research environment. I have always been motivated by his immense energy for research and strive for excellence, which enabled me to overcome numerous obstacles throughout my PhD. His strong emphasis on hard-plus-smart work ethics, flawless research presentations, and high-quality publications has helped me evolve as a researcher. I will be forever indebted to him for letting me steer the course of my PhD during the final years, which has made me learn to be an independent researcher, discover my strengths and gain skills to face the unpredictable challenges of academic problem-solving. I cannot thank him enough for being an outstanding mentor, critic, and companion throughout the course of my PhD. Working under his supervision has been a privilege for me, and this dissertation would certainly not have been possible without him.

Besides my advisor, I would like to thank my dissertation committee members, Prof. Anand Raghunathan, Prof. Zhihong Chen, Prof. Vijay Raghunathan and Prof. Kaushik Roy for their continuous collaboration and support with experiments and analysis. I consider myself extremely fortunate to have constant interaction and guidance from them, who are simply 'world-class' researchers and teachers. Their insightful thoughts and constructive discussions have helped me steer my research towards the right direction. I thank them for their valuable feedback on refining my research proposals in this dissertation.

Next, I would like to thank my fellow lab mates and alumni of the Integrated Circuits and Device Laboratory (ICDL) – Prof. Ahmedullah Aziz (University of Tennessee), Atanu Kumar Saha, Niharika Thakuria, Karam Cho, Chunguang Wang and Xinkang Chen. It was a great experience having everyone around. I truly enjoyed and cherish all our conversations over a broad range of topics including heated research debates and fun, non-academic jabbers. In the process, I have learned so much from each one of them. I will forever remember all the memories of the lab.

Special mention to Dr. Arnab Raha (Intel), Dr. Shubham Jain (IBM) and Dr. Terry Hung (TSMC) with whom I have had numerous discussions, which led me to explore a wide range of topics and gain more comprehensive research perspectives. Their constant support and valuable insights have shaped my dissertation significantly. They have played a very important role in my PhD journey and I sincerely thank them for all their contributions.

I am also grateful to Dr. Hongmei Wang and Dr. Fabio Pellizzer for mentoring me and introducing me to the real-world industrial projects, during my internship at Micron Technologies, USA. I had to opportunity to learn several insights and concepts, along with the experience to face challenges which come along the way while bringing an emerging concept to reality.

I would also like to acknowledge the sponsors of my doctoral research, which was supported in part by the New Materials for Logic, Memory and Interconnects (NEWLIMITS), a center in nCORE, a Semiconductor Research Corporation (SRC) program sponsored by National Institute of Standards and Technology (NIST) and in part by the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award (YFA) Grant. I am also extremely grateful for the Bilsland Dissertation Fellowship awarded to me by the Purdue University during the final year of my PhD.

Finally, I would like to thank to all my friends who gave me good advice and made this journey extremely enjoyable. In particular, I would also like to thank Sujata Khandare, for all her motivation and support during my entire graduate studies at the Purdue University. A special thanks goes to my sister Lalenthika Krishna, for constantly supporting me during this chapter of life and for being my loving stress-busters. Also, a shout-out to my four-legged sister Zoey Krishna (Labrador), for being a faithful virtual companion during the course of my studies. I would like to thank my grandparents, for always believing in me achieving greater heights. My deepest gratitude goes to my parents, Shanthy Madhusudan and Madhusudan Thirumala for their guidance, patience and unconditional love throughout, that provided the foundation for this work. This dissertation is a token of my respect and appreciation to all of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

FEFET	Ferroelectric Field Effect Transistor
R-FEFET	Reconfigurable Ferroelectric Field Effect Transistor
NVM	Non-Volatile Memory
NVFF	Non-Volatile Flip-flop
TMD	Transition Metal Dichalcogenides
IPS	Intermittently Powered Systems
STP	Sequential Transient Process
P_{HOLD}	Hold Polarization
GSHE	Giant Spin Hall Effect
VSHE	Valley-coupled-Spin Hall Effect
ML	Machine Learning
DNN	Deep Neural Networks
AI	Artificial Intelligence
STT-MRAM	Spin-Transfer-Torque Magnetic Random-Access Memory
PCM	Phase Change Memory
RRAM	Resistive Ransom Access Memory
FERAM	Ferroelectric Random Access Memory
CRC	Cyclic Redundance Check
RSA	Reconfigurable Sense Amplifier
RCSA	Reconfigurable Current Sense Amplifier
CiM	Computation-in-Memory
GL	Gate Leakage
RDM	Read Disturb Margin
MAC	Multiply-and-Accumulate
ADC	Analog to Digital Converters
$RFEFET_{SYM}$	Symmetric R-FEFET
$RFEFET_{ASYM}$	Asymmetric R-FEFET
ICT	Information and Communication Technology

ABSTRACT

Over the last decade, there has been an immense interest in the quest for emerging memory technologies which possess distinct advantages over the traditional silicon-based memories. The unique opportunities ushered by these technologies such as high integration density, near-zero leakage, non-volatility and, in some cases, excellent CMOS compatibility, has triggered the development of design techniques, enhancing the computation capabilities of various systems. Further, in the era of big data, the emerging memory technologies offer new design opportunities to address a pressing need of achieving close integration of logic and memory sub-systems with an objective to overcome the von-Neumann bottleneck associated with the humungous cost of data transfer between logic and memory. Such a logic-memory coupling not only enables low power computation in standard systems, but also promises high energy efficiency in unconventional compute platforms such as the brain-inspired deep neural networks (DNNs) which have transformed the field of machine learning (ML) in recent years. However, in order to exploit the unique properties of the emerging memory technologies for efficient logic-memory integration, there exists a strong need to explore cross-layer design solutions which can potentially enable efficient computation for current and future generation of systems. Motivated by this, in this dissertation, we harness the benefits offered by the emerging technologies and propose novel devices and circuits which exhibit an amalgamation of logic and memory functionalities. We propose two variants of memory devices: (a) Reconfigurable Ferroelectric transistors (R-FEFET) and (b) Valley-Coupled-Spin Hall (VSH) effect based magnetic random-access memory (VSH-MRAM), which exhibit unique logic-memory unification. Exploiting the intriguing features of the proposed devices, we carry out a cross-layer exploration from device-to-circuits-to-systems for energy efficient computing. We investigate a wide spectrum of applications for the proposed devices including embedded memories, non-volatile logic, compute-in-memory circuits and artificial intelligence (AI) systems.

The first technology of our focus is ferroelectric transistor (FEFET), which is being actively explored for logic and memory applications. Experimental studies have showcased volatile (logic) or non-volatile (memory) characteristics for FEFET by employing static/design time optimizations. However, if run-time tuning of non-volatile and volatile modes can be achieved, several new

avenues for circuit design will open. Inspired by this, we propose Reconfigurable FEFET (R-FEFET), which has the capability to dynamically modulate its operation between volatile and non-volatile modes, enabling true logic-memory synergy at the device level. Utilizing these unique features of the R-FEFET, we propose an embedded non-volatile flip-flop design (R-NVFF) featuring a fully automatic backup operation (during power shut down) without the need of any external circuitry or signals. Compared to a previously proposed FEFET based NVFF, the proposed R-NVFF exhibits 69% lower check-pointing energy (which includes backup and restore operation). We also propose non-volatile memory (NVM) with highly energy-efficient read and write operations enabled by the dynamic reconfigurability in R-FEFETs. Our proposed NVM exhibits 55% lower write power, 37%-72% lower read power and 33% lower area compared to an FEFET-NVM. Finally, we implement the proposed NVM and R-NVFF in a state-of-the-art intermittently-powered platform and show up to 40% energy savings at the system-level.

Another technology, which has sparked immense interest in spintronic applications, is the Valley-coupled-Spin Hall (VSH) effect in two-dimensional Transition Metal Dichalcogenides (2D TMDs). The unique generation of out-of-plane spin currents in monolayer TMDs can potentially enable efficient switching of nano-magnets. In this dissertation, we propose an emerging spin-based memory device featuring close logic-memory integration utilizing the VSH effect in 2D TMD transistors, where the information is stored in nano-magnets (which are unified with the transistor), for energy efficient computing. We propose two variants of NVM designs, namely single-ended VSH-MRAM and differential DVSH-MRAM. We show that the integrated gate feature exclusive to 2D TMDs, facilitates access transistor-less memory array designs, resulting in ultra-high integration density. We compare the proposed memory designs with the standard Giant Spin Hall (GSH) effect-based memories and showcase 35%-67% lower energy consumption at the circuit-level and up to 3.14X energy efficiency at the system-level in the context of general-purpose computing systems as well as targeted system applications such as energy harvesting platforms.

In addition to traditional computing architectures, the logic-memory synergy in the proposed device technologies, showcase an immense potential for energy-efficient in-memory computation, especially for AI specific hardware running DNN/ML algorithms. We propose R-FEFET and DVSH-MRAM based design of novel compute-enabled memory fabrics. We custom design memory bit-cells which enable massively parallel Boolean and non-Boolean in-memory computations using minimal array accesses. For example, we propose R-FEFET and DVSH-MRAM based NVM cells which performs natural and simultaneous computation of bit-wise Boolean AND and NOR logics in a single array access. We also propose a compact compute module, attached to the array peripherals, for carrying out other logic and arithmetic operations such as addition. The proposed in-memory computation technique shows up to 71% lower energy consumption compared to existing FEFET and GSH-MRAM based compute-in-memory solutions. Moreover, for targeted energy-autonomous system workloads, we propose application-specific, FEFET inspired CiM fabric, which demonstrate 32X and 40X improvement in energy consumption and performance during edge-sensing, when compared to conventional computing architectures. Lastly, for energy-efficient computing in edge devices, we propose compute-enabled memory cells with ternary-precision, which achieves a sweet spot between accuracy and energy-efficiency for DNN workloads. With optimal encoding scheme for the computing elements in synergy with device-circuit co-design, we achieve efficient ternary in-memory dot-product computation with minimal number of transistors per cell. The proposed ternary compute-in-memory arrays show up to 3.4X reduction in energy and 7X improvement in performance when compared to optimized near-memory DNN accelerators. Overall, evaluation results of the proposed CiM techniques in this dissertation, show significant reduction in system energy along with system performance improvement over conventional von-Neumann architecture-based approaches for a wide range of application workloads, thus addressing the critical need for energy efficient logic-memory synergy in future computing platforms.

1. INTRODUCTION

1.1 Data Intensive Computing: Memory Bottleneck and Climate Burden

The memory and storage system in modern day computers consists of a hierarchy of devices with various density, speed, and cost (Fig. 1.1) [1]. The hierarchy includes: registers for holding temporary results and variables; caches which act as staging areas for the data and instructions stored in the main memory; main memory to stage data stored in large but slow storage entities, such as hard disk drives (HDDs) or NAND flash. As we enter the era of big data, many emerging data-intensive workloads become pervasive, and mandate very high bandwidth and heavy data movement between the computing units and the memory. Storing and manipulating such a humungous amount of data raises significant challenges in designing high-performance, energy-efficient memory hierarchy. This has been a major concern over the recent years as the systems are expected to be more compact and powerful with data-intensive computing [2], [3].

Unfortunately, technology scaling has further aggravated the aforementioned problem of memory and storage systems, significantly slowing the performance improvements of computing systems. Over the years, it has been observed that microprocessor speeds are increasing, but not at the same rate that memory access latencies have decreased. This increasing gap between the pace of processor and the memory has created the “memory wall” problem in which the data movement between the logic and memory component of a system is becoming the bottleneck in present computing architectures (Fig. 1.1-1.2) [4]–[6]. Sources state that the increase in the processing

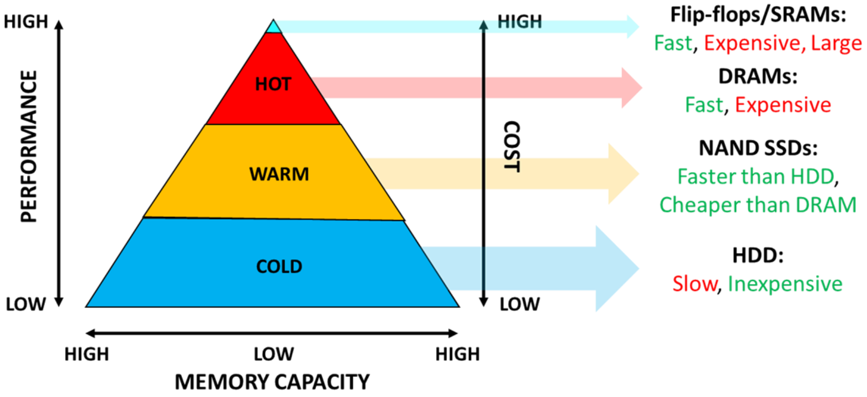


Fig. 1.1 Memory and storage system in a computing system showcasing the hierarchy of elements involved with a wide spectrum of performance, cost and density.

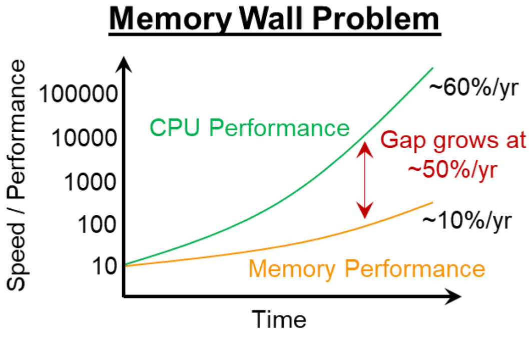


Fig. 1.2 Processor and Memory speed over time illustrating the memory wall problem.

speed is around 60% every year, while the rate of memory operation has only improved by less than 10% per year, resulting in doubling of the gap between the performance of the processor and memory every 1 to 2 years [6], [7]. Moreover, commodity memory technologies, such as SRAM and DRAM, are facing scalability challenges due to several constraints. Firstly, both SRAM and DRAM are leaky at advanced technology nodes and the leakage power starts dominating for high memory capacity. Results in [8], [9] suggests that 25–40% of total power is attributed to the memory system [8] and the embedded processor caches consume over 40% of the total chip power budget [9]. This can significantly degrade the energy efficiency of systems based on scaled technologies when static leakage power is considered. Secondly, the SRAM/DRAM architectures are facing many difficulties while being scaling down. One challenge is in regard to the large increase in process variations [10], [11]. With the continuous scaling of CMOS devices, variations in key parameters such as threshold voltage (V_{TH}), oxide thickness (T_{OX}), etc. are growing at an alarming rate [12]. Subsequently, the performance of different die on the same wafer can vary widely, resulting in a significant parametric yield loss, which directly translates into higher manufacturing costs. Another challenge is that its intrinsically hard to scale down. For example, scaling down DRAM below a 20nm process node is extremely challenging due to the difficulty in keeping an adequate amount of cell capacitance [13]. Therefore, the memory and storage systems are becoming a fundamental performance and energy bottleneck in various advanced generation of electronic systems, ranging from cloud servers to end-user devices.

Another harmful aspect of data-intensive computing is their impact on climate change. The current and future generation of computing largely relies on data centers which store and process humungous amounts of data. The data centers themselves consume a lot of power and as per recent reports, nearly 2% of electrical energy utilization in the U.S alone is from the data centers [14]. It

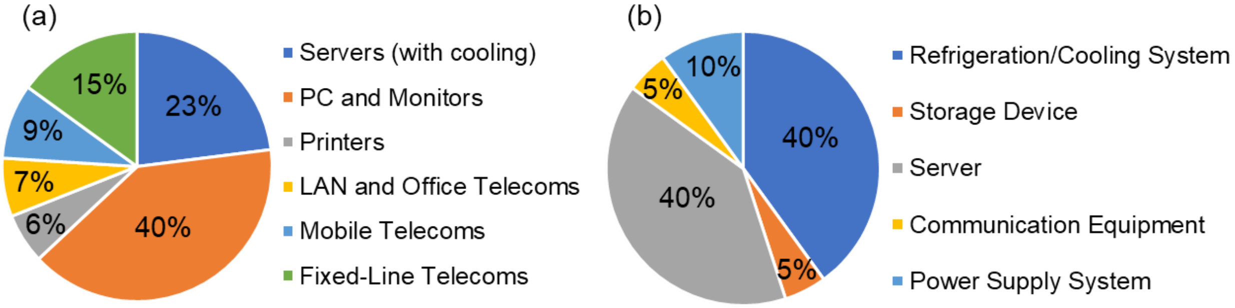


Fig. 1.3 (a) Estimated ICT CO_2 emissions (b) Energy consumption distribution of data centers

is expected that the CO_2 emissions from information and communication technology (ICT) will account for 12% of world-wide emissions by the end of 2020 [15]. These numbers are alarming, especially while considering the explosion in the data demands observed over the recent years.

The energy consumption of data centers may be divided into two categories: computing resources and physical resources (Fig. 1.3). The statistics show that the energy consumption of computing resources accounts for about 50% of the total energy consumption [15]. On the other part, the energy consumption of refrigeration/cooling systems is a major part of energy consumption by physical resources, which accounts for about 40% of the total energy consumption. Scientists believe that this number would vastly increase in the very near future considering the estimate of the 175 zettabytes (175 trillion gigabytes) of data generated by 2025 across the globe [16]. Therefore, we can conclude that servers with the processing elements (which fetch, compute and store data in the memory) and cooling systems are the most substantial energy draining facilities in the data centers. They account for a dominant portion of the total operating costs. Therefore, reducing energy consumption for servers and cooling systems is a key issue to address the sustainable development of data centers.

There are two ways to tackle this problem of carbon emissions from the data centers. The first approach is by efficiently controlling the heat generated as a result of the high computing demands in modern workloads. For example, innovating cooling strategies which consume less energy or re-utilization of the thermal energy generated in partially powering towns and cities would reduce the carbon footprint of data centers [17]. Another possibility is to build data centers in eco-friendly locations (cold regions), thereby partially or completely eliminating the requirements of cooling systems [18]–[20]. However, with the increasing demand and stringent federal constraints on the privacy of data, which requires it to be stored in the home country of the institution managing the data center, setting up data centers in remote/distant locations such as, Iceland, or even Antarctica, for the entirety of world population, might not be feasible. The second approach to overcome the energy inefficiency of the data centers is to intrinsically develop materials, devices, circuits and systems which are intrinsically energy-efficient. Several researchers across the globe are constantly innovating novel design architectures in this aspect. Such an approach can help solve the root cause of the problem and drastically reduce the power consumption of the data centers, which in turn will reduce the impact on climate change. One of the major constituents of a data center is the memory, which stores the data. Accessing data to-and-fro, the memory consumes significant component of the server energy. Therefore, in the following section, we discuss the importance of exploring emerging memories for future generation of electronic systems.

1.2 Emerging Memories to the Rescue

With technology scaling the traditional CMOS memories (SRAM/DRAM) exhibit a fundamental limitation for high performance, energy-efficient computing as detailed in the previous section. To overcome this drawback, various emerging non-volatile memory (NVM) technologies have been proposed to replace/complement SRAMs and/or DRAMs because of their appealing advantages such as high density, zero standby power, fast access speed, non-volatility, etc., [21]. They showcase the potential to efficiently overcome the drawbacks of the traditional memory hierarchy design which makes them important technology enablers for high-performance and intelligent hardware systems in the near future [22], [23].

Both prototypical and emerging non-volatile memories are based on novel materials and mechanisms which are drastically different from the traditional CMOS counterparts [24]. In the following, we briefly describe the advantages and drawbacks of some of the widely explored emerging memories:

- *Phase Change Memories (PCMs)* are based on reversible transition between the amorphous phase (high resistance) and the crystalline phase (low resistance) of chalcogenides (Fig. 1.4(a)) [21]. PCMs use special alloys, including Germanium Antimony Tellurium (GST), which have innovative characteristics that enable the non-volatile storage of their material phase by manipulating the heat inside the material. Micron's X100 NVMe SSD [25] and Intel's Optane

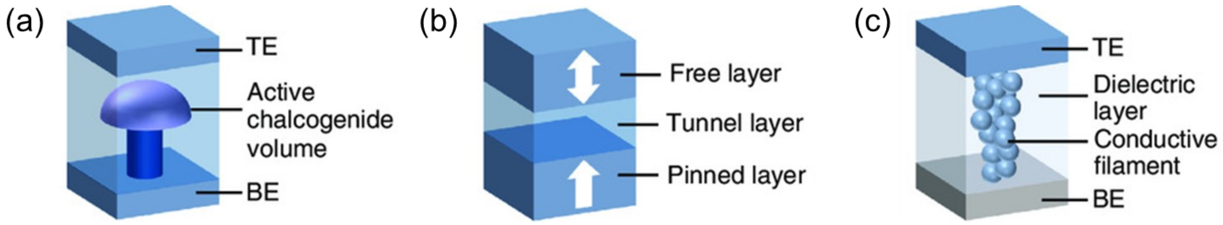


Fig. 1.4 (a) Phase Change Memory cell (b) Magnetic Tunnel Junction used in STT-MRAMs and (c) Resistive RAM cell. Figure adapted from [50].

technology [26] are few examples of well-established and mass-produced products using PCM. They demonstrate desirable characteristics such as longer retention and improved endurance compared to NAND Flash and also showcase functionality at scaled dimensions. However, the large latency of transition between the two phases (~100ns) and their high write power consumption hinder their direct implementation in circuit design [27]. Moreover, the ability of Flash to store and detect multiple bits per cell still gives flash a memory capacity advantage over PCMs. Although, multi-level storage is a possibility in PCMs, it is yet to be demonstrated in the industry products.

- *Spin-Transfer Torque Magnetic RAM (STT-MRAM)* is based on a magnetic tunnel junction (MTJ) cell with the most popular architecture composed of one transistor and one MTJ-based resistor (1T-1R; Fig. 1.4(b)) [23]. The resistance of the MTJ depends on the relative magnetization of the free layer (FL) with respect to the pinned layer (PL). They showcase faster read/write latencies (~10ns) and very high endurance (~10^{15}) when compared to PCMs, and also smaller device footprint for high integration densities, close to that of DRAMs [21]. Samsung’s STT-MRAM in 28nm FDSOI platform [28] and Intel’s FinFET based MRAM technology [29] are some industrial efforts on the implementation of spintronic memory. However, the major disadvantage appears to be the high write current which increases the power consumption [21]. Moreover, the resistance-based distinguishability between the two logic states is low when compared to other emerging technologies (~orders of magnitude), leading to challenges associated with data sensing [30]–[32].
- *Resistive RAMs (RRAMs)* are another class of NVM technologies built on the resistance changing mechanisms. An example of RRAM is conductive bridge RRAM where the switching is generally attributed to the formation and rupture of conductive filaments in insulating oxides (Fig. 1.4(c)) [33]. It involves a simple 3-layer memory cell: metallic top

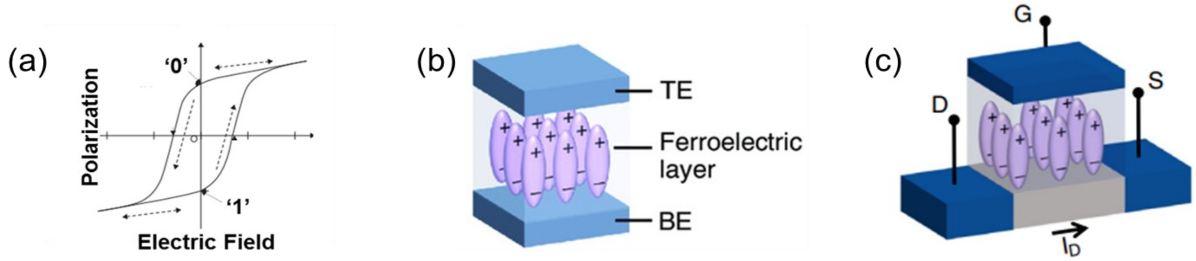


Fig. 1.5 (a) Polarization vs voltage hysteresis loop of FE illustrating the bi-stable states (b) FE capacitor structure used in FeRAMs and (c) FEFET device structure with FE integrated in the gate stack if a transistor. Figure adapted from [50].

electrode, resistive switching medium and metallic/non-metallic bottom electrode. Both uni-polar and bi-polar variants of RRAM have been explored in the past. 1T-1R array architectures have also been built and explored to demonstrate CMOS compatible, high density memory arrays. The P-series from Crossbar Inc. [34], TSMC's 22nm ReRAM technology [35] and Fujitsu's world's largest density and mass produced, 8Mb ReRAM product- MB85AS8MT [36] are some of the leading industrial endeavor on this technology. However, tradeoffs exist among key RRAM parameters, e.g., speed-retention, power-speed, endurance, retention, etc. A major challenge of RRAM is reliability and variability [37]. The switching process is not controlled microscopically and is intrinsically stochastic, which is reflected in the large variation of device resistance and switching voltage from cycle to cycle and from device to device [21]. Also, the memory cells might require large forming voltage which might not be supported by scaled access transistors.

- *Ferroelectric RAMs (FERAMs)* are ferroelectric (FE) capacitor-based memories which have been proposed and industrially implemented using the unique property of FE's polarization retention in the absence of an external electric field Fig. 1.5(a) [38]. The memory follows a 1T-1C architecture, where the binary states are encoded in the polarization state of the ferroelectric as shown in Fig. 1.5 (b). They offer high endurance along with high integration densities close to DRAMs. Texas Instrument's FRAM microcontrollers [39] and Cypress's ExcelonTM FRAM [40] are examples of state-of-the-art efforts to leverage the FE properties for memory applications. However, their read operation is destructive and requires a write back operation, leading to large energy overheads [41]. Also, they employ voltage-based sensing, whose speed is limited by the bit line/ plate line capacitance and the low capacitance distinguishability between their bi-stable states [31]

- *Ferroelectric Transistors (FEFETs)* are field-effect transistor that can serve as a form of non-volatile memory (Fig. 1.5(c)) [42], [43]. The device includes a FE layer being integrated in the gate stack of a transistor. The resistance state of the device is defined based on the polarization stored in the FE, which can be sensed by applying a drain-to-source voltage. Due to the polarization retention in the absence of electric field, even the resistance state of the transistor is retained. Such a device offers separation of read and write paths unlike many of the above mentioned emerging NVMs, which relaxes several design constraints. They offer a more robust sensing compared to FERAMs [31] and come with excellent CMOS compatibility and scalability (particularly those based on hafnium zirconium oxide of HZO) [44]. However, gate leakage in certain device architectures and interface variability can severely degrade the device and memory functionalities [45]. Moreover, large scale realization of this technology in memory products is yet to be achieved.

As mentioned above, every new memory technology comes with its own set of advantages and challenges. In order to realize their implementation in the modern-day memory and storage systems, there is a pressing need to harness their properties to the maximum extent. Therefore, it is important to come up with novel non-volatile memory design solutions using the unique attributes of the emerging memories to counter the limitation with the existing technologies. In this dissertation, we propose devices and circuits based on ferroelectric and spintronic technologies and explore their implications for different classes of applications. Note, the use of emerging NVMs might be necessary but not sufficient to over the processor-memory bottleneck in traditional computing architectures. Therefore, in addition to designing energy efficiency NVMs, there is a need for a radical shift in the computing paradigm to address this issue by enabling tight logic memory synergy, as discussed next.

1.3 Moving Logic Closer to Memory

The closer integration of compute and memory is another promising approach to alleviate the processor-memory bottleneck and reduce the frequent and inefficient memory accesses, enabling substantial improvement in system performance and energy [46]–[51]. The idea is partly inspired from the human brain where logic and memory elements are synergistically integrated with one-another to perform computations in a seamless fashion (Fig. 1.6) [52], [53]. Existing

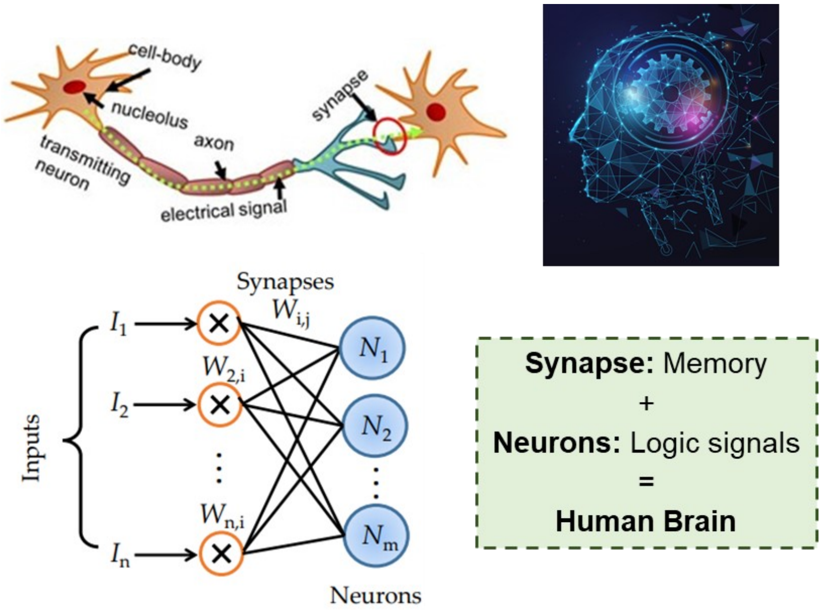


Fig. 1.6 Brain-inspired computing involving closer integration of logic and memory.

efforts to achieve this may be classified into two broad categories based on the degree of integration between logic and memory: (a) Near-Memory Computing [54]–[57] which involves logic placed right next to the memory, e.g., within the same package and (b) Computing-in-Memory (CiM) [58]–[66] which refers to performing computations within a memory array itself. Although both near- and in-memory computing alleviate the processor-memory bottleneck, the latter blurs the distinction between computation and storage. In-memory computing involves the computation of certain tasks to be in the memory itself, which is organized as a computational memory unit. Such an approach reduces the number of memory accesses and the amount of data transferred between processor and memory, and exploits the wider internal bandwidth available within the memory sub-systems to achieve high computing performance and efficiency beyond traditional von-Neumann systems. As discussed in detail later, in-memory computing can be achieved by exploiting in tandem the physical attributes of the memory devices, their array organization, peripheral circuitry and control logic. The concept of integrating the computational and storage functions of the chip on one unit was proposed as early as 1969 [56]. However, benefited from the monetization of Moore's Law and the convenience of separate design of memory and calculator, people paid no attention beyond the von Neumann structure in those days. Only until recently, in-memory logic operations [[58]–[60], [64], [66]] and vector-matrix multiplication targeted for artificial intelligence and machine learning workloads [67]–[76], have demonstrated the potential for improved power/time efficiency. There is a large body of work which involves the development of different schemes to enable in-memory computing for current and future technologies. In this dissertation, we focus on in-memory computing using the proposed memory technologies, for both general purpose and application specific architectures considering arithmetic, Boolean and non-Boolean computations.

Another aspect discussed as part of this dissertation is the enablement of non-volatile computing especially for systems which have severe energy constraints. Non-volatile computing is extremely essential for the emerging IoT and edge devices which are resource-constrained and are required to perform a wide spectrum of data-intensive tasks [77]–[79]. In this computing paradigm, the memory is brought closer to the logic unlike what was discussed earlier for CiM, where logic is brought closer to the memory [80]–[85]. Such an approach enables local backup of data within the processing elements, resulting in the design of non-volatile logic. This mitigates the long-distance data transmission overheads and the complexities of moving computed elements from the volatile processor/registers/caches to the non-volatile storage system.

1.4 Motivation for this Work

The well-established CMOS transistors are volatile, i.e., they lose their information (in terms of their resistance; ON/OFF state) when the power supply and hence, the corresponding gate voltage is removed. In order to preserve the information computed by the logic circuits (information computed by a set of transistors), there arises a need to store the information in a separate memory storage system. The problem with the performance bottleneck as technology scales is because of the inefficient data movement between the processor (for logic computation) and memory (for non-volatile data storage). But what if we can eliminate the requirement for data shuttling between the logic and memory? Can we integrate non-volatility into a transistor to overcome the Von-Neumann bottleneck?

Logic-Memory synergy is the key to alleviate the aforementioned problems with traditional and some of the emerging device technologies. If non-volatility can be embedded into the transistor, then several new avenues for circuit design will open for energy-efficient computing. To that end, in this work, two variants of integrated non-volatile transistor technologies have been proposed whose intriguing features are enabled by harnessing the potentials of the emerging technology of (a) Ferroelectric transistors and (b) Valley-coupled spintronic devices.

The recent discovery of ferroelectricity in hafnium-based oxides has led to the possibility of direct integration of ferroelectric material in the gate stack of a transistor [44], [86]. It has been shown that, by changing the capacitance matching between the ferroelectric and the underlying transistor, for instance by varying the FE thickness (T_{FE}), one can operate the FEFETs between non-volatile or volatile mode [31], [87]. Note that, for either of the operation modes, the device structure remains the same. However, if run time modulation between the logic and memory modes can be achieved, a true logic-memory synergy element can be achieved which can alleviate the above-mentioned issues associated with memory wall problem, technology scaling, big data computing, non-volatile computing, etc.

Another interesting emerging transistor technology is based on two-dimensional Transition Metal Dichalcogenides (TMDs) [88]. As an example, monolayer WSe_2 based transistor exhibits a unique attribute of the generation of out-of-plane spin currents due to the recently discovered Valley-coupled Spin hall effect [89]–[94]. They also showcase higher charge to spin conversion efficiency compared to the traditional Giant Spin Hall effect-based devices [93]. If the unique spintronic properties of these TMD transistors can be effectively harnessed, then logic-memory synergy can be enabled with non-volatility integrated in close proximity with the transistor.

The integrated non-volatile transistors can enable us to explore new computing paradigms for data-intensive applications. Computing-in-memory is an attractive technique to eliminate the memory bottleneck and thrust the future generation of machine learning workloads. On the other hand, non-volatile computing with the help of non-volatile logic can benefit energy-constrained systems with in-situ backup of the processing elements. Therefore, general purpose and application specific circuits and architectures based on emerging memory technologies are needed which can benefit from computing-in-memory and non-volatile computing to drastically push the performance and energy efficiency of electronic systems in the modern technological era. The key contributions of this dissertation are as follows:

- We propose a novel Reconfigurable ferroelectric transistor (R-FEFET) which can dynamically modulate its operation between the logic and memory modes. We explain the possible device geometry of the proposed device using both FinFET and planar technologies. We explain the physics of device operation and the fundamental reason for the reconfigurability. Such inherent device level reconfigurability opens several new avenues for circuit designs.

- Utilizing the unique device level properties of the proposed R-FEFET, we propose energy-efficient non-volatile memory designs. We discuss the memory operation, stability margins, retention and endurance properties, variation analysis, etc. and conduct a energy/performance comparison with the traditional FEFET based memory designs based on read-write-area metrics.
- We also propose two variants of R-FEFET based non-volatile flip-flops where memory is brought closer to the logic for the design of non-volatile computing systems. RNVFF-1 is designed with a completely automatic backup, without the need of any external circuitry, which is enabled by the true embedding of the proposed R-FEFET in a flip-flop and RNVFF-2 involved on-demand backup with a compact and energy-efficient external module. We perform a comprehensive circuit level analysis and understand the implication of the proposed NVMs and NVFFs when implemented in a state-of-the-art intermittently powered platform which performs non-volatile computing.
- We propose another transistor technology where intrinsic logic-memory coupling is achieved by coupling nano-magnets with 2D TMD transistors. We utilize the Valley-coupled-Spin hall effect in these devices to energy-efficiently manipulate the magnetization switching dynamics of the magnet. We propose two flavors of non-volatile memory design and compare its circuit and system-level performance with the traditional giant spin-hall effect-based devices.
- We propose computing-in-memory fabrics for Boolean and Arithmetic operations using the proposed emerging memory devices to alleviate the processor-memory bottleneck in traditional von-Neumann architectures. We explore the design of enhanced sense amplifier and compute modules to enable rich logic functionalities. We evaluate their energy and performance by implementing them for general purpose systems as well as application specific platforms such as intermittently powered systems.
- We also propose low-precision artificial intelligence hardware for accelerating deep neural networks (DNN) with the design of novel ternary compute-enabled memory fabrics based on the proposed memories. We propose novel encoding schemes and generic designs applicable

for a broad range of memory technologies. We evaluate their benefits and trade-offs with respect to near-memory DNN accelerators.

1.5 Organization of the Dissertation

This dissertation is organized as follows: Chapter-2 familiarizes the reader with the basics of ferroelectric transistor technologies. It describes the device structure and explains the two possible device variants achieved by the coupling of ferroelectric with the transistor. It also discusses the advantages and drawbacks of standard FEFETs. Chapter-3 describes the device structure and operation of the proposed reconfigurable ferroelectric transistor with built-in logic-memory synergy. The device design, analysis and influence of gate leakage is also discussed in this chapter. Chapter-4 introduces the non-volatile memory designs based on R-FEFET and their performance-energy evaluation. Non-volatile flip-flops based on R-FEFETs are proposed in Chapter-5. This chapter also investigates the implementation of the proposed flip-flops and memories in a state-of-the-art intermittently powered platform. In Chapter-6, we propose computing-in-memory fabrics based on R-FEFET and FEFET, with the ability to perform Boolean and arithmetic operations for general purpose and application specific workloads. Artificial intelligence hardware with ternary precision in-memory computing capability is proposed in Chapter-7. Chapter-8 introduces to another emerging transistor technology called 2D TMD based spin devices which enable inherent coupling of logic and memory. We propose two variants of non-volatile memory array designs and compute enabled arrays for performing Boolean and non-Boolean computing. Chapter-9 concludes this dissertation and provides possible future directions for enhancing next-generation computing using the proposed devices, circuits and systems.

2. BACKGROUND TO FERROELECTRIC TRANSISTORS

2.1 Introduction

Ferroelectrics are ideal for low power digital information storage since they can be switched purely using electric field (with significantly lower current consumption compared to current-driven NVMs [95]–[97]) and at the same time are non-volatile. Ferroelectric materials are characterized by at least two stable polarization states at zero electrical field that can be switched from one value to the other by applying an electrical field that is larger than the coercive field, E_C [98]. The coercive field, E_C is the field at which the effective ferroelectric polarization is zero. Broadly classifying, three generations of ferroelectric memories have been worked on, starting from 1950s until now (Fig. 2.1).

The first generation of ferroelectric memories were based on barium titanate (BaTiO3) ferroelectric single crystal, proposed back in the 1950s [99]. The memory arrays were built with 1C memory cells in a cross-point fashion. This approach resulted in memory matrices that were not yet integrated with the selector devices and had very low densities. These memory devices were also highly unstable and had a very short life time. Moreover, this simple array of cross-point capacitors suffered severe disturb problems, i.e., reading or writing one cell influenced the memory state of neighboring cells. To a certain extent, this is an issue similar to what is observed for resistive memories which necessitates an additional selector element. Due to the above-mentioned issues, the work on first generation ferroelectric memories was discontinued without reaching commercialization.

With the vast advancements in DRAM technology by 1990s, the interest in ferroelectric memories revived because 1T-1C architecture could be fabricated using ferroelectric capacitors which also offer non-volatility unlike DRAMs. This development led to the second generation of ferroelectric memories with the first commercial products released in the early 1990s [100], [101]. Polycrystalline lead zirconium titanate (PZT) was the ferroelectric material of choice used in these memory devices. PZT exhibited lower processing temperatures, improved reliability and easy CMOS integration compared to other ferroelectric materials available back then such as Strontium bismuth tantalate (SBT). Its operation differs from the DRAM cell and an additional plate-line has to be used for proper functionality. In order to write a ‘0’ or ‘1’ into the capacitor either bit-line or plate-line are driven to a positive potential, while keeping the other line on GND before activating the access transistor via the word-line. The read operation is performed by first pre-charging the bit-line to GND, activating the word-line and applying a positive pulse to the plate-line. Depending on the polarization state of the ferroelectric, either a linear dielectric response is sensed (no polarization switching) or a sharp current-spike response is sensed (polarization switching from -P → +P). Therefore, the read operation is destructive and a write-back is required unlike DRAM (which, in contrast, requires timely refresh). Also, the read signal generated by a ferroelectric capacitor is limited by the available polarization charge [41]. Traditionally DRAMs have used 3-dimentional capacitors with technology scaling for increasing charge for sensing. However, doing the same for ferroelectric materials like PZT has proven to be an extremely challenging due to its complex composition. Consequently, the development work was deprioritized. Therefore, with the 130 nm technology that is in production today [102], [103], the second generation of ferroelectric memories is already close to its scaling limit and will stay limited to niche applications.

One alternative would be to integrate the ferroelectric into the gate stack of a metal-insulator-semiconductor (MIS) transistor. The polarization state could then be read using the drain current of the transistor rather than sensing the switched charge directly, leading to a non-destructive read operation. This concept, the third and current generation of ferroelectric memory is called the ferroelectric field effect transistor (FeFET), which actually dates back to the late 1950s [104], [105]. However, common ferroelectrics like PZT and SBT are incompatible with CMOS process, which led to the decay of the concept. Moreover, they cannot be scaled below 130nm node as also observed in the second generation of memories due to low coercive fields and loss in ferroelectricity. In 2012, first reports indicated that hafnium oxide, which is a standard dielectric

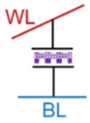
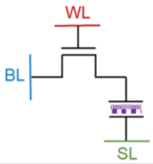
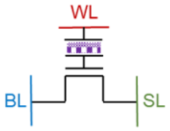
	1st Generation	2nd Generation	3rd Generation
Timeline	1950s	1990s	2000s
Material	$BaTiO_3$	$Pb[Zr_xTi1-_x]O_3$	$Hf_XZr_{1-X}O_2$
Cell Architecture	1C Cross-point 	1T-1C 2T-2C 	1T, 2T 3T, 4T 
Scalability	>130nm	Planar: 90-130nm	Planar and FinFET<28nm
Status	Prototypes	Commercial Products	Research and Development

Fig. 2.1 Three generations of ferroelectric memory technologies

material in modern CMOS processes, can be transformed into a ferroelectric phase. Moreover, the specific properties such as lower permittivity compared to classical perovskite-based ferroelectrics and high coercive fields enabled the realization of scaled ferroelectric field effect transistors that show non-volatile retention of the resistance state of the transistor [98]. Therefore, ferroelectric hafnium oxide can help to finally fully exploit the potential of ferroelectric memories. In this chapter, we discuss the background of FEFET device design and operation followed by their existing application in non-volatile circuits and systems for energy-efficient storage and computing.

2.2 Ferroelectric Field Effect Transistors - FEFETs

2.2.1 Device design

Ferroelectric field effect transistors (FEFETs) are structurally similar to a regular MOSFET or FinFET, with an additional ferroelectric and an optional metal layer integrated in the gate stack [44]. The device structure with and without the metal layer between the FE and dielectric (DE) for a FinFET-based architecture is shown in Fig. 2.2(a, b). (The benefits and drawbacks of both these device structures are discussed later in Section-2.3). The unique non-linear capacitance of ferroelectric [31], [106]–[108] interacts with the underlying MOS capacitance (CMOS) resulting in distinctive characteristics as described later. Previously investigated ferroelectric materials such as PZT for FERAMs tend to show incompatibility with the CMOS process flow, for its direct integration [41]. In 2011, ferroelectric behavior in doped hafnium oxide (HZO) was discovered for the first time [44]. Since hafnium oxide was already used as the high-K dielectric in scaled transistors, integrating FE in the gate stack became CMOS process flow friendly. Soon, preliminary works were published which could verify the functionality of HZO-based ferroelectric

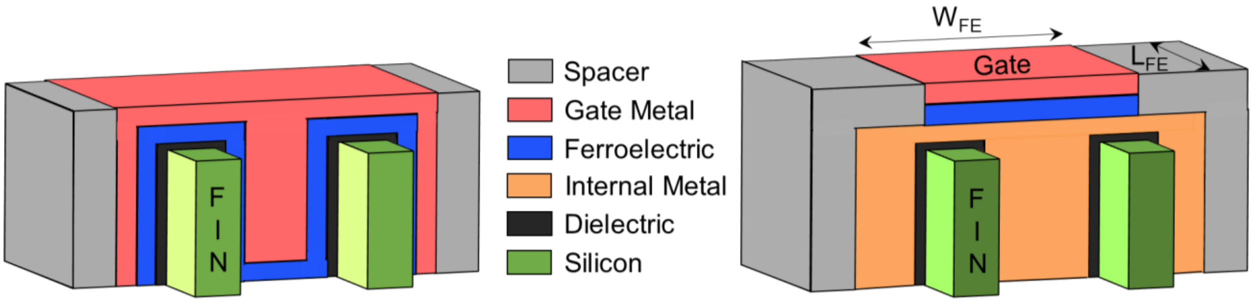


Fig. 2.2 FEFET devices based on FinFET architecture: (a) without an internal metal layer (IML) between FE and DE, (b) with IML between FE and DE.

field effect transistors (FEFETs) [44]. The FE layer can be employed with both n-type and p-type transistors. Hence, depending on the circuit requirements in terms of employing the FEFET in the pull-down or pull-up network of the circuit, one or the other type of FEFET can be used. Over the years, rigorous research on ferroelectric materials, has led to the possibilities of achieving ferroelectric properties at thickness < 2nm, resulting in realization of FEFETs even at scaled technologies, such as FinFETs [109], [110]. The high compatibility of HZO with CMOS processes has mitigated concerns regarding large-scale demonstrations of FEFETs at scaled technologies, which might have impeded industrial-scale realizations earlier. In the following sub-section, let us look into the unique device characteristics exhibited by FEFETs.

2.2.2 Device operation

(a) Steep-switching operation: Negative-capacitance field-effect transistors (NC-FETs) are one of a number of recently developed steep-slope transistor technologies using ferroelectrics, which is being explored for lower-power electronics. This steep switching mode of FEFET was proposed conceptually in 2008 by Salahuddin et al. [111]. It was envisioned that the ferroelectric layer present in the gate stack of transistor follows an ‘S curve’ trajectory for polarization switching. The ‘snap back’ of the ‘S curve’ exhibits a negative permittivity, and thus a corresponding negative capacitance (dQ/dV) is also obtained (due to the negative slope as shown in Fig. 2.3 (a)) [111]. In reality, this NC-effect is difficult to measure directly as it corresponds to an unstable region of operation for ferroelectrics. However, with the capacitive coupling of the underlying transistor, it has been argued that this negative capacitance of ferroelectric can be stabilized [112]. This

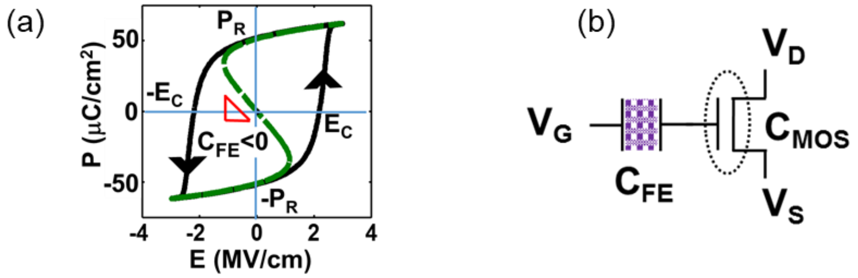


Fig. 2.3 (a) Polarization vs applied electric field of FE capacitor illustrating the negative capacitance region used for steep-switching applications; figure from [108]. (b) Schematic of FE capacitor connected to the MOS capacitance of the underlying transistor.

is made possible by ensuring that the net gate capacitance, which is the series combination of ferroelectric capacitance and the underlying MOS capacitance, is positive (Fig. 2.3 (b)). Stabilizing the operation of the ferroelectric in the negative capacitance region, achieves a voltage step-up action, which reduces the sub-threshold swing below the theoretical limit of 60 mV/ decade limit, enabling low-voltage/low-power operations [111], [112]. Over the last decade, various experimental works have showcased the possibility of negative capacitance effect with steep <60mV/decade subthreshold slope in the transistor's transfer characteristics [109], [112]–[121]. Most of these works have been achieved with lower FE thickness where the FE exhibit a single stable polarization state at V_{GS}=0V. Although, the concept has been investigated intensely in recent years, the topic remains controversial both theoretically [107], [122]–[124] as well as in the context of experimental observations [125], [126]. Further, the experimental device demonstrations, whilst suggesting some steep switching, are yet to show the characteristics required for large scale deployment of NCFETs [113], [122], [125], [127]–[129], [298]. Nevertheless, the concept of NC-FET has elicited an immense interest and besides steep switching, some works suggest that if the existing reliability issues can be addressed and the polarization characteristics of the ferroelectric can be optimized, then NC-effect in ferroelectrics could help mitigate a number of undesirable short-channel effects observed in current technologies [130], [131].

(b) Non-volatile operation: Apart from the NC-effect based operation, FEFET can also be operated in a non-volatile mode with proper optimization of the ferroelectric capacitor and the underlying CMOS capacitance [31], [132]. In this mode, the ferroelectric operates with

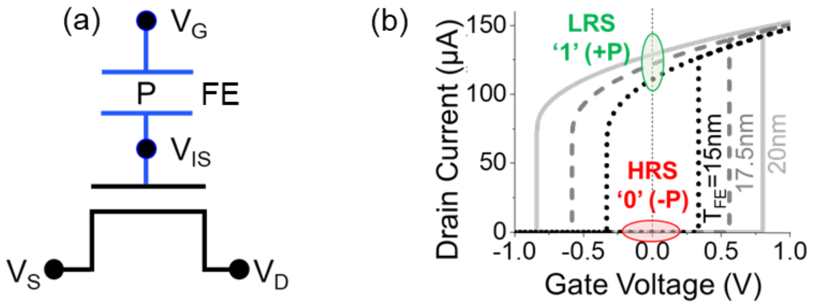


Fig. 2.4 (a) Schematic of FEFET showing the internal potential and (b) Current-Voltage characteristics of an FEFET illustrating the sensed currents, resistance states for the two bi-stable states; Drain-to-source voltage, V_{DS}=0.2V. V_{IS}<V_{TH} for -P and V_{IS}>V_{TH} for +P.

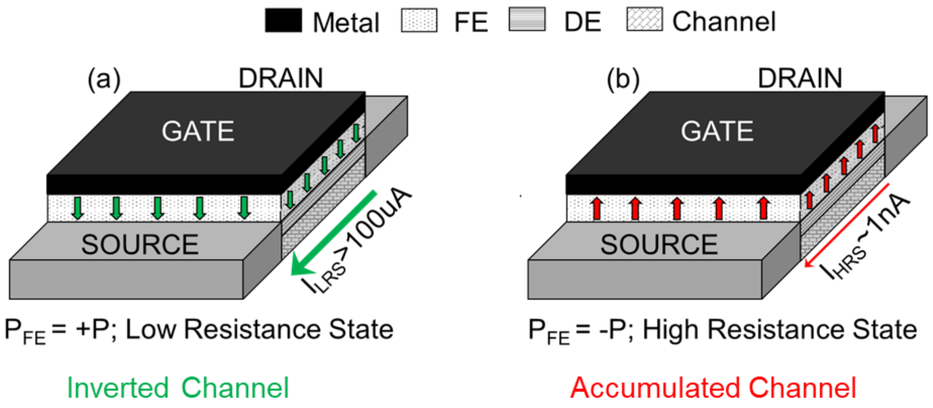


Fig. 2.5 Planar FEFET device structure showing (a) inversion caused in the channel due to positive polarization (+P) and (b) accumulation due to negative polarization (-P).

hysteretic polarization versus gate voltage characteristics spanning over positive and negative gate voltages (unlike Negative Capacitance FETs, where the FE is operated in the non-hysteretic 'S curve'/negative capacitance region). It has been shown that in the FEFET with the inter-layer metal, the interaction of the MOS capacitance (when FE is placed in series with a transistor) leads to a reduction in the effective hysteresis of the polarization versus gate voltage of the FEFET when compared to that of a standalone ferroelectric capacitor [31], [133]. Nevertheless, sufficiently thick layer of ferroelectric preserves sufficient hysteresis and bi-stable polarization states needed for non-volatility. Positive polarization in FE corresponds to a positive internal metal potential, V_{IS}, which is greater than the transistor threshold voltage (V_{TH}) on an n-channel FEFET. This causes inversion of the channel, leading to the FEFET turning ON (low resistance state/ LRS; Fig. 2.4(b) and Fig. 2.5(a)). A negative polarization corresponds to a $V_{IS} <0$ (which is also $<V_{TH}$), resulting in accumulation in the channel of the underlying FET. This corresponds to the OFF state (high resistance state/ HRS) of the FEFET as shown in Fig. 2.4(b) and Fig. 2.5(b) [86], [96], [134]. In the absence of any electric field, due to the intrinsic property of the ferroelectric material, the FE layer retains its polarization [98]. Polarization retention directly corresponds to the resistance state of the FEFET also being retained, leading to built-in non-volatility in the transistor. This non-volatile behavior of FEFETs is extremely appealing for designing non-volatile logic and memory circuits, which can potentially eliminate the memory-bottleneck discussed before in the context of traditional computing systems. In the

following sub-section, we present the existing and ongoing efforts in the design of circuits and systems utilizing the unique attributes of the non-volatile FEFETs.

2.3 FEFET based Circuits and Systems

2.3.1 Non-volatile memory (NVM) design

The hysteretic characteristics of FE capacitor naturally leads to its direct application in NVM designs [135], [136]. As discussed before, the primitive implementation of FE for memories was with FERAMs, where a 1T- 1C architecture was employed with coupled read-write paths (through the bit line). This design was widely implemented in the memory industry till the 130nm technology node. Scaling below 130nm node posed severe challenges as the FE itself, which was based on crystalline materials such as PZT, was difficult to design for polarization retention [98]. Apart from this, since the read operation occurs by sensing P stored in FE through capacitive coupling between the FE and bit-line, FERAMs exhibit low distinguishability between their bi-stable states. Moreover, the read is destructive [136]. Now, with the discovery of ferroelectricity in doped hafnium oxides, FEFET based memory designs have been explored to mitigate the drawbacks of FERAMs [44], [86], [97], [134]. Positive/negative polarization (+P/-P) stored in the FE layer corresponds to low/high resistance state (HRS/LRS) of the FEFET. Moreover, the read-write paths are decoupled leading to robust and disturb free read operation. Also, the currents sensed through the FEFET showcases excellent distinguishability (~$10^{5\text{-}6}$) [31], [97]. However, their write operation at iso-retention results in degradation of energy efficiency compared to FERAMs [31], [133].

In an attempt to effectively harness the FEFETs in memory arrays, several flavors of bit-cells have been proposed. The 2T memory (Fig. 2.6(b)) uses a write access transistor in addition

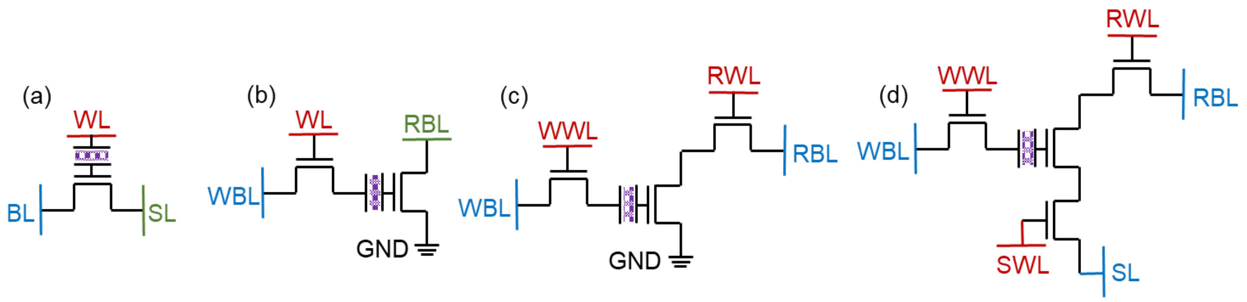


Fig. 2.6 (a) 1T, (b) 2T, (c) 3T and (d) 4T memory designs based on standard FEFETs.

to the FEFET to enable selective write in an array [31]. This memory offers excellent improvements in power and speed with respect to FERAM but requires negative write voltages, which lowers energy efficiency. Also, 2T memory needs unconventional biasing during read. 3T memory (Fig. 2.6(c)) employs an additional transistor for read access thereby reducing the complexity in the read sensing scheme. However, they face similar energy overheads due to the requirement of negative voltages [133]. A 4T memory (Fig. 2.6(d)) avoids the use of negative voltages for writing into the bit cell with the help of a fourth transistor connected to the source of FEFET [137]. However, large number of switching lines due to additional transistors make them energy inefficient. Recently, a highly dense 1T memory using a single FEFET bit cell has been proposed (Fig. 2.6(a)) [138], [139]. However, their robustness for scaled technologies with large parasitic capacitances affecting the device characteristics needs to be further evaluated [140].

The aforementioned issues of various cells may further degrade when FEFETs without IML are considered in the NVM design. This is due to charge trapping/de-trapping effects which affects the endurance and the internal depolarization fields across the FE which degrades the retention characteristics [141]. FEFETs with IML (also referred to as FEMFETs [142]) can potentially relax such concerns [142], as discussed in detail later. However, FEMFETs may suffer from GL triggered loss in the resistance state corresponding to P stored [45]. To overcome the influence of GL, a modified memory function with 2-step read operation has been proposed which establishes the functionality, albeit at the cost of access energy [45].

2.3.2 Non-volatile flip-flop (NVFF) design

Another class of circuits which have greatly benefitted by the innovation in emerging NVMs are non-volatile flip-flops. NVFFs are a crucial element in the design of non-volatile processors [77], [79], [143]. They mitigate the need for moving the data from registers all the way to storage system, by performing a local backup/restore operation in a non-volatile element. With the development of IoT and energy harvesting techniques, power supply disturbance can be frequent and such NVFFs are critical to sustain the computation progress with such non-volatile computing methodology. Fig. 2.7 (a) illustrates a conceptual non-volatile flip-flop (NVFF) where the backup-restore module attached to the storage nodes of standard flip-flop stores the information in case of a power outage. Several design with various emerging non-volatile elements have been explored

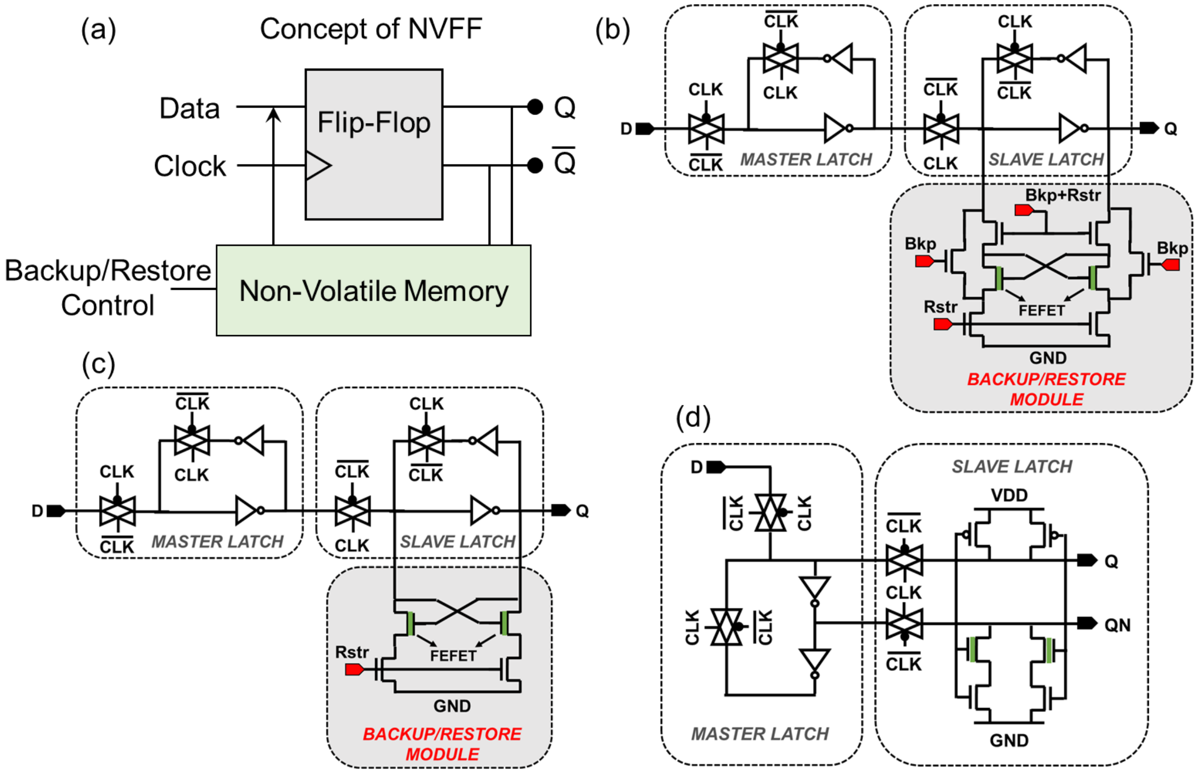


Fig. 2.7 (a) Concept with in situ NVM as the state backup storage; (b) nvDFF1 [146]; (c) nvDFF2 [145]; (d) nvDFF3 [147].

to achieve such an operation [80]–[85], [144]–[146]. Most of the designs, contain current driven memory elements such as RRAMs, MTJs, etc. which lead to high power consumption. Recently, over the last few years, there has been an interest in FEFET based NVFFs which can sustain the flip-flop state during power-off periods [80], [145]–[147] (Fig. 2.7(b-d)). Their unique property of field driven information storage (unlike current driven storage in other memories) makes them promising candidates for energy-efficient computation especially for energy-constrained intermittently powered systems [95], [96], [148].

Most of the previously proposed NVFFs based on FEFETs (as well as and other emerging non-volatile elements) consist of a back-up/restore (B/R) module which backs up the information state of the storage nodes, during a power outage and restores storage node voltages when power turns back ON [146], [147]. This improves the computation progress in systems by overcoming the von Neumann bottle neck. However, the additional backup-restore module attached to every flip-flop leads to area overheads. Moreover, since the backup-restore module is directly attached to the storage nodes, it becomes challenging to ensure minimal energy-latency overheads during the normal operation. Hence, there is a need to optimize the B/R module by exploring new device technologies that leverage the opportunities offered by FE.

2.3.3 Artificial intelligence hardware / compute-in-memory (CiM) fabric for brain-inspired computing

Computing in memory (CiM) is an emerging paradigm explored to eliminate the von-Neumann bottleneck (or memory-bottleneck) in traditional computing systems (Fig. 2.8). Recent advances in memory design enable the opportunity for architects to avoid costly data movement by performing CiM. The idea of CiM has been proposed for at least four decades [54]–[57], but earlier efforts were not widely adopted due to the difficulty of integrating processing elements for computation with the main memory, DRAM. Innovations such as three-dimensional (3-D)-

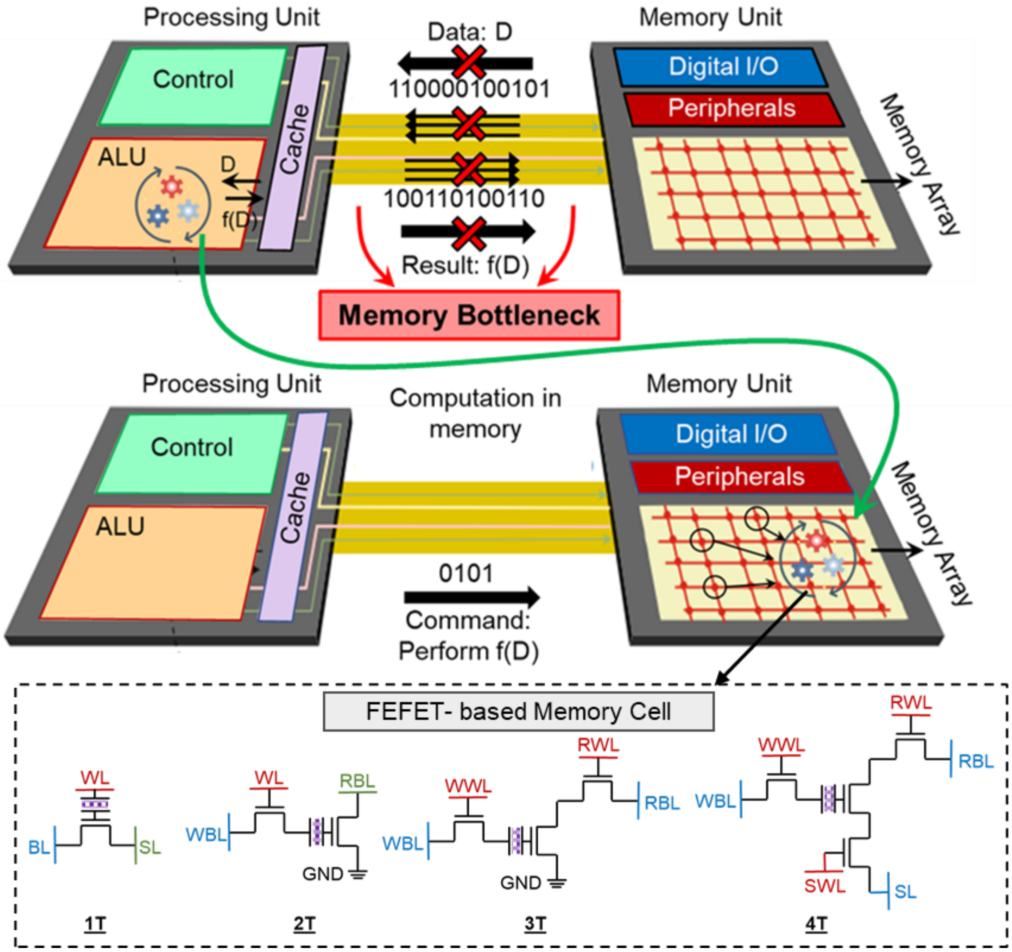


Fig. 2.8 Concept of computing-in-memory using ferroelectric memory technologies. Adapted from [156]

stacked memory dies that combine a logic layer with DRAM layers [149]–[152], the ability to perform logic operations using memory cells themselves inside a memory chip [58], [62], [66]–[68], and the emergence of potentially more computation-friendly emerging NVM technologies [49], [50], [59], [64], [153]–[156] provide new opportunities to embed general purpose computation directly within the memory. FEFETs in particular show an immense promise in this regard, due to the inherent logic-memory synergy at the device level, with non-volatility integrated into a transistor.

Various FEFET based NVM fabrics have been proposed with the capability of CiM operations. [64] proposes FEFET based CiM where computations are enabled with a modified sense amplifier. This design utilizes both voltage and current based sensing techniques to achieve in-memory operations for Boolean and Arithmetic computations. Although intriguing, the design in [64] faces possible challenges with respect to the write disturbs in the unaccessed cells in a column. Moreover, the write operation is two-phased and requires dual bias applied to several metal lines of an accessed word. This leads to significant energy overheads for current generation of workloads. Lastly, the use of both current and voltage-based computing schemes leads to design complexity and energy inefficiency mainly due to the constant DC reference current flowing during sensing operations.

FEFET based CiM designs have also been proposed to perform neuromorphic computation [142], [157]–[162], using multi-domain switching in FEFETs. For ferroelectric with multiple domains, the partial polarization switching - where different portions of domain distribution is switched - can be harnessed to realize multiple intermediate polarization states. Those polarization states will result in different output conductance in FEFET, which can serve as synaptic weight cell to store the neural network weights. Investigation in [161] suggests that by arranging FEFETs in a pseudo-crossbar architecture, in-memory multiply-and-accumulate operation (which constitutes >90% of operations in modern DNN workloads) can be performed in the analog domain, obviating the costly data movement in the conventional von-Neumann systems. Ferroelectric based neurons have also been realized using quasi-leaky integrate and fire approach, where the leakage rate is determined by the polarization of the ferroelectric [162]. These innovations also enable the possibility of realizing an all ferroelectric based biologically inspired spiking neural network [161]. Although the concept of utilizing multi-level states of FE is very attractive for high density and low energy implementations of deep neural networks (DNNs), experimental studies have shown the degradation in stability of the analog states with geometric scaling [163]–[165] due to the challenges with domain control and variability. Therefore, the application of such technique requires further exploration.

Other types of in-memory computing systems have been proposed using FEFETs, for example one that harnesses the dynamics of polarization switching for statistical correlation detection, which is widely used in signal processing and event detection [166]. Ferroelectric ternary content addressable memory (TCAM) is another example, where computations are directly performed in the memory [167], [168]. Although TCAMs have been widely realized using CMOS based SRAMs (16T cell design), FEFET based approach leads to ultra-high density with 2T cells.

To conclude, FEFETs showcase enormous potential for the implementation of the energy/performance efficient CiM architectures for various classes of applications. FEFET-based CiM are being investigated to perform Boolean, non-Boolean and arithmetic operation and mitigate the challenge of memory bottleneck in current generation of electronic systems.

2.4 Drawbacks of Existing Approaches

2.4.1 Requirement of negative voltages for polarization switching

One of the major limitations of some of the previously-proposed FEFETs based circuits is the requirement of negative voltages for storing negative polarization in the FE layer (‘0’ bit). This leads to higher energy consumption for example in non-volatile memory operations (as discussed in Chapter-4). Moreover, generation of negative voltage requires additional bias circuits which increases the design complexity. Recently, techniques such as drain-erase biasing scheme has been proposed where positive voltage at both drain and source terminals (with gate at 0V) are applied to store bit-‘0’ in the FEFETs (Fig. 2.9 (a-d)). However, in such a scheme, charging of all the unaccessed metal lines (connected in a cross-point fashion) and the requirement of multiple voltages lead to energy overheads and design complexities. Another approach involves the use of 2-phase biasing to overcome the use of negative voltage but comes with performance penalty. Fig. 2.9(e-h) illustrates and example of 2-phase scheme using 3T FEFET based memory. Therefore, to effectively tap the qualities offered by FE based technology, there is a need for novel device and circuit designs which can mitigate the requirement of negative voltage at minimal or zero penalty.

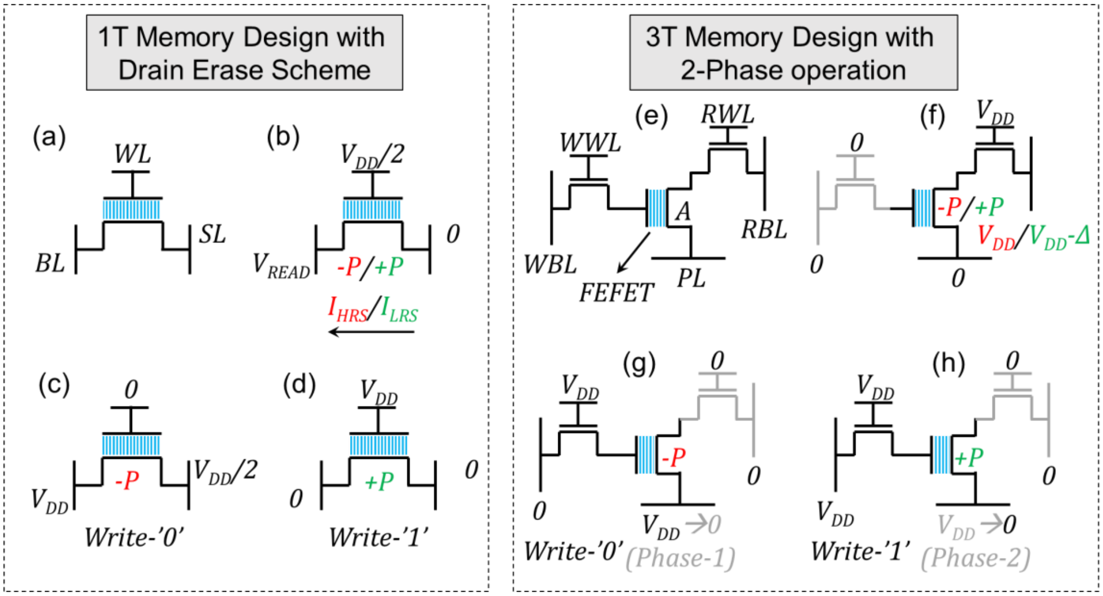


Fig. 2.9 Techniques to mitigate the use of negative bias for polarization switching: (a-d) Drain-erase bias scheme used for 1T memory cells placed in a cross-point fashion and (e-h) example of 2-phase scheme used in a 3T memory cell to write bit- '0' and '1' in different phases.

2.4.2 Retention and endurance tradeoffs

As mentioned before, FEFETs can be realized with or without the internal metal layer (IML) in between the FE and DE layers. However, both the device architectures come with their own drawbacks. In FEFETs without IML, the electrostatic interactions lead to constant depolarization fields (E_{DEP}) directed opposite to the polarization stored in the ferroelectric layer (Fig. 2.10), degrading the retention characteristics [123], [142], [169]. At the same time, trapping/de-trapping effects at the ferroelectric-dielectric interface deteriorates the endurance of these devices [141], [142]. On the other hand, FEFETs with IML (also referred to as FEMFETs) has an advantage that the cross-sectional area of FE and the underlying FET can be independently designed, which helps

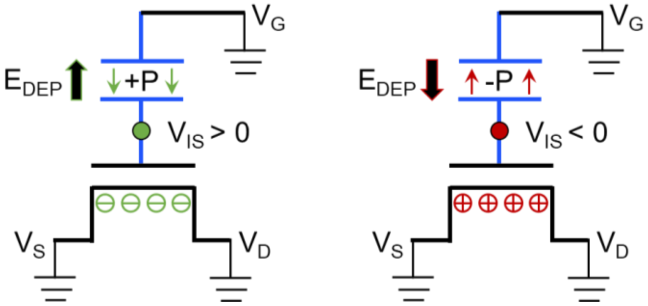


Fig. 2.10 Depolarization issues in FEFET/FEMFET due to $V_{IS} \neq 0V$.

in achieving a desired memory window, polarization switching at scaled voltages and lower depolarization fields compared to FEFETs without IML [142]. They also alleviate the issues corresponding to trapping and de-trapping effects, due to the presence of the intermediate metal layer [142]. However, since the IML is floating, its potential can discharge over time due to GL (Fig. 2.11 (a)), as discussed later. Therefore, GL can cause degradation of resistance-based distinguishability in FEMFETs. Experimental studies have shown (2~3hrs) of retention for FE-Metal-FET (FEMFET) [142], while ~10 years for FEFET (without IML). On the other hand, endurance of ~10^7 cycles have been reported for HZO based FEFETs (without IML) [170], while FEMFETs (FEFETs with IML) showcase higher endurance of ~10^{11} cycles [142].

2.4.3 Write voltage Scalability

A majority of the existing works on FEFETs without IML have shown the requirement of large write voltages (~ ±3V-4 V). This is because of the field distribution across the gate stack being non-uniform, with most of the voltage dropping across the oxide plus semiconductor channel [142]. This increases the write voltage for complete polarization switching in the FE. This challenge has been overcome by FEFETs with IML (FEMFETs) because now, the FE and transistor can be independently optimized with different geometrical aspect ratios [142]. Such an approach can help in achieving maximum of the applied voltage dropping across the FE, which in turn reduces the write voltages for polarization switching (~ ±1V-2V). This makes FEMFETs more favorable and appealing for CMOS integration in scaled technologies.

2.4.4 Gate leakage

Note that, GL through the FE and DE layers in FEFETs with the floating IML, results in discharge of V_{IS} over time as shown in Fig. 2.11 (a) and discussed in detail in [45]. If GL can be controlled, sensing the bit-information is similar to FEFETs without IML. However, if GL is significant, it can have detrimental effects, as the resistance-based distinguishability (determined by IML potential) can degrade and eventually disappear with time. Internal metal potential, V_{IS} discharge due to GL results in loss of distinguishability between positive polarization (+P; ON state) and negative polarization (-P; OFF state). GL also results in polarization-dependent shifting of device characteristics with time which reduces the design margins [45], [171]. However, the

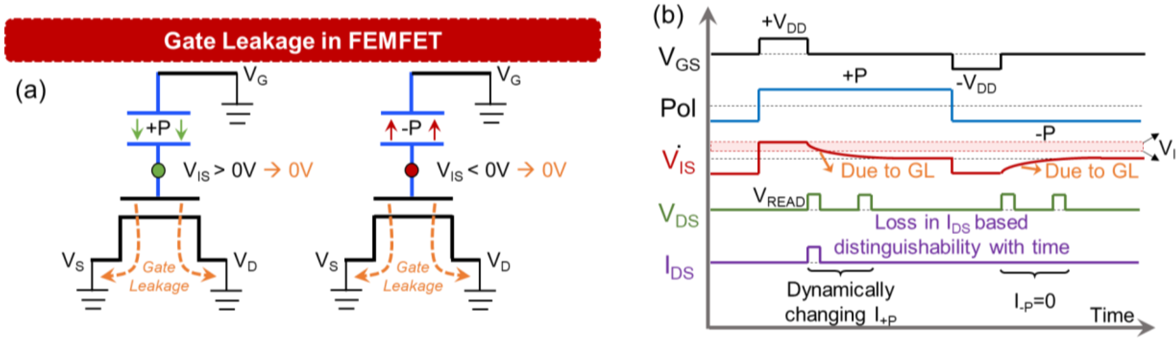


Fig. 2.11 (a) Gate leakage in FEMFET due to floating IML resulting in (b) loss in resistance based/ drain current based distinguishability.

polarization is still retained which stores the bit information (Fig. 2.11(b)). Therefore, it is important to note that the bit-information stored as polarization is still intact and is not lost due to GL. However, sensing this polarization is non-trivial, incurring energy/design overheads [45]. Note, a positive by-product of IML potential going to 0V due to GL is that the depolarization field E_{DEP} goes to zero over time and the polarization retention can be superior to FEFETs (without IML) [169].

Now, as discussed in this chapter, all the previous experimental and theoretical studies have shown volatile (i.e., NCFET) or non-volatile characteristics for FEFET by employing static/design time optimizations, for instance, low T_{FE} show promise for NCFET-logics while high T_{FE} is more suitable for FEFET-NVMs [86], [172]. However, if run-time tuning of non-volatile and volatile modes can be achieved, several new avenues for circuit and system design will open.

2.5 Summary

We introduced ferroelectric transistors, an emerging device technology which has gained immense interest in recent years. We discussed the two variants of device design, i.e., with and without an internal metal layer between FE and DE layers. We explained the two possible modes of device operation: (a) steep-switching and (b) non-volatile, which is determined by the capacitance matching of FE and the underlying transistor. We mentioned the different flavors of existing non-volatile memory and flip-flop designs, implemented using the unique characteristics of FEFETs. Previously proposed computing-in-memory designs using FEFETs to overcome the von-Neumann bottleneck and perform Boolean, non-Boolean and arithmetic operations were also mentioned. Finally, we discussed the drawbacks associated with FEFET based circuits in regards to the requirement of negative voltages, retention and endurance tradeoffs, scalability of write voltages and gate leakage’s negative influence on circuit operation.

3. RECONFIGURABLE FERROELECTRIC TRANSISTOR -- R-FEFET: A NON-VOLATILE MEMORY DEVICE WITH DYNAMICALLY TUNABLE HYSTERESIS

3.1 Introduction

As discussed in the previous chapter, experimental studies on FEFETs have shown volatile or non-volatile characteristics by employing design time device optimizations, for instance by modifying the thickness of the FE layer [31], [87]. However, if run-time tuning of non-volatile and volatile modes can be achieved, several new avenues for circuit design will open with logic-memory synergy. Moreover, such a feature can potentially mitigate the limitations of FEFETs discussed in the Chapter-2. Previously, a method was proposed to shift the FEFET device characteristics, to achieve dynamic logic programmability [173]. While this method is very appealing for such an application, it may not be suitable for reconfigurability between logic and memory modes, because of (i) low current distinguishability (~10x) between the ON-OFF states and (ii) minimal modulation of hysteresis, requiring large voltages for logic operations. To address the need for efficient logic-memory coupling, we propose a reconfigurable FEFET (R-FEFET) which has the capability to dynamically modulate its hysteresis and tune its operation between volatile and non-volatile modes. In addition to such a unique reconfigurability, R-FEFETs can overcome the disadvantages of FEFETs due to gate leakage (GL). In this chapter, we introduce R-FEFETs and comprehensively analyze their device characteristics. The proposed R-FEFETs are then used in several classes of circuits including, non-volatile memory non-volatile logic with data back-up capability, logic-in-memory, etc., which are discussed in the following chapters.

3.2 R-FEFET Device Structure

The proposed R-FEFET comprises of two sets of FE stacks which are regulated by the Gate and the Control terminals. Both the FE stacks are connected by an internal metal layer (IML), which also serves as the gate of the underlying transistor (Fig. 3.1(a, b)). The cross section of one of the fins of the R-FEFET based on FinFET is shown in Fig. 3.1(c), illustrating the gate stack with FE and IML. The equivalent schematic of R-FEFET is shown in Fig. 3.1(d) where the IML of the R-FEFET is shown to be connected with Gate and Control terminals through their respective

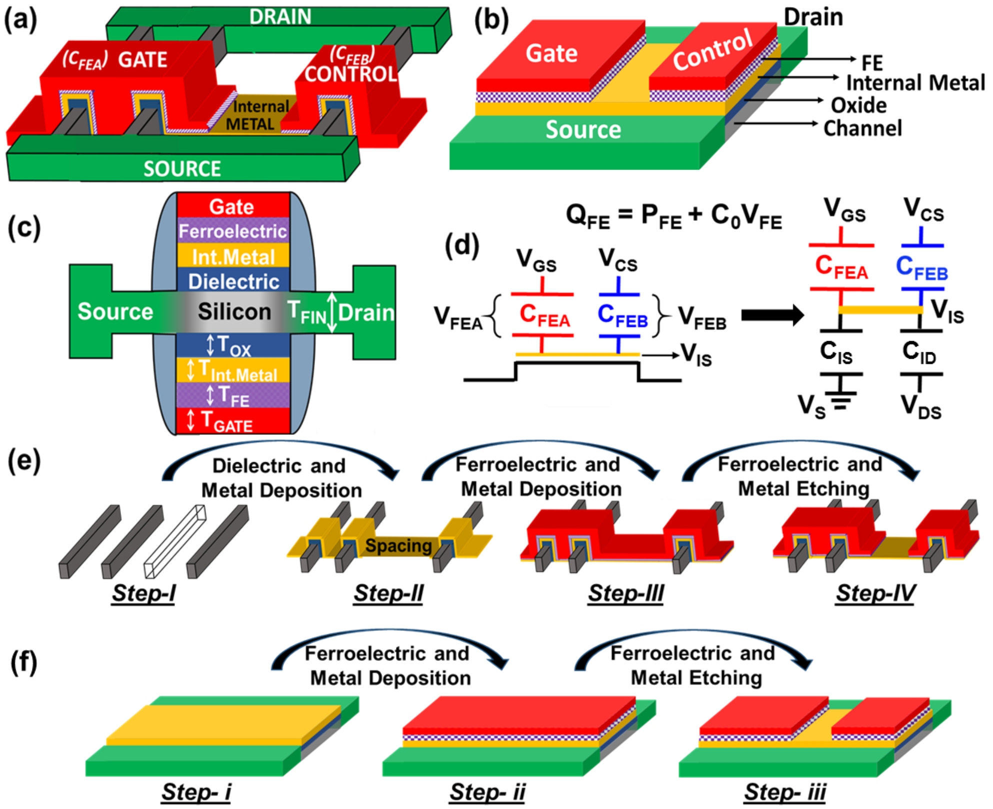


Fig. 3.1 (a) Proposed R-FEFET with 3 fins (b) R-FEFET realized using planar FET (c) Cross Section of (a). (d) Schematic and capacitance network of R-FEFET. Process flow for R-FEFET based on (e) FinFET and (f) planar structures. Note, spacing between FE_A and FE_B can be changed according to design rules of technology node.

FE capacitors (FE_A and FE_B). *(Note, $FE_{A/B}$ and $FE_{G/C}$ are interchangeably used in this dissertation)*. The Gate terminal (controlling FE_A) acts as a regular gate of the transistor, while the Control terminal (controlling FE_B) dictates the mode of operation, as explained later. Note, both the stacks have equal T_{FE}. The Gate and Control signals interact via FE_A and FE_B with IML whose voltage (V_{IS}), in turn, controls the underlying FET. The proposed concept of coupling two FEs with a transistor is applicable for FinFET as well as planar technologies [171], [174]. Following are the details of transistor technologies used in this work:

3.2.1 FinFET technologies

We use 10nm FinFET models [175] with silicon (Si) as the substrate and gate length (L_G) =14nm, fin thickness=8nm, fin height=21nm for our designs. The gate oxide dielectric used in our simulations is high-k (HfO_2), in order to be consistent with the advanced technology nodes. The fabrication of R-FEFET can be achieved with small changes to the standard FinFET process flow (Fig. 3.1(e)). First, the standard process is used to obtain a FinFET with one (or two) of the fins in the middle etched (step-I; to provide spacing for a later process). Then, the regular gate stack is formed up till the IML (step-II). This is followed by FE and gate metal deposition (step-III), and a selective etch process (enabled by the spacing from step-I) to isolate the two FE stacks (step-IV). Note that the IML is kept intact during the etch process to enable coupling in-between the two FE stacks and the transistor. The structure of R-FEFET is quite different from standard FEFET or FinFETs. R-FEFET contains two gate stacks (namely, Gate and Control) unlike only one present in FEFETs or FinFETs. The intrinsic coupling between the two FE stacks and the underlying transistor leads to unique device characteristics as explained later.

Due to the sizing constraints in FinFETs (for e.g. width quantization [176]), the design flexibility with respect to the FE stacks is reduced. Also, the addition of FE and IML in the gate stack, may require an increase in the fin pitch. Therefore, as an alternate method, we can use the processing technique used for FE-Metal-FET (FEMFET) [142], where the FE is integrated in the back-end with electrical connection to the gate of the conventional FinFET. Such a device structure, where the FE sizing is decoupled from the fin geometry, can potentially overcome the aforementioned issues with the sizing of FE and also avoid etching of additional fins (for improved density). Moreover, FEMFET device architectures also exhibit inherent advantages of lower depolarization field and higher endurance (compared to FEFETs), which is explained later. In this work, we discuss and analyze circuits with both variants of FinFET based R-FEFETs discussed above.

3.2.2 Planar technologies

The proposed technique can also be employed for planar technologies in a similar manner as FinFETs. We employ 22nm planar MOSFET models [175] with Si as substrate, high-k gate oxide (HfO_2), and L_G=22nm, gate width (W) =110nm. We ensure a spacing of at least 3λ (~33nm) between the FE stacks (in accordance with the scalable CMOS rules [177]). Note, the planar devices offer much larger design flexibility due to the absence of width quantization (as in FinFETs). The process flow of planar R-FEFETs is illustrated in Fig. 3.1(f). It involves (i) fabrication of standard planar transistor (ii) deposition of FE and metal layers on IML (iii) selective etching to form two FE stacks employing flow similar to contact over active gate process [178]. Note, selective etching is performed along the width (not along the channel length).

It may be mentioned that parasitic capacitive (C_P) coupling exists between Gate and Control, which is considered in our simulations. However, our calculations show that, considering the minimum distance between Gate and Control to be 68 nm (33nm) for 10nm (22) nm node (as per the layout rules [177]), the capacitance, C_P is ~0.3aF, which is much smaller than other capacitances (~ fF). Therefore, C_P has negligible impact on the device operation and is not shown in Fig. 3.1(d) for simplicity. The ratio of the capacitances of FE_A and FE_B (C_{FEA}:C_{FEB}) is a key design knob in device operation (as discussed later).

We also design R-FEFETs at 45nm node to gain advantages of low leakage power consumption while designing circuits to enable in-memory computing paradigms. For this, we employ 45nm planar MOSFET models [175] with Si as substrate, HfO_2 as gate oxide, and L_G=45nm, gate width (W) =11λ. We ensure a spacing of at least 3λ (~33nm) between the FE stacks (in accordance with the scalable CMOS rules [177]), whose widths are identical and equal to 4λ. Identical widths of the FE stacks enable the possibility of eliminating the influence of gate leakage on memory operations as discussed later. With the understanding of the device structure, let us now discuss the modeling and simulation methodology used in this work.

3.3 Simulation Methodology

To analyze the proposed R-FEFETs in this dissertation, we employ a physics-based circuit compatible SPICE model for FE based transistors [108], based on the well-established Landau Khalatnikov (LK) theory [179]. The underlying transistor is modeled using predictive technology [175]. The two FE stacks (FE_A and FE_B) based on the time-dependent LK model are self-consistently coupled in SPICE with each other and the underlying transistor, based on the equivalent circuit shown in Fig. 3.1(d). The effects of depolarization fields due to non-ideal contacts are captured in our model [108]. To model FEFETs, we use a similar approach coupling FE and the underlying transistor, as described in detail in [108]. In this work, we perform our analysis on R-FEFETs based on FinFETs (10nm node) as well as planar (22nm node, 45nm node) technologies [175]. To calibrate our model (Fig. 3.2), we use experimental results for $Hf_{0.7}Zr_{0.3}O_2$ (remnant polarization =20μC/cm^2, coercive field =1MV/cm) [113], which has showcased CMOS

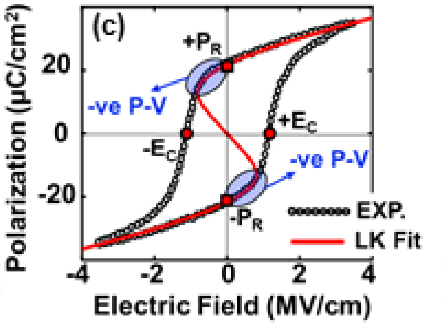


Fig. 3.2. Calibration of the model with experiments from [113].

process compatibility at scaled technologies [86], [134], [173]. The LK parameters of FE derived from calibration are: α=-0.7x10^9 m/F, β=6x10^8 $m^5/F/C^2$, γ=3x$10^{11}$$m^9/F/C^4$. The kinetic coefficient, ρ= 0.025Ω-m, was calculated for polarization switching time of hundreds of picoseconds [180]. Note, the results mentioned in this dissertation are obtained using quasi-static simulations, similar to the experimental methodology for FEFETs [134]. In the following sub-section, let us discuss the device operation and reconfigurability between non-volatile and volatile modes in the proposed R-FEFET.

3.4 Dynamic Reconfigurability in R-FEFETs

To understand the reconfigurability in R-FEFETs, let us consider the following Equation-3.1, derived from the capacitance network shown in Fig. 3.1(d).

$$A_{FEA} * \ (P_{FEA} + C_0 * V_{FEA}) + A_{FEB} * (P_{FEB} + C_0 * V_{FEB}) - (C'_{IS} * V_{IS}) - (C'_{ID} * V_{ID}) = 0 \qquad (3.1)$$

$$Where\ C' = \frac{1}{V}\int dV \ ;\ C_0 = \varepsilon_{FE} * \varepsilon_0 / T_{FE}$$

$P_{FEA\ (B)}$, $A_{FEA\ (B)}$ and $V_{FEA\ (B)}$ are the polarization, area and voltage across the $FE_{A\ (B)}$ capacitor respectively. $C'_{IS\ (D)}$ and $V_{IS\ (D)}$ are the average capacitance and potential difference between IML and source (drain) terminal of the R-FEFET. ε_0 and ε_{FE} (~20) are the permittivity of free space and dielectric constant of FE respectively. Note that the capacitances, C_{IS} and C_{ID} (Fig. 3.1(d)) include the oxide, overlap and fringe capacitances associated with the underlying transistor. Equation-3.1 can be simplified further as:

$$V_{IS} = \frac{(A * P)_{FEA}}{C'_T} + \frac{(A * P)_{FEB}}{C'_T} + \ (A_{FEA} * V_{GS} + A_{FEB} * V_{CS}) * C_0/C'_T + \ V_{DS} * C'_{ID}/C'_T \qquad (3.2)$$

$$Where\ C'_T = C'_{IS} + C'_{ID} + (A_{FEA} + A_{FEB}) * C_0$$

We will describe the device operation qualitatively with the help of this equation. Note, all the results have been obtained from proper simulations as per the model discussed above. We will describe the characteristics in context of the n-type device. This discussion can be extended to the p-type device as well.

The control terminal of the R-FEFET interacts with the regular gate to enable the dynamic reconfigurability between volatile and non-volatile modes of operation. V_{IS} is designed to be affected to a larger extent by the polarization (P) of the regular gate stack (P_{FEA}) compared to P of the control stack (P_{FEB}) by ensuring that A_{FEA} is larger than A_{FEB} (C_{FEA}>C_{FEB}). Note that because of the presence of the common IML, FE_B interacts with FE_A, thereby influencing V_{IS}, yielding unique characteristics as discussed later. In the volatile mode of the R-FEFET, P_{FEA} in the OFF state ('0') corresponds to negative polarization (-P), while the ON state ('1') corresponds to a positive polarization (+P). When switched from the volatile to the non-volatile mode, the bi-stable state of the R-FEFET is defined by P_{FEA}. We discuss the volatile and non-volatile modes in detail in the following paragraphs. For simplicity, we neglect gate leakage (GL) to illustrate the basic concept of device operation. In the subsequent sections, we consider GL and show its impact on the characteristics of R-FEFETs. We first start with the discussion on FinFET based R-FEFET, with C_{FEA}:C_{FEB} = 2:1.

3.4.1 Non-volatile ('NV') mode

For the 'NV' mode of operation, we drive the control voltage to 0 (V_{CS} = 0V). First, we consider drain voltage, V_{DS} = 0V to simplify the discussion. The effect of V_{DS} is considered subsequently. Let us begin our analysis by sweeping the gate voltage (V_{GS}) from a negative to positive value and back. To start with, application of a negative V_{GS} (-0.8V), yields a negative P_{FEA} (-P) and negative V_{IS} (Fig. 3.3(a, c); Fig. 3.4(b)). Negative V_{IS} yields a positive voltage across FE_B (V_{FEB} =V_{CS}-V_{IS} =-V_{IS}; since V_{CS}=0V). This leads to positive P_{FEB} (+P) (see Fig. 3.3(b); Fig. 3.4(b)). As V_{GS} is swept from negative to positive value (-0.8V→0.8V), P_{FEA} switches first from -P→+P when voltage across FE_A (V_{FEA}=V_{GS} –V_{IS}) exceeds the coercive voltage (V_C) of FE_A (step-1). Note, the gate voltage at which P switches is called the critical (or coercive) gate voltage ($V_{GS,C}$; point-A in Fig. 3.3(a)). P_{FEA} switching results in an increase in V_{IS} to a positive value due to capacitive

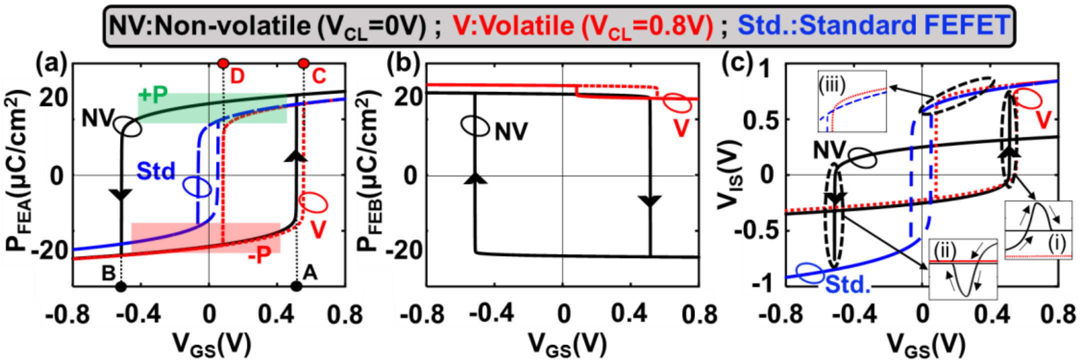
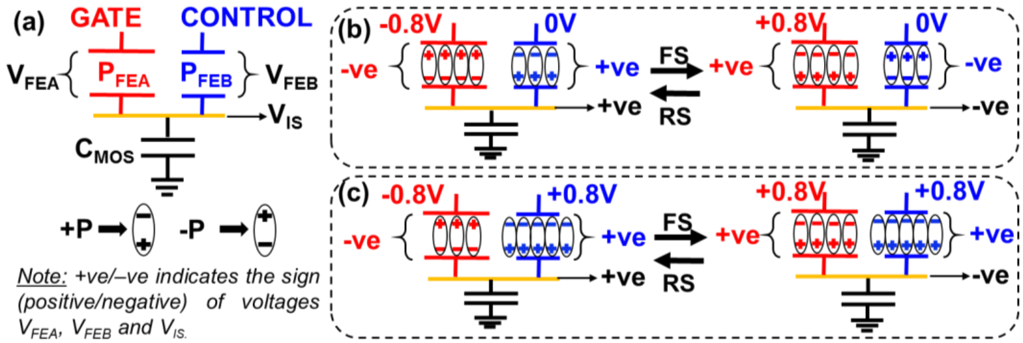


Fig. 3.3 (a) Polarization of FE_A stack, (b) Polarization of FE_B stack and (c) internal metal voltage versus gate voltage. V_{DS}=0V and T_{FE}=8nm.

coupling (step-2). Hence, V_{FEB} (= $-V_{IS}$) becomes negative which yields switching in P_{FEB} from +P→-P (Fig. 3.3(b); step-3). Consequently, switching in FE_B lowers the magnitude of V_{IS} (Fig. 3.3(c)-inset (i); step-4). Therefore, a sequential transient process (STP) from step-1 to step-4 occurs during the device operation as illustrated in Fig. 3.5. Similarly, during the reverse sweep of V_{GS} from positive to negative values (0.8V→-0.8V), P_{FEA} switches from +P→-P when V_{FEA} reaches $-V_C$ (step-1; point-B in Fig. 3.3(a)). This, in turn, yields negative V_{IS} (step-2), such that FE_B reaches its coercive voltage (V_C) for P_{FEB} switching from -P→+P (step-3). P_{FEB} switching eventually leads to decrease in the magnitude of V_{IS} (step-4; Fig. 3.3(c)-inset (ii)). Thus, in the 'NV' mode (V_{CS}=0V), P_{FEB} is always opposite to P_{FEA} (Fig. 3.3(a, b)). As a result, FE_B reduces the effect of FE_A on $|V_{IS}|$ in the 'NV' mode (due to step-4 of STP; Fig. 3.5). Hence, compared to FEFET, lower $|V_{IS}|$ is obtained in R-FEFETs for the same $|V_{GS}|$ (Fig. 3.3(c)). This increases $|V_{GS,C}|$ required for P_{FEA} switching (since $V_C = |V_{FEA}| = |V_{GS,C}–V_{IS}|$). Therefore, larger HW is observed in R-FEFETs compared to FEFETs, leading to larger V_{GS} margins for holding P (or hold margins).



Note: +ve/–ve indicates the sign (positive/negative) of voltages V_{FEA}, V_{FEB} and V_{IS}.

Fig. 3.4 (a) Schematic of R-FEFET. Device operation in (b) non-volatile mode and (c) volatile mode. FS: Forward Sweep, RS: Reverse sweep.

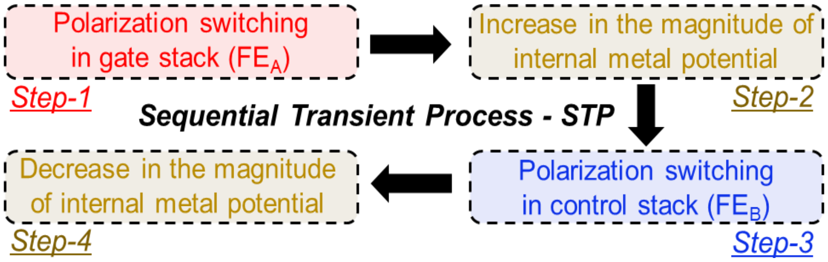

Polarization switching in gate stack (FE$_A$)
Step-1
Increase in the magnitude of internal metal potential
Step-2
Sequential Transient Process - STP
Decrease in the magnitude of internal metal potential
Step-4
Polarization switching in control stack (FE$_B$)
Step-3

Fig. 3.5 Illustration of the step-by-step sequential transient process during the non-volatile mode of operation in the proposed R-FEFETs.

To describe the device operation in the 'NV' mode further, let us consider Equation-3.2. Since V_{CS} =0V and considering V_{DS} =0V, the terms corresponding to V_{CS} and V_{DS} vanish and we obtain the following expressions for V_{FEA} and V_{FEB}:

$$V_{FEA} = V_{GS} - V_{IS} = -(A*P)_{FEA}/C'_T - (A*P)_{FEB}/C'_T + \left(1 - A_{FEA} * \frac{C_0}{C'_T}\right) * V_{GS} \qquad (3.3)$$

$$V_{FEB} = V_{CS} - V_{IS} = -V_{IS} \qquad (3.4)$$

And the critical V_{GS} for P switching ($V_{GS,C}$) is given by:

$$V_{GS,C} = \frac{(\pm V_C + (A*P)_{FEA}/C'_T + (A*P)_{FEB}/C'_T)}{(1 - A_{FEA} * C_0/C'_T)} \qquad (3.5)$$

Let us first consider P_{FEA} switching from -P→ +P (coercive voltage = +V_C) by sweeping V_{GS} from a negative to positive value (-0.8→ 0.8V). The application of negative V_{GS} (-0.8V), leads to P_{FEA} = -P and P_{FEB} = +P in R-FEFETs (as discussed before). Now, for standard FEFETs, the 2nd term in the numerator of Equation-3.5 is negative, since initial value of P_{FEA}= -P, while the 3rd term doesn't exist. On the other hand, for the proposed R-FEFET, the 2nd term is negative along with a positive 3rd term, since the initial value of P_{FEB}= +P. Therefore, because of the opposite P in the two FE stacks of the R-FEFET, the magnitude of numerator in Equation-3.5 is larger for R-FEFETs compared to FEFETs (in which P_{FEB} is not present), resulting in larger $V_{GS,C}$ for P_{FEA} switching from -P → +P. Similarly, during the reverse sweep (V_{GS} = 0.8→-0.8V), $V_{GS,C}$ becomes more negative, with respect to FEFETs, for P_{FEA} switching from +P → -P (coercive voltage = -V_C) due to the presence of the additional term (corresponding to FE$_B$) in Equation-3.5. Therefore, higher |$V_{GS,C}$| for both -P→+P and +P→-P switching results in larger hysteresis (or hold margins) in 'NV' mode of the proposed R-FEFETs compared to standard FEFETs (~10X for

T_{FE}=8nm). Note that switching in P_{FEA} occurs before P_{FEB} (due to STP, Fig. 3.5). This results in a transient spike in the V_{IS} versus V_{GS} plots (Fig. 3.3(c)-inset (i, ii)), obtained using quasi-static analysis.

It is also important to note that P_{FEA} exhibits bi-stability in the 'NV' mode at V_{GS}=0V, which corresponds to the non-volatile state stored in R-FEFET. Even P_{FEB} exhibits bi-stability; however, we define the state of R-FEFET with P_{FEA} since C_{FEA} > C_{FEB}. Moreover, P_{FEA} = +P and –P corresponds to the ON and OFF state of the R-FEFET, respectively [181].

3.4.2 Volatile ('V') mode

To operate R-FEFETs in the 'V' mode, we apply V_{CS}=V_{DD} (0.8V). Due to the positive and large V_{CS}, P_{FEB} remains at a positive value for the entire range of V_{GS} (-0.8V to 0.8V) as shown in Fig. 3.3(b) and Fig. 3.4(c). When V_{GS} is negative (-0.8V), V_{IS} is negative despite positive P_{FEB}. This is because C_{FEA}>C_{FEB}, which results in higher influence of FE_A on the V_{IS} compared to FE_B. Negative V_{GS} yields a negative P_{FEA} as show in Fig. 3.4(c). Now, due to the positive P_{FEB} which tries to pull V_{IS} up (since V_{CS} = V_{DD}), V_{IS} is slightly less negative than the 'NV' mode (Fig. 3.3(c)). This results in a slightly larger $V_{GS,C}$ required for P_{FEA} switching from –P→+P compared to the 'NV' mode (since V_C of FE_A = $V_{GS,C}$ - V_{IS}; point-C in Fig. 3.3(a)). After –P→+P switching (step-1), V_{IS} increases to a positive value (step-2; similar to 'NV' mode). However, absence of P_{FEB} switching (due to V_{CS}=V_{DD} as explained before), results in non-occurrence of step-3 and step-4 of the STP (Fig. 3.5). Now, both FE_A and FE_B store +P, which yields significantly higher V_{IS} compared to the 'NV' mode (Fig. 3.3(c)-inset (iii)). Therefore, during the reverse V_{GS} sweep (from 0.8V→-0.8V), the coercive voltage (-V_C) required for switching P_{FEA} is achieved even for a positive value of $V_{GS,C}$. In other words, at the onset of P_{FEA} switching from +P→-P, V_{FEA}=-V_C =$V_{GS,C}$ -V_{IS} for $V_{GS,C}$>0V as shown in Fig. 3.3(a) (point-D). This results in the 'V' mode of operation with single stable state at V_{GS}=0V. Therefore, in this mode, R-FEFET can be utilized for low-power operations with improved disturb margins utilizing the inherent hysteresis. Note, P_{FEA}=+P and -P correspond to the states '1' (ON) and '0' (OFF), respectively.

We can also explain the 'V' mode of operation using Equation-3.2. In this mode, the simplified equation (considering V_{DS} = 0V) is:

$$V_{FEA} = -(A*P)_{FEA}/C'_T - (A*P)_{FEB}/C'_T + \left(1 - A_{FEA} * \frac{C_0}{C'_T}\right) * V_{GS} \; - A_{FEB} * V_{CS} \; * C_0/C'_T \quad (3.6)$$

$$V_{FEB} = \; V_{CS} - V_{IS} \qquad\qquad (3.7)$$

By design (discussed later), we ensure that $V_{CS}=V_{DD}$ corresponds to $V_{FEB} > -V_C$, so that P_{FEB} never switches from +P→–P. In other words, P_{FEB} remains at +P for the entire V_{GS} sweep (-0.8→0.8V). From Equation-3.6, $V_{GS,C}$ for P_{FEA} switching is:

$$V_{GS,C} = (\pm V_C + (A*P)_{FEA}/C'_T + \; A_{FEB} * (+P)/C'_T + A_{FEB} * V_{DD} \; * C_0/C'_T)/(1 - A_{FEA} * C_0/C'_T) \quad (3.8)$$

With proper design, the sum of the last 3 terms of the numerator in Equation-3.8 is highly positive when compared to $|V_C|$. Hence, $V_{GS,C}$ for +P→-P switching (coercive voltage = $-V_C$) is >0 (Fig. 3.3(a)). For -P→+P switching (coercive voltage=$+V_C$), $V_{GS,C}$ >0 due to the sum of all terms in Equation-3.8 being positive (as discussed for the 'NV' mode). Moreover, this $V_{GS,C}$ (for –P→+P switching) is slightly larger than its corresponding value for 'NV' mode because of the additional positive term, V_{CS} (=V_{DD}) in Equation-3.8. Hence, with $V_{GS,C}$ >0 for P switching in both directions, 'V' mode of operation is achieved.

3.4.3 Effect of drain-to-source voltage

A positive V_{DS} leads to an increase in V_{IS} because of the capacitive coupling between drain and IML [140] (see Equation-3.2). This results in an increase in $V_{GS,C}$ required to reach the respective coercive voltages, shifting the 'V' and 'NV' curves towards the right (Fig. 3.6). Note, the proposed R-FEFET shows drain voltage independent reconfigurability between the two modes, enhancing the circuit design flexibility.

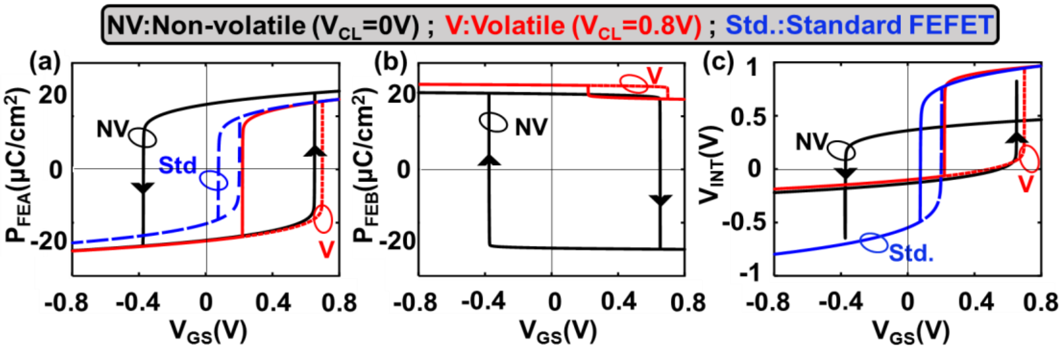


Fig. 3.6 (a) Polarization of FE_A stack, (b) Polarization of FE_B stack and (c) internal metal voltage versus gate voltage. V_{DS}=0.3V; T_{FE}=8nm.

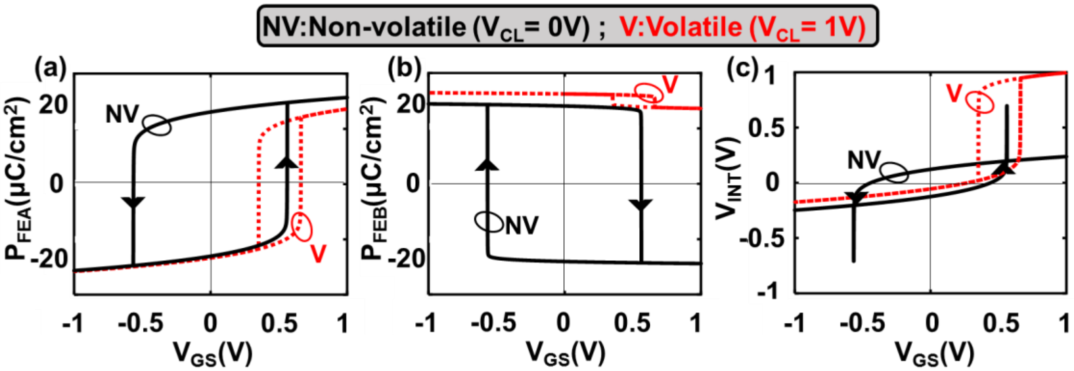


Fig. 3.7 (a) Polarization of FE_A stack, (b) Polarization of FE_B stack and (c) internal metal voltage versus gate voltage for planar R-FEFETs.

In a similar fashion, planar R-FEFETs with 22nm technology node have also been simulated with T_{FE}=7nm, width of FE_A (W_{FEA}) and FE_B (W_{FEB}) equal to 44nm and 33nm respectively and transistor gate width (W) of 110nm ($W_{FEA}+W_{FEB}+3\lambda$). Planar R-FEFETs also show similar characteristics with respect to the dynamic reconfigurability (Fig. 3.7). We achieve HW modulation from 1.1V in 'NV' mode to 0.3V in 'V' mode.

It is important to note from the figures above that the R-FEFET is OFF when V_{GS}=0V, despite $V_{CS}=V_{DD}$. Moreover, its ON state ($V_{GS}=V_{DD}$) is maintained even at V_{CS}=0V. This is because V_{IS} is designed to be impacted more by V_{GS} than V_{CS} with $C_{FEA} > C_{FEB}$ (as discussed before). Also note, we show all the device characteristics (P_{FEA}, P_{FEB}, V_{IS}) versus V_{GS}, to understand the device properties with respect to the applied signal V_{GS}, which will be useful for the discussion in Chapter-4.

The above discussion indicates that the device characteristics are a function of several device parameters such as V_{CS}, T_{FE} and capacitance ratio of the FE stacks. To understand how R- FEFETs need to be designed, we present device analysis considering the aforementioned parameters, next.

3.5 Device Design and Analysis

To carry out the analysis of R-FEFETs in this section, we perform simulations for planar R-FEFETs. Since FinFETs exhibit width quantization, the analysis of capacitance ratio C_{FEA}:C_{FEB} is restricted. Therefore, for a general comprehensive analysis, we focus on planar R-FEFETs. The same trends also hold for FinFETs based structures. We look into three important parameters: (a) T_{FE}, (b) V_{CS} and (c) A_{FEB} (controlled by W_{FEB}).

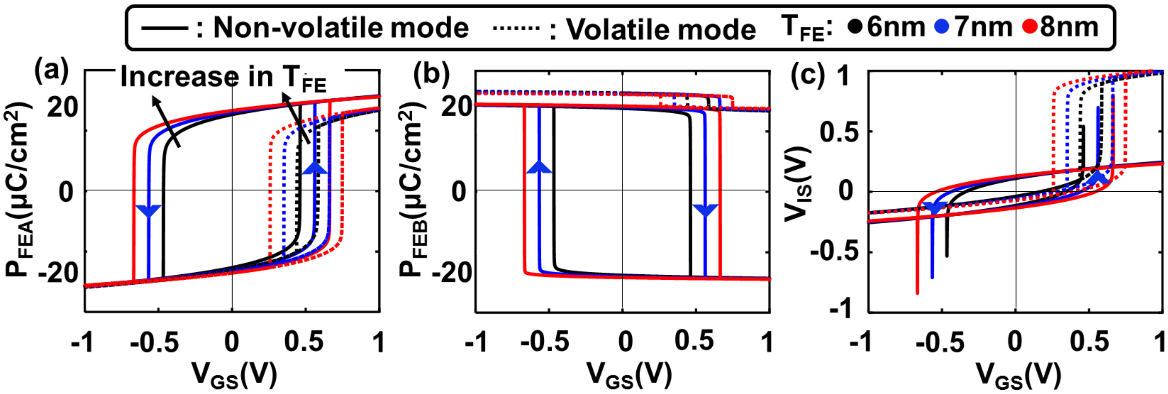


Fig. 3.8 (a) Polarization of FE_A stack, (b) Polarization of FE_B stack and (c) internal metal voltage versus gate voltage for planar R-FEFETs for T_{FE} = 6, 7, 8nm. V_{DS}=0V; W_{FEA}=44nm; W_{FEB}=33nm.

3.5.1 Thickness of ferroelectric (T_{FE})

The device characteristics for different T_{FE} are shown in Fig. 3.8. Note, both stacks have the same T_{FE}. With increase in T_{FE} from 6nm to 8nm, HW widens for both 'V' and 'NV' modes due to increase in $V_{GS,C}$ [31], [133], resulting in larger V_{GS} margins for holding P (or hold margins; discussed further later). T_{FE} optimization in conjunction with proper choice of supply voltage (V_{DD}) can be a key design methodology since HW can be tuned as per the circuit needs.

3.5.2 Control terminal voltage (V_{CS})

V_{CS} plays a crucial role in determining the mode of operation as described before. To achieve the 'V' mode of operation, proper magnitude of V_{CS} is required. To understand the dependency of V_{CS}, we show the characteristics of R-FEFETs for different V_{CS} in Fig. 3.9. As V_{CS} is increased, V_{IS} also increases (Fig. 3.9(c)-inset). This is because of the pull-up action of FE_B due to capacitive coupling, as discussed before. For small V_{CS} (<0.5V), R-FEFET does not exhibit 'V' mode. This is attributed to the fact that P_{FEB} undergoes opposite switching along with P_{FEA} due to low V_{CS}. P_{FEB} switching leads to reduction in V_{IS}, as explained by STP (step-4) (Fig. 3.5). The lower V_{IS}, in turn, prevents +P→-P switching in FE_A for V_{GS} >0V (since V_{FEA}=V_{GS}-V_{IS}), yielding 'NV' characteristics. As V_{CS} is increased, a point is reached where V_{FEB} remains greater than -V_C, thus preventing P_{FEB} switching. In other words, P_{FEB}=+P for the entire range of V_{GS} (-0.8→0.8), resulting in high V_{IS} (non-occurrence of step-3 and step-4 during STP; Fig. 3.5). Now, due to the

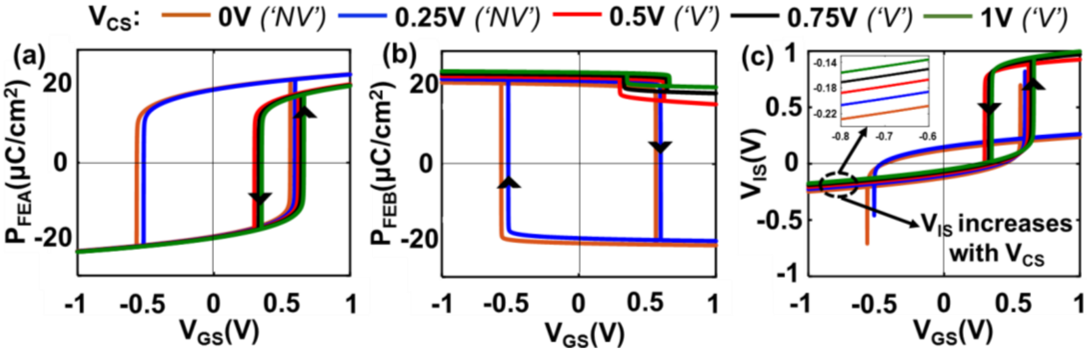


Fig. 3.9 (a) Polarization of FE_A stack, (b) Polarization of FE_B stack and (c) internal metal voltage versus gate voltage for V_{CS} varying from 0V→1V. V_{DS}=0V; T_{FE}=7nm; W_{FEA}=44nm; W_{FEB}=33nm.

high V_{IS}, $V_{GS,C}$ required for P_{FEA} switching from +P→-P becomes >0V, i.e., $V_{FEA}=-V_C=V_{GS,C}-V_{IS}$ for $V_{GS,C}$ >0V. Therefore, at this particular V_{CS} (~0.5V), R-FEFET reconfigures to the 'V' mode. Hence, for V_{CS} >0.5V, absence of P_{FEB} switching, keeps V_{IS} high enough to achieve P_{FEA} switching (+P→-P) at V_{GS} >0V ('V' mode).

3.5.3 Area of control stack (A_{CS})

Area of the FE_B (controlled by W_{FEB}) can also be used as a design knob for device optimization. We discuss its effect in context of (a) V_{IS} (b) HW below:

(a) Internal Metal Potential (V_{IS})**:** As described before, the influence of FE_B is to decrease the effect of the P_{FEA} on IML in the 'NV' mode, leading to lowering of $|V_{IS}|$ compared to FEFETs. As the area of the FE_B is increased by increasing W_{FEB} from 22nm to 44nm (maintaining W_{FEA} = 44nm and $W=W_{FEA}+W_{FEB}+3\lambda$), the influence of P_{FEB} on V_{IS} also increases. Since FE_A and FE_B always have opposite P in the 'NV' mode, $|V_{IS}|$ decreases with increase in FE_B area as shown in Fig. 3.10(b). In fact, when W_{FEB} is 44nm, due to the equal and opposite effect of the control and gate stacks, $|V_{IS}|$ reaches ~0V. If $W_{FEB}>W_{FEA}$, then the roles of the FE stacks change and P_{FEB} dominates the influence on V_{IS}. In the 'V' mode, since P_{FEB} remains +P over the entire V_{GS} region (-1V to 1V), increase in FE_B area results in increase in V_{IS} due to higher influence of P_{FEB} (Fig. 3.10(d)). Increase in V_{IS} results in the P characteristics shifting to the right similar to the effect of V_{DS} (Fig. 3.10(c)).

(b) Hysteresis width (HW): As described before, increase in FE_B area corresponds to a decrease in the influence of P_{FEA} on IML. As a result, R-FEFET requires higher voltage for

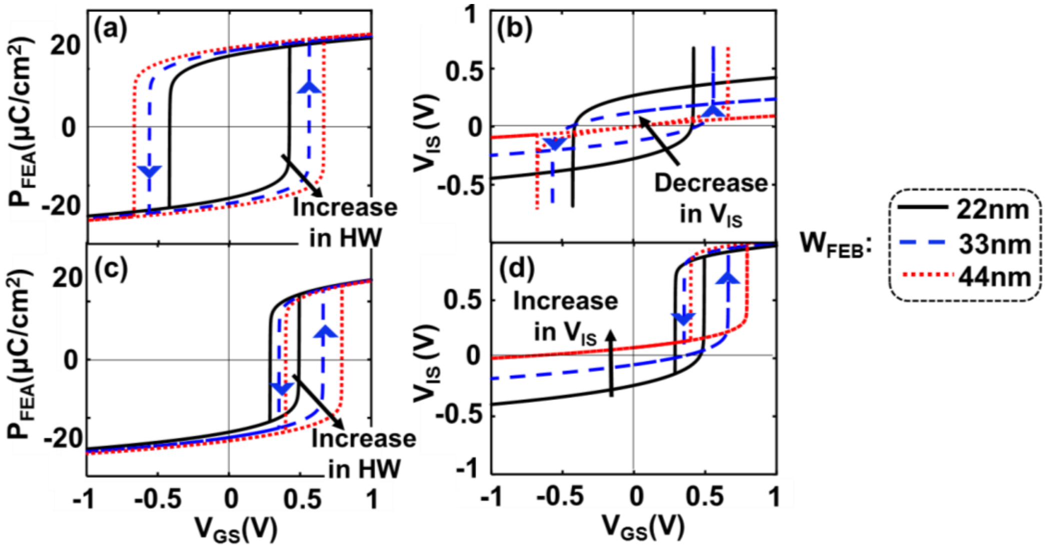


Fig. 3.10 Polarization of FE_A stack and internal metal voltage versus gate voltage for (a, b) *'NV'* and (c, d) *'V'* mode for W_{FEB} = 22, 33, 44nm. V_{DS}=0V; W_{FEA}=44nm; T_{FE}=7nm; HW: Hysteresis Width.

P switching which leads to widening of the hysteresis (Fig. 3.10(a, c)). This plays an important role in determining the stability margins for circuit applications. Our simulation analysis indicates that to achieve reconfigurability between the 'V' and 'NV' modes considering W_{FEA}=44nm, C_{FEA}:C_{FEB} (=W_{FEA}:W_{FEB}) must be <44:5 (with, W=W_{FEA}+W_{FEB}+3λ). This is because, at capacitance ratios above this limit, FE_B is not capable of influencing V_{IS} in order to achieve volatile operation. On the other hand, if C_{FEA}:C_{FEB} <1 then the roles of FE stacks interchange as explained before.

3.5.4 Current-voltage characteristics

With the understanding of the trends above, we optimize R-FEFETs to achieve proper reconfigurability (parameters listed later) and present the current-voltage transfer characteristics in the 'V' mode (Fig. 3.11). Note, the 'V' mode will be used for read/write operations in non-volatile circuits based on R-FEFET. The relevant characteristics (P-V) in the 'NV' mode illustrating P retention (for hold operation) have already been illustrated in Fig. 3.7. Shifting of hysteresis due to V_{DS} is reflected in Fig. 3.11. The ON current (I_{ON}) of planar R-FEFET (W_{FEA}=44nm, W_{FEB}=33nm, W = 110nm, and T_{FE}=7nm) is 13% higher than standard FET (W=110nm; Fig. 3.11(a)). On the other hand, I_{ON} of FinFET based R-FEFET (C_{FEA}:C_{FEB}=2:1, and

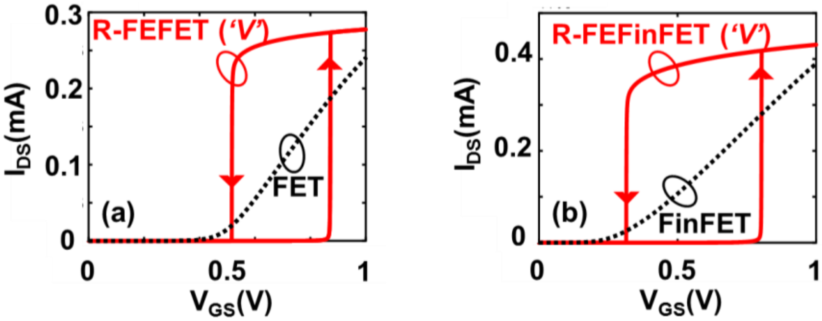


Fig. 3.11 Transfer characteristics of (a) planar and (b) FinFET based R-FEFET with V_{DS}=0.5V.

T_{FE}=8nm) is 10% higher than standard FinFET (# fins = 3; Fig. 3.11(b)). The increase in I_{ON} with respect to standard FET (or FinFETs) is because of FE_A operating in the negative P-V region (V_{FEA}<0, P>0; shaded region in Fig. 3.11(c)) in the ON state, leading to a voltage step-up action, i.e., V_{IS}>V_{GS} since V_{IS}=V_{GS}-V_{FEA} [108]. Also, the I_{ON}/I_{OFF} ratio for R-FEFETs in the 'V' mode is ~ 10^6, which indicates excellent distinguishability. Note, during the circuit implementation of R-FEFETs, we perform a non-destructive sensing of P, which is discussed extensively in Chapter-4 of this dissertation.

It may be noteworthy to mention that the retention properties of the FE in the proposed R-FEFET will be similar to that of standalone FE capacitors. This is due to the presence of IML used for coupling the two FE stacks with the transistor. The floating IML undergoes GL (discussed in the next section), which brings V_{IS} to 0V during the hold state, over time. This results in minimum depolarization fields across the FE, enhancing its retention properties. Experimental studies have shown retention of ~10 years for HZO based FE [141]. On the other hand, as discussed in the recently proposed HZO based FE-Metal-FET (FEMFET) device architecture, the presence of IML result in higher endurance of ~10^{11} cycles [142]. This is attributed to the lower charge trapping effects in FE and IML, when compared to the device structures without IML (~10^7 cycles [141]). The proposed R-FEFETs takes the inherent advantage of the presence of IML (as in FEMFETs) and therefore we expect similar benefits with respect to reliability and endurance of the device as observed in FEMFETs.

3.6 Analysis Considering Gate Leakage

With the understanding of the operation of R-FEFETs, we now discuss the impact of GL on R-FEFETs. As in the previous section, we focus on planar R-FEFETs with W_{FEA}=44nm,

W_{FEB}=33nm, T_{FE}=7nm and V_{CS}=1V. Note, we show trends only for a planar technology to avoid repetition. However, similar characteristics are achieved for FinFETs as well.

3.6.1 Gate leakage in standard FEFETs

As mentioned earlier, FEFETs can be realized with or without an IML in between the FE and DE. Recent studies have revealed the pros and cons of each structure. FEFETs without the IML are more suitable for steep switching applications but for memory applications they may undergo deterioration in performance and retention due to (a) sensitivity to charge trapping effects in FE leading to lower endurance [141], [142] and (b) non-zero depolarization field across FE affecting the retention properties [169], [182]. On the other hand, as discussed in [142], FEFETs with IML are expected to show good endurance due to minimized charge trapping in FE and IML. However, they are adversely affected by the GL [45], [183]. GL through the FE and DE layers (modeled and captured by R_{FE} and R_{DE}; Fig. 3.12(a)) in FEFETs with the floating IML, results in discharge of V_{IS} over time as shown in Fig. 3.12(b) and also discussed in Chapter-2. V_{IS} discharge due to GL results in loss of distinguishability between +P (ON state) and –P (OFF state). However, P is still retained which stores the bit information. This retained P will be referred to as the hold polarization, P_{HOLD} (i.e. the retained P after V_{IS} has discharged to 0V). To read the state of FEFETs, unconventional read techniques are required which incur design overheads [45]. GL issue is also present in the proposed R-FEFETs due to the requirement of IML. However, the presence of the control terminal mitigates the read overheads observed in FEFETs with IML. Moreover, GL enhances the retention properties in the stand-by state (similar to standalone FE capacitors) due to minimized depolarization fields across the FE, as discussed previously [169], [182]. We discuss

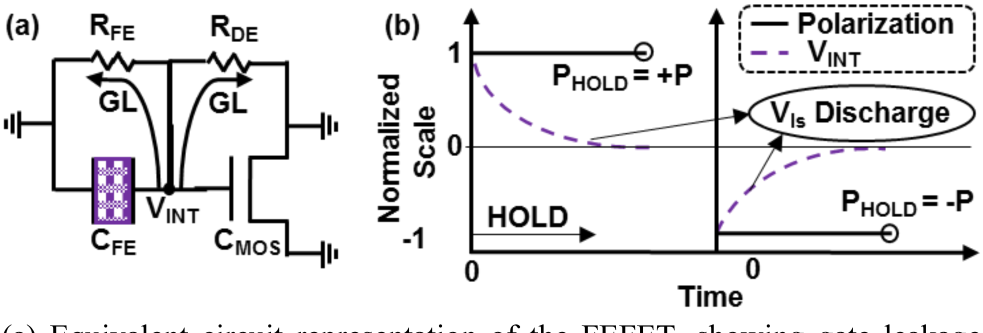


Fig. 3.12 (a) Equivalent circuit representation of the FEFET, showing gate leakage (GL). (b) Transient simulations showing polarization retention with V_{IS} discharge due to GL. P_{HOLD}= Retained Polarization

these aspects later, along with the memory design in Chapter-4. For now, we focus on the impact of GL and the benefits that R-FEFETs possess over FEFETs.

3.6.2 Impact of GL in R-FEFETs

Before we begin our discussion, it is important to mention that P_{HOLD} in R-FEFETs corresponds to P_{FEA}, since P_{FEA} defines state of the R-FEFET, as discussed before. Let us consider that GL has discharged V_{IS} to 0V. Two possible cases of P_{HOLD}, (=P_{FEA}) exist: (i) P_{HOLD} = +P (P_{FEB}= -P) and (ii) P_{HOLD} = -P (P_{FEB} = +P). In the following, we explain these cases for both 'NV' and 'V' modes:

(a) **P_{HOLD} = +P:** Since P_{HOLD} is defined corresponding to the R-FEFET being in the hold state for a long time, the bias conditions for this case are: P_{FEA} = +P, P_{FEB} = -P, V_{IS} = 0V at V_{GS} = 0V and V_{CS} = 0V. The major difference between a device with GL and the one neglecting GL is that the initial V_{IS} for the former is =0V while that of the latter is >0V. This has the following consequence: (i) Since $V_{GS,C}$ =V_{IS}-V_C (for +P→-P switching), $V_{GS,C}$ considering GL is more negative compared to the case neglecting GL (Fig. 3.7). (ii) After +P→-P switching V_{IS} is more negative compared to the case neglecting GL. This results in a lower $V_{GS,C}$ (=V_{IS}+V_C) for the subsequent -P→+P switching (see Figs. 3.7(a), 3.13(a)). Therefore, the net effect of GL in this case (P_{HOLD}=+P) is to shift the device characteristics towards the left with respect to the case neglecting GL. Note, at V_{CS} =0V, R-FEFET retains non-volatility despite this shifting.

Now, let us consider the effect of GL on the 'V' mode (V_{CS}=V_{DD}; Fig. 3.13). Similar to the 'NV' mode GL results in lowering of V_{IS}. This leads to shifting of the 'V'

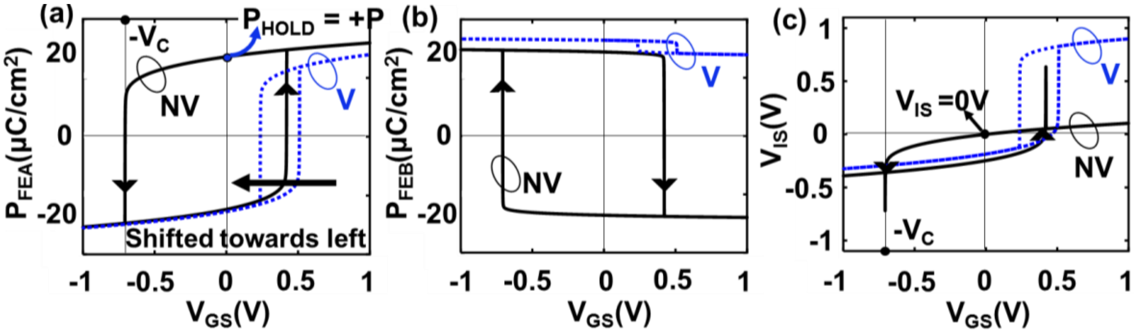


Fig. 3.13 (a) Polarization of FE_A, (b) Polarization of FE_B and (c) internal metal voltage versus gate voltage when P_{HOLD} = +P. V_{DS}=0V; T_{FE}=7nm; W_{FEA}=44nm; W_{FEB}=33nm; HM: Hold Margin.

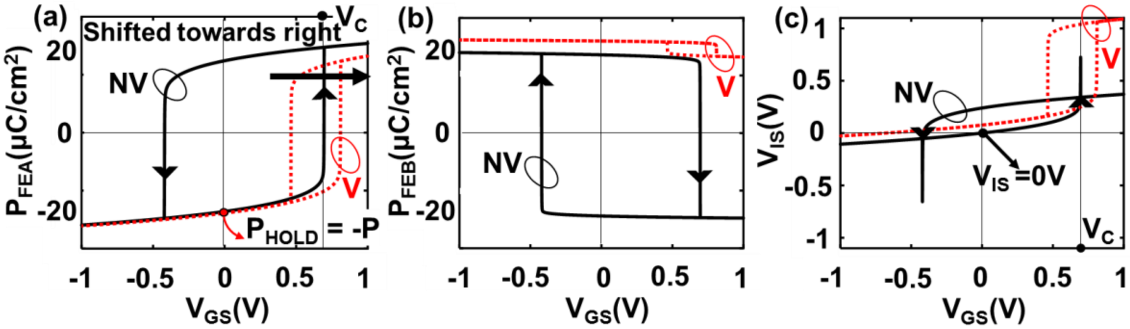


Fig. 3.14 (a) Polarization of FE_A, (b) Polarization of FE_B and (c) internal metal voltage versus gate voltage when P_{HOLD} = -P. V_{DS}=0V; T_{FE}=7nm; W_{FEA}=44nm; W_{FEB}=33nm; HM: Hold Margin.

characteristics towards the left (see Figs. 3.7, 3.13). Note, while shifting to the left, 'V' mode retains its property i.e., single stable state at V_{GS}=0V.

(b) P_{HOLD} **= -P:** In this case the initial conditions are: P_{FEA} = -P, P_{FEB} = +P, V_{IS} = 0V at V_{GS} = 0V and V_{CS} = 0V. Similar to P_{HOLD} = +P, the main difference between the characteristics considering and neglecting GL is in the initial V_{IS}. While initial V_{IS} = 0 with GL, it is <0 without GL (see Figs. 3.7(c), 3.14(c)). Thus, $V_{GS,C}$ (=V_{IS}+V_C) for –P→+P switching becomes more positive when GL is considered. Also, after –P→+P switching, V_{IS} is more positive for the case with GL. Hence, for subsequent –P→+P switching, $V_{GS,C}$ (=V_{IS}-V_C) increases. Thus, the effect of GL for P_{HOLD}=-P is to shift the device characteristics towards the right. Note, non-volatility is preserved when V_{CS}=0V. Also note, the device characteristics further shift towards the right when V_{DS} >0. Our simulations reveal that for V_{DS} up to 1V, the R-FEFET still retains its 'NV' behavior at V_{CS}=0V. Similarly, in the 'V' mode (V_{CS}=V_{DD}), V_{IS} increases due to GL resulting in shifting of device characteristics towards right (see Figs. 3.7, 3.14).

In the presence of GL, the proposed R-FEFET retains dynamic reconfigurability between 'NV' mode (two stable states at V_{GS}=0V) and 'V' mode (OFF state at V_{GS}=0V) for both P_{HOLD}=+P and -P. Moreover, in the presence of GL, V_{DS} dependence remains the same as discussed in the previous sections. Also, all the trends with respect to different design parameters (T_{FE}, W_{FE}, AFE) holds, except for the shift in the characteristics depending on P_{HOLD}, as mentioned above. It is important to mention that, R-FEFETs based on FinFETs also follow similar trends and device characteristics in the presence of GL.

From the discussion in this section, we observe that, in the presence of GL, R-FEFET exhibits 4 modes of operation in total: 2 'V' and 2 'NV', corresponding to P_{HOLD} = +P and –P. Therefore, while designing circuits, we need to ensure correct functionality by design optimizations for these 4 corner cases and verify that, for any transient state in between, the functionality is retained. The implications of the 4 modes and the transient states on non-volatile circuits has been analyzed extensively later. Note that, due to transient effects of GL, the application of the proposed R-FEFET (in the 'V' mode) is not in generic logic computation. However, R-FEFETs have immense potential in the field of energy-efficient non-volatile computing as discussed in [174] and Chapter-5 of this dissertation.

3.6.3 Impact on current-voltage characteristics

For the case with P_{HOLD}=+P (-P), V_{IS} decreases (increases), resulting in decrease (increase) in I_{ON}. For planar R-FEFET in 'V' mode, I_{ON} with P_{HOLD} = +P is similar to standard FETs while for the case with P_{HOLD}=-P, is 15% higher (Fig. 3.15(a)). For FinFET based R-FEFET, I_{ON} with P_{HOLD} = +P (or -P) is 27% lower (or 17% higher) than standard FinFETs (Fig. 3.15(b)). Moreover, even in the presence of GL, R-FEFETs exhibits excellent I_{ON}/I_{OFF} ratio ~10^4 in the 'V' mode.

3.6.4 Comparison with standard FEFETs

The major advantages of R-FEFETs over standard FEFETs are the following: (a) they exhibit larger intrinsic hysteresis (P vs V_{GS}) due to additional capacitive coupling of the control stack. Therefore, R-FEFETs show more robustness for NV operations in the presence of GL and

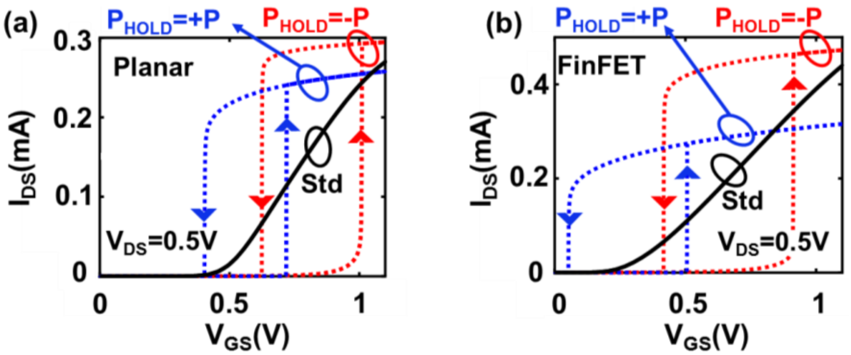


Fig. 3.15 Transfer characteristics of (a) planar and (b) FinFET based R-FEFET in the presence of gate leakage with V_{DS}=0.5V.

variations. (b) Easier distinguishability of the bi-stable P states. Standard FEFETs lose their current based distinguishability (depending on P stored) in the presence of GL [45]. Work-function engineering along with a 2-step read operation has been proposed to mitigate the effects of GL in standard FEFET based memories [45]. However, this leads to higher read power. R-FEFETs overcome these challenges utilizing the unique feature of dynamic reconfigurability. These aspects, in context of the memory design, are further discussed in Chapter-4.

3.7 Transition between different Modes during Circuit Operation

Before discussing the transition between different modes of R-FEFET, let us briefly describe a few terminologies. Let us consider that we write a bit (+P or –P) into the R-FEFET. If the R-FEFET is accessed a short time after the write, GL does not get sufficient time to change V_{IS}. However, if the R-FEFET is accessed after a long time, V_{IS} discharges to 0V. We refer to these cases as 'access after short time' and 'access after a long time', respectively. Considering access after a long time, we have states 1 and 5 corresponding to P_{HOLD}=P_{FEA}= +P and -P respectively (see Fig. 3.16(b, d)). The critical gate voltage ($V_{GS,C}$) required to switch P from these states is

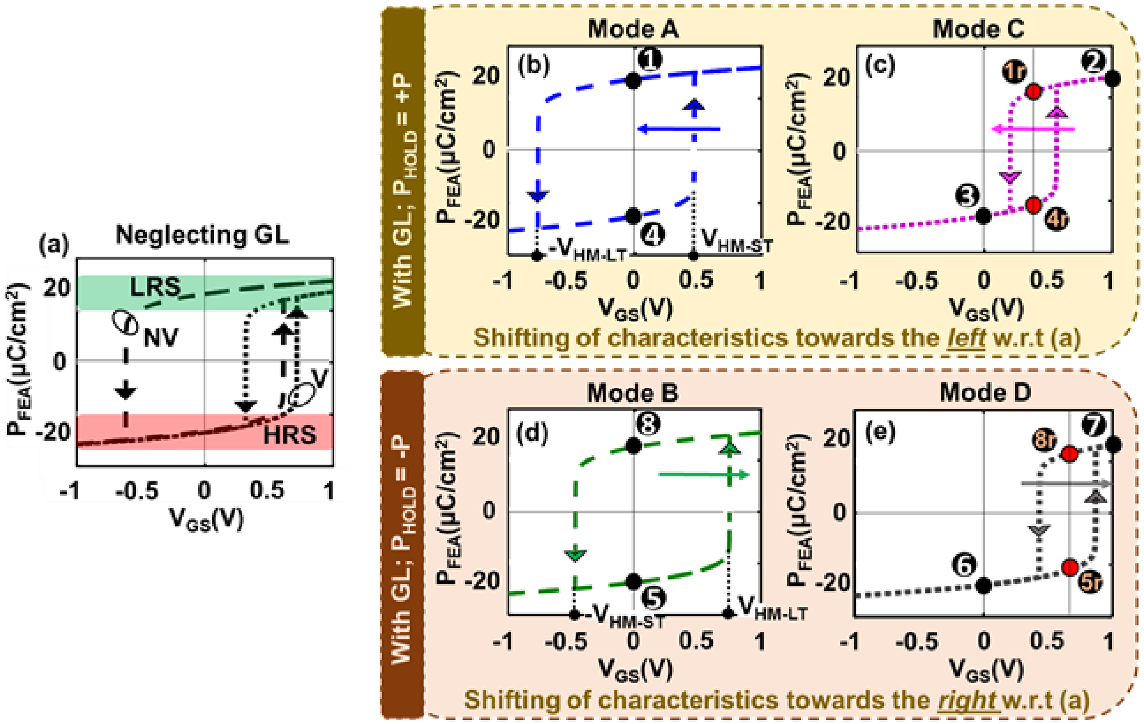


Fig. 3.16 R-FEFET device characteristics (a) without GL, with GL for (b, c) P_{HOLD} = *+P* and (d, e) P_{HOLD} = *-P*. Modes A and B are non-volatile (*NV;* V_{CS}=*0V*); modes C and D are volatile (*V;* V_{CS}=V_{DD}).

±V_{HM-LT}. State 4 (P_{FEA}=-P) and state 8 (P_{FEA}=+P) correspond to the case when the R-FEFET is accessed after a short time. For these states, $V_{GS,C}$ for P switching is ±V_{HM-ST}. $V_{HM-LT/ST}$ correspond to long/short term hold voltage margins. Note, when we design the memory, we need to ensure that the functionality is achieved, irrespective of access after long/short time and we discuss that in the subsequent sections.

As discussed before, GL yields four modes of operation for the R-FEFET. Therefore, from a circuit design perspective, it is important to understand the dynamics of the different modes of operation, which is discussed in this section. The four modes (two 'NV' and two 'V') have been referred to as A, B, C and D from here on (Fig. 3.16(b-e)). A brief description of these modes is given below. Note, the analysis done in this section is generic for R-FEFET implementation in any non-volatile circuit design. In the next chapter, we will particularly focus on the memory application.

Mode A represents the 'NV' mode with P_{HOLD} = +P and initial V_{IS} = 0V (Fig. 3.16(b)). This corresponds to the hold state for bit '1'. Mode B corresponds to the 'NV' mode with P_{HOLD} = -P and initial V_{IS} = 0V (Fig. 3.16(d)). This corresponds to the hold state for bit '0'. Mode C and D represents the 'V' mode corresponding to P_{HOLD} = +P and -P respectively, after the assertion of V_{CS} (Fig. 3.16(c, e)). The states 1-8 labeled in Fig. 3.16(b-e) are used in this section to understand the transitions between the different modes during circuit operation. In the following subsections, we explain the dynamic operation of R-FEFETs with respect to these four modes, which will help in understanding the memory operation discussed in the next chapter.

(a) P_{HOLD}**=+P:** Let us consider that the device is in the 'NV' mode for a long time such that V_{IS} discharges to 0V and P_{HOLD}= P_{FEA} =+P (Mode A). As detailed earlier, P_{FEB} is opposite to P_{FEA} in the 'NV' mode. Therefore, the initial conditions for this state are P_{FEA}=+P, P_{FEB}=-P and V_{IS}=0V which is represented as state 1 in Fig. 3.16(b). Now, writing into the R-FEFET is achieved by changing its operation to 'V' mode (Mode C). Let us consider write '1' and '0' individually:

(i) To write '1', we apply V_{GS}=V_{DD} and then assert V_{CS} to V_{DD}. This leads to operation of the R-FEFET in mode C ('V' mode). Recall, P_{FEB} =+P for the entire range of V_{GS} in the 'V' mode. In other words, P_{FEB} switches from –P→+P with the assertion of V_{CS}. P switching in FE_B results in V_{IS} increasing to a positive value. This corresponds to state 2 (in mode C; Fig. 3.16(c)) with the conditions P_{FEA}=+P, P_{FEB}=+P and V_{IS}>0V. After

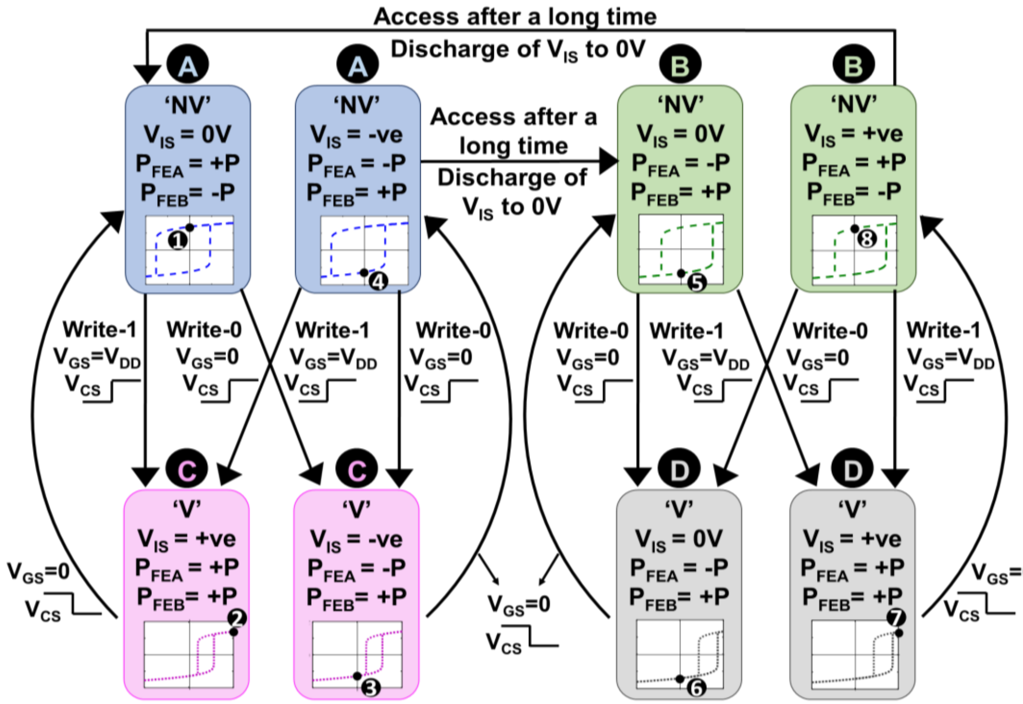


Fig. 3.17 Self-consistent dynamic device operation of R-FEFET while transitioning between the *'V'* and *'NV'* modes (A, B, C and D)

writing '1', let us de-assert V_{CS} to 0 in order to hold the value of P_{FEA} in the 'NV' mode. During this process, P_{FEB} switches back from +P→-P (since, in 'NV' mode P_{FEA} and P_{FEB} are always opposite), leading to discharge of V_{IS} to 0V. This corresponds to the initial state 1 (in mode A) that we started with, i.e., P_{HOLD} (=P_{FEA}) =+P, P_{FEB}=-P and V_{IS}=0V. Further operations from this point would lead to the same steps as explained in this sub-section (Fig. 3.16; state 1).

(ii) To write '0', we keep V_{GS} at 0V and assert V_{CS} to V_{DD}. This yields P_{FEB} switching from –P→+P and P_{FEA} from +P→-P. Since, C_{FEA} > C_{FEB}, P switching (dominated by P_{FEA}) leads to V_{IS}<0. This corresponds to state 3 (in mode C; Fig. 3.16(c)). Now, consider V_{CS} is de-asserted in order to hold the bit information in the 'NV' mode i.e., P_{FEA}=-P ('0'). During this process there is no P switching due to both P_{FEA} and P_{FEB} already being opposite to each other (which corresponds to 'NV' mode), resulting in no charging or discharging of V_{IS}. This corresponds to state 4 (in mode A) with the conditions, P_{FEA}=-P, P_{FEB}=+P and V_{IS}<0. Now, if the device is accessed within a short time such that V_{IS} does not discharge to 0V, then writing in the subsequent cycles follow the same steps as explained above (see Fig. 3.17; state 4). However, if the R-FEFET is accessed after a long time such that V_{IS} =0V, then the 'NV' mode changes from mode A to mode B with the conditions P_{FEA}=-P, P_{FEB}=+P and V_{IS}=0V (state 5; Fig. 3.17(d)). Subsequent device operation from this point is described in the next.

(b) P_{HOLD}**=-P:** Let us consider that the device is in the 'NV' mode for a long time such that the initial conditions are P_{HOLD} (P_{FEA}) =-P, P_{FEB}=+P and V_{IS}=0V. This corresponds to state 5 in mode B (Fig. 3.16(d)). Similar to the previous case (with P_{HOLD}=+P), let us consider the following possibilities:

(i) To write '0' (redundant write), we keep V_{GS}=0V and assert V_{CS} to V_{DD}. Assertion of V_{CS} brings the R-FEFET to the 'V' mode (mode D). However, this corresponds to the same conditions we started with, i.e, P_{FEA} = -P, P_{FEB}=+P and V_{IS}~0V (state 6 in mode D; Fig. 3.16(e)). This is because of no P switching in FE_A or FE_B, and therefore V_{IS} remains at ~0V. After writing '0', V_{CS} is de-asserted in order to hold the value of P_{FEA}. This leads to the conditions P_{FEA} = -P, P_{FEB} = +P and V_{IS} = 0V which are nothing but the initial conditions i.e, state 5 which we started with. Therefore, any further operation would follow the same steps as explained this sub-section (state 5).

(ii) To write '1' we drive V_{GS} to V_{DD} and assert V_{CS} to V_{DD} (mode D). Writing '1' corresponds to P_{FEA} switching from -P→+P. However, P_{FEB} remains at +P because of the device operating in the 'V' mode (as mentioned before). This leads to charging of V_{IS} to a positive value, corresponding to the conditions P_{FEA}=+P, P_{FEB}=+P and V_{IS}>0V (state 7 in mode D; Fig. 3.16(e)). Now, when V_{CS} is de-asserted to operate the R-FEFET in the 'NV' mode, P_{FEB} undergoes switching from +P→-P (since P_{FEA}=+P and in 'NV' mode P_{FEB} is always opposite to P_{FEA}). This P switching corresponds to a slight decrease in the magnitude of V_{IS} but it still remains positive due to the higher influence of P_{FEA}=+P (C_{FEA}>C_{FEB}). Therefore, the device conditions after de-assertion of V_{CS} are P_{FEA}=+P, P_{FEB}=-P and V_{IS}>0V which is state 8 in mode B (Fig. 3.16(d)). Now, if the device is accessed within a short time such that V_{IS} doesn't discharge to 0V, then writing in the subsequent cycles follow the same steps as explained above. However, if the R-FEFET is accessed after a long time such that V_{IS} discharges to 0V, then the 'NV' mode of operation changes from mode B to mode A with the conditions P_{HOLD}(P_{FEA})=-P, P_{FEB}=+P and V_{IS}=0V (state-1). And, subsequent access of the device follows same steps as explained earlier (Fig. 3.17; state 1).

From the above discussion, we notice that the R-FEFET device operation considering various possibilities of P_{HOLD} (+P/-P), writing bit information ('1'/'0') and access after long time or short time from a circuit implementation perspective, leads to a self-consistent operation in between the four modes A, B, C and D. They also showcase modulating hysteresis in between the 'V' and 'NV' modes, even in the presence of GL. This leads to significant advantages over FEFETs during their circuit applications.

3.8 Symmetric R-FEFET (R-FEFET$_{SYM}$) Design to Mitigate the Influence of Gate Leakage

As discussed in the previous sections, the R-FEFET design can be used in the 'NV' mode for sensing only if GL can be controlled or mitigated. For example, this can be achieved by increasing the dielectric thickness to reduce gate tunneling current as widely done for the NAND flash technology [184]. However, this comes at the cost of high depolarization fields since due to the presence of an internal metal potential creating an opposing field to the polarization stored. If GL cannot be reduced/eliminated, then polarization-dependent hysteresis shifts reduces the design margins [171], [185]. Now, to overcome the aforementioned problem, in this section, we propose a symmetric R-FEFET (with A_{FEG} = A_{FEC}) which eliminates depolarization fields in the 'NV' mode and also overcomes the impact of GL on the device characteristics. *(Note, $FE_{A/B}$ and $FE_{G/C}$ are interchangeably used in this dissertation which corresponds to the gate/control FE stack of R-FEFET).* We explain its operation vis-à-vis the asymmetric R-FEFET (A_{FEG} > A_{FEC}) proposed in the previous sub-sections. The following analysis has been caried out using 10-nm FinFET node using the harnessing the FEMFET configuration [142]. For the symmetric R-FEFET (referred to as R-FEFET$_{SYM}$) design in this work, we consider FE thickness (T_{FE}) = 9nm (unless stated otherwise), width of FE_G (W_{FEG}) and FE_C (W_{FEC}) = 4λ each (unless stated otherwise), where λ is half the gate length and number of fins = 4. For the regular R-FEFET (asymmetric; also referred to as R-FEFET$_{ASYM}$) proposed in the previous sections, T_{FE} = 9nm, W_{FEG} = 4λ, W_{FEC} = 3λ and number of fins = 4. Fig. 3.18 illustrates the device characteristics of R-FEFET$_{ASYM}$.

3.8.1 Device design and operation

The proposed R-FEFET$_{SYM}$ features both Gate and Control FE stacks with equal area (A_{FEG} (=W_{FE}*L_{FE}):A_{FEC} (=W_{FE}*L_{FE}) = 1:1) (Fig. 3.19(a)). The device foot-print remains the same for R-

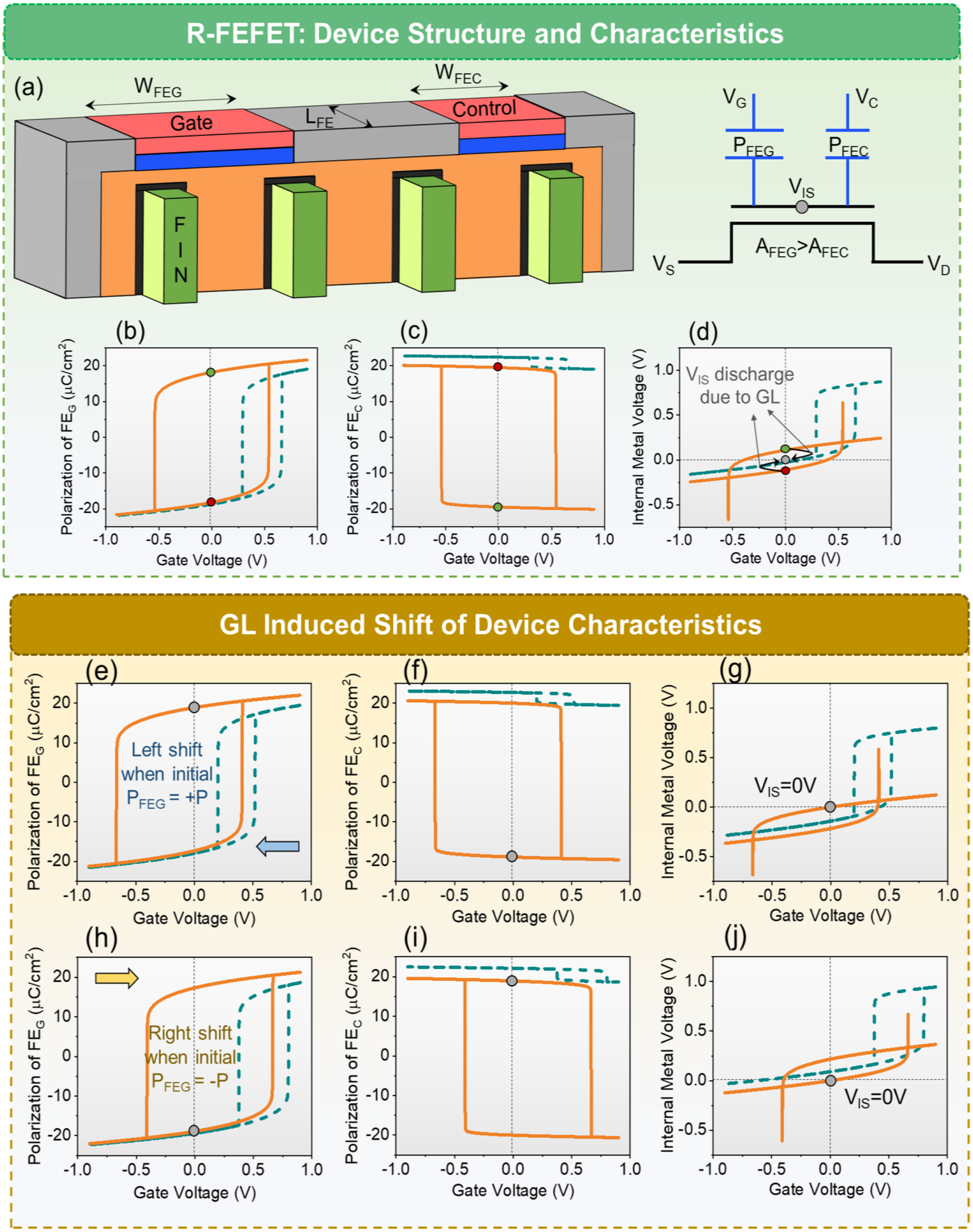


Fig. 3.18 R-FEFET device design and schematic proposed with A_{FEG} (W_{FEG}*L_{FE}) : A_{FEC} (W_{FEC}*L_{FE}) >1. P_{FEG}, P_{FEC} and V_{IS} vs V_{GS} (b, c, d) without GL consideration, (e, f, g) with GL and initial P_{FEG}=+P and (h, i, j) with GL and initial P_{FEG}= -P, respectively.

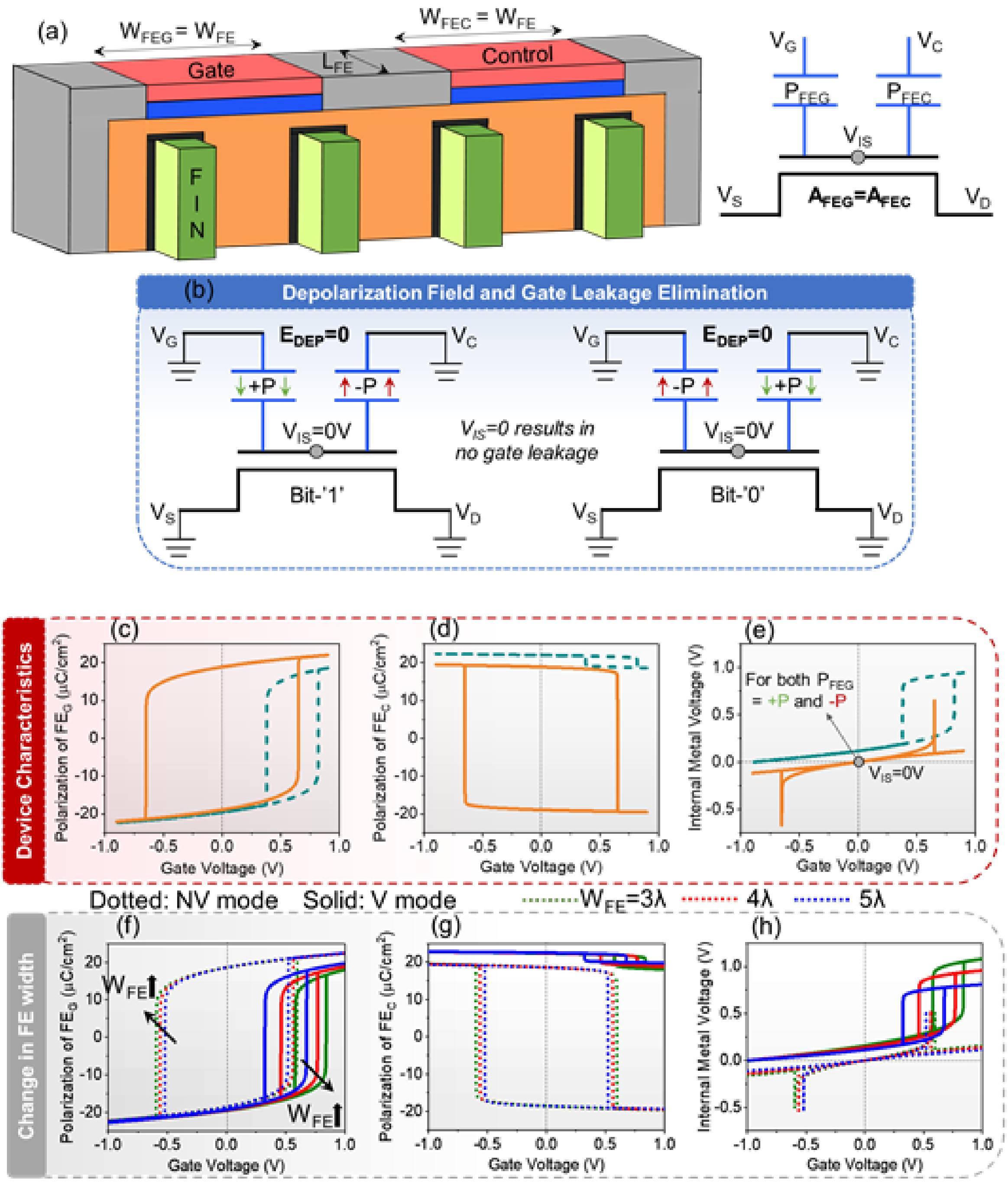


Fig. 3.19 (a) R-FEFET$_{SYM}$ device design and schematic with A_{FEG}: A_{FEC} = 1. (b) Elimination of depolarization fields in 'NV' mode at V_{GS}=0V. (c) P_{FEG}, (d) P_{FEC}, and (e) V_{IS} vs V_{GS} irrespective of initial P_{FEG}. (f) P_{FEG}, (g) P_{FEC}, and (h) V_{IS} vs V_{GS} for varying W_{FE}.

FEFET$_{SYM}$ and R-FEFET$_{ASYM}$ even though they are designed with different A_{FEC}, since both variants are number of fins limited at 4-fins. Now, during the 'NV' mode (V_{CS}=0V), the P_{FEG} is always opposite to the polarization of control stack (P_{FEC}), and they exhibit bi-stability at V_{GS}=0V.

This is because, V_{CS}=0V always results in opposite sign of voltage across FE_C when compared to FE_G [171] (Fig. 3.11(c, d); similar to the explanation in Section-3.3). The value of P_{FEG} is used to store the bit-information in a non-volatile fashion. Now, the symmetric structure of the proposed R-$FEFET_{SYM}$ results in equal and opposite induction of charges (Q/cm^2) on the IML, unlike the R-$FEFET_{ASYM}$. Due to this, the induced IML potential is nullified by both these polarizations, which leads to a situation where V_{IS}=0V at V_{GS}=0V (Fig. 3.11(e)). Such a feature has unique advantages in NVM applications as discussed later. Moreover, as V_{GS} is swept from 0 to +V_{DD} or -V_{DD} (V_{DD}=0.9V), $|V_{IS}|$ increases marginally due to the dielectric response of FE+DE stack. However, for $-V_{DD} < V_{GS} < +V_{DD}$, V_{IS} is mostly < the threshold voltage of the underlying transistor (V_{TH}). Note, during the 'NV' mode, the device undergoes a sequential transient process of polarization switching in FE_G and FE_C as discussed in Section-3.3, which can lead to $|V_{IS}| > V_{TH}$, momentarily (during P_{FE} switching; Fig. 3.19(e)). However, in the steady-state, V_{IS} is always < V_{TH}. In contrast, $V_{IS}\neq$0V (and $|V_{IS}|$ can be > V_{TH}) at V_{GS}=0V in R-$FEFET_{ASYM}$ (Fig. 3.18(d)). Also, the hysteresis window in 'NV' mode (HW_{NV}) is ~1.22X larger in the proposed design due to higher A_{FEC} (with constant A_{FEG}) compared to R-$FEFET_{ASYM}$.

Now, during the 'V' mode, V_{CS}=V_{DD}, P_{FEC} is always = +P. This results in single stable state corresponding to P_{FEG}=-P at V_{GS}=0V (details in Section-3.3). As a result, polarization can be written in FE_G stack based on the V_{GS} applied (-P/+P when V_{GS}=0/V_{DD}). This is unlike FEFET/FEMFETs which require V_{GS}=-V_{DD}/+V_{DD} to store -P/+P. Moreover, similar to the discussion before, the hysteresis window in 'V' mode (HW_V) is larger (~1.2X) than R-$FEFET_{ASYM}$ due to higher A_{FEC}. Also note that, V_{IS} in 'V' mode is higher across $-V_{DD}<V_{GS}<V_{DD}$ when compared to R-$FEFET_{ASYM}$, due to the larger effect of +P in FE_C (as a result of higher A_{FEC}). This results in larger ON state current, I_{ON} (I_{DS} at V_{GS}=V_{DD}), in R-$FEFET_{SYM}$ when compared to R-$FEFET_{ASYM}$ as discussed later. It is important to mention that, the proposed device design inherits the reconfigurability between 'V' and 'NV' modes of operation from R-$FEFET_{ASYM}$ (Fig. 3.19(c)), which results in energy-efficient NVM design, as discussed in Chapter-4.

The drain-to-source voltage (V_{DS}) also plays an important role in the device characteristics. V_{DS} results in a right shift of the device characteristics, which is also observed in R-$FEFET_{ASYM}$ [171]. This is attributed to the fringe capacitance between the IML and the drain-terminal, which results in the increase of V_{IS} with V_{DS}. We observe similar shifts (by ~80mV) for R-$FEFET_{SYM}$ and R-$FEFET_{ASYM}$ at an applied V_{DS} of 0.2V.

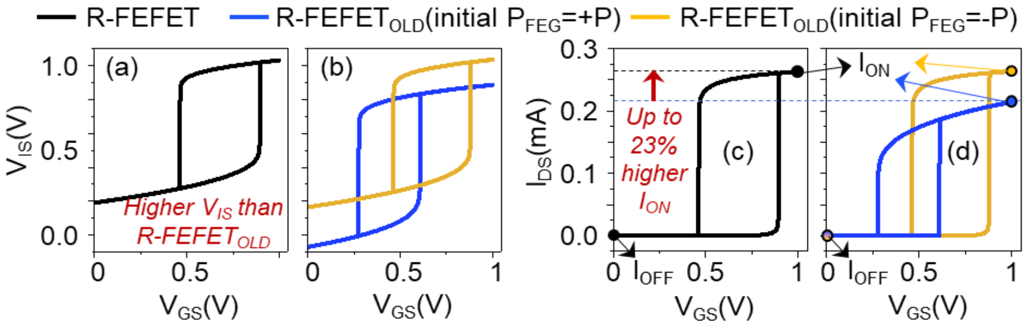


Fig. 3.20 (a, b) V_{IS} vs V_{GS} and (c, d) I_{DS} vs V_{GS} for R-FEFET and R-FEFET$_{OLD}$, considering initial P_{FEG} dependent shifts. V_{DS}=0.2V.

The transfer characteristics in the 'V' mode are illustrated in Fig. 3.20 (c, d). While considering the worst-case GL-induced shift in R-FEFET$_{ASYM}$, R-FEFET$_{SYM}$ achieve up to 23% higher I_{ON}. This is because: (a) higher A_{FEC} in R-FEFET$_{SYM}$ leads to larger influence of P_{FEC}=+P on increasing V_{IS} (and therefore I_{DS}) and (b) in the worst-case scenario, R-FEFET$_{ASYM}$ undergoes left shift (initial P_{FEG}=+P - Fig. 3.18(e-g); Fig. 3.20(b, d)), which further degrades I_{ON} [171]. On an average, i.e., considering both the left and right GL-induced shifts in R-FEFET$_{ASYM}$ (initial P_{FEG}=+P and -P; Fig. 3.20(c, d)), we observe 12% improvement in I_{ON} for R-FEFET$_{SYM}$. The I_{ON}/I_{OFF} ratio (i.e. I_{DS} @ V_{GS}=0V/V_{DD}) is ~10^5 for R-FEFET$_{SYM}$ and ~10^5-10^6 for R-FEFET$_{ASYM}$. The larger I_{ON}/I_{OFF} for the latter is because of its lower A_{FEC}, resulting in lower effect of P_{FEC} = +P in increasing V_{IS} (and hence I_{OFF}) during the OFF state.

3.8.2 Advantages of R-FEFET$_{SYM}$ vs R-FEFET$_{ASYM}$

Since the proposed R-FEFET$_{SYM}$ exhibits V_{IS}=0V in the hold state (all terminal voltages = 0V), GL is reduced in the 'NV' mode of operation. Thus, GL-induced hysteresis shift no longer occurs, which decreases the circuit design complexity. In contrast, in FEFETs and R-FEFET$_{ASYM}$, V_{IS} may ≠ 0V at V_{GS}=0V. In such cases, when gate tunneling current becomes significant, we observe transient shifting of device characteristic based on the initial value of polarization stored, as discussed in the previous section. This results in the consideration of multiple transient states and modes for non-volatile circuit design, thereby affecting the design flexibility. The elimination of GL also enhances the RDM of the proposed R-FEFET$_{SYM}$ NVM as discussed in Chapter-4.

The proposed device structure for R-FEFET$_{SYM}$ also inherits the advantage of minimizing E_{DEP} across FEs from the spilt-gate architecture proposed in [182]. This is because, in the hold mode when all terminals are biased to 0V, V_{IS} is also =0V. This results in zero E_{DEP} (Fig. 3.19(b)) and retention properties can be as high as that of standalone FE capacitors. However, the proposal in [182] is limited to improving the retention of polarization/bit stored in FEFET devices. In contrast, the proposed R-FEFET utilizes the device structure to also perform dynamic reconfigurability between 'V' and 'NV' modes of operation, which enables high density and energy-efficient NVM design.

It is important to note that the proposed R-$FEFET_{SYM}$ operates in the sub-threshold region in the NV mode due to V_{IS}=0V at V_{GS}=0V, resulting in a scenario where the stored data cannot be read-out in the 'NV' mode. However, the polarization which corresponds to the bit-information is retained. We utilize the dynamic reconfigurability offered by R-$FEFET_{SYM}$ and perform sensing of bit-information in the 'V' mode, as discussed later. Using the unique attributes of the proposed R-$FEFET_{SYM}$ design presented in this section, we propose compact, energy-efficient and robust NVM design, in Chapter-4.

3.9 Summary

We proposed a reconfigurable ferroelectric transistor, a unique device technology which has exhibits amalgamation of logic and memory operation modes. We discuss its device structure considering FinFET as well as planar architectures. We explained the unique dual-mode device characteristics including non-volatile and volatile modes of operation, which can be dynamically reconfigured using an external voltage signal. We showed that a hysteresis window modulation from 1.1V in 'NV' mode to 0.3V in 'V' mode can be achieved using the R-FEFETs when designed for 22nm technology node. We illustrated the device design and analysis considering different parameters such as FE thickness, control gate voltage and area of FE stacks. We also showcased the influence of drain voltage on the device characteristics and 10-13% higher ON current achieved by the R-FEFETs compared to regular FETs. We extensively discussed the influence of gate leakage on the device characteristics, which depends on the initial polarization stored in the ferroelectric. We illustrated the multiple modes and states of operation of the R-FEFET in the presence of gate leakage. In the end, we also proposed another device topology called symmetric R-FEFET which exhibits zero depolarization fields in the 'NV' mode and eliminates the issues in regards to the complex influence of gate leakage on the device characteristics. We also discussed the advantages and trade-offs associated with R-FEFET$_{SYM}$ vis-a-vis the initially proposed R-FEFET, and showed that the former achieves 12% higher ON current on average at the cost of distinguishability with respect to the latter. In the following chapters, we discuss how the proposed R-FEFETs can be utilized for the design of energy efficient non-volatile memories and logic.

4. R-FEFET BASED NON-VOLATILE MEMORIES

4.1 Introduction

Conventional silicon based static random-access memory (SRAM) has been used for on-chip applications for the past few decades. However, they face growing challenges with scaling such as short channel effects, increasing leakage and low integration densities [23], [186], [187]. As an alternate, NVMs are poised to revolutionize storage systems to enable efficient, high-performance computing. NAND flash memory has observed an exponential growth in interest over the last decade. This memory technology has partially closed the performance gap that exists between the main memory and secondary storage systems. However, these devices have certain shortcomings such as slow programming compared to S/DRAMs, requirement of large voltages for memory operation and relatively lower endurance. Emerging NVMs such as PCMs, STT-MRAMs and RRAMs are very exciting as they all can offer orders of magnitude higher performance and endurance than what the Flash based storage systems can deliver. However, their current-driven memory operation becomes a bottleneck when considering energy efficient memory-subsystems. Ferroelectric based emerging NVM technologies are an attractive alternate due to their field-driven polarization switching mechanism being extremely energy efficient. Several FE based NVMs have been proposed in the past with their own advantages and drawbacks as discussed in Chapter-2.

In this chapter, we propose compact NVM designs utilizing the intriguing features of the emerging R-FEFET device proposed in Chapter-3. When compared to current-driven memories such as STT-MRAM, PCMs and RRAMs, R-FEFET-NVMs exhibit electric-field driven write, leading to significant energy efficiency [96], [98], [148]. Moreover, due to the possibility of dynamic reconfigurability between the ‘V’ and ‘NV’ modes of operation, the proposed memories avoid the use of negative voltages when compared to some existing FEFET-NVM designs, resulting in low power operation. Their appealing characteristics also help in simplifying the read operation and improve the robustness when compared to the FEFET-NVMs (those with inter-layer metal such as FEMFET) in the presence of gate leakage. These aspects are discussed in detail in this chapter.

4.2 3T-R NVM Design and Operation

The schematic and layout of the proposed R-FEFET based 3T memory (3T-R) are shown in Fig. 4.1(a, b). To begin with, we implement this design employing R-FEFET$_{ASYM}$ designed at 22nm technology node proposed in the previous chapter. The drain and gate of the R-FEFET are connected to the read and write access transistors (RA and WA) respectively. The memory array is formed by connecting WWL and RWL of the cells in a row and WBL, RBL and GND of the cells in a column as shown in Fig. 4.1(c). The control line (CL) shared amongst bits of a word (64-bit in our analysis) is controlled via an inverter driven by Word Enable (WEN) and Enable Bar (ENB). The memory operations are explained below (bias conditions in Table. 4.1).

4.2.1 Memory operation

(a) Write: To write into the cell, the write access transistor (WA; Fig. 4.1(a)) is turned ON (WWL=V$_{DD}$), and WBL=V$_{WRITE}$ = 0/V$_{DD}$ is applied for writing '0'/'1' (HRS/LRS), across the bit cells in the same word. After this, all the R-FEFETs in the accessed word are configured to the 'V' mode by turning ON WEN and setting ENB to 0, which asserts CL of

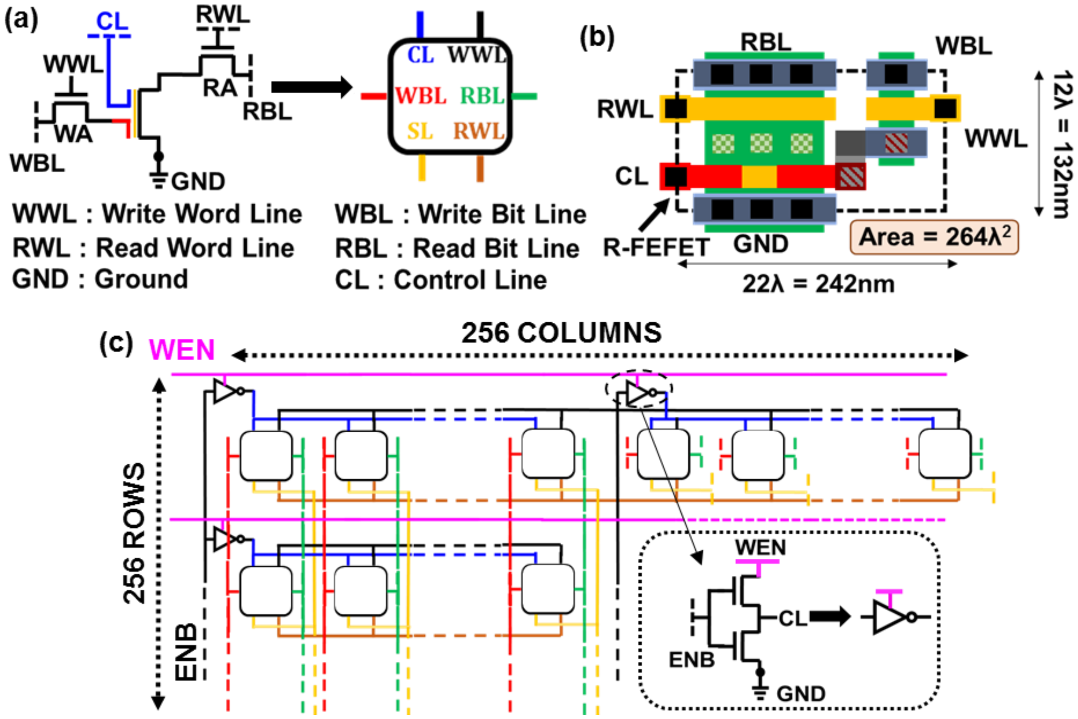


Fig. 4.1 Proposed 3T-R memory: (a) Schematic, (b) Layout and (c) 256X256 Array with 64-bit word.

Table. 4.1 Operating Bias Conditions for 3T-R NVM

	WWL	WBL	RWL	RBL	ENB	WEN
WRITE	V_{DD}	V_{WRITE}	V_{DD}	0	0	V_{DD}
READ	V_{DD}	V_{GREAD}	V_{DD}	V_{READ}	0	V_{DD}
HOLD	0	0	0	0	0	0

the accessed word. Now, R-FEFETs of the accessed word will either configure into mode C or D ('V' modes) depending on the initial state of the R-FEFET as explained in in Chapter-3. The write voltage (V_{WRITE}) is determined by mode D (considering worst case scenario) because of its shift towards the right side due to GL. Depending on the value of WBL (0 or V_{WRITE}), P_{FEA} is set to -P (state 3 or 6) or +P (state 2 or 7) in 'V' mode (as discussed in Chapter-3), which represents the logic state ('0' or '1') stored in the bit-cell. Note, irrespective of the initial conditions (access after short/long time, P_{HOLD}=+P/-P, etc.) the write operation ensures successful programming with simple bias conditions. After every write cycle, CL is de-asserted by driving ENB to V_{DD}, in order to operate the R-FEFET in the 'NV' mode and retain the value of its polarization (P_{FEA}). Note that, the proposed design does not need negative voltages unlike the standard FEFET based 2T and 3T memories [31], [133].

(b) Hold: For holding the value of the bit cell (stored as P), all the signals in the memory are de-asserted which corresponds to the 'NV' mode of operation. In the presence of GL, due to the shifting of device characteristics, we need to consider the following two scenarios for hold stability margin:

(i) Long term hold stability margin: This is defined as $|V_{GS,C}|$ required to switch P when the device has not been accessed for a long time (V_{IS}=0V). From Fig. 3.16, this corresponds to V_{GS} required to disturb state 1 (P_{HOLD}=+P; Mode A) and state 5 (P_{HOLD}=-P; Mode B), whose magnitudes are equal to V_{HM-LT}. This can be viewed as the best case hold margin because, V_{GS} required to switch P is as high as that of a standalone FE capacitor (coercive voltage, $V_C=V_{FE}=V_{GS,C}-V_{IS}=V_{GS,C}$; V_{IS}=0V).

(ii) Short term hold stability margin: This is defined as $|V_{GS,C}|$ required to switch P when a bit cell was accessed previously within a short period of time (such that V_{IS} does not discharge much). From Fig. 3.16, we observe that the $|V_{GS}|$ required to disturb state 4 (P_{FEA}=-P; Mode A) and state 8 (P_{FEA}=+P; Mode B) corresponds to the hold margin for this case, V_{HM-ST}. Recall, immediately after an access, $|V_{IS}|$ is high. This leads to lower $V_{GS,C}$ for P switching, since $V_{FE}=\pm V_C=V_{GS,C}-V_{IS}$. In other words, $V_{HM-ST}<V_{HM-LT}$ (Fig. 3.16). Therefore, V_{HM-ST} can be viewed as the worst-case hold stability margin of the bit-cell.

(c) Read: Read operation of the proposed 3T-R memory is performed by sensing the resistance state of the R-FEFET. Now, if GL can be controlled, i.e., the internal metal potential doesn't discharge over time, then one can perform the read operation using the non-volatile mode of R-FEFETs. However, in the presence of GL and as explained [45], [171], V_{IS} can potentially discharge to 0V over time. This results in loss in the resistance-based distinguishability of the R-FEFET for P_{HOLD} = +P ('1') and -P ('0'), resulting in R-FEFET always operating in HRS in the 'NV' mode (V_{IS} = 0V < threshold voltage of the underlying transistor, V_{TH}). Therefore, access after a long time is the worst case for read distinguishability, which is considered for designing the read biasing scheme. However, recall that the bit information in the form of P remains intact. We use the unique property of dynamic modulation between the 'V' and 'NV' modes in the proposed R-FEFETs to re-establish the resistance-based distinguishability between the states '1' (P_{FEA} = +P) and '0' (P_{FEA} = -P) by restoring V_{IS}. We assert WWL and drive WBL to a voltage V_{GREAD} (explained later). This is followed by assertion of CL and turning ON the read access transistor (RA; Fig. 4.1(a)). First, let us discuss the selection of V_{GREAD} which yields read disturb free operation and then we will explain how V_{IS} is restored and distinguishability is established.

Assertion of CL configures the R-FEFETs of the accessed word in either of the two 'V' modes (C or D) as explained in detail in Chapter-3. The value of V_{GREAD} must be selected such that the desired functionality is achieved for both modes C and D without disturbing the P_{FEA} stored, in spite of being in the 'V' mode. Therefore, for mode C, V_{GREAD} must lie in between V_1 and V_2 (as shown in Fig. 4.2(a)) to enable read disturb free operation and avoid any P switching. Similarly, considering mode D, V_{GREAD} must lie in between V_3 and V_4 (Fig. 4.2(a)). In order to satisfy the above-mentioned requirements for both the 'V' modes (C and D), we choose V_{GREAD} to lie in between V_2 and V_3 ($V_2< V_{GREAD}< V_3$; see Fig. 4.2(b, c)).

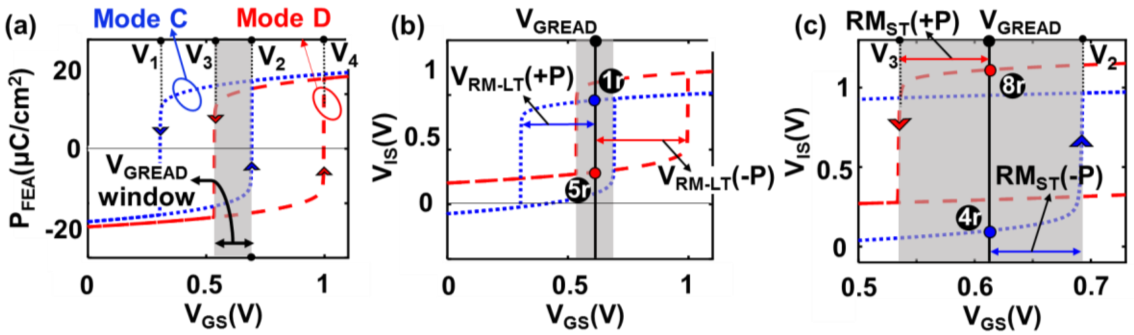


Fig. 4.2 (a) Polarization and (b, c) Internal metal potential versus gate voltage for the volatile mode of operation with P_{HOLD}=+*P* and -*P*

Now, it is important to note that, by configuring the R-FEFET in the 'V' mode and applying V_{GREAD} at the gate terminal, V_{IS} is restored back corresponding to the initial resistance state of the R-FEFET in the bit-cell. To understand this, let us consider the four extreme cases, States 1 and 5 corresponding to when the memory cell is accessed (read) after a long time, such that V_{IS}=0V (due to GL) and States 4 and 8, corresponding to when the cell is accessed immediately after a write operation, where $|V_{IS}| \neq 0V$ (Fig. 3.17).

(i) State-1; Mode-A and State-8 Mode-B (P_{FEA}=+P, P_{FEB}=-P): When CL is asserted along with the application of V_{GREAD} (as discussed before), R-FEFET switches to Mode-C and Mode-D with States '1r' and '8r' respectively (as shown in Fig. 3.16 and Fig. 4.2). During this process, FE_B switches from –P→+P, (due to the 'V' mode since P_{FEB} is always equal to +P; see Fig. 4.3(b)). P_{FEB} switching leads to charging of V_{IS} to a value > V_{TH}, bringing the transistor to LRS corresponding to the bit information stored when P_{FEA}=+P (bit state '1'). This can now be sensed (I_{LRS}) by turning ON the read access transistor and applying a drain voltage (V_{DS}).

(ii) State-4; Mode-A and State-5; Mode-B (P_{FEA}=-P, P_{FEB}=+P): In this case, assertion of CL along with the application of V_{GREAD} brings the R-FEFET to Mode-C and Mode-D with State '4r' and '5r' respectively (see Fig. 3.16 and Fig. 4.2). During this process, there is no switching in FE_B as P_{FEB} is already = +P (Fig. 4.3(a)). Due to the absence of any P switching, V_{IS} remains at a value < V_{TH}, corresponding to the bit state '0' (P_{FEA}=-P). Therefore, the transistor is in HRS which can be sensed (I_{HRS}) by applying V_{DS}.

The ratio between the sensed currents, $I_{LRS}/I_{HRS} \sim 10^4$, thereby giving excellent distinguishability between the bi-stable states. Once read operation in done, we de-assert all the signals along with CL (by driving ENB to V_{DD}) in order to bring it back the R-FEFET to the 'NV' mode, thereby retaining the value of polarization (P_{FEA}) after every read cycle. Fig. 4.3(a) illustrates the simulation waveforms of 3T-R. The transients of the restoration of V_{IS} during read operation is shown in Fig. 4.3(b).

For quantifying the robustness of read operation, we define the read disturb margins. Similar to the hold stability margins defined in the previous section, read disturb margins can be categorized for access after long time and short time. The read disturb margin of state '1r' (P_{FEA}=+P) and state '5r' (P_{FEA}=- P) corresponds to access after long time (V_{RM-LT}; Fig. 4.2(b)), since V_{IS}=0V before the read operation. For these two States ('1r' and '5r') V_{RM-LT} is defined as $|V_{GREAD} - V_{1(4)}|$. The long-term retention properties of the proposed device is similar to isolated FE capacitors. This is because of V_{IS} driving to 0V in the presence of GL. Therefore, for applications in the context of intermittently powered systems (IPS), which operate at low frequencies (~25MHz; [174], [188]), the proposed R-FEFET based memory can utilize the FE retention properties to the maximum extent. On the other hand, for access after short time, the read disturb margin (V_{RM-ST}; Fig. 4.2(c)) corresponds

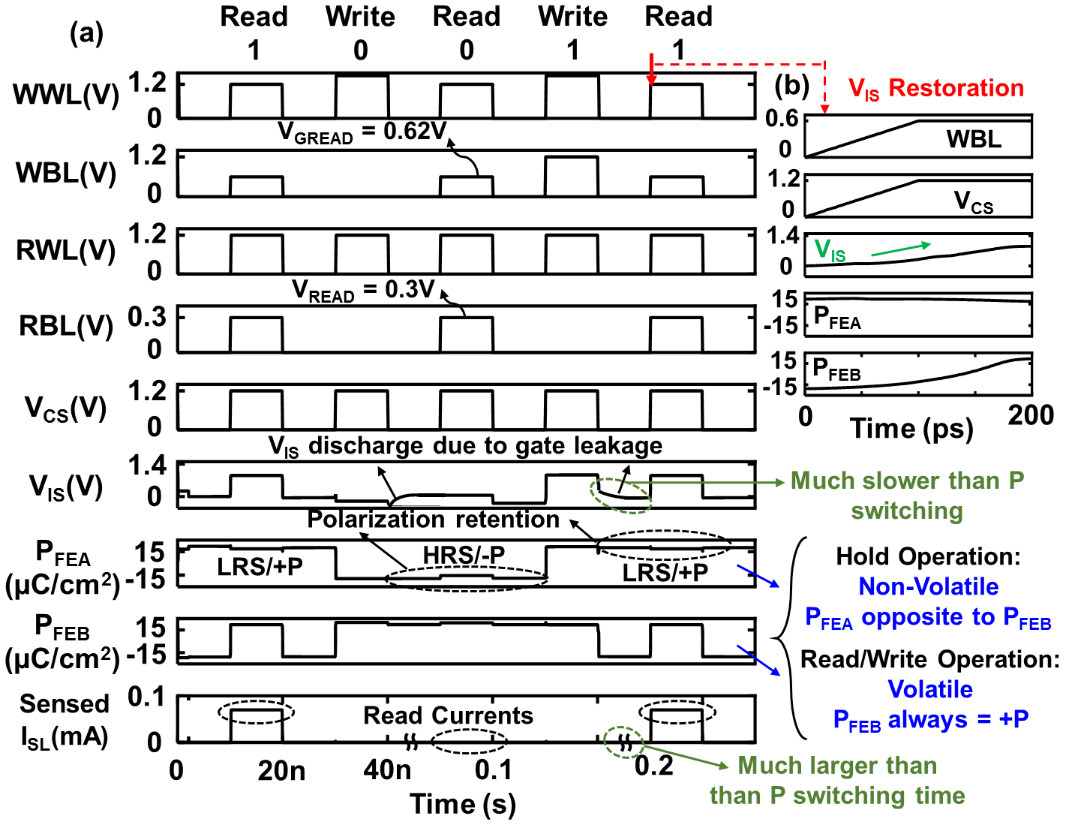


Fig. 4.3 (a) Transient waveforms of the proposed 3T-R memory design (b) V_{IS} restoration transient when P_{FEA}=+*P* and initial V_{IS}=*0V*.

to the states ‘4r’ (P_{FEA}=-P) and ‘8r’ (P_{FEA}=+P). Here, V_{RM-ST} =|V_{GREAD} –$V_{2(3)}$|. Note, it is important to consider both V_{RM-LT} and V_{RM-ST} as they represent the two extreme cases considering GL. Any transient state in between (i.e., 0<|V_{IS}|<V_{IS-MAX}) will have a read-disturb margin in between V_{RM-ST} and V_{RM-LT} (V_{RM-ST}<V_{RM}<V_{RM-LT}). In this section we use V_{GREAD} = 0.62V to achieve the maximum short term and long term read disturb margins (V_{RM-ST}=125mV and V_{RM-LT}=392mV) for FE layer thickness (T_{FE}) =8nm.

Note, V_{RM-ST}<V_{RM-LT} as illustrated in Fig. 4.2 and therefore short-term margins (worst case) should be considered to design the cell for general purpose applications. However, for systems such as IPS, which operate at low frequencies, the subsequent read access after a write operation may happen when the gate leakage has already discharged V_{IS} to 0V. For such systems, LT margins will drive the cell design. Now, recall that the hysteresis window (in ‘NV’ and ‘V’ modes) increases with increasing T_{FE}. This property directly improves the read disturb margins for higher T_{FE} (as shown in Table. 4.2), which can be used to tackle the effects of device dispersion [170]. Moreover, increasing T_{FE} also results in higher V_{HM-ST} and V_{HM-LT} during the hold operation (‘NV’ mode; Table. 4.2). Another design knob for improving margins is the selection of V_{GREAD} bias. For T_{FE}=8nm, we achieve maximum read disturb margins for V_{GREAD}=0.62V as mentioned before. Note, C_{FEA}:C_{FEB} =4:3 for the results shown in Table. 4.2. Decreasing the capacitance ratio has the same influence on the hysteresis width as increasing T_{FE} does (which has been discussed next).

4.2.2 Probability of failures

T_{FE} plays an important role in determining the margins of write and read stability. With increasing T_{FE}, the hysteresis width increases which allows for a larger window (V_2-V_3; Fig. 4.2) for the selection of V_{GREAD} during the read operation as shown in Fig. 4.4(a). On the other hand, increasing T_{FE} also changes the margins associated with the write operation, i.e, V_{P+}= V_{DD}-V_4 for –P→+P switching and V_{P-}= V_1-0V for +P→-P switching (as illustrated for T_{FE}=7nm in Fig. 4.4(a)). To quantitatively understand the influence of variations on the functionality of 3T-R, we perform a variation analysis to determine the probability of failures (P_{FAIL}), where the failure for write is defined as the instance when V_4 > V_{DD} (V_{P+} <0) or V_1 < 0V (V_{P-} <0) and read failure is defined as the instance when V_{GREAD} lies outside the volatile hysteresis window leading to accidental

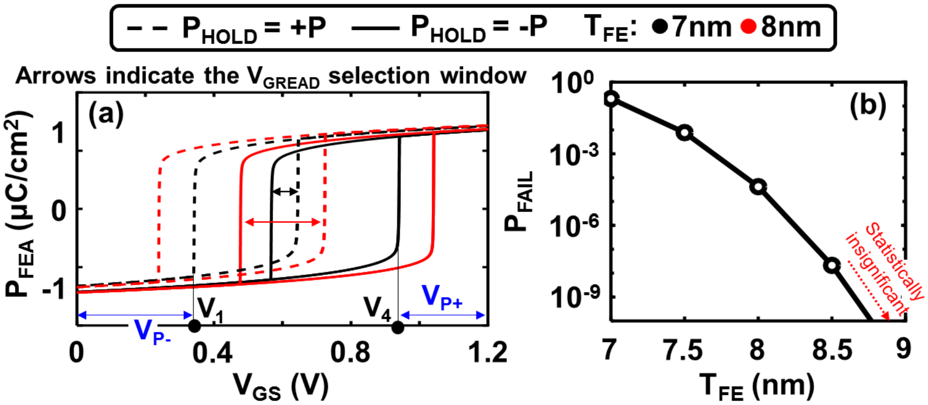


Fig. 4.4 (a) Polarization of FE_A for varying T_{FE} considering P_{HOLD}=+P and –P (b) Probability of failure during memory access considering T_{FE} and V_{TH} variations.

switching of polarization. We consider: (a) variation in threshold voltage (V_{TH}) of the transistor as well as (b) variation in T_{FE} to determine P_{FAIL} (with C_{FEA}: C_{FEB}=4:3). We perform our analysis considering a deviation (6σ) of 180mV for the V_{TH} variations (as reported in [189]), and have assumed 6σ of 1nm for T_{FE} variations for T_{FE} ranging from 7nm to 8.5nm. (Note, that ferroelectric (Hafnium Zirconium Oxide, HZO) layer which is deposited using atomic layer deposition (ALD) process, can achieve "angstrom level precision" [190] due to the self-limiting layer-by-layer process. Therefore, we have assumed 6σ of 1nm). In our work, we design our R-FEFETs to ensure sufficient margins are achieved during the write operation (V_{P+}/V_{P-} ~200mV) for P switching and therefore the dominant component of failure mechanism for the proposed 3T-R is associated with the read operation. Our analysis shows that for T_{FE}=7nm, due to the small read stability margins as shown in Table. 4.2, the P_{FAIL}~0.2 when compared to T_{FE}=8nm which exhibits P_{FAIL}~ 1e-5, due to larger stability margins (as a result of larger hysteresis, Fig. 4.4(a, b)). Note, that with increase in T_{FE}, we achieve lower P_{FAIL}, however this comes at the cost of higher energy for P switching (larger voltage required for P switching [31]). Therefore, while designing memories, the trade-off between P_{FAIL} and energy should be considered.

Table. 4.2 Hold and Read stability margins for various T_{FE}

C_{FEA}:C_{FEB}=4:3	T_{FE}=7nm	7.5nm	8nm	8.5nm
V_{HM-LT}	710mV	765mV	830mV	875mV
V_{HM-ST}	420mV	465mV	510mV	560mV
V_{RM-LT}	292mV	345mV	392mV	442mV
V_{RM-ST}	35mV	80mV	125mV	167mV

4.2.3 Endurance and retention

It may be important to mention that field cycling effects [42] also exist in the proposed R-FEFETs. As discussed in the recently proposed HZO based FE-Metal-FET (FEMFET), the presence of IML enhances the endurance properties of the ferroelectric transistor (~10^{11} cycles) [142]. This is attributed to the minimization of charge trapping in FE and IML, when compared to the device structure without IML. The proposed R-FEFET uses the inherent advantage of the presence of IML (as in FEMFETs) and therefore we expect similar benefits with respect to the endurance of the device as observed in FEMFETs, which is better than HZO based FEFETs (without IML) whose endurance has been reported to be around 10^7-10^8 cycles [170].

Although FEMFETs are very attractive for application in neuromorphic computing as discussed in [142], their application in NVM design is hindered by their low retention property (~2-3 hours [142]). This is because of the discharge of V_{IS} over time due to GL, leading to loss in resistance-based distinguishability as discussed in Chapter-3. To read the bit information stored, unconventional read techniques are required which incur design overheads [45]. On the other hand, the proposed R-FEFET based memory, overcomes this inefficiency by utilizing the dynamic reconfigurability during the read operation. Moreover, the floating IML in the proposed R-FEFET, also undergoes GL, which brings V_{IS} to 0V after a while, during the hold state. Therefore, the voltage across the ferroelectric (V_{FE}) becomes 0 in the stand-by state, resulting in minimal depolarization fields across the FE's present in R-FEFET [123], [182]. Therefore, we expect the retention properties to become as good as that of FE capacitors (since, V_{FE}=0V for FE capacitors in the stand-by state [169], [182]). Several experimental results on HZO based FE have showcased retention of around 10 years [191]. The impact of dynamic modulation on the retention properties and cycling phenomenon requires additional study.

4.2.4 Advantages over standard FEFET based memories

Standard FEFETs with IML also showcase shifting of device characteristics, corresponding to the value of P_{HOLD} due to GL [45]. With proper device design to mitigate the effect of GL, (for instance, designing with higher T_{FE} and/or lower FE capacitor area), standard FEFETs can be used for NVM design, as explained earlier and in [45], [171], [185]. Although the write operation is relatively straightforward (based on WL assertion and driving BL to appropriate voltages), the read operation can be more complex [45]. Unlike R-FEFETs which uses its unique property of dynamic reconfigurability, to switch to the ‘V’ mode, for restoration of V_{IS} (as discussed above), FEFETs need a 2-step operation for reading the bit information [45]. Moreover, since the 2-step read is destructive, a write-back is required. All these steps lead to energy overheads in FEFET based NVMs. On the other hand, read operation in R-FEFETs is disturb free and reconfiguring the device in the ‘NV’ mode after reading, retains the P information, thereby mitigating the complexities faced by FEFETs. Furthermore, the proposed R-FEFETs offer low V_{DD} operation by leveraging gate work-function engineering (GWE) to achieve its ‘V’ mode to operate within 0 to V_{DD} window, which can reduce the energy consumption of various non-volatile circuits. However, GWE doesn’t improve the energy metrics of FEFET based circuits due to the requirement of bi-polar voltages for P switching. For instance, reduction in the voltage for –P → +P switching (with GWE) results in an increase in the voltage for +P → –P switching. In the next section, we analyze the energy metrics of the proposed 3T-R (without GWE for a conservative analysis) and compare it with FEFET based memories.

4.2.5 Memory array analysis

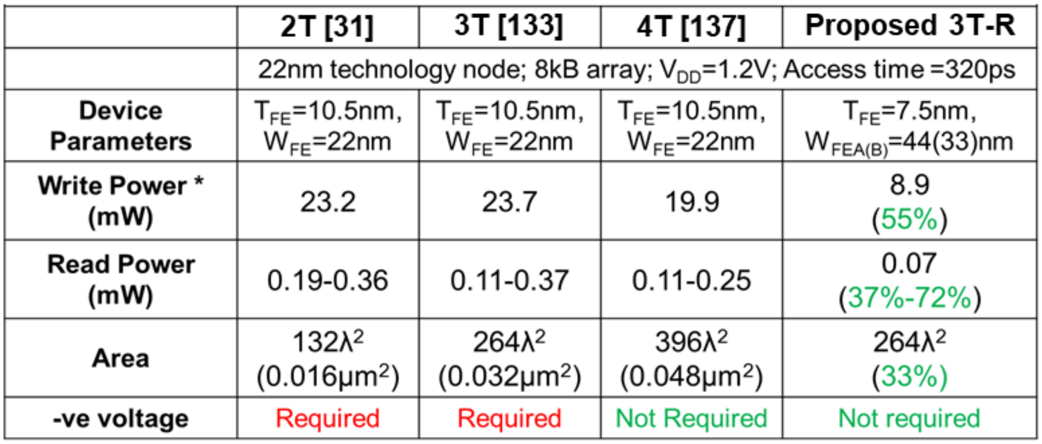
In this section, we compare the proposed 3T-R memory with 2T, 3T and 4T standard FEFET memories [31], [133], [137] with respect to cell area, write power and read power. For fair comparison between the various memories, we ensure that the hysteresis widths are equalized for FEFETs and R-FEFETs (in ‘NV’ mode). At iso-T_{FE} and iso-gate stack area (determined by width of FE stack; W_{FE}), hysteresis width of R-FEFET is larger than FEFET (as explained in [31]). Therefore, to equalize the hysteresis width for this analysis, we optimize T_{FE} and W_{FE} for 2T, 3T, 4T and 3T-R considering iso-width of the underlying transistor, W=110nm (device parameters in Table. 4.3). Note, in our simulations, we scale the bit line/ word line capacitances according to the array area. The 2T and 3T follow the same write bias as mentioned in [31], [133]. Write bias conditions for 4T are mentioned in [137]. Due to GL, the read operation is performed by a 2-step process along with a write back for FEFET memories as mentioned in [45]. Table. 4.3 summarizes the comparison of the proposed 3T-R with the FEFET based 2T, 3T and 4T memory designs, which are discussed as follows.

(a) Write Power: 3T-R offers significant improvement in write power over other designs due to (a) absence of negative voltages for writing into the bit cells unlike 2T and 3T designs and (b) lower number of switching lines compared to 4T memory. Under iso-access time conditions of 320ps, write power reduction of 55%-63% is observed compared to 2T, 3T and 4T FEFET memories.

(b) Read Power: The read operation in 3T-R is a 1-step operation, leading to increase in energy efficiency compared to FEFET memories which require 2-step read. The read power of 3T-R is 37% lower than 4T for reading the state '1'. For reading state '0', the read power of 3T-R achieves 72% improvement over 4T. This is due to the additional energy consumed during the write back operation in standard FEFET memory, because of its destructive read operation. This corresponds to an overall average read power decrease of 46%-86% for the proposed 3T-R compared to the 2T, 3T and 4T memory designs.

(c) Cell area: The 3T-R memory shows 33% lower area compared to the 4T memory design and similar footprint with respect to 3T. Although 3T-R showcases 50% higher area than 2T, their advantages in terms of low read/write power are enormous. Moreover, for certain applications such as IPS, where the main target is to achieve ultra-low power consumption with no major constraint on the hardware footprint [174], the proposed 3T-R can show significant advantages over FEFET based memories.

Table. 4.3 Performance metrics of 2T, 3T, 4T and 3T-R memories



	2T [31]	**3T [133]**	**4T [137]**	**Proposed 3T-R**
	22nm technology node; 8kB array; V_{DD}=1.2V; Access time =320ps			
Device Parameters	T_{FE}=10.5nm, W_{FE}=22nm	T_{FE}=10.5nm, W_{FE}=22nm	T_{FE}=10.5nm, W_{FE}=22nm	T_{FE}=7.5nm, $W_{FEA(B)}$=44(33)nm
Write Power * (mW)	23.2	23.7	19.9	8.9 (55%)
Read Power (mW)	0.19-0.36	0.11-0.37	0.11-0.25	0.07 (37%-72%)
Area	132λ^2 (0.016μm^2)	264λ^2 (0.032μm^2)	396λ^2 (0.048μm^2)	264λ^2 (33%)
-ve voltage	Required	Required	Not Required	Not required

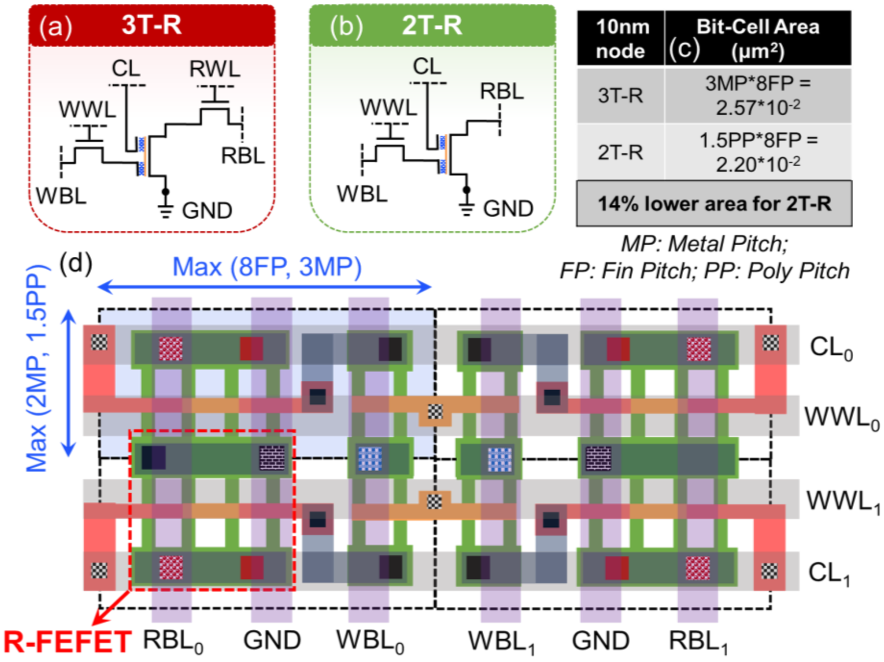
Note: % improvements of 3T-R are mentioned with respect to 4T because both the designs require all positive voltages for their memory operations.
**Write power calculated for a word (32 write-1 and 32 write-0 operations)*

4.3 2T-R NVM Design and Operation

The R-FEFET based memory 3T-R proposed in the previous section, consists of two access transistors in addition to the R-FEFET. Such a design is most beneficial when GL doesn't exist so that the read operation can be performed using the non-volatile mode. Even in the presence of GL, the 3T-R NVM design achieves proper read/write functionalities, but involves stricter and complex design considerations as discussed in the previous section. Therefore, to simply the NVM design, especially when gate leakage cannot be controlled, we proposed another NVM design, 2T-R, which is based on the proposed R-FEFET$_{SYM}$. 2T-R inherently diminishes the effect of GL leading to improved design margins. In addition, we exploit the feature of V_{IS}=0 at V_{GS}=0 in the 'NV' mode to eliminate the read access transistor, yielding a more compact NVM solution than 3T-R. Note, the following evaluations have been considering FinFET based device architecture for R-FEFETs (see- Chapter-3; Section-3.7).

4.3.1 Memory design

We propose an R-FEFET$_{SYM}$ based NVM design, 2T-R, which comprises of one n-type standard FET and one n-type R-FEFET$_{SYM}$ device proposed in Section-3.7. The schematic of 2T-

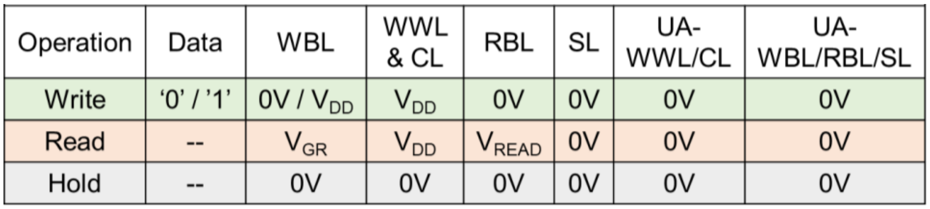


10nm node	(c) Bit-Cell Area (μm²)
3T-R	3MP*8FP = 2.57*10^{-2}
2T-R	1.5PP*8FP = 2.20*10^{-2}
14% lower area for 2T-R	

MP: Metal Pitch; FP: Fin Pitch; PP: Poly Pitch

Fig. 4.5 Schematic of proposed (a) 3T-R and (b) 2T-R NVMs. (c) Bit-cell area comparison of 3T-R and 2T-R. (d) 2X2 layout of the 2T-R array.

Table. 4.4 Operating bias conditions for 2T-R NVM



Operation	Data	WBL	WWL & CL	RBL	SL	UA-WWL/CL	UA-WBL/RBL/SL
Write	'0' / '1'	0V / V_{DD}	V_{DD}	0V	0V	0V	0V
Read	--	V_{GR}	V_{DD}	V_{READ}	0V	0V	0V
Hold	--	0V	0V	0V	0V	0V	0V

R is shown in Fig. 4.5(b). The gate of the R-FEFET$_{SYM}$ is connected to the write access transistor (WA). WA isolates the unaccessed bit-cells in a column and avoids write disturbances. We eliminate the requirement of read access transistor compared to 3T-R (Fig. 4.5(a)). This is because, during read, V_{IS} is 0V at V_{GS}=0V for all the unaccessed bit-cells of a column which are biased in the 'NV' mode. This results in no/insignificant contribution to the current being sensing at RBL, from the unaccessed cells. This is unlike 3T-R, where a read access transistor is necessary to avoid read access failures due to current contribution of the unaccessed cells with V_{IS} > 0V in the 'NV' mode [185].

The layout of a 2x2 memory array layout is shown in Fig. 4.5(d). The write bit-lines (WBL), read bit-lines (RBL) and sense-lines (SL) are shared along the columns. The write word-lines (WWL) are shared along the rows while the control line (CL) is shared for a word using a segmented architecture as discussed in [185].

4.3.2 Memory operation

The operating bias conditions of the proposed 2T-R are mentioned in Table. 4.4.

(a) Read: For reading the bit-information from 2T-R NVM, we assert CL of the accessed row and activate the 'V' mode of the R-FEFET$_{SYM}$. This is because, in the 'NV' mode, the device operates in the sub-threshold region (V_{IS}<V_{TH}; as discussed before) and therefore, the data cannot be read-out using drain current of R-FEFET$_{SYM}$. Now, while asserting CL, it's important to bias V_{GS} of R-FEFET$_{SYM}$ in-between the HW_V (say V_{GR}), in order to avoid any disturbance to the bit-stored. This is achieved by turning ON WWL and asserting WBL to V_{GR}. Biasing the gate of R-FEFET$_{SYM}$ at V_{GR} also determines V_{IS} based on P_{FEG}/P_{FEC} stored in the 'NV' mode, before the start of sensing operation. When bit-'0' is stored (P_{FEG}=-P and P_{FEC}=+P), asserting CL and biasing the V_{GS}=V_{GR} does not change any of the FE polarizations (because they are already in their desired states), resulting in V_{IS}~0V (<V_{TH})

and R-FEFET$_{SYM}$ in HRS. However, when bit stored is '1' (P_{FEG}=+P and P_{FEC}=-P), asserting CL switches the P_{FEC} to +P, resulting in a buildup of V_{IS} to a value > V_{TH} (by design). This brings the R-FEFET$_{SYM}$ to the LRS. The P_{FEG}-dependent resistance state thus re-established, can be read by applying V_{READ} at RBL and sensing the current (I_{LRS} (I_{HRS}) for P_{FEG} = +P (-P)). This enables a non-destructive read unlike FEFET-NVMs as also discussed in the previous sub-section. For the unaccessed cells in the column, we apply 0V to CL and WWL, and for the unaccessed cells in the row, we apply 0V to CL and WBLs. These bias conditions retain P_{FEG} in the 'NV' mode of their R-FEFET$_{SYM}$ from 2T-R NVM, we assert CL of the accessed row and activate the 'V' mode of the R-FEFET$_{SYM}$.

Let us now present a qualitative comparison of the proposed R-FEFET$_{SYM}$ based 2T-R with R-FEFET$_{ASYM}$ based 3T-R NVM. Firstly, the read current in 2T-R is higher than 3T-R due to the following reasons:

- R-FEFET$_{SYM}$ intrinsically offers higher I_{ON} due to larger A_{FEC} and lack of GL induced shift in the hysteresis (as discussed in Chapter-3) when compared to R-FEFET$_{ASYM}$.
- Only one transistor in the read path of 2T-R versus two transistors present in 3T-R NVM.
- The RBL voltage in 3T-R undergoes a resistive-divider action (due to two transistors in series), resulting in the drain-terminal of R-FEFET$_{ASYM}$ receiving a voltage <V_{READ}. In contrast, the drain voltage of R-FEFET$_{SYM}$ in 2T-R = V_{READ}, which also contributes to the higher current being sensed.

The higher read current sensed in 2T-R also translates to higher sense margin (at iso-V_{READ}). This allows for lowering of V_{READ} to achieve read energy savings, as we discuss later. Secondly, since the V_{DS} of R-FEFET$_{SYM}$ in 2T-R is higher than that of R-FEFET$_{ASYM}$ in 3T-R at iso-V_{READ} applied at RBL, the device characteristics undergo a larger right-shift for the former than the latter (see Chapter-3). This results in the requirement of higher V_{GR} for 2T-R in comparison with 3T-R (Fig. 4.6). Lastly, the RDM which is defined as the minimum margin between V_{GR} and the edge of polarization switching in the 'V' mode [185] (Fig. 4.6), is larger for 2T-R compared to 3T-R. This is because, in 3T-R, due to GL, the overlapped region of 'V' modes considering initial P_{FEG} = +P and -P determines the RDM (Fig. 4.6(b, c)). While in the proposed 2T-R, the entire 'V' mode hysteresis window is

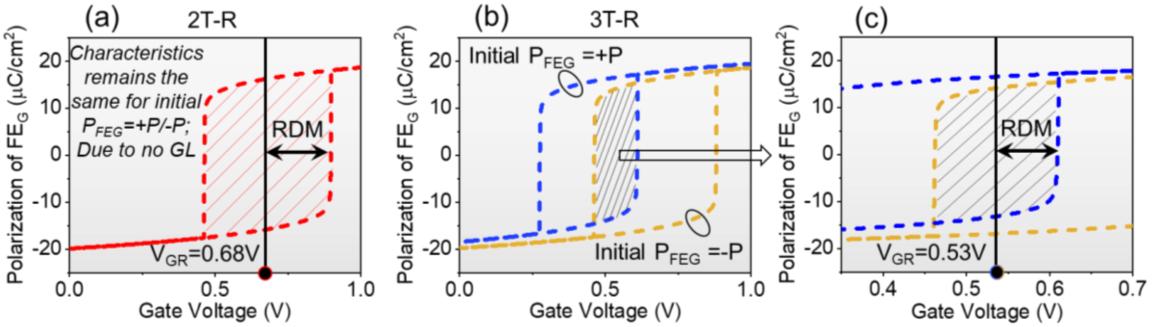


Fig. 4.6 P_{FEG} vs V_{GS} in the volatile mode for R-FEFET in (a) 2T-R and (b, c) 3T-R NVMs (for both initial P_{FEG} = +P and -P). Shaded region is the V_{GR} selection window. V_{READ}=0.2V.

considered for RDM evaluation, since GL is zero (Fig. 4.6(a)). In other words, 2T-R exhibits a larger window for V_{GR} selection when compared to 3T-R, and therefore exhibits improved robustness as discussed quantitatively later.

(b) Write: The write operation of the proposed 2T-R NVM is similar to 3T-R NVM. We first assert CL of the accessed row in order to bring the R-$FEFET_{SYM}$ to the 'V' mode. WBL is driven to a voltage corresponding to P_{FEG} to be stored in the accessed R-$FEFET_{SYM}$ (WBL=0/V_{DD} for P_{FEG}=-P/+P (bit-'0'/'1')), followed by WWL assertion. The feature of dynamic reconfigurability to the 'V' mode of operation enables unipolar write voltages, leading to energy efficient write operations for 2T-R (as well as 3T-R), when compared to bi-polar voltage requirements/two-phase write in FEFET-NVMs, as discussed quantitatively later. For the unaccessed cells in the column, we apply 0V to WWLs and CLs, and for the unaccessed cells in the row we apply 0V to CL and WBLs. This results in their R-$FEFET_{SYM}$ operating in the 'NV' mode and avoids write disturbs. Note, during write, all RBLs are driven to 0V.

In comparison with 3T-R, 2T-R exhibits relaxed design constraints for the selection of V_{DD} during write operation. This is because, in 3T-R, minimum V_{DD} depends on the maximum of the coercive voltage required for P_{FEG} switching from -P→+P considering the worst-case GL-induced hysteresis shifts [171], [185] (Fig. 3.18). On the other hand, in 2T-R, no hysteresis shifting occurs and hence, (Fig. 3.19) minimum V_{DD} for successful write operation is a function of the intrinsic device characteristics only (unaffected by GL). Now, considering the worst-case GL-induced hysteresis shift (initial P_{FEG}=-P) of R-$FEFET_{ASYM}$ in 3T-R and also the larger HW_V of R-$FEFET_{SYM}$ in 2T-R (at iso-T_{FE}), we observe that the minimum V_{DD} for both the memory designs are similar to each other (Fig. 3.18, 3.19).

(c) Hold: In In the hold/stand-by mode of operation, all metal-lines in the 2T-R array are brought to 0V. This ensures 'NV' mode for R-FEFET$_{SYM}$ of all the bit-cells, with V_{GS}=0V. In this condition, P_{FEG} which stores the bit information, is retained in a non-volatile fashion as discussed earlier. Also, due to the property of V_{IS}=0V at V_{GS}=0V in the 'NV' mode of R-FEFET$_{SYM}$, they experience lower E_{DEP} resulting in improved retention characteristics in 2T-R. In contrast, 3T-R designed with R-FEFET$_{ASYM}$ can experience E_{DEP} until GL (if present) reduces V_{IS} to 0V, which impacts the retention properties [182], [185].

4.3.3 Comparison with previous FE(M)FET/R-FEFET NVMs

The proposed NVM has minimal design complexity because of the elimination of GL unlike in 3T-R which requires the consideration of multiple intermediate states and operation nodes for NVM design as discussed in the previous section, due to GL-induced hysteresis shifts (Fig. 3.18 (e-j)). Moreover, the RDM of 2T-R has no dependency on the initial P_{FEG}, leading to robust read operations when compared to 3T-R. Furthermore, the compact bit-cell footprint of 2T-R leads to higher density. When compared to FEMFET-NVMs, which require complex and power hungry 2-step read operation [45], 2T-R overcomes the challenges with a simple 1-step operation as discussed before in this section. Additionally, 2T-R requires unipolar and single-phase write operation unlike several other FE(M)FET-NVMs [31], [64], [133], [192]. All these attributes lead to significant benefits compared to previously proposed FEFET, FEMFET and R-FEFET based NVMs. In the following sections, we present the quantitative analysis of the memory operations, robustness of sensing and impact of variation in 2T-R.

4.3.4 Memory array analysis

In this section, we evaluate the proposed 2T-R versus 3T-R in terms of performance, energy and the robustness of read operation. Following this, we perform the evaluation of 2T-R versus 4T FEFET-NVM [137].

(a) 2T-R versus 3T-R (Fig. 4.7(b)): The geometry and device parameters used for this analysis are illustrated in Fig. 4.7(a). Note, V_{GR} (i.e., WBL voltage during read) is fixed considering the maximum RDM (at iso-V_{READ}) possible for both the designs. (V_{GR}=0.53V for 3T-R and 0.68V for 2T-R). 32kB NVM array is considered with V_{DD}=0.9V and V_{READ}=0.2V.

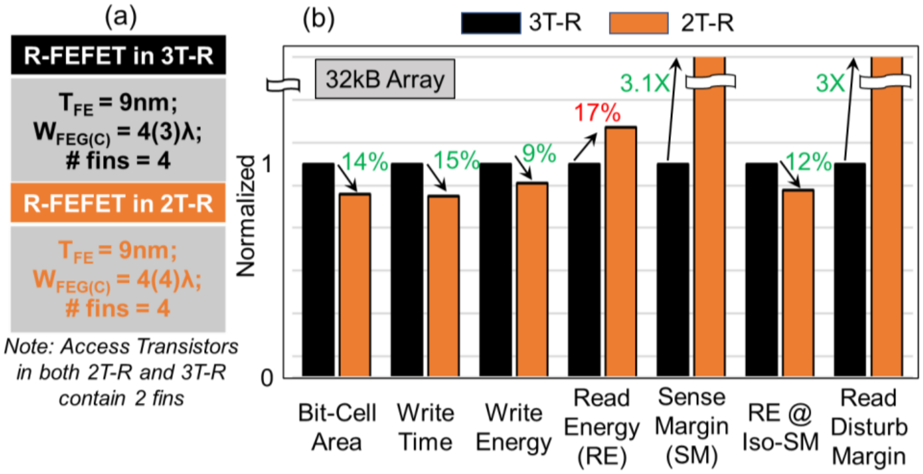
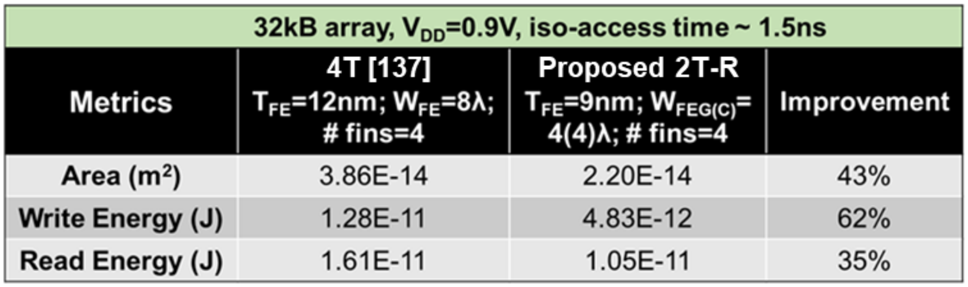


Fig. 4.7 (a) Device parameters used for 3T-R and 2T-R NVMs. (b) Array-level results of 3T-R and 2T-R for various metrics.

Due to the elimination of the read access transistor, 2T-R exhibits 14% lower bit-cell area compared to 3T-R (Fig. 4.5(c)). The write delay and energy of 2T-R is 15% and 9% lower respectively compared to 3T-R. This is attributed to the reduced energy associated with bit-line capacitance charging due to lower NVM array area. The read energy of 2T-R is 17% worse than 3T-R (for the same V_{READ}). This is because of (a) higher V_{GR} (for maximum RDM) and (b) larger current during the sensing operation in 2T-R (as explained earlier). However, it is important to note that the higher read current also results in 3.1X improvement in sense margin for the proposed 2T-R. Therefore, considering iso-sense margin analysis, which is achieved by reducing the V_{READ} for 2T-R, we observe 12% improvement in read energy for the proposed NVM. A key point to note is that the RDM of the proposed NVM is ~3X larger than 3T-R. This is again attributed to the absence of GL-induced hysteresis shift, which reduces the RDM in 3T-R. Higher RDM ensures more robustness of the bit-cell to unwanted flipping of bit-information, as discussed in more detail in the next section.

(b) 2T-R versus 4T: The benefits of 3T-R over FEFET-NVMs for planar technology has already been discussed eariler. Here, we briefly carry out the analysis for 2T-R vs 4T FEFET-NVM in the context of FinFETs. We consider 4T FEFET-NVM for fair comparison because of its unipolar voltage-based array design and single-phase write operation similar to R-FEFET based NVMs. The following analysis has been performed considering a 32kB array with V_{DD}=0.9V, V_{READ}=0.2V. The geometry/device parameters are illustrated in

Table. 4.5 Array-level comparison of 2T-R vs 4T NVM



32kB array, V_{DD}=0.9V, iso-access time ~ 1.5ns			
Metrics	4T [137] T_{FE}=12nm; W_{FE}=8λ; # fins=4	Proposed 2T-R T_{FE}=9nm; $W_{FEG(C)}$= 4(4)λ; # fins=4	Improvement
Area (m^2)	3.86E-14	2.20E-14	43%
Write Energy (J)	1.28E-11	4.83E-12	62%
Read Energy (J)	1.61E-11	1.05E-11	35%

Table. 4.5. Note, different geometries are used to obtain the same HW in the 'NV' mode of R-FEFET$_{SYM}$ and FEFET.

Due to the dynamic reconfigurability between 'V' and 'NV' modes exhibited by the proposed R-FEFET$_{SYM}$, which enables a 2T-R design with all positive V_{GS} for polarization switching, they present 43% lower area compared to the 4T design. Moreover, 2T-R achieves 62% higher write energy efficiency compared to 4T due to (a) lower bit-cell area and (b) the dynamic reconfigurability, which enables the possibility of biasing V_{GS} (of R-FEFET$_{SYM}$) to 0V for writing bit- '0' in the 'V' mode. The write access time remains similar for both NVMs (as shown in Table. 4.5). On the other hand, due to the non-destructive, 1-step read operation in 2T-R, it exhibits 53% lower read energy and 65% lower read access time over 4T.

4.4 Variation Analysis of R-FEFET based NVMs

The key characteristics of the proposed R-FEFET$_{SYM}$ rely on the fact that there exists equal and opposite effect of FE$_G$ and FE$_C$ on the induction of charge in the IML. Recall, this is because the two FE stacks are designed with equal area. However, the question is, *what if the effect of these two stacks do not cancel each other (i.e., $V_{IS}\neq 0V$ at $V_{GS}=0V$) due to variation in FE$_{G/C}$ stack areas?* Then, if $V_{IS} \neq$ 0V, it is important to understand the impact of current contribution from the unaccessed (UA) cells of a column during sensing operation. On the other hand, RDM, which depends on the selection of V_{GR} and HW$_V$, may get degraded when the device-level variations are considered, since variations in V_{TH} and $W_{FEG/C}$ can shift or modify the device characteristics. We study all these effects for the proposed 2T-R NVM in this section. Note, we neglect T_{FE} variations in HZO since they can be controlled very well with the recent advancement in atomic layer deposition (ALD) techniques [190].

(a) Leakage from unaccessed cells during read: The advantage of eliminating read access transistor in 2T-R and achieving high memory density compared to 3T-R comes from the fact that V_{IS}=0 at V_{GS}=0V in the 'NV' mode of R-FEFET$_{SYM}$. This enables us to perform read operation without the unwanted contributions to the read current from the UA cells. However, if $W_{FEG/C}$ variations are considered then V_{IS} might be > 0V, leading to the sullying of the read current due to the contributions from the UA cells. This limits the number of bit-cells which can be placed in a column. To understand this, we consider a Gaussian distribution of $W_{FEG/C}$ variations with standard deviation (σ) of 15% of its mean value (4λ). We also consider V_{TH} variations with σ(V_{TH}) = 20mV (Gaussian; [193]) in the following analysis. Fig. 4.8 (a) illustrates the drain current (I_{DS} at V_{GS}=0V and V_{DS}=0.2V) of R-FEFET$_{SYM}$ in the 'NV' mode for 1000 Monte Carlo samples, with maximum/minimum current (I_{WORST}/I_{BEST}) = 0.27nA/ 5.89pA. Fig. 4.8(b) illustrates the distinguishability with respect to number of cells in a column (#CC) where distinguishability is defined as the ratio of I_{LRS}/I_{HRS} considering the contributions from the UA cells (see Equation - 4.1).

$$Distinguishability = \left(\frac{I_{LRS}}{I_{HRS} + (\#CC - 1) * I_{UA}}\right) \qquad (4.1)$$

The shaded region in Fig. 4.8(b) illustrates the range of distinguishability bounded by worst- and best-case scenarios of leakage from UA cells (I_{UA}) i.e. I_{UA}=I_{WORST} and I_{UA}=I_{BEST},

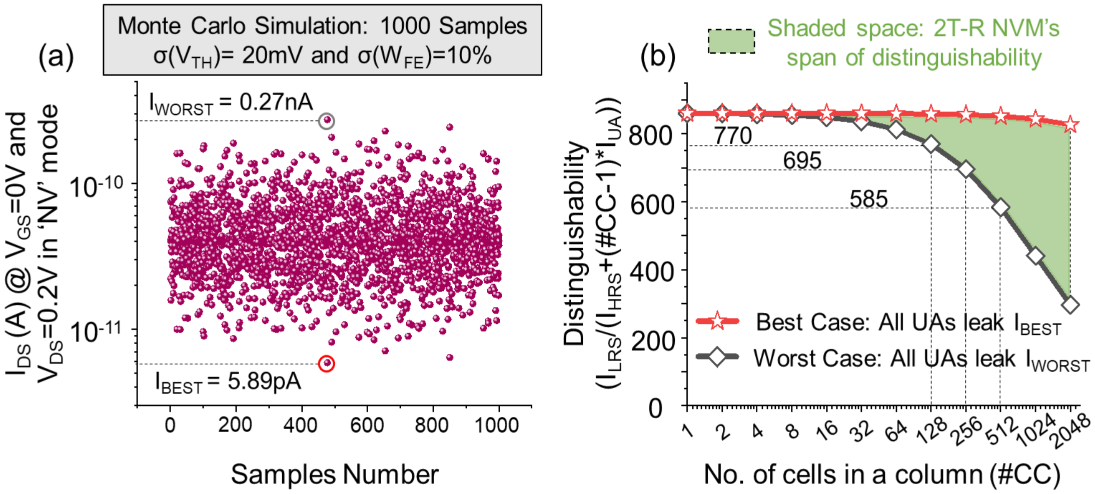
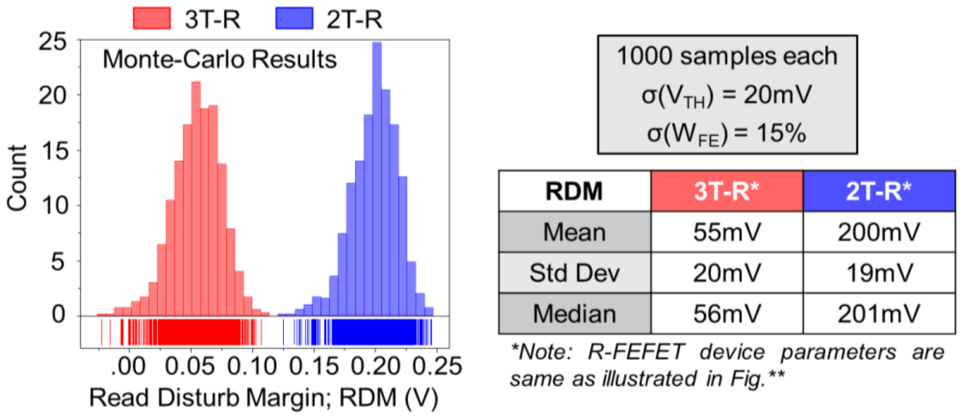


Fig. 4.8 Monte Carlo simulation result of I_{DS} for 1000 R-FEFET$_{SYM}$ device samples with σ(V_{TH})=20mV and σ(W_{FE})=15% (b) Plot of distinguishability vs #CC depicting the proposed 2T-R NVM's span of distinguishability.

respectively. Without variations, distinguishability = 825 is obtained. For the worst-case scenario, where all UA cells in a column sink in I_{WORST} from RBL, distinguishability decreases with increase in #CC. For #CC=128, distinguishability = 770 and for #CC=512, distinguishability = 585. Therefore, even in the presence of the variations, the proposed design offers sufficient distinguishability and the leakage in the UA cells does not affect the read functionality.

(b) Impact on read disturb margins: Here, focus on the RDM, which is important to understand because the read operation critically relies on the WBL biasing to V_{GR} in-between the 'V' mode hysteresis. Firstly, V_{TH} variations can cause the device characteristics to shift and reduce RDM (since V_{GR} is fixed). Secondly, as also discussed in the previous sub-section, $W_{FEG/C}$ variations can result in a situation where $V_{IS} \neq 0V$, causing mild GL-induced hysteresis shift and RDM degradation. We evaluate the impact of these device-level variations on RDM for 2T-R and compared it with 3T-R, next.

Fig. 4.9 illustrates the Monte-Carlo analysis performed considering 1000 samples of 3T-R and 2T-R with variation in V_{TH} of the transistors present in the bit-cell and W_{FE} of R-FEFETs. As before, we consider a Gaussian distribution of V_{TH} with $\sigma(V_{TH})$ = 20 mV and $\sigma(W_{FE})$ = 15% of its mean value. We observe that the 2T-R exhibits much higher RDM (>124mV) compared to 3T-R (RDM <110mV). This is attributed to (a) the mitigation of GL-induced hysteresis shifts and (b) intrinsically larger HW_V of R-FEFET$_{SYM}$ (in 2T-R) compared to R-FEFET$_{ASYM}$ (in 3T-R). Note that, RDM of 3T-R can be increased using larger T_{FE} (which increases HW_V [31]). However, this comes with the requirement of larger



1000 samples each
$\sigma(V_{TH})$ = 20mV
$\sigma(W_{FE})$ = 15%

RDM	3T-R*	2T-R*
Mean	55mV	200mV
Std Dev	20mV	19mV
Median	56mV	201mV

Note: R-FEFET device parameters are same as illustrated in Fig.*

Fig. 4.9 Monte Carlo simulation results of RDM for 1000 samples of each 2T-R and 3T-R NVMs with $\sigma(V_{TH})$=20mV and $\sigma(W_{FE})$=15%.

V_{DD} and therefore higher write energy. In contrast, for 2T-R, similar RDM can be achieved with lower T_{FE}, leading to reduced constraints on V_{DD} selection and energy costs.

4.5 Summary

We proposed two variants of non-volatile memory designs using the proposed R-FEFETs. The first design, 3T-R was based on R-FEFET$_{ASYM}$. We described the memory operations of 3T-R in detail considering the presence of gate leakage. We discussed how 3T-R elegantly overcomes the drawbacks of gate leakage in FEFET based NVM using its unique feature of dynamic reconfigurability. We showed that 3T-R achieves 55-63% and up to 46-86% lower write and read power, respectively, when compared to existing FEFET based 2T, 3T and 4T memory designs. To further enhance the energy efficiency, especially in cases when GL cannot be controlled, we proposed another NVM 2T-R using R-FEFET$_{SYM}$ device design. We discussed the unique benefits of 2T-R design over 3T-R such as lower depolarization fields in the NV mode of operation and diminishing gate leakage effects for less complex read operation. We showed that 2T-R exhibits up to 12% higher energy efficiency along with 3X increase in the RDM and 14% lower area compared to 3T-R. We also performed a detailed variation analysis and studied the influence of threshold voltage and FE width variations on the RDM for both 2T-R and 3T-R, and showcased the higher robustness of the former over the latter.

The appealing attributes of R-FEFETs translated to superior performance of NVMs when compared to standard FEFETs. However, their applications can be much broader due to their unique logic-memory coupling. With that in mind, let us turn our attention to integrating non-volatility within logic (specifically, flip-flops) using R-FEFETs in the next chapter. In Chapter 6, we will explore how logic computations can be performed within R-FEFET based NVMs.

5. R-FEFET BASED NON-VOLATILE FLIP-FLOPS

5.1 Introduction

Energy harvesting from ambient sources has been extensively studied as a promising candidate to enable energy autonomous systems. In the near future, it is predicted that a number of systems will be powered using harvested energy including, toxic gas sensors, portable gadgets and medical implants [194], [195]. However, scavenged energy from ambient sources such as solar, thermal and vibration exhibits an erratic nature with intermittent power supply (V_{DD}). Such power failures have a drastic impact on standard CMOS logic, suffering from inefficient reboots and rollbacks [196]. Therefore, it becomes important to back up the state of a logic system to alleviate the loss in computation progress.

Non-volatile computing is an emerging solution to mitigate computation progress loss due to unexpected power failures [196]. Systematic consistency-aware check-pointing mechanisms have been proposed to avoid data inconsistency and computation errors due to power failures [197]. This is achieved by backing up the states of a circuit such as, on-chip memory, flip-flops (FFs) and registers into an off-chip non-volatile memory (NVM). However, this incurs significant energy/delay overheads due to long distance data transmissions and constrained parallelism.

Embedded non-volatile computing is an attractive alternate method to backup the computation states into a local on-chip NVM, eliminating the transmission overheads. Several NVFF designs using memristors [83], magnetic tunnel junctions [81] and resistive RAMs [84] as local non-volatile elements have been proposed with on-demand backup/restore (B/R). However, they may incur area overheads due to the incorporation of a B/R module[81]. Also, high write current during backup increases their power consumption [84]. Ferroelectric (FE) capacitor-based designs have also been proposed, utilizing their property of polarization (P) retention in the absence of electric field (E) [85]. However, low distinguishability between their non-volatile states degrades the robustness during restore operation.

The recent advent of ferroelectric transistors (FEFETs) with the possibility to integrate FE in the gate stack of a transistor has led to a new era for logic-memory synergy [44]. Lately, FEFET based NVFFs have been proposed with a potential to overcome the challenges in FE capacitor-based designs [80], [145]–[147]. The innovation stems from utilizing the three-terminal non-volatile transistor to improve distinguishability and simplify the restore operation. However, these designs also contain a B/R module driven by control signals, leading to higher switching energy/delay. Hence, there is a need to optimize the B/R module by exploring new device technologies that leverage the opportunities offered by FE.

To that end, the R-FEFET we presented in Chapter-3 exhibits a unique characteristic of dynamic tuning between volatile and non-volatile modes. Exploiting such distinct features, we propose 2 variants of energy-efficient NVFF designs (referred to as RNVFF-1 and RNVFF-2).

5.2 Intermittently Powered Systems

An intermittently powered systems (IPS) possess the ability to execute a program across multiple power-ON periods in the presence of sporadic and highly unreliable energy supplied from the energy harvester [196]. IPS executes long running computations in small increments by checkpointing the system state (processor registers, peripheral registers, and program state in local on-chip memory) to a non-volatile memory during power loss and later restoring the checkpointed

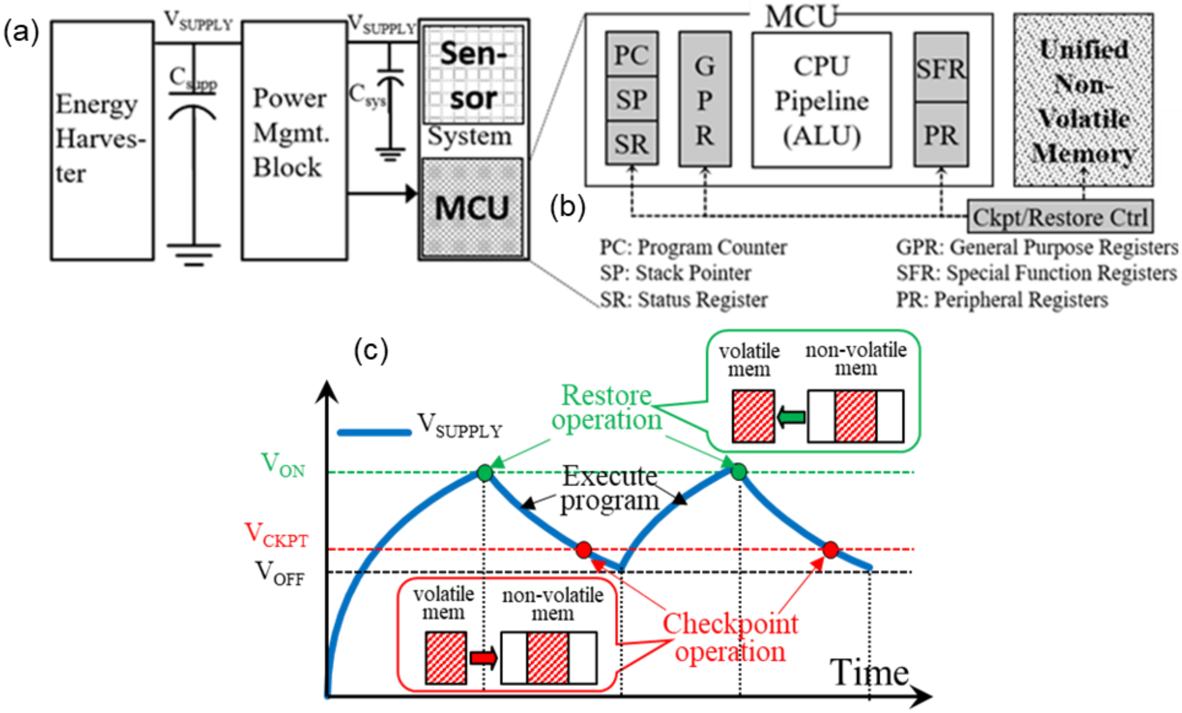


Fig. 5.1 (a) Conceptual diagram of an intermittently powered system (IPS) (b) MCU core registers and unified NVM mapped to R-FEFET-based NVFFs (RNVFF-1 and RNVFF-2) and NVM (3T-R) respectively (c) Energy profile from an energy harvesting source.

system state from the NVM when the system has sufficient supply energy. A high-level conceptual diagram of an IPS is shown in Fig. 5.1(a) and Fig. 5.1(c) shows a typical power/voltage profile obtained from the energy harvesting source. Here, C_{SUPP} is the input capacitor that stores ambient energy. Whenever, the V_{SUPP} reaches V_{ON} (ON voltage as show in Fig. 5.1(c)), the system has enough energy to begin operation. However, when the V_{SUPP} goes below V_{OFF}, the system switches off due to insufficient energy. Note that V_{CKPT} is the supply voltage at which a checkpoint is triggered and is conservatively set to a value that can enable checkpointing the least amount of system states that is required to enable forward progress. As can be seen in Fig. 5.1(c), a restore operation is required to be performed at the beginning of each power ON or execution cycle before the system can resume executing from the previously checkpointed state without restarting the entire process all over again. Note that the power management block shown in Fig. 5.1(a) has dual functions of providing V_{SUPP} to the IPS whenever $V_{SUPP} > V_{ON}$. In addition, it is also responsible to initiate the checkpoint operation at $V_{SUPP} = V_{ON}$. A hysteresis is always built into the system architecture such that system turns on only whenever there is enough energy for the system to restore, execute, and checkpoint successfully. The value of C_{SUPP} is critical design parameter for an IPS. C_{SUPP} determines the total amount of energy available in each power cycle, i.e., $½*C_{SUPP}*(V_{ON}^2-V_{OFF}^2)$. Thus, for a given power profile, a smaller value of C_{SUPP} results in smaller energy per power cycle leading to larger number of checkpoints/restores required to execute a fixed computational task.

Fig. 5.1(b) provides a more detailed look within the microcontroller unit (MCU) architecture of the IPS. It highlights the elements that needs to be mapped to NVFFs and to NVM for checkpointing and restoring of system state. As shown, MCU core registers such as the Program Counter (PC), Status Register (SR), Stack Pointer (SP), General Purpose Registers (GPR), Special Function Registers (SFR), and Peripheral Registers (PR) are all mapped to NVFFs and the local on-chip unified memory is mapped to an NVM to ensure forward progress even in the presence unreliable energy supply. Note that, lower the access and the checkpointing/restore energies of NVFF and NVM, higher is the forward progress, as the saved energy can be used for execution. We already demonstrated low-power NVM design using R-FEFETs in Chapter-4 and in the following we propose R-FEFET based non-volatile flip-flops (RNVFFs). The R-FEFET based NVFFs and NVMs are then integrated together to build energy-efficient IPS.

5.3 NVFF with Auto-Backup: RNVFF-1

5.3.1 Circuit design

The ability of R-FEFETs to serve as a switch in the ‘V’ mode, enables its direct embedding in logic circuits. RNVFF-1 consists of a standard volatile flip-flop (FF) with transistors M4 and M6 (nMOS) replaced by R-FEFETs (Fig. 5.2(a)). Such natural embedding of the non-volatile element leads to an automatic backup without any external circuitry or signals. Note that, during normal operation, the embedding of R-FEFETs can affect the Clock-to-Q (CLK-Q) delay ($T_{CLK\text{-}Q}$), since P_{FE} switching time may not be as fast as standard FETs. To mitigate this issue, inverter INV2 is used to bypass INV1 and obtain output Q (Fig. 5.2(a)). The RNVFF-1 architecture is unlike previous approaches [80]–[85], [145]–[147] where the non-volatile element is in a distinct backup/restore (B/R) module. The operation of the RNVFF-1 is discussed next.

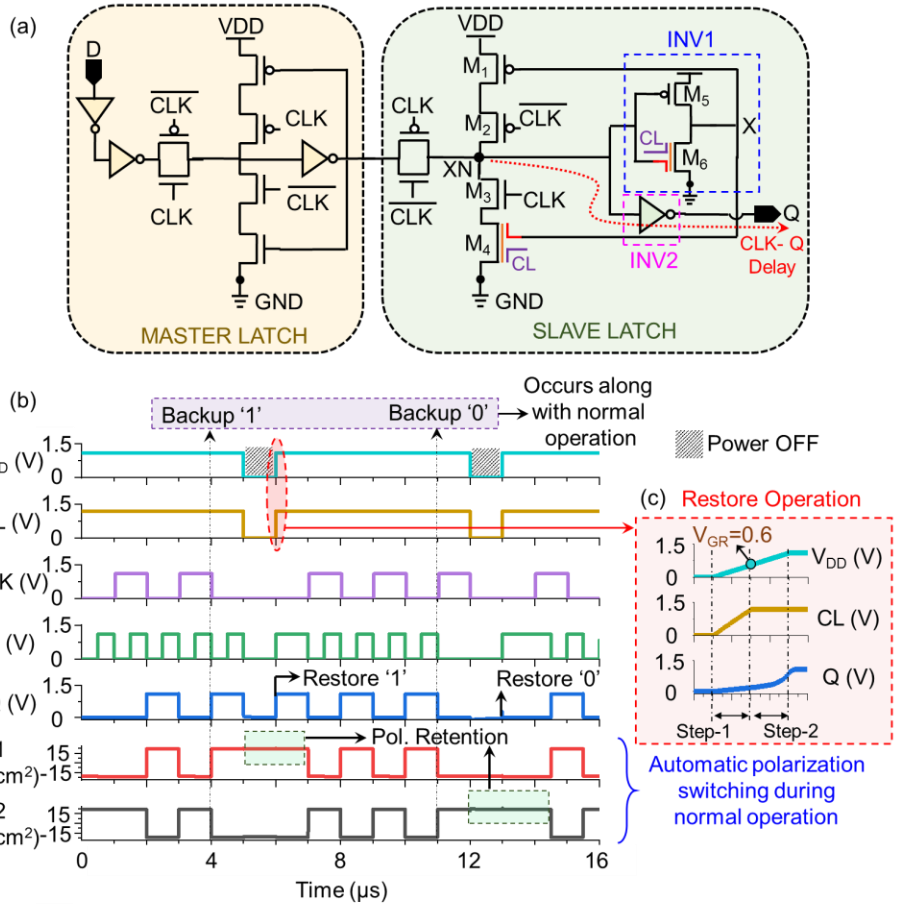


Fig. 5.2. (a) Schematic of RNVFF-1 (b) Transient waveforms showcasing auto-backup operation with polarization switching along with normal operation, (c) two step restore operation.

5.3.2 Circuit operation

(a) Normal operation: CL=V_{DD} is applied for R-FEFETs, in order to operate them as a switch, i.e., R-FEFET is OFF/ON at V_{GS} = 0/V_{DD} (similar to an nMOS FET; see Chapter-3). Therefore, the normal operation of RNVFF-1 is similar to the standard volatile FFs. Moreover, P_{FEG} follows the voltage at the storage nodes X and XN. As an example, when Q = V_{DD}, X and XN are at V_{DD} and 0, and P_{FEG} of M4 and M6 (R-FEFETs) store +P and -P respectively, corresponding to the state of the FF. This is also illustrated in the transient waveforms shown in Fig. 5.2(b). Now as expected, the normal operation energy (E_{OP}) increases compared to volatile FFs due to P_{FE} switching in the embedded R-FEFETs. However, interestingly, CLK-to-Q delay decreases in the proposed RNVFF-1. This is attributed the following two reasons: Firstly, the effective oxide thickness of FE+DE stack is higher in R-FEFETs due to the large T_{FE}, resulting in lower gate capacitance (when no PFE switching occurs). Secondly, since the P_{FE} switching is much slower than switching of logic gates, the total capacitance at X and XN (storage nodes) which drive the gates of the R-FEFETs is lower than standard FF, during the time window when the data is being propagated to Q. This unique attribute of the proposed RNVFF-1 results to lower CLK-to-Q delay compared to standard volatile FF.

(b) Auto backup operation: The most appealing feature of RNVFF-1 is the direct embedding of R-FEFETs in the FF architecture, which enables a completely automatic backup, without any external B/R module. The value of P_{FEG} in R-FEFETs correspond to the X and XN (storage nodes), as discussed above. During a power outage, V_{CS} reduces to 0V and the R-FEFETs automatically switch from the 'V' to 'NV' mode of operation. This leads to M4/M6 retaining P_{FEG} corresponding to the output Q of the FF [174].

(c) Restore operation: The restore operation, similar to backup, occurs without any additional signals or external circuitry. As discussed previously, although the storage node information is retained as P_{FEG} of the R-FEFETs in the 'NV' mode (i.e, when power supply of OFF), they cannot be distinguished due to the device optimization performed to eliminate the effect of gate leakage (see Chapter-3). Therefore, to sense the information stored in P_{FEG} for restoring the state of the FF, we utilize the built-in reconfigurability of R-FEFETs from 'NV' to 'V' mode as done for 3T-R NVM and perform the following 2-step operation (Fig.

11(c)): Step-1: CLK and V_{CS} are asserted immediately after the power supply is turned ON, bringing the R-FEFETs to the ‘V’ mode. Then, V_{DD} of RNVFF-2 is raised till V_{GR} (0.6V) to bias the nodes X and XN (gates of the R-FEFETs; M4 and M6) at an intermediate voltage within HW_V. This brings the R-FEFET storing ‘+1’/’0’ (P_{FEG}=+P/-P) to LRS/HRS, similar to the sensing operation performing for 3T-R. Step-2: After the resistance is established in R-FEFETs, V_{DD} is raised to 1.1V and the FF state is restored by virtue of the large distinguishability (>10^4) of the two R-FEFETs and cross-coupled action in the slave latch.

As mentioned above, RNVFF-1 achieves a completely automatic backup without the need of any external signals or circuitry. However, this comes with a penalty in EOP due to P_{FE} switching in every cycle. As discussed later, such an FF design can be extremely beneficial for systems/applications which require frequent check-pointing. However, to overcome the operational energy drawbacks for a general class of IPS, we also present another variant of NVFF with on-demand backup and restore operations, as discussed next.

5.4 NVFF with Gated Backup: RNVFF-2

5.4.1 Circuit design

RNVFF-2 features a gated backup, and thus involves the use of an external B/R module (Fig. 5.3(a)). In this design, only when the power supply turns OFF, the back-up is performed and therefore it overcomes the large E_{OP} cost observed with RNVFF-1. The B/R module of RNVFF-2 is designed with 3 standard FETs and 2 R-FEFETs. Transistors M1 and M2 connect/dis-connect the slave latch with the B/R module. The state of the flip-flop is stored in M4 and M5 (R-FEFETs). M3 is used to ensure that during normal operation, P_{FEG} of M4/M5 is always = -P, as explained in the following sub-section.

5.4.2 Circuit operation

(a) Normal operation: The B/R module is completely isolated from the FF by turning OFF M1 and M2. The CL is asserted for R-FEFETs (M4 and M5 in ‘V’ mode) and M3 is turned ON. Now, just before the normal operation (i.e. after restore operation), one of M4/M5 is ON (LRS; discussed later). Since M3 is ON during normal operation, the drains/gates of

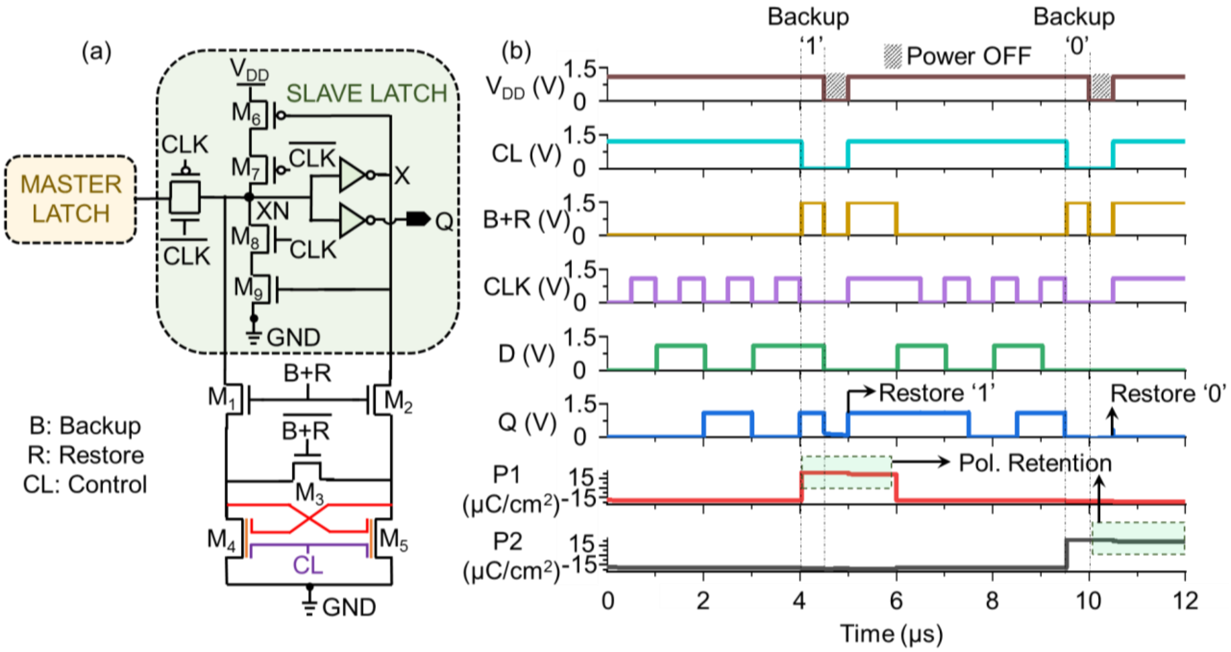


Fig. 5.3 (a) Schematic and (b) transient waveforms of RNVFF-2 illustrating the normal, backup and restore operations.

both M4 and M5 are driven to 0V resulting in P_{FEG}=-P (HRS). This is useful during backup as discussed later. Note, the B/R module has near-zero impact on the energy/performance of RNVFF-2 during normal operation because of the isolation of capacitances of the B/R module from X and XN (due to M1 and M2 being OFF). The transient waveforms of RNVFF-2 are illustrated in Fig. 5.3(b)).

(b) Backup operation: For backup, we turn ON M1 and M2 and connect the B/R module to the slave latch. M3 is turned OFF and CL is driven to 0V to configure the R-FEFETs in the 'NV' mode. At the onset of backup (i.e. just after normal operation), both M4 and M5 are in the HRS (discussed above). Now, corresponding to the storage nodes (X and XN) voltage, one of the R-FEFETs obtains V_{GS}=0 and the other obtains V_{GS}=V_{DD}. For example, when X=0 and XN=V_{DD} (Q=0), V_{GS} = 0/V_{DD} for M4/M5). This leads to M4 remaining in HRS (P_{FEG}=-P); while M5 switches to LRS (P_{FEG} =+P). The opposite occurs when the FF output, Q=V_{DD}. Therefore, the state of the flip-flop is as P_{FEG} of the R-FEFETs before a power shut down.

(c) Restore operation: When power supply is re-established, the B/R module is connected to slave latch by turning ON M1 and M2. After which, a 2-step restore scheme is employed similar to RNVFF-1 (Fig. 5.2(c)). The large distinguishability between the bit-stable states stored in R-FEFETs, along with cross-coupled action in slave latch, restores the output, Q.

5.5 Circuit-Level Analysis

In this section, we analyze the proposed RNVFFs and compare them to the NVFF in [146]. Since the design in [80] employs a different mechanism for B/R (independent of FF topology), we focus on comparison with [146] (which uses FEFETs in conjunction with cross-coupled inverter action similar to our RNVFFs). We perform our analysis for 45nm node and the results are tabulated in Table. 5.1.

(a) Clock-to-Q delay (T_{CLK-Q}) and operation energy (E_{OP}): Due to the minimal capacitance overheads at X and XN due to no external B/R module, T_{CLK-Q} of RNVFF-1 is 8% lower, compared to [146]. T_{CLK-Q} of RNVFF-2 is similar to NVFF due to both designs containing a B/R module. Compared to standard volatile FF (STD FF), T_{CLK-Q} of RNVFF-2 is similar but that of RNVFF-1 is lower by 5%. This is due to the fact that, nodes XN and Q transitions before the switching of P_{FEG} in M6. And before P_{FEG} switching, R-FEFETs exhibit lower capacitance than standard FETs. E_{OP} of RNVFF-1 includes P_{FEG} switching energy, which results in 10% higher E_{OP} with respect to [146]. RNVFF-2 mitigates the E_{OP} overheads of RNVFF-1 with no P_{FEG} switching during normal operation and achieving E_{OP} mildly lower than [146] (Table. 5.1).

(b) Check-pointing delay (T_{CKPT}) and energy (E_{CKPT}): A check-pointing operation involves one backup and one restore operation. RNVFF-1 exhibits 69% lower E_{CKPT} compared to NVFF. This is mainly attributed to the automatic backup enabled by the direct embedding of R-FEFETs, resulting in ~0 backup energy. Although, the restore energy is higher due to

Table. 5.1 Comparison of energy-delay metrics for flip-flop designs

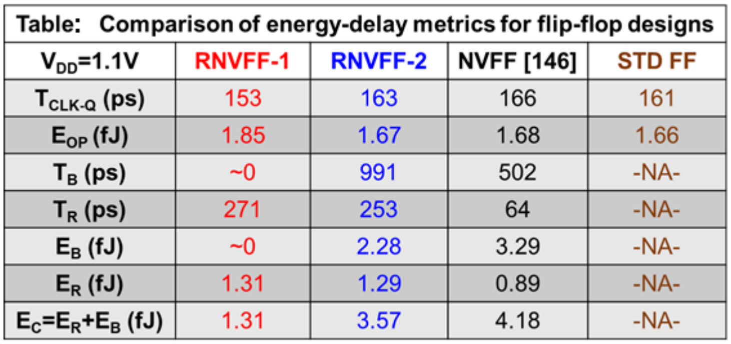


Table: Comparison of energy-delay metrics for flip-flop designs

V_{DD}=1.1V	RNVFF-1	RNVFF-2	NVFF [146]	STD FF
T_{CLK-Q} (ps)	153	163	166	161
E_{OP} (fJ)	1.85	1.67	1.68	1.66
T_B (ps)	~0	991	502	-NA-
T_R (ps)	271	253	64	-NA-
E_B (fJ)	~0	2.28	3.29	-NA-
E_R (fJ)	1.31	1.29	0.89	-NA-
E_C=E_R+E_B (fJ)	1.31	3.57	4.18	-NA-

the 2-step operation, E_{CKPT} (sum of backup and restore energy) is collectively lower than NVFF. RNVFF-2 exhibits 15% lower E_{CKPT}, due to only one P_{FEG} switching involved during backup, unlike two FE switching in NVFF. ~52% lower T_{CKPT} is observed for RNVFF-1 due to the complete automatic backup while for RNVFF-2, 2.2X increase is observed, mainly attributed to 2-step restore scheme. (Note, the comparison is made with NVFF [146]; FEFET w/o IML). However, higher restore time, T_R of RNVFF-2 might not be much of a concern for IPS, which operates at low frequencies (~ MHz) [188], [197], [198].

The R-FEFET based NVM presented in Chapter-4 and NVFFs presented in this chapter, exhibits unique characteristics which make them suitable for IPS which are energy-constrained platforms. The 3T-R exhibits lower write and read energy attributed to the unipolar write voltages and non-destructive read operation respectively, unlike FEFET (with IML) based NVMs [45]. The RNVFFs exhibit energy-efficient check-pointing operation for storing/re-storing the state of the FFs. However, this comes with certain tradeoffs, such as, high operational energy of RNVFF-1 and slower check-pointing operation in RNVFF-2 compared to FEFET based NVFFs. In the following section we design intermittently powered systems using the R-FEFET based NVMs and NVFFs and evaluate their benefits and trade-offs for a wide range of benchmarks.

5.6 Implementation in Intermittently Powered Systems

5.6.1 System-level simulation methodology

In order to evaluate the system-level energy benefits of the proposed R-FEFET-based NVFF and NVM designs in the context of an intermittently powered system, we collaborated with Prof. Vijay Raghunathan and Dr. Arnab Raha, and use the device/circuit/architecture/system co-design simulation framework presented in [198]. The simulation setup is shown in detail in Fig. 5.4. Our system-level experiments are based on the TI MSP430FR5739 microcontroller [39] (MCU)-based edge device that runs at 24 MHz [198] and uses a unified NVM of 32KB based on 3T-R. The R-FEFET NVFFs are used for implementing the volatile MCU core registers as shown in Fig. 5.1. The system is powered using an energy harvesting source that charges a supply capacitor, C_{SUPP} (we evaluate the system for two different values of C_{SUPP}, 10 nF and 1 nF). In this section, we re-

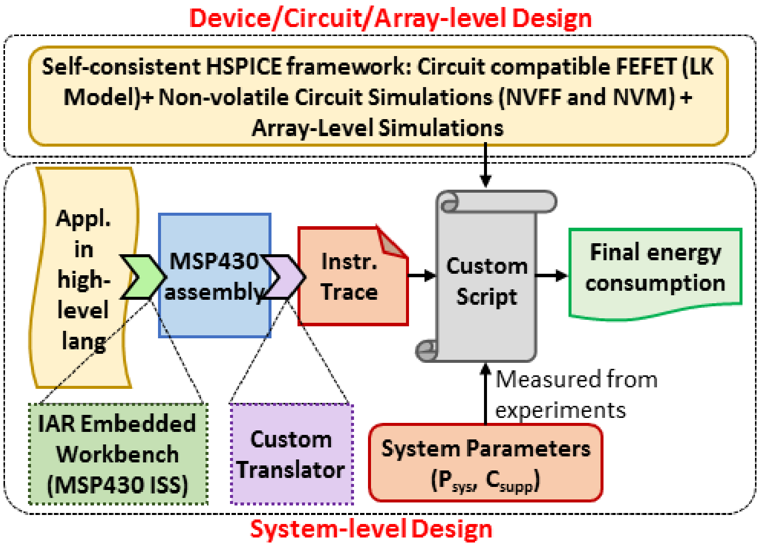

Device/Circuit/Array-level Design
Self-consistent HSPICE framework: Circuit compatible FEFET (LK Model)+ Non-volatile Circuit Simulations (NVFF and NVM) + Array-Level Simulations
Appl. in high-level lang
MSP430 assembly
Instr. Trace
Custom Script
Final energy consumption
Measured from experiments
IAR Embedded Workbench (MSP430 ISS)
Custom Translator
System Parameters (P_{sys}, C_{supp})
System-level Design

Fig. 5.4 Experimental Methodology for IPS evaluation

design the proposed R-FEFET-based NVFFs and NVM as well as the baseline FEFET-based NVFF [146] and 2T-FEFET NVM [31] at 45nm, as that also enables us to synthesize the MSP430 microcontroller core (soft IP core obtained from OpenMSP430 [199]) using OpenNangate 45nm technology logic library [200]. Note, we use a unified NVM for our simulation studies as compared to a conventional SRAM+NVM for this case. As described in detail in previous works such as [188], [197], [198], the checkpointing operation in unified NVM is performed in situ that avoids any explicit transfer of program state data (data, bss, stack) in the NVM which forms a major portion of the overall checkpointing state. However, this still requires us to checkpoint (backup and restore) the MCU processor and peripheral states such as the program counter (PC), Status Register (SR), Stack Pointer (SP), General Purpose Registers (GPR), Special Function Registers (SFR), and Peripheral Registers (PR) as shown in Fig. 5.1 using the RNVFF-based registers. Therefore, the unified NVM requires much smaller checkpointing energy overhead compared to the conventional SRAM+NVM configuration.

Without any loss of generality, we calculate the energy savings with respect to either the total NVFF-based register (Reg) energy consumption, the total memory energy (Mem+Reg) consumption, or the total system energy consumption (Proc+Mem+Reg). Note that the Reg energy includes both the read and writes to the NVFFs as well as the backing up and restore energies due to checkpointing. In this work, we consider the total memory consumption (represented by tot mem) to be the sum of register energy consumption (represented by Reg) and the NVM energy consumption (represented by Mem). The total system energy consumption (represented by tot sys) is assumed to be the sum of the total memory energy (tot mem = Reg+Mem) consumption, and the MSP430 microcontroller execution energy (represented by Proc) [thus, tot sys = Reg+Mem+Proc]. The average power consumption of the MSP430 microcontroller is calculated by synthesizing the OpenMSP430 microcontroller RTL [200] using Synopsys Design Compiler and taking the synthesized RTL through Synopsys Power Compiler and then the total execution energy was calculated by multiplying the average power with the total number of execution cycles for the program. Note that for all the cases, the energy consumption is normalized with respect to the system that consists of the existing FEFET-NVFF presented in [146] and the 2T-FEFET NVM [31]. In this work, most energy improvement and consumption numbers are represented as a range as they vary due to the change in supply capacitance value from 10 nF (lower energy savings due to lower number of checkpoints) to 1 nF (higher energy savings due to larger number of checkpoints). Note that the amount of energy improvement is directly proportional to the forward progress of the IPS, an alternative metric that is used to show the improvement in the lifetime/energy consumption in an IPS.

5.6.2 System-level simulation results

In this sub-section, we first quantitatively compare just the register-level energy consumption of RNVFF-1, RNVFF-2, and FEFET-NVFF [146], based on application-level energy consumption while running various real and synthetic benchmarks in 45 nm technology. Note that for this work, we omitted comparing against the standard volatile FF (STD) as it performed significantly worse under all conditions in [174]. We show the relative impact of improving the NVFFs and NVM separately on the total memory as well as the total system energy consumption. We show these energy savings for both set of real and synthetic benchmarks.

(a) Register-level energy results: The register-level energy consumption is shown in Figs. 5.5 -5.6 where we see that both RNVFF-1 and RNVFF-2 outperform the baseline FEFET-NVFF [146]. RNVFF-1 show significantly higher energy savings compared to RNVFF-2. With smaller values of C_{SUPP}, the register-level energy savings increases for RNVFF-1 but decreases for RNVFF-2 since the checkpoint energy becomes the most prominent register

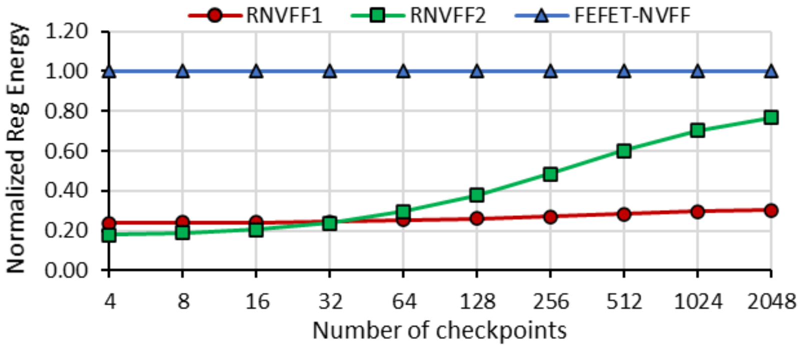


Fig. 5.5 Normalized core register energy consumption of the different NVFF designs for synthetic benchmarks generated by varying the number of checkpoints.

energy component at larger number of checkpoints. Note that for all these results, the register energy consumption is normalized with respect to baseline [146]. The details area as follows.

(i) Synthetic Benchmarks: Since the energy benefits of using RNVFF-1, RNVFF-2 over NVFF [146] depend significantly on the total number of system-checkpoints while executing a specific application, we constructed a synthetic application benchmark that has 25% of all instructions to be register reads, 25% to be register writes, and the

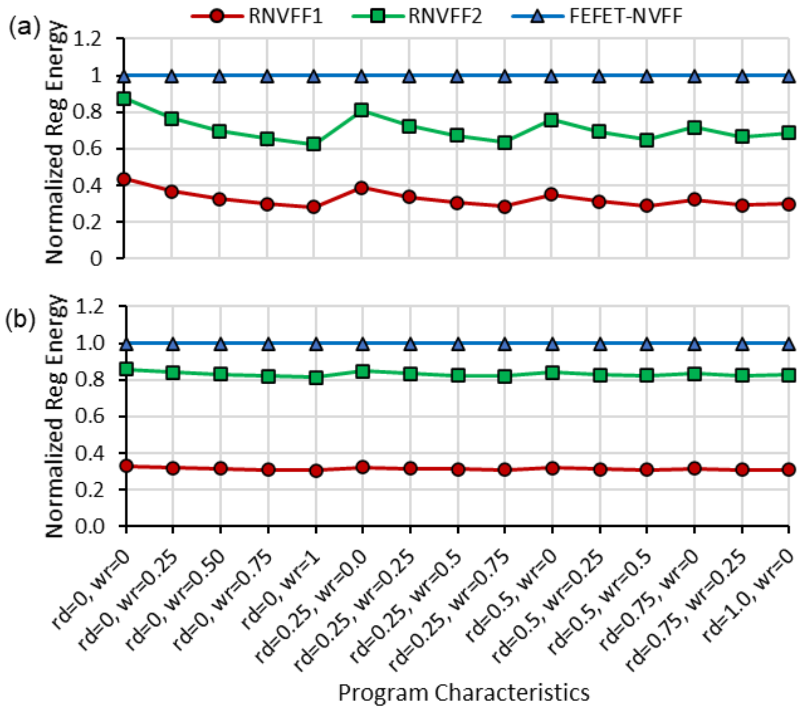


Fig. 5.6. Normalized core register energy consumption of different NVFF designs for synthetic benchmarks using (a) C_{SUPP} = 10 nF and (b) C_{SUPP} = 1 nF.

remaining 50% to be memory bound instructions. In addition, we varied the total number of checkpoints while keeping the checkpoint size and the total number of instructions constant (100K). The results are shown in Fig. 5.5. Note that in Fig. 5.5, the energy numbers are normalized with respect to [146]. In this case, we observe that as the number of checkpoints per application execution increases, the energy savings due to RNVFF-1, RNVFF-2 increase rapidly compared to [146]. This is because as the number of checkpoints increases, RNVFF-1 performs exceedingly well compared to either RNVFF-2 or NVFF due to a much lower checkpoint/restore energy as seen from the circuit results presented in earlier. It is important to note that RNVFF-2 performs better than [146] irrespective of the number of checkpoints as it has both lesser checkpointing/restore as well as normal read-write energy consumption. Further, the energy savings from RNVFF-2 remains almost invariant due to a constant difference in the read/write energies and a relatively small benefit in checkpoint/restore energies as shown in Table. 5.1. On the other hand, RNVFF-1 outperforms [146] only at higher checkpoint sizes (or #checkpoints) due to a significant advantage only in checkpoint/restore energy consumption (but energy overheads for normal operation - see Table. 5.1) where the energy savings become larger with the number/size of checkpoint. Fig. 5.6 shows a second set of synthetic benchmarks where we show the energy benefits arising from the variation in the nature of program characteristics, i.e., total number of reads and writes during program execution while executing a specific application. We constructed a set of synthetic benchmarks where we vary the fraction of total memory

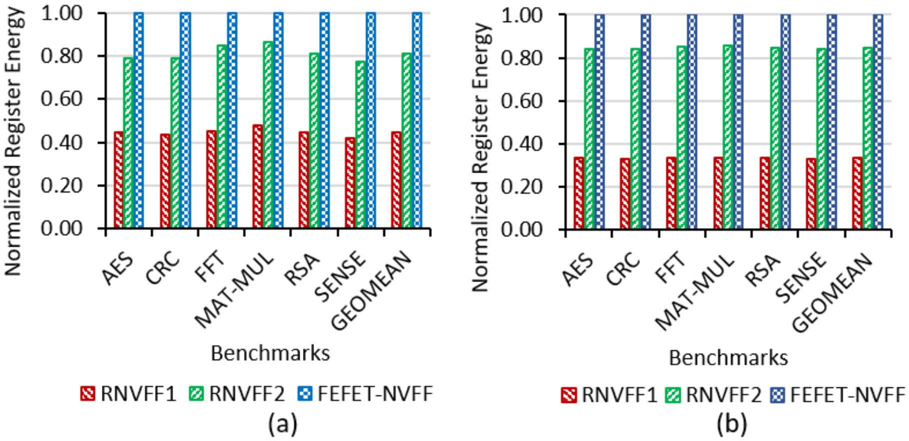


Fig. 5.7 Normalized core register energy consumption of different NVFF designs for real benchmarks using (a) C_{SUPP} = 10 nF and (b) C_{SUPP} = 1 nF.

read and write instructions subject to C_{SUPP} = 10 nF and 1 nF with a constant checkpoint size of 896 B and total number of instructions (100K). Here, the expression {r:0.25, w:0.25} represents that 25% of the total instructions are non-volatile memory reads, 25% are non-volatile memory writes, and the rest are normal computational operations. The latter 50% computational operations have a fixed 25% each of register reads and writes and the rest 50% of compute instructions does not involve any registers. Irrespective of the program characteristics, we see that RNVFF-1 and RNVFF-2 outperforms [146] for various benchmark types. Note that RNVFF-1 performs better than RNVFF-2 due to the low cost of checkpointing and restore. Fig. 5.6 also shows larger energy benefits of RNVFF-1 compared to RNVFF-2 and NVFF with higher memory read instructions as well as smaller C_{SUPP} that results in higher number of checkpoints. This is in accordance to the different circuit-level energy values show in Table. 5.1. As seen earlier, we observe that RNVFF-2 always perform better than [146] irrespective of the program characteristics and C_{SUPP} value. However, RNVFF-1 perform worse than the FEFET-NVFF at larger C_{SUPP} values when checkpointing and restoring energies are small compared to the access energies. However, at moderate to smaller C_{SUPP} values, RNVFF-1 will have significant lower energy consumption compared to either RNVFF-2 or FEFET-NVFF [146].

(ii) Real benchmarks: For real application benchmarks, Fig. 5.7 shows the core register energy savings (normalized to the FEFET-NVFF) achieved by the RNVFF-1 and RNVFF-2 designs over the existing NVFF [146] design using real benchmarks for two different C_{SUPP} values. Fig. 5.7(a) demonstrates register-level energy savings for RNVFF-1 and RNVFF-2 of 55% and 18.3% on average, compared to [146] for C_{SUPP} = 10 nF. Fig. 5.7(b) shows that RNVFF-1 and RNVFF-2 result in 67% and 15.3% on average register-level energy savings, respectively, compared to [146] for C_{SUPP} = 1 nF (the savings are higher in this case because the smaller capacitor leads to higher system checkpoints). As mentioned previously, RNVFF-2 will always perform better than [146] irrespective of the benchmark or C_{SUPP} value as it has both lower checkpoint/restore and operation energies compared to [146].

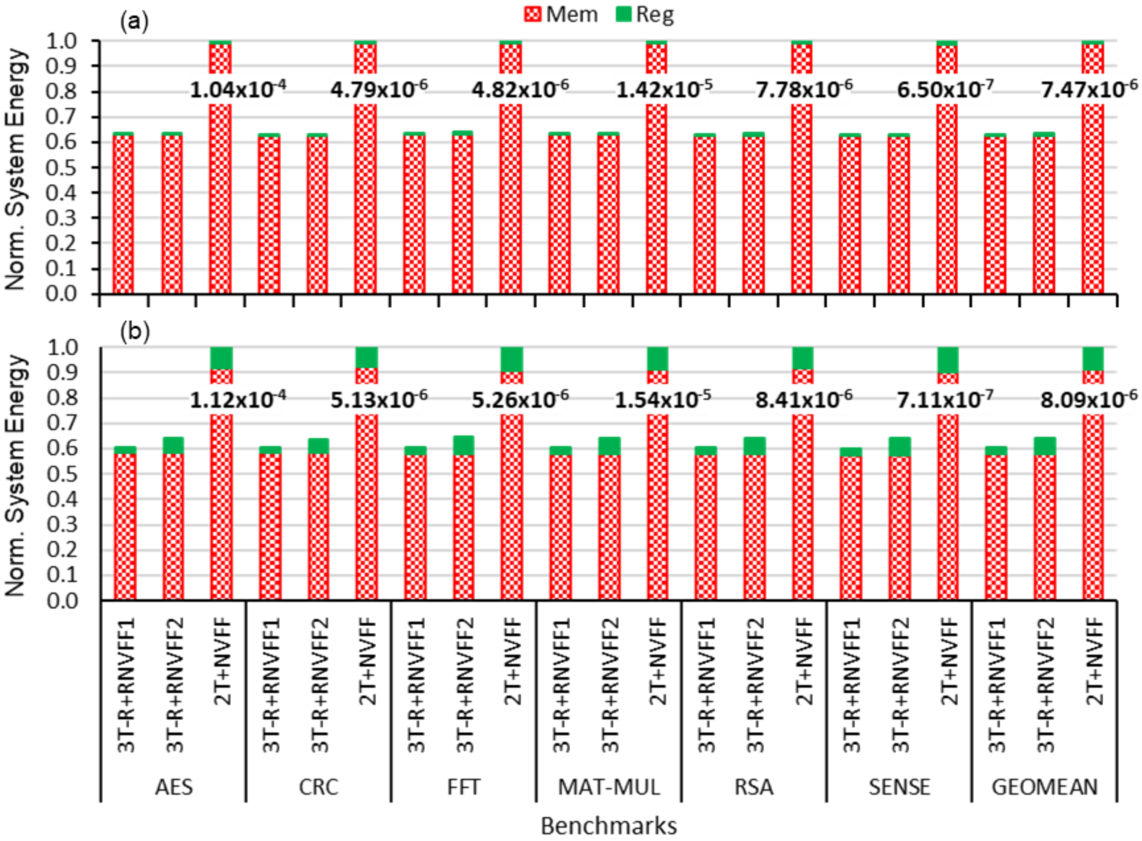


Fig. 5.8 Normalized total memory (*tot mem*) energy consumption of different NVFF+NVM designs for real benchmarks using (a) C_{SUPP} = 10 nF and (b) C_{SUPP} = 1 nF.

(b) Memory and system-level energy results: The system-level energy results demonstrate the impact of integrating the proposed 3T-R NVM along with the R-FEFET-based NVFFs on system-level energy consumption.

(i) Real benchmarks: For real application benchmarks, we present the total memory energy in Fig. 5.8 and the total system energy in Fig. 5.9 corresponding to C_{SUPP} values of 10 nF and 1 nF, respectively. Fig. 5.10 summarizes the energy savings achieved by the RNVFF-1 and RNVFF-2 +3T-R NVM designs (represented as 3T-R+RNVFF) over the existing FEFET-based NVFF [146] and NVM [31] (represented as 2T+NVFF) design using real benchmarks for these two different C_{SUPP} values. It demonstrates the maximum energy savings achieved by either combination (NVFF+NVM) of the proposed RNVFF-1+3T-R or RNVFF-2+3T-R over the baseline FEFET-NVFF +2T-FEFET NVM combination.

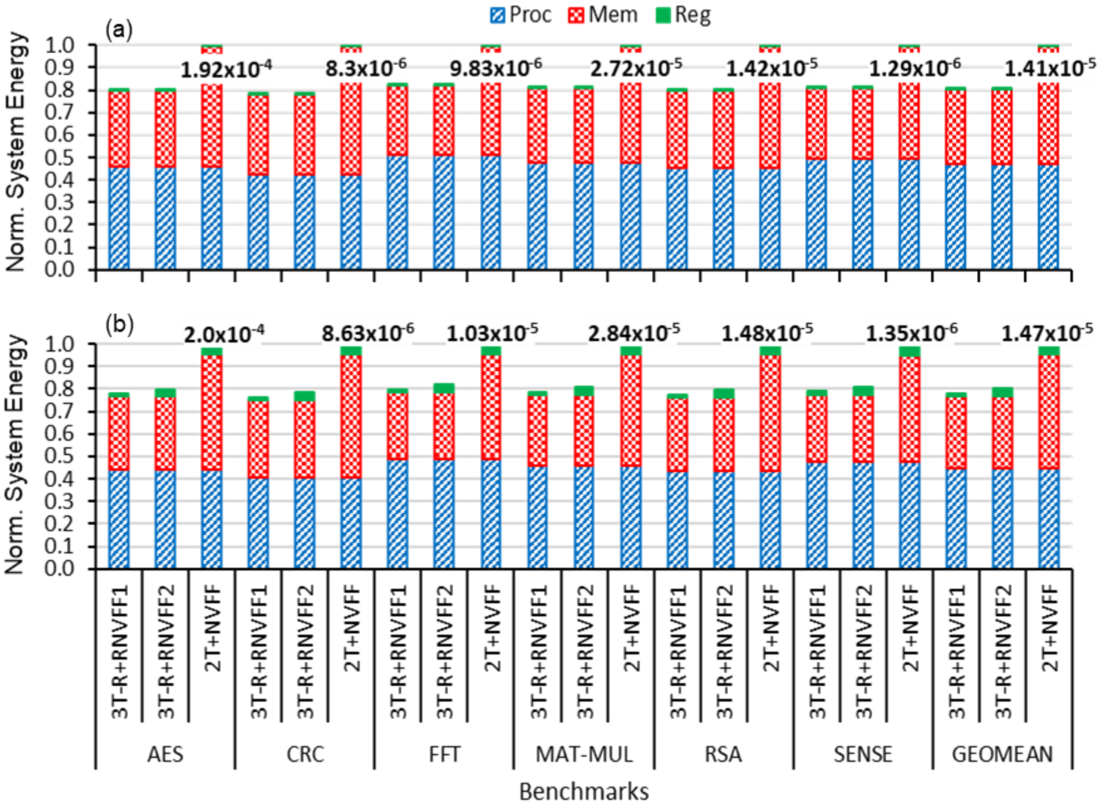


Fig. 5.9 Normalized system energy (*tot sys*) consumption of different NVFF+NVM designs for real benchmarks using (a) C_{SUPP} = 10 nF and (b) C_{SUPP} = 1 nF.

As we can see, our proposed RNVFF-1/RNVFF-2+ 3TR memory combination results in memory-level and system-level energy savings around 37% and 20% on average, compared to [146] for C_{SUPP} = 10 nF. For C_{SUPP} = 1 nF, these energy savings

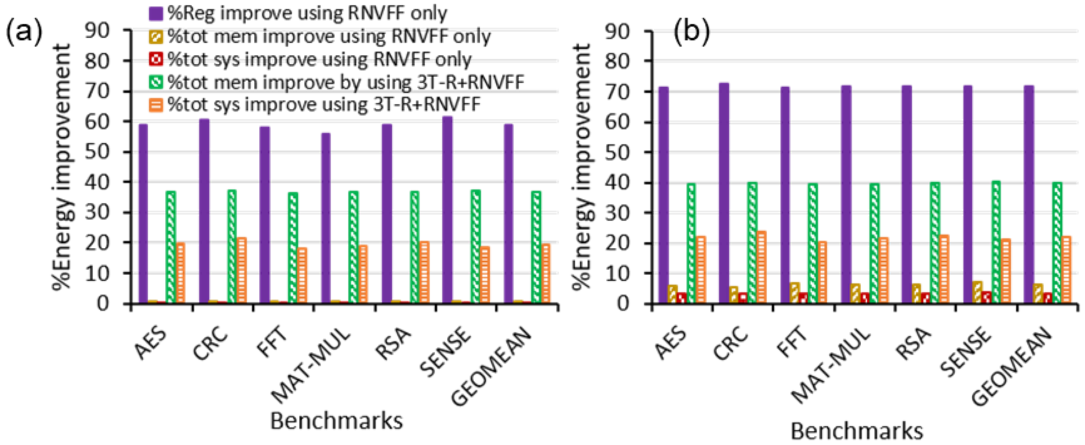


Fig. 5.10 Register, memory and system-level energy improvements of different NVFF+NVM designs for real benchmarks using (a) C_{SUPP} = 10 nF and (b) C_{SUPP} = 1 nF.

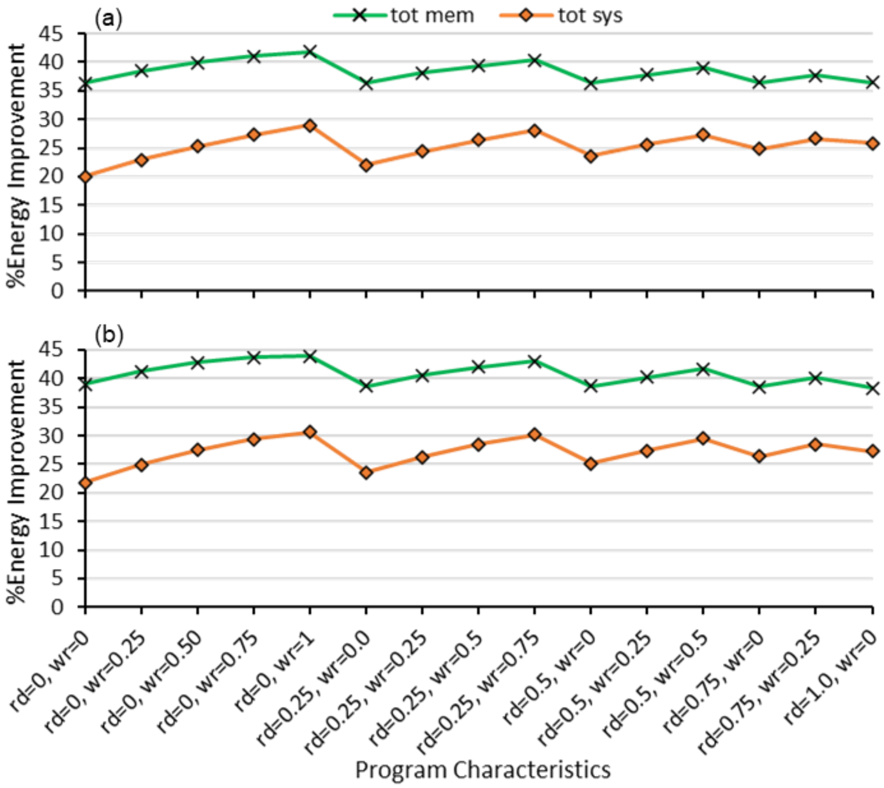


Fig. 5.11 Memory and system-level energy improvement of the proposed NVFF+NVM design for synthetic benchmarks using (a) C_{SUPP} = 10 nF and (b) C_{SUPP} = 1 nF.

increase to be 40% and 22%, respectively due to higher number of checkpoints. As Figs. 5.8 and 5.9 show, the majority of the energy benefits arise from integrating a better NVM (3T-R) as compared to the RNVFFs. Compared to the case where we just improved the NVFFs, the proposed RNVFF-1/RNVFF-2+3TR combination provides almost 63X and 7X more system-level energy savings for C_{SUPP} = 10 nF and 1 nF, respectively. These are significant improvements over the 2T-FEFET based NVM design [31]. As stated before, the system-level energy consumption of IPS is indirectly proportional to its relative forward progress that is metric to measure IPS energy-efficiency [31]. In this case, the system-level energy savings stemming due to the use of 3T-R NVM results in a forward progress of 1.25X-1.29X for C_{SUPP} = 10 nF-1nF.

(ii) Synthetic benchmarks: In this sub-section, we demonstrate that the benefits of improving the NVM are significant irrespective of the benchmark characteristics or the supply energy (that determines the number of checkpoints). This is mainly due to the significant improvements in the proposed NVM (3T-R compared to 2T-FEFET NVM

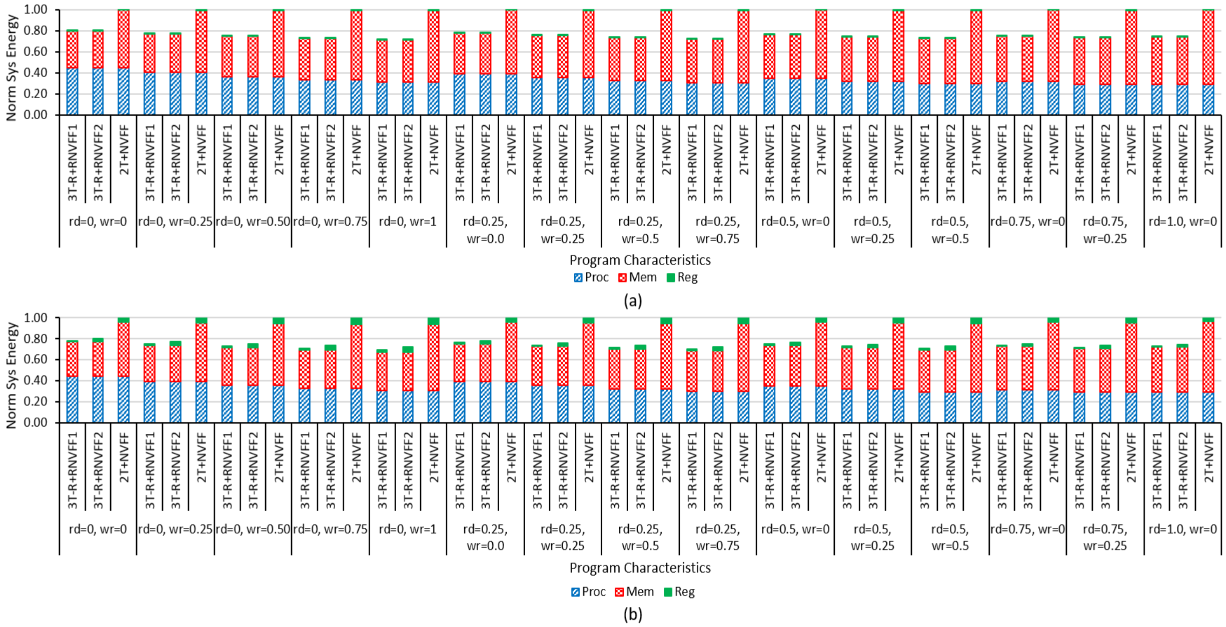


Fig. 5.12 Normalized system energy consumption of different NVFF+NVM designs for synthetic benchmarks using (a) C_{SUPP} = 10 nF and (b) C_{SUPP} = 1 nF.

[31]) energy consumption. Fig. 5.11 shows that across different program characteristics, we achieve memory and system-level energy savings in the range of 37-42% and 21-29%, respectively for C_{SUPP} = 10 nF. This increases to a range of 39-44% and 24-31%, respectively, for C_{SUPP} = 1 nF. The split of the different energy contributions in Fig. 5.12 shows that the proposed 3T-R NVM is the primary reason for this energy improvement. Note that RNVFFs also demonstrate decent memory/system-level energy savings at a smaller C_{SUPP} that result in a larger number of system checkpoints as evident from Fig. 5.12. Even in the case where we vary the number of checkpoints, Fig. 5.13 shows that we obtain significant system-level energy savings about 28% on average. Note that for Fig. 5.13, where we vary the number of checkpoints, i) the memory energy consumption almost remain the same the NVM checkpoint is performed in situ due to the unified NVM architecture that results in a very small size of checkpoint data (only the very small MCU state needs to be checkpointed as mentioned before); ii) the energy savings vary and are higher at larger checkpoints as there the energy consumption due to NVFF (Reg) checkpoints start to be more prominent than before. To conclude, the experimental results obtained from both real and synthetic benchmarks demonstrate a significant boost in memory and system-level energy savings due to the use of 3T-R.

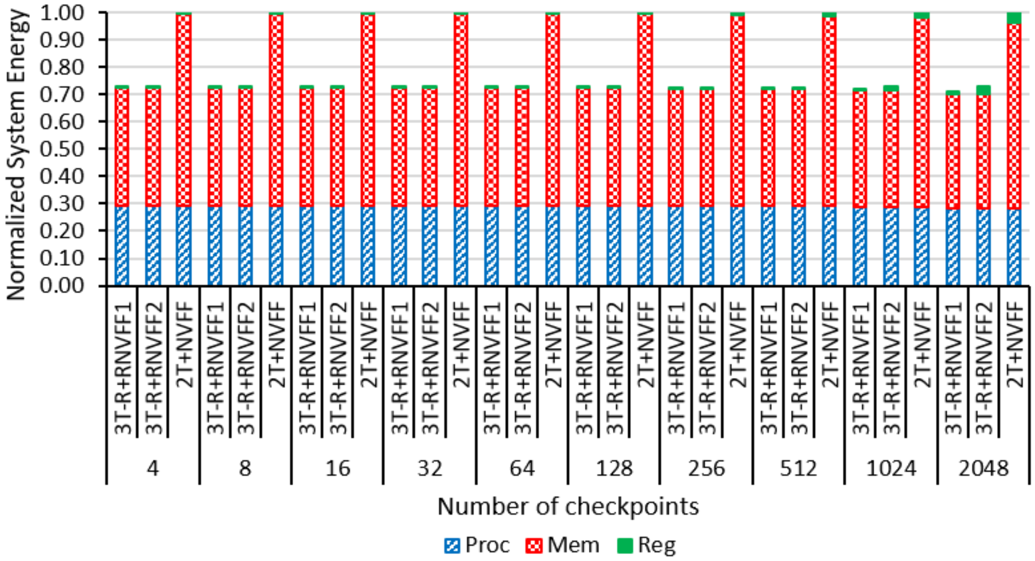


Fig. 5.13 Normalized system energy consumption of different NVFF+NVM designs for synthetic benchmarks generated by varying the number of checkpoints.

5.7 Summary

We proposed two variants of non-volatile flip-flop designs utilizing the dynamic tunability offered by R-FEFETs. We presented (a) RNVFF-1, which exhibited a completely automatic backup without the need of any external circuity or signals and (b) RNVFF-2 which involved an on-demand based backup operation enabled by a backup/restore module. When compared to existing FEFET based NVFF, we showed that RNVFF-1 exhibited high backup energy efficiency, resulting in 69% lower total checkpointing energy, RNVFF-2 had 15% lower checkpointing energy due to its low-power, compact backup/restore module. We also discussed how RNVFF-2 overcomes the high operational energy costs of RNVFF-1. In the end, using the proposed R-FEFET based NVMs in Chapter-4 and NVFFs in this chapter, we explored the design of energy efficient intermittently powered system in the context of TI MSP430 microcontroller. We performed our analysis, considering real and synthetic benchmarks, considering different supply capacitances powering the system. We showed the different components of energy being spent on processor, memory and register operations. Our simulations demonstrated a total system-level energy savings in the range of 37-40% and 20-22% for synthetic and real benchmarks, respectively.

6. R-FEFET/FEFET BASED ARCHITECTURES FOR BOOLEAN/ARITHMETIC COMPUTING-IN-MEMORY

6.1 Introduction

Memory speed has not kept up with processor speed over the last few decades leading to the so-called Memory Wall problem [4], [6]. It is estimated that the gap between the improvement in processor and memory speeds is increasing by more than 50% every year [5], [7]. Furthermore, in the era of big data, cloud computing and artificial intelligence, data-intensive applications have come to the forefront. This has led to restricted processor-memory bandwidth, resulting in overall performance/energy degradation in modern day systems [201].

One solution to this problem is to perform Computing-in-Memory (CiM), which can mitigate the aforementioned issues [201] (Fig. 6.1). Despite CiM being a decades old concept, interest in it has been rekindled in recent years. This is largely driven by advances in non-volatile memory (NVM) [23] and 3D monolithic technologies [149]–[151], [202], but also to the meet the demands of extensive data processing for current/next generation computing.

Various CiM architectures have been proposed to enable array level compute operations within memories. Several works on SRAM based CiM have shown bit-wise Boolean functions [59], [60], [64], [66], [203] and dot-product computations [58], [62], [68]–[70], [73], [159], [161], [204]–[206]. However, the fundamental challenge with these designs is the robustness of data storage during compute operations due to multi-word-line assertions [59], [66]. Monolithic 3D integration of SRAMs has been proposed for enhancing the stability of in-memory computation

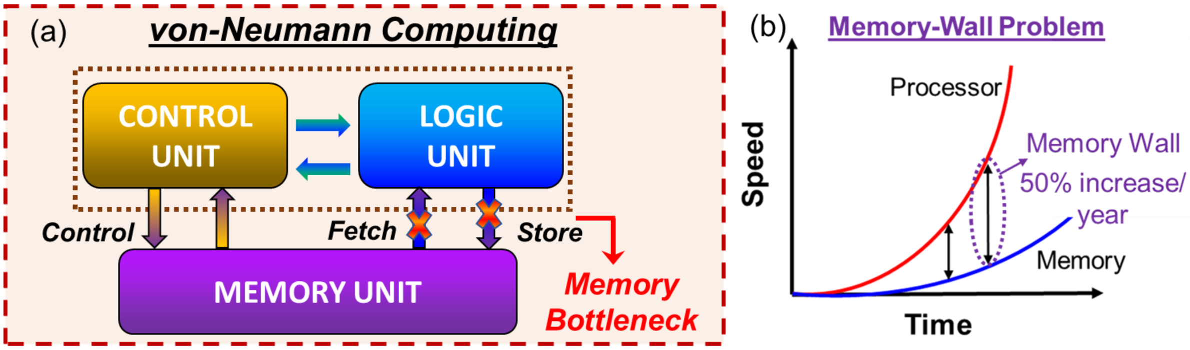


Fig. 6.1 (a) Memory bottleneck in tradition von-Neumann computing systems (b) Memory wall problem

[202]. However, such an approach requires multiple references and a two-phase compute operation scheme, leading to performance/energy penalty.

Efforts have been made to design CiM architectures using NVM technologies such as RRAMs [60], [203], STT-MRAMs [59] and FEFETs [64] to overcome the inefficiencies and robustness issues associated with SRAM based CiM. Memristor-Aided Logic (MAGIC) based on RRAMs enables computation in cross-bar arrays [203]. However, such an architecture requires large number of intermediate access cycles leading to performance and energy inefficiencies [64], [203]. In contrast, the approach based on assertion of multiple word-lines, such as in STT-CiM [59], overcomes the energy/delay bottleneck of MAGIC by performing a range of compute operations using a single memory access. However, a major challenge in STT-CiM is the reliable sensing due to low distinguishability of its bi-stable states [23], [59], which is further aggravated during compute due to the assertion of multiple wordlines. Moreover, the requirement of two current-based references for computation, leads to significant design overheads in STT-CiM.

Recently, FEFET-CiM [64] was proposed to overcome the drawbacks of MAGIC and STT-CiM. The large distinguishability between the bi-stable states achieved by storing the bit information as polarization, enables improved robustness while asserting single/multiple word-lines. Moreover, due to electric field driven write in the FEFET, along with the voltage based read sensing, FEFET-CiM outperforms STT-CiM design, albeit at the cost of area. However, FEFET based memory requires bi-polar voltages in order to encode the bit with a single-phase operation, resulting in energy-inefficiency. Although, recent studies have shown the possibility of using positive voltages with a two-phase write [192], such an approach leads to ~2X performance penalty. Moreover, similar to the other emerging NVM-CiM designs [59], [64], FEFET-CiM also requires a current based compute scheme and multiple references for sensing and logic functions, leading to design/energy overheads.

To address the issues with existing FEFET based CiM designs, in the first part of the chapter, we propose R-FEFET based CiM architecture utilizing its unique characteristics of dynamic reconfigurability in the modes of operation. We look at general purpose computing to evaluate the proposed CiM technique using R-FEFET, and compare it with FEFET based design for a wide range of application workloads. In the second part of the chapter, we explore the opportunities of designing a CiM engine specifically targeted for intermittently powered systems (IPS) performing edge computing. Since these platforms are extremely energy-constrained, they can benefit largely from the integration of richer functions into the memory arrays which is enabled by the novel IPS-CiM technique proposed in this chapter. IPS-CiM can be implemented using a wide range of memory technologies, and in this work, we focus on using ferroelectric transistors due to their benefits of low power electric field driven memory operations.

6.2 R-FEFET based Energy-Efficient CiM Engine using Differential NVM

Utilizing the intriguing features of R-FEFET, we propose a novel non-volatile memory, 4T-R, based on cross-coupled R-FEFETs featuring (a) differential read, (b) positive write voltages for both write-0 and write-1 and (c) low power in-memory computing. To the best of our knowledge, 4T-R is the first differential memory based on cross-coupled ferroelectric transistors using unipolar voltages, and therefore, is able to embrace the best features of SRAM and NVM, synergistically coupling them and mitigating the existing issues of CiM.

6.2.1 Differential 4T-R NVM

In this sub-section, we present our 4T differential NVM cell based on R-FEFETs (4T-R NVM). Two access transistors (standard FETs) are used to drive the cross coupled R-FEFETs to design the 4T cell as shown in Fig. 6.2(a). (Note, R-FEFETs used in this section is based on 45nm node R-FEFET$_{\mathrm{SYM}}$ device architecture). The inherent reconfigurability between 'V' and 'NV' modes (and the resultant all-positive write voltages) enable the possibility of cross-coupling of the

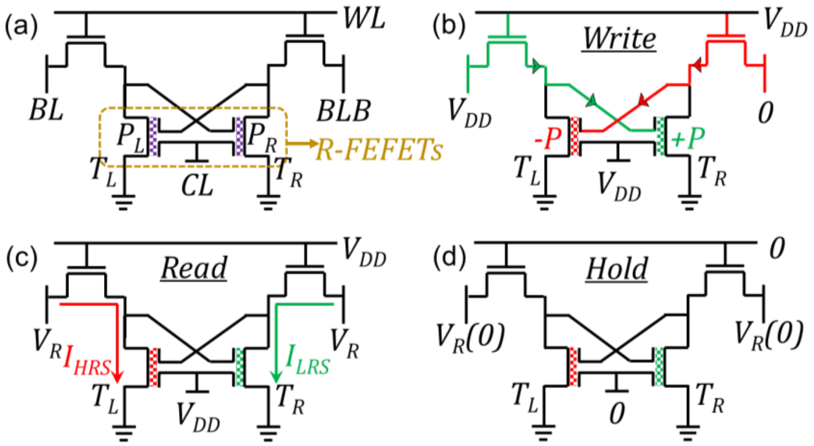
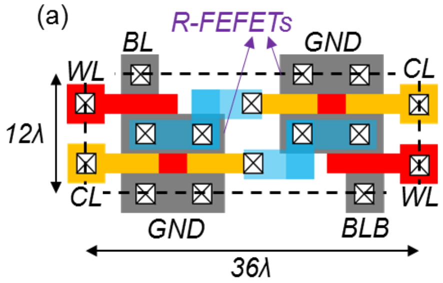


Fig. 6.2. (a) Proposed 4T-R memory cell schematic with cross-coupled R-FEFETs. Schematic with biasing for (b) Write, (c) Read and (d) Hold operations of the proposed 4T-R memory.



(b) **Operating bias conditions for 4T-R**

	WRITE	***READ***	***HOLD***	***COMPUTE***
Mode	Volatile	Volatile	Non-Volatile	Volatile
WWL	V_{DD}=1.1V	V_{DD}	0V	V_{DD}
BL	V_{DD} (0V)	V_R=0.7V	V_R /0V	V_R
BLB	0V (V_{DD})	V_R	V_R /0V	V_R
CL	V_{DD}	V_{DD}	0V	V_{DD}

Fig. 6.3 (a) Layout and (b) Bias conditions for the proposed 4T-R memory

R-FEFETs. This may not be feasible with standard FEFETs which need negative gate-to-source voltages for +P→-P switching [31], [133]. Our proposed bit cell design requires just one additional transistor to achieve differential operation compared to standard FEFET based 3T memory [64]. Although, previous attempts have been made to achieve differential storage using standard FEFETs by duplicating the 3T cells [168], i.e., by using two de-coupled single ended bit cells, they lead to ~2X area overhead with respect to 3T design. On the other hand, the proposed 4T-R NVM (layout in Fig. 6.3(a)) yields only ~1.6X penalty in footprint compared to minimum sized 3T. Note, the mirroring of 3T cells to achieve differential storage with FEFETs still requires bi-polar voltages for their operation [168]. However, the proposed 4T-R cell design uses uni-polar voltages, leading to significant improvement in energy savings during memory operations (as discussed extensively earlier). The array organization consists of the bit-lines (BL and BLB) shared amongst the cells in the same column. The word-lines (WL) and control-lines (CL, which drives the control terminal of both R-FEFETs, see Fig. 6.2(a)) are shared amongst the cells in the same row. As discussed before, CL determines the mode of operation of the R-FEFETs (‘NV’: V_{CL}=0V; ‘V’: V_{CL}=V_{DD}). The FE_G polarizations stored as PR in the right-side R-FEFET (T_R), is driven by BL and that as PL, in the left-side R-FEFET (T_L), is driven by BLB (Fig. 6.2(a)), as we discuss later. The cross-coupled R-FEFETs always store opposite P_{FEG} (similar to SRAMs storing complementary voltages), enabling fast and low power differential access. We discuss these aspects in the 4T-R memory operations, next (bias conditions in Fig. 6.3(b)).

6.2.2 Memory operations

(a) Write operation: To perform write, we use the ‘V’ mode of R-FEFETs. In the ‘V’ mode, V_{GS} of the R-FEFET =0V (V_{DD}) corresponds to P_{FEG}= -P (+P). We use this feature to

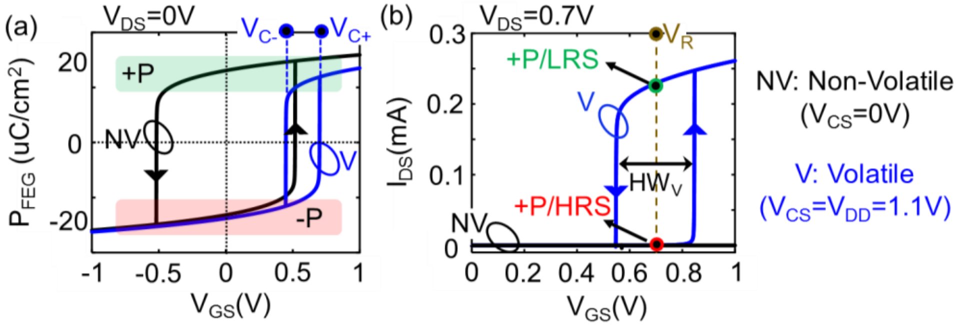


Fig. 6.4. (a) Polarization of FE_G (P_{FEG}) and (b) Drain Current (I_{DS}) vs Gate voltage (V_{GS}) for Non-volatile and Volatile modes of operation.

simultaneously write the true and complementary values in our differential memory. To write '0' ('1'), BL and BLB are driven to 0V (V_{DD}) and V_{DD} (0V) respectively (V_{DD}=1.1V in this work). (Note that, V_{DD} can be scaled with material-device co-design, for example by tuning T_{FE} [31] or performing work-function engineering). After this, the corresponding WL and CL of the accessed row are asserted. Now, during the write '0' ('1') operation, P_{FEG} stored in the R-FEFETs are P_R= -P (+P) and P_L= +P (-P) (Fig. 6.2(b)). Therefore, P_R stores the bit information and P_L stores the complement of bit information. In order to ensure full voltage swing at the gate of the R-FEFETs we boost WL voltage, a technique used in common [31]. Since, the memory access is performed for a word (32/64 bits), which has a shared WL, the dominant factor of energy consumption comes from the driving of bit-lines (32/64). Therefore, the increase in energy due to word-line boosting is negligible when compared to the total energy. The unaccessed cells in the same column are isolated by driving the WLs and CLs to 0V (R-FEFETs in 'NV' mode). The BL and BLB of the unaccessed cells in the same row are driven to V_R, so that we operate the R-FEFETs within the volatile hysteresis window (HW_V), in order to avoid accidental switching of polarization (Fig. 6.4(b)). After the write operation, WL and CL are de-asserted, which brings all R-FEFETs to the 'NV' mode and by virtue of bi-stability at V_{GS}=0V (Fig. 6.4(a)), the bit information is stored as P_{FEG} (P_R and P_L), in a non-volatile fashion.

(b) Read operation: As mentioned before, the bit information is encoded as P_{FEG} of the R-FEFETs (P_R and P_L). To perform read (as well as CiM operations), we choose a voltage-based sensing scheme for better energy efficiency than current sensing (as claimed in [64]), using RSA [66]. To read, we first pre-charge the bit-lines to V_R (=0.7V). Next, the R-

FEFETs are configured to the ‘V’ mode by asserting CL along with WL (boosted). This results in V_R to appear as V_{GS} of the R-FEFETs. Now, depending on the P_{FEG} stored, the resistance state of the R-FEFET is re-established to either LRS (P_{FEG}=+P) or HRS (P_{FEG}=-P) (Fig. 6.4(b)). Subsequently, we use the high magnitude difference in the resistance states (>10^4) for sensing the bit information as described next. (Note that, the selection of V_R is critical for the stability of read operations which is discussed later.)

In order to read the bit/word information, differential sensing is used. BL and BLB are initially precharged to V_{READ}=0.7V. Let us consider the case when the bit information stored is ‘1’ i.e., P_R= +P (T_R: LRS - Fig. 6.2(c)) and P_L= -P (T_L: HRS). Now, during read, BLB starts discharging (T_R in LRS), while BL remains at the pre-charged value (T_L in HRS) as shown in Fig. 6.2(c). After BLB discharges by ~50mV, the sense amplifier (SA) amplifies the difference and brings the nodes OUT1 to 0V and OUT2 to 0.7V (transients in Fig. 6.5(a)). Similarly, when the bit is ‘0’ (P_R=-P and P_L=+P), the read operation brings OUT1 to 0.7V and OUT2 to 0V.

It is noteworthy that, during the ‘V’ mode of operation, the R-FEFETs exhibits bi-stability only within HW_V (Fig. 6.4(b)). Therefore, the choice of V_R should be within HW_V in order to avoid accidental switching of the bit stored. In this work, we design our R-

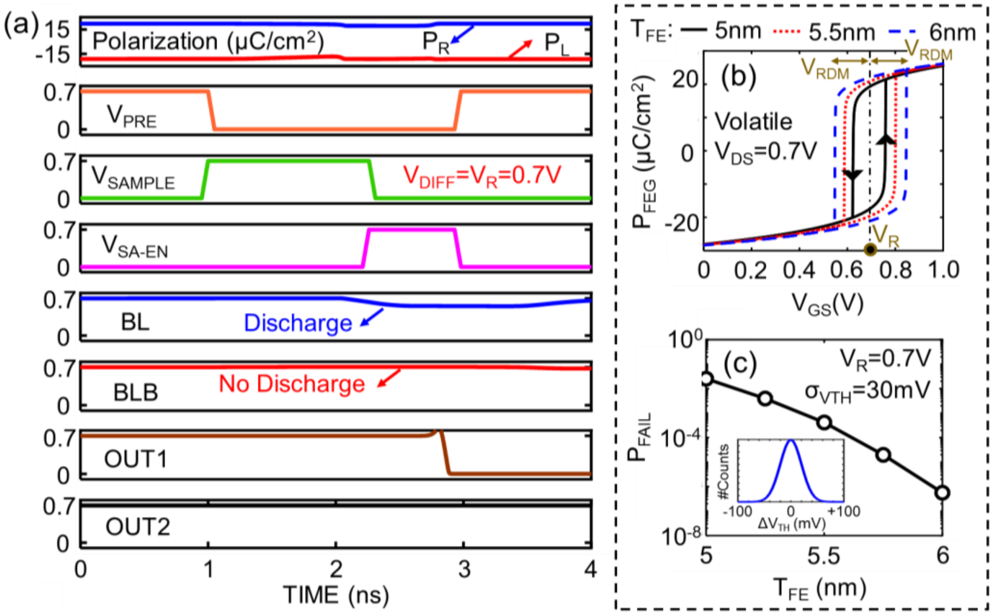


Fig. 6.5. (a) Transient waveforms during read. (b) P_{FEG} vs V_{GS} in the ‘V’ mode for various T_{FE}. (c) Probability of failure for the 4T-R memory, considering V_{TH} variations (inset shows the V_{TH} variation histogram).

FEFETs such that HW_V is sufficiently high (~300mV; Fig. 6.4(b)). We also perform variation analysis, in the next sub-section to determine the probability of failures for the proposed 4T-R. Moreover, note that bit-line discharging during the read does not disturb the state of the cell. This is because the discharging bit-line is connected to the gate of the R-FEFET storing P_{FEG}= -P. Therefore, discharging bit-line (reducing V_{GS} of R-FEFET) is conducive for –P in the FE and reinforces the stored polarization. This ensures high robustness during read, in the proposed 4T-R.

(c) Hold operation: For the hold mode, i.e., when the memory is not being accessed, the power supply is completely shut down (all signals are de-asserted; Fig. 6.2(d)). This brings V_{CS} of the R-FEFETs to 0V leading to storage of the bit information (P_{FEG}) in the bi-stable 'NV' mode. As a result, the proposed 4T-R exhibits zero stand-by leakage. However, before shutting OFF the supply, we pre-charge the bit-lines to V_R (=0.7V; Fig. 6.2(d)), and let them float. This reduces the energy associated with bit-line charging during the subsequent access (similar to SRAMs). Since no current is drawn from the supply, the advantage of zero standby leakage remains.

6.2.3 Variation analysis

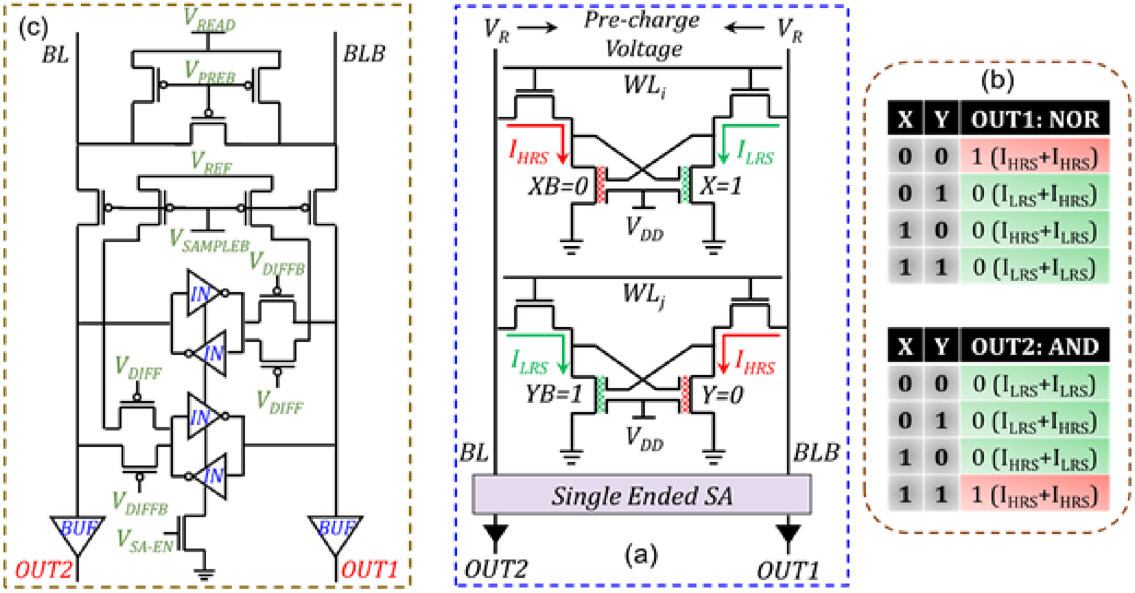
In order to understand the influence of variations on the 4T-R memory operation, we analyze the probability of failure (P_{FAIL}), considering variation in V_{TH} of the transistor for T_{FE} ranging from 5nm to 6nm (Fig. 6.5(b, c)). We consider a probability distribution function with standard deviation (σ) of 30mV for the V_{TH} variations. Note that, ferroelectrics such as HZO are deposited using atomic layer deposition, which achieves high precisions [190]. Therefore, we do not consider the variations in T_{FE} in our analysis.

For the write operations, failure is defined as the instance when V_{C+} >V_{DD} or V_{C-} <0V (see Fig. 6.4(a)). On the other hand, read disturb failure occurs when V_R lies outside HW_V (Fig. 6.4(b)). In this work, we design the R-FEFETs to ensure that during write, V_{DD} -V_{C+} and V_{C-} both are > 300mV (see Fig. 6.4(a)). Therefore, the probability of write failures is minimal with V_{TH} variations (as validated by our analysis). Moreover, read decision failures are also minimal due to the large distinguishability of the bi-stable states (>10^4). On the other hand, due to the selection of V_R being critical to ensure read stability (as mentioned before), the dominant factor of failure mechanism associated with 4T-R is the read disturb failure. It is well understood that increase in T_{FE} leads to larger HW_V which can lead to higher read-disturb margins (V_{RDM}; Fig. 6.5(b)) as discussed in Chapter-4 [171], and therefore lower P_{FAIL}. Our results show that for T_{FE} ~ 5.5nm to 6nm, P_{FAIL} as low as ~10^{-4} to 10^{-7} can be achieved for the proposed 4T-R (Fig. 6.5(c)). Based on this analysis and to ensure good cell stability, we use T_{FE}=6nm for our evaluations and results in this work.

6.2.4 Computation-in-memory using 4T-R NVM

In this section, we utilize the simultaneous true and complementary bit storage of the proposed 4T-R to enable energy efficient in-memory computation. The differential storage in 4T-R leads rise to the natural generation of bit-wise AND and NOR logic functions, by using the single ended configuration of the RSA. Utilizing the outputs of these logics in conjunction with CMOS logic gates, we propose a low power compact compute module to perform Boolean logic functions as well as arithmetic operations.

(a) Reconfigurable sense amplifier (RSA): As with previous CiM designs, the proposed CiM requires a modified sense amplifier (SA). In our work, we use a reconfigurable sense amplifier (RSA) proposed in [66] (Fig. 6.6) and integrate it with the proposed low power compact compute module (discussed later) to realize our R-FEFET-CiM. The RSA achieves



X	Y	OUT1: NOR
0	0	1 (I_{HRS}+I_{HRS})
0	1	0 (I_{LRS}+I_{HRS})
1	0	0 (I_{HRS}+I_{LRS})
1	1	0 (I_{LRS}+I_{LRS})

X	Y	OUT2: AND
0	0	0 (I_{LRS}+I_{LRS})
0	1	0 (I_{LRS}+I_{HRS})
1	0	0 (I_{HRS}+I_{LRS})
1	1	1 (I_{HRS}+I_{HRS})

Fig. 6.6. (a) Reconfigurable sense amplifier design proposed in [66]. (b) Example of multi-word line assertion and currents through R-FEFETs during compute operations (c) Truth tables for the naturally generated NOR (OUT1) and AND (OUT2) functions.

run-time reconfigurability between the differential and single-ended sensing. During differential mode, the RSA is self-referenced (with differential bit-lines as input), while in single-ended mode, a voltage reference, V_{REF} is used for sensing the bit-lines individually.

(b) Natural generation of bit-wise NOR and AND logics: The compute operation in the proposed architecture is based on the simultaneous assertion of two WLs in order to connect multiple bit-cells to the bit-lines (BL and BLB; Fig. 6.6(b)). Similar to the read operation of the memory, we first pre-charge the bit-lines to V_R=0.7V but now use the single ended configuration of the RSA by applying V_{DIFF}=0V (V_{DIFFB}=0.7V; Fig. 6.6(a)). The reference voltage for the single ended sensing is V_{REF}=0.65V. Depending on the bit values stored, we naturally generate bit-wise NOR (OUT1) and AND (OUT2) logic functions at the two ends of the RSA as shown in Fig. 6.6, without any additional circuitry. For example, when the two bits stored are X= ‘1’ and Y= ‘0’ (corresponding word-lines are WLi and WLj in Fig. 6.6(b)), BL and BLB (which are initially pre-charged) start discharging after the assertion of the WLs and CLs (since there exists an LRS path; Fig. 6.6(b)). And once they discharge to 0.6V, the sense enable is turned ON (V_{SA-EN}=0.7V) and since 0.6V< V_{REF}, the RSA brings OUT1 to 0V and OUT2 to 0V. The truth tables considering all other input combinations for OUT1 (NOR) and OUT2 (AND) are shown in Fig. 6.6(c). Therefore, we achieve natural and simultaneous generation of bit-wise Boolean AND and NOR logic functions with just one voltage reference. The generated outputs are then integrated with the compact compute module for the computation of other functions as discussed next.

(c) Compute module integrated with the RSA: We propose a low power and compact compute module (CM) as shown in Fig. 6.7(a), in order to realize in-memory computing, which includes bit-wise Boolean operations such as (N)AND, (N)OR, X(N)OR as well as arithmetic operations such as addition (ADD). First, the natively generated AND and NOR functions are simultaneously inverted (using standard CMOS inverters; Fig. 6.7(a)) to compute NAND and OR functions. The XOR functionality is achieved by using a CMOS based NOR logic with input operands being the naturally generated AND and NOR functions as shown in Fig. 6.7(a). Utilizing the bitwise Boolean operations discussed above, we also implement an in-memory ripple carry adder (RCA) using three additional CMOS logic gates. The carry-out (C_{OUT}) from the previous stage is propagated as carry-in (C_{IN}) to the next stage. In our evaluations carried out later, we consider a 32-bit word where the

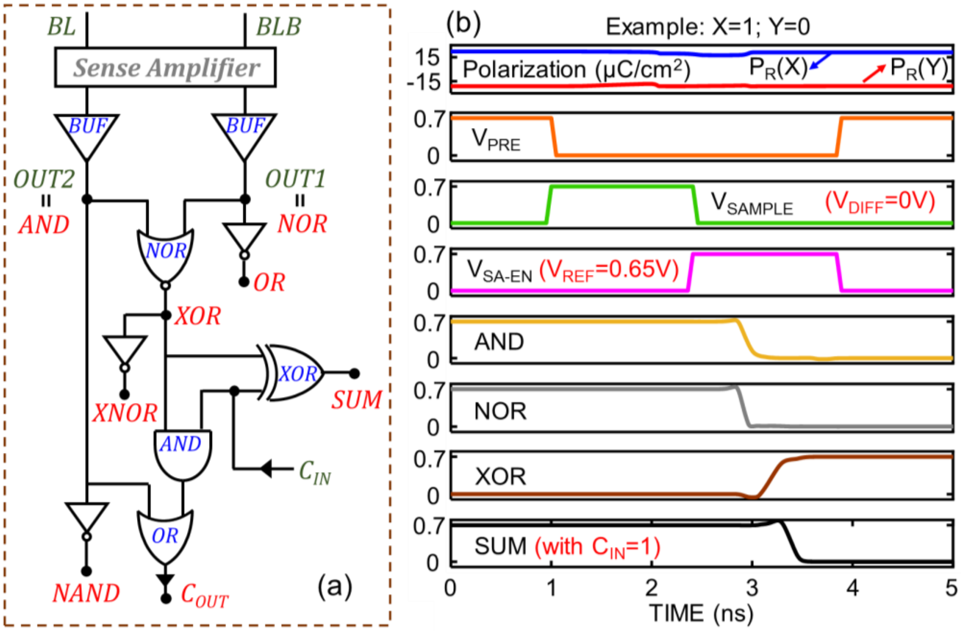


Fig. 6.7. (a) Proposed compute module/CM which is integrated with RSA. (b) Simulation waveforms showing various logic functionalities.

C_{OUT} to C_{IN} routing is performed in adjacent bit's SA within a word. The transient waveforms of compute operations for the two bits storing X= '1' and Y= '0' (example in Fig. 6.6(b)) are shown in Fig. 6.7(b) with C_{IN}= '1' (0.7V). Note that the area and energy overheads associated with CM are minimal, since it constitutes a small fraction of the total memory area/energy (total memory area/energy is typically dominated by the core array [59]).

(d) Comparison with other emerging NVM-CiM designs: Our R-FEFET-CiM architecture offers many useful features for compute operations compared to previously proposed NVM-CiMs. With respect to MAGIC based on RRAMs [203], we achieve a wide range of compute operations within a single access cycle, which holds true even for STT-CiM [59] and FEFET-CiM designs [64]. Compared to STT-CiM, we achieve high robustness while enabling two word-lines, due to large distinguishability (>10^4) between the bi-stable states. We also achieve significant energy-savings due to the natural and simultaneous generation of AND and NOR logics with one voltage reference, while STT-CiM requires two current references to compute (N)AND and (N)OR functions. Moreover, due to electric field driven information storage and computation in R-FEFET-CiM (vs. current based in MAGIC and STT-CiM), the overall energy efficiency is improved [64]. With respect to FEFET-CiM, our R-FEFET-CiM enhances the energy-efficiencies for performing logic operations due to: (i) natural and simultaneous generation of AND and NOR logic functions and (ii) low power CM with single voltage reference for in-memory computing (along with the other benefits achieved during standard memory operation). Moreover, FEFET-CiM requires a mix of voltage and current sensing schemes for read and compute (resulting in design overheads). On the other hand, R-FEFET-CiM utilizes voltage-based sensing for both read and compute. (Note that, although we use voltage-based sensing, the same operations can also be accomplished using current-based sensing.) We perform an extensive array and system-level analysis to quantitatively compare the proposed R-FEFET-CiM with FEFET-CiM, next.

6.2.5 Array-level results

We consider 1MB array (8 banks, each bank with 1024 rows and 1024 columns) with 32-bit words and evaluate the write, read and compute operations metrics for the proposed R-FEFET based CiM with standard FEFET-CiM architecture. Note, considering electric field driven memory storage (and in the interest of space), we perform our evaluations with respect to FEFET based design only. The benefits of FEFET-CiM over other NVM-CiM has already been shown in [64].

(a) **Write (Fig. 6.8(b)):** Due to the unique feature of dynamic reconfigurability in the R-FEFETs embedded in the proposed 4T-R, which allows the use of uni-polar voltage for the write operations, we achieve 50% lower write energy (WE) compared to the standard FEFET memory (3T). However, this comes at the cost of 14% higher cycle-to-cycle write time (WT). This is due to larger WL capacitance in the proposed 4T-R memory because of higher area. Iso-WT analysis (by increasing V_{DD} for 4T-R) shows 33% lower WE for 4T-R.

(b) **Read (Fig. 6.8(c)):** Due to the differential sensing enabled by the cross-coupling of the R-FEFETs in the proposed 4T-R, the cycle-to-cycle read time (RT) is 12% lower when compared to standard FEFET based 3T, which uses single-ended sensing scheme. Moreover, 25% lower read energy (RE) is achieved for 4T-R which is attributed to the lower voltage drop on the bit-lines (50mV BL discharge) due to differential sensing and self-referencing

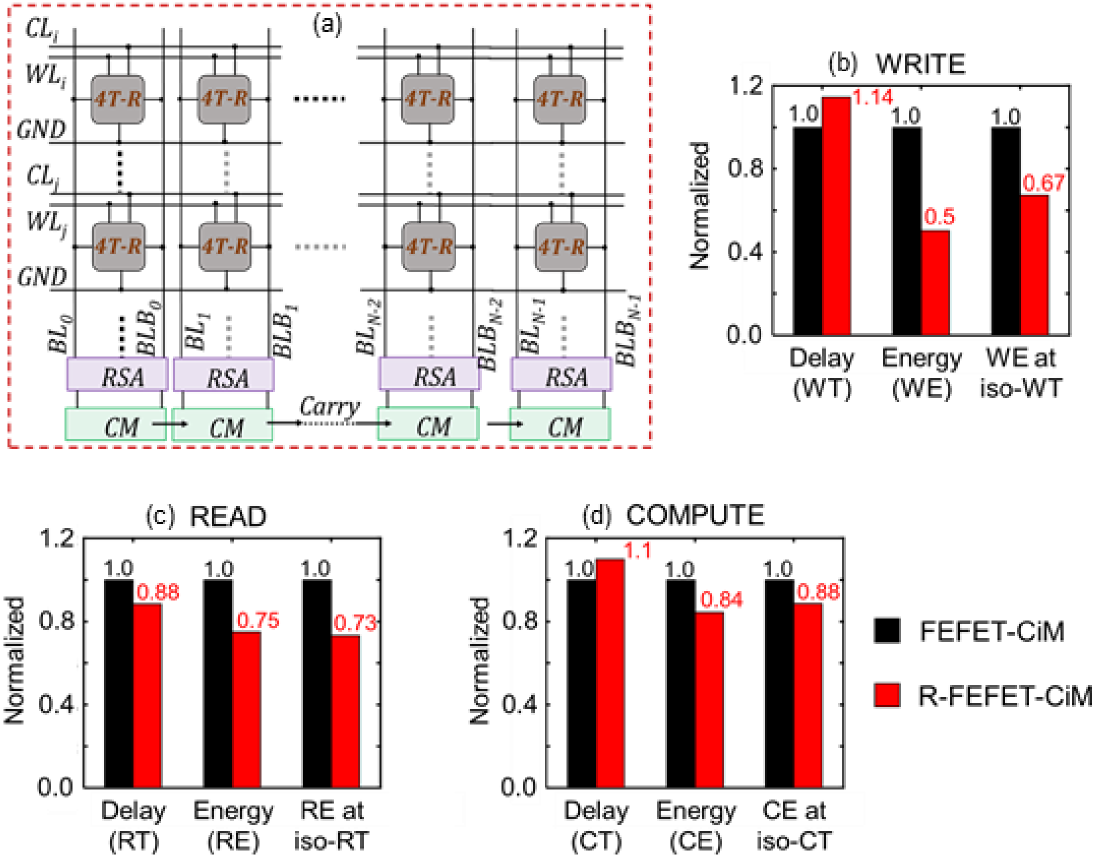


Fig. 6.8. (a) R-FEFET-CiM array architecture. Array-level comparison of the normalized delay, energy and energy at iso-delay for the (b) write, (c) read and (d) compute operations of the proposed R-FEFET-CiM with respect to FEFET-CiM.

in 4T-R, compared to the 100mV BL discharge required for single-ended sensing in 3T. At iso-RT (achieved by increasing V_R for 3T), RE of 4T-R is 27% lower than 3T.

(c) Compute (Fig. 6.8(d)): We consider the worst-case configuration to analyze the energy and delay metrics for compute operations. X=1 and Y=0 (or X=0 and Y=1) is the worst-case scenario since in this case both BL and BLB discharge during the compute (Fig. 6.6). Therefore, the cycle-to-cycle compute time (CT) is 10% higher than 3T, which requires only one bit-line pre-charging after compute operation in the worst case. However, due to the proposed low power compact compute module (Fig. 6.7(a)), the proposed 4T-R exhibits 16% lower compute energy (CE) compared to 3T [64], which requires a more complex compute circuitry along with a mix of voltage and current based compute operation schemes. At iso-CT (by decreasing V_R for 3T), RE of the proposed 4T-R is 12% lower when compared to 3T.

6.2.6 System-level results

For the system-level evaluations discussed in this sub-section, we collaborated with Prof. Anand Raghunathan and Dr. Shubham Jain to understand the utility of the proposed R-FEFET based 4T-R compute enabled NVM, for general purpose computing.

(a) Simulation framework: Fig. 6.9 shows the system-level framework used for our evaluations, wherein the proposed R-FEFET-CiM is integrated as a 1-MB scratchpad for the Intel Nios II processor. To expose CiM operations to software, we add custom instructions to the Nios II processor's instruction set. We also extend the Avalon on-chip bus to support CiM operations. We perform cycle-accurate RTL simulation to obtain the execution time and the memory traces for our benchmark applications [59]. Using these traces and the array-level results, we estimate the system-level energy and performance benefits. We compare R-FEFET-CiM with two baselines (iso-capacity): (i) a standard FEFET memory without CiM support (FEFET-Non-CiM), and (ii) a standard FEFET memory with CiM support (FEFET-CiM). Further, we design FEFET-CiM to be iso-latency with the proposed R-FEFET-CiM (see iso-WT/RC/CT in Fig. 6.8 (b, c, d)).

(b) Performance analysis: Fig. 6.10(a) details the normalized execution time for the R-FEFET-CiM, FEFET-CiM, and FEFET-Non-CiM designs for various applications. We observe that CiM designs reduce all facets of execution time, i.e., memory accesses, data transfers over the on-chip bus, and instructions executed in the processor, and achieve 5% to 23% speedups over the Non-CiM FEFET baseline across our benchmarks.

(c) Energy analysis: Here, we present the total system energy benefits for various benchmarks (Fig. 6.10(b)). We show all major components of energy, viz., memory, interconnect, and processor. R-FEFET-CiM achieves total system energy savings of 17% to 27% over the FEFET-Non-CiM baseline and 8% to 24% over FEFET-CiM. The benefits primarily arise

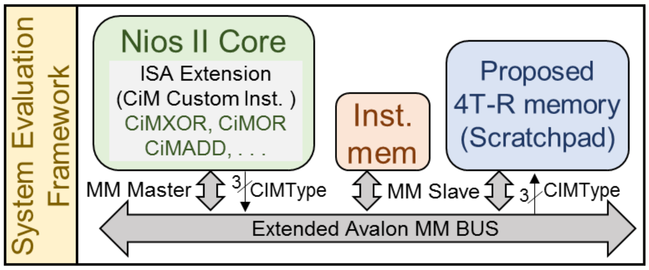


Fig. 6.9. Simulation framework used for system-level evaluation

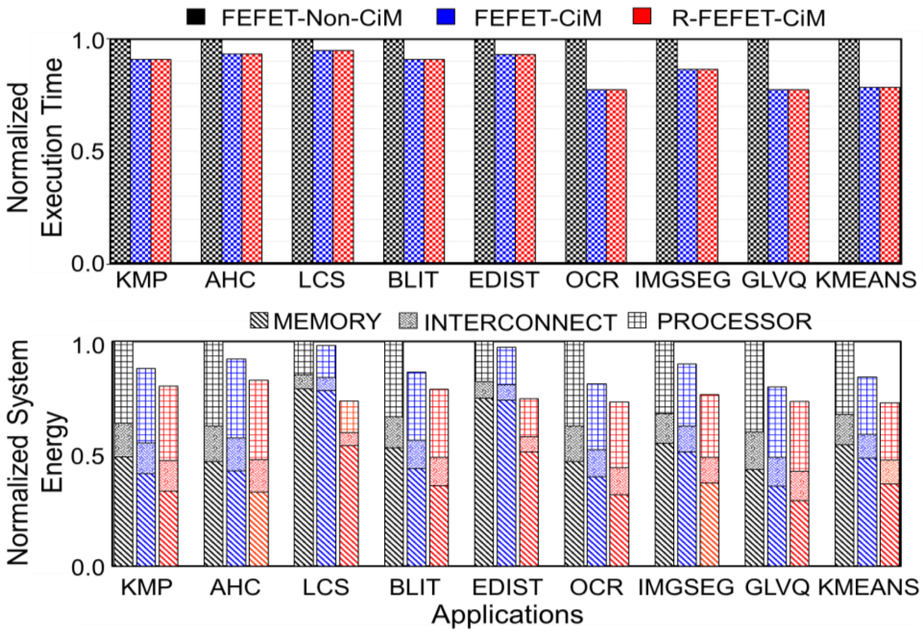


Fig. 6.10. (a) Execution time and (b) total system energy consumption of the proposed R-FEFET-CiM in comparison with FEFET-CiM and FEFET-Non-CiM for various application benchmarks. Iso-performance for FEFET-CiM and R-FEFET-CiM has been achieved by tuning V_{DD}.

due to energy-efficient CiM operations based on the low power CM. CiM operations reduce memory accesses, bus transfers and processor instructions leading to savings across all energy components. Further, R-FEFET-CiM also benefits from its superior read due to differential sensing and improved write due to the uni-polar voltage design.

6.3 IPS-CiM: FEFET based CiM Hardware for Performing IPS Workloads

So far, we have looked at utilizing R-FEFETs for enhancing CiM architectures for performing general purpose workloads which have their own specific requirements. Next, we turn our attention to intermittently powered systems which have gained a lot of traction in recent years due to the uprising demands of IoT devices and wearables. Now, applications such as IPS need more complex CiM operations, for example: in-memory comparison (as discussed later in this section), which requires the design of a novel IPS specific CiM technique. This has largely remained unexplored in the community and in this section, we address this challenge with the proposal of the IPS-CiM architecture. As mentioned earlier, the techniques proposed in this section can be applicable to a wide range of memory technologies and our focus is primarily on evaluating the merits of the proposed IPS-CiM. Since these platforms are extremely energy-constrained, we use low power, electric field driven FEFETs for the discussion and evaluation in this section. It is important to note that, R-FEFETs can further enhance the benefits mainly due to their device-level advantages, as discussed in the previous chapters.

6.3.1 Need for energy-efficiency in IPS

Energy autonomous systems have gained an immense popularity in recent years due to the advent of Internet of Things (IoT) [207] and Body-Area-Networks [208]. These systems are battery-less and their operation depends on the energy harvested from ambient sources [194] such as solar, thermal, RF, motion, etc. (Fig. 6.11). Compared to systems which run on batteries, energy-harvesting systems have larger lifetimes and are more environment friendly. However, these systems face several challenges such as: (a) sporadic nature of ambient sources leads to frequent power failures, (b) low output power due to the small size factor and limited efficiency of the harvesters and (c) unpredictable harvested energy pattern. These problems lead to repeated system shut-downs, resulting to loss in computation progress [194].

As discussed in the previous chapter, non-volatile computing enables the system to overcome the drawback of loss in computation progress by backing-up the states of the processor (stored in flip-flops/ registers) as well as the SRAM (on-chip memory) to a non-volatile memory (NVM) before a power shut down occurs [197]. When the power supply is re-established, the states of the processor and SRAM data is restored leading to zero loss in the computation progress. Such a systematic consistency-aware check-pointing mechanism have been explored to avoid data inconsistency and irreversible computation errors due to erratic power failures [209]. However, the sequential long-distance data movement between processor and NVM in standard von-Neumann architectures creates energy and performance bottlenecks [210].

Another important requirement for IPS is to achieve reduced energy consumption for various workloads which is critical for the small power budgets associated with ambient sources. An example of a major energy component in IPS is the transmission of data to cloud/ host processor for computing [211]. Previous works suggest large inefficiency with wireless transmission for computation when compared to in-situ computation [211]. However, with the advent of data-intensive workloads, which require mammoth processing, the need to reduce computation energy in IPS has never been greater.

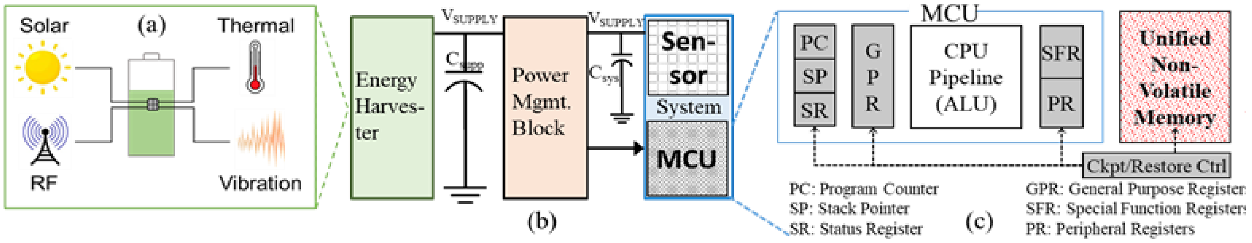


Fig. 6.11. (a) Sources for energy harvesting (b) Conceptual diagram of an intermittently powered system (IPS) (c) MCU core registers and unified NVM.

In order to address the needs for mitigating the von-Neumann bottleneck as well as achieving low power computation, several works have proposed compute-in-memory (CiM) for energy harvesting systems. Resistive RAM based CiM design has been proposed which performs non-volatile logic and neuromorphic computations [212]. Although intriguing, such a design mainly focuses on machine learning based workloads. Similarly, SONIC in [213] introduces compressed neural networks in IPS to perform inference of vision-based workloads. This work is significant as it improves the error-resiliency of machine-learning workloads on an energy harvesting platform. Other CiM designs have been proposed for SRAM with Neural Cache [214], DRAM with Compute-DRAM [215], for NVMs with Pianotube [60], FEFET-CiM [64], etc. It is noteworthy that these technologies are meant to be integrated into the memory hierarchy of traditional CPUs and have not been considered specifically for energy harvesting applications. Compute-DRAM and Neural Cache based on DRAM and SRAM may not be preferable for IPS due to their volatile storage. Pianotube uses current-driven emerging NVMs, which are energy inefficient compared to voltage-based information storage. FEFET-CiM (voltage-driven storage) uses both voltage and current based sensing, leading to high energy expenditure and design complexity, which may not be suitable for IPS. Hence, there is a need to explore CiM for energy harvesting systems, capable of computing a wide range of workloads within an energy efficient NVM.

In this section, we address this critical need by proposing IPS-CiM based on FEFETs to enhance the energy efficiency of IPS. To the best of our knowledge, this is the first effort on introducing CiM operations in IPS with processing of wireless sensory network (WSN) workloads for edge-sensing and error detection using cyclic redundancy check.

6.3.2 Transient computing workloads

In this work, we leverage two of the most popular WSN–based applications to show benefits of using IPS-CiM viz., (i) CRC and (ii) SENSE, used frequently during communication and sensing [197], [198].

(a) **Error detection using cyclic redundancy check:** Cyclic Redundancy Code (CRC) is a popular error-detection algorithm that is used to determine the correctness of data transmission or storage [216]. The fundamental mathematics behind the CRC is modulo polynomial division. In CRC-n code, a message is augmented with n parity bits. The parity bits represent the remainder of the division of the message appended with n 0s with a pre-determined nth-order polynomial. A received code word (original message + parity bits) evenly divisible by the same nth-order polynomial implies no single bit transmission errors. The native implementation for computing and checking a CRC is bit-based which typically makes it suitable for hardware implementation [216]. For ultra-low power MCUs, CRC is implemented either using a bitwise algorithm (low memory, low cost) or a table-based algorithm (low instructions/second, low power). In this work, we use the table-based CRC solution that trades off execution cycles for memory accesses allowing a processor to operate on bytes rather than bits.

(b) **Sensing and data aggregation (SENSE):** WSN is an ensemble of small computing devices whose main task is to acquire data through its various sensors, process it and send it to the requestor. Due to their miniature size and frequent remote deployments, these devices have very limited energy budget and network bandwidth and may operate using energy harvested from the environments, thus functioning as an IPS. One of the most important applications of WSNs is in-network sensor data aggregation [217]. Data acquired from one or more sensors gets aggregated in a host (sink) node using various statistical methods and the aggregated value is sent back to the requestor/user. Since there are hundreds of such deployed sensors, any query to the sensors leads to flooding of the entire network with a large number of packets from the individual nodes to the host, each containing only a small sensor value. This leads to rapid depletion of network energy causing a drastic reduction of network lifetime. To conserve both energy and bandwidth, it is essential to filter and condition the sensor data within the network itself using in-network aggregation [217] where sensor readings are accumulated at the intermediate nodes. In-network data aggregation has been shown to increase the accuracy of results (by eliminating faulty outlier sensor values) while reducing the number of packets, the probability of packet collisions and data redundancy. Common data aggregation functions that show desirable properties such as duplicate sensitivity, summary, monotonicity, and partial state requirements are average, min, max, sum, and variance [217].

6.3.3 Proposed FEFET based compute enabled memory

In this section, we present a 3T NVM (Fig. 6.12(a), which utilizes the polarization of FEFET to store the bit-information. Our memory utilizes a bit-cell which is inspired from the design proposed for NVM storage in [192] and comes with the enablement of CiM. Let us briefly discuss the memory design and operations. The bit-cell exhibits separate read-write paths which allows for immense design flexibility while performing CiM as discussed later. Read and write access transistors are connected to the drain and gate terminal of the FEFET. The schematic and layout of the 3T NVM are shown in Fig. 6.12(a) and Fig. 6.12(e) respectively. The array organization consists of the write/read bit-lines (RBL/WBL) connected across cells in a column. The write/read word-lines (RWL/WWL) and the plate-line (PL) are shared amongst cells in a row (as in [192]) unlike the design in [64] which shares PL in column. This leads to energy efficiency and avoids

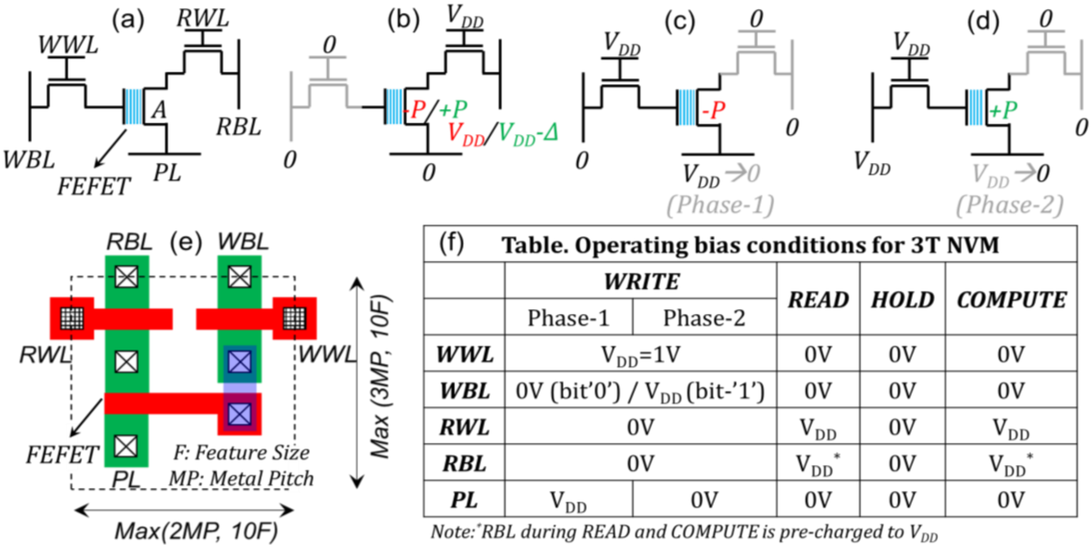


Table. Operating bias conditions for 3T NVM

	WRITE		READ	HOLD	COMPUTE
	Phase-1	Phase-2			
WWL	V_{DD}=1V		0V	0V	0V
WBL	0V (bit'0') / V_{DD} (bit-'1')		0V	0V	0V
RWL	0V		V_{DD}	0V	V_{DD}
RBL	0V		V_{DD}^*	0V	V_{DD}^*
PL	V_{DD}	0V	0V	0V	0V

Note: *RBL during READ and COMPUTE is pre-charged to V_{DD}

Fig. 6.12. (a) 3T FEFET memory cell schematic. Schematic with biasing for (b) Read, (c) Write ‘0’ in phase-1 and (d) Write ‘1’ in phase-2. (e) Layout and (f) operating bias conditions of 3T FEFET NVM.

write disturbs in unaccessed cells as discussed later. Fig. 6.12(f) shows the biasing conditions and the array is shown in Fig. 6.14(b). Note, although more compact 1T/2T FEFET-NVMs have been proposed previously [192], [218], they require (a) negative voltages (additional bias circuitry) or/and (b) charging of all the unaccessed WLs and BLs (connected in a cross-point fashion). These lead to energy overheads and design complexities due to the need for multiple voltages, which may not be suitable for the energy-constrained IPS, targeted in this work. Therefore, we design 3T FEFET-NVM to avoid multiple voltages (as discussed next) and other energy/design costs associated with 1T/2T NVMs.

(a) Standard read-write operations:

(i) Write (Fig. 6.12(c,d)): For storing bit-‘0’ or bit-‘1’, we drive the polarization of the FEFET to negative (-P) or positive (+P) state respectively. This is achieved by a 2-phase operation [192], where the first and second phase is used to write ‘0’ and ‘1’ respectively. WWL is asserted and WBLs of the accessed cells are driven to either 0V or V_{DD} to write ‘0’ or ‘1’ respectively. This is followed by PL being driven to V_{DD} during the 1st phase and then to 0V in the 2nd phase. In the 1st phase, bit-cells with WBL = 0V store -P (‘0’) as V_{GS} of FEFET=-V_{DD}. In the 2nd phase, bit-cells with WBL=V_{DD} have V_{GS} of FEFET=V_{DD}, which writes +P. Note, V_{GS} of FEFETs which are supposed to store ‘1’ and ‘0’ remains at 0V in phase-1 and phase-2 respectively. This avoids write disturbs because FEFET exhibits bi-stability at V_{GS}=0V. After write, all the lines are brought to 0V and the polarization retention in FEFET in the absence of electric field enables non-volatile storage.

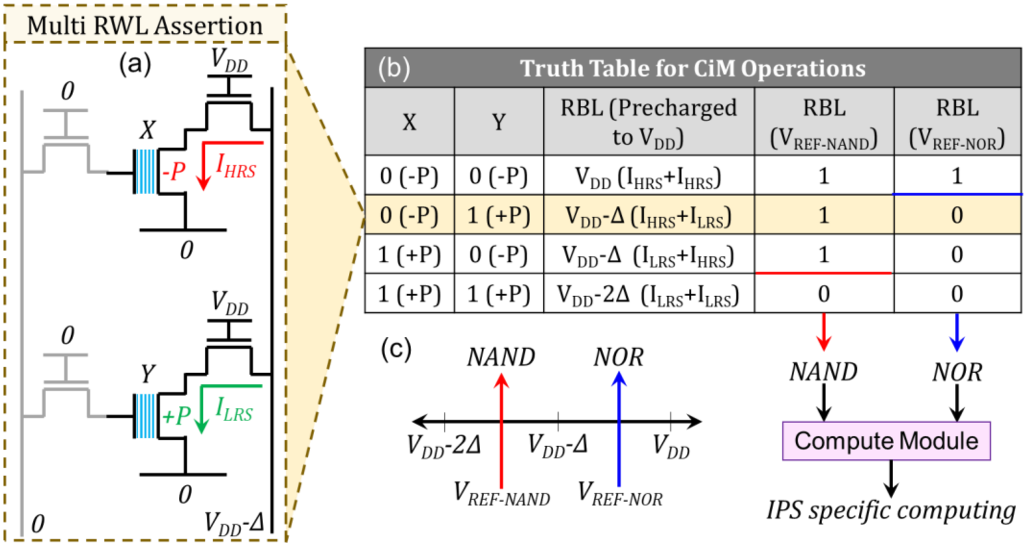
The unaccessed cells in the column avoid write disturbs with WWL and PL being driven to 0V. This might not be true with FEFET-CiM design in [64] where PL is shared across a column which can potentially disturb the unaccessed cells. Moreover, the design in [64] expends larger write energy compared to the proposed cell due to the requirement to drive multiple PLs with the 2-phase voltage scheme. Also, the system which we target in this sub-section (IPS) can support write access of all bit-cells in a row at once and therefore, we avoid the overheads associated with biasing WBL at an intermediate voltage (V_{DD}/2) for unaccessed cells, which is required to eliminate write disturbs in standard NVM storage [192].

Although FEFET-NVMs have been proposed with single phase write previously, this comes with either the requirement of (a) negative voltages or (b) additional access transistor (increase in cell area), both of which lead to energy overheads. Therefore, we use of two-phase operation along with a compact bit-cell for the targeted energy-constrained IPS. Note, a two-phase scheme inherently comes with higher latency compared to single-phase for write operation (both of which are in the orders of ~ns). However, since the targeted IPS runs at ~MHz, there is no significant impact of two-phase operation on system performance (discussed later).

(ii) Read (Fig. 6.12(b): For sensing, RWL is asserted with RBL pre-charged to V_{DD} and PL driven to 0V. Depending on the polarization stored, there exists an HRS (-P; '0') or LRS (+P; '1') current through the bit-cell. For HRS, the current is negligible [64], [192] and RBL remains at V_{DD} (=1V). For LRS, current is significant which discharges RBL to V_{DD}-Δ (=0.9V), where Δ is drop in RBL voltage (=100mV in this work). Single-ended voltage sensing is performed with a reference voltage, V_{REF}=0.95V to read the bit stored. Note, during read, WWL and WBL are de-asserted.

Both current and voltage-based sensing can be used to read the bit state. In view of voltage-based sensing being more efficient, we implement it for both read and compute operations as discussed later. This also enables easy integration of the compute module with our modified sense amplifier to perform computing-in-memory, which is discussed below.

(b) Compute-in-memory: The computation within memory array is based on the technique of multi-wordline assertion, used in several previous works [59], [64], [66], [219]. This technique is coupled with our proposed voltage-based modified sense amplifier (MSA) design, specifically targeted for the single-ended FEFET NVM. The outputs of the MSA are coupled with the low power compute module (CM) capable of performing a wide range of Boolean and arithmetic functions. Based on the requirement of the workloads in IPS, we (a) implement addition/subtraction/ standard logic operations with the CM and (b) propose for the first time a low-power in-memory comparison operation. Before discussing the various in-memory computations, we briefly discuss the concept of multi word-line assertion in the FEFET NVM, next.



(b) Truth Table for CiM Operations

X	Y	RBL (Precharged to V_{DD})	RBL ($V_{REF\text{-}NAND}$)	RBL ($V_{REF\text{-}NOR}$)
0 (-P)	0 (-P)	V_{DD} (I_{HRS}+I_{HRS})	1	1
0 (-P)	1 (+P)	V_{DD}-Δ (I_{HRS}+I_{LRS})	1	0
1 (+P)	0 (-P)	V_{DD}-Δ (I_{LRS}+I_{HRS})	1	0
1 (+P)	1 (+P)	V_{DD}-2Δ (I_{LRS}+I_{LRS})	0	0

Fig. 6.13 (a) Example of multi RWL assertion for in-memory computation. (b) Truth table for RBL discharge during multi-RWL assertion. (c) Voltage reference location for achieving bit-wise NAND and NOR logics.

(i) Multi-RWL assertion: By simultaneously asserting two read-word lines, bit-wise logic operations between the binary states stored in the two bit-cells of the same column can be achieved (Fig. 6.13). Similar to read, RBL is initially precharged to V_{DD} (=1V) and PL is driven to 0V. Depending on the polarization stored in the two bit-cells, we observe either 0, Δ or 2Δ drop in the RBL voltage, where Δ (=100mV) is the voltage drop corresponding to activating one bit-cell storing +P (as discussed for read in the previous sub-section). An example of RBL discharge from two simultaneously activated bit-cells X and Y storing -P and +P, is depicted in Fig. 6.13(a) and the complete truth table is shown in Fig. 6.13(b).

Now, depending on the positioning of the reference voltage for sensing, one can achieve either NAND or NOR operation between the two bit-cells activated as shown in Fig. 6.13. V_{REF} = $V_{REF\text{-}NOR}$ = 0.95V (V_{DD}- Δ/2; same as reference voltage for read operation) results in bit-wise NOR operation. On the other hand, V_{REF} = $V_{REF\text{-}NAND}$ = 0.85V (V_{DD} - 3Δ/2) leads to bit-wise NAND operation between bit-cells X and Y (Fig. 6.13(c)). Previous work on FEFET based CiM require the use of voltage and current based sensing to perform (N)AND and (N)OR logic. This requirement of mixed sensing schemes leads to design complexity and energy overheads (due to the requirement of constant DC current during computations). However, in this work, we present an all voltage-based modified sense amplifier design which performs bit-wise (N)AND and (N)OR operations as discussed next.

(ii) Modified Sense Amplifier: To perform the above-mentioned bit-wise operations, we propose to duplicate the RBL voltages across two single ended, cross-coupled inverter-based SAs as shown in Fig. 6.14(a). The other ends of the two cross-coupled sense amplifiers are connected to $V_{REF-NAND}$ and $V_{REF-NOR}$ respectively (Fig. 6.13(c)) during the sampling phase (V_{SAMPLE}=V_{DD} - Fig. 6.14(a)). After sampling, the sense amplifiers are enabled to obtain the output.

The final voltages from the MSA correspond to (N)AND and (N)OR operation between the activated bit-cells (Fig. 6.14(a)). We use AND along with NOR for performing other Boolean logic as well as arithmetic operations with an integrated CM (discussed next). Note, during read operation, i.e, when one RWL is asserted, we enable only the sense amplifier with V_{REF}=0.95V, to sense the bit stored. The inverted end of NOR output is used to achieve READ (or OR) output which corresponds to the bit stored (Fig. 6.14(a)).

(iii) Compute Module (CM): Here, we propose a low power compute peripheral (Fig. 6.14 (c, d)) which is directly integrated with the above-mentioned MSA outputs: AND, NOR and READ. We design two variants of the CM where, Fig. 6.14(c) is implemented for the MSB bit of the word while, Fig. 6.14(d) is used for all other bits from LSB to MSB-1. The requirement of a different CM for MSB is discussed in the next section. The two variants exhibit no significant difference in energy /performance (verified by our simulations). In the following, we discuss the key arithmetic operations in our CiM design. Note that, all the CiM operations discussed below, harness the advantage of standard IPS, where an entire row can be read-out simultaneously (due to high access granularity [220]). This enables the proposed IPS-CiM to perform parallel computations across all columns of the array.

- *Addition and Subtraction:* We utilize the approach presented in several previous works [59], [64], [66], [155], [219] for in-memory addition and subtraction using our proposed MSA and CM. Firstly, in both variants of the afore-mentioned CMs, we perform bit-wise XOR between bit-cells X and Y using a CMOS NOR gate with input operands being the natively generated AND and NOR logics in the MSA as

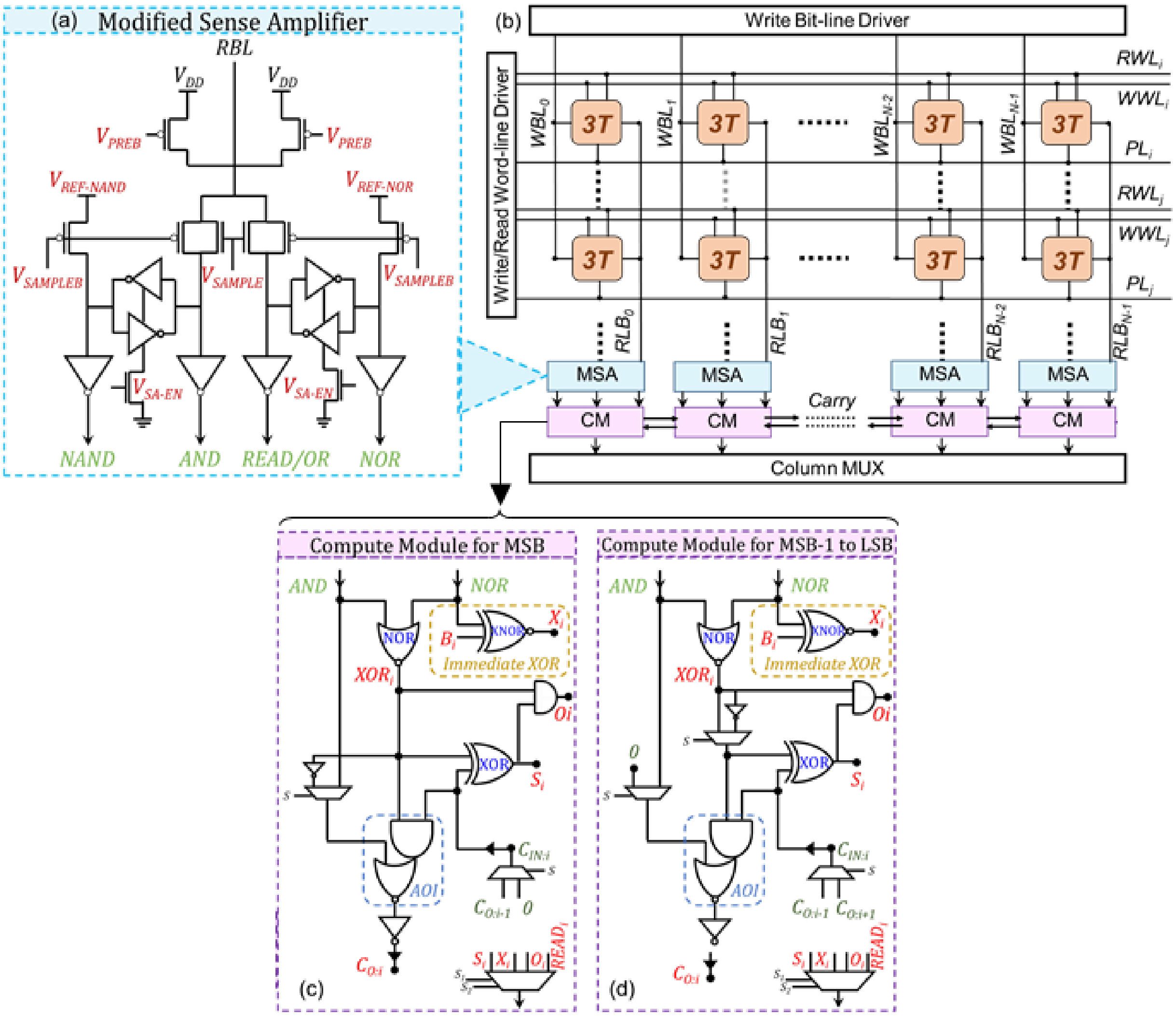


Fig. 6.14 Modified sense amplifier used to simultaneously generate AND and NOR logics. (b) Schematic of the 3T NVM array (c) Compute module used in the column of (c) MSB bit and (d) all other bits of a word. S is the select signal for 2:1 MUXs in the compute modules.

described before. Now, using the generated Boolean functions we use additional CMOS gates to perform 1-bit addition (Fig. 6.14(c, d)). In order to execute the addition of two words present in the same column, the C_{OUT} and C_{IN} of adjacent CM are connected together (C_{OUT} of right bit's CM to C_{IN} of left bit's CM; Fig. 6.14(b)) to perform a ripple-carry addition. Note, the entire ripple carry addition occurs in just a single array access. For in-memory subtraction (say, A-B), we perform in-memory addition of A and 2's complement of B. For that, we obtain B' (i.e., read and write-back inverted bits of B) and then perform addition with carry-in initialized to '1' at the LSB [214]. Note, select signal of 2:1 MUXs in CM is set to '0' while performing addition or subtraction ((Fig. 6.14(c, d)). Compared to [64] which requires hybrid current- and voltage-based sensing for CiM, the proposed CM utilizes an all voltage-based sensing, resulting in simpler and energy-efficient CiM design. With respect to bit-serial CiM proposed in [214], [221], the CiM design in this sub-section performs bit-parallel computing resulting in performance efficiency.

- *Immediate XOR:* The CiM fabric proposed also exhibits the feature of evaluating immediate XOR function, where one input operand (A) is stored in the array while the other operand (B) is applied directly at CM as an external input (see Fig. 6.14(c, d)). We use an XNOR gate in the CM to compute immediate A XOR B. This computation is useful for performing intermediate steps during computation of algorithms within the memory as discussed next.
- *Comparison (maximum or minimum or equal; Fig. 6.15):* We propose a novel method for comparison of two numbers and its low power hardware implementation using the proposed CiM engine. To achieve in-memory comparison, we define a new operation called Pseudo-ADD in which carry propagation takes place from MSB to LSB, as discussed later. Conventionally, near-memory standard comparators have been employed or multiple memory read/write operations are required to perform in-memory comparison [214]. However, this comes with severe area/energy-overheads and/or large near-memory peripheral design complexity. We overcome such issues by proposing an unconventional method for in-memory comparison, with minimal hardware additions to the CM employed for addition/subtraction. The proposed in-memory comparison operation is executed in

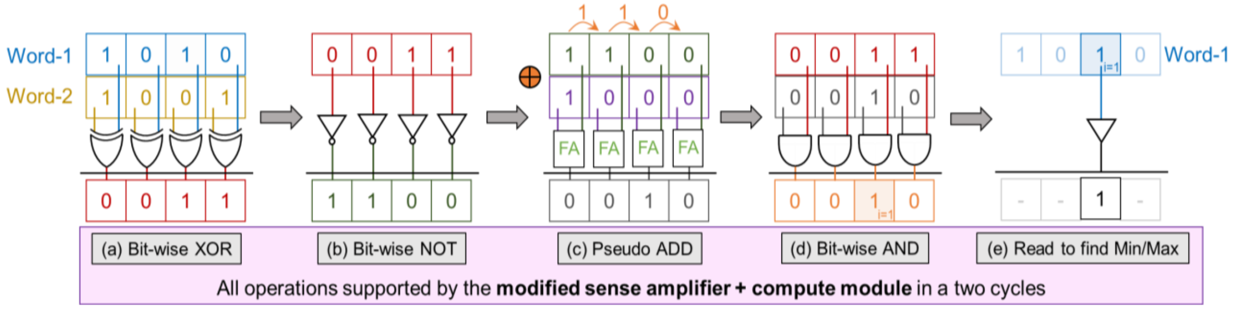


Fig. 6.15 CiM flow for performing minimum/maximum operations. (a) Bit-wise XOR (b) bit-wise NOT of the XOR result (c) pseudo-ADD operation of the NOT result (d) bit-wise AND of XOR and pseudo-ADD results (e) Read bit of any word corresponding to the set bit location of the AND result.

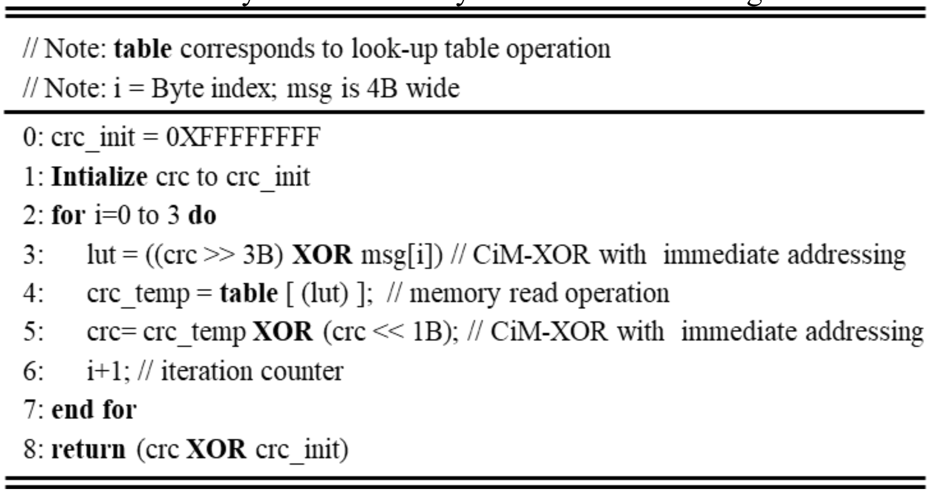
2 cycles. First, the bit-wise XOR of two words is computed (Fig. 6.15(a)). The XOR outputs corresponding to 0 indicate that same bit is stored in the respective bit indices, while XOR outputs =1 correspond to the locations where there exists a mis-match. The location of the first mis-match in the direction from MSB to LSB can reveal the comparison between the two words.

Now, to determine the location of the left-most set bit (= 1) in the XOR output vector, we first invert all its bits (Fig. 6.15(b)). After this, we perform a Pseudo-ADD of the resultant vector with another word W (where W is a constant with MSB=1 and the rest of the bits = 0; (Fig. 6.15(c))). In contrast to standard addition, the proposed Pseudo-ADD corresponds to the ripple-carry progressing from left to right instead of right to left [59], [64], [219], which is enabled by setting the select signal of the 2:1 MUXs in CM to '1' (Fig. 6.14(c, d)). After this, we perform bit-wise AND of initial XOR and the result of pseudo-ADD (Fig. 6.15(d)). The resulting value will have only one bit set to '1' which corresponds the bit-location of the first mismatch (from left to right) between the two words being compared. The pseudo-ADD operation discussed above, is the reason for having separate CMs for MSB and rest of the bits (Fig. 6.14(c, d)). Note, all of the aforementioned computations occur in a single cycle with our CM, with pseudo-ADD being performed in the CM itself.

In the second cycle of the comparison process, we read the bit information of one of the words corresponding to the left most set bit found in the 1st cycle (Fig. 6.15(e)). If the sensed value is '1' then the corresponding word is the maximum, else, it's the minimum of the two. If all bits after 1st cycle are 0 then the two words are equal.

The proposed method for comparison is highly energy- and area- efficient compared to traditional approaches (non-CiM) considered in IPS using standard 16/32-bit CMOS comparators. The proposed in-memory comparison only requires an inclusion of 2 or 3 MUXs (depending on CM), 1 NOT and 1 AND gate in the CM (Fig. 6.14(c, d)), on top of what is required for addition and subtraction. Moreover, we achieve massive parallelism in performing several comparisons simultaneously, unlike pipelining performed in standard architectures (non-CiM).

Table. 6.1 Cyclic Redundancy Code: CiM-based Algorithm



```
// Note: table corresponds to look-up table operation
// Note: i = Byte index; msg is 4B wide
crc_init = 0XFFFFFFFF
Intialize crc to crc_init
for i=0 to 3 do
    lut = ((crc >> 3B) XOR msg[i]) // CiM-XOR with  immediate addressing
    crc_temp = table [ (lut) ];  // memory read operation
    crc= crc_temp XOR (crc << 1B); // CiM-XOR with  immediate addressing
    i+1; // iteration counter
end for
return (crc XOR crc_init)
```

When compared to [221] which requires 2N+1 cycles (where N is the word-length) to compare two words in a bit-serial fashion, the proposed bit-parallel CiM comparison is performed in just 2 cycles (independent of the word-length). The performance-energy evaluation has been discussed quantitively further later.

6.3.4 Enabling transient computing workloads in-memory

We now propose the application of CiM discussed previously to perform two important transient computing workloads: (a) CRC for error detection and (b) edge-sensing in WSNs.

(a) Cyclic redundancy code (CRC): CRC is commonly used to ensure the correctness of data storage and transmission during communication protocols. IoT devices require the transmission of data to cloud for intensive computation which cannot be performed at the edge due to constrained power budget or for storage in the database because of the limited on-chip memory availability. Table. 6.1 depict the algorithm used for CRC.

Fig. 6.16 illustrates the CRC implementation on a message of 32-bits. Because the message is 4 bytes (4B) long, the computation is performed in 4 iterations. In the following, CRCBj-i refers to jth byte of CRC in the ith iteration. Now, in the 1st iteration, we initialize CRC to CRC_{INIT} = 0XFFFFFFFF. Then immediate XOR (8-bit) is performed with 1st byte of the message (stored in the memory) and the 1st Byte of CRC (=CRC_{B3-0}=11111111). The resulting 8-bit output is sent to an address decoder of pre-stored memory with 256 words (each with 32-bit width). Thus, the look-up table operation is performed and the 32-bit

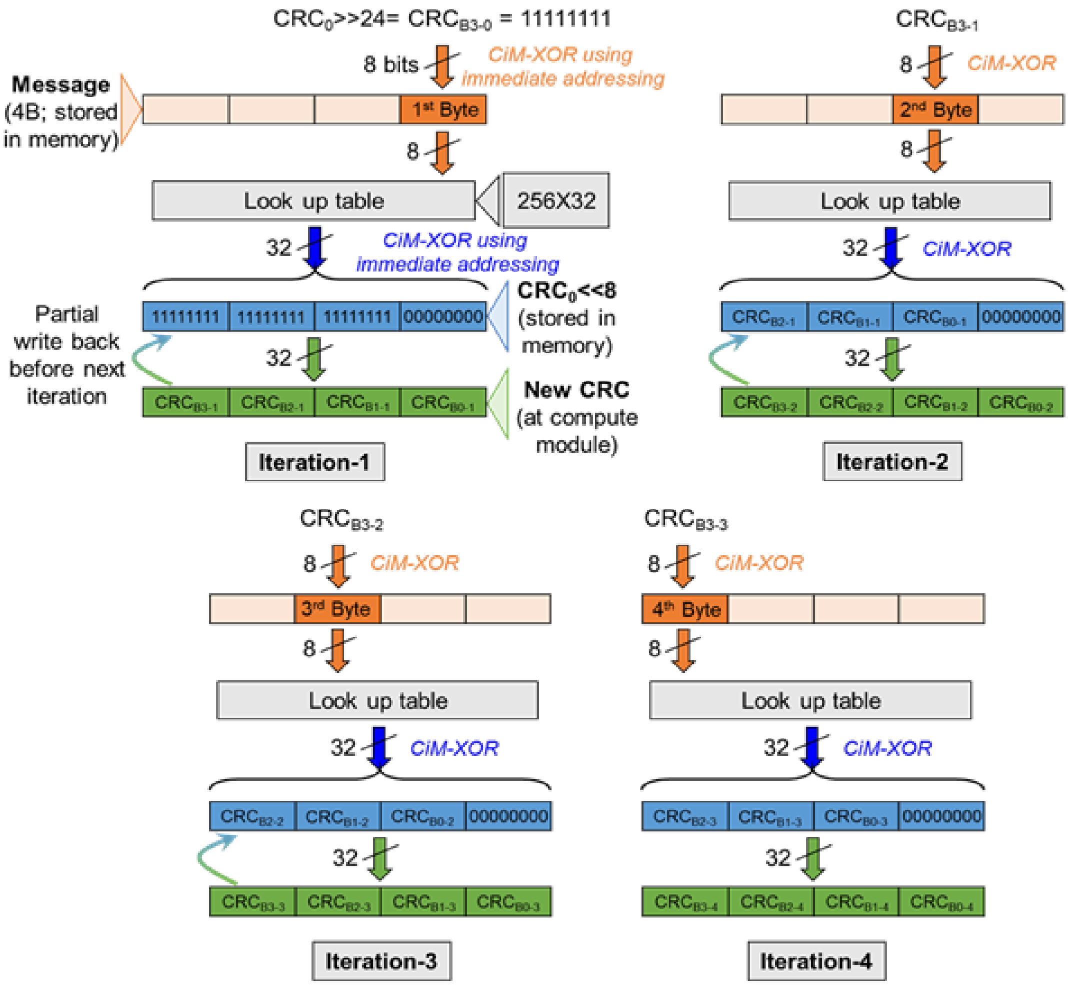


Fig. 6.16 CiM flow for performing cyclic redundancy check. 32-bit message has 4 iterations, each corresponding to each byte of the message

output undergoes immediate XOR operation with CRC after left shifting by 1B. The resultant output is the new CRC which is used in the next iteration.

In the 2nd iteration, we start with the $CRC_{B3\text{-}1}$ computed in the previous step which undergoes immediate XOR operation with the 2nd byte of the message. The rest of the process remains the same as discussed before. Similarly, iteration 3 and 4 are carried out as shown in Fig. 6.16. After the 4th iteration, the CRC undergoes another immediate XOR (32-bit) with CRC_{INIT} which results in the final 32-bit CRC value for the 32-bit message.

Along with the CiM of CRC, we also achieve parallelism with computation of multiple CRCs at once for messages stored in the same row. This appealing feature enables the proposed IPS-CiM in achieving significant performance improvement over traditional IPS architectures which have limited processing elements in MCU.

(b) Edge-sensing (SENSE): Edge-sensing in the IoT era is mainly performed by WSNs. They possess some key properties such as ease of deployment and self-organization. In the following, we discuss the implementation of the proposed CiM fabric for efficient edge-sensing with key tasks after aggregation of data being evaluation of average, maximum, minimum and variance. A small fraction of operations needed for SENSE is performed in the processor as discussed later.

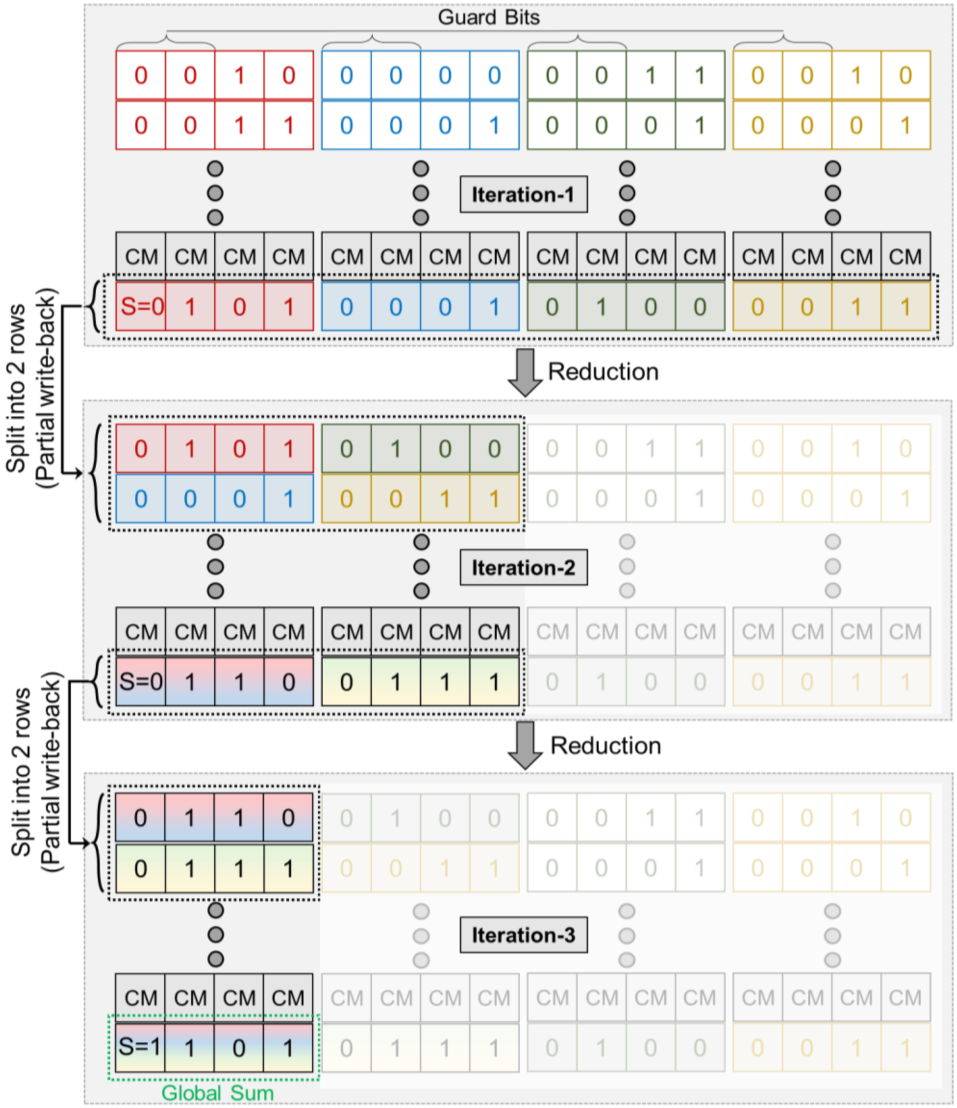


Fig. 6.17. Example of CiM flow for performing global sum operation which is required for average and variance computation.

(i) For computing the average of aggregated sensed data stored in the memory, we first perform a global sum operation. Fig. 6.17 illustrates an example for computing sum of 8 words (each of 2-bit length). Note, we consider 2 bits as guard bits for accumulations in this example (Fig. 6.17), which can be extended depending on the word length and number data points being computed on. We consider 4 pairs of words and perform four parallel in-memory additions between 2 words of each pair by activating the corresponding RWLs (as discussed earlier). The computed sum of the 4 different additions is re-arranged within the memory array as shown in Fig. 6.17. This rearrangement can be performed at overlapping addresses of the initial data stored or at a different address location depending on the requirement of the re-usability of data. After this, in-memory addition is again performed between 2 words in the same column and the results are stored back into the array for the final addition to achieve the global sum of 8 words as illustrated in Fig. 6.17. In this example, since the number of words being added together are 8, we have $\log_2(8)=3$ iterations. If N words are being added together, we will have $\log_2(N)$ iterations, considering the memory allows for storing N/2 pair of words in a row. If not, then we pipeline the in-memory addition operations. After global sum is evaluated, we divide by the number of input data points to find the average. This is performed in-memory with truncation by right shifting the global sum by $\log_2(N)$ to obtain the average. For this work, we consider a fixed sample space of N=256 ($\log_2(N)=8$) and therefore, we perform a fixed shifting by considering the most significant NW-8 bits as output (quotient), where NW is word-length.

(ii) For finding maximum of several data points from the sensor, we utilize the proposed computing fabric to perform in-memory evaluation of comparison operation in 2 cycles. We consider an example of 8 words (4-bit each) to discuss the operation as shown in Fig. 6.18. Note, we consider four pairs of words and perform four comparisons between each pair in parallel (see Fig. 6.18). Recall that, in the first cycle, we start with finding the XOR and its left most set bit (bit = '1') for the two words being compared. After finding the left most set bit, the corresponding bit location of word in row-1 is read-out in the second cycle. If the read-out=1 then the word in row-1 is maximum else word in row-2 is maximum. In the next step, the maximum values found in the previous iteration are re-written into the array to perform another iteration. This is followed by the final

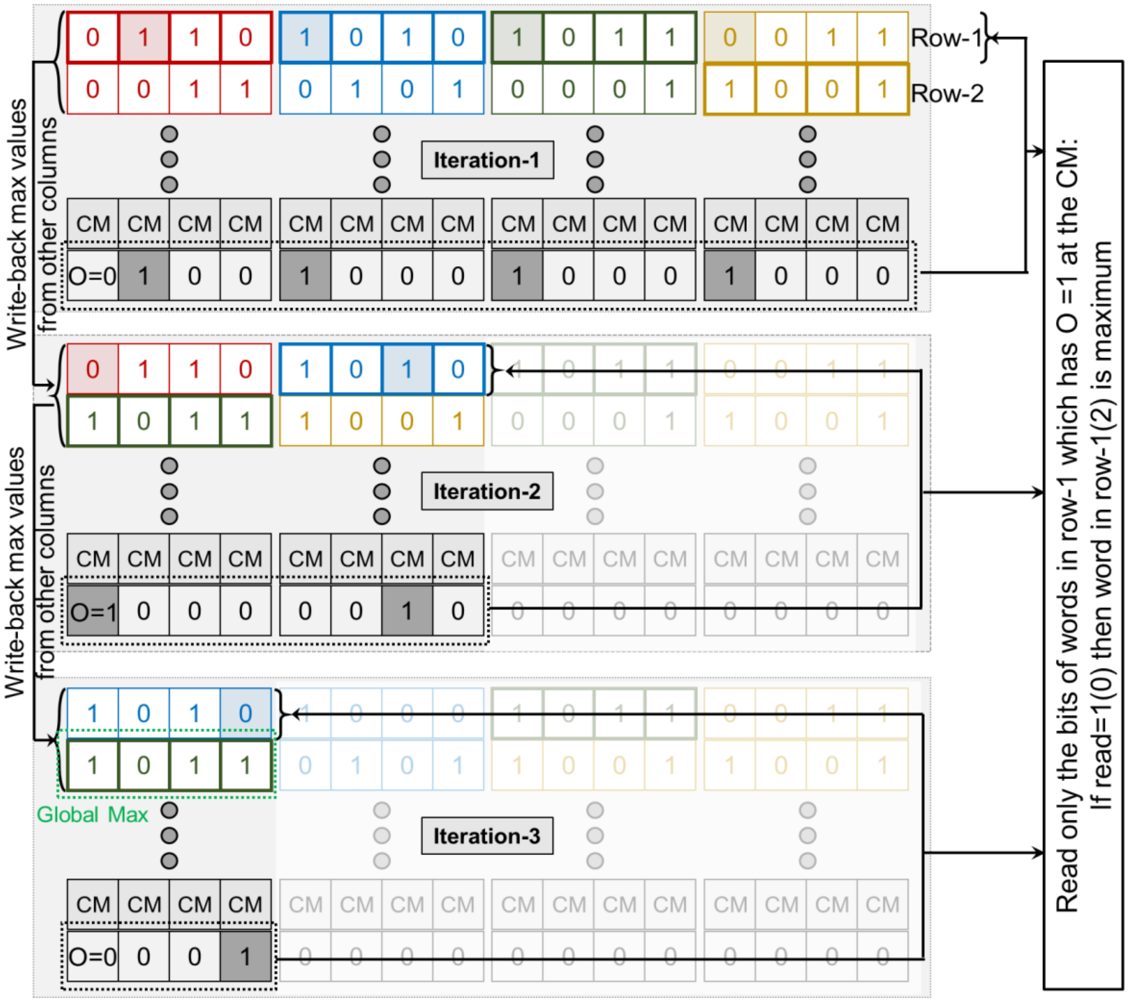

Row-1
Row-2
Iteration-1
Iteration-2
Iteration-3
CM
O=0
O=1
Global Max
Write-back max values from other columns
Write-back max values from other columns
Read only the bits of words in row-1 which has O =1 at the CM:
If read=1(0) then word in row-1(2) is maximum

Fig. 6.18. Example of CiM flow for performing global maximum operation

iteration to produce global maximum of the 8 words. For finding minimum, words with read-out = 0 are written back into the array during the iterative process. Similar to average, number of iterations for maximum/minimum evaluation depends on memory dimension and sample space, as discussed before.

(iii) For evaluating the variance, a combination of in-memory and in-processor operations is used. We first subtract the average (found previously) from the data points in the sample space, using in-memory subtraction discussed earlier. Note, we perform this efficiently by inverting the average in the array only once and duplicating/storing this value in multiple words of a row to enhance parallelism during subtractions. Following this, the results are sent to processor to perform square operations and division with the number of elements in sample space to achieve the final variance.

It is important to mention that IPS powered by harvested energy can have a sporadic power supply. This requires checkpointing operations to be performed in between a computation in order to have minimal loss in computation progress. Our proposed CiM designs can seamlessly implement this feature. For example, in between the iterations being performed during CRC or SENSE, if there is a power outage detected, we check-point the intermediate states to NVM and recall them when the power supply is restored to carry on with the computations without any loss in progress [197].

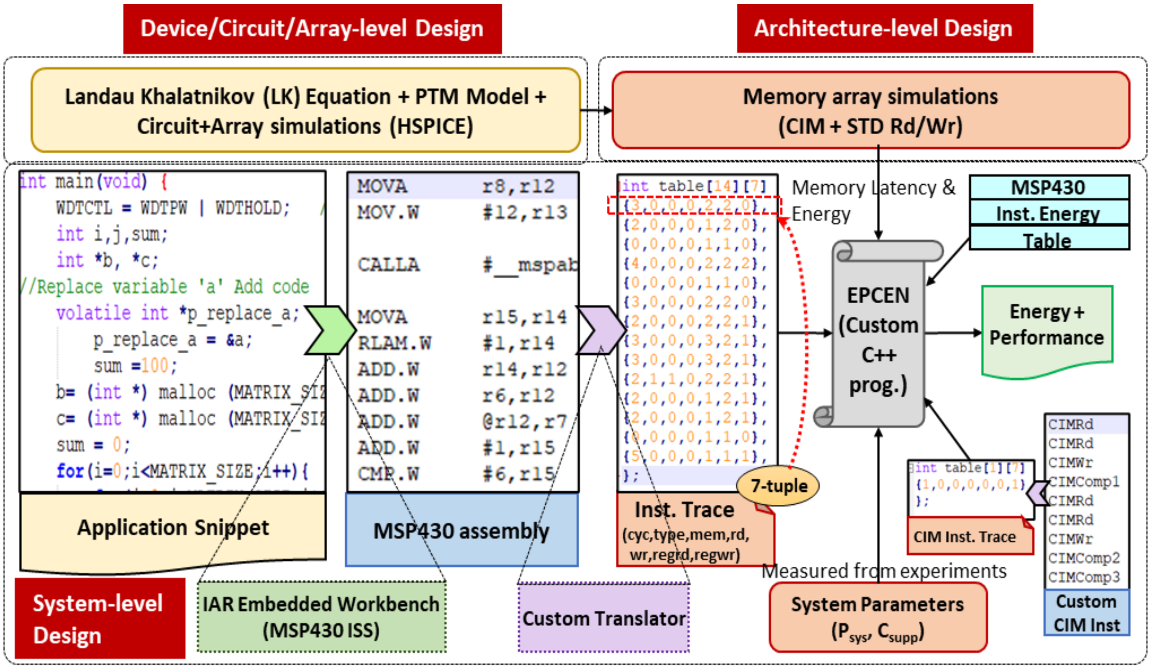
6.3.5 Evaluation

(a) Simulation framework:

(i) Device-Circuit-Array: For simulating the FEFET device, as discussed in Chapter-3, we employ a SPICE-based circuit compatible model based on time-dependent Landau Khalatnikov (LK) equation which is self-consistently coupled with an underlying transistor based on 45-nm technology PTM [108]. Thickness of FE used is 15nm and the experimentally calibrated LK parameters are α=-0.7x10^9m/F; β=6x10^8 $m^5/F/C^2$; γ=3x10^{11} $m^9/F/C^4$ [171]. We design the 3T NVM cell with minimum-sized FEFETs and standard access FETs for high density. We perform circuit and array evaluation in SPICE for read/ write and CiM operations. We implement an 32kB unified memory array with a line size of 16B. V_{DD} used is 1V and the bit-lines are kept precharged to V_{DD} in the idle/hold state to minimize the bit-line charging energy during read/write/compute.

(ii) System: For the system-level evaluations discussed here, we collaborated with Prof. Vijay Raghunathan and Dr. Arnab Raha to understand the implications of the IPS specific CiM techniques proposed in this section, in the context of a state-of-the-art intermittently powered platform. We employ the TI MSP430-based MCU architecture (which was also used in Chapter-5) as the baseline system due to its extreme low power consumption and several features suitable for IPS [188], [197], [198]. In this work, we implement and use a modified version of the IPS energy simulation framework that has been previously been used in [188], [198]. The overall flow diagram is depicted in Fig. 6.19.

Note that the memory organization and their sizes (SRAM and NVM) for the system are selected empirically based on currently available COTS MSP430 MCU modules.


Device/Circuit/Array-level Design
Architecture-level Design
Landau Khalatnikov (LK) Equation + PTM Model + Circuit+Array simulations (HSPICE)
Memory array simulations (CIM + STD Rd/Wr)
Memory Latency & Energy
MSP430 Inst. Energy Table
EPCEN (Custom C++ prog.)
Energy + Performance
7-tuple
Inst. Trace (cyc,type,mem,rd, wr,regrd,regwr)
CIM Inst. Trace
Measured from experiments
System Parameters (P_{sys}, C_{supp})
Custom CIM Inst
Application Snippet
MSP430 assembly
System-level Design
IAR Embedded Workbench (MSP430 ISS)
Custom Translator

```
int main(void) {
    WDTCTL = WDTPW | WDTHOLD;   /
    int i,j,sum;
    int *b, *c;
//Replace variable 'a' Add code
    volatile int *p_replace_a;
        p_replace_a = &a;
        sum =100;
    b= (int *) malloc (MATRIX_SIZ
    c= (int *) malloc (MATRIX_SIZ
    sum = 0;
    for(i=0;i<MATRIX_SIZE;i++){
```

```
MOVA      r8,r12
MOV.W     #12,r13

CALLA     #__mspab

MOVA      r15,r14
RLAM.W    #1,r14
ADD.W     r14,r12
ADD.W     r6,r12
ADD.W     @r12,r7
ADD.W     #1,r15
CMP.W     #6,r15
```

```
int table[14][7]
{3,0,0,0,2,2,0},
{2,0,0,0,1,2,0},
{0,0,0,0,1,1,0},
{4,0,0,0,2,2,2},
{0,0,0,0,1,1,0},
{3,0,0,0,2,2,0},
{2,0,0,0,2,2,1},
{3,0,0,0,3,2,1},
{3,0,0,0,3,2,1},
{2,1,1,0,2,2,1},
{2,0,0,0,1,2,1},
{2,0,0,0,1,2,1},
{0,0,0,0,1,1,0},
{5,0,0,0,1,1,1},
};
```

```
int table[1][7]
{1,0,0,0,0,0,1}
};
```

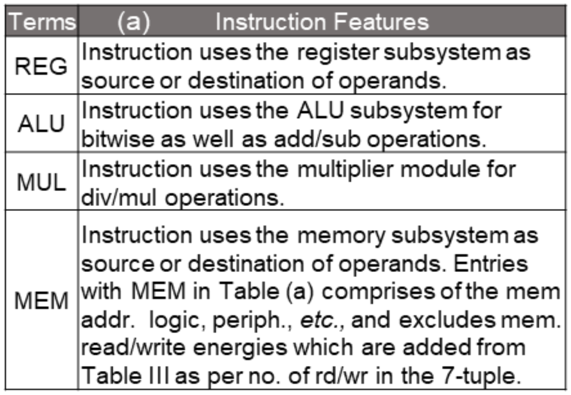
```
CIMRd
CIMRd
CIMWr
CIMComp1
CIMRd
CIMRd
CIMWr
CIMComp2
CIMComp3
```

Fig. 6.19 IPS-CiM Energy and Performance Simulation Framework

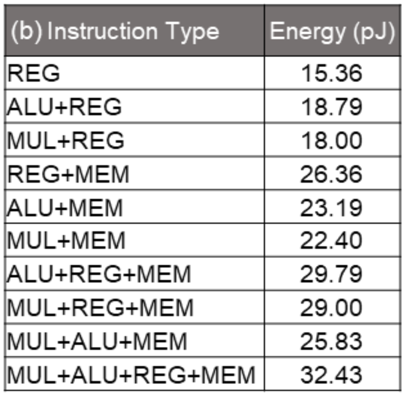
Without any loss in generality, we selected two modes of memory mapping for our experiments [188], [197]: (i) 8KB of RAM and 32KB of FEFET-based NVM, and (ii) a 32KB Unified FEFET-based NVM for our evaluation. Similar to the previous works [188], [222], for the conventional SRAM+NVM mode, we map the various program sections (data (heap), bss, stack) of the benchmarks to the SRAM while for the unified NVM mode, all parts of the program (data, bss, stack) are mapped to the NVM. Evidently, whenever a checkpoint operation is triggered due to power loss, an explicit checkpoint is triggered for the SRAM+NVM case that requires a large number of cycles and energy consumption to perform data migration from the SRAM to the NVM. On the other hand, the unified NVM mode performs in situ checkpointing with almost negligible energy overhead compared to SRAM+NVM due to the absence of data migration from the volatile SRAM (although there exists the need to migrate minimal processor states) [188], [197], [198]. On the contrary, the advantages of SRAM+NVM system are low power writes as well as low read/write latencies. Note, depending on the application characteristics, either of these modes can outperform the other [188], and hence, we present our results with both baselines.

- *Energy and performance computation engine (EPCEN):* For energy estimation, we first synthesized the TI MSP430 microcontroller core RTL (soft IP core obtained from

Table. 6.2 (a) Description of terms used in Table (b). (b): Instr. feature-type energy breakdown from MPS430 synthesis at 45 nm. Note, FEFET-NVM atomic read/write energies are excluded and are to be added separately.



Terms	(a) Instruction Features
REG	Instruction uses the register subsystem as source or destination of operands.
ALU	Instruction uses the ALU subsystem for bitwise as well as add/sub operations.
MUL	Instruction uses the multiplier module for div/mul operations.
MEM	Instruction uses the memory subsystem as source or destination of operands. Entries with MEM in Table (a) comprises of the mem addr. logic, periph., *etc.*, and excludes mem. read/write energies which are added from Table III as per no. of rd/wr in the 7-tuple.



(b) Instruction Type	Energy (pJ)
REG	15.36
ALU+REG	18.79
MUL+REG	18.00
REG+MEM	26.36
ALU+MEM	23.19
MUL+MEM	22.40
ALU+REG+MEM	29.79
MUL+REG+MEM	29.00
MUL+ALU+MEM	25.83
MUL+ALU+REG+MEM	32.43

OpenMSP430 [199]) using Synopsys Design Compiler and then mapped the design to 45 nm based on the Nangate OpenCell library. We use Synopsys Power Compiler for generating the power numbers corresponding to different types of instructions (except memory access energy) as shown in Table. 6.2. For memory instructions, the memory operation energies derived from the circuit simulation are added to the base energies in Table. 6.2. Note that different types of instructions consume different amounts of energy as they use different elements of the MCU core (such as the adder for ADD/SUB operations, multiplier for MUL/DIV operations, a relatively simpler ALU core for bitwise operations, etc.). In addition, the number of checkpointing (and restore) operations, which is a function of the system power and the available energy to the system, also contribute to the final energy consumption of IPS. The system energy consumption mainly consists of the power of the MCU execution core and associated peripheral energies. For our analysis, the supply capacitance, C_{SUPP}, is selected in the range of 10nF - 1μF that supply energies in the range of 4.2×10^{-7}J to 4.2×10^{-9}J per power cycle to the system as given by $(1/2)*C_{SUPP}*(V_{ON}^2 - V_{OFF}^2)$, where V_{ON} and V_{OFF} are 2.2V and 2.0V, respectively corresponding to a TI MSP430FR5739 based microcontroller. We assume that the frequency of operation is 25 MHz. Note that the C_{SUPP} for an IPS is generally set at a value that can provide sufficient energy to ensure significant forward progress across a variety of workloads while incurring a small area footprint. For our experimental system and set of benchmarks in this work, C_{SUPP} of 10nF - 1μF proved to be adequate. The MCU energy consumption for each instruction type along with the memory read and write energies for SRAM and FEFET-

NVM and their corresponding latency numbers are fed into a custom IPS energy and performance computation engine (EPCEN) which is a C++-based program that uses assembly instruction traces in the form of septuple (7-tuples) of the target application for generating the overall system-level energy consumption. We use IAR Embedded Workbench-based MSP430 Instruction Set Simulator (ISS) to generate the assembly-level instruction traces for an application. As part of the EPCEN framework, we also built a custom assembly instruction-level translator that identifies the type of each assembly instruction and subsequently, translates each of them to a septuple (7-tuple) having the format (cycles, type, mem, rd, wr, regrd, regwr). Here, cycles denote the total number of cycles required to execute an instruction, mem denotes whether an instruction accesses memory or not. The fields rd and wr denote whether the mem operation is a memory read or write operation, respectively, or both. Similarly, regrd and regwr show whether MCU general purpose and special registers are used as source or destination, respectively, or both. Finally, type defines the type of ALU operation such as bitwise OR, AND, XOR, NOT, ADD, SUB, MUL, etc. Note that except MUL, all other operations fall under the type ALU in Table. 6.2. Demarcation of these features are essential in calculating the correct energy consumption of the instructions for the base-IPS (Table. 6.2) and the IPS-CiM systems. Note that EPCEN also takes into the account various system parameters (C_{SUPP}, V_{ON}, V_{OFF}, f_{MCU}) to evaluate available energy per power cycle and simulates the IPS behavior to compute the total number of checkpoints and the energy associated with them based on the specified memory mapping used in the IPS.

In order to calculate the energy consumption for the IPS-CiM, we use the same EPCEN albeit with a modified instruction trace. For CiM, the instruction trace for the application consists of sequences of CiM-specific instructions such as CiMRd, CiMWr, CiMComp1, CiMComp2 that are used to refer to the CiM reads, writes, and compute operations, respectively. Note that there are multiple CiMComp instructions because applications can use different types of CiM computations (such as single-step addition and two-step comparison) consuming different amounts of energy. We obtain the array-level energy and latency values for each of them from circuit-level analysis and feed them to the EPCEN. Since these CiM instructions are highly application-specific, IPS-CiM requires the existing compiler to be augmented manually with these new instructions based on the applications executed using CiM. Fig. 6.19 shows a snippet of the new CiM-based instruction trace that is constructed manually. The EPCEN outputs both the total energy and total cycles required to execute target application.

(iii)Data Mapping: For IPS-CiM, the MCU has an additional function of organizing the input operands in the memory in a format suitable for the CiM to operate. This is achieved either after the inputs are acquired through a sensor or are aggregated and stored in the memory. Now, since the process of generating sequence of addresses and the storage of input operands needs to be performed even for our baseline designs, the only additional requirement for IPS-CiM is performing this in a more sequential order (such as writing parallel inputs in a single row). And our analysis indicates that this leads to minimal energy overhead compared to the baseline cases. Moreover, such a data mapping approach is a common requirement for most CiM architectures [59], [60], [215]. Note that, this data mapping is accomplished by having a pre-determined address map for storing the input values which is fixed for a specific application and is stored as part of the application program.

(b) Results:

(i) Array-level: The write, read and compute energy(delay) of the proposed 3T FEFET NVM CiM architecture are 15.95pJ (3.41ns), 2.6pJ (2.06ns) and 4.36pJ (2.48ns) respectively. All the energy and delays are calculated for 16B operation. The compute energy/delay is calculated for ADD/Boolean operations. For the evaluation of the energy/delay of other complex CiM operations such as subtraction, comparison etc., the computations are mapped to a sequence of ADD and Boolean operations as per the discussions above, which are fed into the IPS framework to account for compute energy during each iteration (if any) and accumulate them to calculate the total energy consumption. The above-mentioned results are integrated with IPS-CiM for system-level evaluation, next.

(ii) System-level: Here, we report the system-level energy and performance improvements of using IPS-CiM compared to the two baseline-IPS cases that uses (i) SRAM + FEFET NVM and (ii) Unified FEFET-NVM memory organizations. Fig. 6.20 depicts the normalized energy consumption (a-b), speedup (c-d), and the number of checkpoints (e-f) shown by IPS-CiM over the baseline IPS systems for CRC and SENSE applications, respectively. Fig. 6.20 (a) and (b) show energy improvements for the IPS-CiM in the range of 400X – 3275X

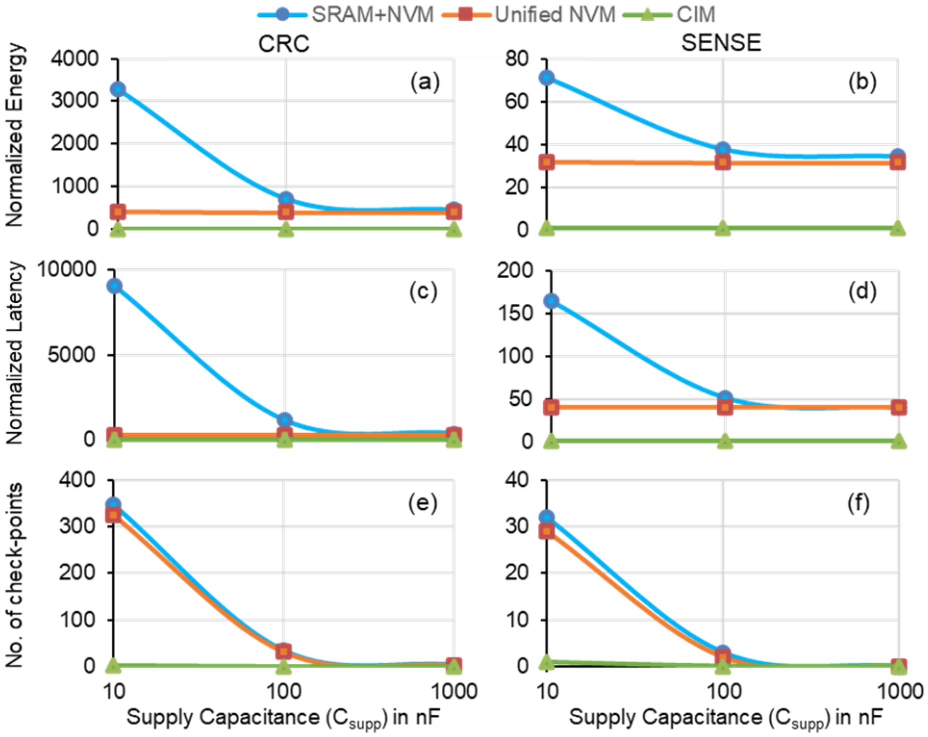


Fig. 6.20. (a and b) Normalized system energy consumption, (c and d) Speedup achieved by IPS-CiM, (e and f) Number of checkpoints for SRAM+NVM, Unified NVM, and IPS-CiM systems for CRC and SENSE applications respectively.

for CRC application and 32X-71X for SENSE application corresponding to supply capacitances (C_{SUPP}) in the range of 10nF - 1µF. The improvements are attributed to (i) the reduction in data movement to-and-fro between memory and processor and (ii) massive parallelism enabled by IPS-CiM. The energy savings achieved by the proposed IPS-CiM is higher with SRAM+NVM baseline (450X-3275X for CRC and 35-71X for SENSE) compared to the Unified NVM baseline (400X for CRC and 32X for SENSE) as the former performs explicit checkpointing that incurs significantly higher checkpointing overhead than the Unified NVM case which performs in situ checkpointing [188], [197], [198]. Note that the energy savings of IPS-CiM over the SRAM+NVM baseline system increases for smaller value of supply capacitances as it results in larger number of checkpoints due to lesser energy delivered per power-ON period ($0.5*C_{SUPP}*(V_{ON}^2 – V_{OFF}^2)$). The increase in number of checkpoints leads to higher energy consumption for migrating data from volatile SRAM to NVM and vice versa. It also results in large number of active MCU cycles as well as memory access energies. Note that the energy savings over the Unified NVM case

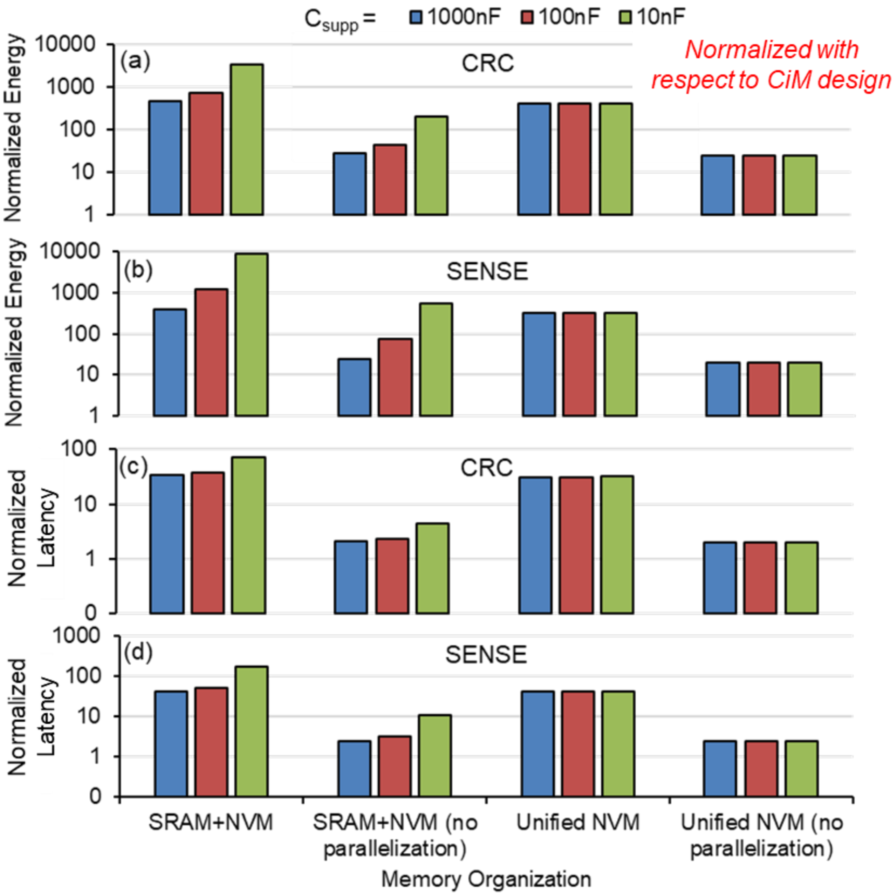


Fig. 6.21 (a, b) Normalized energy consumption and (c, d) Speedup for SRAM+NVM, Unified NVM, and their non-parallel versions for (a, c) CRC and (b, d) SENSE applications. All plots are in logarithmic scale.

remains constant (independent of C_{SUPP}) as it performs implicit (in situ) checkpointing within the NVM incurring negligible checkpointing overhead due to the absence of data migration between memories [188], [197], [198]. The higher energy consumption in both the baselines results in a higher number of checkpoints as shown in Fig. 6.20 (d) and (e) compared to the IPS-CiM, inhibiting forward progress in computation.

We also evaluate speedup, which is defined as the ratio of the total number of clock cycles used by the baseline IPS to the total number of clock cycles required by the IPS-CiM to execute the same application. Fig. 6.20 (c, d) show that for CRC and SENSE, the speedups obtained are in the range of 325X- 9100X and 40X-165X, respectively, for C_{SUPP}=10nF-1µF. Note that higher the number of checkpoints, larger is the number of execution cycles for the baseline systems resulting in higher speedup for IPS-CiM. SENSE results in relatively lesser energy and performance benefits compared to CRC as SENSE uses the energy-intensive MUL/DIV operator of the MCU in addition to CiM while calculating sensor value statistics. On the other hand, CRC is much simpler and only uses bitwise and arithmetic operations such as XOR, ADD, SUB, etc., which are performed in-memory in the proposed IPS-CiM.

An inherent reason of achieving order of magnitude higher energy/performance benefits in the proposed IPS-CiM is due to the parallelization of computations in the memory array. For both CRC and SENSE, IPS-CiM processes 16 and 32 input words (16B line size in the NVM array with words of 32-bit each) simultaneously compared to just a single compute operation per clock cycle achievable by the single threaded single MCU execution core. In order to isolate the benefits of having multiple processing elements working simultaneously from other algorithmic and architecture benefits arising due to IPS-CiM, Fig. 6.21 shows the energy and performance improvements for IPS executing only CRC and SENSE operations with and without parallel processing. Fig. 6.21 reports energy and performance improvements in the range of 2X-205X and 2X-568X without parallel processing for CRC and SENSE applications. It is evident that while a portion of the benefit is attributed to parallelism, elimination of data movement between memory and processor also contributes to the energy savings.

6.4 Summary

In this chapter, we proposed two compute-in-memory architectures for general purpose and targeted application workloads. First, we proposed 4T R-FEFET non-volatile memory (4T-R) with differential read, energy-efficient write as well as in-memory computation support of performing Boolean and arithmetic logic functions using a low power and compact compute module (CM). Based on an extensive array analysis, we quantified the performance and energy benefits of the proposed R-FEFET-CiM (over existing FEFET-CiM) arising due to (i) uni-polar voltage design, (ii) differential sensing and (iii) low power CM. We carried out system-level analysis of the R-FEFET-CiM using an Intel Nios II processor-based system for various application benchmarks and showed up to 27% energy savings can be achieved when compared to FEFET-CiM baseline. For addressing the requirement of CiM techniques targeted for IPS specific workloads, we proposed a 3T compute-enabled FEFET NVM featuring voltage-based operations for both data reading and computing, making it highly energy efficient. We showed how various Boolean and arithmetic in-memory operation can be performed along with a novel technique for energy-efficient in-memory comparison operation with minimal hardware overheads. Utilizing the proposed FEFET based CiM engine designed for IPS, we proposed an architecture to perform two major transient computing workloads: (a) cyclic redundancy check (CRC) for error detection and (b) edge-sensing (SENSE) in wireless sensory networks. We integrated the proposed CiM engine with IPS to build the IPS-CiM architecture. We evaluated the energy and performance of the proposed IPS-CiM and compare it to two baselines: (a) hybrid SRAM+NVM and (b) unified NVM architectures, both of which perform out-of-memory computing. IPS-CiM system showed excellent energy savings in the range of 400X-3275X and 32X-71X with respect to the baseline systems for CRC and SENSE applications, respectively.

7. FEFET BASED ARTIFICIAL INTELLIGENCE HARDWARE FOR TERNARY PERICION COMPUTING-IN-MEMORY TO ACCELERATE DEEP NEURAL NETWORKS

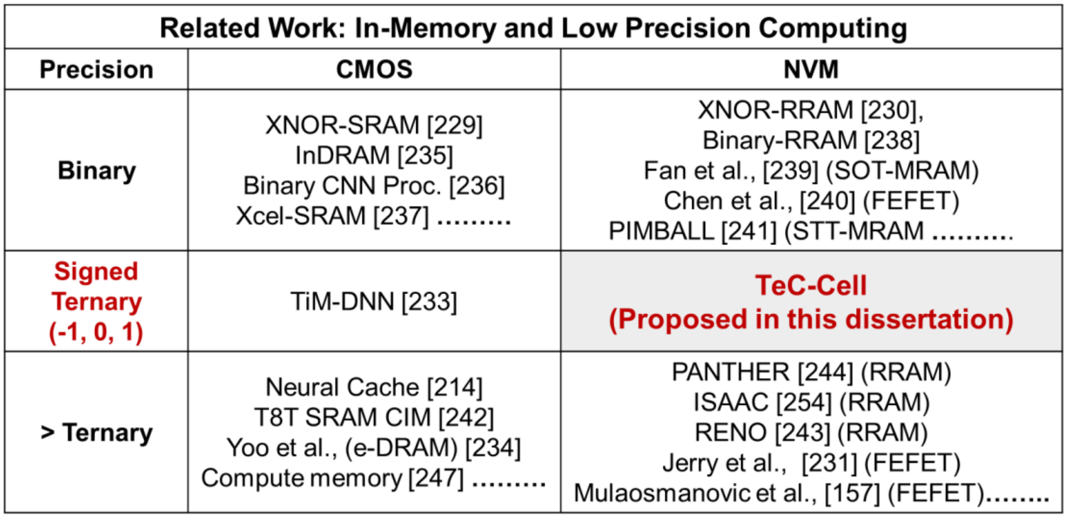
7.1 Introduction

Deep Neural Networks (DNNs) have gained immense popularity in recent years due to their ability to achieve remarkable accuracies in a wide range of cognitive tasks [223]. However, the high computation and storage demands pose key challenges to their ubiquitous adoption. An important scenario that exemplifies this challenge is low-power inference, wherein DNN models are executed on deeply embedded IoT devices and wearables that have severe energy and area constraints [224].

To deploy DNNs on cost-constrained systems, low-precision is of great interest as it lowers all aspects of energy usage, viz., compute, interconnect, and memory. Recent studies suggest that ternary precision networks are especially promising as they offer accuracy significantly higher than binary networks but with a moderate degradation compared to full precision networks [225], [226]. Ternary networks drastically reduce the complexity of matrix multiplication which constitutes >90% of DNN computations, thereby facilitating reductions in computation time and energy. In this work, we explore the design of efficient hardware for ternary DNNs.

Traditional CPUs, GPUs and specialized DNN accelerators suffer from frequent memory accesses, limiting their energy efficiency and performance [227]. To address this issue, various works have proposed in-memory computing, wherein computations are performed within the memory array, eliminating the memory access overheads associated with traditional von-Neumann architectures [66], [68], [157], [214], [228]–[247] (Table. 7.1). Most existing designs perform in-memory multiplication of binary operands [228]–[230], binary activations with ternary weights [229], [234], or target higher-than-ternary precisions for analog vector-matrix multiplication [157], [248]. Recently, a CMOS based ternary in-memory DNN (TiM-DNN) architecture was proposed for pure signed ternary computation (ternary inputs and weights: '-1', '0', '+1') [233]. Such an approach enables massively parallel signed ternary vector-matrix multiplications in a single array access, for efficient realization of ternary DNNs.

Table. 7.1 Related work exploring the synergy between in-memory and low-precision computing



Related Work: In-Memory and Low Precision Computing		
Precision	**CMOS**	**NVM**
Binary	XNOR-SRAM [229] InDRAM [235] Binary CNN Proc. [236] Xcel-SRAM [237]	XNOR-RRAM [230], Binary-RRAM [238] Fan et al., [239] (SOT-MRAM) Chen et al., [240] (FEFET) PIMBALL [241] (STT-MRAM ....
Signed Ternary (-1, 0, 1)	TiM-DNN [233]	**TeC-Cell (Proposed in this dissertation)**
> Ternary	Neural Cache [214] T8T SRAM CIM [242] Yoo et al., (e-DRAM) [234] Compute memory [247]	PANTHER [244] (RRAM) ISAAC [254] (RRAM) RENO [243] (RRAM) Jerry et al., [231] (FEFET) Mulaosmanovic et al., [157] (FEFET)........

Although CMOS-based in-memory computing designs are promising for achieving energy and performance improvements compared to traditional CPU/GPU architectures, they face some major drawbacks. For instance, in 6T SRAMs, coupling of read-write paths may lead to cell disturbances during computations with multi-word-line assertion [66], [214]. Moreover, static leakage due to technology scaling offsets the efficiency gain achieved during in-memory compute operations [249]. Lastly, large bit-cell area limits their on-chip capacity and in-memory computation bandwidth. Emerging non-volatile memories (NVMs) such as spin-transfer-torque magnetic RAM (STT-MRAM), Resistive RAMs (RRAMs) and FEFETs have showcased great potential to replace or complement CMOS based memories by overcoming their drawbacks. FEFETs, in particular, are extremely promising due to their electric-field-driven low-power write operation compared to current-driven write in STT-MRAMs and RRAMs [250]. These desirable properties have driven recent interest towards in-memory computing with NVM. However, to the best of our knowledge, ternary in-memory computation using any emerging NVM has not been previously explored.

In this work, we propose a non-volatile ternary compute-enabled memory cell (TeC-Cell) that can perform massively parallel in-memory matrix multiplication in the signed ternary regime. The proposed TeC-Cell is designed by utilizing CMOS compatible FEFETs coupled with a judicious selection of input, weight and output encodings, which enable a compact cell design compared previous SRAM-based in-memory computing designs.

7.2 Ternary Precision Networks

Ternary networks have emerged as an attractive option in the quest for low-precision DNNs. However, the performance and energy efficiency of near-memory accelerators for ternary networks are bottlenecked by the on-chip memory due to the sequential row-by-row access. The closest prior efforts on in-memory computing involve dot product computation of either ternary inputs with binary weights [229] or vice-versa [234]. Although these are attractive design choices to achieve improved energy efficiency, a pure ternary network with signed ternary weights and inputs ('-1', '0', '+1') can achieve substantially better accuracy compared to the binary networks [225], [226]. Furthermore, techniques presented in [229], [234] can only enable simultaneous activation of a limited numbers of rows due to sensing constraints. This limits the parallelism achieved in vector-matrix multiplication. A recent CMOS-based design, TiM-DNN [233] overcomes such limitations by performing massively parallel in-memory dot product computations in the signed ternary regime.

This chapter proposes a novel compute-enabled ternary cell (TeC-Cell) using emerging NVM based on FEFETs, featuring non-volatility, higher integration density and near-zero stand-by leakage power compared to SRAMs. Notably, the input, weight and output encodings that we propose here enable in-memory dot product computation of weight and input vectors with the addition of just two more transistors to a pair of FEFET NVM cells. The compact TeC-Cell design enables a higher degree of parallelism for CiM compared to other memories at iso-area (as discussed later). The built-in non-volatility of TeC-Cell, along with its low power operation can potentially enable energy-efficient realization of DNNs for edge computing devices such as IoT sensors. Note that our design technique is not limited to FEFETs but can also be applied to other memories with separate read-write paths (such as Spin Orbit Torque MRAMs (SOT-MRAMs) [251], eDRAMs [234], etc.), to enable ternary in-memory computation.

7.3 FEFET based Ternary Compute Enabled Memory

7.3.1 Memory Design

To enable ternary in-memory computation, we propose a non-volatile ternary cell (TeC-Cell) which consists of 2 FEFETs and 6 standard FETs. The schematic and layout are shown in Fig. 7.1.

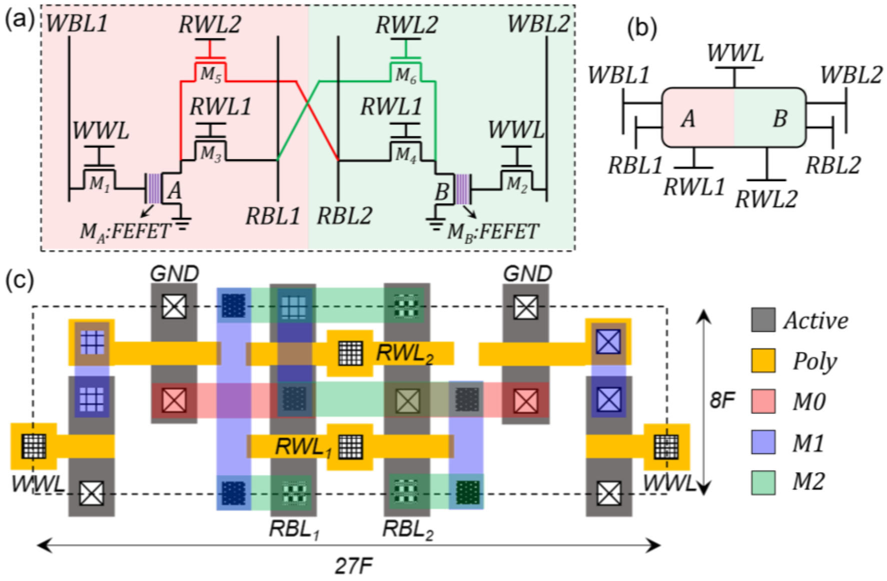


Fig. 7.1. (a) Circuit design, (b) schematic and (c) layout of ternary compute-enabled non-volatile memory cell (TeC-Cell)

The core of the TeC-Cell involves two 3T-FEFET based memories [133] (for ternary storage), which are cross-coupled with each other using just 2 additional transistors per cell (M_5 and M_6). Transistors M_1 and M_2 are the write access transistors, which enable selective writing of the data in the array as the polarization of the two FEFETs (P_A in MA and P_B in MB). M_3 and M_4 are the read access transistors, used to sense the data without disturbances from the unaccessed cells. Cross-coupled transistors M_5 and M_6 (along with M_3 and M_4) enable in-memory ternary scalar multiplication as discussed later. The proposed technique of designing a TeC-Cell can also be employed in other memories with separate read-write paths (such as SOT-MRAMs [251], eDRAMs [234] etc.), using just 2 additional cross-coupled transistors. In this work, we focus our discussion on FEFETs since they demonstrate appealing properties such as non-volatility, near-zero leakage energy and low-power write. Moreover, their high distinguishability (Fig. 7.1(b)) is particularly useful for robust in-memory computation, as explained later.

7.3.2 Ternary read-write operation

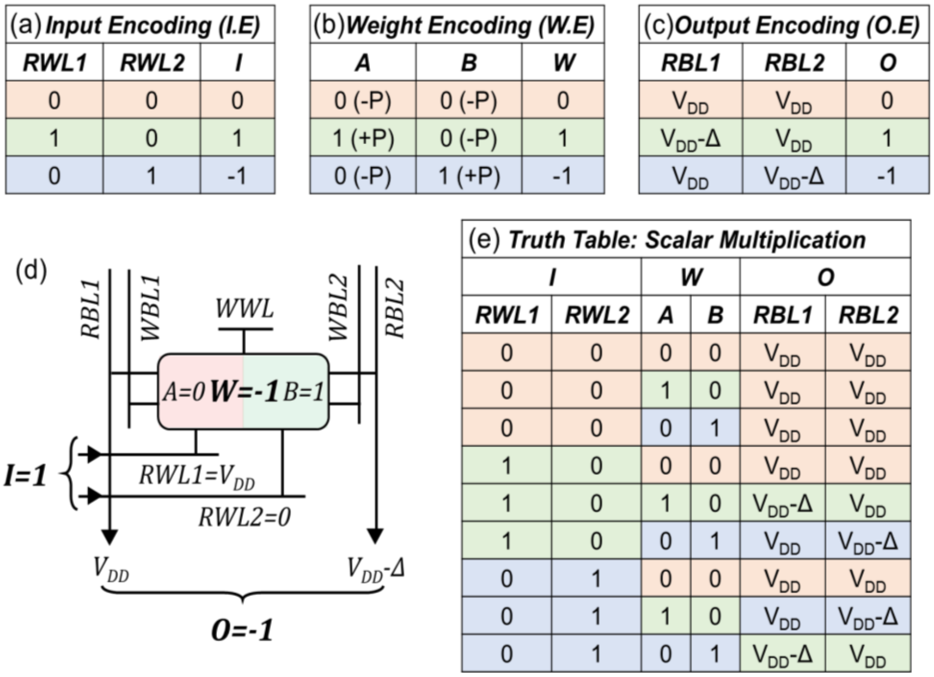
For storing ternary data (storage/weight encoding in Fig. 7.2 (b)), we assert the write word-line (WWL=V_{DD}=1V) and drive the write bit-lines (WBL1 and WBL2) to appropriate values. To write '+1' ('-1'), WBL1= V_{DD} (-V_{DD}) and WBL2= -V_{DD} (V_{DD}) is applied. This brings the polarization of the FEFETs to P_A= +P (-P) and P_B= -P (+P). To write ‘0’, WBL1 and WBL2 are driven to -V_{DD}, resulting in P_A=P_B=-P. After write, WWL is de-asserted with WBL1=WBL2=0V, resulting in storage of the bit-information in M_A and M_B as P_A and P_B in a non-volatile fashion. Note that, polarization stored in the FEFETs corresponds to its resistance states (+P: LRS; -P: HRS), which is used for the read operation as discussed next.

For sensing the bit stored, the read word-line (RWL1) is asserted with the read bit-lines (RBL1 and RBL2) pre-charged to V_{DD}. Now, based on the polarization stored, RBL1 and RBL2 will either discharge or remain at V_{DD}, due to high LRS and low HRS currents, respectively (Fig. 7.1). For the case when the bit stored is ‘+1’ (P_A=+P; P_B=-P), RBL1 discharges (M_A in LRS), while RBL2 remains at V_{DD} (M_B in HRS). The opposite occurs when the bit stored is ‘-1’ (P_A=-P; P_B=+P). When the TeC-Cell stores a ‘0’ (P_A=-P; P_B=-P), both RBL1 and RBL2 remain at V_{DD}. We use a voltage sense amplifier to compare the RBL1 and RBL2 voltages with a reference voltage (0.95V in our analysis). Note that, during read and write, RWL2 is always de-asserted. Additionally, the proposed TeC-Cell can also be used as a 2-bit binary memory where P_A and P_B correspond to independent bits, without any circuit modifications.

7.3.3 In-memory ternary scalar multiplication using TeC-Cell

In this section, we propose in-memory scalar multiplication of ternary weight (stored in the TeC-Cell) with ternary input to obtain a ternary output. Initially, the read bit-lines (RBL1 and RBL2) are pre-charged to V_{DD}. The ternary inputs are encoded as read word-line (RWL1 and RWL2) voltages as shown in Fig. 7.2(a). Depending on the ternary weight (encoded as P_A and P_B; see Fig. 7.2(b)), the final RBL1 and RBL2 voltages represent the multiplication output (output encoding in Fig. 7.2(c)). We explain this further with the following examples:

- When input I= +1 (RWL1=V_{DD}; RWL2=0) and weight W= -1 (A=0; B=1), transistors M_3, M_4, M_B are ON and M_5, M_6 and M_A OFF. This condition results in a discharge path for RBL2, resulting in a voltage drop of Δ=100mV (which is sensed with the sense amplifier), while RBL1 remains pre-charged at V_{DD}. This corresponds to output (O=I*W) = -1. Note that the output is inferred with single-ended sensing of RBL1 and RBL2 (see Fig. 7.2(d)). The same voltage conditions of RBL1 and RBL2 hold true for the case when I= -1 (RWL1=0; RWL2=1) and W= +1 (A=1; B=0) as shown in Fig. 7.2(d).



(a) *Input Encoding (I.E)*

RWL1	RWL2	I
0	0	0
1	0	1
0	1	-1

(b) *Weight Encoding (W.E)*

A	B	W
0 (-P)	0 (-P)	0
1 (+P)	0 (-P)	1
0 (-P)	1 (+P)	-1

(c) *Output Encoding (O.E)*

RBL1	RBL2	O
V_{DD}	V_{DD}	0
V_{DD}-Δ	V_{DD}	1
V_{DD}	V_{DD}-Δ	-1

(e) *Truth Table: Scalar Multiplication*

I		W		O	
RWL1	RWL2	A	B	RBL1	RBL2
0	0	0	0	V_{DD}	V_{DD}
0	0	1	0	V_{DD}	V_{DD}
0	0	0	1	V_{DD}	V_{DD}
1	0	0	0	V_{DD}	V_{DD}
1	0	1	0	V_{DD}-Δ	V_{DD}
1	0	0	1	V_{DD}	V_{DD}-Δ
0	1	0	0	V_{DD}	V_{DD}
0	1	1	0	V_{DD}	V_{DD}-Δ
0	1	0	1	V_{DD}-Δ	V_{DD}

Fig. 7.2 (a) Input, (b) Weight and (c) Output encoding for ternary compute operations. (d) Example of scalar multiplication in TeC-Cell. (e) Truth table for all permutation of I.E and W.E.

- When I= +1 (RWL1 =V_{DD}; RWL2 =0V) and W= +1 (A=1; B=0), transistors M_3, M_4, M_A are ON while M_5, M_6 and M_B remain OFF. This corresponds to a discharge path for RBL1 (resulting in Δ drop) with RBL2 remaining at its pre-charged voltage, V_{DD}. This voltage condition corresponds to O= I*W= +1. This condition also holds true when I= -1 and W= -1.
- When W or I=0, RBL1 and RBL2 remain pre-charged at V_{DD}, corresponding to O= I*W= 0.

The truth table for all permutations is shown in Fig. 7.2(e). Note that the proposed TeC-Cell exhibits isolation of read-write paths and therefore, in-memory scalar multiplication has no effect on the information stored as polarization in the FEFETs.

7.3.4 Ternary dot-product computation

We next discuss how TeC-Cells enable in-memory ternary dot product computation for vector-matrix multiplication. This is achieved by simultaneously asserting the read word-lines of TeC-Cells present in a single column as illustrated in Fig. 7.3(a) [233]. The weight vector with ternary elements Wi is stored in the TeC-Cells, while the input vector with elements Ii is encoded using the voltages of RWL1i and RWL2i (Fig. 7.2(a). With RBL1 and RBL2 (which are pre-charged to V_{DD}) connected to cells in the same column, the scalar products from each TeC-Cell (as discussed in previous sub-section) add up through cumulatively discharge of RBL1 and RBL2, resulting in a multiply-and-accumulate (MAC) operation. The final RBL1 (RBL2) voltages correspond to the number of TeC-Cells producing +1 (-1) as the scalar product. For example, if 'a' scalar multiplication produced an output of '+1' and 'b' scalar multiplications produced an output of '-1', then the final RBL1 and RBL2 voltages are V_{DD} - aΔ and V_{DD} - bΔ respectively. Flash analog-to-digital converters (ADCs) are employed to yield the digital value corresponding to 'a' and 'b'. The final dot product given by $\sum_{i=1}^{n} I_i * W_i = a - b$, is achieved by subtracting 'b' from 'a' using a digital CMOS subtractor. Fig. 7.3(b) illustrates the sensing circuit required to realize the final dot product computation.

It is important to mention that the sense margin reduces as 'a' or 'b' increase, due to the exponential nature of the bit-line capacitance discharging. For example, if 'a' or 'b' increase from

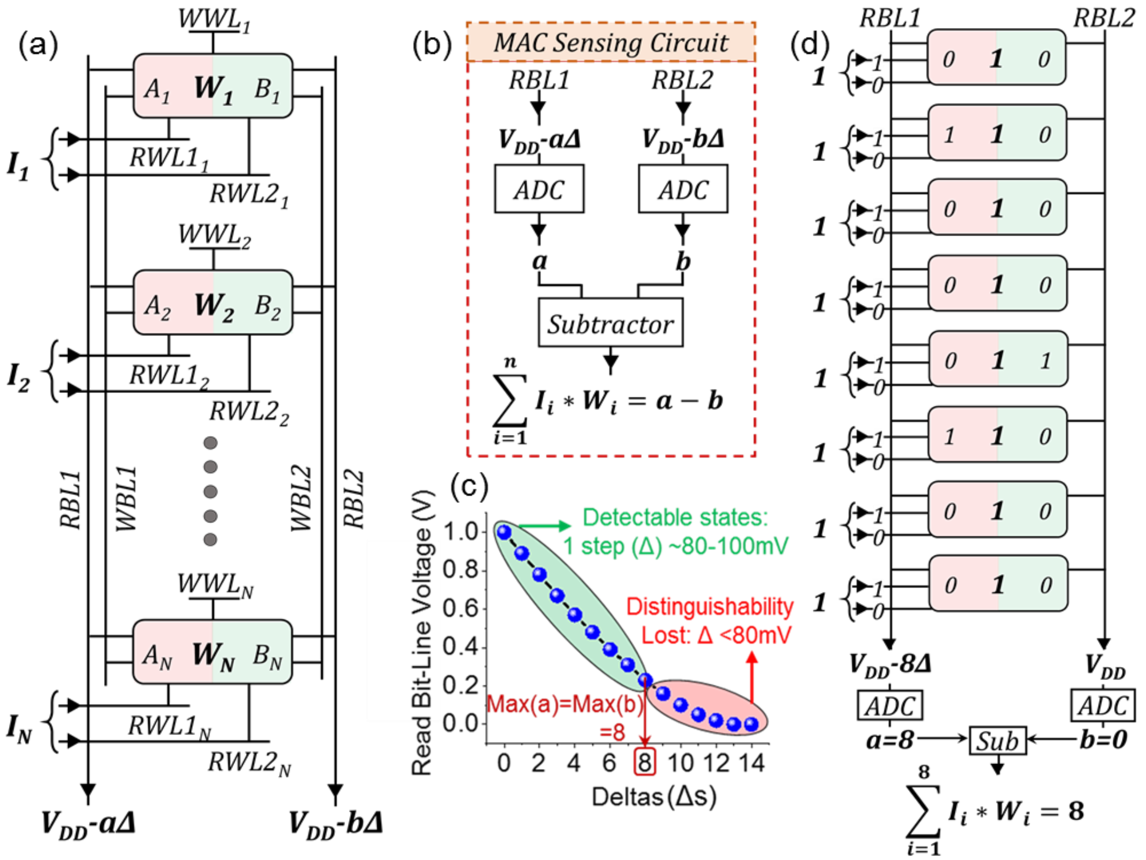


Fig. 7.3 (a) Ternary dot-product computation of input vector I and weight vector W (b) MAC sensing unit consisting of ADCs and subtractor. (c) RBL voltage vs number of Δ drops. (d) An example of worst-case input-weight scenario for sensing, resulting in *a*=8.

1 to 8, Δ reduces from 100mV to 80mV as shown in Fig. 7.3(c). This limits the number of cells that can be simultaneously activated during dot-product computation. Fig. 7.3(d) illustrates an example of a worst-case input-weight vector scenario for a stack of eight TeC-Cells. However, the statistics of the data, specifically the prevalence of zero values in weights and activations, also plays a role in determining the design choice, as discussed later. Before undertaking this discussion, we first analyze the implications of process variations on the degradation of sense margins in the next sub-section.

7.3.5 Variation analysis

We study the impact of transistor threshold voltage (V_{TH}) variation on the in-memory dot-product operation. We consider 6σ = 120mV for V_{TH} of all the transistors (where σ is the standard deviation). We perform Monte-Carlo SPICE simulations considering 1000 samples each, for cases ranging from 1Δ discharge to 8Δ discharge (states >8Δ are not considered since they are not sufficiently distinguishable). As the amount of discharge increases, the probability of sensing error also increases as shown in Fig. 7.4(a) (higher overlapping of RBL voltages between adjacent Δ states). However, it is also important to note that the probability of occurrence of the states decrease with increasing discharge values [233] (due to data statistics). The probability of an error in the dot-product is equal to the product of sensing error probability and the occurrence probability of a particular discharge state (number of Δs; #Δ). Fig. 7.4(b) illustrates the dot product error probability as a function of #Δ, exhibiting a non-monotonic behavior. Moreover, the total probability of error (P_T) during the dot-product operation is the sum of errors observed for each #Δ (Fig. 7.4(b)), is $3.10e^{-3}$. In other words, for every 1000 MAC operations we have ~3 errors with

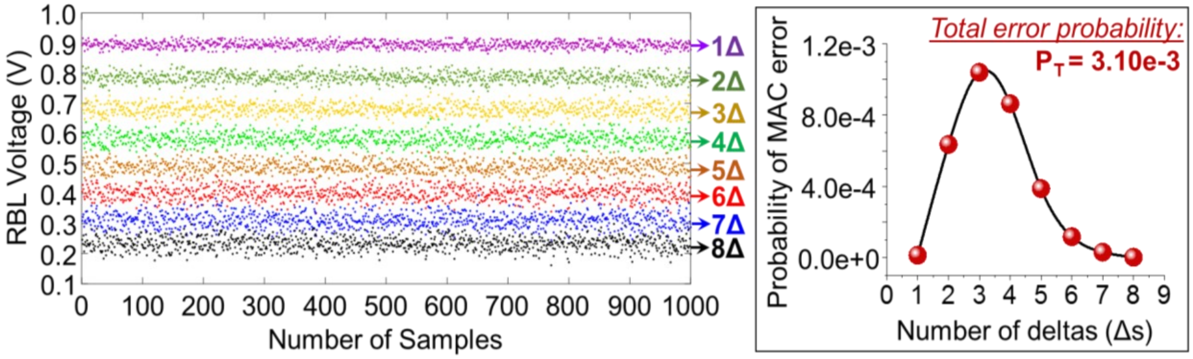


Fig. 7.4 (a) Variation analysis with 1000 Monte Carlo sample for each state varying from 1Δ to 8Δ. (b) Probability of MAC error with varying Δs.

magnitude ±1 (since only adjacent Δ states overlap, as seen in Fig. 7.4(a)). Our system-level evaluations reveal that P_T of 3.10e^{-3} has negligible impact on accuracy of DNNs, attributed to the low magnitude of errors and resiliency of DNNs to computational errors [252]. Note that FEFETs may encounter variability due to variation in FE parameters such as domain size/distribution [157], whose implications on the proposed ternary computation requires additional study.

7.4 TeC-Array Design

In this section, we present an array architecture using the proposed TeC-Cells for accelerating ternary DNNs. The TeC-Array can perform massively parallel vector-matrix multiplication (or in-memory dot product computation) between ternary inputs and weights. The maximum number of simultaneously accessed cells in a column is determined by two factors: (a) Sensing failure: As discussed in the previous section, increase in 'a' or 'b' results in reduced sense margins and higher errors. (b) Sparsity: At the same time, the occurrence probability of large 'a' or 'b' is also low [233]. This is due to >40% of vector elements being zeros as discussed in [225], [226], [233] (known as sparsity in DNNs). Therefore, considering the above-mentioned factors, the optimal number of TeC-Cells which can be accessed simultaneously is N=16. It is important to note that, although we can only detect a maximum of 8 states reliably (Fig. 7.3(c)), we are able to use N=16 by harnessing the advantages of sparsity in DNNs [233]. However, having only N=16 TeC-Cells in each column of an array may not be practical. Therefore, we designed a blocked 2D array with TeC-Cells, grouped into M=16 blocks, with each block containing K=256 columns, and each column having N=16 rows of TeC-Cells. Thus, the proposed array consists of N*M*K TeC-Cells (Fig. 7.5). We use a block decoder to access the N rows of a block simultaneously. WWLs, RWL1s and RWL2s of TeC-Cells in a row are connected together, while WBL1s, WBL2s, RBL1s and RBL2s of TeC-Cells in the same column are shared. K TeC-Cells in a row are accessed together for the read/write operations. On the other hand, in-memory ternary dot product computation is achieved at the block granularity where K dot-product operations of vector length N are performed in parallel. 3-bit Flash ADCs connected to RBL1 and RBL2 along with a 3-bit subtractor are employed for determining the dot-product (since maximum number of detectable Δs =8Δ; see Fig. 7.3(c)). Therefore, in one block access, the array can perform ternary multiplication of input vector I (with N elements) and weight matrix W (of size N*K).

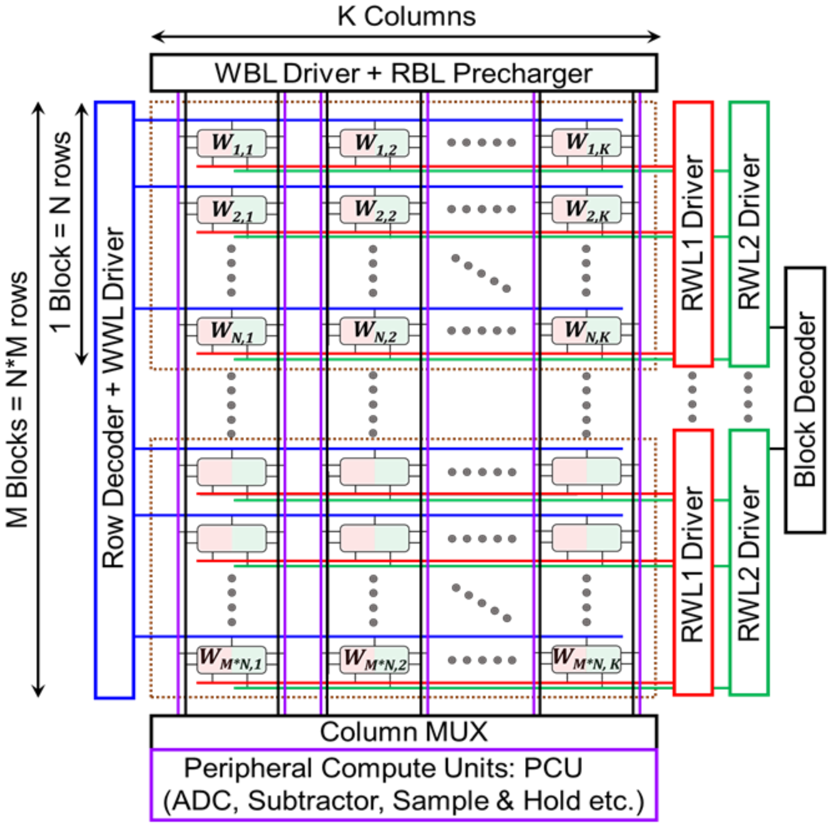


Fig. 7.5. TeC-Cell array design with N-rows and K-columns in a block and M-blocks in a column.

In order to perform ternary dot products on vector lengths N=16, we utilize the technique proposed in [233] of storing partial sums in a peripheral compute unit (PCU) using a sample and hold circuitry. After multiple block accesses (in the same column), we accumulate all the partial sums to determine the final dot products. The dot products are then quantized, and passed through an activation function to derive inputs to the next DNN layer. Moreover, as discussed in [233] we utilize L=32 PCUs for the entire array (where L<K=256) in order to amortize area energy overheads of the peripheral circuits.

7.5 Results

7.5.1 Array-level

In this sub-section, we compare the write, read and MAC performance and energy of the proposed TeC-Array with respect to two baselines: 6T-SRAM and 3T-FEFET NVM. We design near-memory ternary accelerators for the baselines, where the accelerators access scratchpad memories row-by-row before performing vector-matrix multiplication. We note that the gains shown for our design are pessimistic as we do not include the energy and latency of the processing elements in the near-memory compute baselines. All the memory arrays are designed with the same capacity (=128Kb).

(a) Layout Area (Fig. 7.6(a)): The proposed TeC-Cell exhibits 33% lower area compared to two 6T-SRAM cells (which can store a ternary bit) due to 4 less transistors. With respect to two 3T-FEFET-NVM cells, the proposed TeC-Cell exhibits 34% higher area attributed to the additional M_5 and M_6 transistors (Fig. 7.1) that are added to enable ternary in-memory computation. Note that, although two 6T-SRAM or 3T-FEFET cells can store a ternary weight, they do not support in-memory ternary compute offered by the proposed TeC-Cells.

(b) MAC Operation (Fig. 7.6(b)): The major advantage of the proposed TeC-Array is massively parallel in-memory computation of ternary dot-products. This results in 91% and 89% higher performance for the TeC-Array in comparison with the SRAM and 3T-FEFET NVM array baselines. At the same time, the MAC operation using TeC-Arrays exhibits 72% and 74% improved energy efficiency compared to SRAM and 3T-FEFET NVM, respectively. This is attributed to the simultaneous assertion of multiple-word-lines unlike the near-memory compute baselines which require row-by-row access. For DNNs, the predominant contributor to energy/delay is the MAC operation. Hence, the energy savings achieved at array-level are expected to translate to system-level energy efficiency, as discussed subsequently.

(c) Read/Write Operations (Fig. 7.6(c, d)): The enablement of ternary in-memory computation in the proposed TeC-Cells comes at the cost of some overhead for the read/write operations. Compared to 3T FEFET-NVM, we observe 19%, 12%, 19% and almost similar read delay, write delay, read energy and write energy, respectively, for the proposed TeC-Cells. This is mainly attributed to the larger cell area and additional BL capacitances due to the drain capacitances of M_5 and M_6 (Fig. 7.1). When compared to SRAM, we observe similar trends with one exception in read delay which is 7% lower. This is due to lower WWL capacitance in TeC-Cell (due to smaller area). Note that write energy of FEFET memories is ~2X compared to SRAM, mainly due to the overheads associated with negative voltages needed for polarization switching.

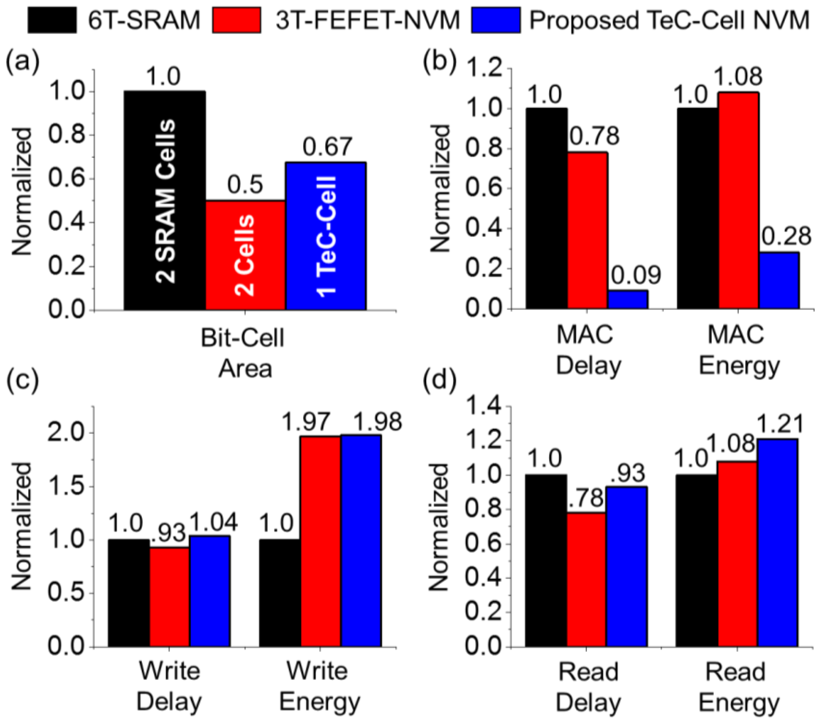


Fig. 7.6. (a) Cell layout area and normalized energy-delay metrics for (b) MAC, (c) Write and (d) Read operations for TeC-Cell with in-memory computation, FEFET-NVM and SRAM with near memory computation.

It is important to note that in DNN applications, more than 90% of operations are MACs. Therefore, even in the presence overheads in standard read and write operations, the total system performance and energy is drastically improved for ternary DNNs implemented using TeC-Arrays, as discussed next.

7.5.2 System evaluation

(a) Simulation framework: In this sub-section, we evaluate the system-level performance/energy efficiency of TeC-Cells. We collaborated with Prof. Anand Raghunathan and Dr. Shubham Jain to evaluate the TeC-Arrays in the context of a state-of-the-art ternary DNN accelerator. To that end, we utilize the TiM-DNN accelerator architecture proposed in [233] and design an TeC-Cell based system (TeC-System) with 32 TeC-Arrays (256x256). We compare the TeC-System with near-memory DNN accelerators to quantify the system-level benefits due to in-memory operations enabled by the proposed TeC-Cell. The baseline accelerators are designed using memories with near-memory computation units to execute ternary dot-products. (Note, baseline memories considered here cannot perform in-memory computation). We use two memory technologies SRAM and FEFET, and design two types of baseline systems: (i) iso-area and (ii) iso-weight storage capacity (2 Mega ternary words) as the TeC-System. TeC-Arrays (256x256) are 0.89X smaller than 6T SRAM arrays (256x512) and 1.5X larger than FEFET arrays (256x512) (including the overheads of peripherals). Therefore, the SRAM based iso-area design uses 28 arrays and the FEFET based iso-area baseline utilizes 48 arrays. We use an in-house architectural simulator to obtain the energy/performance of the TeC-System compared to the baselines using a suite of DNN benchmarks [233].

(b) Performance benefits: Fig. 7.7(a) shows the normalized execution time for various DNN benchmarks executed on the baseline and the proposed designs. We also show the breakdown of the execution time into two components – TMAC-Ops (Ternary vector-matrix multiplication operations) and Non-TMAC-Ops (other DNN operations). On

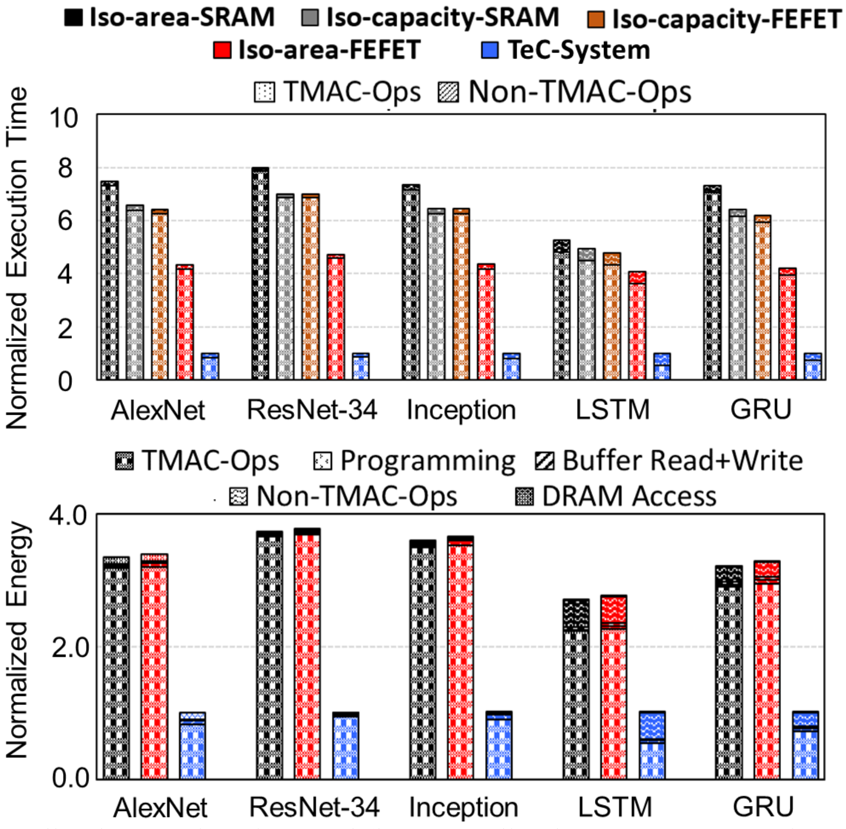


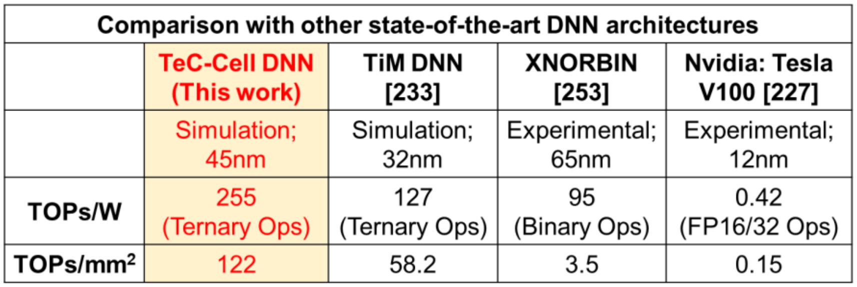
Fig. 7.7. (a) Normalized execution time and (b) Normalized energy consumption of the proposed TeC-System with respect to iso-capacity and iso-area baselines using SRAM and FEFET NVM based near-memory compute architectures, for a suite of DNN benchmarks.

average, we achieve 7X and 6.3X speedup over SRAM based iso-capacity and iso-area baselines, respectively, and 6.1X and 4.3X speedup over FEFET-based iso-capacity and iso-area baselines, respectively. Across our baselines, the FEFET-based iso-area design achieves the best performance due to the higher-level of parallelism available from the extra 16 arrays. For the proposed design, the performance benefits arise due to ternary in-memory operations in TeC-Arrays, wherein we activate and compute on 16 memory rows simultaneously. The application-level speedup depends on the fraction of the execution time spent on TMAC-ops, and therefore, benchmark applications with higher TMAC-Ops/Non-TMAC-Ops ratio achieve higher speedups.

(c) Energy benefits: We present the system-level energy benefits of the TeC-System over the iso-area SRAM and FEFET baselines. Note that, the iso-capacity baselines will exhibit similar energy consumption as iso-area baselines because, the total system energy consumption depends on the number of TMAC-Ops and Non-TMAC-Ops, which remains constant for an iso-capacity or iso-area baseline. Fig. 7.7(b) shows that the major components of energy consumption are TMAC-Ops, programming (writing weights into arrays), DRAM accesses, buffer reads and writes, and Non-TMAC-Ops. On an average, we achieve 3.3X and 3.4X reduction in the application-level energy over the SRAM and FEFET baselines, respectively. Across our benchmark applications, the factors indicating higher speedup are also predictive of higher energy savings, i.e., larger fraction of TMAC-Ops leads to superior energy benefits. This is because the proposed TeC-Array utilizes massively parallel in-memory TMAC-Ops which are more energy efficient than near-memory computing baselines (Fig. 7.6). We also observe that the FEFET-iso-area baseline consumes slightly more energy than the SRAM-iso-area design, due to high write energy of FEFETs.

Table. 7.2 shows the comparison of the proposed architecture with other state-of-the-art approaches. With respect to TiM-DNN [233], we achieve ~2X improvement in TOPS/W and TOPS/mm2 due to TeC-Cell's compact layout footprint. With respect to experimental findings in XNORBIN [253] and Tesla V100 [227], which are traditional computing architectures (not in-memory), we observe 2.7X-607X and 35X-813X improvements in TOPS/W and TOPS/mm2, respectively.

Table. 7.2 Comparison of TeC-Cell DNN with other state-of-the-art DNN architectures



Comparison with other state-of-the-art DNN architectures				
	TeC-Cell DNN (This work)	**TiM DNN [233]**	**XNORBIN [253]**	**Nvidia: Tesla V100 [227]**
	Simulation; 45nm	Simulation; 32nm	Experimental; 65nm	Experimental; 12nm
TOPs/W	255 (Ternary Ops)	127 (Ternary Ops)	95 (Binary Ops)	0.42 (FP16/32 Ops)
TOPs/mm^2	122	58.2	3.5	0.15

7.6 Summary

We proposed a ternary compute-enabled NVM cell (TeC-Cell), which can perform scalar multiplication of the stored value (weight) and an external input, where both the weight and the inputs are signed ternary numbers. Utilizing the TeC-Cell, we designed an array (TeC-Array) that performs massively parallel signed ternary dot-products in-memory. We demonstrated that the TeC-Array achieves significant energy-delay benefits compared to near-memory designs based on FEFET-based NVM and SRAM. Finally, we incorporated the proposed TeC-Array in a ternary DNN accelerator to evaluate its performance and energy benefits across a wide range of state-of-the-art DNN benchmarks including both deep convolutional and recurrent neural networks. We achieved 3.3X-3.4X energy efficiency and 4.3X-7X performance boost compared to SRAM and FEFET-based near-memory DNN accelerator.

8. 2D TRANSITION METAL DICHALCOGENIDE BASED SPIN-DEVICES EXHIBITING LOGIC-MEMORY SYNERGY

8.1 Introduction

So far, we have considered ferroelectric technologies in the implementation of non-volatile memories, non-volatile logic and compute-in-memory architectures for Boolean, Non-Boolean and arithmetic operations. They exhibit some unique characteristics and advantages, primarily attributed to the energy-efficient electric field driven memory operations. On the other hand, spin-based memories using magnetic tunnel junctions (MTJs) [254] are equally attractive due to their industrial demonstration of extremely high integration densities along with excellent endurance and retention properties. Specifically, spin-transfer-torque magnetic RAM (STT-MRAM) has attracted immense interest. Samsung's STT-MRAM in 28nm FDSOI platform [28] and Intel's FinFET based MRAM technology [29] are some industrial efforts on the implementation of spintronic memory. Although they have their own application space in a wide-range of systems, there are several challenges which still need to be addressed for improving and further advancing spin-based technologies.

STT-MRAMs exhibit low distinguishability between their bi-stable states making them prone to sensing failures [255]. Also, due to their two-terminal cell-design, the write and read paths are coupled, leading to design challenges. Recent advancements with the possibility of generating spin polarized current using charge current in heavy metals has led to the realization of the Giant Spin Hall (GSH) effect based MRAM [251], [256], [257] (also known as spin-orbit-torque MRAM; SOT-MRAM). Compared to STT-MRAMs, GSH-MRAM showcase significant improvement in write energy along with the possibility to independently co-optimize the read and write operations due to their decoupled read and write current paths. GSH effect also enables the possibility of achieving a differential storage due to the simultaneous generation of opposite polarized spin currents [32]. However, both the single ended and differential memory designs based on GSH effect require multiple access transistors leading to a significant area penalty [32], [251]. Also, the spin injection efficiency which is directly proportional to the spin hall angle (θ_{SH} <0.3) is low for these heavy metals [256], [258]. This results in performance degradation and energy inefficiency. Another drawback with GSH-MRAMs is that they can only switch in-plane magnetic anisotropy

(IMA) magnets without the presence of any external magnetic field or geometrical changes to the ferromagnet [259], [260]. As perpendicular magnetic anisotropy (PMA) magnets are known to be more energy efficient in switching and thermally stable than IMA [261], GSH-MRAMs offer limited performance and energy benefits. Therefore, there arises a need to explore new memory technologies to harness the full potential of spin-based storage.

On the application front, as also discussed earlier, data-intensive workloads have come to the forefront in recent years. This has led to frequent and humongous number of data accesses from the memory system to the processor. As a result, larger amount of storage is required, which demands the exploration for high density memory solution. On the other hand, due to the larger delay associated with memory access compared to the processing time (also known as memory-wall problem [4], [6]), the data movement to and from the memory cell (across the bit-lines, memory interface and interconnects) is a major performance and energy bottleneck in standard computing architectures. Therefore, there is also a need to explore alternate computing paradigms such as Computation-in-Memory (CiM), where computations are performed inside the memory array [62], [66]–[68], [203], [219]. This reduces the data transfer between memory and processor, thereby improving the performance and energy efficiency.

Most of the prior efforts on CiM designs using spin-based information storage involve the use of single ended STT-MRAM or SOT/GSH-MRAMs [59], [262]–[264]. Multi-word-line assertion along with a modified sense amplifier and peripheral circuitry enables Boolean and arithmetic operations to be performed within the memory array [59]. Although STT-CiM [59] benefits form the high density and good endurance of STT-MRAM, and GSH-MRAM based CiM [263] overcomes the drawbacks of high write energy consumption in STT-MRAMs, they both suffer from degraded robustness during in-memory compute. This is attributed to the poor distinguishability between their bi-stable states yielding deteriorated sense margins during compute operations [59], [262]–[264]. Therefore, it is important to explore robust and energy-efficient CiM designs utilizing the benefits offered by spin-based information storage for current and future generation of compute systems involving large amount of data.

To that end, in this chapter, we propose non-volatile memory devices based on Valley-coupled-Spin Hall (VSH) effect with the ability to naturally switch PMA magnets, leading to higher energy efficiency compared to GSH-MRAM. Moreover, exploiting the spin generation through a semiconductor (WSe_2) rather than a metal (in GSH), we propose an integrated gating in our NVM devices, which enable access transistor less bit-cell design leading to large integration density. Furthermore, the proposed devices inherently lead to differential read functionality. Leveraging this attribute, we present an array design with energy efficient computation-in-memory capabilities, which can potentially alleviate the von-Neumann bottleneck and overcome the drawbacks of existing spin-based CiM designs.

8.2 Background

8.2.1 Giant spin hall (GSH) effect

The Giant Spin Hall effect is an efficient mechanism for generating spin polarized currents. A charge current passing through a heavy metal layer such as Ta, Pt or W have been experimentally demonstrated to generate in-plane spin polarized currents [256], [258] (Fig. 8.1(a)). The GSH effect is mainly used for switching magnetization of IMA magnets. Deterministic switching of PMA magnets using GSH effect requires externally assisted magnetic field to break the symmetry [259] or complex design modifications to the MTJ geometry [260]. The major advantage with GSH effect-based magnetization switching is the low write current/energy when compared to the STT-based magnetization switching [32], [256], [258], [265], [266].

The generated spin current (I_S) to charge current (I_C) ratio which is also known as the spin injection efficiency is directly proportional to the spin hall angle, θ_{SH} [251]. Experiments have shown θ_{SH} ~ 0.1-0.3 for heavy metals, resulting in low spin injection efficiency [256], [258] . Furthermore, the efficiency of GSH effect is impacted by the spin-flip length (λ_S), which characterizes the mean distance between spin-flipping collisions. λ_S has been calculated to be around ~1-2nm [256], [258] for heavy metals with large GSH effect.

8.2.2 GSH effect based non-volatile memories

The three terminal device structure of the GSH effect-based spin device (Fig. 8.1(a)) mitigates the read-write conflict of the two terminal STT-MRAM due to the separation of read-write paths. Moreover, such an approach has shown to be promising for energy-efficient storage compared to STT-MRAM [32], [251], [256], [258]. Several bit-cell designs have been proposed using the GSH effect [32], [251], [265]. Fig. 8.1(b) shows circuit schematic of GSH-MRAM which consist of a read and write access transistor for single-ended memory [251]. Write operation is achieved by turning ON the write access transistor and depending on the direction of charge current flow (which determines the spin current polarization), the MTJ state is stored. The spin current interacting with the MTJ to deterministically switch the magnetization is calculated as:

$$I_S = \frac{A_{MTJ}}{A_{HM}} * \theta_{SH} * I_C \qquad (8.1)$$

where A_{MTJ} and A_{HM} are the cross-sectional area of MTJ and heavy metal, respectively [251]. The read operation is carried out by turning ON the read access transistor and sensing the resistance state of the MTJ (parallel (P) or anti-parallel (AP)). As the read and write paths are decoupled, they can be optimized independently [251], [267].

Utilizing the opposite spin generation at the top and bottom surfaces of the heavy metal, a differential GSH-MRAM (DGSH-MRAM) was proposed in [32] with two MTJs placed on either side of the heavy metal layer (Fig. 8.1(c)). This leads to true and complimentary bit storage in the memory cell. The write operation remains the same as GSH-MRAM while the read is achieved using differential sensing, leading to higher sense margins. However, compared to GSH-MRAM, two more additional transistors are required to selectively access a bit cell in an array without disturbing the unassessed cells. Furthermore, fabrication of true and differential MTJs on the top and bottom sides of the heavy metal may increase processing complexities and costs.

The above mentioned GSH effect-based memory designs have been proposed to switch IMA based MTJs. This is because, only in-plane spin polarized currents are generated in the heavy metals. IMA magnets are not suitable for ultra-scaled dimensions due the limit on the aspect ratio of the free layer as well as low thermal stability [259]–[261]. In comparison, PMA magnets are more stable and robust at scaled dimensions with high packing density, which is mainly attributed to the absence of in-plane shape magnetic anisotropy [267].

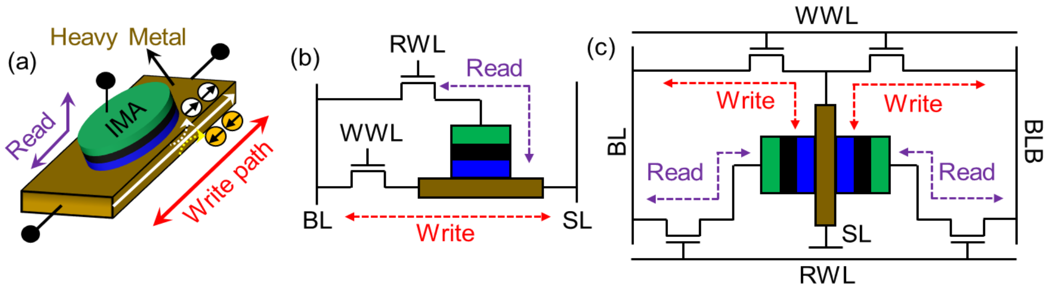


Fig. 8.1. (a) GSH effect in heavy metal leading to magnetization switching in MTJ. (b) Single ended GSH-MRAM and (c) Differential DGSH-MRAM bit-cell schematics.

Moreover, due to the absence of de-magnetization fields, lower energy is required for magnetization switching in PMA magnets compared to IMA, even at iso-thermal stability [267]. Although, GSH effect-based PMA switching has been demonstrated with external magnetic field [259], [268], or GSH assisted STT switching [269], [270] or a local di-polar field [271] or introducing tilted anisotropy in the ferromagnet [260], the feasibility of achieving such a design change in scaled, high density technologies is yet to be explored. Moreover, the requirement of additional access transistors for GSH effect-based bit-cell designs leads to large area overheads which also increases the energy consumption for bit-line and word-line charging.

To address the aforementioned challenges associated with GSH-effect based devices, we propose to utilize the valley-coupled-Spin Hall (VSH) effect in monolayer WSe_2 to design MRAMs based on PMA magnets. The VSH effect is naturally suited to switch PMA magnets, which promises higher energy efficiency in the proposed designs. Before we discuss our memory designs and the associated benefits and trade-offs, let us briefly review the VSH effect, next.

8.2.3 Valley-coupled-spin hall (VSH) effect

Monolayer transition metal dichalcogenides (TMDs) are multi-valley 2D semiconductors (Fig. 8.2(a)) with inherent broken inversion and preserved time reversal symmetries. Time reversal symmetry requires that the spin polarization in the K and K' valley must be opposite (illustrated

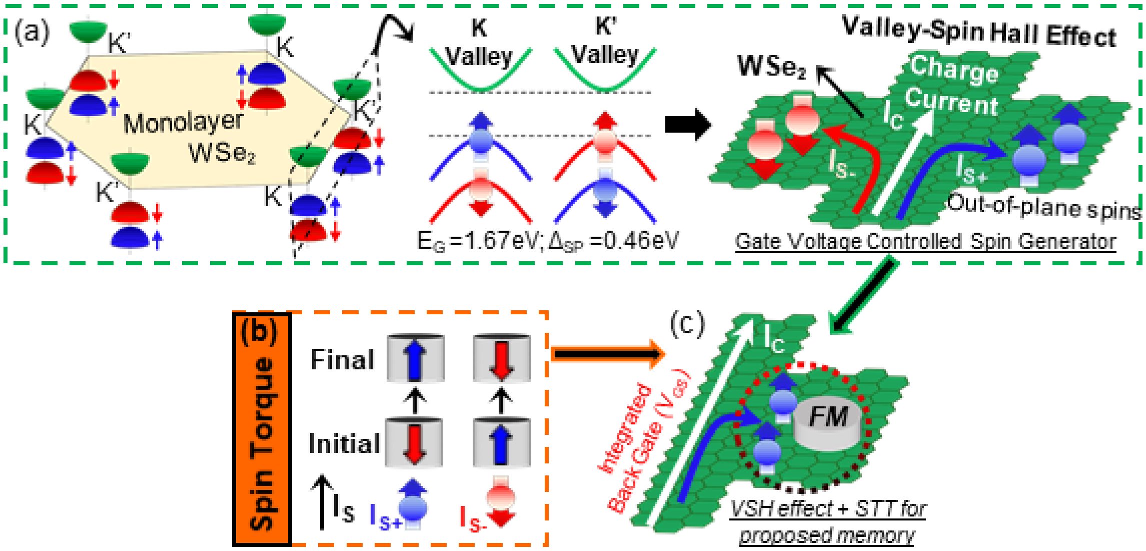


Fig. 8.2 (a) Band structure of WSe_2 showcasing spin-valley coupling resulting in VSH Effect (b) STT for switching PMA magnet. (c) Proposed idea of coupling VSH effect and spin torque for NVM design.

as blue and red arrows in Fig. 8.2(a)), which in combination with the large spin splitting (Δ_{SP}) in the valence band for TMDs such as WSe_2 [272], gives rise to holes in the K and K' valley with opposite signs of spin polarization at the Fermi level. As a result, carriers in the K and K' valleys of the valence band (p-type) possess nonzero Berry curvature (Ω) such that $\Omega(K)= -\Omega(K')$. The resultant transverse carrier velocity leads to valley-coupled spin currents on the application of electric field. This phenomenon is called the VSH effect [93], [273], [274]. VSH effect in WSe_2 generates out-of-plane spin polarized currents (I_{S+}/I_{S-}; Fig. 8.2(a)) which can switch PMA magnets without any external magnetic field (B_{EXT}) or complex changes to the MTJ structure, unlike GSH-effect based memory devices [259], [260], [268]–[271]. It has been experimentally demonstrated that monolayer TMDs exhibit a large valley-hall angle, θ_{VH} ~1 [93] at 25°C. Now, due to the existence of strong spin-valley coupling in monolayer WSe_2 [273], [274] (as a result of large Δ_{SP}), the θ_{SH} is expected to be equal to θ_{VH}, i.e., θ_{SH} ~1. The large θ_{SH} corresponds to high spin injection efficiency which can potentially lead to enhanced energy efficiency during magnetization switching. In contrast, GSH effect exhibit relatively much smaller θ_{SH} ~ 0.1-0.3. Moreover, VSH effect resulting in out-of-plane spin generation exhibits λ_S of 0.5-1µm [93], [273] (unlike GSH effect in heavy metals; λ_S ~ 1-2nm). The large θ_{SH} and λ_S in monolayer WSe_2 opens up new opportunities for information storage.

Utilizing the unique attributes of VSH effect in conjunction with spin torque physics (Fig. 8.2(b, c)), we propose valley-coupled spintronic devices in this chapter and showcase their application in (a) high-density non-volatile memories, (b) computation-in-memory for Boolean/arithmetic workloads and (c) non-Boolean computing for accelerating neural networks.

8.2.4 Fabrication and experimental results

For supporting our exploration in this chapter, we collaborated with Prof. Zhihong Chen and Dr. Terry Hung for experimental evaluation of the valley-coupled-spin-hall effect in monolayer WSe_2 material. The insights from their experimental results (as discussed below) have been used in the development of a comprehensive simulation framework (discussed in the next section), which is implemented for the evaluation of the proposed non-volatile memory arrays in the context of general-purpose systems as well as targeted application platforms, along with the study of Boolean and non-Boolean CiM designs for data-intensive computing.

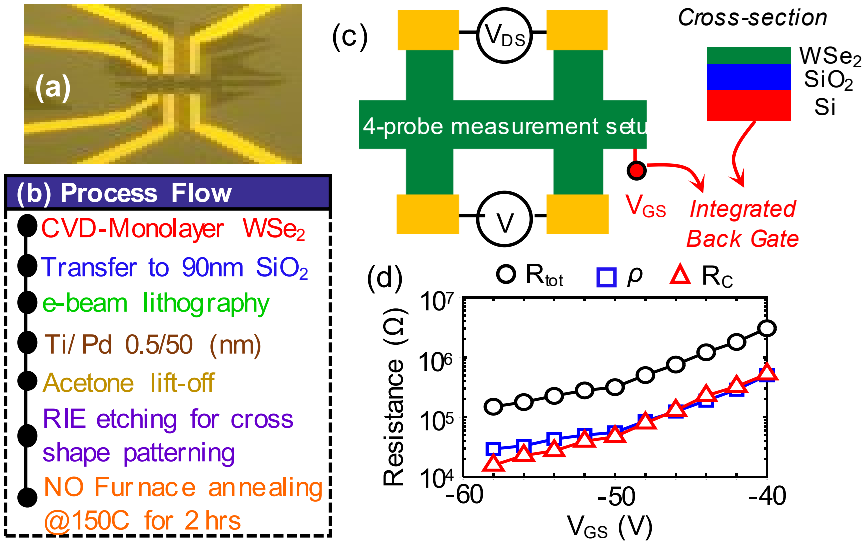


Fig. 8.3 (a) Optical microscope image of the hall bar device structure (b) Fabrication process flow (c) Cartoon of 4-probe measurement setup (d)Total, sheet and contact resistances versus gate voltage, V_{GS}.

(a) Device fabrication: The process flow for the fabrication of a double Hall cross device structure (Fig. 8.3(a)) is illustrated in Fig. 8.3(b). Different from a conventional Hall bar structure, it has two crosses and that enables the separation of local charge and non-local charge for our study. Chemical vapor deposition (CVD) grown WSe_2 films were transferred to 90nm SiO_2 substrates with highly doped silicon on the back side. Doped Si serves as the integrated back gate for controlling the flow of I_C and I_S (as explained later). Standard e-beam lithography using PMMA A4 950 resist was employed to pattern electric contacts on the CVD WSe_2 flakes. Ti/Pd (0.5/50nm) was deposited in an e-beam evaporator followed by a lift-off process in acetone. CVD grown BN film was transferred from Cu foil onto the devices through a process that involves etching the Cu foil with iron chloride ($FeCl_3$) and immersing it in diluted HCl and DI water alternatingly for few times before scooping up. This BN layer was inserted to minimize any effect induced by PMMA residues due to RIE etching process, but not necessary for general (D)VSH-MRAM fabrication. RIE etching mask was defined by e-beam lithography using PMMA A4 950 resist and BN/WSe_2 flakes were etched using Ar/SF6 for 10 seconds. The final devices underwent vacuum annealing ($\sim 10^{-8}$ torr) at 250°C for four hours to remove PMMA residue and nitric oxide (NO) furnace annealing at 150°C for two hours to achieve p-doping [275].

(b) Parameter extraction: Two types of measurements were performed, as illustrated in Fig. 8.3 and 8.4 [93]. A conventional four probe measurement (Fig. 8.3 (c)) was conducted to extract sheet resistance (ρ), contact resistance (R_C) and total resistance (R_{TOT}) (Fig. 8.3(d)). The non-local (NL) measurements were performed to probe the Hall voltage induced by any carrier distributions due to the VSH and its reciprocal effect (Fig. 8.4(a)). Note that only the ON state of the WSe_2 device will be considered for valley-coupled-spin transport in our discussions below, since access to holes in the valence band is necessary. It is worth mentioning that the so-called Ohmic contribution was clearly excluded [93]. We ensured that by plugging in the channel resistance extracted from four-probe measurements in Fig. 8.3 (c) and the device geometry (Fig. 8.4(a)) into Equation-8.2 [276].

$$V_{ohmic} = I_C * \rho * \frac{W}{W_1} * e^{-\frac{\pi L_A}{W}} \qquad (8.2)$$

Unambiguously, ohmic contribution (not shown here) is orders of magnitude smaller than the non-local signal we measured in the following. Fig. 8.4(b, c) shows gate control of charge current (I_C) and NL resistance (R_{NL}= V_{NL}/I_C) for different device samples with arm lengths (L_A) equal to 2μm, 3μm and 5μm. R_{NL} vs L_A at V_{GS}= -60V was used to extract the spin flip length, λ_S = 550nm from the fitting of $R_{NL} \propto e\ (-L_A/\lambda_S)$ [93], [273] (Fig. 8.4(d)). It is important to note that the spin-generator is a p-type device and therefore it requires negative gate-to-source voltages to turn it ON.

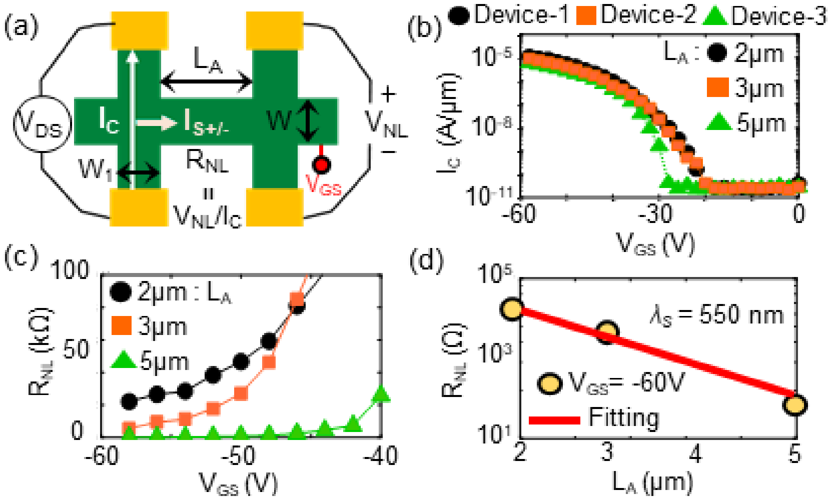


Fig. 8.4 (a) Non-local measurement setup (b) Charge current (I_C) and (c) Non-local resistance (R_{NL}) vs V_{GS} for different L_A and (d) R_{NL} vs L_S to extract spin flip length, λ_S.

8.3 Simulation Framework

We have built a self-consistent simulation framework in SPICE for the proposed valley-coupled spintronic memory device/array evaluation (Fig. 8.5(a)) [277], [278]. Monolayer WSe_2 electrostatics is modelled using the capacitance network model suggested in [279], albeit with modification for back-gated device used in this work. Further, we model the charge current using the continuity equations for drift-diffusion transport as proposed in [279] (calibration in Fig. 8.5(b)). The charge current is then used in conjunction with the Valley Spin Hall effect model, which calculates the spin current based on the experimental θ_{SH} and λ_S values [93], [273]. I_S interacting with the free layer of MTJ is calculated as

$$I_S = \frac{D_{MTJ}}{L_G} * \theta_{SH} * I_C \qquad (8.3)$$

where D_{MTJ} is the diameter of MTJ (circular) and L_G is gate length of the transistor (see Fig. 8.6). Spin diffusion and interface scattering are considered in the monolayer WSe_2 channel while calculating the spin current flow as per the method proposed in [93], [269]. Landau-Lifshitz-Gilbert- Slonczewski (LLGS) equation is used to model the switching dynamics of the PMA magnet, which serves as the free layer (FL) of a magnetic tunnel junction (MTJ) formed on top of the TMD (as described later). For sensing, the MTJ resistance (R_{MTJ}) model is obtained from [280]. Further, as we will discuss in detail later, the read path is 'T'/ 'H' shaped. To properly account for

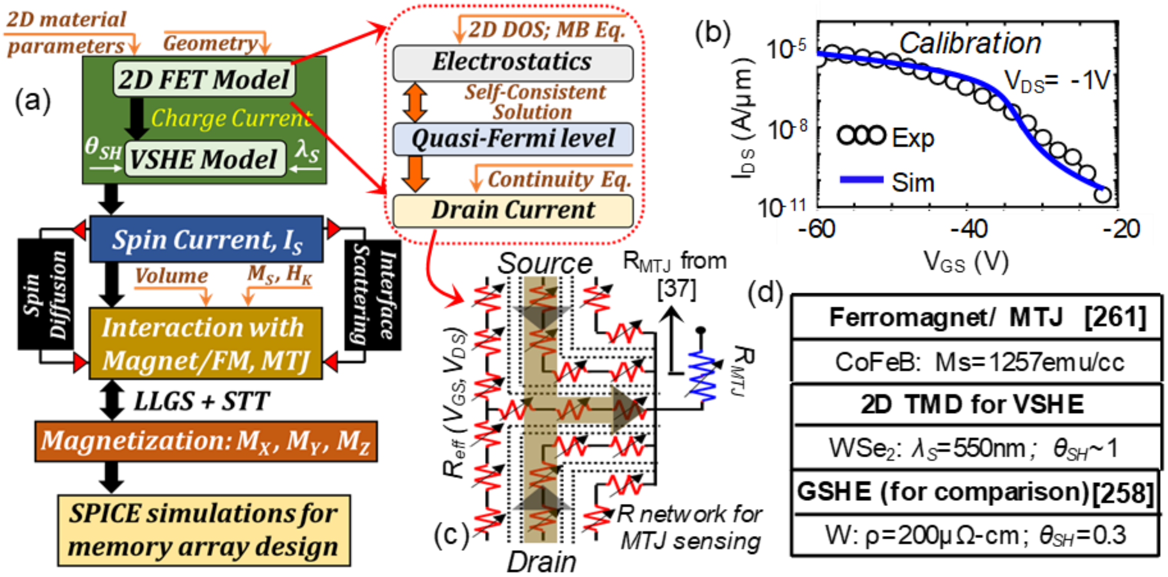


Fig. 8.5 (a) Self-consistent simulation framework (b) Calibration of the monolayer WSe_2 FET (c) SPICE-based distributed resistance network for sensing MTJ resistance (d) Material parameters

the sensed currents, we use a distributed resistive network (Fig. 8.5(c)) based on the conductance of WSe_2 layer and the shape of the read path. Therefore, the read path includes the resistance of the MTJ as well as that of conducting WSe_2 layer. The sensed currents are used to read the bit-information stored and also perform computation in memory, as discussed extensively later. We incorporate contact resistances at the drain terminal, source terminal and MTJ-TMD interface based on [281] (accounting for Schottky barrier). The contact resistances play a crucial role and require further investigation for performance/energy optimization. The simulations parameters used in this chapter are shown in Fig. 8.5(d). Note, in all of our analysis, we evaluate the proposed memory devices and circuits considering a minimum gate length (L_G) of 45nm (for system compatibility). With the understanding of the simulation methodology, let us now present the proposed VSH effect based spintronic memory.

8.4 VSH Effect based Non-Volatile Memories

We propose single-ended and differential variants of non-volatile memories using the VSH effect, namely VSH-MRAM (single-ended) and DVSH-MRAM (differential). We discuss the memory device structures and their characteristics followed by array design in this section.

8.4.1 Structure and operation of VSH memory devices

Fig. 8.6(a, b) illustrates the proposed single ended and differential memory device structures. Single-ended VSH-MRAM consists of only one arm along which the transverse spin current flows. On the other hand, the differential DVSH-MRAM contains two arms for complementary spin current flow. In the single ended design, a PMA MTJ is integrated on top of the arm of the monolayer TMD spin generator as shown in the Fig. 8.6(a, b), whose free layer (FL) is used for non-volatile magnetic storage. In the differential design, two PMA MTJs storing true and complementary values are integrated on the two arms of the spin generator. The read terminals of the memory devices (connected to the pinned layer of the read MTJs - Fig. 8.6(a, b)) are used to sense the bit-information stored. By virtue of VSH-based write and MTJ-based read (discussed in detail later), the proposed memory devices feature decoupled read-write paths.

The VSH effect in monolayer WSe_2 generates out-of-plane spin current (I_S), which interacts with the MTJ through spin torque to switch the FL magnetization (Fig. 8.2). Since VSH effect

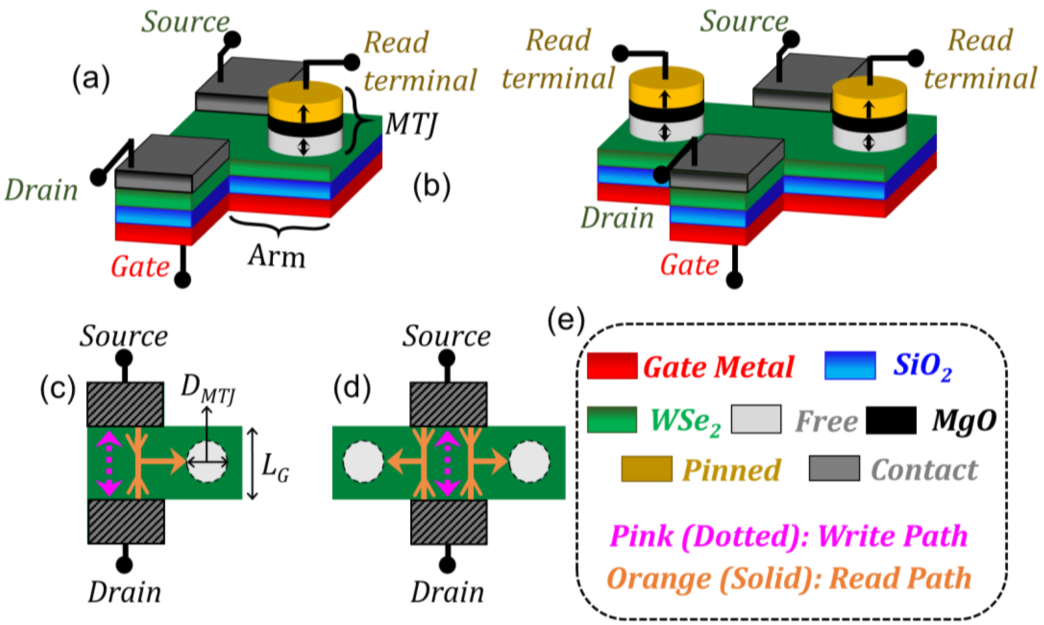


Fig. 8.6 Proposed (a) single ended (VSH-MRAM) and (b) differential (DVSH-MRAM) non-volatile memories. Read and write paths for (c) VSH-MRAM and (d) DVSH-MRAM. (e) Legend.

leads to the flow of opposite spin currents in divergent directions, the proposed DVSH-MRAM is able to seamlessly store and switch both true and complementary bits. The direction of the charge current (I_C) (controlled by the polarity of drain-to-source voltage (V_{DS})) determines the polarization of the spin current (I_{S+}/I_{S-}) flowing towards the MTJ(s) (see- Fig. 8.2 and Fig. 8.6 (c, d)). When the current flows from the drain to source terminals, I_{S+} flows towards the MTJ in VSH-MRAM and I_{S+}/I_{S-} flow towards the right/left MTJ ($MTJ_{R/L}$) in DVSH-MRAM. These spin currents generate spin torque leading to parallel (P) state in the MTJ of VSH-MRAM, and P and anti-parallel (AP) states in MTJ_R and MTJ_L respectively, in DVSH-MRAM. Current is passed in the opposite direction to store the opposite states. This corresponds to the write operation of the proposed memory devices. For reading the bit-information, we use the resistance difference between the P and AP states of MTJs. The read current flows from the source and drain terminals of the transistor to the read terminal of MTJs, in a 'T'/ 'H' shape as illustrated in Fig. 8.6 (c, d), for VSH/ DVSH-MRAMs. The biasing conditions to achieve this is explained later. Based on the current sensed at the read terminals (which depend on the state of MTJ, P/AP), the bit-information stored is retrieved. It is important to note that VSH-MRAM achieves single-ended sensing using a reference current source, while DVSH-MRAM enables differential sensing leading to higher sense margins and self-referenced operation. These aspects are discussed in detail later.

A unique feature of our devices is the integrated back gate, which enables modulation of the I_C, IS and hence the switching characteristics of the PMA magnets (gate controllability quantified later). While this feature can be appealing for several applications, in this work, we utilize it for compact memory design, as discussed in the subsequent sections.

8.4.2 Memory device characteristics

As mentioned previously, charge current flowing through the monolayer WSe_2 generates transverse spin currents. Fig. 8.7(a) illustrates the simulated gate voltage (V_{GS}) modulated charge and spin current flow. The polarity of V_{DS} determines the polarization for the spin current flowing towards the MTJ resulting in corresponding magnetization switching as illustrated in Fig. 8.7(b). Since, the gate voltage controls the carrier density in the WSe_2 layer, the magnetization switching time is gate controllable as shown in Fig. 8.7(c). Higher $|V_{GS}|$ corresponds to larger I_C (or I_S) which in turn results in smaller magnetization switching time. For our proposed (D)VSH-MRAM cells, we achieve switching time ranging from 3.2ns to 1.5ns for V_{GS}=-1.0V to -1.2V. Note, the magnetization switching time for VSH and DVSH-MRAMs devices remain similar because of the inherent and concurrent generation of the complementary spin currents (I_{S+} and I_{S-}) due to VSH effect. In comparison, GSH-MRAM cells (Fig. 8.1(b)) exhibit a switching time ranging from 11.6ns to 4.0ns for V_{DD} = 1.0V to 1.2V. The benefits are attributed to the easier switching of PMA magnets in VSH-MRAMs when compared to IMA in GSH-MRAMs, mainly due to lower switching current requirement for a given thermal stability [261]. (MTJ parameters in Table. 8.2).

It is important to mention that when V_{GS}=0V, the VSH memory device is OFF and the magnetization state is retained due to the non-volatility of the ferromagnet. To read or change the magnetization state stored, the device has to be turned ON (negative V_{GS}). As we discuss later,

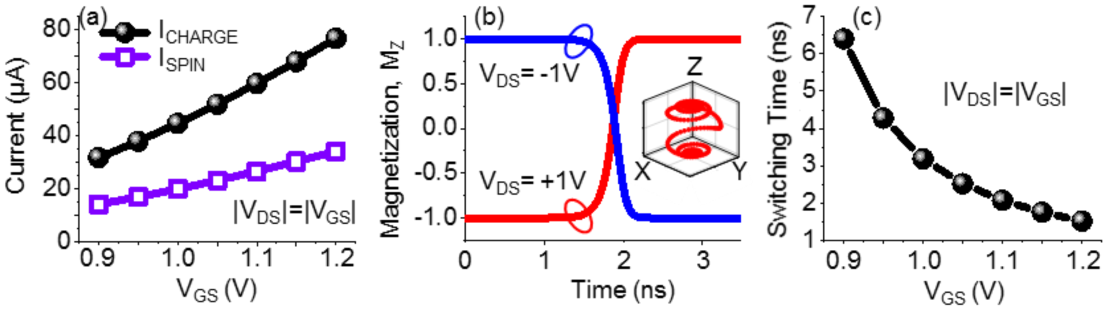


Fig. 8.7 (a) V_{GS} modulation of generated spin currents (b) Magnetization (M_Z) switching for different V_{DS} polarity (inset: magnetization trajectory) (c) Magnetization switching time vs V_{GS}.

during read, even though the device is ON, we ensure that no charge current flows from the drain to source (to avoid generation of spin current due to VSH effect), thereby safeguarding the magnetization state from any VSH-induced disturbance. Let us now discuss the proposed memory array design and operation.

8.4.3 Memory array design and operation

Utilizing the integrated back-gate of our devices, we propose VSH and DVSH-MRAM arrays which feature access-transistor-less cells by virtue of the integrated gate of the proposed devices (Fig. 8.8(a, b)). The integrated gates of all the memory cells in the same row are connected to the word-line (WL). The source, drain and the read-ports of all the cells in the same column are connected to bit-line (BL), bit-line-bar (BLB) and sense-line/sense-line-bar (SL/SLB) respectively. The integrated back gate provides selective word access in the array as discussed later. This feature leads to compact layouts as shown in Fig. 8.8(c, d). The memory operations are discussed next (bias conditions in Table. 8.1).

(a) Write: For writing into the proposed memory cell, we apply 0V to WL of the accessed word (Recall that the proposed devices are p-type). We then assert BLs and BLBs according

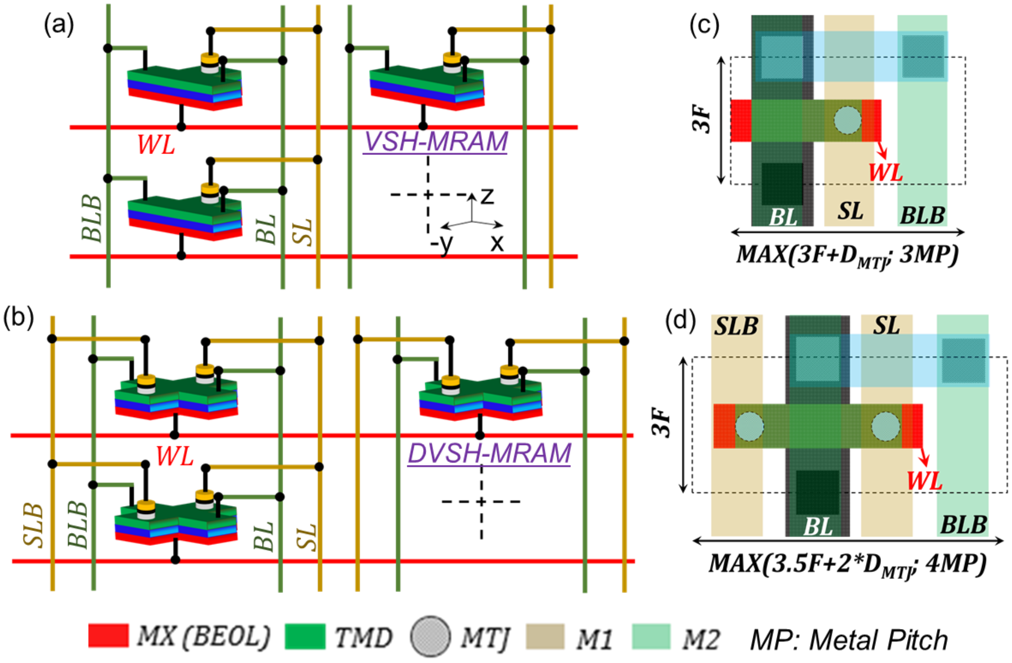


Fig. 8.8 Memory array architecture and bit-cell layout (with L_G=45 nm) for the proposed (a, c) VSH-MRAM and (b, d) DVSH-MRAM.

to the bit-information which is to be stored (direction of charge current determines the bit stored). SLs and SLBs are kept pre-charged (and floating) at V_{DD} (1.0V). This creates a high impendence path for the charge current to flow through MTJ, avoiding accidental magnetization switching due to STT effect. Now, let us first consider the case where we write bit- ‘0’. 0V/V_{DD} is applied to BL/BLB in both VSH- and DVSH-MRAMs (V_{DD}=1.1V). VSH effect flips the FL of MTJ in VSH-MRAM to positive magnetization state (M_Z=+1) and the MTJ comes to the P configuration. While for DVSH-MRAM, FL of MTJ_R and MTJ_L flip to positive and negative magnetization states (M_Z =+1 and –1) which brings them to P and AP configurations respectively, corresponding to bit-‘0’. On the other hand, for writing bit-‘1’, V_{DD}/0V is applied to BL and BLB, and the VSH effect leads to storage of M_Z=-1 in FL of MTJ (AP) of VSH-MRAM and M_Z= -1/+1 in FL of MTJ_R(AP)/MTJ_L(P) of DVSH-MRAM. Note, in DVSH-MRAM, the true bit value is stored in MTJ_R while the complementary bit is stored in MTJ_L. After write, all lines are pre-charged to V_{DD}. Note, the BLs/BLBs, SLs/SLBs of the unaccessed cells are precharged to V_{DD}, while the WLs are driven to V_{DD} to avoid any unintentional M_Z switching. This corresponds to V_{GS}=V_{DS}=0V in the unaccessed memory devices resulting in insignificant charge/spin current flow (no write disturbance).

(b) Read: For reading the bit-information, we apply 0V to WL and V_{DD} to BLs and BLBs of the accessed word. The SLs and SLBs are driven to V_{DD}-V_{READ}. This brings the memory devices of the accessed word to the ON state and there exists a read current flow between the sense line(s) and source/drain terminals of the memory cell (due to the voltage difference, V_{READ}=0.4V). The read current (I_{SL}/I_{SLB}) depends on the resistance of the MTJ storing P or AP configuration. For VSH-MRAM, I_P is the current sensed at SL when the memory cell stores bit-‘0’ (parallel configuration of MTJ) and I_{AP} is the current sensed when bit-‘1’ is stored (anti-parallel MTJ), where $I_P > I_{AP}$. For DVSH-MRAM, I_P (I_{AP}) and I_{AP} (I_P) are the

Table. 8.1 Operating bias conditions of VSH-MRAMs and DVSH-MRAMs

Red: Pre-charged	WL	BL	BLB	SL	SLB
WRITE	0	V_{DD}/0	0/V_{DD}	V_{DD}	V_{DD}
READ	0	V_{DD}	V_{DD}	V_{DD}-V_{READ}	V_{DD}-V_{READ}
HOLD	V_{DD}	V_{DD}	V_{DD}	V_{DD}	V_{DD}

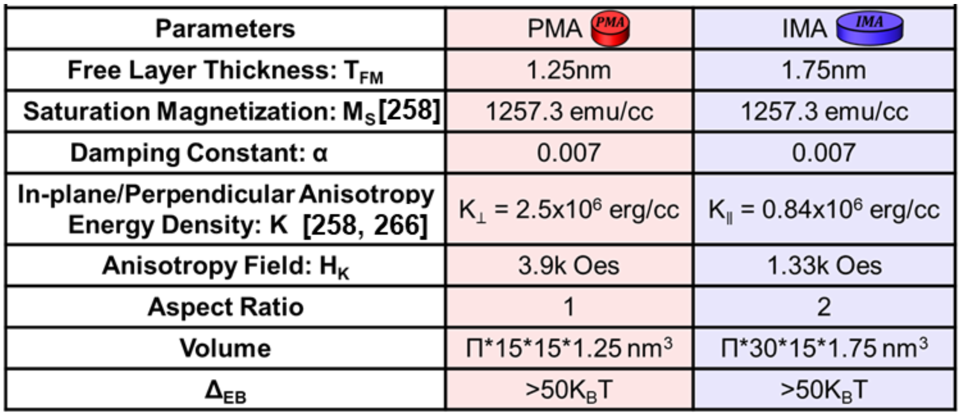
currents sensed at SL and SLB when the bit stored is ‘0’ (‘1’). VSH-MRAMs employs single-ended sensing, where a reference cell current, I_{REF}= (I_P+I_{AP})/2 is used to compare the current flowing through SL (I_{SL}). On the other hand, DVSH-MRAM is self-referenced. After the read operation, all lines are pre-charged to V_{DD}. Note, similar to the write operation, the BLs/BLBs and SLs/SLBs of the unaccessed cells are precharged to V_{DD} and the WLs are driven to V_{DD} to avoid any disturbances.

(c) Hold/Sleep: During the hold operation, all the lines of the memory array are precharged to V_{DD}. This process also ensures minimal energy consumption during charging/dis-charging of bit-lines for memory’s read/write operations. On the other hand, during the sleep mode, i.e., when the power supply is completely shut down for a long time, all lines (BL/BLB, SL/SLB and WL) are driven to 0V. In both these cases (hold and sleep modes), the non-volatility of the magnetization in FL of MTJ ensures storage of the bit-information even in the absence of any external power supply leading to zero stand-by leakage power.

8.4.4 Results

(a) Array-level analysis: We perform memory array analysis of the proposed VSH/DVSH-MRAMs in comparison with existing GSH/DGSH-MRAMs [32], [251]. We consider 1MB array (8 banks, each bank with 1024 rows and 1024 columns) with 32-bit words and evaluate the area, write and read metrics. Iso-energy barrier of ~55K_BT (>10 years of retention [261]) for PMA MTJs in the proposed VSH/DVSH-MRAMs and IMA MTJs in GSH/DGSH-MRAMs is considered for a fair evaluation. This is achieved by tuning the

Table. 8.2 MTJ parameters for (D)VSH-MRAMs and (D)GSH-MRAMs



Parameters	PMA PMA	IMA IMA
Free Layer Thickness: T_{FM}	1.25nm	1.75nm
Saturation Magnetization: M_S[258]	1257.3 emu/cc	1257.3 emu/cc
Damping Constant: α	0.007	0.007
In-plane/Perpendicular Anisotropy Energy Density: K [258, 266]	$K_\perp$ = 2.5x10^6 erg/cc	$K_\parallel$ = 0.84x10^6 erg/cc
Anisotropy Field: H_K	3.9k Oes	1.33k Oes
Aspect Ratio	1	2
Volume	Π*15*15*1.25 nm^3	Π*30*15*1.75 nm^3
Δ_{EB}	>50K_BT	>50K_BT

device geometry (MTJ parameters listed in Table. 8.2). Fig. 8.9 illustrates the array level comparison of the memory designs.

(i) Layout (Fig. 8.8(c, d)): The proposed VSH/DVSH-MRAMs achieve 66/62% lower bit-cell area compared to GSH/DGSH-MRAMs. This is attributed to the access transistor less array design (Fig. 8.8), achieved due to the unique integrated back gate feature. The lower bit-cell area leads to reduced metal-line capacitances (for word-lines/bit-lines) in the memory array. This feature, along with other properties of the VSH effect, enhances the energy efficiencies for memory operations for VSH/DVSH-MRAM as discussed next.

(ii) Write: The write metrics of the proposed VSH and DVSH-MRAMs remain similar because of the inherent and concurrent generation of I_{S+} and I_{S-} due to the VSH effect. However, the same property doesn't hold true for the GSH and DGSH-MRAMs because of different number of access transistors (one and two respectively) driving the write operation (Fig. 8.1). Our analysis shows that VSH/DVSH-MRAMs achieve 59%/ 67% lower write energy (WE) and 50%/ 11% lower write time (WT) compared to the GSH/DGSH-MRAM. This is attributed to two factors. First, the unique generation of out-of-plane spin currents with VSH-effect enables the switching of PMA magnets, unlike GSH effect which can only switch IMA magnets. It is well established that IMA switching is relatively less energy-efficient than PMA switching due to

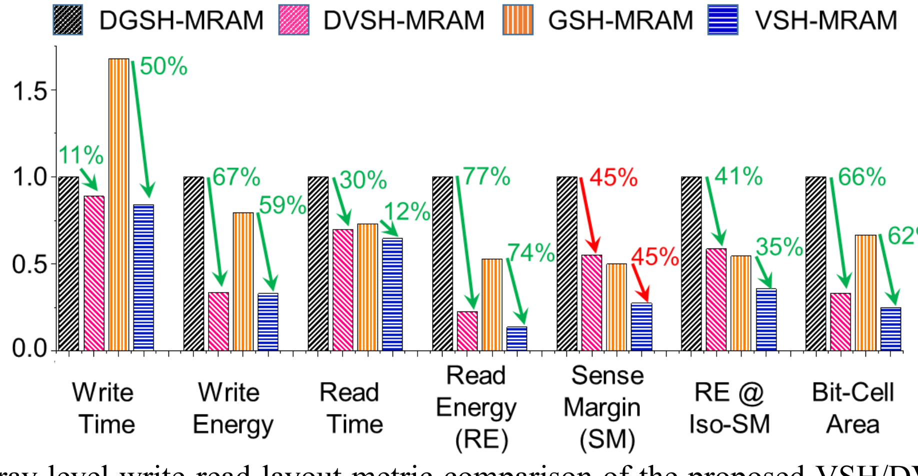


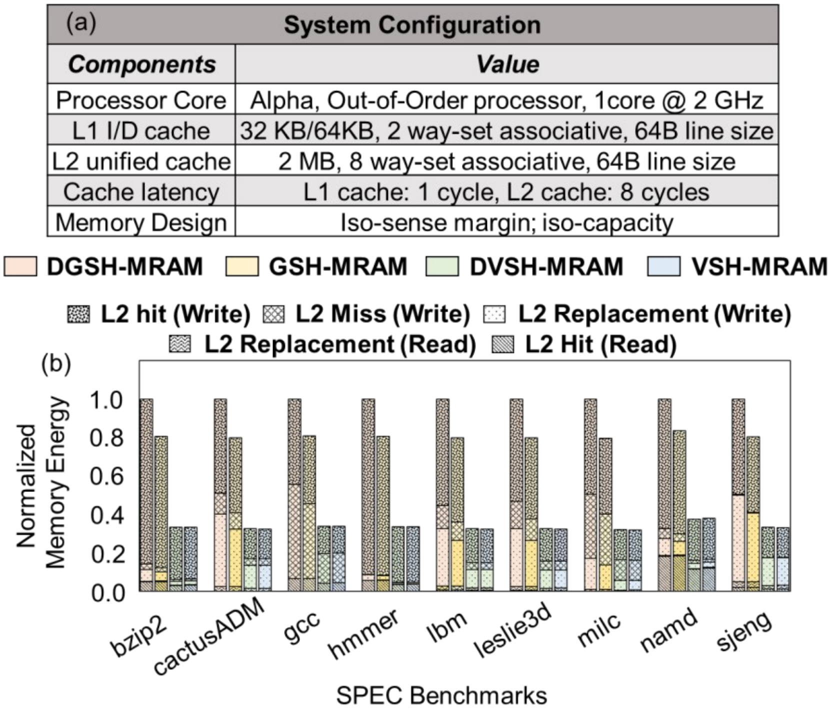
Fig. 8.9 Array level write-read-layout metric comparison of the proposed VSH/DVSH-MRAMs with GSH/DGSH-MRAMs (normalized). *Note: Iso-SM analysis has been performed individually for single-ended and differential designs.*

demagnetization fields [261]. Second, lower cell area in the proposed memories results in reduced time and energy consumption for bit-line charging/dis-charging during the write operation.

(iii) Read: The PMA MTJs in VSH/DVSH-MRAMs exhibit higher resistance due to its smaller area compared to IMA MTJs in GSH/DGSH-MRAMs at iso-energy barrier (Table. 8.2). Moreover, the WSe_2 FET is more resistive than a silicon-based FET used in (D)GSH-MRAMs due to lower mobility. This results in lower sensing currents in VSH/DVSH-MRAMs during the read operation. At the same time, lower area of the proposed memory array due to the integrated back gate feature reduces the bit-line charging/dis-charging energy. Both these factors lead to 74%-77% lower read energy consumption in the proposed memories. However, the lower sensed currents result in 45% lower sense margin for VSH/DVSH-MRAMs compared to GSH/DGSH-MRAMs, at V_{READ}=0.4V. At iso-sense margin (achieved by reducing V_{READ} for (D)GSH-MRAM to 0.15V), 35%/ 41% lower read energy is achieved by VSH/DVSH-MRAMs. The read time of VSH/DVSH-MRAMs is 12%/30% lower than GSH/DGSH-MRAMs. Even though (D)VSH-MRAMs are more resistive in nature (resulting in lower read currents) which leads to longer sensing delays at the sense amplifier, the major benefits come from lower metal-line charging delays due to lower area attained by the proposed array design, which results in overall reduction of the read time.

With respect to the single-ended VSH-MRAMs, differential DVSH-MRAMs exhibit 50% improved sense margin with a penalty of 64% increase in read energy, attributed to the additional sense-line (SLB) charging energy. However, at iso-sense margin, achieved by reducing V_{READ} of DVSH-MRAM to 0.2V, we observe similar read energies for VSH and DVSH-MRAMs.

(b) System-level analysis: With the understanding of the array-level benefits and trade-offs for the proposed memories, we now evaluate the application-level memory energy benefits of the proposed (D)VSH-MRAMs compared to the existing (D)GSH-MRAMs in the context of (a) general purpose processor (in collaboration with Prof. Anand Raghunathan and Dr. Shubham Jain) and (b) intermittently-powered system (in collaboration with Prof. Vijay Raghunathan and Dr. Arnab Raha).



(a)

System Configuration	
Components	***Value***
Processor Core	Alpha, Out-of-Order processor, 1core @ 2 GHz
L1 I/D cache	32 KB/64KB, 2 way-set associative, 64B line size
L2 unified cache	2 MB, 8 way-set associative, 64B line size
Cache latency	L1 cache: 1 cycle, L2 cache: 8 cycles
Memory Design	Iso-sense margin; iso-capacity

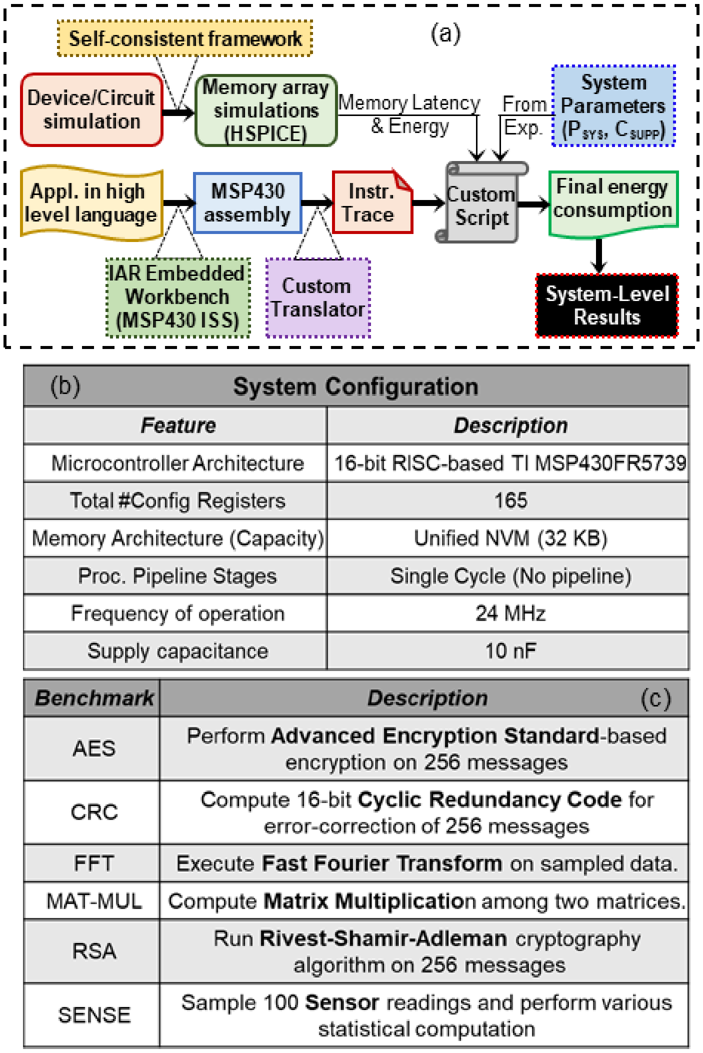
Fig. 8.10 (a) System configuration used in the general-purpose processor-based system analysis (b) Normalized memory energy for various SPEC benchmarks for DGSH, GSH, DVSH and VSH-MRAMs.

(i) General purpose systems: We evaluate the system-level benefits of the proposed VSH-MRAM and DVSH-MRAM designs when used as an L2 cache (unified memory) in a general-purpose processor. Fig. 8.10(a) details the system configuration, wherein we design a 2-MB, 8-way set- associative cache using the baseline (GSH-MRAM and DGSH-MRAM) and the proposed (VSH-MRAM and DVSH-MRAM) memory designs. We use gem5 [282], a cycle-accurate architecture simulator, to perform the full-system simulation and generate memory access traces. We estimate the total L2 cache energy for baseline and proposed designs using the memory traces and the array-level energy results discussed in the previous sub-section.

Fig. 8.10(b) shows the normalized L2 cache energy for the baseline (GSH/DGSH) and proposed (VSH/DVSH) designs. It also shows the energy consumed by the major L2 cache operations, which are, reads during L2-hits, reads and writes during L2-replacements, and writes during L2-misses and L2-hits. Across a suite of SPEC2K6 benchmarks, VSH-MRAM and DVSH-MRAM exhibit similar L2 cache energy due to similar write and read energies (at iso-sense margin). In comparison with DGSH-

MRAM and GSH-MRAM, the proposed VSH-MRAMs and DVSH-MRAMs show 2.63-3.14X and 2.19-2.50X reduction in the total L2 cache energy, respectively. Further, the applications (e.g., milc) with a lower read/write ratio show higher benefits. This is because the proposed designs can perform writes far more efficiently compared to the baseline designs.

(ii) Intermittently Powered-Systems: Due to the tight energy constraints of intermittently powered systems, we choose the more energy-efficient design for GSH memory for this analysis (single ended GSH-MRAM consumes less energy than the differential design- Fig. 8.9). Also, for fair comparison, our analysis covers only VSH-MRAM (although both VSH and DVSH MRAMs show similar energy efficiency at iso-sense


(a)
Self-consistent framework
Device/Circuit simulation
Memory array simulations (HSPICE)
Memory Latency & Energy
From Exp.
System Parameters (P_{SYS}, C_{SUPP})
Appl. in high level language
MSP430 assembly
Instr. Trace
Custom Script
Final energy consumption
IAR Embedded Workbench (MSP430 ISS)
Custom Translator
System-Level Results

(b) System Configuration	
Feature	***Description***
Microcontroller Architecture	16-bit RISC-based TI MSP430FR5739
Total #Config Registers	165
Memory Architecture (Capacity)	Unified NVM (32 KB)
Proc. Pipeline Stages	Single Cycle (No pipeline)
Frequency of operation	24 MHz
Supply capacitance	10 nF

Benchmark	***Description*** (c)
AES	Perform **Advanced Encryption Standard**-based encryption on 256 messages
CRC	Compute 16-bit **Cyclic Redundancy Code** for error-correction of 256 messages
FFT	Execute **Fast Fourier Transform** on sampled data.
MAT-MUL	Compute **Matrix Multiplication** among two matrices.
RSA	Run **Rivest-Shamir-Adleman** cryptography algorithm on 256 messages
SENSE	Sample 100 **Sensor** readings and perform various statistical computation

Fig. 8.11 (a) Simulation framework of IPS to evaluate the memory designs. (b) System configuration and (c) application benchmarks used for evaluation.

margin as discussed before). We use a simulation framework shown in Fig. 8.11(a), similar to the one used in Chapter-5. Our system-level simulations are based on the TI MSP430FR5739 microcontroller-based edge platform running at 24MHz [198] and use a unified 32kB NVM based on the proposed VSH-MRAM (with iso-sense margin of 1.85µA compared to the baseline GSH-MRAM; see Fig. 8.9). The system is powered using an energy harvesting source that charges a supply capacitor of 10nF. The system configuration and set of real benchmarks used are shown in Fig. 8.11(b, c). All results discussed below and showcased in Fig. 8.12 depict total memory energy consumption for iso-work conditions. Note, the energy numbers in Fig. 8.12 are normalized to GSH-MRAM energy consumption.

The energy savings obtained from using VSH-MRAMs compared to GSH-MRAMs depend primarily on the program characteristics, i.e., total number of reads and writes during program execution while executing a specific application. We

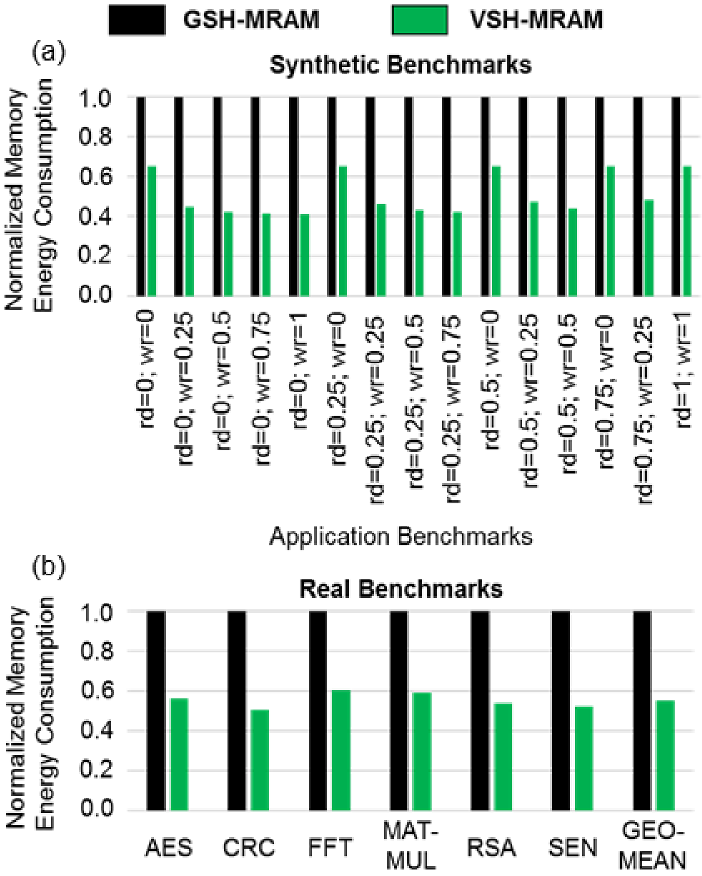


Fig. 8.12 Normalized system energy consumption of VSH and GSH-MRAM for (a) synthetic and (b) real application benchmarks.

constructed a set of synthetic benchmarks where we vary the fraction of total memory read and write instructions with a constant checkpoint size of 128B and total number of instructions (100K). Here, the expression {rd:0.25, wr:0.25} represents that 25% of the total instructions are memory reads, 25% are memory writes, and the rest are normal computational operations. In Fig. 8.12(a), we observe that the proposed VSH-MRAMs achieve significant energy benefits over GSH-MRAMs, ranging from 35% to 59% for a wide spectrum of synthetic memory instructions. This is attributed to the improved read-write energy (at iso-sense margins). For real application benchmarks, we observe that VHS-MRAMs exhibit energy savings in the range of 40% - 49% (1.66X-1.98X) and 45% (1.80X) on an average compared to GSH-MRAMs (Fig. 8.12(b)).

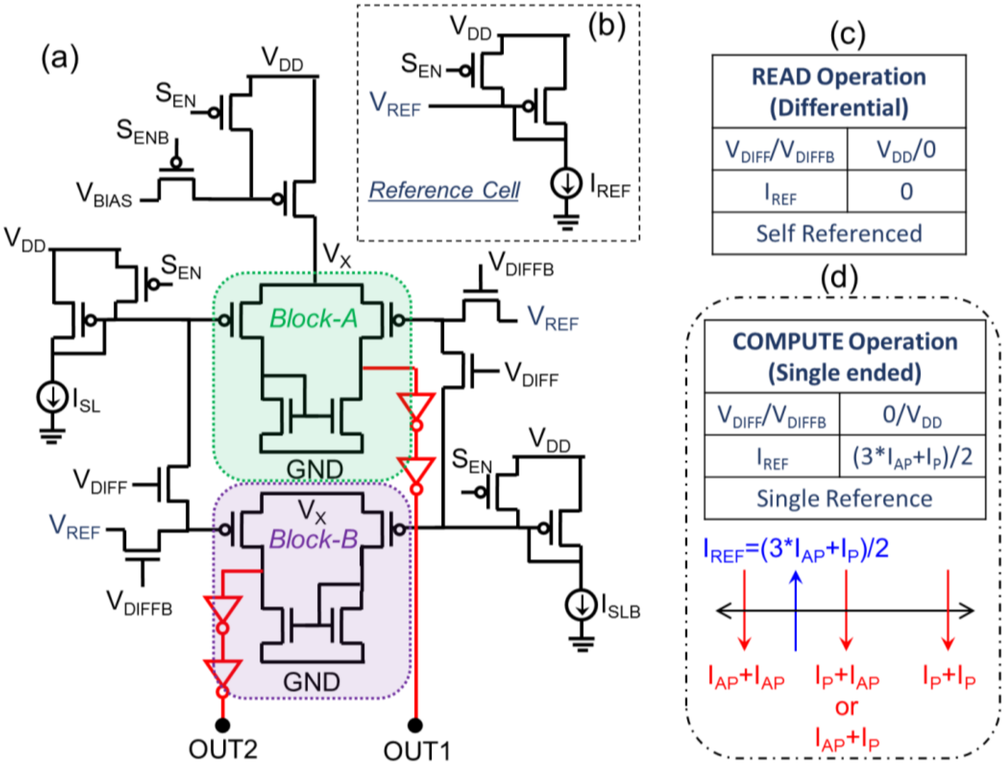
8.5 Boolean/Arithmetic Computation-in-Memory

In the previous sections, we discussed how the proposed device-circuit design techniques enable single-ended and differential memories which yield significant improvements in area and read/write energies compared to their GSH counterparts. In this section, we go beyond the standard memory operation and utilize the simultaneous true and complementary bit storage of the proposed DVSH-MRAM to enable energy efficient computation in memory (DVSH-MRAM: CiM). As discussed later, by utilizing the multi-wordline assertion (as proposed for R-FEFET/FEFET CiM proposed in Chapter-6), natural and simultaneous generation of bit-wise AND and NOR logic functions is achieved in DVSH MRAMs. Utilizing the outputs of these logics in conjunction with other logic gates, we propose a compact compute module to perform computation-in-memory of Boolean logic functions and arithmetic addition (ADD). We also evaluate the proposed single-ended VSH-MRAM for computation in-memory (VSH-MRAM: CiM) based on multi-word-line assertion [59]. To enable computations within the memory array along with standard memory operations, we present a reconfigurable sense amplifier which switches between memory and compute modes as discussed next.

8.5.1 Reconfigurable current sense amplifier

We present a current based reconfigurable sense amplifier which can dynamically switch its operation between differential sensing mode (for memory-read) and single ended sensing mode (for computation in-memory; discussed later). RCSA is designed for the DVSH-MRAM design where the complementary bit-storage can be efficiently harnessed to enable computation in memory. For VSH-MRAM, which is single-ended, we use a standard current-mirror based sense amplifier along with a reference current generation circuit.

The circuit diagram of the RCSA is shown in Fig. 8.13(a). It contains a pair of core amplifiers (block-A and block-B) based on current-mirroring. During the standard memory-read mode, the two amplifiers are connected together by applying V_{DIFF}=V_{DD} and V_{DIFFB}=0. This results in self-referenced differential sensing of the bit stored, resulting in OUT1 and OUT2 that correspond to the currents through SL and SLB. The reference generation circuit (Fig. 8.13(b)) is turned OFF in this mode. On the other hand, during the compute-in-memory mode, where bit-wise computations are carried out individually at SL and SLB (as discussed later), we apply V_{DIFF} (B) = 0 (V_{DD}) which decouples the two amplifiers. This results in individual single-ended sensing of SL and SLB based on the reference cell current.



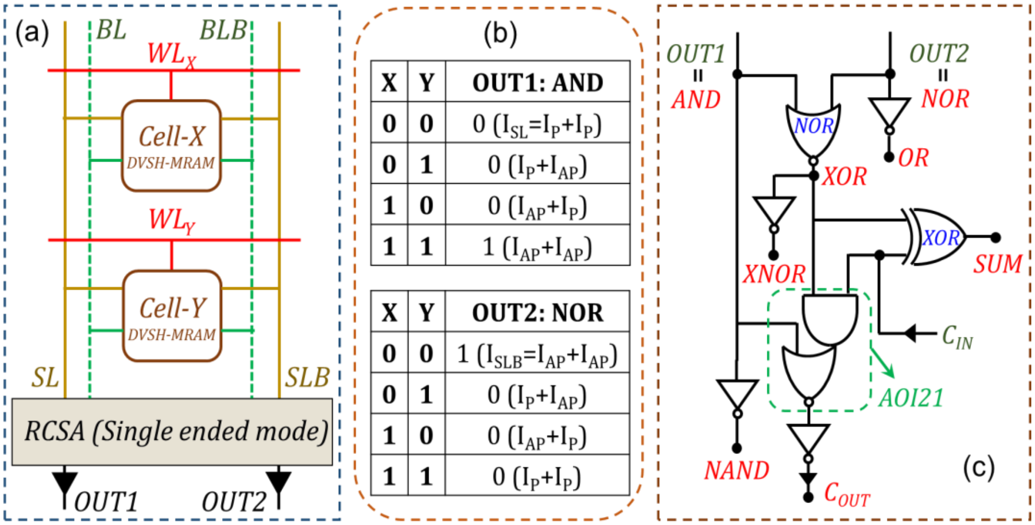
READ Operation (Differential)	
V_{DIFF}/V_{DIFFB}	V_{DD}/0
I_{REF}	0
Self Referenced	

COMPUTE Operation (Single ended)	
V_{DIFF}/V_{DIFFB}	0/V_{DD}
I_{REF}	$(3*I_{AP}+I_P)/2$
Single Reference	

Fig. 8.13 (a) Reconfigurable current sense amplifier along with (b) reference current generation. (c) Bias conditions for single ended and differential sensing. (d) Reference current used for in-memory compute.

8.5.2 Bit-wise AND and NOR logics

The compute operation in the proposed technique is based on the simultaneous assertion of two WLs (Fig. 8.14(a)) [64], [66], [219], [262]–[264]. The compute operation follows the same bias conditions as read operation of DVSH-MRAM with RCSA being operated in the single-ended mode (V_{DIFF}=0V). The reference current for the single-ended sensing during the compute operations is $I_{REF}=(3*I_{AP}+I_P)/2$ (Fig. 8.13(d)). Let us consider two bit-cells storing X: bit-‘0’ and Y: bit-‘1’. When WL_X and WL_Y are asserted (see Fig. 8.14(a)), the currents through SL, I_{SL} is equal to I_P+I_{AP} (from MTJ_R of X (P) and MTJ_R of Y (AP)) and that through SLB, I_{SLB} is equal to $I_{AP}+I_P$ (from MTJ_L of X (AP) and MTJ_L of Y (P)). Now, since $I_{SL}=I_{SLB}>I_{REF}$, OUT1 and OUT2 are brought to 0V. The truth table for all other input combinations (bit-information stored) is given in Fig. 8.14(b). Therefore, we naturally and simultaneously generate bit-wise AND (OUT1) and NOR (OUT2) logic functions at the two ends of the RCSA, with only one reference and without any additional circuitry. This is similar to previous proposals using SRAMs and ferroelectric NVMs for compute-in-memory [66], [219]. The major difference compared to [66], [219] is the use of differential spintronic devices for performing CiM operations and the implementation of the RCSA which enables dynamic switching between current based standard memory-read mode and compute-in-memory operation mode. The generated outputs, OUT1 and OUT2 are integrated with the compute module for the computation of other functions as discussed in the next sub-section.



X	Y	OUT1: AND
0	0	0 ($I_{SL}=I_P+I_P$)
0	1	0 (I_P+I_{AP})
1	0	0 ($I_{AP}+I_P$)
1	1	1 ($I_{AP}+I_{AP}$)

X	Y	OUT2: NOR
0	0	1 ($I_{SLB}=I_{AP}+I_{AP}$)
0	1	0 (I_P+I_{AP})
1	0	0 ($I_{AP}+I_P$)
1	1	0 (I_P+I_P)

Fig. 8.14 (a) Example of multi-word line assertion (b) Truth tables for the natively generated AND (OUT1) and NOR (OUT2) functions (c) compute module attached to RCSA for Boolean logic and ADD.

For VSH-MRAM: CiM, we use a single ended current sense amplifier as in [59] and utilize two reference current schemes i.e., I_{REF}= (3*I_{AP}+I_P)/2 and (I_{AP}+3*I_P)/2 for achieving the logics AND and OR similar to what has been discussed extensively for STT-MRAMs [59].

The advantages of having a differential memory functionality (like in DVSH-MRAM) compared to single- ended design are (a) the use of only one reference current source for performing the compute operations and (b) natural and simultaneous generation of AND and NOR logic at SL and SLB, as discussed above. These unique attributes lead to significant energy savings which is discussed later.

8.5.3 Compute model integrated with RCSA

In order to realize computing in-memory (as also discussed in Chapter-6), which includes bit-wise Boolean operations such as (N)AND, (N)OR, X(N)OR as well as arithmetic operations such as addition (ADD), we present a low power and compact compute module as shown in Fig. 8.14(c). The naturally generated bit-wise AND and NOR functions of DVSH-MRAM:CiM are simultaneously inverted using standard inverter to compute NAND and OR functions. Using AND and NOR as the input operands for a standard NOR logic gate, we achieve the bit-wise XOR function as shown in Fig. 8.14(c) which is also simultaneously inverted to achieve the XNOR function. We also implement an in-memory ripple carry adder (RCA) utilizing the bitwise Boolean operations discussed above along with three additional standard logic gates (Fig. 8.14(c)). The carry-out (C_{OUT}) from the previous stage is propagated as carry-in (C_{IN}) to the next stage. In our evaluations (next sub-section), we consider a 32-bit word where the C_{OUT} to C_{IN} routing is performed in compute module of the adjacent bit. For VSH-MRAM:CiM, we use the approach proposed in [59]. In the following sub-section, we evaluate the array and system-level implications of the CiM design for VSH and DVSH-MRAMs. For our baselines, we use the same methodology for compute-in-memory in GSH and DGSH-MRAMs, as discussed in the previous sub-section

8.5.4 Results

(a) Array-level analysis (Fig. 8.15): Similar to the analysis performed earlier, we evaluate a 1MB array for CiM. During compute operations, due to (a) lower currents through the sense line and (b) lower charging/discharging energy of the bit/sense-lines (similar to read

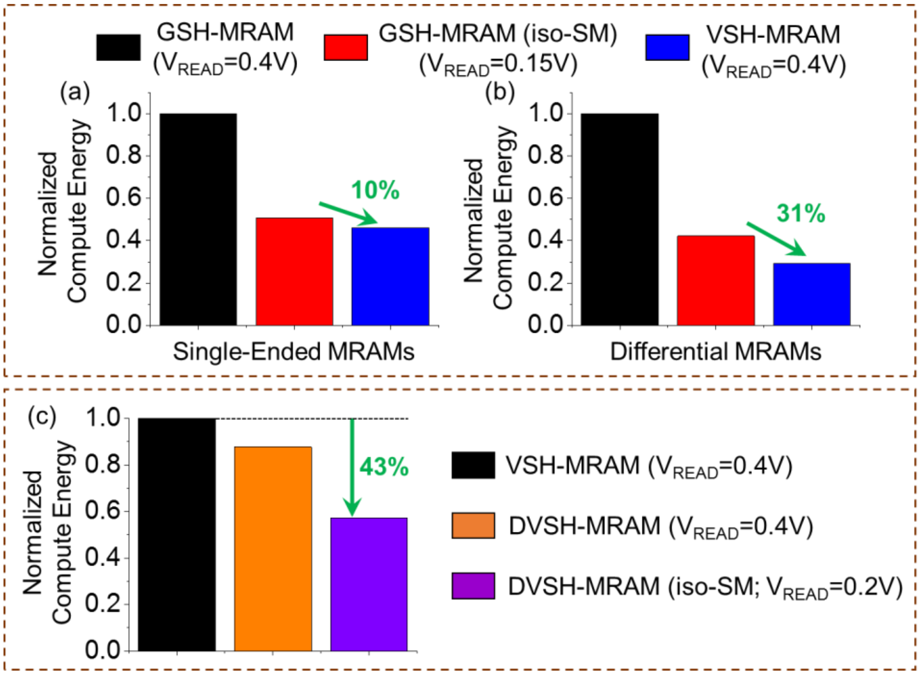


Fig. 8.15 Array-level normalized compute energy for (a) single-ended and (b) differential memory designs based on VSH and GSH effect. (c) Normalized compute energy consumption of VSH/DVSH-MRAMs.

operation discussed earlier) the proposed VSH/DVSH-MRAM: CiM exhibits is 54%/ 71% lower compute energy consumption for ADD operation when compared to GSH/DGSH-MRAM: CiM design. The lower sensed currents for compute operations are attributed to (a) the higher resistance of the PMA MTJs over IMA MTJs at iso-thermal energy barrier (~55KBT) and (b) higher resistance of 2D TMD channel compared to Silicon FET. However, due to this, the sense margin for compute in the proposed VSH/DVSH-MRAMs is 45% lower compared to GSH/D-GSH MRAMs. At iso-sense margin (achieved by reducing V_{READ} (to 0.15V) for GSH/DGSH-MRAM: CiM), the compute energy for ADD operation is 10%/31% lower for the proposed VSH/DVSH-MRAM: CiM. We also compare the proposed differential DVSH-MRAM: CiM design with single-ended VSH-MRAM: CiM and observe that the former achieves 43% lower compute energy consumption at iso-sense margin. This is mainly attributed to (a) natural and simultaneous generation of AND and NOR functions and (b) single current reference for all logic operations.

Due to the superior energy efficiency of differential CiM architectures, we omit the analysis of single-ended VSH/GSH-MRAM CiM design in our system level evaluations (discussed in the next sub-section). Before, we move on, it is important to establish the

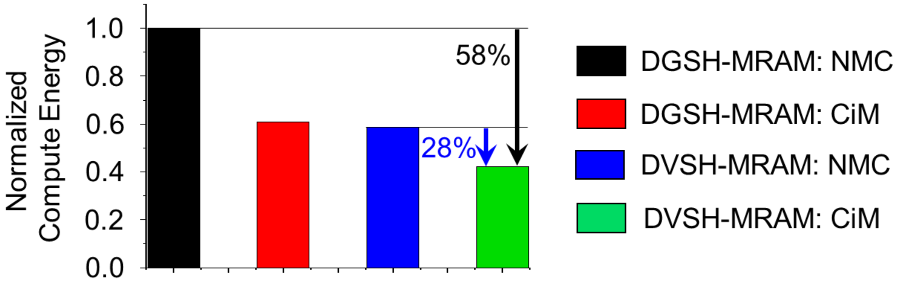


Fig. 8.16 Array-level normalized compute energy using near-memory compute (NMC) and compute in-memory (CiM) architectures for DGSH-MRAM and DVSH-MRAM.

benefits of CiM over a standard near-memory compute (NMC) architecture where we perform two read operations and then compute in a near-memory logic module. Fig. 8.16 illustrates the benefits of CiM design over NMC for addition. DVSH-MRAM: CiM exhibits 28% energy benefits compared to DVSH-MRAM: NMC. Compared to DGSH-MRAM: NMC, the proposed DVSH-MRAM:CiM shows 58% energy benefit. In the following, we evaluate the proposed CiM design at the system-level

(b) System-level analysis (Fig. 8.17): We follow the system-level framework used in Chapter-6 (in collaboration with Prof. Anand Raghunathan and Dr. Shubham Jain) for our

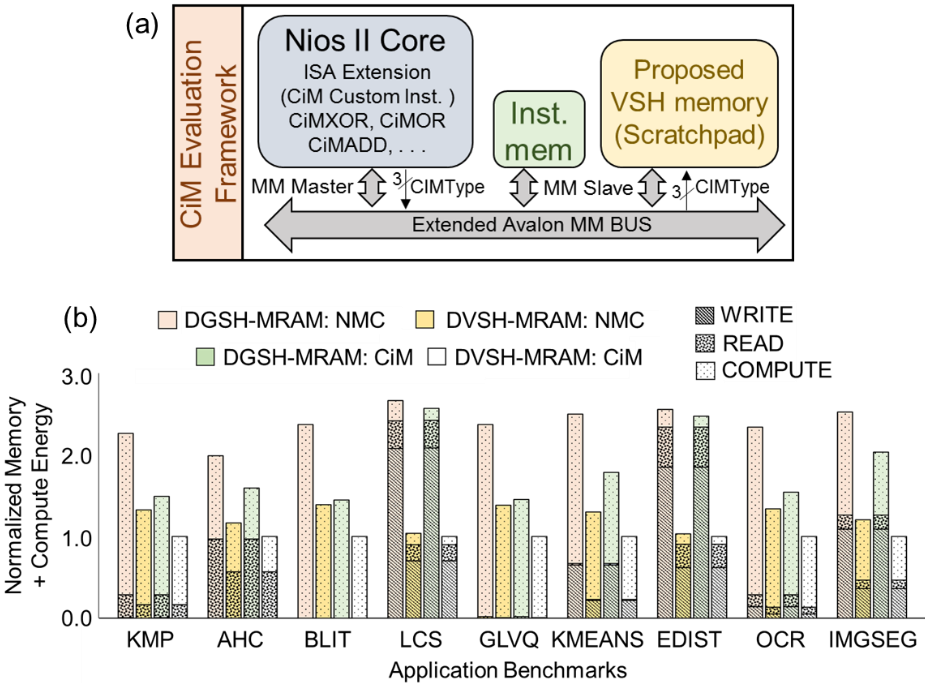


Fig. 8.17 (a) Simulation framework for system-level evaluation (b) total system energy consumption of the proposed DVSH-MRAM: CiM in comparison with DGSH-MRAM: CiM and DVSH/DGSH-MRAM: NMC for various application benchmarks.

evaluations, wherein the proposed DVSH-MRAM: CiM architectures are integrated as a 1-MB scratchpad for the Intel Nios II processor. To expose CiM operations to software, we add custom instructions to the Nios II processor's instruction set, as discussed in detail in [59]. We also extend the Avalon on-chip bus to support CiM operations [59]. Using the array-level results, we estimate the system-level memory energy benefits. We compare DVSH-MRAM:CiM with respect to DGSH-MRAM:CiM, DVSH-MRAM:NMC and DGSH-MRAM:NMC, all with iso-capacity. Further, we design all memories with iso-sense margin of 3.7μA as discussed earlier (see Fig. 8.9).

We present the total memory energy benefits for various application benchmarks [59] in Fig. 8.17 (b). We show all components corresponding to write, read and compute operations for the given application set. We observe that the proposed DVSH-MRAM: CiM achieves total system energy savings of 2.00X to 2.66X over the DGSH-MRAM: NMC, 1.04X to 1.39X over DVSH-MRAM: NMC and 1.45X to 2.57X over DGSH-MRAM: CiM. The benefits primarily arise due to energy- efficient compute operations along with superior read and write operations due to the unique attributes of VSH effect in the propose DVSH-MRAMs. CiM operations reduce memory accesses, bus transfers and processor instructions leading to significant energy savings when compared to the NMC baselines.

8.5.5 Other CiM architectures using (D)VSH-MRAMs

The proposed memory designs can also be implemented with other existing spin-based CiM techniques as discussed in several works [264], [283]–[285], to enhance the energy efficiency with respect to GSH-based memories. Let us broadly classify some of the existing techniques into three major categories: (i) Design-1: Spin-based CiM where input operands are stored in the memory array and multi-wordline assertion-based sensing operation is used to compute Boolean logic. This approach was presented in the previous sub-sections in this chapter and by other works in [59], [262], [263]. (Note, similar approach is followed in Chapter-6 using R-FEFETs and [66] using SRAMs). (ii) Design-2: Reconfigurable spin-logic based on control bits stored in the memory array as shown in [264], [283], which is similar to a computing a majority logic and (iii) Design-3: CiM using external signals such as current and voltages as input operands and Boolean logic is computed and stored in the spin-based memory [284], [285]. The benefits of VSH-MRAM in

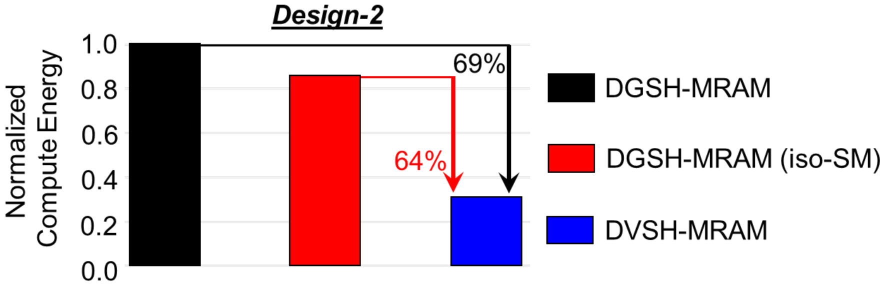


Fig. 8.18 Array-level normalized compute energy using Design-2 based CiM implementation for DVSH-MRAM and DGSH-MRAM.

Design-1 has been discussed in the previous sub-section. In the following, we discuss their implications in Design-2 and Design-3.

Design-2 is based on differential spin-based memories where CiM operations are performed without modifying the sense amplifier. This reduces the design complexity of the memory array and sensing operation. This design requires a control bit value to be written into the memory array every time before a logic is computed. Now, due to the superior write operation of our proposed DVSH-MRAM, implementing Design-2 with DVHS-MRAMs significantly reduces the energy consumption for in-memory computations, compared to DGSH-MRAMs. Based on our array-level analysis, CiM based on Design-2 shows 64% energy benefits for DVSH-MRAM compared to DGSH-MRAM design at iso-sense margin (Fig. 8.18).

Similar to Design-2, Design-3 also needs a write operation before every compute. It involves initialization of the memory state and writing of the Boolean logic into the ferromagnet, before the logic output is read out. Such an approach can be useful for applications where immediate addressing is required [286]. Voltage controlled magnetic anisotropy (VCMA) is used to achieve the desired in-memory computations. In principle, even the proposed VSH-memories can be employed in conjunction with VCMA proposed in [284], [285] to achieve CiM. However, the coupling of VSH effect with VCMA is yet to be understood comprehensively. This requires further experimental/theoretical studies. Therefore, a quantitative evaluation of Design-3 using VSH-MRAMs is outside the scope of this work. However, qualitatively, we do expect potential energy benefits for VSH-MRAMs over GSH-MRAMs when implemented in CiM architectures based on Design-3, mainly due to their energy efficient write operations.

8.6 Ternary Precision Non-Boolean Computation-in-Memory

As discussed in Chapter-7, several works have proposed CiM even for DNN workloads with various choices of precision [157], [228], [229], [234] . Most existing designs perform in-memory multiplication of binary operands [228]–[230], binary activations with ternary weights [234], or target higher-than-ternary precisions for analog vector-matrix multiplication [51]. Recently, a CMOS based ternary in-memory DNN architecture was proposed for pure signed ternary computation (ternary inputs and weights: '-1', '0', '+1') [233]. Also, in Chapter-7, we proposed a novel FEFET-based ternary compute enabled cell to perform in-memory computations in the signed ternary regime [287]. These (CMOS and FEFET based) approaches enable massively parallel signed ternary vector-matrix multiplications in a single array access, for efficient realization of ternary DNNs.

Although CMOS/FEFET-based ternary-precision CiM are promising for achieving energy/performance improvements compared to traditional CPU/GPU architectures, they might face some challenges. In 6T SRAMs, (a) coupling of read-write paths may lead to cell disturbances during CiM, (b) static leakage offsets the CiM efficiency gain and (c) large bit-cell area limits their capacity and bandwidth. FEFET-based approach is exciting due to their high density and non-volatility. However, gate-leakage, scalability and retention issues (due to depolarization fields) are some of the challenges which need to be addressed. Spin-based memories offer various benefits in terms of high density, large retention and high endurance. They have also been industrially adopted for large scale manufacturing. Therefore, if a ternary compute-enabled cell can be designed with spin-based NVMs, then one can harness richer advantages than the existing FEFET/CMOS implementations for DNN acceleration.

8.6.1 Ternary compute-enabled VSH Cell (TVC)

To enable ternary precision based in-memory computation, we propose a non-volatile ternary cell (TVC) containing 2 single ended VSH-MRAMs at its core, as shown in Fig. 8.19. The VSH-MRAMs store magnetizations M_A and M_B (for ternary storage). It also consists of 2 read access transistors attached to the MTJs of the respective cells (T1 and T2). Two additional transistors are cross-coupled with each other (T3 and T4; Fig. 8.19), following the technique presented in Chapter-7. Cross-coupled T3 and T4 (along with T1 and T2) enable in-memory ternary scalar

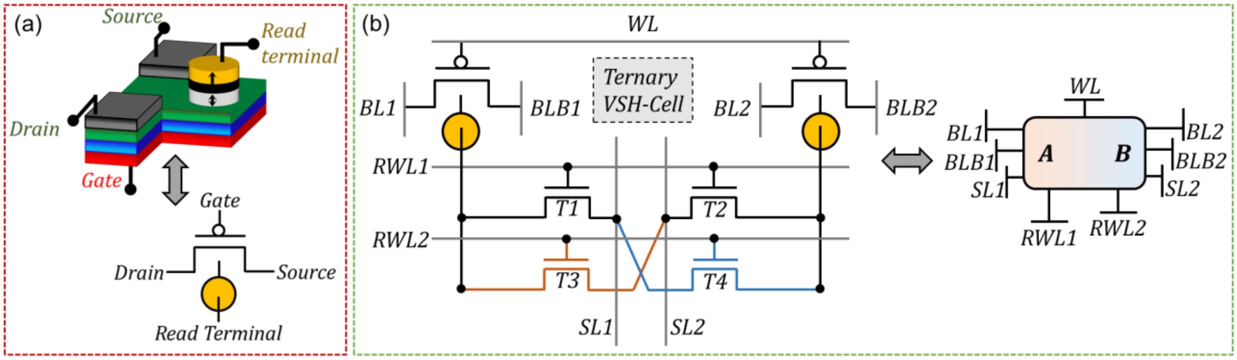


Fig. 8.19 (a) Schematic of VSH-MRAM and (b) proposed ternary compute-enabled VSH memory cell (TVC).

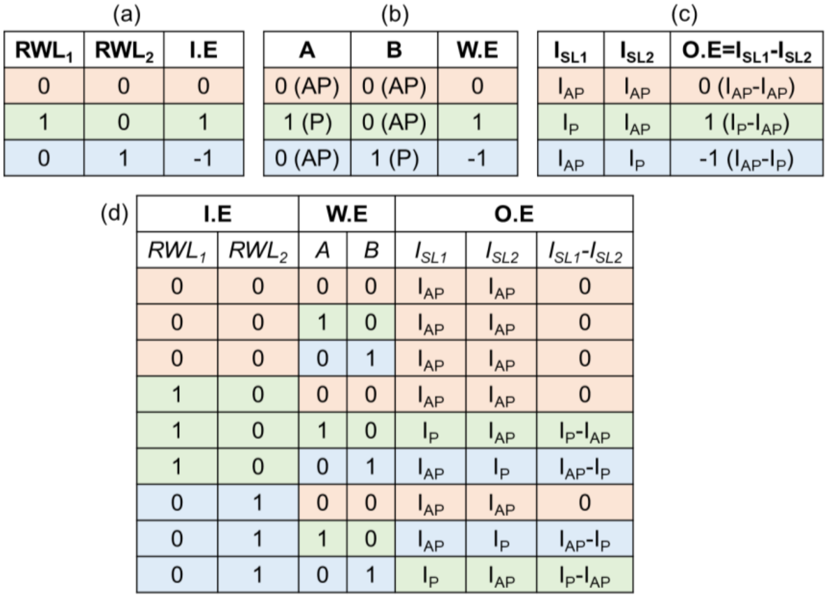
multiplication. Note, to reduce the area overhead, transistors T1, T2, T3 and T4 can be shared for multiple cells in a column, as discussed later.

For storing ternary data (storage/weight encoding in Fig. 8.20 (b)), we follow the same bias conditions used for VSH-MRAM and write M_A and M_B [277]. RWL1 and RWL2 are driven to 0V during write operation. Note, as mentioned earlier, magnetization stored in the MTJs corresponds to its resistance states (P: LRS; AP: HRS), which is used for the read operation. For sensing the ternary bit stored, the read word-line (RWL1) is asserted with the sense-lines (SL1 and SL2) driven to V_{DD}-V_{READ}. Rest of the biasing conditions remain the same as done for reading VSH-MRAMs. Note that, during read, RWL2 is always de-asserted.

8.6.2 In-memory ternary multiplication using TeC-Cell

In this section, we propose in-memory scalar multiplication of ternary weight (stored in the TVC) with ternary input to obtain a ternary output. Initially, the sense-lines (SL1 and SL2) are driven to V_{DD}-V_{READ}. BL1, BL2, BLB1, BLB2 are driven to V_{DD} with WWL driven to 0V. Similar to the discussion in Chpater-7, the ternary inputs are encoded as read word-line (RWL1 and RWL2) voltages as shown in Fig. 8.20(a). Depending on the ternary weight (encoded as M_A and M_B; see Fig. 8.20(b)), the final SL1 and SL2 currents represent the multiplication output (output encoding in Fig. 8.20(c)). We explain this further with an example, next.

When input I= +1 (RWL1=V_{DD}; RWL2=0V) and weight W= -1 (A=0; B=1), transistors T1, T2 are ON and T3, T4 OFF. This condition results in I_{AP} following at SL1 and I_P at SL2. This corresponds to output (O=I*W) = -1. Note that the output is inferred with the subtraction of currents at SL1 and SL2 (see Fig. 8.20(d)), which in this case is I_{SL1}-I_{SL2} = I_{AP}-I_P, corresponding



(a)

RWL_1	RWL_2	I.E
0	0	0
1	0	1
0	1	-1

(b)

A	B	W.E
0 (AP)	0 (AP)	0
1 (P)	0 (AP)	1
0 (AP)	1 (P)	-1

(c)

I_{SL1}	I_{SL2}	O.E=I_{SL1}-I_{SL2}
I_{AP}	I_{AP}	0 (I_{AP}-I_{AP})
I_P	I_{AP}	1 (I_P-I_{AP})
I_{AP}	I_P	-1 (I_{AP}-I_P)

(d)

I.E		W.E		O.E		
RWL_1	RWL_2	A	B	I_{SL1}	I_{SL2}	I_{SL1}-I_{SL2}
0	0	0	0	I_{AP}	I_{AP}	0
0	0	1	0	I_{AP}	I_{AP}	0
0	0	0	1	I_{AP}	I_{AP}	0
1	0	0	0	I_{AP}	I_{AP}	0
1	0	1	0	I_P	I_{AP}	I_P-I_{AP}
1	0	0	1	I_{AP}	I_P	I_{AP}-I_P
0	1	0	0	I_{AP}	I_{AP}	0
0	1	1	0	I_{AP}	I_P	I_{AP}-I_P
0	1	0	1	I_P	I_{AP}	I_P-I_{AP}

Fig. 8.20 (a) Input, (b) Weight and (c) Output encoding for the proposed TVC. (d) Truth table for all combinations of input and output.

to output '-1'. The same current conditions of SL1 and SL2 hold true for the case when I= -1 (RWL1=0; RWL2=1) and W= +1 (A=1; B=0) as shown in Fig. 8.20(d) and Fig. 8.21(a). Similarly, the truth table for all permutations can be derived is shown in Fig. 8.20(d). Note that the proposed TVC exhibits isolation of read-write paths and therefore, in-memory scalar multiplication has no effect on the information stored as magnetization in the VSH-MRAMs. Now that we understand scalar ternary multiplications in-memory let us discuss how TVCs enable in-memory ternary dot product computation for vector-matrix multiplication, next.

8.6.3 Ternary dot product computation

For dot product computation we simultaneously assert multiple read word-lines of TVCs present in a single column as illustrated in Fig. 8.22(a) [233]. The weight vector with ternary elements Wi is stored in the TVCs (as A/M_A and B/M_B), while the input vector with elements I_i is encoded using the voltages of RWL1_i and RWL2_i (Fig. 8.20(a)). Currents at SL1 and SL2 add up cumulatively. This is followed by subtraction of the two current I_{SL1} and I_{SL2}, which eventually results in a multiply-and-accumulate (MAC) operation. The multiple 'n' in the final subtracted currents: I_{SL1} - I_{SL2} = n*(I_P - I_{AP}) correspond as the vector dot product of the input and weight vectors. Flash analog-to-digital converters (ADCs) are employed to yield the digital value

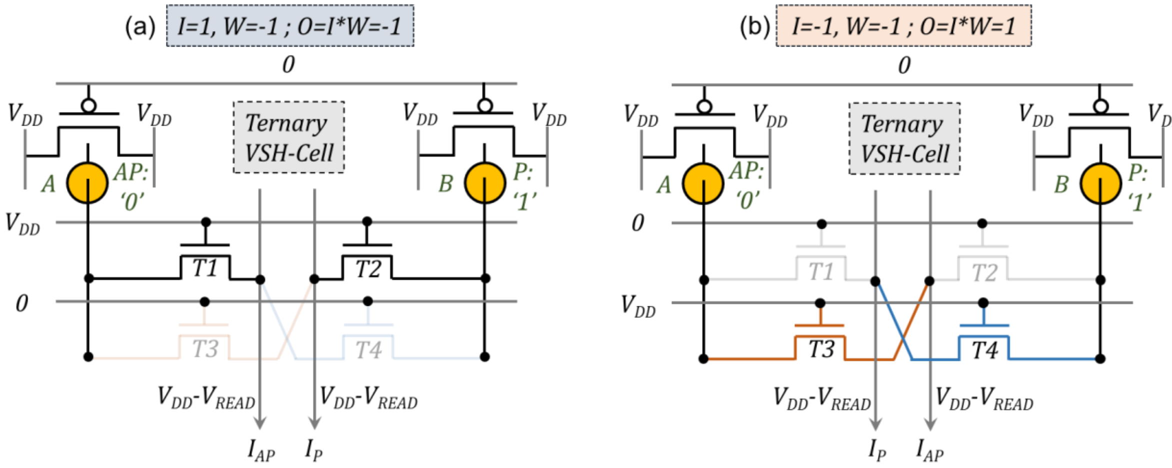


Fig. 8.21 Examples of scaler ternary precision in-memory multiplication for (a) I = 1, W = -1 and (b) I = -1, W = -1.

corresponding 'n'. The sign of 'n' is evaluated based on the comparator output (between I_{SL1} and I_{SL2}), which is fed to the subtractor and ADC (Fig. 8.22 (a, b)). Fig. 8.22(b, c, d) illustrates the sensing circuit required to realize the final dot product computation.

Now the questions is: *how many cells can be accessed together robustly, while performing the MAC operation?* This depends on the ADC precision and sparsity of input and weight vectors. Considering these two factors, and as discussed in Chapter-7 and [233], we assert N=16 cells simultaneous for performing ternary dot-product operation, while considering a 3-bit ADC. We design the TVC-Array based on this feature to perform massively parallel vector-matrix multiplication (or in-memory dot product computation) between ternary inputs and weights. Now, because we access only 16 ternary cells at once (in a column), this allows us to share the 4 sensing transistors (T1, T2, T3, T4) across blocks of bits in a column. For example, in a TVC-Array with K=1024 bits in a column, we have M=K/N=64 bits in a block, of which only one is asserted at a time and each block contains its own 4 transistor module. This technique can drastically reduce the area overheads of the TVC-Array. The value of N can be further increased, to support larger dot-product vector sizes. In our proposed TVC-Array, we have L=512 columns which can be accessed parallelly. Therefore, in one block access, the array can perform ternary multiplication of input vector I with N elements and weight matrix W of size N*L.

In order to perform ternary dot products on vector lengths N>16, we utilize the technique proposed in [233] of storing partial sums in a peripheral compute unit using a sample and hold circuitry. After multiple block accesses (in the same column), we accumulate all the partial sums

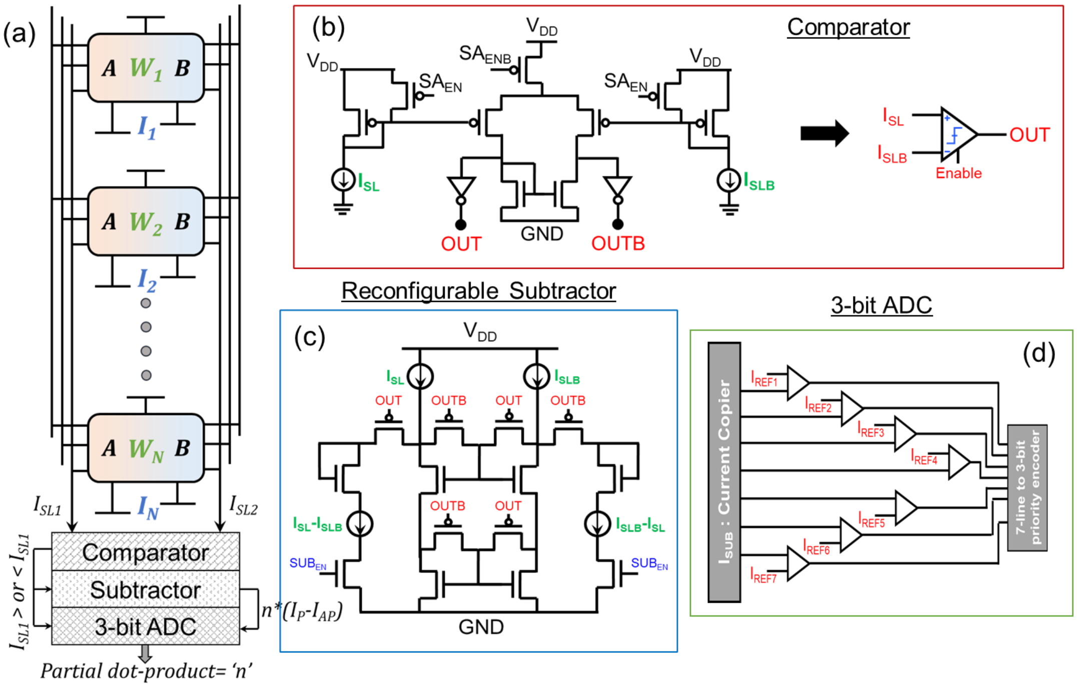


Fig. 8.22 (a) Schematic of a column in the proposed TVC-Array with peripherals. Schematics of (b) comparator, (b) reconfigurable subtractor and (d) 3-bit ADC designs used in the vector dot-product computation.

to determine the final dot products. The dot products are then quantized, and passed through an activation function to derive inputs to the next DNN layer. Moreover, we utilize only Q=64 PCUs for the entire array (where Q<L=512) in order to amortize area/energy overheads of the peripheral circuits.

8.6.4 Results

In this sub-section, we compare the write, read and MAC performance and energy of the proposed TeC-Array with respect to our baseline which is a near-memory ternary precision accelerator based on VSH-MRAMs, where the accelerator accesses scratchpad memories row-by-row before performing vector-matrix multiplication. We note that the gains shown for our design are pessimistic as we do not include the energy of the processing elements in the near-memory compute (NMC) baselines. All the memory arrays are designed with the same capacity (=1Mb).

The major advantage of the proposed TVC-Array is massively parallel in-memory computation of ternary dot-products. This results in 84% energy efficiency compared to our baseline with near-memory computing (NMC) as shown in Fig. 8.23(a). This is attributed to the

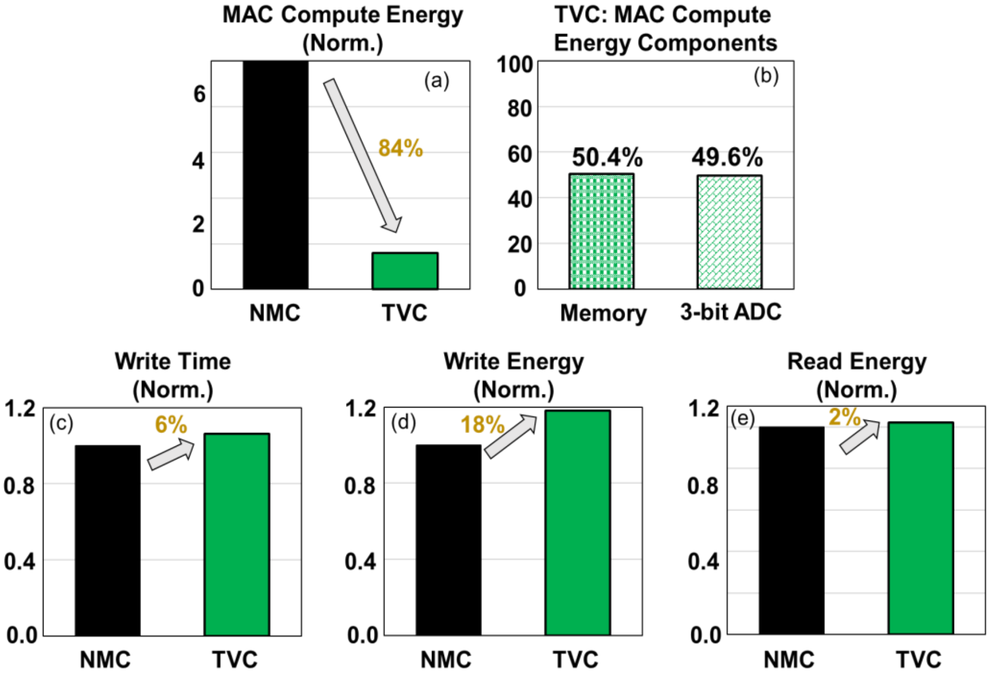


Fig. 8.23 Array-level results for (a) compute energy of MAC operation vs baseline near-memory computing design (NMC). (b) Split of energy between memory and ADC component for the TVC based design. (c)Write time, (d) write energy and (e) read energy evaluation of the proposed TVC based ternary-in-memory computing vs baseline NMC.

simultaneous assertion of multiple-word-lines unlike the near-memory compute baselines which require row-by-row access. The energy-split between array and ADC is illustrated in Fig. 8.23(b), which shows that ADC energy is nearly ~50% of the total compute energy at the array-level. This justifies the use of a 3-bit ADC since increasing the precision further would result in an exponential increase in ADC-energy [288]. Apart from the MAC energy benefits, it is important to note that the larger array area (resulting due to the 4T transistor modules) leads to 6%, 18% and 2% larger write time, write energy and read energy for the proposed TVC-Array compared to the baseline NMC (Fig. 8.23 (c, d, e)). However, as discussed in previous works on DNNs, the predominant contributor to energy is the MAC operation. Hence, the energy savings achieved at array-level for MAC in TVC is expected to translate to system-level energy efficiency as discussed in Chapter-7 and [233].

8.7 Summary

We proposed energy-efficient VSH effect based single-ended (VSH-MRAM) and differential (DVSH-MRAM) spintronic memory devices and their access-transistor-less non-volatile memory arrays. We developed a physics-based simulation framework in SPICE for the proposed VSH based memory devices and calibrated it with experiments. We employed this framework to design and evaluate our memory devices and arrays. At the array-level, the proposed VSH/DVSH-MRAMs achieve 50%/ 11% lower write time, 59%/ 67% lower write energy, 12%/30% lower read time and 35%/ 41% lower read energy at iso-sense margin, compared to single-ended/differential Giant-Spin Hall (GSH/DGSH)-MRAMs. System level evaluation in the context of general-purpose processor and intermittently-powered system shows up to 3.14X and 1.98X better energy efficiency for the proposed (D)VSH-MRAMs over (D)GSH-MRAMs respectively. We proposed computation of Boolean logic and arithmetic addition operations within the memory array (CiM) with simultaneous assertion of multiple word-line using the proposed DVSH-MRAM. We designed a reconfigurable current sense amplifier which can dynamically switch its operation between differential mode for memory read and single-ended sensing mode for in-memory compute in the proposed DVSH-MRAM. We also carried out system-level evaluation of the proposed DVSH-MRAM based CiM for various application benchmarks and observed up to 2.57X total system energy savings. In the end, we proposed ternary compute enabled memory cells utilizing VSH-MRAMs for DNN acceleration, which exhibited up to 84% energy efficiency at the array-level when compared to near-memory computing baseline.

9. CONCLUSION

9.1 Synopsis

Data seems to be the key for a broad spectrum of emerging workloads such as artificial intelligence, machine learning, genomics, particle physics experiments, etc. The challenge for all these applications is the requirement to rapidly and efficiently process large amounts of data. However, the problem and the crude reality is that, data is ever increasing and the human race is generating more data than what we can process. Apart from this, the conventional von-Neumann computing architectures suffer from the so-called memory bottleneck where the rapid emergence of data-intensive workloads have clogged the pipeline between processor and the memory subsystem. These challenges are even more critical for energy-constrained systems, such as energy autonomous platforms, IoT sensor nodes etc.

Fig. 9.1 illustrates the problem and a potential solution lucidly. To summarize, conventional approaches used for the current-era of computing systems involving GPUs and CPUs implemented in CMOS technologies are facing several roadblocks such as memory wall, Moore's wall and the

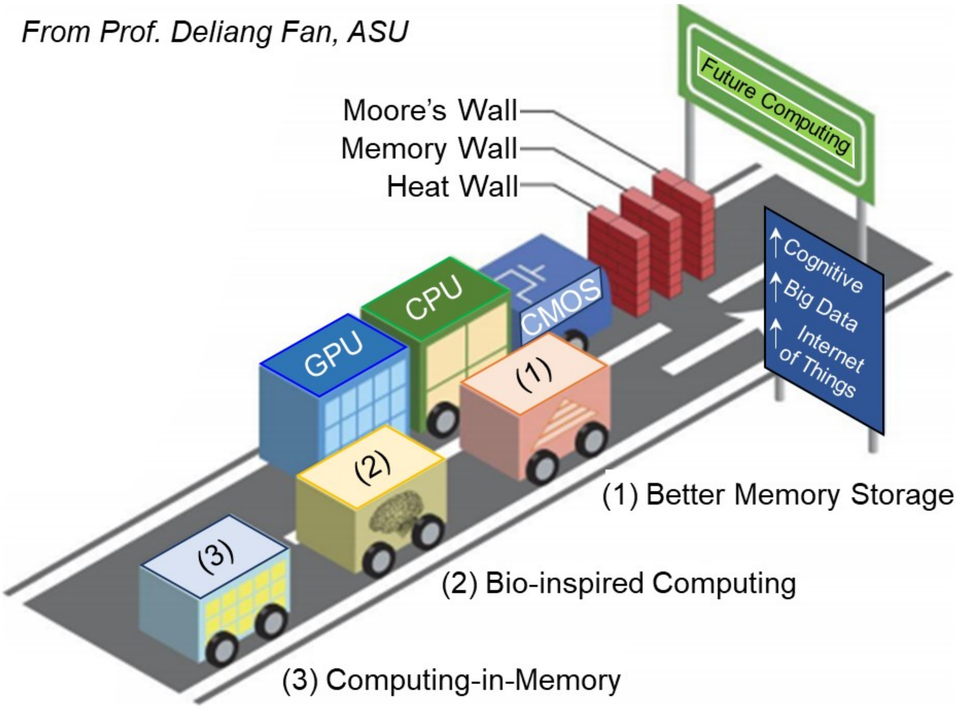


Fig. 9.1 Illustration of the problem and solution to tackle the challenges of future computing. Adapted from [297].

heat wall when processing future generation of data-intensive application workloads. To overcome these challenges as well as to cater to the new demands of emerging cognitive, big data and IoT tasks, bringing memory and logic closer to each other is the key, and for that, there is a need to explore novel memory/storage technologies (inherently suitable for such logic-memory coupling) along with beyond von-Neumann computing architectures that can take advantage of such new devices.

To that end, in this dissertation, we proposed integrated non-volatile transistor technologies namely, reconfigurable ferroelectric transistor (R-FEFET) and valley-coupled-spin (VSH) memory device, which exhibit efficient logic-memory coupling. Utilizing these novel devices, we proposed efficient techniques catered towards different class of applications, wherein the key objective is to utilize the unique attributes of the proposed devices to overcome the memory bottleneck. We explored novel device-circuit design techniques targeted towards non-volatile computing for intermittently powered systems. We also presented compute-enabled memories, where processing is performed within the memory arrays resulting in extreme energy efficiency and performance boost for various big-data workloads, when compared to the conventional computing systems. In the following section, we briefly summarize the major contributions of this dissertation.

9.2 Dissertation Summary

9.2.1 Integrated non-volatile transistor technologies

To address the critical challenges associated with the current technologies to meet the demands of emerging applications, we proposed two variants of integrated non-volatile transistors, which exhibit unique features of tight logic-memory coupling. First, we proposed a reconfigurable ferroelectric FET (R-FEFET), which can dynamically reconfigure its operation between volatile and non-volatile modes during run-time by modulating its hysteresis window. The R-FEFET comprised of two ferroelectric (FE) stacks, one modulated by the gate terminal and the other by the control terminal. We showed that by changing the control voltage between 0 and a high voltage, the hysteresis width (HW) of polarization in gate stack can be dynamically modulated between a volatile (logic) and non-volatile (memory) mode of operation. We presented the unique device characteristics and performed extensive design analysis with respect to different parameters. We also showed the various advantages that R-FEFETs possess when compared to regular FEFETs, a few being (a) larger hold margins, (b) higher drive current strengths and (c) minimal impact due to gate leakage.

At the same time, considering the immense potential of spintronic devices which exhibit large endurance and high integration density, we propose another flavor of an integrated non-volatile transistor harnessing the unique phenomenon of valley-coupled-spin hall effect. This device is based on monolayer WSe_2, where time-reversal symmetry along with broken inversion symmetry results in the generation of transverse, out-of-plane spin currents in the presence of a longitudinal charge current. These spin currents were utilized to store information in perpendicular magnetic anisotropy-based magnets, which exhibits high performance and energy efficiency. We presented the unique device characteristics and how an applied voltage at an integrated back gate can modulate the charge and spin current flow along with the magnetization switching dynamics.

Due to such unique device features of the proposed R-FEFETs and VSH devices, they exhibit an immense potential to open new avenues for several applications including non-volatile computing and brain-inspired computing which were proposed as a part of this dissertation.

9.2.2 Non-volatile circuits and their system implications

Using the unique attributes of R-FEFETs as well as VSH devices, we proposed different variants of non-volatile memory (NVM) designs. We proposed single-ended 2T-R and 3T-R memories with R-FEFETs which offer significant power savings over FEFET based memories by virtue control terminal driven dynamic modulation of hysteresis. The highlight of the proposed memories is a single-phase unipolar voltage based write operation due to the enablement of a volatile mode in R-FEFETs (which is challenging to achieve with standard FEFETs). Our analysis shows that the proposed R-FEFET memories exhibit up to 55% lower write power and 37-72% lower read power at iso access time and 33% lower area compared to existing FEFET based memory designs.

Similarly, using the VSH devices, we proposed two variants of VSH memories: single ended VSH-MRAM and differential DVSH-MRAM. The key features of these memories included (a) the ability to switch magnets with perpendicular magnetic anisotropy and (b) an integrated gate that can modulate the charge/spin current (I_C/I_S) flow. The former attribute resulted in high energy efficiency (compared to the Giant-Spin Hall (GSH) effect-based devices with in-plane magnetic anisotropy magnets). The latter feature leads to a compact access-transistor-less memory array design. The proposed VSH/DVSH-MRAMs achieved 50%/ 11% lower write time, 59%/ 67% lower write energy, 12%/30% lower read time and 35%/ 41% lower read energy at iso-sense margin, compared to existing single-ended/differential Giant-Spin Hall (GSH/DGSH)-MRAMs.

Along with the design of NVMs using R-FEFETs, we also design non-volatile flip-flops (NVFF) for intermittently-powered non-volatile processors. Utilizing the unique attributes of the R-FEFET device, we proposed two variants of NVFFs: (a) RNVFF-1 with completely automatic backup without the need of any external circuity and (b) RNVFF-2 with a need-based backup/restore module. While the former offers high back-up energy efficiency, the latter offers low normal operation energy. Our circuit-level analysis indicated up to 69% improvement in checkpointing (backup and restore) energy when compared to an existing FEFET based design. Using the proposed NVMs and NVFFs based on R-FEFET, we also explored the design of energy efficient intermittently powered systems. Our system-level evaluations demonstrated energy savings up to 40% in the context of a state-of-the-art IPS. We also explored the implications of the single-ended VSH-MRAM in the design of a unified memory system for IPS, which exhibited ~2X energy efficiency compared to their GSH counterparts, for various benchmarks.

9.2.3 Exploration of the proposed devices and circuits for advanced computing architectures

Apart from the exploration of the proposed devices in standard general-purpose computing architectures as well as targeted applications such as intermittently-powered systems, this dissertation also explored the attractive possibilities of utilizing the emerging FEFET, R-FEFET and VSH devices for advanced architectures in the field of brain-inspired computing. First, we showed how R-FEFETs can be exploited for the design of compute-enabled differential non-volatile memory, 4T-R. We presented a technique that enables natural computation of AND and NOR logic functions between two bits stored in the 4T-R array, with the assertion of two word-lines. Using this feature, we proposed a compute-in-memory (CiM) architecture involving the use of a compact compute module integrated to a sense amplifier which performs Boolean logic as well as arithmetic operations between two words with a single array access. Unlike existing non-volatile CiM designs, our proposals featured: (i) a self-referenced read operation due to differential access and (ii) a single universal voltage reference for all compute operations. System analysis performed by integrating our R-FEFET-CiM in the Nios II processor across various benchmarks showed total system energy savings of 24% and 14% compared to near-memory computing and previously-proposed FEFET-CiM, respectively. Furthermore, we also presented the possibility of utilizing the differential sensing in DVSH-MRAM to enable similar in-memory computation of Boolean and arithmetic functions, albeit by using a novel current-based sensing technique. The system-level analysis performed by integrating our DVSH-MRAM: CiM in the Nios II processor showed up to 2.57X total energy savings, compared to DGSH-MRAM: CiM.

We proposed another flavor of CiM engine which alleviates the memory-processor bottleneck and enhances energy-efficiency, specially designed for the transient computing workloads in IPS. We presented an FEFET-based memory architecture which supports (a) non-volatile memory (NVM) storage, (b) standard Boolean and arithmetic operations, (c) cyclic redundancy check for error detection and (d) edge-sensing for wireless sensory networks. Using the proposed CiM engine as a unified NVM, we constructed an integrated IPS-CiM architecture based on the TI MSP430 microcontroller system. We observed that IPS-CiM exhibited energy and performance benefits in the range of 35X-450X and 32X-400X respectively, over conventional microcontroller-based systems.

Lastly, we looked into another aspect of artificial intelligence or brain-inspired hardware in the context of deep neural networks (DNNs) which have gained significant attention in recent years. In particular, we designed low precision ternary DNN hardware, which employ signed ternary precision for weights and activations. Such an approach has shown immense promise due to their capability of energy efficiency close to that of binary networks with only a moderate loss of accuracy compared to the full-precision networks. We proposed a custom designed non-volatile ternary compute-enabled memory cell (TeC-Cell) based on FEFETs for in-memory computing. In particular, the proposed cell enables storage of ternary weights and employs multi-word-line assertion to perform massively parallel signed dot-product computations between ternary weights and ternary inputs. We evaluated the proposed design at the array level and showed 72% and 74% higher energy efficiency for multiply-and-accumulate (MAC) operations compared to standard near-memory computing designs based on SRAM and FEFET, respectively. Furthermore, we evaluated the proposed TeC-Cell in an existing ternary in-memory DNN accelerator and our results showed up to 3.4X reduction in system energy and up to 7X improvement in system performance over SRAM and FEFET based near-memory accelerators, across a wide range of DNN benchmarks including both deep convolutional and recurrent neural networks. We also showcased the possibility of designing ternary networks using the VSH-MRAMs and their benefits at the array-level when compared to a near-memory computing design. Moreover, the proposed technique of cross-coupling memory cells to achieve a TeC-Cell can also be implemented using several other memories based on CMOS and post-CMOS technologies.

In the following section, we discussed a few possible extensions to the research presented in this dissertation for the advancement of current and future generation of computing systems.

9.3 Future Outlook

As briefed in the previous section, in this dissertation we have explored a wide range of topics including novel devices and circuits for the emerging data-intensive applications. Going forward, it is important to continue the exploration of architectures with tight coupling between the processing and storage elements with the main goal of achieving energy efficiency and performance improvements for emerging workloads. As an extension to the research presented in this dissertation, the following two topics can further accelerate next-generation computing.

9.3.1 Development of low-precision AI specific hardware

To enable edge intelligence in current and future IoT devices, reducing the precision is a popular approach which alleviates the huge computational and storage costs associated with full precision architectures [289]. For instance, current state-of-the-art hardware employs 8-bit precision [290], [291]. This raises the prospect of reducing the data precision further to achieve energy savings. Although, lowering the precision inherently comes with the accuracy degradation, recent algorithmic advances have led to the achievement of acceptable accuracies with low precision inference for many cognitive tasks. In general, the regime of ultra-low precision (ULP) with 3 to 8 levels (2-3 bits) of precision, can be highly beneficial for any application demands [292]–[294]. However, their hardware implementation (especially for signed input/weight/output representations) is largely unexplored. To that end, it is important to explore novel device-circuit co-design solutions for ULP synaptic arrays with an aim to achieve high energy efficiency in conjunction with unique properties of various non-volatile technologies. In Chapter-7 of this dissertation, we proposed FEFET-based signed ternary synaptic arrays using cross-coupled FEFET bit-cells, which in concurrence with optimal encoding of weights, inputs and outputs (with values {-1, 0, +1}) elegantly performs in-memory dot product computation in the signed ternary regime.

Now, without tuning the ternary cell hardware proposed in Chapter-7 but modifying the input encoding scheme, where twice the pulse width of the RWL_1 and RWL_2 is encoded as {+2 and -2}, one can increase the input precision to 5-levels which can help in improving the accuracy when compared to the pure ternary design. On the other hand, if one can relax the constraints on the cell-footprint, we can exploit more from the same idea by incorporating multiple FEFET devices to achieve pure quaternary, quinary and other ULP networks, where the weight encoding is based on polarization stored in the FEFETs. All of the above-mention ideas can potentially improve the accuracy for various cognitive tasks when compared to the ternary networks proposed in Chapter-7, however they might come with additional area overhead and/or design complexities.

Another attractive alternative to realize a ULP hardware is by exploring the multi-domain behavior of ferroelectrics to achieve multiple resistance states corresponding to different weight levels in the synapse design without the need for increasing the cell footprint. This along with new circuit techniques (including peripheral circuit designs) to support multiple and programmable precision levels, one can achieve efficient ULP AI hardware. Moreover, in a similar fashion, other memory technologies can also be explored to achieve the benefits of ULP inference for the emerging DNN workloads.

9.3.2 Neural network fabric for approximate edge computing in intermittently powered systems

Advancements in deeply embedded battery-less IoT devices and wearables have led to their ubiquitous adoption in everyday life. They are powered by energy harvesting technologies and have found their application in medical implants, wildlife monitoring, defense sector, sensory networks, crop imaging and many more. Now, the next wave of IoT applications involve pushing inference of cognitive tasks to the edge, using DNNs. However, running artificial intelligence (AI) workloads and performing decision making on the edge has become extremely challenging in IPS due to their frequent power failures and severe resource constraints such as low bandwidth, insubstantial processing capabilities and limited on-chip memory [78], [194].

An attractive alternate approach to realize seamless implementation of DNNs in cost-constrained systems is the utilization of binary-precision data. Recent studies have demonstrated that aggressive reduction in precision to even 1-bit for achieving Binary Neural Networks (BNNs). This has enabled excellent performance and energy efficiency with controlled accuracy penalty for many cognitive tasks [74], [204], [206], [228]–[230], [232], [253], [295], [296]. Such an approach is tailor made for the resource-constrained IPS. However, due to the limitations on the on-chip memory capacity, even implementing small network topologies in IPS might be challenging. Apart from this, long-distance data transition between processor and memory also serves as the bottleneck in the traditional von-Neumann computing systems.

In Chapter-6, we showed how beneficial CiM-based techniques can be for IPS and their applications. As a research extension to this technique, an approximate BNN hardware accelerator targeted for IPS can be designed using the emerging FEFET/R-FEFET/VSH-based NVM cells proposed in this dissertation with extreme energy efficiency and minimal loss in accuracy. In-memory computing macros can be explored with the above-mentioned NVM cells, which perform low-power in-memory vector-matrix multiplications, especially in the unsigned binary precision domain {0, +1}, to evaluate approximate dot-products. Scalar multiplication in binary values chosen i.e., {0, +1} is nothing but a standard read operation which corresponds to AND logic operation between input and weight. This approach will require only one storage element for a synaptic weight. In contrast, previously proposed XNOR based binary precision dot-products use signed inputs and weights {-1, +1} which requires two NVM cells for storing one synaptic weight [228]–[230]. Therefore, this can help in reducing the overheads of memory capacity by half in a resource-constrained IPS. The output of the macros should undergo sensing operation using smart quantization techniques along with the optimal choice of ADC precision. After this, the accelerators can be designed to handle DNN operations which can be integrated with IPS for performing ultra-low-power edge-inference.

The future works proposed in this section gives insights on the directions for propelling the advancement of next-generation computing systems. There is a critical need to address the challenge of the humungous amounts of data being generated and that will be generated in the coming few years. One of the major solutions to this challenge as mentioned in this section is utilizing low precision for computing. Such an approach is being actively explored in the industry, especially for the data-intensive AI workloads [290]–[292]. To conclude, with the emergence of data-intensive applications, efficient devices, circuits and systems need to be designed using novel techniques with a radical shift in the computing paradigms to continue the marvelous achievements of the electronic industry.

REFERENCES

[1] Steven A. Przybylski, “Cache and memory hierarchy design: a performance-directed approach” *Morgan Kaufmann Publishers Inc*., 1990.

[2] H. M. Makrani *et al.*, “A Comprehen-sive Memory Analysis of Data Intensive Workloads on Server Class Ar-chitecture,” *MEMSYS*, p. 12, doi: 10.1145/3240302.3240320.

[3] R. M. Clapp, *et al.,* “Quantifying the Performance Impact of Memory Latency and Bandwidth for Big Data Workloads,” *2015 IEEE International Symposium on Workload Characterization*, doi: 10.13140/RG.2.1.2677.2562.

[4] W. A. Wulf and S. A. McKee, “Hitting the memory wall,” *ACM SIGARCH Computer Architecture News*, vol. 23, no. 1, pp. 20–24, Mar. 1995, doi: 10.1145/216585.216588.

[5] D. A. Patterson and J. L. Hennessy “Computer Architecture: A Quantitative Approach.” *Morgan Kaufmann Publishers Inc*., 1989.

[6] S. A. Mckee, “Reflections on the Memory Wall,” *Proceedings of the 1st conference on Computing frontiers*, 2004, doi: 10.1145/216585.216588.

[7] P. Machanick, “How Multithreading Addresses the Memory Wall.” School of IT and Electrical Engineering, University of Queensland,” Technical Report, 2002. (accessed Dec. 08, 2020).

[8] A. N. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, and N. P. Jouppi, “Rethinking DRAM Design and Organization for Energy-Constrained Multi-Cores,” *in the annual international symposium on Computer architecture (ISCA '10*), 175–186, 2010. doi: https://doi.org/10.1145/1815961.1815983.

[9] Y. Meng, T. Sherwood, and R. Kastner, “On the limits of leakage power reduction in caches,” in *Proceedings - International Symposium on High-Performance Computer Architecture*, 2005, pp. 154–165, doi: 10.1109/HPCA.2005.23.

[10] P. M. Zeitzoff and H. R. Huff, “MOSFET Scaling Trends, Challenges, and Key Associated Metrology Issues Through the End of the Roadmap,” *AIP Conference Proceedings*, vol. 788, p. 203, 2005, doi: 10.1063/1.2062964.

[11] P. M. Zeitzoff and J. E. Chung, “A perspective from the 2003 ITRS: MOSFET scaling trends, challenges, and potential solutions,” *IEEE Circuits and Devices Magazine*, vol. 21, no. 1, pp. 4–15, 2005, doi: 10.1109/MCD.2005.1388764.

[12] Z. Tarawneh, "The Effects of Process Variations on Performance and Robustness of Bulk CMOS and SOI Implementations of C-Elements," *PhD thesis,* School of Electrical, Electronic and Computer Eng., Newcastle University, 2011. (accessed Dec. 08, 2020).

[13] S. H. Lee, "Technology scaling challenges and opportunities of memory devices," in *Technical Digest - International Electron Devices Meeting, IEDM*, 2017, pp. 1.1.1-1.1.8, doi: 10.1109/IEDM.2016.7838026.

[14] "United States Data Center Energy Usage Report | Energy Technologies Area." https://eta.lbl.gov/publications/united-states-data-center-energy (accessed Dec. 08, 2020).

[15] H. Rong, H. Zhang, S. Xiao, C. Li, and C. Hu, "Optimizing energy consumption for data centers," *Renewable and Sustainable Energy Reviews*, vol. 58, pp. 674–691, 2016, doi: 10.1016/j.rser.2015.12.283.

[16] D. Reinsel, J. Gantz, and J. Rydning, "The Digitization of the World From Edge to Core," https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. (accessed Dec. 08, 2020).

[17] "DigiPlex plugs into district heating in Stockholm - DCD." https://www.datacenterdynamics.com/en/news/digiplex-plugs-into-district-heating-in-stockholm/ (accessed Dec. 08, 2020).

[18] "Iceland's data centers are booming—here's why that's a problem | MIT Technology Review." https://www.technologyreview.com/2019/06/18/134902/icelands-data-centers-are-booming-heres-why-thats-a-problem/ (accessed Dec. 08, 2020).

[19] "Facebook builds 'green' datacentre in Sweden | Environment | The Guardian." https://www.theguardian.com/environment/2011/oct/27/facebook-green-datacentre-sweden-renewables (accessed Dec. 08, 2020).

[20] "Hamina, Finland – Data Centers – Google." https://www.google.co.uk/about/datacenters/locations/hamina/ (accessed Dec. 08, 2020).

[21] A. Chen, "Emerging nonvolatile memory (NVM) technologies," in *2015 45th European Solid State Device Research Conference (ESSDERC)*, 2015, pp. 109–113, doi: 10.1109/ESSDERC.2015.7324725.

[22] M. Asadinia and H. Sarbazi-Azad, "Introduction to non-volatile memory technologies," in *Advances in Computers*, vol. 118, Academic Press Inc., 2020, pp. 1–13.

[23] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," *Solid-State Electronics*, vol. 125, pp. 25–38, 2016, doi: 10.1016/j.sse.2016.07.006.

[24] "ITRS Reports - International Technology Roadmap for Semiconductors." http://www.itrs2.net/itrs-reports.html (accessed Dec. 08, 2020).

[25] "Micron's X100 SSD." https://www.micron.com/products/advanced-solutions/3d-xpoint-technology/x100 (accessed Dec. 08, 2020).

[26] "Intel® Optane^{TM} Memory." https://www.intel.com/content/www/us/en/architecture-and-technology/optane-memory.html (accessed Dec. 08, 2020).

[27] B. C. Lee *et al.*, "Phase-change technology and the future of main memory," *IEEE Micro*, vol. 30, no. 1, pp. 131–141, 2010, doi: 10.1109/MM.2010.24.

[28] Y. K. Lee *et al.*, "Embedded STT-MRAM in 28-nm FDSOI logic process for industrial MCU/IoT application," in *Digest of Technical Papers - Symposium on VLSI Technology*, 2018, pp. 181–182, doi: 10.1109/VLSIT.2018.8510623.

[29] O. Golonzka *et al.*, "MRAM as Embedded Non-Volatile Memory Solution for 22FFL FinFET Technology," in *Technical Digest - International Electron Devices Meeting, IEDM*, 2019, pp. 18.1.1-18.1.4, doi: 10.1109/IEDM.2018.8614620.

[30] A. Aziz, N. Shukla, S. Datta, and S. K. Gupta, "COAST: Correlated material assisted STT MRAMs for optimized read operation," in *Proceedings of the International Symposium on Low Power Electronics and Design*, 2015, pp. 1–6, doi: 10.1109/ISLPED.2015.7273481.

[31] S. George *et al.*, "Nonvolatile memory design based on ferroelectric FETs," in *Proceedings of the 53rd Annual Design Automation Conference on - DAC '16*, 2016, pp. 1–6, doi: 10.1145/2897937.2898050.

[32] Y. Kim, S. H. Choday, and K. Roy, "DSH-MRAM: Differential Spin Hall MRAM for On-Chip Memories," *IEEE Electron Device Letters*, vol. 34, no. 10, pp. 1259–1261, Oct. 2013, doi: 10.1109/LED.2013.2279153.

[33] M. Kund *et al.*, "Conductive bridging RAM (CBRAM): An emerging non-volatile memory technology scalable to sub 20nm," in *Technical Digest - International Electron Devices Meeting, IEDM*, 2005, vol. 2005, pp. 754–757, doi: 10.1109/IEDM.2005.1609463.

[34] "ReRAM IP Cores for Embedded NVM in MCU & SOCs | Crossbar." https://www.crossbar-inc.com/products/p-series/ (accessed Dec. 08, 2020).

[35] "TSMC offers 22nm RRAM, taking MRAM on to 16nm." https://www.eenewsanalog.com/news/tsmc-offers-22nm-rram-taking-mram-16nm (accessed Dec. 08, 2020).

[36] "Fujitsu Launches an 8Mbit FRAM – the Largest Memory Density in the FRAM Family" https://www.fujitsu.com/global/products/devices/semiconductor/memory/reram/spi-8m-mb85as8mt.html (accessed Dec. 08, 2020).

[37] C. Xu *et al.*, "Overcoming the challenges of crossbar resistive memory architectures," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture, HPCA 2015*, Mar. 2015, pp. 476–488, doi: 10.1109/HPCA.2015.7056056.

[38] D. Takashima, "Overview of FeRAMs: Trends and perspectives," in *2011 11th Annual Non-Volatile Memory Technology Symposium Proceeding*, 2011, pp. 1–6, doi: 10.1109/NVMTS.2011.6137107.

[39] "MSP430FR5739 data sheet, product information and support | TI.com." https://www.ti.com/product/MSP430FR5739 (accessed Dec. 08, 2020).

[40] "Excelon^{TM} Ferroelectric-RAM (F-RAM): The Lowest-Power Nonvolatile Memory." https://www.cypress.com/products/excelon-fram (accessed Dec. 08, 2020).

[41] Y. Kato *et al.*, "Overview and Future Challenge of Ferroelectric Random Access Memory Technologies," *Japanese Journal of Applied Physics*, vol. 46, no. 4B, pp. 2157–2163, 2007, doi: 10.1143/JJAP.46.2157.

[42] S. Mueller *et al.*, "From MFM Capacitors Toward Ferroelectric Transistors: Endurance and Disturb Characteristics of HfO_2-Based FeFET Devices," *IEEE Transactions on Electron Devices*, vol. 60, no. 12, pp. 4199–4205, 2013, doi: 10.1109/TED.2013.2283465.

[43] S. Dunkel *et al.*, "A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 19.7.1-19.7.4, doi: 10.1109/IEDM.2017.8268425.

[44] T. S. Boscke, J. Muller, D. Brauhaus, U. Schroder, and U. Bottger, "Ferroelectricity in hafnium oxide: CMOS compatible ferroelectric field effect transistors," in *2011 International Electron Devices Meeting*, 2011, pp. 24.5.1-24.5.4, doi: 10.1109/IEDM.2011.6131606.

[45] S. K. Thirumala and S. K. Gunta, "Gate Leakage in Non-Volatile Ferroelectric Transistors: Device-Circuit Implications," in *2018 76th Device Research Conference (DRC)*, Jun. 2018, pp. 1–2, doi: 10.1109/DRC.2018.8442186.

[46] "Moving Data And Computing Closer Together." https://semiengineering.com/moving-data-and-computing-closer-together/ (accessed Dec. 08, 2020).

[47] S. Ghose, A. Boroumand, J. S. Kim, J. G Omez-Luna, and O. Mutlu, "Processing-in-memory: A workload-driven perspective," 2019, doi: 10.1147/JRD.2019.2934048.

[48] G. Singh *et al.*, "Near-Memory Computing: Past, Present, and Future," *in Microprocessors and Microsystems*, vol. 71, p. 102868, 2019, doi: 10.1016/j.micpro.2019.102868

[49] S. Bavikadi, P. R. Sutradhar, K. N. Khasawneh, A. Ganguly, and S. M. P. Dinakarrao, “A review of in-memory computing architectures for machine learning applications,” in *Proceedings of the ACM Great Lakes Symposium on VLSI, GLSVLSI*, Sep. 2020, pp. 89–94, doi: 10.1145/3386263.3407649.

[50] D. Ielmini and G. Pedretti, “Device and Circuit Architectures for In-Memory Computing,” *Advanced Intelligent Systems*, vol. 2, no. 7, p. 2000040, 2020, doi: 10.1002/aisy.202000040.

[51] X. Huang, C. Liu, Y. G. Jiang, and P. Zhou, “In-memory computing to break the memory wall,” *Chinese Physics B*, vol. 29, no. 7, p. 078504, 01, 2020, doi: 10.1088/1674-1056/ab90e7.

[52] J. D. Kendall and S. Kumar, “The building blocks of a brain-inspired computer,” *Applied Physics Reviews*, vol. 7, no. 1, p. 011305, 2020. doi: 10.1063/1.5129306.

[53] G. Karunaratne, M. le Gallo, G. Cherubini, L. Benini, A. Rahimi, and A. Sebastian, “In-memory hyperdimensional computing,” *Nature Electronics*, 2020, doi: 10.1038/s41928-020-0410-3.

[54] E. Riedel, C. Faloutsos, G. A. Gibson, and D. Nagle, “Active disks for large-scale data processing,” *Computer*, vol. 34, no. 6, pp. 68–74, 2001, doi: 10.1109/2.928624.

[55] Q. Zhu, K. Vaidyanathan, O. Shacham, M. Horowitz, L. Pileggi, and F. Franchetti, “Design automation framework for application-specific logic-in-memory blocks,” in *Proceedings of the International Conference on Application-Specific Systems, Architectures and Processors*, 2012, pp. 125–132, doi: 10.1109/ASAP.2012.21.

[56] M. Oskin and F. T. Chong, “Active Pages: A Computation Model for Intelligent Memory,” in *Proceedings of the 25th annual international symposium on computer architecture*, p. 192-203, 1998, doi: 10.1145/279358.279387.

[57] D. Patterson *et al.*, “Intelligent RAM (IRAM): Chips that remember and compute,” in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 40, pp. 224–225, 1997, doi: 10.1109/isscc.1997.585348.

[58] F. Gao *et al.,* “ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs,” in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture* (MICRO), p. 100–113, 2019, doi: https://doi.org/10.1145/3352460.3358260.

[59] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, “Computing in Memory With Spin-Transfer Torque Magnetic RAM,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 3, pp. 470–483, Mar. 2018, doi: 10.1109/TVLSI.2017.2776954.

[60] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, and Y. Xie, "Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories," in *Proceedings - Design Automation Conference*, 2016, vol. 05-09-June-2016, doi: 10.1145/2897937.2898064.

[61] "US20140334216A1 - General Structure for Computational Random Access Memory (CRAM) - Google Patents." https://patents.google.com/patent/US20140334216 (accessed Dec. 08, 2020).

[62] M. Kong, M. S. Keel, N. R. Shanbhag, S. Eilert, and K. Curewitz, "An energy-efficient VLSI architecture for pattern recognition via deep embedding of computation in SRAM," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014, pp. 8326–8330, doi: 10.1109/ICASSP.2014.6855225.

[63] P. Dlugosch, D. Brown, P. Glendenning, M. Leventhal, and H. Noyes, "An efficient and scalable semiconductor architecture for parallel automata processing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 12, pp. 3088–3098, 2014, doi: 10.1109/TPDS.2014.8.

[64] D. Reis, M. Niemier, and X. S. Hu, "Computing in memory with FeFETs," in *Proceedings of the International Symposium on Low Power Electronics and Design*, 2018, pp. 1–6, doi: 10.1145/3218603.3218640.

[65] S. Aga, S. Jeloka, A. Subramaniyan, S. Narayanasamy, D. Blaauw, and R. Das, "Compute Caches," in *Proceedings - International Symposium on High-Performance Computer Architecture*, 2017, pp. 481–492, doi: 10.1109/HPCA.2017.21.

[66] S. Jeloka, N. B. Akesh, D. Sylvester, and D. Blaauw, "A 28 nm Configurable Memory (TCAM/BCAM/SRAM) Using Push-Rule 6T Bit Cell Enabling Logic-in-Memory," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 1009–1021, 2016, doi: 10.1109/JSSC.2016.2515510.

[67] M. Kang, S. Gonugondla, and N. R. Shanbhag, *Deep In-memory Architectures for Machine Learning*. Springer International Publishing, 2020.

[68] J. Zhang, Z. Wang, and N. Verma, "In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, 2017, doi: 10.1109/JSSC.2016.2642198.

[69] A. Jaiswal, I. Chakraborty, A. Agrawal, and K. Roy, "8T SRAM Cell as a Multi-bit Dot Product Engine for Beyond von-Neumann Computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol.27, no.11, p. 2556-2567, 2019, doi: 10.1109/TVLSI.2019.2929245.

[70] H. Jiang, X. Peng, S. Huang, and S. Yu, “Cimat: A compute-in-memory architecture for on-chip training based on transpose sram arrays,” *IEEE Transactions on Computers*, vol. 69, no. 7, pp. 944–954, 2020, doi: 10.1109/TC.2020.2980533.

[71] Y. Long, E. Lee, D. Kim, and S. Mukhopadhyay, “Flex-PIM: A Ferroelectric FET based Vector Matrix Multiplication Engine with Dynamical Bitwidth and Floating-Point Precision,” *in International Joint Conference on Neural Networks (IJCNN),* 2020, doi: 10.1109/IJCNN48605.2020.9206672.

[72] Y. Long *et al.*, “A Ferroelectric FET-Based Processing-in-Memory Architecture for DNN Acceleration,” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 5, no. 2, pp. 113–122, 2019, doi: 10.1109/JXCDC.2019.2923745.

[73] X. Ma, L. Chang, S. Li, L. Deng, Y. Ding, and Y. Xie, “In-memory multiplication engine with SOT-MRAM based stochastic computing.”, *arXiv:*1809.08358, 2018.

[74] T. Ziegler, R. Waser, D. J. Wouters, and S. Menzel, “In-Memory Binary Vector–Matrix Multiplication Based on Complementary Resistive Switches,” *Advanced Intelligent Systems*, vol. 2, no. 10, p. 2000134, 2020, doi: 10.1002/aisy.202000134.

[75] A. Amirsoleimani *et al.*, “In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal–Oxide–Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives,” *Advanced Intelligent Systems*, vol. 2, no. 11, p. 2000115, 2020, doi: 10.1002/aisy.202000115.

[76] Y. Liao *et al.*, “Novel in-memory matrix-matrix multiplication with resistive cross-point arrays,” in *Digest of Technical Papers - Symposium on VLSI Technology*, pp. 31–32, 2018, doi: 10.1109/VLSIT.2018.8510634.

[77] K. Ma *et al.*, “Nonvolatile Processor Architectures: Efficient, Reliable Progress with Unstable Power,” *IEEE Micro*, vol. 36, no. 3, pp. 72–83, 2016, doi: 10.1109/MM.2016.35.

[78] Y. Liu *et al.*, “Ambient energy harvesting nonvolatile processors: From circuit to system,” in *Proceedings - Design Automation Conference*, 2015, doi: 10.1145/2744769.2747910.

[79] F. Su, K. Ma, X. Li, T. Wu, Y. Liu, and V. Narayanan, “Nonvolatile Processors: Why is it Trending?” in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, doi: 10.23919/DATE.2017.7927131.

[80] D. Wang, S. George, A. Aziz, S. Datta, V. Narayanan, and S. K. Gupta, “Ferroelectric Transistor based Non-Volatile Flip-Flop,” in *Proceedings of the 2016 International Symposium on Low Power Electronics and Design - ISLPED ’16*, 2016, pp. 10–15, doi: 10.1145/2934583.2934603.

[81] N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, “Nonvolatile Magnetic Flip-Flop for standby-power-free SoCs,” in *2008 IEEE Custom Integrated Circuits Conference*, Sep. 2008, pp. 355–358, doi: 10.1109/CICC.2008.4672095.

[82] Y. Wang *et al.*, “A 3us wake-up time nonvolatile processor based on ferroelectric flip-flops,” in *European Solid-State Circuits Conference*, 2012, pp. 149–152, doi: 10.1109/ESSCIRC.2012.6341281.

[83] C.-M. Jung, K.-H. Jo, E.-S. Lee, H. M. Vo, and K.-S. Min, “Zero-Sleep-Leakage Flip-Flop Circuit With Conditional-Storing Memristor Retention Latch,” *IEEE Transactions on Nanotechnology*, vol. 11, no. 2, pp. 360–366, 2012, doi: 10.1109/TNANO.2011.2175943.

[84] I. Kazi *et al.*, “Energy/Reliability Trade-Offs in Low-Voltage ReRAM-Based Non-Volatile Flip-Flop Design,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 11, pp. 3155–3164, 2014, doi: 10.1109/TCSI.2014.2334891.

[85] H. Kimura *et al.*, “A 2.4 pJ ferroelectric-based non-volatile flip-flop with 10-year data retention capability,” in *2014 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2014, pp. 21–24, doi: 10.1109/ASSCC.2014.7008850.

[86] J. Muller *et al.*, “Ferroelectricity in HfO2 enables nonvolatile data storage in 28 nm HKMG,” in *2012 Symposium on VLSI Technology (VLSIT)*, 2012, pp. 25–26, doi: 10.1109/VLSIT.2012.6242443.

[87] A. I. Khan, C. W. Yeung, C. Hu, and S. Salahuddin, “Ferroelectric negative capacitance MOSFET: Capacitance tuning & antiferroelectric operation,” *International Electron Devices Meeting, IEDM*, 2011, doi: 10.1109/IEDM.2011.6131532.

[88] F. Schwierz, J. Pezoldt, and R. Granzner, “Two-dimensional materials and their prospects in transistor electronics,” *Nanoscale*, vol. 7, no. 18, p. 8261–8283, 2015, doi: 10.1039/c5nr01052g.

[89] K. F. Mak, K. L. McGill, J. Park, and P. L. McEuen, “The valley Hall effect in MoS_2 transistors.,” *Science*, vol. 344, no. 6191, pp. 1489–92, 2014, doi: 10.1126/science.1250140.

[90] J. Zou, Y. Yuan and J. Kang “Spin and spin-valley Hall effects in a honeycomb lattice with antiferromagnetism and spin-orbit couplings,” Phy*sics Letters A,* vol. 383, no.25, p. 3162-3166, 2019, doi: 10.1016/j.physleta.2019.07.001.

[91] B. T. Zhou, K. Taguchi, Y. Kawaguchi, Y. Tanaka, and K. T. Law, “Spin-orbit coupling induced valley Hall effects in transition-metal dichalcogenides,” *Communications Physics*, 2, 26, 2019, doi: 10.1038/s42005-019-0127-7.

[92] Z. Wang, J. Shan, and K. F. Mak, “Valley- and spin-polarized Landau levels in monolayer WSe_2,” *Nature Nanotechnology*, vol. 12, 2017, doi: 10.1038/NNANO.2016.213.

[93] T. Y. T. Hung, K. Y. Camsari, S. Zhang, P. Upadhyaya, and Z. Chen, “Direct Observation of Valley Coupled Topological Current in MoS2.” *Science Advances*, Vol. 5, no. 4, 2019, doi: 10.1126/sciadv.aau6478.

[94] D. Xiao, G.-B. Liu, W. Feng, X. Xu, and W. Yao, “Coupled Spin and Valley Physics in Monolayers of MoS_2 and Other Group-VI Dichalcogenides,” *Physical Review Letters*, vol. 108, no. 19, p. 196802, May 2012, doi: 10.1103/PhysRevLett.108.196802.

[95] T. Schenk, M. Pešić, S. Slesazeck, U. Schroeder, and T. Mikolajick, “Memory technology- A primer for material scientists,” *Reports on Progress in Physics*, vol. 83, no. 8. Institute of Physics Publishing, p. 086501, Aug. 01, 2020, doi: 10.1088/1361-6633/ab8f86.

[96] A. I. Khan, A. Keshavarzi, and S. Datta, “The future of ferroelectric field-effect transistor technology,” *Nature Electronics*, vol. 3, no. 10, pp. 588–597, Oct. 2020, doi: 10.1038/s41928-020-00492-7.

[97] M. Trentzsch *et al.*, “A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs,” in *Technical Digest - International Electron Devices Meeting, IEDM*, Jan. 2017, pp. 11.5.1-11.5.4, doi: 10.1109/IEDM.2016.7838397.

[98] T. Mikolajick, U. Schroeder, and S. Slesazeck, “The Past, the Present, and the Future of Ferroelectric Memories,” *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1434–1443, Apr. 2020, doi: 10.1109/TED.2020.2976148.

[99] “Ferroelectrics for Digital Information Storage and Switching - Dudley Allen Buck.” https://books.google.com/books/about/Ferroelectrics_for_Digital_Information_S.html?id=ya0MHQAACAAJ (accessed Dec. 08, 2020).

[100] D. W. Bondurant and F. P. Gnadinger, “Ferroelectrics for nonvolatile RAMs,” *IEEE Spectrum*, vol. 26, no. 7, pp. 30–33, Jul. 1989, doi: 10.1109/6.29346.

[101] F. P. Gnadinger, “High speed nonvolatile memories employing ferroelectric technology,” 1989, doi: 10.1109/cmpeur.1989.93335.

[102] K. R. Udayakumar *et al.*, “Manufacturable high-density 8 Mbit one transistor-one capacitor embedded ferroelectric random access memory,” *Japanese Journal of Applied Physics*, vol. 47, no. 4 PART 2, pp. 2710–2713, Apr. 2008, doi: 10.1143/JJAP.47.2710.

[103] S. R. Summerfelt *et al.*, “High-density 8Mb 1T-1C ferroelectric random access memory embedded within a low-power 130nm logic process,” in *IEEE International Symposium on Applications of Ferroelectrics*, 2007, pp. 9–10, doi: 10.1109/ISAF.2007.4393151.

[104] “US2791760A - Semiconductive translating device - Google Patents.” https://patents.google.com/patent/US2791760A/en (accessed Dec. 08, 2020).

[105] J. L. Moll and Y. Tarui, “A new solid state memory resistor,” *IEEE Transactions on Electron Devices*, vol. 10, no. 5, pp. 338–338, Sep. 1963, doi: 10.1109/T-ED.1963.15245.

[106] Nobuhito Ogata and Hiroshi IshiwaraHiroshi Ishiwara, “A model for high frequency C-V characteristics of ferroelectric capacitors,” *IEICE Transactions on Electronics*, vol. E84-C, no. 6, pp. 777–784, 2001.

[107] A. K. Saha, S. Datta, and S. K. Gupta, "‘negative capacitance’ in resistor-ferroelectric and ferroelectric-dielectric networks: Apparent or intrinsic?,” *Journal of Applied Physics*, vol. 123, no. 10, p. 105102, Mar. 2018, doi: 10.1063/1.5016152.

[108] A. Aziz, S. Ghosh, S. Datta, and S. Gupta, “Physics-Based Circuit-Compatible SPICE Model for Ferroelectric Transistors,” *IEEE Electron Device Letters*, pp. 1–1, 2016, doi: 10.1109/LED.2016.2558149.

[109] M. H. Lee *et al.*, “Physical thickness 1.x nm ferroelectric HfZrOx negative capacitance FETs,” in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 12.1.1-12.1.4, doi: 10.1109/IEDM.2016.7838400.

[110] S. S. Cheema *et al.*, “Enhanced ferroelectricity in ultrathin films grown directly on silicon,” *Nature*, vol. 580, no. 7804, pp. 478–482, 2020, doi: 10.1038/s41586-020-2208-x.

[111] S. Salahuddin and and S. Datta, “Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices,” *Nano Lett.,* 8, 2, 405–410, 2008, doi: 10.1021/NL071804G.

[112] A. I. Khan *et al.*, “Negative capacitance in a ferroelectric capacitor,” *Nature Materials*, vol. 14, no. 2, pp. 182–186, 2015, doi: 10.1038/nmat4148.

[113] M. Kobayashi, N. Ueyama, K. Jang, and T. Hiramoto, “Experimental study on polarization-limited operation speed of negative capacitance FET with ferroelectric HfO2,” in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 12.3.1-12.3.4, doi: 10.1109/IEDM.2016.7838402.

[114] P. Sharma, J. Zhang, A. K. Saha, S. Gupta, and S. Datta, “Negative capacitance transients in metal-ferroelectric Hf0.5Zr0.5O2-Insulator-Semiconductor (MFIS) capacitors,” in 75th Annual Device Research Conference (DRC), 2017, doi: 10.1109/DRC.2017.7999477.

[115] J. Gomez *et al.*, “Hysteresis-free negative capacitance in the multi-domain scenario for logic applications,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2019, vol. 2019-December, doi: 10.1109/IEDM19573.2019.8993638.

[116] Z. Wang *et al.*, “Direct Observation of Stable Negative Capacitance in $SrTiO_3$ @$BaTiO_3$ Heterostructure,” *Advanced Electronic Materials*, vol. 6, no. 2, p. 1901005, 2020, doi: 10.1002/aelm.201901005.

[117] M. Hoffmann *et al.*, “Direct Observation of Negative Capacitance in Polycrystalline Ferroelectric HfO2,” *Advanced Functional Materials*, vol. 26, no. 47, pp. 8643–8649, 2016, doi: 10.1002/adfm.201602869.

[118] J. C. Wong and S. Salahuddin, “Negative Capacitance Transistors,” *Proceedings of the IEEE*, vol. 107, no. 1, pp. 49–62, 2019, doi: 10.1109/JPROC.2018.2884518.

[119] M. Si *et al.*, “Steep-slope hysteresis-free negative capacitance MoS2 transistors,” *Nature Nanotechnology*, vol. 13, no. 1, pp. 24–28, 2018, doi: 10.1038/s41565-017-0010-1.

[120] M. Si, C. Jiang, W. Chung, Y. Du, M. A. Alam, and P. D. Ye, “Steep-Slope WSe 2 Negative Capacitance Field-Effect Transistor,” *Nano Letters*, vol. 18, no. 6, pp. 3682–3687, 2018, doi: 10.1021/acs.nanolett.8b00816.

[121] M. A. Alam, M. Si, and P. D. Ye, “A critical review of recent progress on negative capacitance field-effect transistors,” *Applied Physics Letters*, vol. 114, no. 9, p. 090401, 2019, doi: 10.1063/1.5092684.

[122] W. Cao and K. Banerjee, “Is negative capacitance FET a steep-slope logic switch?,” *Nature Communications*, vol. 11, no. 1, p. 196, 2020, doi: 10.1038/s41467-019-13797-9.

[123] T. P. Ma and Jin-Ping Han, “Why is nonvolatile ferroelectric memory field-effect transistor still elusive?,” *IEEE Electron Device Letters*, vol. 23, no. 7, pp. 386–388, 2002, doi: 10.1109/LED.2002.1015207.

[124] H. Wang *et al.*, “New Insights into the Physical Origin of Negative Capacitance and Hysteresis in NCFETs,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2019, pp. 31.1.1-31.1.4, doi: 10.1109/IEDM.2018.8614504.

[125] C. Jin, K. Jang, T. Saraya, T. Hiramoto, and M. Kobayashi, “Experimental Study on the Role of Polarization Switching in Subthreshold Characteristics of HfO 2 -based Ferroelectric and Anti-ferroelectric FET,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2019, pp. 31.5.1-31.5.4, doi: 10.1109/IEDM.2018.8614486.

[126] X. Li and A. Toriumi, “Direct relationship between sub-60 mV/dec subthreshold swing and internal potential instability in MOSFET externally connected to ferroelectric capacitor,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2019, pp. 31.3.1-31.3.4, doi: 10.1109/IEDM.2018.8614703.

[127] M. Hoffmann, S. Slesazeck, U. Schroeder, and T. Mikolajick, “What’s next for negative capacitance electronics?,” *Nature Electronics*, vol. 3, no. 9, pp. 504–506, 2020, doi: 10.1038/s41928-020-00474-9.

[128] “Rethinking negative capacitance research,” *Nature Electronics*, vol. 3, no. 9, pp. 503–503, 2020, doi: 10.1038/s41928-020-00483-8.

[129] Z. C. Yuan *et al.*, “Switching-Speed Limitations of Ferroelectric Negative-Capacitance FETs,” *IEEE Transactions on Electron Devices*, vol. 63, no. 10, pp. 4046–4052, 2016, doi: 10.1109/TED.2016.2602209.

[130] S. Thomas, “Guiding the design of negative-capacitance FETs,” *Nature Electronics*, vol. 3, no. 2, pp. 72–72, 2020, doi: 10.1038/s41928-020-0377-0.

[131] W. X. You, C. P. Tsai, and P. Su, “Short-Channel Effects in 2D Negative-Capacitance Field-Effect Transistors,” *IEEE Transactions on Electron Devices*, vol. 65, no. 4, pp. 1604–1610, 2018, doi: 10.1109/TED.2018.2805716.

[132] A. I. Khan, C. W. Yeung, Chenming Hu, and S. Salahuddin, “Ferroelectric negative capacitance MOSFET: Capacitance tuning &amp; antiferroelectric operation,” in *2011 International Electron Devices Meeting*, 2011, pp. 11.3.1-11.3.4, doi: 10.1109/IEDM.2011.6131532.

[133] S. K. Gupta *et al.*, “Harnessing ferroelectrics for non-volatile memories and logic,” in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, 2017, pp. 29–34, doi: 10.1109/ISQED.2017.7918288.

[134] S. Dunkel *et al.*, “A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond,” in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 19.7.1-19.7.4, doi: 10.1109/IEDM.2017.8268425.

[135] R.S. Lous, “Ferroelectric Memory Devices How to store the information of the future?,” *Master’s thesis*, 2011. https://www.rug.nl/research/zernike/education /topmasternanoscience/ns190lous.pdf

[136] A. Sheikholeslami and P. G. Gulak, “A survey of circuit innovations in ferroelectric random-access memories,” *Proceedings of the IEEE*, vol. 88, no. 5, pp. 667–689, 2000, doi: 10.1109/5.849164.

[137] S. George *et al.*, “Symmetric 2-D-Memory Access to Multidimensional Data,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 6, pp. 1040–1050, 2018, doi: 10.1109/TVLSI.2018.2801302.

[138] A. Sharma and K. Roy, “1T Non-Volatile Memory Design Using Sub-10nm Ferroelectric FETs,” *IEEE Electron Device Letters*, vol. 39, no. 3, pp. 359–362, 2018, doi: 10.1109/LED.2018.2797887.

[139] D. Reis *et al.*, “Design and Analysis of an Ultra-Dense, Low-Leakage, and Fast FeFET-Based Random Access Memory Array,” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 5, no. 2, pp. 103–112, 2019, doi: 10.1109/JXCDC.2019.2930284.

[140] S. Gupta, M. Steiner, A. Aziz, V. Narayanan, S. Datta, and S. K. Gupta, “Device-Circuit Analysis of Ferroelectric FETs for Low-Power Logic,” *IEEE Transactions on Electron Devices*, vol. 64, no. 8, pp. 3092–3100, 2017, doi: 10.1109/TED.2017.2717929.

[141] K. Ni *et al.*, “Critical Role of Interlayer in Hf $_{0.5}$ Zr $_{0.5}$ O $_2$ Ferroelectric FET Nonvolatile Memory Performance,” *IEEE Transactions on Electron Devices*, vol. 65, no. 6, pp. 2461–2469, 2018, doi: 10.1109/TED.2018.2829122.

[142] Ni, K. et al., “SoC Logic Compatible Multi-Bit FeMFET Weight Cell for Neuromorphic Applications,” *IEEE International Electron Devices Meeting (IEDM),* 2018, doi: 10.1109/IEDM.2018.8614496.

[143] K. Ma *et al.*, “Nonvolatile Processor Architecture Exploration for Energy-Harvesting Applications,” *IEEE Micro*, vol. 35, no. 5, pp. 32–40, 2015, doi: 10.1109/MM.2015.88.

[144] X. Li *et al.*, “Advancing Nonvolatile Computing With Nonvolatile NCFET Latches and Flip-Flops,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 11, pp. 2907–2919, 2017, doi: 10.1109/TCSI.2017.2702741.

[145] X. Li *et al.*, “Lowering Area Overheads for FeFET-Based Energy-Efficient Nonvolatile Flip-Flops,” *IEEE Transactions on Electron Devices*, vol. 65, no. 6, pp. 2670–2674, 2018, doi: 10.1109/TED.2018.2829348.

[146] X. Li *et al.*, “Enabling Energy-Efficient Nonvolatile Computing with Negative Capacitance FET,” *IEEE Transactions on Electron Devices*, vol. 64, no. 8, pp. 3452–3458, 2017, doi: 10.1109/TED.2017.2716338.

[147] X. Li *et al.*, “Advancing Nonvolatile Computing with Nonvolatile NCFET Latches and Flip-Flops,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 11, pp. 2907–2919, 2017, doi: 10.1109/TCSI.2017.2702741.

[148] M. Bibes, “Nanoferronics is a winning combination,” *Nature Materials 2012 11:5*, 2012, doi: 10.1038/nmat3318.

[149] R. Hadidi, B. Asgari, A. Mudassar, S. Mukhopadhyay, S. Yalamanchili, and H. Kim, “Demystifying the Characteristics of 3D-Stacked Memories: A Case Study for Hybrid Memory Cube,” *IEEE International Symposium on Workload Characterization (IISWC),* 2017, doi: 10.1109/IISWC.2017.8167757.

[150] S. H. Pugsley *et al.*, “NDC: Analyzing the impact of 3D-stacked memory+logic devices on MapReduce workloads,” in *2014 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2014, pp. 190–200, doi: 10.1109/ISPASS.2014.6844483.

[151] Q. Zhu *et al.*, “A 3D-stacked logic-in-memory accelerator for application-specific data intensive computing,”, in *IEEE International 3D Systems Integration Conference (3DIC),* 2013, doi: 10.1109/3DIC.2013.6702348.

[152] W. W. Shen *et al.*, “3-D Stacked Technology of DRAM-Logic Controller Using Through-Silicon Via (TSV),” *IEEE Journal of the Electron Devices Society*, vol. 6, no. 1, pp. 396–402, 2018, doi: 10.1109/JEDS.2018.2815344.

[153] K. Roy, M. Sharad, D. Fan, and K. Yogendra, “Brain-inspired computing with spin torque devices,” in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2014*, 2014, pp. 1–6, doi: 10.7873/DATE.2014.245.

[154] A. Sengupta, P. Panda, A. Raghunathan, and K. Roy, “Neuromorphic Computing Enabled by Spin-Transfer Torque Devices,” in *2016 29th International Conference on VLSI Design and 2016 15th International Conference on Embedded Systems (VLSID)*, 2016, pp. 32–37, doi: 10.1109/VLSID.2016.117.

[155] F. Parveen, S. Angizi, Z. He, and D. Fan, “Low power in-memory computing based on dual-mode SOT-MRAM,” in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2017, pp. 1–6, doi: 10.1109/ISLPED.2017.8009200.

[156] A. Sebastian, M. le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, “Memory devices and applications for in-memory computing,” *Nature Nanotechnology*, vol. 15, no. 7, pp. 529–544, 2020, doi: 10.1038/s41565-020-0655-z.

[157] H. Mulaosmanovic *et al.*, “Novel ferroelectric FET based synapse for neuromorphic systems,” in *Digest of Technical Papers - Symposium on VLSI Technology*, 2017, pp. T176–T177, doi: 10.23919/VLSIT.2017.7998165.

[158] S. Moon, J. Shin, and C. Shin, “Understanding of Polarization-Induced Threshold Voltage Shift in Ferroelectric-Gated Field Effect Transistor for Neuromorphic Applications,” *Electronics*, vol. 9, no. 5, p. 704, 2020, doi: 10.3390/electronics9050704.

[159] S. Jot, A. Zyarah, S. Kurinec, K. Ni, F. T. Zohora, and D. Kudithipudi, “FeFET-Based Neuromorphic Architecture with On-Device Feedback Alignment Training,” in *Proceedings - International Symposium on Quality Electronic Design, ISQED*, 2020, vol. 2020-March, pp. 317–322, doi: 10.1109/ISQED48828.2020.9137035.

[160] S. Oh, H. Hwang, and I. K. Yoo, “Ferroelectric materials for neuromorphic computing,” *APL Materials*, vol. 7, no. 9, p. 091109, 2019, doi: 10.1063/1.5108562.

[161] S. Dutta, C. Schafer, J. Gomez, K. Ni, S. Joshi, and S. Datta, “Supervised Learning in All FeFET-Based Spiking Neural Network: Opportunities and Challenges,” *Frontiers in Neuroscience*, vol. 14, 2020, doi: 10.3389/fnins.2020.00634.

[162] C. Chen *et al.*, “Bio-Inspired Neurons Based on Novel Leaky-FeFET with Ultra-Low Hardware Cost and Advanced Functionality for All-Ferroelectric Neural Network,” in *Digest of Technical Papers - Symposium on VLSI Technology*, 2019, vol. 2019-June, pp. T136–T137, doi: 10.23919/VLSIT.2019.8776495.

[163] P. Wurfel and I. P. Batra, "Depolarization-field-induced instability in thin ferroelectric films experiment and theory," *Physical Review B*, vol. 8, no. 11, pp. 5126–5133, 1973, doi: 10.1103/PhysRevB.8.5126.

[164] Zheng Wang, Muhammad Mainul Islam, Panni Wang, Shan Deng, Shimeng Yu, A. I. Khan, and K. Ni, "Depolarization Field Induced Instability of Polarization States in HfO2 Based Ferroelectric FET," in *IEEE International Electron Devices Meeting (IEDM)*, 2020.

[165] K. Ni, W. Chakraborty, J. Smith, B. Grisafe, and S. Datta, "Fundamental Understanding and Control of Device-to-Device Variation in Deeply Scaled Ferroelectric FETs," in *Digest of Technical Papers - Symposium on VLSI Technology*, Jun. 2019, pp. T40–T41, doi: 10.23919/VLSIT.2019.8776497.

[166] R. Khachaturyan, J. Wehner, and Y. A. Genenko, "Correlated polarization-switching kinetics in bulk polycrystalline ferroelectrics: A self-consistent mesoscopic switching model," *Physical Review B*, vol. 96, no. 5, p. 054113, 2017, doi: 10.1103/PhysRevB.96.054113.

[167] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, no. 11, pp. 521–529, 2019, doi: 10.1038/s41928-019-0321-3.

[168] X. Yin, M. Niemier, and X. S. Hu, "Design and benchmarking of ferroelectric FET based TCAM," in *Proceedings of the 2017 Design, Automation and Test in Europe, DATE 2017*, 2017, pp. 1444–1449, doi: 10.23919/DATE.2017.7927219.

[169] H. Ishiwara, "Recent progress in FET-type ferroelectric memories," in *IEEE International Electron Devices Meeting 2003*, pp. 10.3.1-10.3.4, doi: 10.1109/IEDM.2003.1269274.

[170] S. D. Hyun *et al.*, "Dispersion in Ferroelectric Switching Performance of Polycrystalline Hf $_{0.5}$ Zr $_{0.5}$ O $_2$ Thin Films," *ACS Applied Materials & Interfaces*, vol. 10, no. 41, pp. 35374–35384, 2018, doi: 10.1021/acsami.8b13173.

[171] S. K. Thirumala and S. K. Gupta, "Reconfigurable Ferroelectric Transistor—Part I: Device Design and Operation," *IEEE Transactions on Electron Devices*, vol. 66, no. 6, pp. 2771–2779, 2019, doi: 10.1109/TED.2019.2897960.

[172] Z. Krivokapic *et al.*, "14nm Ferroelectric FinFET technology with steep subthreshold slope for ultra low power applications," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 15.1.1-15.1.4, doi: 10.1109/IEDM.2017.8268393.

[173] E. T. Breyer, H. Mulaosmanovic, T. Mikolajick, and S. Slesazeck, "Reconfigurable NAND/NOR logic gates in 28 nm HKMG and 22 nm FD-SOI FeFET technology," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 28.5.1-28.5.4, doi: 10.1109/IEDM.2017.8268471.

[174] S. K. Thirumala *et al.*, "Dual Mode Ferroelectric Transistor based Non-Volatile Flip-Flops for Intermittently-Powered Systems," in *Proceedings of the International Symposium on Low Power Electronics and Design - ISLPED '18*, 2018, pp. 1–6, doi: 10.1145/3218603.3218653.

[175] "PTM - Latest models." http://ptm.asu.edu/latest.html (accessed May 27, 2019).

[176] J. Gu, J. Keane, S. Sapatnekar, and C. Kim, "Width Quantization Aware FinFET Circuit Design," in *IEEE Custom Integrated Circuits Conference,* 2006, pp. 337–340, doi: 10.1109/CICC.2006.320916.

[177] "MOSIS." https://www.mosis.com/files/scmos/scmos.pdf (accessed Dec. 09, 2020).

[178] https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Kaizad-Mistry-2017-Manufacturing.pdf (accessed Dec. 09, 2020).

[179] S. Sivasubramanian and A. Widom, "Physical Kinetics of Ferroelectric Hysteresis," 2001. Accessed: Dec. 10, 2018. [Online]. Available: https://arxiv.org/pdf/cond-mat/0106549.pdf.

[180] J. Li, B. Nagaraj, H. Liang, W. Cao, Chi. H. Lee, and R. Ramesh, "Ultrafast polarization switching in thin-film ferroelectrics," *Applied Physics Letters*, vol. 84, no. 7, pp. 1174–1176, 2004, doi: 10.1063/1.1644917.

[181] I. Katsouras *et al.*, "Controlling the on/off current ratio of ferroelectric field-effect transistors," *Scientific Reports*, vol. 5, no. 1, p. 12094, Dec. 2015, doi: 10.1038/srep12094.

[182] Sung-Min Yoon and H. Ishiwara, "Memory operations of 1T2C-type ferroelectric memory cell with excellent data retention characteristics," *IEEE Transactions on Electron Devices*, vol. 48, no. 9, pp. 2002–2008, 2001, doi: 10.1109/16.944189.

[183] A. I. Khan, U. Radhakrishna, K. Chatterjee, S. Salahuddin, and D. A. Antoniadis, "Negative Capacitance Behavior in a Leaky Ferroelectric," *IEEE Transactions on Electron Devices*, vol. 63, no. 11, pp. 4416–4422, Nov. 2016, doi: 10.1109/TED.2016.2612656.

[184] https://stacks.stanford.edu/file/druid:ds551wt1033/SVerma_Thesis-augmented.pdf (accessed Dec. 09, 2020).

[185] S. K. Thirumala and S. K. Gupta, "Reconfigurable Ferroelectric Transistor–Part II: Application in Low-Power Nonvolatile Memories," *IEEE Transactions on Electron Devices*, vol. 66, no. 6, pp. 2780–2788, 2019, doi: 10.1109/TED.2019.2912562.

[186] C. J. Xue, Y. Zhang, Y. Chen, G. Sun, J. J. Yang, and H. Li, "Emerging non-volatile memories," in *Proceedings of the seventh IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis - CODES+ISSS '11*, 2011, p. 325, doi: 10.1145/2039370.2039420.

[187] S. Yu and P.-Y. Chen, “Emerging Memory Technologies: Recent Trends and Prospects,” *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43–56, 2016, doi: 10.1109/MSSC.2016.2546199.

[188] H. Jayakumar, A. Raha, and V. Raghunathan, “Energy-Aware Memory Mapping for Hybrid FRAM-SRAM MCUs in IoT Edge Devices,” in *Proceedings of the IEEE International Conference on VLSI Design*, 2016, pp. 264–269, doi: 10.1109/VLSID.2016.52.

[189] R. Li, R. Naous, H. Fariborzi, and K. N. Salama, “Approximate Computing with Stochastic Transistors’ Voltage Over-Scaling,” *IEEE Access*, vol. 7, pp. 6373–6385, 2019, doi: 10.1109/ACCESS.2018.2889747.

[190] R. W. Johnson, A. Hultqvist, and S. F. Bent, “A brief review of atomic layer deposition: from fundamentals to applications,” *Materials Today*, vol. 17, no. 5, pp. 236–246, 2014, doi: 10.1016/J.MATTOD.2014.04.026.

[191] J. Muller, T. S. Boscke, U. Schroder, R. Hoffmann, T. Mikolajick, and L. Frey, “Nanosecond Polarization Switching and Long Retention in a Novel MFIS-FET Based on Ferroelectric HfO_2,” *IEEE Electron Device Letters*, vol. 33, no. 2, pp. 185–187, 2012, doi: 10.1109/LED.2011.2177435.

[192] X. Li *et al.*, “Design of 2T/Cell and 3T/Cell Nonvolatile Memories with Emerging Ferroelectric FETs,” *IEEE Design & Test*, vol. 36, no. 3, pp. 39–45, 2019, doi: 10.1109/MDAT.2019.2902094.

[193] M. D. Giles *et al.*, “High sigma measurement of random threshold voltage variation in 14nm Logic FinFET technology,” in *Digest of Technical Papers - Symposium on VLSI Technology*, 2015, pp. T150–T151, doi: 10.1109/VLSIT.2015.7223657.

[194] H. Jayakumar, K. Lee, W. S. Lee, A. Raha, Y. Kim, and V. Raghunathan, “Powering the Internet of Things,” in *Proceedings of the International Symposium on Low Power Electronics and Design*, 2015, pp. 375–380, doi: 10.1145/2627369.2631644.

[195] C. Lu, S. P. Park, V. Raghunathan, and K. Roy, “Analysis and design of ultra low power thermoelectric energy harvesting systems,” in *Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design - ISLPED ’10*, 2010, p. 183, doi: 10.1145/1840845.1840882.

[196] Y. Liu *et al.*, “Ambient energy harvesting nonvolatile processors,” in *Proceedings of the 52nd Annual Design Automation Conference on – DAC,* 2015, pp. 1–6, doi: 10.1145/2744769.2747910.

[197] H. Jayakumar, A. Raha, and V. Raghunathan, “QUICKRECALL: A Low Overhead HW/SW Approach for Enabling Computations across Power Cycles in Transiently Powered Computers,” in *2014 27th International Conference on VLSI Design and 2014 13th International Conference on Embedded Systems*, 2014, pp. 330–335, doi: 10.1109/VLSID.2014.63.

[198] A. Raha, A. Jaiswal, S. S. Sarwar, H. Jayakumar, V. Raghunathan, and K. Roy, “Designing Energy-Efficient Intermittently Powered Systems Using Spin-Hall-Effect-Based Nonvolatile SRAM,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 2, pp. 294–307, 2018, doi: 10.1109/TVLSI.2017.2767033.

[199] “Overview :: openMSP430 :: OpenCores.” https://opencores.org/projects/openmsp430 (accessed Dec. 09, 2020).

[200] “Open-Cell Library | Silicon Integration Initiative.” https://si2.org/open-cell-library/ (accessed Dec. 09, 2020).

[201] S. Khoram, Y. Zha, J. Zhang, and J. Li, “Challenges and opportunities: From near-memory computing to in-memory computing,” in *Proceedings of the International Symposium on Physical Design*, Mar. 2017, vol. Part F127197, pp. 43–46, doi: 10.1145/3036669.3038242.

[202] S. Srinivasa *et al.*, “ROBIN: Monolithic-3D SRAM for Enhanced Robustness With In-MemoryComputation Support,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–13, 2019, doi: 10.1109/TCSI.2019.2897497.

[203] N. Talati, S. Gupta, P. Mane, and S. Kvatinsky, “Logic Design Within Memristive Memories Using Memristor-Aided loGIC (MAGIC),” *IEEE Transactions on Nanotechnology*, vol. 15, no. 4, pp. 635–650, 2016, doi: 10.1109/TNANO.2016.2570248.

[204] W.-S. Khwa *et al.*, “A 65nm 4Kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3ns and 55.8TOPS/W fully parallel product-sum operation for binary DNN edge processors,” in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 496–498, doi: 10.1109/ISSCC.2018.8310401.

[205] I. Yoon *et al.*, “A FeFET Based Processing-In-Memory Architecture for Solving Distributed Least-Square Optimizations,” in *2018 76th Device Research Conference (DRC)*, 2018, pp. 1–2, doi: 10.1109/DRC.2018.8442235.

[206] W.-H. Chen *et al.*, “A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors,” in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 494–496, doi: 10.1109/ISSCC.2018.8310400.

[207] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010, doi: 10.1016/j.comnet.2010.05.010.

[208] B., Braem, B., Moerman, I. et al., "A survey on wireless body area networks," *Wireless Netw* 17, 1–18, 2011, doi: 10.1007/s11276-010-0252-4.

[209] H. Jayakumar, A. Raha, W. S. Lee, and V. Raghunathan, "QUICKRECALL: A HW/SW approach for computing across power cycles in Transiently Powered Computers," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 12, no. 1, 2015, doi: 10.1145/2700249.

[210] Y. Liu *et al.*, "A 130-nm ferroelectric nonvolatile system-on-chip with direct peripheral restore architecture for transient computing system," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 3, pp. 885–895, 2019, doi: 10.1109/JSSC.2018.2884349.

[211] B. Karthikeyan, M. Velumani, R. Kumar, and S. R. Inabathini, "Analysis of data aggregation in wireless sensor network," in *International Conference on Electronics and Communication Systems, ICECS* 2015, pp. 1435–1439, doi: 10.1109/ECS.2015.7124823.

[212] F. Su *et al.*, "A 462GOPs/J RRAM-based nonvolatile intelligent processor for energy harvesting IoE system featuring nonvolatile logics and processing-in-memory," in *Digest of Technical Papers - Symposium on VLSI Technology*, 2017, pp. C260–C261, doi: 10.23919/VLSIT.2017.7998149.

[213] G. Gobieski, N. Beckmann, and B. Lucia, "Intelligence Beyond the Edge: Inference on Intermittent Embedded Systems," *International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS*, pp. 199–213, 2019, doi: 10.1145/3297858.3304011.

[214] C. Eckert *et al.*, "Neural cache: Bit-Serial In-Cache acceleration of deep neural networks," in *Proceedings - International Symposium on Computer Architecture*, 2018, pp. 383–396, doi: 10.1109/ISCA.2018.00040.

[215] F. Gao, G. Tziantzioulis, and D. Wentzlaff, "ComputeDRAM," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 100–113, doi: 10.1145/3352460.3358260.

[216] "CRC Implementation With MSP430^{TM} MCUs Application Report," 2004. Accessed: Dec. 09, 2020. [Online]. Available: www.ti.com/lit/zip/slaa221.

[217] S. Randhawa and S. Jain, "Data Aggregation in Wireless Sensor Networks: Previous Research, Current Status and Future Directions," *Wireless Personal Communications*, vol. 97, no. 3. Springer, pp. 3355–3425, 2017, doi: 10.1007/s11277-017-4674-5.

[218] K. Ni, X. Li, J. A. Smith, M. Jerry, and S. Datta, "Write Disturb in Ferroelectric FETs and Its Implication for 1T-FeFET and Memory Arrays," *IEEE Electron Device Letters*, vol. 39, no. 11, pp. 1656–1659, 2018, doi: 10.1109/LED.2018.2872347.

[219] S. K. Thirumala, S. Jain, A. Raghunathan, and S. K. Gupta, “Non-Volatile Memory utilizing Reconfigurable Ferroelectric Transistors to enable Differential Read and Energy-Efficient In-Memory Computation,” 2019, pp. 1–6, doi: 10.1109/islped.2019.8824948.

[220] W. Song, Y. Zhou, M. Zhao, L. Ju, C. J. Xue, and Z. Jia, “EMC: Energy-Aware Morphable Cache Design for Non-Volatile Processors,” *IEEE Transactions on Computers*, vol. 68, no. 4, pp. 498–509, 2019, doi: 10.1109/TC.2018.2879103.

[221] J. Wang *et al.*, “A 28-nm Compute SRAM with Bit-Serial Logic/Arithmetic Operations for Programmable In-Memory Vector Computing,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 76–86, 2020, doi: 10.1109/JSSC.2019.2939682.

[222] S. K. Thirumala, A. Raha, V. Narayanan, V. Raghunathan, and S. K. Gupta, “Non-volatile Logic and Memory based on Reconfigurable Ferroelectric Transistors,” 2019, doi: 10.1109/NANOARCH47378.2019.181302.

[223] “Artificial Intelligence Is Driving Huge Changes at Google, Facebook, and Microsoft | WIRED.” https://www.wired.com/2016/11/google-facebook-microsoft-remaking-around-ai/ (accessed Dec. 09, 2020).

[224] S. Venkataramani, K. Roy, and A. Raghunathan, “Efficient embedded learning for IoT devices,” in *Proceedings of the Asia and South Pacific Design Automation Conference, ASP-DAC*, 2016, pp. 308–311, doi: 10.1109/ASPDAC.2016.7428029.

[225] A. Mishra, E. Nurvitadhi, J. J. Cook, and D. Marr, “WRPN: Wide reduced-precision networks,” *arXiv*, 2017, available: https://arxiv.org/abs/1709.01134v1.

[226] P. Wang, X. Xie, L. Deng, G. Li, D. Wang, and Y. Xie, “HitNet: Hybrid Ternary Recurrent Neural Network,” *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, p. 602–612.

[227] “NVIDIA V100 | NVIDIA.” https://www.nvidia.com/en-us/data-center/v100/ (accessed Dec. 09, 2020).

[228] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks,” *ArXiv,* 2016, available: http://arxiv.org/abs/1603.05279.

[229] Z. Jiang, S. Yin, M. Seok, and J. Seo, “XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks,” in *2018 IEEE Symposium on VLSI Technology*, 2018, pp. 173–174, doi: 10.1109/VLSIT.2018.8510687.

[230] X. Sun, S. Yin, X. Peng, R. Liu, J. Seo, and S. Yu, “XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks,” in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2018, pp. 1423–1428, doi: 10.23919/DATE.2018.8342235.

[231] M. Jerry *et al.*, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 6.2.1-6.2.4, doi: 10.1109/IEDM.2017.8268338.

[232] S. Yu *et al.*, "Binary neural network with 16 Mb RRAM macro chip for classification and online training," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 16.2.1-16.2.4, doi: 10.1109/IEDM.2016.7838429.

[233] S. Jain, S. K. Gupta, and A. Raghunathan, "TiM-DNN: Ternary In-Memory Accelerator for Deep Neural Networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 7, pp. 1567–1577, 2020, doi: 10.1109/TVLSI.2020.2993045.

[234] T. Yoo, H. Kim, Q. Chen, T. T. H. Kim, and B. Kim, "A Logic Compatible 4T Dual Embedded DRAM Array for In-Memory Computation of Deep Neural Networks," in *Proceedings of the International Symposium on Low Power Electronics and Design*, 2019, doi: 10.1109/ISLPED.2019.8824826.

[235] H. Choi, Y. Lee, J. J. Kim, and S. Yoo, "A Novel In-DRAM Accelerator Architecture for Binary Neural Network," *IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*, 2020, doi: 10.1109/COOLCHIPS49199.2020.9097642.

[236] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8μJ/86% CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28nm CMOS," in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, 2018, vol. 61, pp. 222–224, doi: 10.1109/ISSCC.2018.8310264.

[237] A. Agrawal *et al.*, "Xcel-RAM: Accelerating Binary Neural Networks in High-Throughput SRAM Compute Arrays," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 8, pp. 3064–3076, 2019, doi: 10.1109/TCSI.2019.2907488.

[238] L. Ni, H. Huang, Z. Liu, R. v. Joshi, and H. Yu, "Distributed in-memory computing on binary RRAM crossbar," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 13, no. 3, pp. 1–18, 2017, doi: 10.1145/2996192.

[239] D. Fan and S. Angizi, "Energy efficient in-memory binary deep neural network accelerator with dual-mode SOT-MRAM," in *Proceedings - 35th IEEE International Conference on Computer Design, ICCD 2017*, 2017, pp. 609–612, doi: 10.1109/ICCD.2017.107.

[240] X. Chen, X. Yin, M. Niemier, and X. S. Hu, "Design and optimization of FeFET-based crossbars for binary convolution neural networks," in *Proceedings of the 2018 Design, Automation and Test in Europe Conference and Exhibition, DATE 2018*, 2018, vol. 2018-January, pp. 1205–1210, doi: 10.23919/DATE.2018.8342199.

[241] S. Resch *et al.*, "PIMBALL: Binary Neural Networks in Spintronic Memory," *arXiv*, 2018, available: https://doi.org/10.1145/1122445.1122456.

[242] X. Si *et al.*, “A Twin-8T SRAM Computation-in-Memory Unit-Macro for Multibit CNN-Based AI Edge Processors,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 189–202, 2020, doi: 10.1109/JSSC.2019.2952773.

[243] X. Liu *et al.*, “RENO: A high-efficient reconfigurable neuromorphic computing accelerator design,” in *Proceedings - Design Automation Conference*, 2015, doi: 10.1145/2744769.2744900.

[244] A. Ankit *et al.*, “PANTHER: A Programmable Architecture for Neural Network Training Harnessing Energy-Efficient ReRAM,” *IEEE Transactions on Computers*, vol. 69, no. 8, pp. 1128–1142, 2020, doi: 10.1109/TC.2020.2998456.

[245] A. Shafiee *et al.*, “ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars,” in *Proceedings - 2016 43rd International Symposium on Computer Architecture, ISCA 2016*, 2016, pp. 14–26, doi: 10.1109/ISCA.2016.12.

[246] P. Chi *et al.*, “PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory,” in *Proceedings - 2016 43rd International Symposium on Computer Architecture, ISCA 2016*, 2016, pp. 27–39, doi: 10.1109/ISCA.2016.13.

[247] “US9697877B2 - Compute memory - Google Patents.” https://patents.google.com/patent/US9697877 (accessed Dec. 09, 2020).

[248] M. Jerry *et al.*, “Ferroelectric FET analog synapse for acceleration of deep neural network training,” in *Technical Digest - International Electron Devices Meeting, IEDM*, 2018, pp. 6.2.1-6.2.4, doi: 10.1109/IEDM.2017.8268338.

[249] A. Agarwal, H. Li, and K. Roy, “DRG-Cache: A data retention gated-ground cache for low power,” in *Proceedings - Design Automation Conference*, 2002, pp. 473–478, doi: 10.1109/dac.2002.1012671.

[250] A. Aziz *et al.*, “Computing with ferroelectric FETs: Devices, models, systems, and applications,” in *Proceedings of the 2018 Design, Automation and Test in Europe Conference and Exhibition (DATE),* 2018, doi: 10.23919/DATE.2018.8342213.

[251] Y. Seo, X. Fong, K. W. Kwon, and K. Roy, “Spin-hall magnetic random-access memory with dual read/write ports for on-chip caches,” *IEEE Magnetics Letters*, vol. 6, 2015, doi: 10.1109/LMAG.2015.2422260.

[252] G. Li *et al.*, “Understand-ing Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications,” *in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis,* 2017 Article No.: 8 Pages 1–12, doi: 10.1145/3126908.3126964.

[253] A. al Bahou, G. Karunaratne, R. Andri, L. Cavigelli, and L. Benini, "XNORBIN: A 95 TOp/s/W hardware accelerator for binary convolutional neural networks," in *21st IEEE Symposium on Low-Power and High-Speed Chips and Systems, COOL Chips 2018 - Proceedings*, 2018, pp. 1–3, doi: 10.1109/CoolChips.2018.8373076.

[254] S. P. Park, S. Gupta, N. Mojumder, A. Raghunathan, and K. Roy, "Future cache design using STT MRAMs for improved energy efficiency," in *Proceedings of the 49th Annual Design Automation Conference - DAC*, 2012, doi: 10.1145/2228360.2228447.

[255] Y. Xie, J. Ma, S. Ganguly, and A. W. Ghosh, "From materials to systems: a multiscale analysis of nanomagnetic switching," *Journal of Computational Electronics*, vol. 16, no. 4, pp. 1201–1226, 2017, doi: 10.1007/s10825-017-1054-z.

[256] L. Liu, C.-F. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, "Spin-torque switching with the giant spin Hall effect of tantalum.," *Science (New York, N.Y.)*, vol. 336, no. 6081, 2012, doi: 10.1126/science.1218197.

[257] A. van den Brink *et al.*, "Spin-Hall-assisted magnetic random access memory," *Applied Physics Letters*, vol. 104, no. 1, p. 012403,2014, doi: 10.1063/1.4858465.

[258] C. F. Pai, L. Liu, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, "Spin transfer torque devices utilizing the giant spin Hall effect of tungsten," *Applied Physics Letters*, vol. 101, no. 12, p. 122404, 2012, doi: 10.1063/1.4753947.

[259] L. Liu, O. J. Lee, T. J. Gudmundsen, D. C. Ralph, and R. A. Buhrman, "Current-induced switching of perpendicularly magnetized magnetic layers using spin torque from the spin hall effect," *Physical Review Letters*, vol. 109, no. 9, p. 096602, 2012, doi: 10.1103/PhysRevLett.109.096602.

[260] G. Yu *et al.*, "Switching of perpendicular magnetization by spin–orbit torques in the absence of external magnetic fields," *Nature Nanotechnology*, vol. 9, no. 7, pp. 548–554, 2014, doi: 10.1038/nnano.2014.94.

[261] S. Ikeda *et al.*, "A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction," *Nature Materials*, vol. 9, no. 9, pp. 721–724, 2010, doi: 10.1038/nmat2804.

[262] Z. He, S. Angizi, and D. Fan, "Exploring STT-MRAM based in-memory computing paradigm with application of image edge extraction," in *Proceedings - 35th IEEE International Conference on Computer Design, ICCD 2017*, 2017, pp. 439–446, doi: 10.1109/ICCD.2017.78.

[263] Z. He, Y. Zhang, S. Angizi, B. Gong, and D. Fan, "Exploring a SOT-MRAM Based In-Memory Computing for Data Processing," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 4, no. 4, pp. 676–685, 2018, doi: 10.1109/TMSCS.2018.2836967.

[264] L. Zhang *et al.*, “A high-reliability and low-power computing-in-memory implementation within STT-MRAM,” *Microelectronics Journal*, vol. 81, pp. 69–75, 2018, doi: 10.1016/j.mejo.2018.09.005.

[265] Y. Seo and K. Roy, “High-Density SOT-MRAM Based on Shared Bitline Structure,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 8, pp. 1600–1603, 2018, doi: 10.1109/TVLSI.2018.2822841.

[266] R. Andrawis, A. Jaiswal, and K. Roy, “Design and Comparative Analysis of Spintronic Memories Based on Current and Voltage Driven Switching,” *IEEE Transactions on Electron Devices*, vol. 65, no. 7, pp. 2682–2693, 2018, doi: 10.1109/TED.2018.2833039.

[267] R. Andrawis, A. Jaiswal, and K. Roy, “Design and Comparative Analysis of Spintronic Memories Based on Current and Voltage Driven Switching,” *IEEE Transactions on Electron Devices*, vol. 65, no. 7, pp. 2682–2693, 2018, doi: 10.1109/TED.2018.2833039.

[268] K. S. Lee, S. W. Lee, B. C. Min, and K. J. Lee, “Threshold current for switching of a perpendicular magnetic layer induced by spin Hall effect,” *Applied Physics Letters*, vol. 102, no. 11, p. 112410, 2013, doi: 10.1063/1.4798288.

[269] K. Y. Camsari, S. Ganguly, and S. Datta, “Modular Approach to Spintronics,” *Scientific Reports*, vol. 5, no. 1, p. 10571, 2015, doi: 10.1038/srep10571.

[270] K. S. Lee, S. W. Lee, B. C. Min, and K. J. Lee, “Thermally activated switching of perpendicular magnet by spin-orbit spin torque,” *Applied Physics Letters*, vol. 104, no. 7, p. 072413, 2014, doi: 10.1063/1.4866186.

[271] P. Debashis and Z. Chen, “Experimental Demonstration of a Spin Logic Device with Deterministic and Stochastic Mode of Operation,” *Scientific Reports*, vol. 8, no. 1, p. 11405, 2018, doi: 10.1038/s41598-018-29601-5.

[272] Z. Y. Zhu, Y. C. Cheng, and U. Schwingenschlögl, “Giant spin-orbit-induced spin splitting in two-dimensional transition-metal dichalcogenide semiconductors,” *Physical Review B - Condensed Matter and Materials Physics*, vol. 84, no. 15, p. 153402, 2011, doi: 10.1103/PhysRevB.84.153402.

[273] T. Y. T. Hung, A. Rustagi, S. Zhang, P. Upadhyaya, and Z. Chen, “Experimental observation of coupled valley and spin Hall effect in p-doped WSe $_2$ devices,” *InfoMat*, vol. 2, no. 5, pp. 968–974, 2020, doi: 10.1002/inf2.12095.

[274] E. Barré *et al.*, “Spatial Separation of Carrier Spin by the Valley Hall Effect in Monolayer WSe2 Transistors,” *Nano Letters*, 2019, doi: 10.1021/acs.nanolett.8b03838.

[275] P. Zhao *et al.*, “Air stable p-doping of WSe2 by covalent functionalization,” *ACS Nano*, vol. 8, no. 10, pp. 10808–10814, 2014, doi: 10.1021/nn5047844.

[276] D. A. Abanin, A. v. Shytov, L. S. Levitov, and B. I. Halperin, "Nonlocal charge transport mediated by spin diffusion in the spin Hall effect regime," *Physical Review B - Condensed Matter and Materials Physics*, vol. 79, no. 3, p. 035304, 2009, doi: 10.1103/PhysRevB.79.035304.

[277] S. K. Thirumala *et al.*, "Valley-Coupled-Spintronic Non-Volatile Memories With Compute-In-Memory Support," *IEEE Transactions on Nanotechnology*, vol. 19, pp. 635–647, 2020, doi: 10.1109/TNANO.2020.3012550.

[278] S. Thirumala *et al.*, "WSe based Valley-Coupled-Spintronic Devices for Low Power Non-Volatile Memories," in *Device Research Conference - Conference Digest, DRC*, 2019, pp. 211–212, doi: 10.1109/DRC46940.2019.9046398.

[279] S. Suryavanshi and E. Pop, "S2DS: Physics-based compact model for circuit simulation of two-dimensional semiconductor devices including non-idealities," *Journal of Applied Physics*, vol. 120, no. 22, p. 224503, 2016, doi: 10.1063/1.4971404.

[280] X. Fong, S. K. Gupta, N. N. Mojumder, S. H. Choday, C. Augustine, and K. Roy, "KNACK: A hybrid spin-charge mixed-mode simulator for evaluating different genres of spin-transfer torque MRAM bit-cells," in *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, 2011, pp. 51–54, doi: 10.1109/SISPAD.2011.6035047.

[281] C. D. English, G. Shine, V. E. Dorgan, K. C. Saraswat, and E. Pop, "Improved Contacts to MoS 2 Transistors by Ultra-High Vacuum Metal Deposition," *Nano Letters*, vol. 16, no. 6, pp. 3824–3830, 2016, doi: 10.1021/acs.nanolett.6b01309.

[282] N. Binkert *et al.*, "The gem5 simulator," *ACM SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1–7, 2011, doi: 10.1145/2024716.2024718.

[283] W. Kang, H. Wang, Z. Wang, Y. Zhang, and W. Zhao, "In-Memory Processing Paradigm for Bitwise Logic Operations in STT–MRAM," *IEEE Transactions on Magnetics*, vol. 53, no. 11, pp. 1–4, 2017, doi: 10.1109/TMAG.2017.2703863.

[284] H. Zhang, W. Kang, L. Wang, K. L. Wang, and W. Zhao, "Stateful Reconfigurable Logic via a Single-Voltage-Gated Spin Hall-Effect Driven Magnetic Tunnel Junction in a Spintronic Memory," *IEEE Transactions on Electron Devices*, vol. 64, no. 10, pp. 4295–4301, 2017, doi: 10.1109/TED.2017.2726544.

[285] H. Zhang, W. Kang, K. Cao, B. Wu, Y. Zhang, and W. Zhao, "Spintronic Processing Unit in Spin Transfer Torque Magnetic Random Access Memory," *IEEE Transactions on Electron Devices*, vol. 66, no. 4, pp. 2017–2022, 2019, doi: 10.1109/TED.2019.2898391.

[286] R. Garlic, "Microprocessor with immediate and indirect addressing," 1978.

[287] S. K. Thirumala, S. Jain, S. K. Gupta, and A. Raghunathan, "Ternary Compute-Enabled Memory using Ferroelectric Transistors for Accelerating Deep Neural Networks," in

Proceedings of the 2020 Design, Automation and Test in Europe Conference and Exhibition, DATE 2020, 2020, pp. 31–36, doi: 10.23919/DATE48585.2020.9116495.

[288] Y. Kim, H. Kim, and J.-J. Kim, "Neural Network-Hardware Co-design for Scalable RRAM-based BNN Accelerators," *ArXiv,* 2018, available: http://arxiv.org/abs/1811.02187.

[289] S. Hashemi, N. Anthony, H. Tann, R. Iris Bahar, and S. Reda, "Understanding the Impact of Precision Quantization on the Accuracy and Energy of Neural Networks," *in DATE: Proceedings of the Conference on Design, Automation & Test in Europe,* 2017, Pages 1478–1483.

[290] "8-Bit Precision for Training Deep Learning Systems | IBM Research Blog." https://www.ibm.com/blogs/research/2018/12/8-bit-precision-training/ (accessed Dec. 09, 2020).

[291] N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan, "Training Deep Neural Networks with 8-bit Floating Point Numbers." *ArXiv,* 2018, available: https://arxiv.org/abs/1812.08011.

[292] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "PACT: Parameterized Clipping Activation for Quantized Neural Networks," *arXiv*, 2018, available: http://arxiv.org/abs/1805.06085.

[293] P. Colangelo, N. Nasiri, E. Nurvitadhi, A. Mishra, M. Margala, and K. Nealis, "Exploration of Low Numeric Precision Deep Learning Inference Using Intel® FPGAs," in *Proceedings - 26th IEEE International Symposium on Field-Programmable Custom Computing Machines, FCCM 2018*, 2018, pp. 73–80, doi: 10.1109/FCCM.2018.00020.

[294] A. Mishra, E. Nurvitadhi, J. J. Cook, and D. Marr, "WRPN: Wide reduced-precision networks," *arXiv*. 2017, available: https://arxiv.org/abs/1709.01134v1.

[295] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training Deep Neural Networks with binary weights during propagations,", *ArXiv,* 2015, available: http://arxiv.org/abs/1511.00363.

[296] X. Sun, X. Peng, P.-Y. Chen, R. Liu, J. Seo, and S. Yu, "Fully parallel RRAM synaptic array for implementing binary neural network with (+1, −1) weights and (+1, 0) neurons," in *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2018, pp. 574–579, doi: 10.1109/ASPDAC.2018.8297384.

[297] "In-Memory Computing – Deliang Fan." https://dfan.engineering.asu.edu/in-memory-computing/ (accessed Dec. 09, 2020).

[298] P. Wu and J. Appenzeller, "Sub-60 mV/decade switching in a metal-insulator-metal-insulator-semiconductor transistor without ferroelectric component", *ArXiv,* 2020, available: https://arxiv.org/abs/2012.00897

VITA

Sandeep Krishna Thirumala is a PhD candidate in the School of Electrical and Computer Engineering at Purdue University, Indiana, USA. He received his Bachelor's degree in Engineering Physics from the Indian Institute of Technology, Madras, India in 2016 and Master's degree in Electrical Engineering from The Pennsylvania State University, Pennsylvania, USA in 2018. He worked as an intern in the Technology Development division of Micron Technologies, Idaho, USA in 2020. His primary research interest includes exploring emerging device technologies for next-generation computing workloads. He explores artificial intelligence hardware for edge and Internet-of-Things (IoT) computing, processing-in-memory solutions for machine learning systems, and emerging technologies for logic and storage such as ferroelectrics, spintronics and complex-oxide electronics. He received several awards and accolades for his research including the best paper award in International Symposium on Nanoscale Architectures (NANOARCH), 2019 in Qingdao, China. He was also nominated for best paper award in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2020 in Grenoble, France and International Symposium on Low Power Electronic Design (ISLPED), 2018 in Washington, USA. He was the recipient of the prestigious *Bilsland Dissertation Fellowship*, awarded by Purdue University in 2020 and the *Merit cum Means scholarship* for academic excellence, awarded by the Indian Institute of Technology Madras, India from 2012-2016.

PUBLICATIONS

- **S. K. Thirumala**, A. Raha, V. Raghunathan and S. K. Gupta, "IPS-CiM: Enhancing Energy Efficiency of Intermittently-Powered Systems with Compute-in-Memory", *IEEE International Conference on Computer Design (ICCD)*, 2020.
- **S. K. Thirumala**, T. Hung, S. Jain, A. Raha, N. Thakuria, V. Raghunathan, A. Raghunathan, Z. Chen and S. K. Gupta, "Valley-Coupled-Spintronic Non-Volatile Memories with In-Memory Compute Support", in *IEEE Transactions on Nanotechnology (TNANO)*, 2020.
- **S. K. Thirumala**, S. Jain, S. K. Gupta and A. Raghunathan., "Ternary Compute-Enabled Memory using Ferroelectric Transistors for Accelerating Deep Neural Networks," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2020, Grenoble, France. ***(Best Paper Award Nominee).***
- K. Cho, **S. K. Thirumala**, et al., "Utilizing Valley-Spin Hall Effect in WSe_2 for Low Power Non-Volatile Flip-Flop Design", in *Device Research Conference (DRC)*, 2020, Columbus, OH.
- N. Thakuria, A. Saha, **S. K. Thirumala**, et al., "Polarization-induced Strain-coupled TMD FETs (PS FETs) for Non-Volatile Memory Applications", in *Device Research Conference (DRC),* 2020, Columbus, OH.
- **S. K. Thirumala**, S. Jain, A. Raghunathan and S. K. Gupta, "Non-Volatile Memory utilizing Reconfigurable Ferroelectric Transistors to enable Differential Read and Energy-Efficient In-Memory Computation," in *Int. Symposium on Low Power Electronic Design (ISLPED)*, 2019, Lausanne, Switzerland.
- **S. K. Thirumala**, et al., "Non-volatile Logic and Memory based on Reconfigurable Ferroelectric Transistors," in *International Symposium on Nanoscale Architectures (NANOARCH)*, 2019, Qingdao, China. ***(Best Paper Award).***
- **S. K. Thirumala**, et al., "WSe_2 based Valley-Coupled-Spintronic Devices for Low Power Non-Volatile Memories" in *Device Research Conference (DRC)*, 2019, Ann Arbor, MI.
- **S. K. Thirumala** and S. K. Gupta, "Reconfigurable Ferroelectric Transistors – Part I: Device Design and Analysis", in *IEEE Transactions on Electron Devices (TED)*, 66(6), pp. 2771-2779, 2019.

- **S. K. Thirumala** and S. K. Gupta, "Reconfigurable Ferroelectric Transistors – Part II: Application in Low Power Non-Volatile Memories", in *IEEE Trans. on Electron Devices (TED)*, 66(6), pp. 2780-2788, 2019.
- N. Thakuria, A. K. Saha, **S. K. Thirumala**, B. Jung and S. K. Gupta, "Oscillators Utilizing Ferroelectric-Based Transistors and their Coupled Dynamics", in *IEEE Trans. on Electron Devices (TED)*, 66(5), pp. 2415-2423, 2019.
- A. A. Saki, S. Lin, **S. K. Thirumala**, S. K. Gupta and S. Ghosh, "A Family of Compact Non-Volatile Flip-Flops with Ferroelectric FET", in *IEEE Trans. on Circuits and Systems I: Regular Papers (TCAS-I)*, 66(11) pp. 4219-4229, 2019.
- **S. K. Thirumala**, A. Raha, K. Ma, V. Narayanan, V. Raghunathan and S. K. Gupta, "Dual Mode Ferroelectric Transistor based Non-Volatile Flip-Flops for Intermittently-Powered Systems," in *Int. Symposium on Low Power Electronic Design (ISLPED)*, 2018, Bellevue, WA. ***(Best Paper Award Nominee).***
- **S. K. Thirumala** and S. K. Gupta, "Gate leakage in non-volatile ferroelectric transistors: Device-circuit implications" in *Device Research Conference (DRC)*, 2018, Santa Barbara, CA.
- A. Aziz, **S. K. Thirumala** et al., "Computing with ferroelectric FETs: Devices, models, systems, and applications", in *Design, Auto. & Test in Europe Conf. & Exhibition (DATE)*, 2018, Dresden, Germany.
- A. De, A. Iyengar, M. N. I. Khan, S. Lin, **S. K. Thirumala**, S. Gosh and S. K. Gupta, "CTCG: Charge-trap based camouflaged gates for reverse engineering prevention", in *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2018, Washington DC, USA
- A. Aziz, **S. K. Thirumala** et al., "Sensing in Ferroelectric Memories and Flip-Flops", In: Ghosh S. (eds) 'Sensing of Non-Volatile Memory Demystified', *Springer Nature*, pp. 47-80, 2018. ***(Book Chapter).***