LOW POWER CIRCUITS DESIGN USING RESISTIVE NON-VOLATILE MEMORIES

HUANG KEJIE

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Huang Kejie

12 July 2014

Acknowledgments

This thesis would not have been possible without the guidance, support, and love of many people to whom I would like to express my deepest gratitude.

First of all, I'd like to sincerely thank my supervisor, Prof. Lian Yong, for the great efforts he has put on my academic and guidance. His insightful and inspiring guidance has helped me grow as an independent researcher and good team player, which will continue to have profound influence on my future endeavor.

I am also very grateful to Prof. Zhao Rong, who provided me a research job in Singapore University of Technology and Design after I left Data Storage Institute. She has given me tremendous help to support my Ph.D study with useful data and insights for my research.

Lastly, I'd like to thank my parents Huang Difu and Sun Yudi, my sister Huang Xvxia, my brother-in-law Chen Quantong for their unconditional love wherever I am. I also want to express my gratefulness to my wife Ming Zhaoyan who gave birth and takes care of our daughter Huang Yuxi, and has helped me a lot in my Ph.D study. Our daughter was born during my Ph.D pursing period, and made my Ph.D study joyful and colorful.

Contents

List of Tables v

List of Figures vii

Chapter 1 Introduction 1

1.1 Motivation . . . 1

1.2 Resistive NVMs . . . 5

1.2.1 STT-MRAM . . . 7

1.2.2 PCM . . . 11

1.2.3 RRAM . . . 12

1.3 Resistive NVMs for Low Power . . . 14

1.3.1 Break Even Point (BEP) . . . 14

1.3.2 Using STT-MRAM as the Retention Register . . . 15

1.3.3 Integrating RRAM/PCM in FPGAs . . . 16

1.4 Related Works . . . 17

1.4.1 Non-volatile Latch/Flip-flop . . . 17

1.4.2 Non-volatile FPGAs . . . 20

1.5 My Contributions . . . 25

1.6 Thesis Organization . . . 27

Chapter 2 Non-volatile Latch/FF for Zero Standby Power Systems 28

2.1 Introduction 28
2.2 Proposed nvLatch/nvFF 29
2.2.1 The State Saving Mode 32
2.2.2 The State Restoration Mode 33
2.2.3 The Normal Latch Mode 34
2.2.4 Non-volatile Flip-flop 34
2.3 Simulation Results 36
2.3.1 Analysis the impact of VDD 37
2.3.2 The performance of the proposed nvFF 38
2.3.3 Analysis the impact of MTJ parameters 45
2.4 Summary 50

Chapter 3 Localized Array for Zero Sleep Power Systems 51

3.1 Introduction 51
3.2 Proposed Scheme 53
3.2.1 Circuit Architecture 55
3.2.2 Minimum Sleep Time 57
3.3 Localized STT-MRAM Array Design 58
3.3.1 Dual-Step-Write for Low VDD 59
3.3.2 Read-before-Write for Low Power 60
3.3.3 Pipelined Quad-Phase Write Scheme for High Speed 62
3.3.4 2σ Write Scheme for Low Power 66
3.3.5 Reference Resistance Generator 70
3.4 Simulation Results 77
3.4.1 Spice Simulation Results of the Proposed Array 78
3.4.2 Analysis of the Reference Resistance Generator 85
3.5 Summary 88

Chapter 4 Non-volatile Switch based FPGA 89

4.1 Introduction 89

4.1.1 Baseline 2D FPGA 90

4.1.2 Access Device 91

4.2 Proposed Storage Element 92

4.3 Proposed non-volatile FPGA 93

4.3.1 Proposed Crossbar Array and Switch Point 96

4.3.2 Proposed Look-Up Table 98

4.4 Layout and Area Estimation 99

4.4.1 Routing of the RRAM cells proposed nvFPGA 99

4.4.2 Area Estimation 101

4.5 Simulation Results 102

4.5.1 Write Power and Reliability 103

4.5.2 RC Delay Simulation Results 108

4.5.3 LUT Comparison 108

4.5.4 VPR Simulation Results 109

4.6 Summary 111

Chapter 5 Non-volatile SRAM-based FPGA 112

5.1 Introduction 112

5.2 Proposed nvSRAM based FPGA 113

5.2.1 Working Modes and Power Advantage 114

5.2.2 Multi-context FPGA and Area Advantage 116

5.3 Proposed Storage Element 117

5.3.1 Single Context nvSRAM 118

5.3.2 Multi-context nvSRAM 120

5.4 Simulation Results 123

5.4.1 Single Context Simulation Results 125

5.4.2 Multi-context Simulation Results 128
5.5 Summary 133

Chapter 6 Conclusions **134**

Acronyms **137**

List of Tables

1.1 Comparison of conventional and emerging memories. Most data other than those of RRAMs were taken from [1]. 6

1.2 Comparison among different approaches in the nvLatches/nvFFs. . 18

2.1 Description of the 90nm embedded MTJs and 45nm CMOS process. 36

2.2 The write energy comparison among different write approaches. . . 36

2.3 The performance of our proposed nvFF. 39

2.4 The performance comparison among the proposed nvFF, conventional nvFFs and the CMOS retention FF during saving operation. 42

2.5 The performance comparison among the proposed nvFF, conventional nvFFs and the CMOS retention FF during normal operation. 44

2.6 The estimated area comparison among the proposed nvFF, conventional nvFFs and the CMOS retention FF during normal operation. 45

3.1 Example of pipelined quad-phase saving scheme. Row clock is used in the table. 62

3.2 Example of pipelined quad-phase saving scheme with the 2σ write approach. Row clock is used in the table. 70

3.3 Description of the 45nm embedded MTJs process. 77

3.4 Per cell area overhead comparison among different retention schemes. The data in the '()' have included 6 transistors for scan chains. The number of transistors are estimated based on M=64 and G=8K. . . 83
3.5 The comparison among non-volatile Flip-flips and proposed schemes. The sleep energy and t_{BEP} are based on M=64. η is set to 10%. . . 85

4.1 The number of RRAM cells and the RRAM area partition of each FPGA block. 101
4.2 The simulation results of the RC delay among our proposed scheme, the conventional '1R' and SRAM schemes. 107
4.3 The speed, power and area comparison among different LUT schemes. 109

5.1 The control logic information of our proposed nvSRAM in different operation modes. 121
5.2 The parameters of the PCM used in the simulation. 123
5.3 The results comparison among the SRAM, proposed nvSRAM, [2] and [3]. 125

List of Figures

1.1 CMOS Front End Process and STT-MRAM Back End Process . . . 7

1.2 (a) Block diagram of a 1T1MTJ structure of an STT-MRAM cell. (b) Writing from P to AP state. (c) Writing from AP to P state . . 8

1.3 Phase change materials reversibly switch between amorphous and poly-crystalline states by electrical pulses. 12

1.4 Possible combinations of set and reset I-V curves. The combinations can be 'positive set, positive reset', 'positive set, negative reset', 'negative set, positive reset' and 'negative set, negative reset'. . . . 13

1.5 Break even point 14

1.6 Existing approaches using nvLatches. (a) Latch is used as write driver; (b) V_{th} drop in the write path; (c) Serial write. 19

1.7 (a) Conventional SRAM storage element to configure FPGAs (S-RAM); (b) non-volatile storage element to configure the switch transistor in FPGAs (1T2R); and (c) non-volatile storage element to replace the switch transistor and SRAM (2T1R, or '1R'). 20

1.8 (a) The high leakage current issue, and (b) the write disturbance issue in the conventional RRAM based non-volatile SP. The en-dash lines are the paths to program the RRAM cells, and dash-dot-dot lines are the sneak paths. 23

1.9 Equivalent circuit of a diode-less crossbar array. R_{cell} is the RRAM cell resistance under programming, R_L is the resistance of RRAM cells in LRS, M is the dimension size of the array, R_{p0} is the input parasitic resistance from the switch, metal, etc., R_{p1} is the parallelled input parasitic resistance, which is $R_{p0}/(M-1)$ for $V/2$ or $V/3$ write scheme and infinite for floating scheme, V_w, V_{b0} and V_{b1} are the writing voltage, and biasing voltages for the unselected word lines and bit lines, respectively. 24

2.1 Proposed STT-MRAM based non-volatile latch with two-phase write approach. 30
2.2 Two-phase write operation control logic to generate $S0$, $S0b$, $S1$ and $S1b$. 31
2.3 (a) Block diagram of the system level controller to save the states of the proposed nvLatches/nvFFs in the MTJs; (b) The four operation modes of the proposed nvLatches/nvFFs. 32
2.4 Proposed STT-MRAM based nvFFs. (a) The nvLatch is used as a master latch in the nvFF; (b) The nvLatch is used as a slave latch in the nvFF. 35
2.5 The supply voltage vs. the nvFF saving speed among three write approaches. 38
2.6 The nvFF saving speed vs. saving energy among three write approaches. 39
2.7 The simulation results of the proposed nvFF. It has two read operations (restoration), one write operation (saving) and two normal FF operations. 40

2.8 The corner simulation results among the proposed nvFF and the conventional nvFFs. Min corner: MTJ size -5%, Jc0 -5%, transistor width +5%; Max corner: MTJ size -5%, Jc0 -5%, transistor width +5%. A: [4]; B: [5]; C: [6]; D: [7]; E: [8]. 41
2.9 Sleep energy comparison among different nvFFs and conventional CMOS FFs. A: [4]; B: [5]; C: [6]; D: [7]; E: [8]; F: [9]. 43
2.10 The supply voltage requirement of the three write approaches vs. (a) $J_{c0}^{P\to AP}$, (b) size of the MTJ cells, (c) TMR, (d) RA, (e) γ, and (f) thermal stability Δ. 46
2.11 The required nvFF saving energy for the three write write approaches vs. (a) $J_{c0}^{P\to AP}$, (b) size of the MTJ cells, (c) TMR, (d) RA, (e) γ, and (f) thermal stability Δ. 48

3.1 Power consumption of (a) CMOS retention registers based approaches, (b) nvFF based approaches, and (c) proposed dedicated NVM array based approach. 52
3.2 (a) MTJ cells are distributed randomly in conventional nvFF schemes; (b) localized NVM arrays in our proposed scheme. 54
3.3 (a) Top diagram of the scan based approach to save the states of the registers in the local dedicated NVM array; (b) The four modes of our proposed low power system. 55
3.4 Proposed architecture with the localized non-volatile memory array. Left side of the diagram is the LSI block. Right side of the diagram is the NVM array with the memory controller. 56
3.5 (a) The access device in conventional write schemes significantly limit the write current passing through the MTJ. (b) Proposed dual-step-write scheme to achieve low VDD. 59

3.6 The sensing and comparing block diagram for the read-before-write scheme. 61
3.7 Proposed pipelined quad-phase control block diagram. 63
3.8 The array diagram of our proposed quad-phase writing approach. . 64
3.9 Block diagrams of our proposed pipelined scheme in the (a) i^{th}, (b) $(i+1)^{th}$, (c) $(i+2)^{th}$ and (d) $(i+3)^{th}$ system clocks. Each time two rows are active simultaneously. The active row addresses are highlighted in the figures. 65
3.10 Distribution of characteristic currents in STT-MRAM array [10]. . . 66
3.11 (a) The relationship between the first write current amplitude and the total write energy with our proposed write scheme. (b) The relationship between the standard deviation of I_{c0} in percentage and the write energy improvement with our proposed write scheme. 68
3.12 The distribution of the 2σ writing. 68
3.13 Proposed pipelined quad-phase control block diagram for the 2σ saving approach. 69
3.14 The block diagram of 8 memory channels for the 2σ saving approach. 69
3.15 Share the reference columns for two adjacent banks, reference1 is from bank1 and put closely to bank1 array, while reference2 is from bank2 and put closely to bank2 array and sense amplifier is shared by two banks of STT-MRAM array. 72
3.16 Example for concept of reference cell folding. (a) Reference cells connected in series before folding. (b) Folding the whole column of reference cells to a $N \times N$ array. (c) Final construction of the $N \times N$ reference array by connecting the folded points. 73

3.17 A circuit implementation of the equivalent $N \times N$ reference circuit when there are 2^{2n} cells in one reference column in which cells are averaged to obtain the equivalent resistance. 75
3.18 (a) The width of the access transistors vs. the write current that can pass through, (b) the VDD of the 1T1R scheme vs. the write current. 78
3.19 The waveform of the read-before-write and verify-after-write functions. 79
3.20 The relationship between the power comparison of our proposed two schemes and switching percentage of registers to be saved. 'Proposed 1' and 'Proposed 2' are the scheme without and with 2σ write approach, respectively. In this simulation, the standard deviations of the intrinsic switching current distribution were set to 5% and 10%, and the saving energy of our proposed scheme without 2σ write approach was set to the same for both intrinsic switching current distributions. The scan chain length is set to 64. 80
3.21 The relationship between the power reduction and operation clock cycles. In this simulation, the averaged switching activities of registers were set to 4% and 16%, and the standard deviation of the intrinsic switching current distribution was set to 10%. The scan chain length is set to 64. 81
3.22 The relationship between the power reduction and the scan chain length. In this simulation, the standard deviations of the intrinsic switching current distribution were set to 5% and 10%, and 50% of the registers were switched. 82
3.23 Normalized area overhead. The area is normalized to the minimum width transistors. 83

3.24 The sleep power consumption comparison among conventional structures and our proposed schemes. η is set to 10%. The sleep energy for MFF and nvFF are based on a single cell. A: [9]; B: [11]. 84

3.25 Python simulation results for distribution and deviation versus different equivalent reference block size. Distribution of the 16×16 equivalent reference array versus σ_P and σ_{AP} (a) without write failure and (b) with one AP cell stuck to P state; (c) Shift of the mean versus different equivalent reference block size; Deviation from the ideal mean versus (d) TMR ($R_{0P} = 4000$) and (e) R_P ($TMR_0 = 1$) with different slope of R_{AP}, where $I_{read} = 20\mu A$, $N = 16$; (f) Circuits simulation results for equivalent 16×16 reference block size. The standard deviations of both R_P and R_{AP} are set to 10% 86

4.1 A simple island style SRAM-based FPGA layout. 91

4.2 (a) The proposed non-volatile element to replace the FPGA routing switch and 6T SRAM. Adjacent non-volatile elements connecting to A or B share the same diodes. (b) A 3D schematic of the proposed non-volatile element. Metal line A or B may be routed at different layers depending on the routing direction. 92

4.3 (a) Top view structure of the proposed stacking RRAM based nvFGPA, (b) schematic diagram of the memory in our proposed nvFPGA system. The RRAM cells are arranged using '1D2R' crossbar array structure. 94

4.4 The schematic of our proposed '1D2R' based non-volatile FPGA. The crossbar structure is used for both CB and local interconnect. . 95

4.5 The schematic view of '1D2R' based (a) non-volatile crossbar array structure; (b) non-volatile switch point (SP). The non-volatile crossbar array is used in the CB and local interconnect. 96

4.6 The SB and CB structures used in the proposed nvFPGA. The switch box is based on Universal architecture. To simplify, the '1D2R' storage elements show only two RRAM cells in the dash line boxes. 97
4.7 Our proposed '1D2R' based non-volatile look-up table. It is an example of a 2-input LUT, and it can be extended to the other LUT size. 99
4.8 (a) The cross-section view of the switch in CB; (b) our proposed crossbar routing architecture to program the RRAM cells. 101
4.9 Area consumptions of the SRAM-based FPGA tile and our proposed '1D2R' based FPGA tile. The switch and SRAM area in our proposed '1D2R' based scheme is negligible because they are placed on top of the CMOS circuits. 103
4.10 A simulation diagram of the diode-less or transistor free crossbar array with parasitic resistance (R_p) in the word lines and bit lines. . 104
4.11 (a) The normalized write voltage across the selected RRAM cell; (b) the normalized required current at the input driver of the bit line or word line; (c) the write current analysis of different RRAM array schemes; (d) the normalized total write power. All results are normalized to the one single RRAM cell. 105
4.12 (a) The write voltage distribution in a 64×64 diode-less crossbar RRAM array due to the parasitic resistance in the word lines and bit lines; (b) the histogram plot of the normalized write voltage distribution in a 64×64 diode-less crossbar RRAM array; (c) the programming results in the 64×64 diode-less crossbar RRAM array. Black color represents successfully programmed cells and white color represents unprogrammed cells. 106

4.13 The write error rate comparison between $V/2$ write scheme and the scheme using diode as the selector. 107
4.14 (a) The delay simulation results; (b) the power simulation results; (c) the power and delay product results. The three schemes are simulated based on 20 MCNC test benches with VPR and the power model in [12,13]. 110

5.1 The proposed nvSRAM based FPGA Architecture. 6T SRAMs are replaced by our proposed nvSRAMs. SB, CB and CLB are switch block, connection block and configurable logic block, respectively. . 114
5.2 The power consumption of the (a) SRAM-based FPGA and (b) our proposed nvSRAM-based FPGA in different operation modes. . . . 115
5.3 (a) Conventional SRAM-based multi-context FPGA; (b) Proposed nvSRAM based multi-context FPGA. 116
5.4 The proposed single-context nvSRAM. The signals BL_p and BL_n are shared with other nvSRAMs in the same column. 118
5.5 The proposed single context in the (a) write mode, (b) read mode, and (d) FPGA execution mode. 120
5.6 The proposed multi-context nvSRAM. The signals BL_p and BL_n are shared with other nvSRAMs in the same column 121
5.7 A schematic of the nvSRAM 3D integration. The phase change material is deposited in the format of thin-film on the top of the CMOS transistors. 122
5.8 The 4-input LUT structure used to evaluate the proposed nvSRAM. 123
5.9 The power and delay simulation results of the proposed nvSRAM when loading the states from PCM cells to the latch. 124
5.10 The power consumption comparison among different LUT architectures. A: [2]; B: [3]. 126

5.11 (a) IV curve of the PCM cell in the amorphous state. (b) the PCM retention of the designs in [2,3], and our proposed nvSRAM. A: [2]; B: [3]. 127
5.12 The RTR simulation results of the proposed 8-context nvSRAM based 4-input LUT. 128
5.13 the 4-input LUT (a) active leakage power and (b) dynamic power comparison among the 6T SRAM, the designs in [2, 3], and the proposed nvSRAM. A: [2]; B: [3]. 129
5.14 The propagation delay comparison among the 6T SRAM, the designs in [2, 3], and the proposed nvSRAM based 4-input LUTs. A: [2]; B: [3]. 130
5.15 4-input LUT loading power comparison among the 6T SRAM, the designs in [2,3], and the proposed nvSRAM. A: [2]; B: [3]. 130
5.16 8-context 4-input LUT power comparison among the designs in [2, 3], and the proposed nvSRAM. All of the results are normalized to the SRAM based 8-context 4-input LUT under the same conditions. The average LUT switching frequency is set to 10MHz. (a) The power consumption versus the ratio of idle time and active time. The active time is set to $1ms$. (b) The power consumption versus the active time. The ratio of idle time and active time is 0.9. A: [2]; B: [3]. 131
5.17 Area comparison among the 6T SRAM, the design in [2] and our proposed nvSRAM. The area is normalized to the single context 6T SRAM. A: [2]; B: [3]. 132

Abstract

The increasing leakage current in the complementary metal oxide semiconductor (CMOS) circuits due to technology nodes scaling down has been one of the critical issues in the current generation digital circuits and field programmable gate arrays (FPGAs). There are growing research effort in the integration of resistive non-volatile memory (NVM) cells to achieve low power high performance circuits. Although the reported circuits help to minimize the sleep power consumption of the system, there are various drawbacks that limit the performance or reliability of the circuits.

This dissertation presents new schemes for both digital circuits and FPGAs to achieve low power and high performance circuits. The new non-volatile flip-flops (nvFFs) and localized NVM array based on spin transfer torque MRAM (STT-MRAM) are proposed to retain the states of registers during standby. Both designs are targeting for the low VDD and low write power. The nvFF can be designed as a standard cell to be compatible with digital design flow thus the design cycle could be greatly reduced. The localized NVM array could further reduce the power consumption with higher density. The non-volatile storage elements proposed for the non-volatile FPGAs (nvFPGAs) are targeting for the high reliability, high density and low power. Compared to the conventional nvFPGAs, the reliability is significantly improved and power is greatly reduced, while compared to the static random access memory (SRAM) based FPGAs, the FPGA area and power could be greatly reduced.

Chapter 1

Introduction

1.1 Motivation

CMOS logic technology nodes have been scaled down for more than 40 years [14–18] to achieve higher density and better performance. According to Moore's law, the transistor dimensions are scaled down by 30% (0.7×) every technology generation, and therefore increases operating frequency by about 40% (1.4×) [19]. To keep electric field constant and maintain a high drive current, supply voltages and threshold voltages have been scaled down in proportion to metal oxide semiconductor field effect transistor (MOSFET) device dimensions, resulting in an exponential increase in sub-threshold leakage [20, 21]. Consequently, the stand-by leakage power dissipation is rapidly becoming a substantial contributor to the total power dissipation in memories or state retention in duty cycled systems. For those standby-power-critical systems, which have long idle times punctuated by bursts of activity, such as cell phones, tablet laptops and wireless sensor networks, this standby power consumption reduces the effectiveness of duty-cycling. Large standby leakage power poses significant challenge to achieve the goal of low power.

To address the high standby leakage power issue in battery powered sys-

tems, increasing battery capacity and harvesting energy from the environment are two possible solutions. However, the energy density of the battery is improved by less than 7% every year [22]. Alternatively, the energy scavenging could compensate the leakage power loss during standby. However, according to the research records from National Renewable Energy Laboratory (NREL), the energy harvest efficiency gains by less than 1% every year [23]. Therefore, other solutions are required to reduce the leakage power.

There are four main sources cause the leakage current in a CMOS transistor [24]: 1. Reverse-biased junction leakage current; 2. Gate induced drain leakage; 3. Gate direct-tunneling leakage; 4. Subthreshold (weak inversion) leakage. Among these four leakage sources, "gate induced drain leakage" is not a component of the leakage of an OFF state transistor. The "subthreshold leakage" is the drain-source current of a transistor operating in the weak inversion region, in which the diffusion current of the minority carriers dominates. The magnitude of the subthreshold current is a function of the temperature, supply voltage, device size, and the process parameters [24]. Among these parameters, the threshold voltage (V_{th}) plays a dominant role.

In current CMOS technologies, the relatively low V_{th} due to scaling makes the subthreshold leakage current (I_{SUB}) much larger than the other leakage current components. I_{SUB} is calculated by using the following formula [24]:

$$I_{SUB} = \frac{W}{L}\mu\nu_T^2 C_{sth} e^{\frac{V_{GS}-V_{th}+\eta V_{DS}}{n\nu_T}}(1 - e^{-\frac{V_{DS}}{\nu_T}}) \tag{1.1}$$

where W and L are the transistor width and length, respectively. $\nu_T = kT/q$ is the thermal voltage at the temperature T, $C_{sth} = C_{dep} + C_{it}$ denotes the summation of the depletion region capacitance per unit area of the MOSFET gate and the interface trap capacitance per unit area of the MOSFET gate, μ and η denote the carrier mobility and the drain induced barrier lowering (DIBL) coefficient [25],

respectively. n is the slope shape factor and is calculated as:

$$n = 1 + \frac{C_{sth}}{C_{ox}} \tag{1.2}$$

where C_{ox} denotes the gate input capacitance per unit area of the MOSFET gate. When a transistor is in the OFF state (V_{GS}=0), the subthreshold leakage can be reduced by increasing V_{th} or reducing V_{DS}. Multiple threshold voltage levels [26, 27], well-bias control [28, 29] have been used to increase V_{th}, and stack effect based method [30], VDD reduction and power gating (PG) [31–34] have been used to reduce V_{DS}. Among these techniques, PG is one of the most effective means, in which inactive blocks are turned off by inserting a high threshold sleep transistor between the power supply and digital circuits. This scheme is efficacious for reducing leakage power when a large scale integrated (LSI) function block is in the sleep state. However, part of the blocks need to be powered on due to the volatile nature of retention registers. Therefore, the leakage still exists in both logic circuits and decoupling capacitors. Moreover, the wake-up process, *i.e.*, transition from sleep to active mode, involves a large rush current through the sleep transistors. Due to the inductance from power rails and packages, this rush current can cause Ldi/dt noise, which is manifested as ground bounce when a footer is used, or as VDD fluctuation when a header is used [35–37]. PG control should be carefully designed so that the integrity of the data in retention elements is guaranteed.

As the counterpart of the application specific integrated circuits (ASICs), FPGAs have been rapidly growing in the integrated circuit (IC) market share due to the post-fabrication reconfigurability, fast time to market, design fault tolerant, and low development cost. Hence SRAM-based FPGA logic circuits have been under focused development in the past 20 years [38–41]. SRAMs are used to configure logics and routing information to realize the required functionalities. FPGA

interconnects including switch blocks (SBs), connection blocks (CBs), and configuration SRAMs account for around $80-90\%$ of the total area, delay and power. In contrast, the logic blocks (LBs) occupy only $10-20\%$ of the total area [42–44]. Thus, reducing the length of interconnects and improving the configuration memory cells are the key of the FPGA design.

Additionally, SRAM-based FPGAs require reprogramming each time when powering on, because SRAMs lose the configuration information after powering down. Moreover, as CMOS technology nodes scale down to $90nm$ and below, the leakage power has rapidly become the dominant component of total power dissipation [45, 46]. As a result, SRAM-based FPGAs suffer from slow power-on speed, high power-on power and leakage power. The high power-on power and slow power-on speed limit the power-off opportunities of the FPGA. In other words, it is not possible to power off the FPGA when the idle time between two events is short. Moreover, additional external NVM is required to store the configuration information.

Integrating NVMs in the CMOS circuits is an effective solution to reduce the leakage current. By replacing the dynamic random access memory (DRAM) or SRAM in FPGAs, or retaining the states of the registers into the NVMs, the whole system can be fully powered off without losing information. However, the conventional nvFF and nvFPGA schemes suffer from various weaknesses including high VDD, high write power, high active leakage power, low read/write reliability, etc. The details of the related works will be discussed in Section 1.4. Therefore, new integration solutions and architectures are required to address various weaknesses in the conventional resistive NVM based flip-flops (FFs) and FPGAs. In this dissertation, we will propose several schemes to design the non-volatile latch (nvLatch) or the localized array to replace the retention registers for the stand-by power free systems. In addition, new FPGA storage elements/architecures are

proposed based on the resistive NVMs to achieve the low power, high performance and high density.

1.2 Resistive NVMs

The conventional FLASH memory has been used to achieve low power systems. Each memory cell in a FLASH memory consists of only one MOSFET with an additional floating gate. In spite of the wide application of FLASH memories in commercial products, e.g. digital cameras, memory sticks and tablets, the current FLASH memory technology has various disadvantages. The primary limitation of FLASH memory is that while their design is superb for $5V$ operation, while the standard logic level has decreased from $5V$ to $3.3V$ to $1V$ and will eventually decrease to $0.5V$ in the coming years. FLASH memories (based on the Fowler-Nordheim tunneling) cannot reliably function at $0.5V$. The remedy by inserting internal 'charge pumps' for programming will decrease yields, increase cost and failure mechanisms [47]. The other disadvantages are much longer write and erase times and much lower write/erase cycles ($1e5$) than DRAM, as shown in Table 1.1). In addition, the FLASH memory technology will touch the miniaturization limit when the lateral feature size of DRAMs and FLASH memories shrinks down to $21nm$ (for DRAM technology 2016 and for FLASH technology 2013) [1, 48, 49]. In a summary, the conventional FLASH, is facing limitations of the scale down, endurance, speed and operation voltage.

Fortunately, the emerging memories may address the limitations of the FLASH memory [1, 48, 49]. There are more than a dozen non-volatile memories have been considered as emerging memories. For example, resistive random access memorys (RRAMs) [50–61], magnetic RAMs (MRAMs) [11, 62–66], phase change memorys (PCMs) [67–74], carbon nanotube memory [75], racetrack memory [76, 77], ferroelectric RAMs (FeRAMs) [78], millipede memory [79], molecu-

Table 1.1: Comparison of conventional and emerging memories. Most data other than those of RRAMs were taken from [1].

Type	Baseline Technologies				Prototypical Technologies		
	SRAM	DRAM	NOR-Flash	NAND-Flash	MRAM	PCM	RRAM
Cell elementes	6T	1T1C	1T	1T	1T1R	1T(D)1R	1T(D)1R
Storage Mechanism	Latch	Stack /trench capacitor	Floating gate /charge trap	Floating gate /charge trap	Magneti - zation	phase-change	resistance change
Feature size	45nm	36nm	90nm	22nm	65nm	20nm	
Cell area	$140F^2$	$6F^2$	$10F^2$	$4F^2$	$20F^2$	$4F^{2\ b}$	$4F^{2\ c}$
Write/ erase time	0.2ns/ 0.2ns	<10ns/ <10ns	1us/ 10m-s	1ms/ 0.1ms	35ns/ 35ns	10ns/ 100ns	5ns/ 5ns [83]
Endurance (Cycles)	$>1e16$	$>3e16$	$>1e5$	$>1e4$	$>1e12$	$1e9$	$>1e10$ [84]
Write Operation Voltage (V)	1	2.5	10	15	1.8	3	
Write Energy (J/bit)	$5e{-}16$	$4e{-}15$	$1e{-}10$	$>2e{-}16$	$2.5e{-}12$	$6e{-}12$	

lar memory [80], programmable metallization cells (PMCs) memory [81], DNA memories [82], etc. Among these memories, RRAMs, MRAMs and PCMs have been considered as emerging memories to potentially overcome the limitations of DRAMs and FLASH memories. Unlike FLASH and DRAM which use charge as the information carrier, RRAMs, MRAMs and PCMs rely on non-volatile, resistive information storage in the memory cells, thus exhibit zero standby power consumption, and hold the potential to scale to much smaller geometries than charge memories. These characteristics, coupled with their CMOS-compatibility, fast read/write speed, high density and write endurance, make resistive memories promising candidates for storing the register information with no off-state leakage current. They also provide an excellent opportunity to achieve high speed, high density, instant power on and superior energy efficiency FPGAs. The comparison between the conventional and emerging memories is given in Table 1.1.

A cross section schematic shown in Fig. 1.1 illustrates the integration process of the resistive NVMs in the CMOS process. The CMOS front end process includes the bottom substrate, CMOS layers, and metal layers. The CMOS-

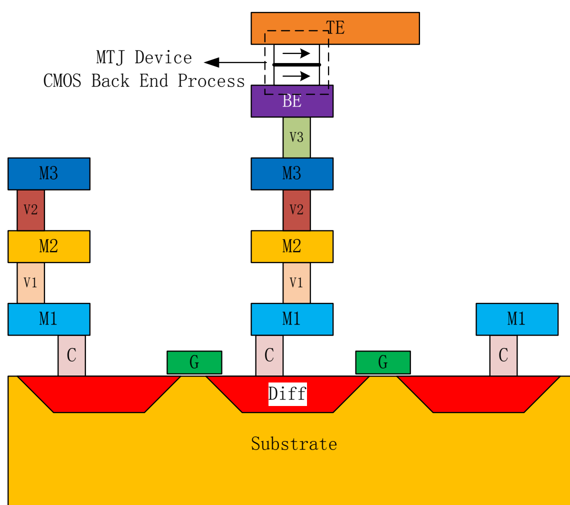


Figure 1.1: CMOS Front End Process and STT-MRAM Back End Process

compatible back end process deposits the resistive NVM layer between two metal layers (top electrode (TE) and bottom electrode (BE)). magnetic tunnel junction (MTJ) is used in this example, but it worths noting that the MTJ layer could be RRAM, PCM or other resistive NVMs.

1.2.1 STT-MRAM

MRAMs that have been considered as possible candidates to replace several types of current memories such as embedded SRAMs, DRAMs and FLASH memories. There are two main types of MRAMs have been developed: field-writing MRAM and STT-MRAM. The field writing MRAM is written by a magnetic field around the current line. The primary issue of field-write MRAM is the high write current, which makes scaling down difficult.

STT-MRAM has combined the advantages of SRAMs (high speed), DRAMs (scalability) and FLASH memories (non-volatility) [85], promising it as a next-generation memory candidate. However, the OFF/ON ratio is big concern since low resistance ratio leads to low read reliability. Another concern is the high

energy dissipation during operation.

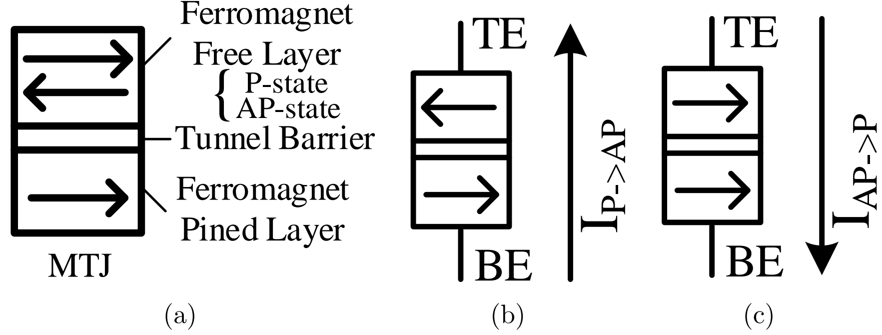


Figure 1.2: (a) Block diagram of a 1T1MTJ structure of an STT-MRAM cell. (b) Writing from P to AP state. (c) Writing from AP to P state

A typical STT-MRAM structure is illustrated in Fig. 1.2(a). The MTJ device has a low resistance of R_P when the magnetic moment of the free layer is parallel to that of the pinned layer (P-state) and a high resistance of R_{AP} when the free layer moment is oriented anti-parallel to the pinned layer moment (AP-state). When the current flows from BE to TE, the MTJ switches from P-state to AP-state ($P \rightarrow AP$), as shown in Fig. 1.2(b). If the current flows in the opposite direction, the MTJ changes from AP-state to P-state ($AP \rightarrow P$), as shown in Fig. 1.2(c). The tunnel magnetoresistance (TMR) ratio of an MTJ cell is defined as $TMR = (R_{AP} - R_P)/R_P$. The resistance of a STT-MRAM cell can be expressed as:

$$R_{MTJ} = |I_{MTJ}| * K_{MTJ} + R_0 \tag{1.3}$$

where I_{MTJ} is the current goes through the MTJ cell in either direction, K_{MTJ} is the slope of R_{MTJ}, R_0 is the zero current resistance. K_{MTJ} has two values K_P and K_{AP}, which are the slope of R_P and R_{AP}, respectively. R_0 also has two values R_{0P} and R_{0AP}, which are the R_P and R_{AP} value when $I_{MTJ} = 0$.

Usually the distributions of the values of R_P and R_{AP} follow a Gaussian

distribution [86, 87] which can be written as

$$f(R) = \frac{1}{\sqrt{2\pi(\sigma_{MTJ} * R_{MTJ})^2}} e^{-\frac{(R-R_{MTJ})^2}{2(\sigma_{MTJ}*R_{MTJ})^2}} \quad (1.4)$$

where σ_{MTJ} is the deviation in percentage for R_{AP} or R_P.

At a finite temperature, thermal agitation plays an important role in reducing the switching current at long switching pulses (>10ns) [88, 89]. In this slow thermal activated switching regime, the switching pulse width is dependent on the switching current amplitude and thermal stability factor $\Delta = K_u V / k_B T$ of the free layer, where k_B is the Boltzmann's constant, T is the temperature, and $K_u V$ is anisotropy energy. A model that describes the correlation of the parameters was proposed by Néel-Brown [90]:

$$J_c = J_{c0}(1 - \frac{1}{\Delta} ln(\frac{T_{WR}}{\tau_0})) \quad (1.5)$$

where T_{WR} is the pulse width of switching current, τ_0 is the inverse of the attempt frequency, and J_{c0} is the intrinsic switching current density. The intrinsic current density J_{c0} required for current driven magnetization reversal in an MTJ with the magnetization in the film plane can be expressed as

$$J_{c0} = (\frac{2e}{\hbar})(\frac{\alpha}{\eta})(t_F M_s)(H_k + 2\pi M_s) \quad (1.6)$$

where M_s and t_F are the magnetization and thickness of the free layer respectively, α is the damping constant, and H_k is the effective anisotropy field including magneto-crystalline anisotropy and shape anisotropy. The spin transfer efficiency η, is a function of the current polarity, polarization, and the relative angle between the free and pinned layers. When $J_c > J_{c0}$, an initial stable magnetization state of the free layer along the easy axis becomes unstable at zero temperature and the magnetization enters a stable precessional state or a complete reversal occurs. From (1.5), one can estimate the critical current density J_{c0} by extrapolating the experimentally observed switching current density J_c at $t = \tau_0$.

For fast precessional switching in nanosecond (ns) regime (less than a few ns), the required switching current is several times greater than the instability current J_{c0} [88, 89]. The switching current density can be estimated as

$$J_c = J_{c0} + \frac{Cln(\pi/2\theta)}{T_{WR}} \tag{1.7}$$

where θ is the initial angle between the magnetization vector of the free layer and the easy axis, and C is the fitting parameter. At finite temperature, θ is a thermal distribution.

The probability that a data of the STT-MRAM is switched for a given time t at least unit time is expressed by using the Poisson distribution [88, 91, 92]:

$$f_{switch}(I, t) = 1 - e^{-\frac{t}{t_p}} \tag{1.8}$$

where t_p is derived from (1.5), $I = J_c \times A$ is the writing current amplitude, and A is the area of the MTJ.

Read disturbance is related to the margin between the read and write currents. The probability that the read disturbance occurs at a given read current I_{read} is given by

$$P = \int_0^{I_{read}} f_{switch}(I)dI \tag{1.9}$$

More intuitively, if the read disturbance rate of a M Gb STT-MRAM is 1ppm, P is smaller than $1/(N \times M \times 1024^3 \times 10^6)$.

To achieve low read disturbance, i.e. accidental writing of a bit while trying to read the bit, the read current has to be much smaller than the median critical current [88]. Assuming that all other parameters remain the same but with 5% deviation in the median critical current, the read disturbance probability is increased by several orders of magnitude at a specified read current [88]. The read current has to be reduced to about 20% the median critical current to maintain the same level of read disturbance error rate.

1.2.2 PCM

It has been more than four decades since the first idea to use phase-change materials in memory devices [93, 94]. However, the low material quality and high power consumption of this technology prevented it from the commercialization. In the last few decades, the great improvement in the semiconductor manufacturing technology and the quality of PCM provides the phase-change material based NVMs a second life.

The PCM provides the benefits of high density [95], high scalability [96], low cost [97] and high resistance ratio (R_H/R_L) [98, 99]. The $4F^2$ small PCM cell size based on $20nm$ technology node has been achieved by Samsung [100, 101]. The high resistance ratio between the Amorphous ($RESET$) and Crystalline (SET) states increases the read reliability and the sense speed. Moreover, PCM also has the potential to achieve nano-second [102–104] and sub micro-ampere current switch [105]. PCMs are expected to replace NOR-FLASH memories in the memory market at present. Recent progress in PCM technology has provided a clear demonstration of the excellent scaling potential to and beyond the $16nm$ generation [70].

The typical PCM structure is a chalcogenide layer (*i.e.*, Ge2Sb2Te5, or GST) sandwiched between a metal contact and a heat electrode. Phase-change materials exhibit an ability for reversible phase transition between the Amorphous and Crystalline phases with the help of Joule heating. This phase transition brings about a change in the resistance as well as the reflectivity. The heat produced by the passage of an electric current through the heating element is used to transform the material between the poly-crystalline and amorphous states. As shown in Fig. 1.3, if the chalcogenide material is quickly heated (melting) and quenched (rapid cooling), it will be reset to the amorphous state (high resistance state, R_H, binary '0'). On the other hand, if the material is held in its crystallization

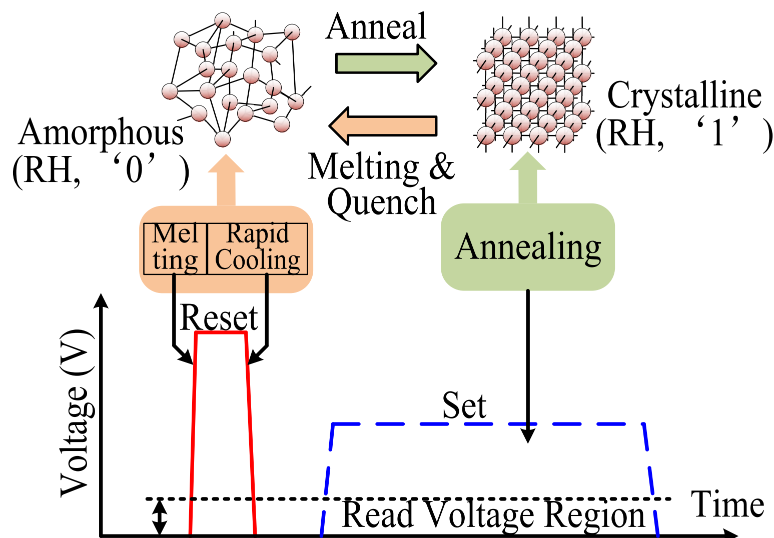


Figure 1.3: Phase change materials reversibly switch between amorphous and poly-crystalline states by electrical pulses.

temperature range for some time (annealing), it will be set to the poly-crystalline state (low resistance state, R_L, binary '1'). The cell resistance between the poly-crystalline and amorphous states may have orders difference. Therefore, as shown in Fig. 1.3, $RESET$ (quickly heating and quenching) requires short pulse and high voltage, while SET (holding in crystallization temperature) requires long pulse and medium voltage. To avoid unintended write, the read voltage should be much lower than the SET voltage.

1.2.3 RRAM

Resistive NVMs generally include all types of NVMs using two or more distinctive resistance states as the binary numbers '0' and '1'. In principle, PCMs and MRAMs could be considered as resistive NVMs as well. The resistive switch in each memory cell consists of a switching layer sandwiched by TE and BE. This capacitor-like switching cell is characterized by two distinctive resistance states: a high resistance state (HRS) and a low resistance state (LRS). The basic idea of the

RRAM switch mechanism is that a dielectric, which is normally insulating, can be made to conductive through a filament or conduction path. The RRAM can be reversibly switched between HRS (filament broken) and LRS (filament reformed) by applying an appropriate voltage. Reversible resistive switching was observed in various materials, such as $Nb2O5$, $Al2O3$, $SiO2$ and $TiO2$ [106–110].

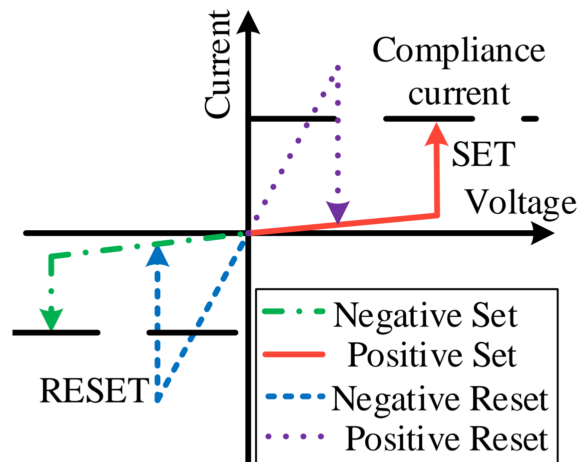


Figure 1.4: Possible combinations of set and reset I-V curves. The combinations can be 'positive set, positive reset', 'positive set, negative reset', 'negative set, positive reset' and 'negative set, negative reset'.

Several possible combinations of set and reset curves are shown in Fig. 1.4. For unipolar switching, the lower voltage acts as set and the higher voltage in the same direction acts as reset, whereas for bipolar switching only 'negative set, positive reset (eightwise)' or 'positive set, negative reset (counter eightwise)' is possible [111, 112].

RRAM has the potential to become the front runner among the emerging NVMs. Compared to PCM, RRAM operates at a faster switching speed (less than $10ns$). Compared to MRAM, it has a simpler process, smaller cell structure ($4F^2$ metal insulator metal (MIM) stack), and higher resistance ratio. Compared to FLASH memory, it has a much lower switching voltage and much higher switching speed. The $30nm$ cell size of the RRAM has been demonstrated by Industrial

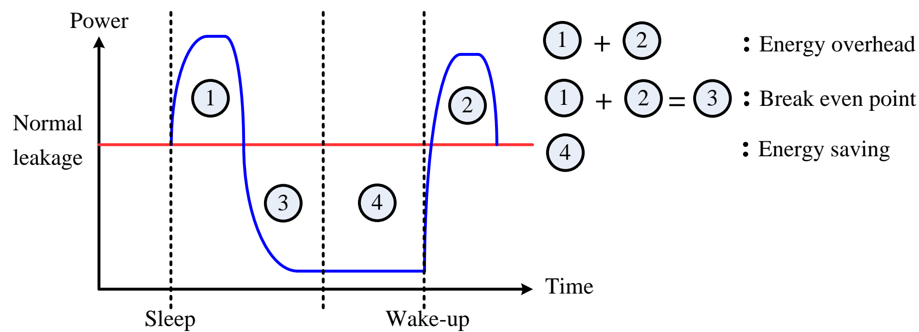


Figure 1.5: Break even point

Technology Research Institute (ITRI) recently [79], and it is believed that the oxygen motion may take place in regions as small as $2nm$ [113].

1.3 Resistive NVMs for Low Power

1.3.1 Break Even Point (BEP)

Before the discussion of the applications of the emerging NVMs, we introduce the concept of break even point (BEP) which is an important merit to judge the power reduction benefit with the new NVMs. Most microelectronic systems spend considerable time in a standby state. The energy consumed by the non-volatile memory to save or restore the information must be considered carefully. If there no cost of transiting to and from a standby power state, the greedy policy of entering the low power state as soon as the system is idle may be adopted. Otherwise, the expected duration of the standby state must be accurately calculated and taken into account when devising a power management policy. When the sleep period is longer than BEP as shown in Fig. 1.5, the system could be power off to reduce the leakage power. BEP is defined by the time when the reduced sleep energy (area 3) equals to the energy required to save and restore the system (area 1 and 2, respectively). Therefore, the standby leakage power in area 4 is reduced.

Otherwise, if the system is powered off when the standby time is short than BEP, the total power increases. Hence, the low saving and restoring energy should be the primary consideration when integrating NVMs in CMOS circuits to achieve zero standby power system.

1.3.2 Using STT-MRAM as the Retention Register

As the discussed in Section 1.2, the intrinsic features of the three NVMs determine their applications in the integration in CMOS circuits. PCM and RRAM have simpler process and lower cost than STT-MRAM. However, the high program voltage of PCM [114, 115] and RRAM [116, 117] limits their integration in digital circuits, especially when the supply voltage scales down to $1V$ and below. Among these three candidates, STT-MRAM exhibits the advantages of fast switching speed between parallel (P) and anti-parallel (AP) states [58, 63, 118], and low switching current [118] or voltage [64], making it a potential candidate to be integrated with deep sub micron CMOS processes without a level shifter. Therefore, STT-MRAM is the best choice among these three candidates to replace the retention registers to achieve zero standby digital systems. This is because the states of the digital systems have to be saved to the NVM cells each time when powering down, and read them back to the digital systems each time when powering on. Hence, fast read/write speed and low read/write power are crucial to reduce the BEP. In other words, STT-MRAM allows the digital systems to be powered off in a much shorter idle period between two activities.

The state-of-art design to retain the states of the FFs during standby is the nvFF scheme, which has combined the FF and NVM in one cell. Hence it could be designed as a standard cell to design cycle. Saving the states to a NVM array is another solution, which could adopt more technique to improve the performance and reduce the BEP as well. But it has to be elaborated upon the size, area,

architecture, etc. Otherwise, the total power may be increased.

1.3.3 Integrating RRAM/PCM in FPGAs

The emerging resistive NVM technologies with the advantages of high density, near zero power-on delay, and superior energy efficiency have provided an excellent platform to advance the FPGA technology. Since FPGAs only need to be programmed once during configuration, the slow write time and high write voltage may not an issue in such applications. In contrast, the low process cost and the high reliability due to high resistance ratio make PCM/RRAM more attractive in the FPGA applications.

Among them, RRAM becomes the front runner among resistive NVMs due to its fast switching speed (less than 10ns [59]), small cell size ($4F^2$ [119]), high resistance ratio [120], low switching voltage [121] and current [122], and compatible to current CMOS processes, etc. The six order resistance ratio of the RRAM has been demonstrated in [123]. These merits enable RRAM as a universal replacement of the SRAM and switch in the SRAM-based FPGAs. The states of the RRAM cells are configured as ON/OFF switches initially in the routing and logic blocks, thus achieving various functions as the conventional SRAM-based FPGAs. The new nvFPGA will achieve much higher density and greater reduction of the RC delay in the routing. Moreover, the RRAM-based switch also addresses the v_{th} drop issue in the SRAM-based FPGAs.

PCM could be a universal NVM [68] as well that provides the benefits of high density [95], high scalability [96], low cost [97] and high resistance ratio [99]. The $4F^2$ small PCM cell size based on $20nm$ technology node has been achieved by Samsung [101]. The high resistance ratio between the amorphous ($RESET$) and poly-crystalline (SET) states increases the read reliability. Moreover, PCM also has the potential to achieve nano-second [102] and sub micro-ampere current

switch [105]. Coupling with its low cost process, it is also a good choice to replace the SRAM in the conventional FPGAs. To replace the switch directly requires high resistance difference between the amorphous state and crystalline state, but it is only $2-3$ orders currently.

Therefore, both RRAM and PCM could be design as non-volatile SRAMs (nvSRAMs) to configure the single-context FPGAs, or even multi-context FPGAs to achieve low power and high density. In addition, the high resistance ratio of the RRAM enables it a universal replacement of the switches and SRAMs to attain high performance and high density nvFPGAs.

1.4 Related Works

1.4.1 Non-volatile Latch/Flip-flop

Integrating the NVM into the digital circuits is an effective solution to retain the states of the FFs, thus the whole system can be fully powered off. In particular, it is only necessary for all FFs to be nonvolatile if the function blocks are clock-synchronized. Employing nvFF can provide a more efficient use of energy in System-on-Chips (SOCs) for standby-power-critical and quick-startup applications, especially the battery powered appliances. The nvFFs could be designed as standard cells to be compatible with the digital design flow, thus the design cycle could be greatly reduced.

Many nvFF works have been reported [4–9, 124, 125] to integrate NVMs in the latches or FFs to achieve zero standby power consumption systems. Though their proposed circuits have efficiently reduced the sleep power consumption of the system, their performance is limited by various weaknesses, such as updating MTJs states every clock cycle, latch is used as write driver, the "source degeneration" effect in the write path, serial write, etc. Table 1.2 summarizes different approaches

Table 1.2: Comparison among different approaches in the nvLatches/nvFFs.

nvFFs	Saving speed	Saving power	Latch speed	Latch size	VDD	Preferred
Update MTJs every clock cycle	Low	High	Low	-	-	
Update MTJs before sleep	High	Low	High	-	-	√
Serial write	High	High	-	Large	High	
Parallel write	High	Medium	-	Medium	Low	
Two-phase write	Low	Low	-	Small	Low	√
MTJs inside the latch	-	-	Low	-	-	
MTJs outside the latch	-	-	High	-	-	√
Latch as the write driver	-	-	Low	Large	-	
Latch as the sense amplifier	-	-	High	Small	-	√

implemented in those nvFFs.

There are growing research efforts in the integration of MTJs in the latches or FFs [4–9]. Although the reported circuits help to minimize the sleep power consumption of the system, there are several drawbacks limit the nvFFs performance as summarized below.

1. **The requirement of updating MTJ states every clock cycle** [4] and [8]. Updating MTJ states every clock cycle does not necessary reduce the sleep power consumption of the system. On the contrary, it increases the power consumption and reduces the speed during normal FF operation. Moreover, it also reduces the endurance of the MTJs. The states of the FFs only need to be retained in the MTJs during sleep mode.

2. **The requirement of latch as write driver** [5, 8, 9]. The use of the latch as part of the write driver may require large size transistors in the latch.

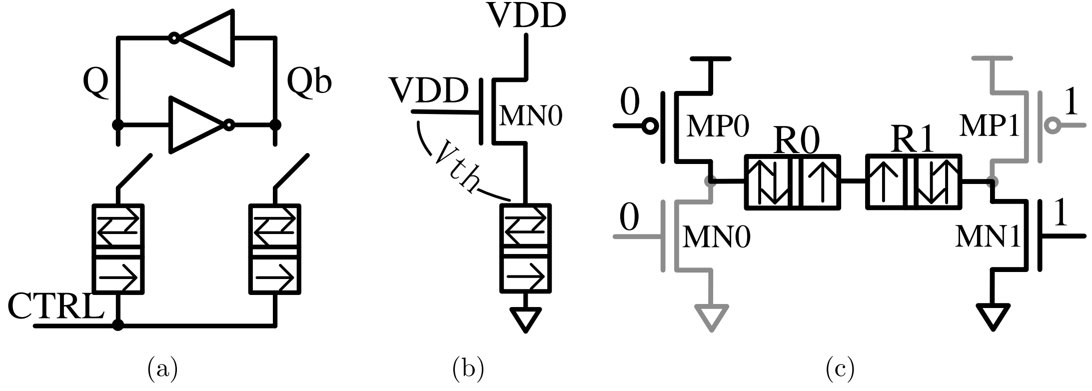


Figure 1.6: Existing approaches using nvLatches. (a) Latch is used as write driver; (b) V_{th} drop in the write path; (c) Serial write.

As a result, it not only slows down the latch operation speed due to the large parasitic capacitances, but also affects data integrity. For example, in Fig. 1.6(a) the write voltage on CTRL may flip the state of the latch before saving the state into the MTJs.

3. **The "source degeneration" effect in the write path** [5,6,8]. As shown in Fig. 1.6(b), the "source degeneration" effect caused by V_{th} drop in the write path limits the write current when the source of the transistor is connected to the MTJ. Therefore, higher VDD is required to pump in sufficient current into the MTJs to switch their states, resulting in high power consumption and area.

4. **The serial write approach** [4,6,7]. The serial write approach to store the states of FFs into the MTJs, as shown in Fig. 1.6(c), requires VDD to be higher than $V_{P\to AP} + V_{AP\to P}$, where $V_{P\to AP}$ and $V_{AP\to P}$ are the $P \to AP$ and $AP \to P$ switching voltages, respectively. Therefore, the serial write approach requires either high VDD or low $V_{P\to AP}$ and $V_{AP\to P}$. The high VDD may result in high power consumption and scaling down difficulty. Low $V_{P\to AP}$ and $V_{AP\to P}$ may face long switching time.

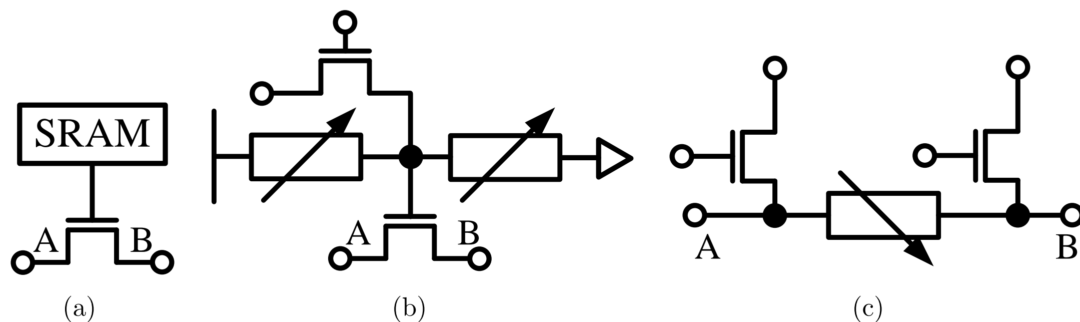


Figure 1.7: (a) Conventional SRAM storage element to configure FPGAs (S-RAM); (b) non-volatile storage element to configure the switch transistor in FP-GAs (1T2R); and (c) non-volatile storage element to replace the switch transistor and SRAM (2T1R, or '1R').

5. **MTJs are embedded in the latch** [4,7–9]. It may slightly reduce the FF operation speed by embedding the MTJ cells inside the latch.

1.4.2 Non-volatile FPGAs

To address the leakage issue in the SRAMs, people are turning their attention to the emerging resistive NVM technologies. With the advantages of near zero power-on delay, dynamic reconfiguration, and superior energy efficiency, the nvFPGAs have been the object of intense development in the past few years. Many works have been reported to integrate RRAM [126], PCM [2,127] or STT-MRAM [128] in the FPGA circuits. FPGAs have the opportunity to significantly reduce the area, power and delay with emerging resistive NVMs. We categorize the conventional FPGA configuration memory technologies into three, i.e. SRAM, 2T1R, 1T2R, as shown in Figs. 1.7(a), 1.7(c) and 1.7(b), respectively.

1. **SRAM**.The SRAM-based FPGA storage element to configure the FPGA function as shown in Fig. 1.7(a) has three key weaknesses. First, SRAM-based FPGAs have to load the configuration information every time when powered on, which reduces the effectiveness of the off/on duty-cycling. Sec-

ond, to keep electric field constant and maintain a high drive current, supply voltages and threshold voltages have been scaled down in proportion to MOSFET device dimensions, resulting in an exponential increase in subthreshold leakage [20, 21]. Hence the leakage power dissipation of SRAM-based FPGAs is rapidly becoming a substantial contributor to the total power dissipation of FPGAs. The last one is the interconnects include SBs, CBs, and configuration SRAMs account for more than 80% of the total area, delay and power of the FPGAs [43, 44].

To improve the performance and reduce the area of FPGA, the NVM-based solutions are under focused development. There are two main solutions: 1T2R scheme and 2T1R scheme. However, both solutions have various weaknesses that limit their feasibility to be integrated in FPGAs. The detailed will be discussed in the following.

2. **1T2R**.The '1T2R' scheme as shown in Fig. 1.7(b) was reported in [2,3,129–131] to replace the conventional SRAM cell with the NVM-based storage element to have the advantages of instant power-on and zero standby power. Unfortunately, it suffers from high active leakage power and low reliability issues, which limit their application in FPGAs. The high active leakage power and low reliability are caused by the insufficient R_H. The low reliability is caused by the low retention of RRAM/PCM cells with a bias voltage of VDD during operation.

   One of the important concerns to integrate the NVM in FPGAs is its retention. The NVM may lose its advantage over other volatile memories if the states can only be retained a few seconds. For example, retention failure of PCM occurs when the phase-change material in the amorphous state is crystallized into the poly-crystalline state. The crystallization process can be accelerated by chip temperature and/or reading bias voltage [132], also

named as thermal disturbance and read disturbance, respectively. The bias voltage on PCM cells will heat up phase change material. The crystallization speed of PCM is dependent on the temperature and increases when the temperature is higher. The elevated temperature due to the bias voltage will result in fast crystallization and hence poor retention. This is also one of the reasons to hold the read voltage much lower than *SET* voltage. Since the read voltage exponentially reduces the retention time [132], it is better to bias PCM cells at 0V during FPGA operation which could greatly improve their retention performance. The read disturbance not only exists in PCM, but is also one of the major issues in RRAM [133] and STT-MRAM [64], since the read operation shares the same current path as the write operation.

3. **2T1R**. The '2T1R' (or '1R') scheme as shown in Fig. 1.7(c) was suggested in [129, 134–136] to replace the NMOS switch and SRAM cell to achieve high speed and density. Although it addresses some of the issues in SRAM solution, it faces problems such as significant low write reliability and high write power due to the high leakage current in the sneak paths. For example, to program RRAM cell R_{NW} between nodes N and W in Fig. 1.8(a), the potential on N is at V_{set} or V_{reset} (where V_{set} and V_{reset} are the RRAM set and reset switching voltages, respectively) and the potential on node W is the ground. However, if R_{NW}, R_{SN} and R_{SW} are at high, low and low resistance states, respectively, the majority current goes through R_{SN} and R_{SW}, resulting extremely large leakage current since the resistance of RRAM cells in HRS and LRS has two to six orders difference. Therefore, the current on R_{NW} may be insufficient to switch the selected cell. The write disturbance may worsen the write reliability. As shown in Fig. 1.8(b), if R_{NW}, R_{SN} and R_{SW} are at high, low and high states, respectively, the potential on R_{NW} and R_{SW} is almost the same. As a result, both R_{NW} and

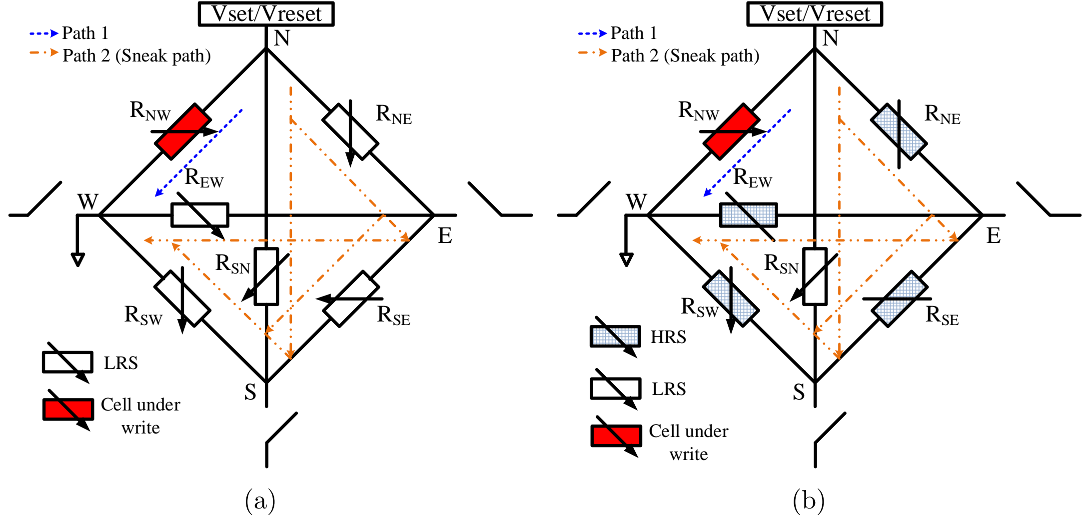


Figure 1.8: (a) The high leakage current issue, and (b) the write disturbance issue in the conventional RRAM based non-volatile SP. The en-dash lines are the paths to program the RRAM cells, and dash-dot-dot lines are the sneak paths.

R_{SW} may be switched.

Though biasing the unselected device at half ($V/2$ scheme) or one-third ($V/3$ scheme) of the programming voltage may reduce the write disturbance, the leakage current may still severely affect the configuration data integrity [137]. As the equivalent circuit illustrated in Fig. 1.9 when unselected RRAM cells are at LRS, the sneak path can be regarded as equivalent resistors paralleled to the cell under programming. For example, if the $V/2$ scheme is used, the paralleled resistance between the write voltage V_w and the ground is about $2(R_L + R_{p0})/(M - 1)$. As a result, the majority of the current goes to the sneak paths, and the parasitic resistance R_{p0} may dominate the total equivalent resistance between V_w and the ground. Increasing V_w to compensate the drop of the write voltage will make the RRAM suffer from high breakdown risk because the voltage on R_{cell} may be excessively high if most of the unselected cells are at HRS. Moreover, the unselected cells may still suffer from high write disturbance, because they are biased at the

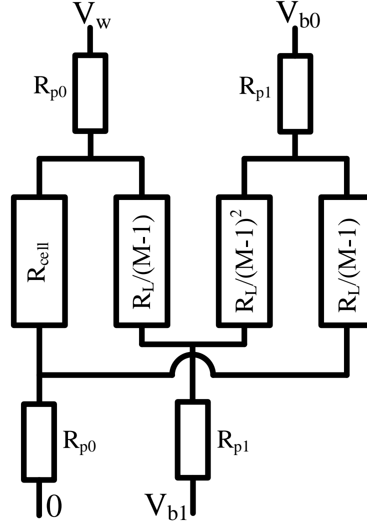


Figure 1.9: Equivalent circuit of a diode-less crossbar array. R_{cell} is the RRAM cell resistance under programming, R_L is the resistance of RRAM cells in LRS, M is the dimension size of the array, R_{p0} is the input parasitic resistance from the switch, metal, etc., R_{p1} is the parallelled input parasitic resistance, which is $R_{p0}/(M-1)$ for $V/2$ or $V/3$ write scheme and infinite for floating scheme, V_w, V_{b0} and V_{b1} are the writing voltage, and biasing voltages for the unselected word lines and bit lines, respectively.

half of the write voltage. The 1D1R or 1T1R structure may help to reduce the sneak path leakage current. However, the diode and transistor cannot be embedded in the FPGA routing path. Otherwise, they will increase the voltage drop and delay. Applying the non-linearity to the RRAM cell or embedded a non-linear selector in series may help to reduce the sneak patch current and voltage drop. However, the potential on the "ON" RRAM cell has to be zero during FPGA operation. Therefore, the "ON" resistance could be significantly large due to the non-linearity, which conflicts the low "ON" resistance requirement to reduce the RC delay of the interconnect in FPGAs.

1.5 My Contributions

In this dissertation, we propose four schemes to address various limitations in the conventional nvFF and nvFPGA designs. The detailed of each contribution is listed in the following.

(1) We propose a new nvFF with two-phase write approach instead of parallel/serial write approach to achieve lower VDD, lower saving/restoring power, and higher FF operation speed. We also analysis the impact of the MTJ parameters on the performance of the nvFF.

(2) A localized dedicated NVM array with '2-σ' and quad-phase pipelined write approaches is proposed to further reduce the saving power and improve the density as well, which may open a new direction of the zero standby leakage power dissipation design. In addition, a new reference resistance generator circuit is proposed to achieve low power and high sense margin.

(3) The '2D1R' storage element is proposed, which works as diode-less crossbar interconnect during operation, and '2D1R' memory array during configuration. The new FPGA architecture based on the proposed storage element is also proposed. Compared to the conventional nvFPGA designs, the proposed scheme significantly improves the write reliability and reduce the write power, while compared to the SRAM-based FPGAs, it achieves much higher density and performance.

(4) The PCM-based nvSRAMs are proposed for single-context and multi-context FPGAs. It greatly simplify the process, and significantly improves the read reliability with much lower active leakage power by biasing the NVM cells at 0V during the FPGA operation.

Contribution (1) has been published by IEEE Transactions on Nanotechnology, the localized NVM array design in Contribution (2) has been submitted to IEEE Transactions on Circuits and Systems: Regular I, and the reference re-

sistance generator in Contribution (2) has been accepted by IEEE Transactions on VLSI Systems, Contribution (3) has been submitted to IEEE Transactions on VLSI Systems as well, and Contribution (4) has been accepted by IEEE Transactions on Circuits and Systems: Regular I. The list of the publications is provided in the following.

Publications

1. Kejie Huang, Ning Ning, Yong Lian. *Optimization Scheme to Minimize Reference Resistance Distribution of Spin-transfer-torque MRAM.* IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol.PP, no.99, pp.1,1, 0. doi: 10.1109/TVLSI.2013.2260365

2. Kejie Huang, Yajun Ha, Zhao Rong, Akash Kumar, Yong Lian. *A Low Active Leakage and High Reliability Phase Change Memory (PCM) based Non-volatile FPGA Storage Element.* IEEE Transaction on Circuits and Systems I: Regular Paper. (Accepted)

3. Kejie Huang, Rong Zhao, Ning Ning, Yong Lian. *A Low Power Localized 2T1R STT-MRAM Array with Pipelined Quad Phase Saving Scheme for Zero Sleep Power Systems.* IEEE Transaction on Circuits and Systems I: Regular Paper. (Minor Revision)

4. Kejie Huang, Yong Lian. *A Low Power Low VDD Non-volatile Flip-Flop using STT-MRAM.* IEEE Transactions on Nanotechnology, vol.12, no.6, pp.1094,1103, Nov. 2013. doi: 10.1109/TNANO.2013.2280338

5. Kejie Huang, Rong Zhao, Wei He, Yong Lian. *High Density and High Reliability Non-volatile Field Programmable Gate Array (FPGA) with Staked 1D2R RRAM Array.* IEEE Transactions on Very Large Scale Integration (VLSI) Systems. (Submitted)

6. Kejie Huang, Rong Zhao, Yong Lian. *Racetrack Memory based Non-volatile Storage Element for Multi-context FPGAs.* IEEE Transactions on Computers. (Submitted)

1.6 Thesis Organization

Chapter 1 is the introduction to this thesis. It provides the motivation to integrate NVMs in CMOS circuits, the background of NVMs and the related works. The organization of the thesis is also provided. Chapter 2 presents the circuit design and simulation of the proposed nvLatch for zero standby power systems. The impact of the parameters are also discussed. Chapter 3 describe an alternative solution - a dedicated NVM array to retain the states of the registers during standby. The detailed analysis and impact of the parameters are also provided. Chapter 4 shows a non-volatile FPGA switch to overcome the low write reliability of the conventional design. Both SPICE and VPR simulation results are provides. Chapter 5 provides a new nvSRAM based solution for the single-context and multi-context FPGAs. Its detailed simulation results are also provided. Finally, the conclusions are drawn in Chapter 6.

Chapter 2

Non-volatile Latch/FF for Zero Standby Power Systems

This chapter is written mainly based on the paper "A Low Power Low VDD Non-volatile Flip-flop Using STT-MRAM".

2.1 Introduction

The NVM is an effective solution to retain the states of the registers thus the whole system can be fully powered off during sleep mode. In particular, it is only necessary for all the FFs to be non-volatile if the function blocks are clock-synchronized. Employing nvFF can provide a more efficient use of energy in SOCs for standby-power-critical and quick-startup applications, especially the battery powered appliances. The ground bounce or VDD fluctuation issues will not affect the retention states that saved to NVM cells. The nvFFs could be designed as the standard cells to maintain the compatibility with digital design flows in order to reduce the design cycle.

The main operational principle of nvFFs based approach is to store the

states of FFs into NVMs during standby, and restore them back to FFs when the system is powered on. Many MRAM based nvFF works have been reported [4–9]. However, the existing works face several issues in various aspects, *i.e.*, updating MTJs every clock cycle, programming two MTJs in series, source degeneration, etc. These issues significantly affect the performance of nvFFs and the integration of MTJs in the deep sub micron CMOS processes.

In this chapter, we propose a novel nvLatch using STT-MRAM technology. The proposed nvLatch can be used as a stage of master latch or slave latch to implement the nvFF circuit. Low VDD and low power are achieved by using two-phase write approach instead of the serial or parallel write approaches, and the complementary PMOS and NMOS pair in the write path rather than one select transistor only. The low VDD also helps to reduce the CMOS feature size and thus increase the latch operation speed. The VDD and saving energy could be further reduced by decreasing MTJ cell size, resistance-area product (RA), MTJ critical current. The proposed nvFF achieves $4.78pJ$ saving energy and the size is only 1.77 times of the conventional CMOS retention FF. The latch and the read/write circuit are connected by two sense NMOS transistors and one inverter only, thus the parasitic loading of the latch is greatly reduced. The setup time, propagation delay time T_{HL} and T_{LH} are $37ps$, $45ps$ and $48ps$, respectively.

2.2 Proposed nvLatch/nvFF

The write circuitry for the nvFF should be carefully designed to reduce the energy of the saving operation while keeping high write reliability. Since longer pulse width results in higher switching possibility [58, 91], the write pulse width should be long enough to achieve sufficient low write bit error rate (BER) at low VDD, especially when there is no error correction code (ECC) modules. Moreover, the asymmetry of MTJ switching at two switching directions [64, 138] results in longer

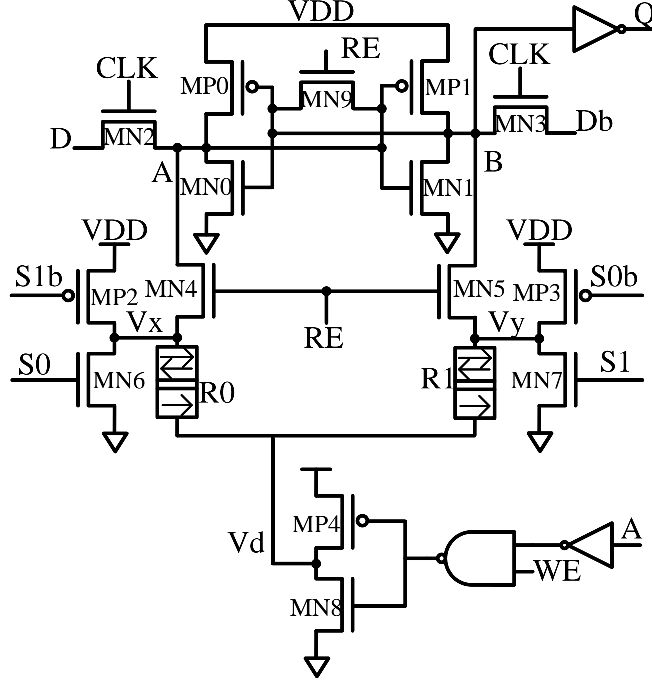


Figure 2.1: Proposed STT-MRAM based non-volatile latch with two-phase write approach.

$P \rightarrow AP$ pulse than $AP \rightarrow P$ pulse. The conventional nvFFs in which the serial or parallel write approaches are used, the write pulse width should follow $P \rightarrow AP$ pulse. This asymmetrical switching is mainly due to the different spin-transfer efficiency at the both sides of the oxide barrier. The MTJ switching threshold current density ratio of $AP \rightarrow P$ to $P \rightarrow AP$ can be calculated as [138]

$$\gamma = \frac{J_{c0}^{AP \rightarrow P}}{J_{c0}^{P \rightarrow AP}} \tag{2.1}$$

where $J_{c0}^{P \rightarrow AP}$ and $J_{c0}^{AP \rightarrow P}$ denote the MTJ switching threshold current density for the $P \rightarrow AP$ and $AP \rightarrow P$ operations, respectively.

To reduce the power consumption and increase the operation speed, the s-tates of MTJs are only updated before sleep mode and restored back to latches/FFs after the system is powered on. Hence during normal operation, the read/write

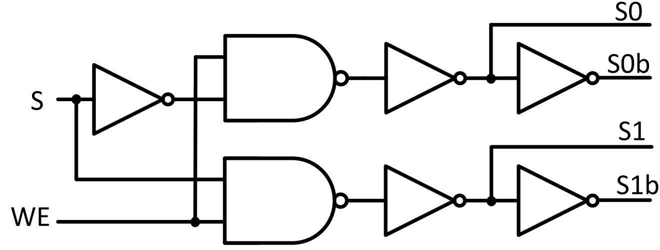


Figure 2.2: Two-phase write operation control logic to generate $S0$, $S0b$, $S1$ and $S1b$.

circuitries are turned off. To maintain data integrity, we prefer complementary MTJ structure, since the TMR of MTJs is only as low as 100% [64, 139]. To further reduce VDD, one NMOS and one PMOS are used as the select transistors rather than one NMOS transistor only, thus the "source degeneration" effect is eliminated. The latch is used as sense amplifier only and MTJs are moved outside the latch to reduce the parasitic RC of the latch caused by the read/write circuit. Another advantage to move the MTJs outside the latch is that the state of the latch can be used to program the MTJs directly. Otherwise, the input data has to be used to program the MTJs, thus the retained state may not be correct if the input data is changing during programming.

Our proposed nvLatch is shown in Fig. 2.1, which includes complete read, write and normal operation functions. V_x and V_y are the TE of R_0 and R_1, respectively. V_d is the BE of R_0 and R_1 connected together. The four global control signals $S0$, $S1$, $S0b$ and $S1b$ are generated by the write enable signal WE and the two-phase write control signal S as shown in Fig. 2.2. The parasitic loading of the latch is minimized, since the connection between the latch and the read/write circuitry is only two small sensing transistors and one inverter.

The block diagram of the proposed nvLatch/nvFF at the system level is shown in Fig. 2.3(a). The power management block determines when to power on or off the system. The data is saved from latches to MTJs and restored from

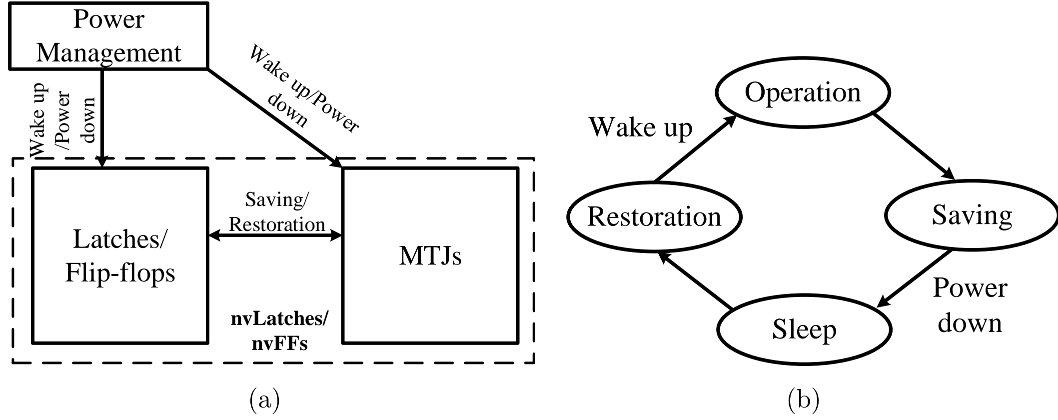


Figure 2.3: (a) Block diagram of the system level controller to save the states of the proposed nvLatches/nvFFs in the MTJs; (b) The four operation modes of the proposed nvLatches/nvFFs.

MTJs to latches when received "power down" and "wake up" instructions from the power management block, respectively. As shown in Fig. 2.3(b), the proposed nvLatches/nvFFs have four modes controlled by the power management block: the operation mode, the sleep mode, the saving mode and the restoration mode. If "power down" instruction is sent by the power management block in the operation mode, the system goes into the saving mode before "Powered down". If "wake up" instruction is sent by the power management block in the sleep mode, the system enters into the restoration mode before "Waken up". In the sleep mode, every blocks are powered off.

2.2.1 The State Saving Mode

The state saving mode is to write the state of the latch into the two complementary MTJs. In this mode, WE is high. Meanwhile, CLK is suggested to be low to isolate the latch from the input data, and RE is also low to avoid the writing operation to disturb the state of the latch. The writing operation controlled by the global signals $S0$, $S1$, $S0b$ and $S1b$ has two phases: the first $P \rightarrow AP$

switching and followed by a second $AP \rightarrow P$ switching. The control signal S is at low and high states in the first and second phases, respectively. Therefore, in the first write phase, $S0$=1, $S0b$=0, $S1$=0 and $S1b$=1 and in the second write phase, $S0$=0, $S0b$=1, $S1$=1 and $S1b$=0. For example, to write data 0 (A=0) to MTJs, the node V_d in Fig. 2.1 is pulled to VDD. Since S is in the low state initially, the states of the four global control signals are $S0$=1, $S0b$=0, $S1$=0 and $S1b$=1, thus V_x=0 and V_y=1. Therefore, only R_0 is under $P \rightarrow AP$ switching since the potential across R_0 and R_1 are $-VDD$ and 0, respectively. Once $P \rightarrow AP$ switching of R_0 is finished, the control signal S is raised to high. In this phase, V_x=1 and V_y=0 because $S0$=0, $S0b$=1, $S1$=1 and $S1b$=0. Therefore, only $R1$ is programmed to P state since the potential on R_0 and R_1 are 0 and $-VDD$, respectively. Similarly, to write data 1 to MTJs, R_1 and R_0 are programmed to AP and P states sequentially.

Writing MTJs in two phases could lead to 50% reduction of write driver size ($MP4$ and $MN8$) as compared to the parallel write approach and 30% reduction of VDD as compared to the serial write approach. This approach enables $P \rightarrow AP$ and $AP \rightarrow P$ pulses to be separately controlled, which is not possible in either the parallel write approach or the serial write approach due to the simultaneous MTJs programming nature.

In our proposed design, only the node V_d is determined by the latch state. The nodes V_x and V_y are controlled by $S0$, $S0b$, $S1$ and $S1b$ globally to save the area. Once the write operation is finished, the latch may be powered off and all signals are disabled.

2.2.2 The State Restoration Mode

In the state restoration mode, the data stored in the MTJs is read back to the latch. In this mode, WE is low to pull V_d to ground and disconnect V_x and V_y

from VDD or ground by turning off $MP2 - MP3$ and $MN6 - MN7$. Once V_d is pulled to ground, the MTJs could be sensed by RE, thus the control signals of the read operation are simplified. Meanwhile, CLK is still low to isolate the latch from input data. To sense the data from MTJs, RE is set to high first to equalize the voltage on nodes A and B, and set the sensing voltages on V_x and V_y. The NMOS transistor pair $MN4$ and $MN5$ is used to reduce the sensing voltage on V_x and V_y by V_{th}. Therefore, the voltages on nodes V_x and V_y are clamped to $VDD - V_{th}$, and the initial sensing current on two MTJs are $(VDD - V_{th})/R_P$ and $(VDD - V_{th})/R_{AP}$, respectively. Hence, R_P side has faster discharge speed than R_{AP} side. When it is stable, there is a voltage difference between nodes A and B. For example, if R_0=R_P and R_1=R_{AP}, node A will be discharged much lower than node B. Once read operation is finished, RE is set back to low to disconnect the MTJs from the latch and amplify the voltage difference on nodes A and B by the latch. Minimizing the pulse width of RE could reduce the static current flows through the MTJs.

2.2.3 The Normal Latch Mode

In the normal latch mode, both WE and RE are low to disable the write and read operations, respectively. The read/write circuitry is disconnected from the latch by turning off the two NMOS sense transistors $MN4$ and $MN5$. Thus, the parasitic loading from the read/write circuitry is small. The design works as a conventional $6T$ SRAM - data is written into the latch through $MN2$ and $MN3$, and stored at the outputs of the two inverters.

2.2.4 Non-volatile Flip-flop

The proposed nvLatch can be used as either a master latch or a slave latch in an nvFF circuit. If the nvLatch is used as a master latch, the output port Q is

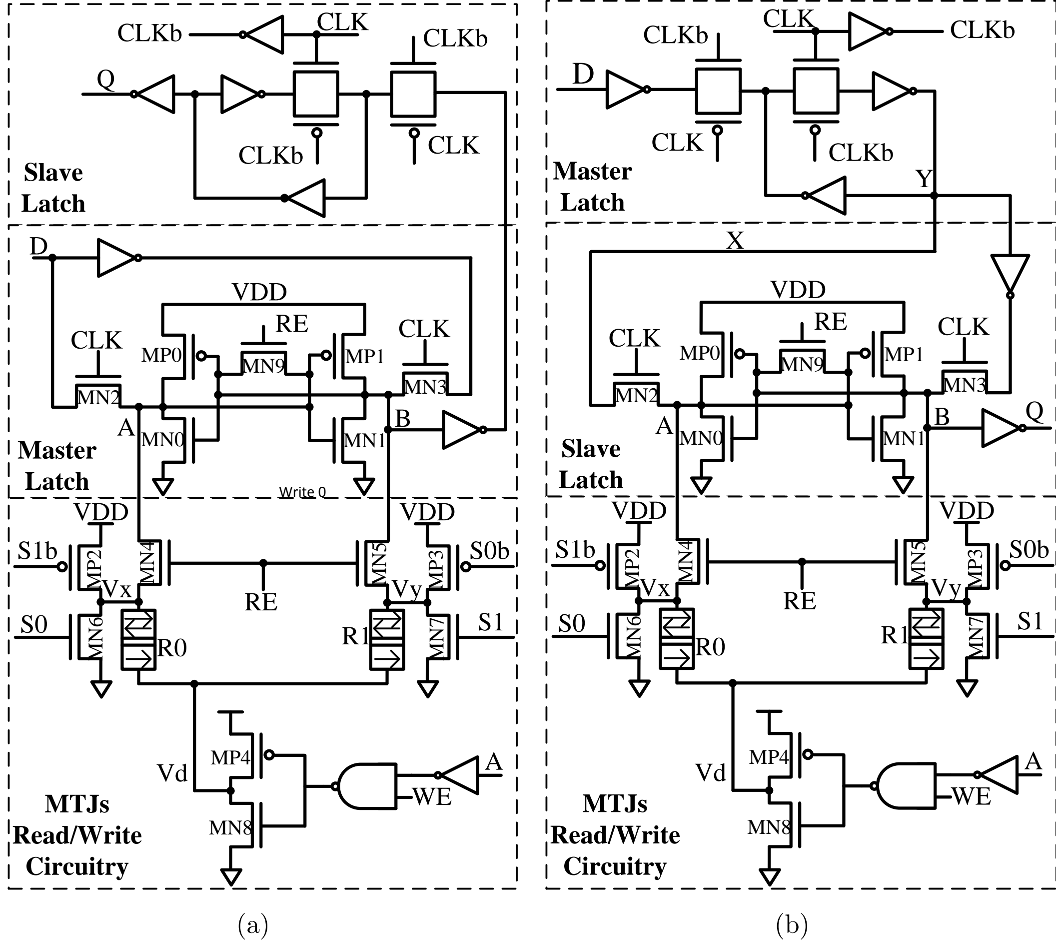


Figure 2.4: Proposed STT-MRAM based nvFFs. (a) The nvLatch is used as a master latch in the nvFF; (b) The nvLatch is used as a slave latch in the nvFF.

connected to the input of the slave latch. If the nvLatch is used as a slave latch, the input ports D and Db are connected to the output of another master latch. Figs. 2.4(a) and 2.4(b) show the configurations where the nvLatch is used as the master latch and slave latch in the nvFF, respectively. The saving and restoration operations are the same as the nvLatch discussed above.

Table 2.1: Description of the $90nm$ embedded MTJs and $45nm$ CMOS process.

Device parameters	value
CMOS	Cadence 45nm generic PDK
VDD	1V
MTJ Size	$90nm \times 90nm \pm 5\%$
TMR	100%
Resistance-area (RA) product	$25.4\Omega{\cdot}\mu m^2$
Thermal Stability (Δ)	65
P to AP intrinsic switching current density ($J_{C0}^{P\rightarrow AP}$)	$2.38MA/cm^2 \pm 5\%$
AP to P intrinsic switching current density ($J_{C0}^{AP\rightarrow P}$)	$1.47MA/cm^2 \pm 5\%$

Table 2.2: The write energy comparison among different write approaches.

Write Approaches	**Write Energy**
Parallel	$(I_{AP\rightarrow P} * VDD * T_{AP\rightarrow P} + I_{P\rightarrow AP} * VDD * T_{P\rightarrow AP} + I_p * VDD * \lvert T_{AP\rightarrow P} - T_{P\rightarrow AP}\rvert$
Serial	$I_{s1} * VDD * T_{s1} + I_{s2} * VDD * T_{s2}$
Two-phase	$I_{AP\rightarrow P} * VDD * T_{AP\rightarrow P} + I_{P\rightarrow AP} * VDD * T_{P\rightarrow AP}$

2.3 Simulation Results

In this section, we firstly evaluate the impact of VDD on the performance of the nvFFs. After that, we evaluate the performance of our proposed nvFF compared to the other reported nvFFs. Finally, we further evaluate impact of the MTJ parameters on the three different write approaches. Table 2.1 tabulates the default design parameters used in the simulation.

The MTJ model in [88, 89] is used in this chapter for the simulation. The detailed description of the model has been provided in Section 1.2.1. In all of the simulations below, the write circuits have been optimized to minimize the write energy for each approaches to achieve close speed performance. For example, in the parallel write approach, if $AP \rightarrow P$ switching occurs before $P \rightarrow AP$ switching, $I_{AP\rightarrow P}$ could be reduced to make both switching equal. Otherwise, the

lower resistance after switching increases the total write energy. In the serial write approach, if $AP \rightarrow P$ switching occurs before $P \rightarrow AP$ switching, the current goes through two MTJs is increased, so that the VDD could be reduced to achieve similar $P \rightarrow AP$ switching speed as the parallel and two-phase write approaches. On the other hand, if $AP \rightarrow P$ switching occurs after the $P \rightarrow AP$, the VDD should be high enough to make both switching succeed. The details of the three write approaches are summarized in Table 2.2, where I_p is the excessive current when one MTJ cell is switched faster than the other one in the parallel write approach; I_{s1} and I_{s2} are the first and second cells switching current in the serial write approach; and T_{s1} and T_{s2} are the first and second cells switch speed in the serial write approach. Therefore, if the two MTJ cells are switched simultaneously, the $I_p * VDD * |T_{AP \rightarrow P} - T_{P \rightarrow AP}|$ part in the parallel write energy equation in Table 2.2 could be eliminated.

2.3.1 Analysis the impact of VDD

We firstly evaluate the impact of the supply voltage on the nvFF saving speed performance. All parameters are set to default value in Table 2.1 except γ, where $\gamma = \frac{J_{C0}^{AP \rightarrow P}}{J_{C0}^{P \rightarrow AP}}$. The γ is set to 0.5 and 1 in this simulation. It can be observed from Fig. 2.5 that the two-phase write approach is much faster than the serial write approach, but slightly slower than the parallel write approach. To achieve $1V$ or lower VDD when γ=0.5, the parallel and two-phase write approaches could finish the saving operation in less than $30ns$. However, to achieve similar speed performance, the VDD of the serial write approach has to be higher than $1.6V$.

Fig. 2.6 shows the required energy to store the nvFF state into the MTJs among three write approaches. To achieve the same saving speed, *i.e.*, $30ns$, the two-phase write approach requires much lower energy than the other two, no matter γ is 0.5 or 1. Increasing the saving speed may require higher VDD. On the

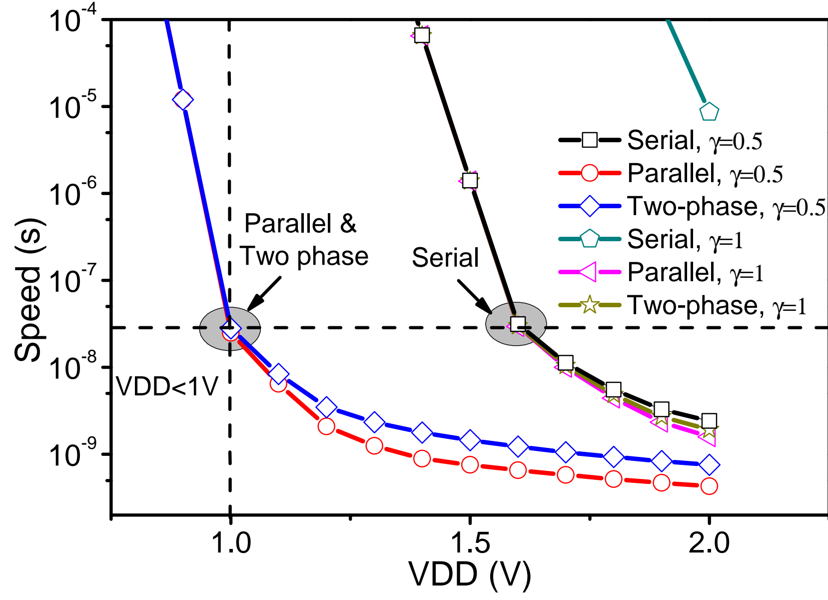


Figure 2.5: The supply voltage vs. the nvFF saving speed among three write approaches.

other hand, reducing the saving speed increases the saving energy.

2.3.2 The performance of the proposed nvFF

As discussed in Section 2.3.1, the VDD of the nvFF is set to $1V$, and the saving speed is set to around $30ns$. Fig. 2.7 shows the simulation results of the proposed nvFF in Fig. 2.4(a). The results show the example of one write operation, two read operations, and two normal FF operations. Initially, R_0 is at AP state, and R_1 is at P state. The output Q of the nvFF is updated to 1 and 0 by the first read and normal FF operations, respectively. The followed write operation synchronizes the states of R_0 and R_1 to R_P and R_{AP}, respectively. Though the second FF operation updates Q to 1 again, the second read operation synchronizes states of the two MTJs and Q, ignoring the input data D. The clock "CLK" should always be 0 to avoid any disturbance from the input data during saving and restoration operations.

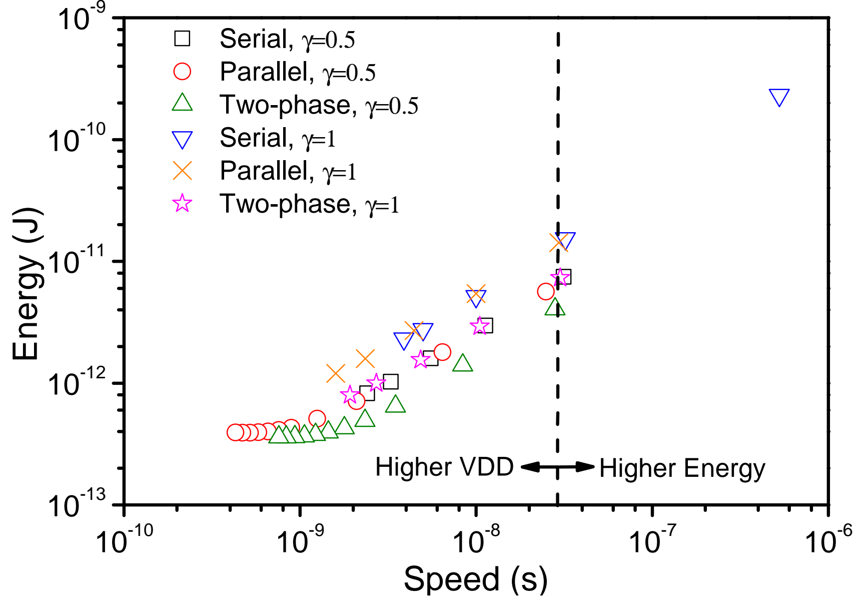


Figure 2.6: The nvFF saving speed vs. saving energy among three write approaches.

Table 2.3: The performance of our proposed nvFF.

Device parameters	value
$T_{P \to AP}$	24.9ns
$T_{AP \to P}$	10.8ns
Restoration speed	>0.1ns
Restoration energy	>0.22fJ
Saving speed	35.7ns
Saving energy	4.78pJ

Table 2.3 summarizes the performance of our proposed nvFF in Fig. 2.4(a). The nvFF provides $91\mu A$ $AP \to P$ current ($I_{AP \to P}$) and $151\mu A$ $P \to AP$ current ($I_{P \to AP}$) in the two write phases, respectively, which finish the two write phases in $25ns$ and $12.5ns$, respectively. By finely controlling the pulse width, the write energy could be greatly reduced. The simulation results show that the states of the latch are restored $100ps$ after RE is enabled, and the restoration power is $2.2\mu W$. The restoration energy of our proposed nvFF is escalated with the restoration pulse RE. Therefore, the restoration pulse RE should be minimized to reduce the

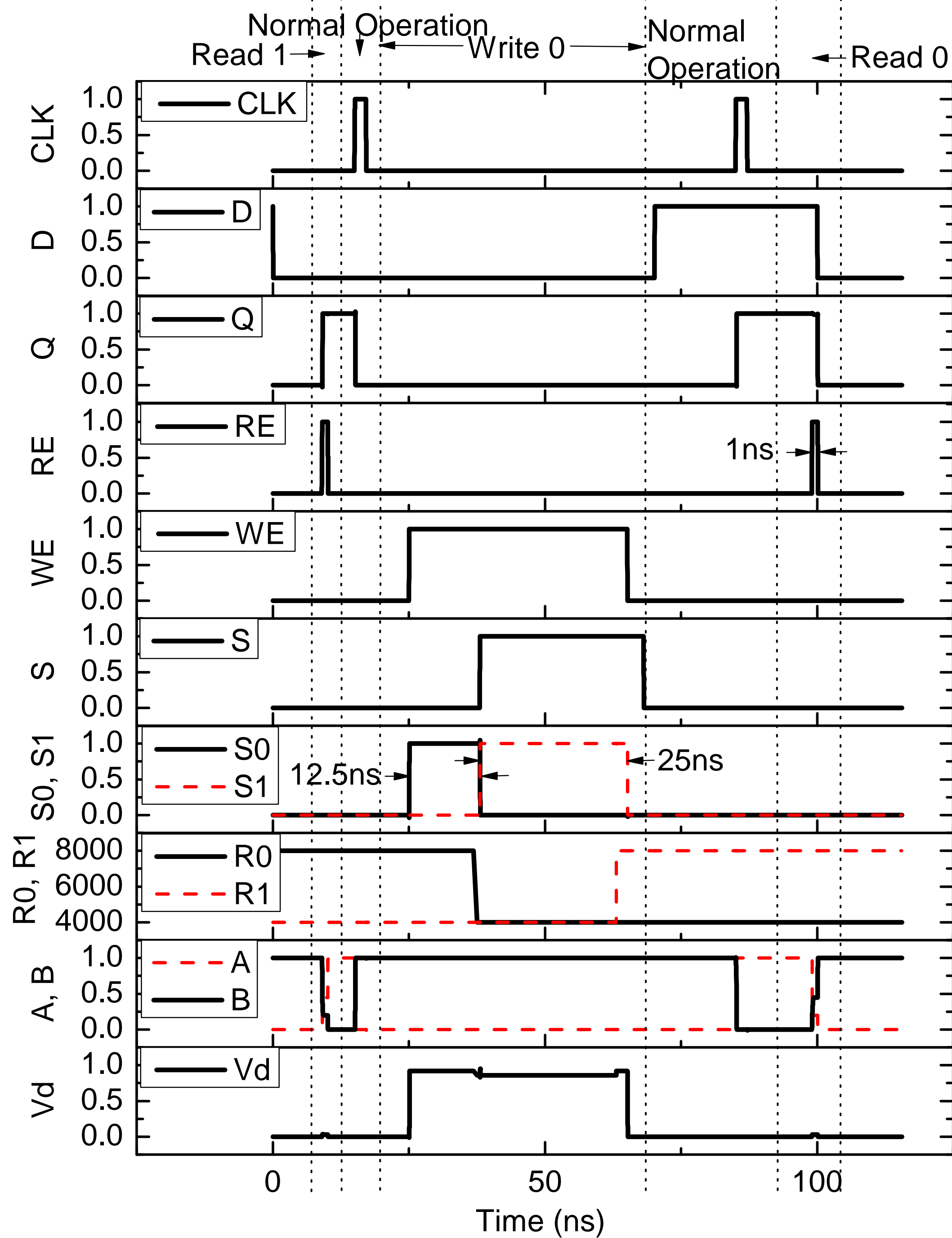


Figure 2.7: The simulation results of the proposed nvFF. It has two read operations (restoration), one write operation (saving) and two normal FF operations.

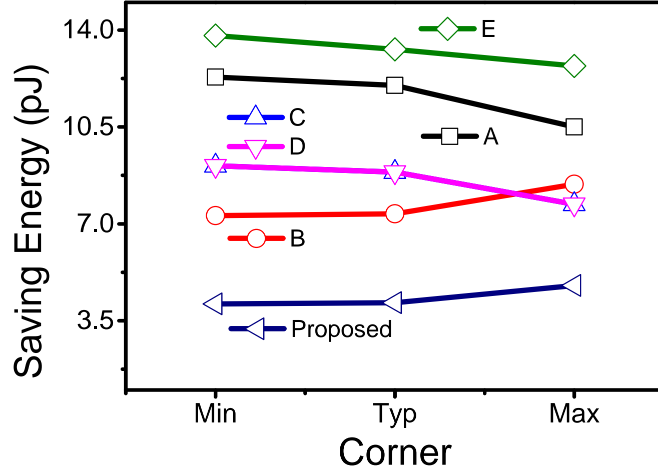


Figure 2.8: The corner simulation results among the proposed nvFF and the conventional nvFFs. Min corner: MTJ size -5%, Jc0 -5%, transistor width +5%; Max corner: MTJ size -5%, Jc0 -5%, transistor width +5%. A: [4]; B: [5]; C: [6]; D: [7]; E: [8].

restoration energy.

The nvFF is designed at its worst corner to ensure the states could be successfully saved to MTJs in all corners. The worst corner here is defined as smallest write current, *i.e.*, highest MTJ resistance and smallest transistor width, and highest J_{c0}. In this simulation, only MTJ size, transistor width and J_{c0} are considered, and all these variables are set to ±5% variation from its typical value. The other corners have higher writing current than its worst corner. Therefore, the write reliability can be guaranteed. Compared to the conventional nvFFs, our proposed nvFF could save more than 38% power in all corners as shown in Fig. 2.8. In this simulation, the same switching periods are set for all corners.

Table 2.4 provides the comparison among different nvFFs and the CMOS retention FF. The saving power of [9] is estimated based on $200MHz$, $2.5V$ and $1mA$ write energy, allowing the cell to be successfully programmed. Other MTJs are using the same MTJ model as the proposed one. Compared to the reported

Table 2.4: The performance comparison among the proposed nvFF, conventional nvFFs and the CMOS retention FF during saving operation.

Structures	**Required VDD**	**Saving Energy**	**Saving Speed**	t_{BEP}
Porposed	1V	4.78pJ	35.7ns	0.956ms
CMOS FF	1V	3fJ	0.1ns	0.6us
[4]	2.4V	10.5pJ	32.7ns	2.1ms
[9]	2.5V	12.5pJ	5ns	2.5ms
[5]	1.7V	8.43pJ	36ns	1.69ms
[6]	1.6V	7.71pJ	30ns	1.54ms
[7]	1.6V	7.71pJ	30ns	1.54ms
[8]	1.8V	12.7pJ	25ns	2.54ms

nvFFs, our proposed nvFF has the smallest saving energy, which is only $4.78pJ$. The restoration speed and energy are ignored in the comparison since they are much smaller than the saving speed and energy of the nvFFs. The required VDD of our proposed nvFF is only $1V$ and the energy of the saving operation has been reduced by more than 30% compared to the other nvFF structures. The saving time is slightly longer since it has to sequentially program the two MTJs. However, the BEP [140] is a more important value than the saving speed, which represents the time when the nvFFs have the sleep energy reduction to store the states into the MTJs. We define t_{BEP} as

$$t_{BEP} = \frac{E_{retain} + E_{restore}}{P_{FF}} \tag{2.2}$$

where P_{FF} is the leakage power of the flip-flop; E_{retain} and $E_{restore}$ are the energy of the saving and restoration operations, respectively. The leakage power of the proposed nvFF without leakage power reduction techniques is $5nW$ at room temperature based on the simulation result, hence t_{BEP} of our proposed nvFF is around $1ms$. Therefore, the saving and restoration time as shown in Table 2.3 is much smaller than t_{BEP}. The smaller t_{BEP} allows the system to be powered on/off more frequently. Reducing t_{BEP} relies on the energy reduction of the saving

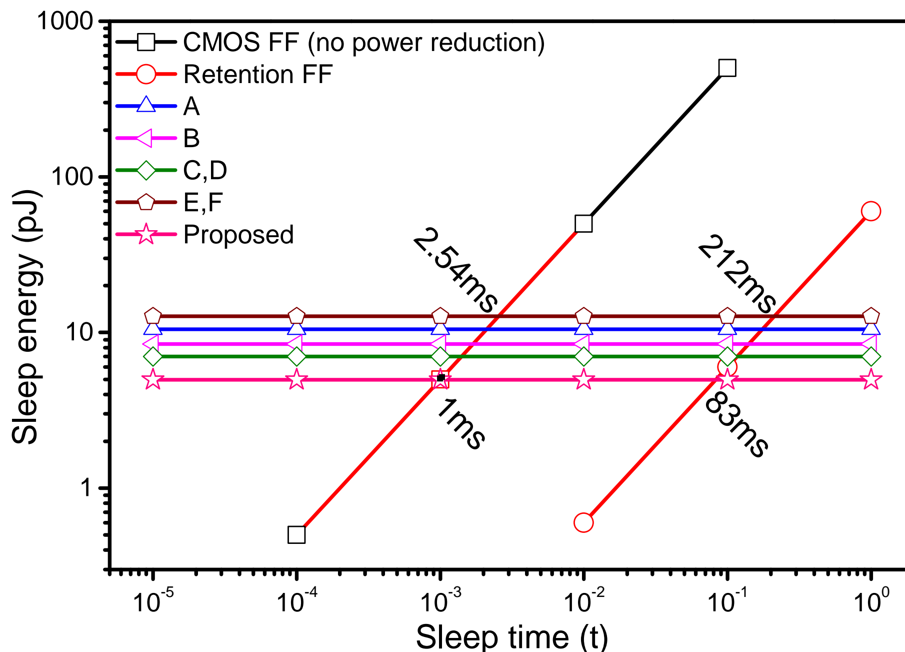


Figure 2.9: Sleep energy comparison among different nvFFs and conventional CMOS FFs. A: [4]; B: [5]; C: [6]; D: [7]; E: [8]; F: [9].

and restoration operations, especially the saving operation, which is determined by the STT-MRAM technology. For example, reducing the write pulse width or current.

Fig. 2.9 shows the sleep energy comparison among different states retention technologies. The sleep of the conventional CMOS retention FF is proportional to the time. Even with the power reduction technique, the total sleep energy will exceed the nvFF technologies after a long standby time. The leakage power of the CMOS retention FF with sleep transistor off is $60pW$ at room temperature from the simulation. Thus as shown in Fig. 2.9, when the sleep time is longer than $80ms$, our nvFF has the advantage of the energy reduction compared to the CMOS retention FF. When the sleep time is $1s$, the energy reduction is around 92%. In the system, the sleep energy reduction is much larger since most of the FFs and all of the combinational do not need to retain their states [37, 141]. This principle

Table 2.5: The performance comparison among the proposed nvFF, conventional nvFFs and the CMOS retention FF during normal operation.

Structures	**Propagation delay**(L→H /H→L)	**Setup time**	**FF state update energy**
Porposed	45ps/48ps	37ps	5fJ
CMOS FF	33ps/32ps	47ps	2.4fJ
[4]	63ps/68ps	30ns	10.5pJ
[9]	57ps/84ps	79ps	100fJ
[5]	63ps/94ps	67ps	46fJ
[6]	77ps/72ps	77ps	20fJ
[7]	81ps/95ps	74ps	24fJ
[8]	0ps/447ps	25ns	12.7pJ

also applies to t_{BEP}. For example, if the leakage power of the retention registers only occupy 10% of total system standby power, then t_{BEP} of our proposed nvFF is only $95.6\mu s$.

The performance comparison among the proposed nvFF, conventional n-vFFs and the CMOS retention FF during the normal operation in listed in Table 2.5. The setup time of [4] and [8] are the minimum time period to successfully program the MTJ cells, and the propagation delay is the sense time of the MTJs. As shown in Table 2.5, the setup time, rising and falling propagation delays (CLK-to-Q) of our proposed nvFF are $37ps$, $45ps$ and $48ps$, respectively, which is much better than the other nvFFs. The energy to update the state of our nvFF is only $5fJ$, reducing more than 70% from the conventional nvFFs. The higher energy during normal FF operation compared to the conventional CMOS retention FF is due to the SRAM style latch is used in our nvFF. The small propagation delay and state updating energy are achieved by the low VDD and small parasitic loading on the latch.

The 1.77 normalized area is also much smaller than the reported nvFFs as shown in Table 2.6. The normalized area is estimated by

Table 2.6: The estimated area comparison among the proposed nvFF, conventional nvFFs and the CMOS retention FF during normal operation.

Structures	**Total transistors**	**Write transistors**	**Estimated FF Size**
Porposed	37	6	1.77
CMOS FF	31	0	1
[4]	44	4	10.4
[9]	36	4	9.68
[5]	29	9	5.22
[6]	38	4	4.13
[7]	41	4	4.38
[8]	42	12	6.9

$$AREA = \frac{(M - N + \alpha * N) * VDD^2}{T} \tag{2.3}$$

where M and N are the number of the total transistors in the nvFF and the transistors in the write path, respectively; T is the number of the transistors of the CMOS retention FF, and α is the magnified ratio of the transistor size in the write path, which is around 4 from the simulation. It is a conservative estimation since the scaling speed of the transistor feature size is much faster than that of VDD [17, 142].

2.3.3 Analysis the impact of MTJ parameters

We further evaluate the impact of MTJ parameters on the minimum VDD requirement and saving energy of the nvFFs.

The VDD requirement of the three write approaches are evaluated with different MTJ parameters as shown in Fig. 2.10. Three common features can be summarized from Fig. 2.10: (1) under the same conditions, the parallel/two-phase write approaches reduce more than 30% VDD requirement compared to the serial write approach; (2) reducing VDD requires smaller $J_{c0}^{P \to AP}$, MTJ size, TMR, RA, γ and $Delta$; (3) high speed requires high VDD, and the gap between

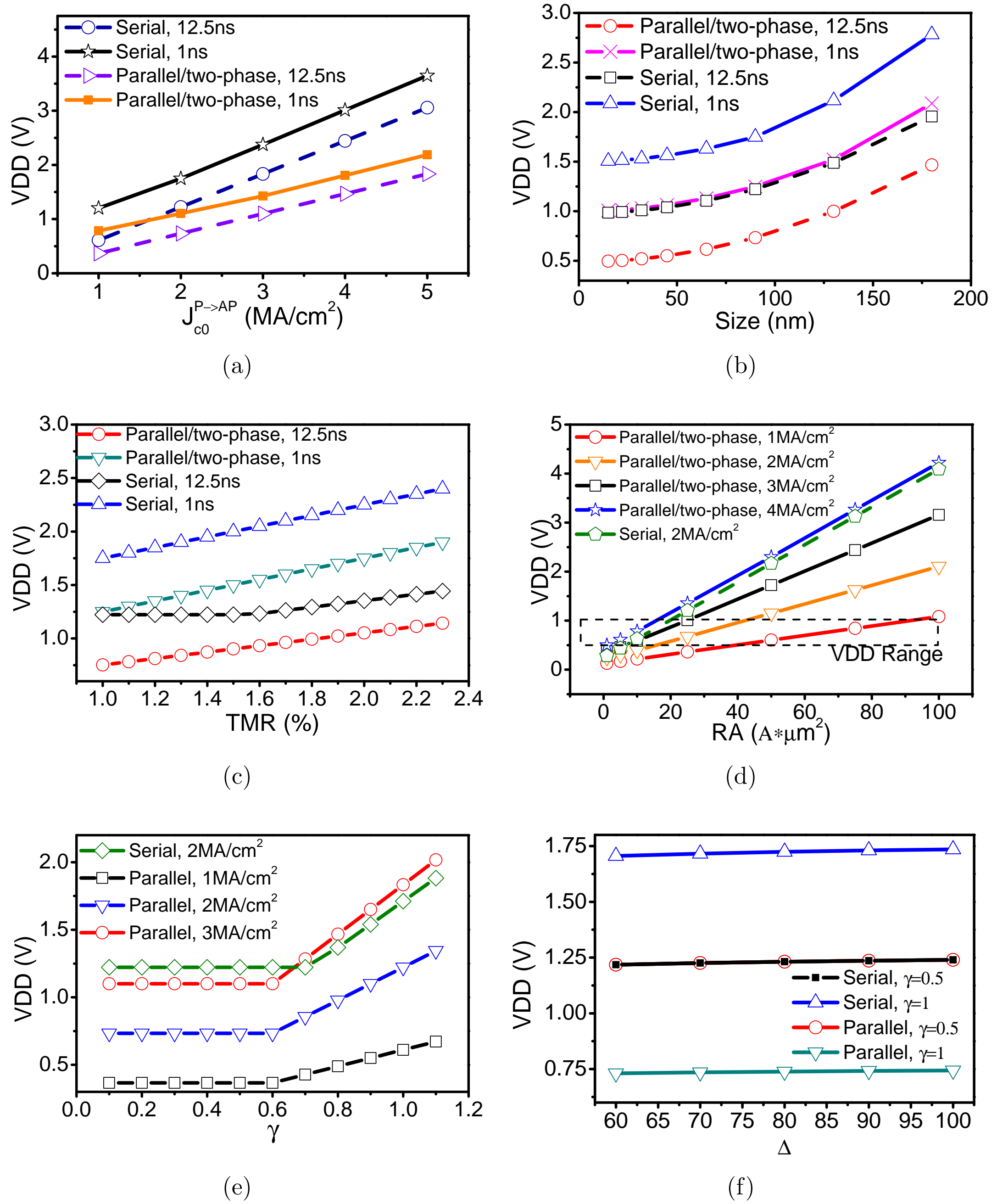


Figure 2.10: The supply voltage requirement of the three write approaches vs. (a) $J_{c0}^{P\to AP}$, (b) size of the MTJ cells, (c) TMR, (d) RA, (e) γ, and (f) thermal stability Δ.

different switching speeds is almost constant. Moreover, the VDD requirement is proportional to the $J_{c0}^{P\rightarrow AP}$, TMR (except low TMR of the serial write approach), RA, Δ and square of the MTJ size. The VDD is determined by R_P and R_{AP} at low TMR and high TMR, respectively. Figs. 2.10(a) and 2.10(d) illustrate the relationship among $J_{c0}^{P\rightarrow AP}$, RA and VDD of the serial and parallel/two-phase write approaches. The $J_{c0}^{P\rightarrow AP}$ and RA should be appropriately chosen in order to achieve targeted VDD of the system. Fig. 2.10(e) shows the required VDD of the serial and parallel/two-phase write approaches versus γ. With the same $J_{c0}^{P\rightarrow AP}$, the required VDD of the two-phase write approach is proportional to γ when γ is larger than 0.6. On the other hand, the required VDD is constant when γ is smaller than 0.6. This phenomenon is due to the required write voltages for $P \rightarrow AP$ and $AP \rightarrow P$ switching dominate the regions of γ>0.6 and γ<0.6, respectively. The serial write approach has a similar phenomenon, but much higher VDD at the same $J_{c0}^{P\rightarrow AP}$. As shown in Fig. 2.10(f), the effect of the Δ is much smaller than the other parameters. It also shows that the VDD requirement of the serial write approach when γ=0.5 almost overlaps the VDD requirement of the two-phase write approach when γ=1. This phenomenon also can be observed from Fig. 2.10(e) that the serial write approach when γ=0.5 and the parallel/two-phase write approaches when γ=1 require almost the same VDD level.

Fig. 2.11 shows the simulation results of the required nvFF saving energy for the three write approaches with different MTJ parameters. It can be observed from Fig. 2.11 that two-phase write approach requires the lowest energy in all conditions. Moreover, the fast precessional switching requires much less switching energy than the thermal activated switching. Figs. 2.11(a) and 2.11(b) show that the high $J_{c0}^{P\rightarrow AP}$ and MTJ cell size exponentially increase the saving energy. In other words, the low $J_{c0}^{P\rightarrow AP}$ and the MTJ cell size are important to achieve low nvFF saving energy. As shown in Fig. 2.11(c), the effect of TMR on the

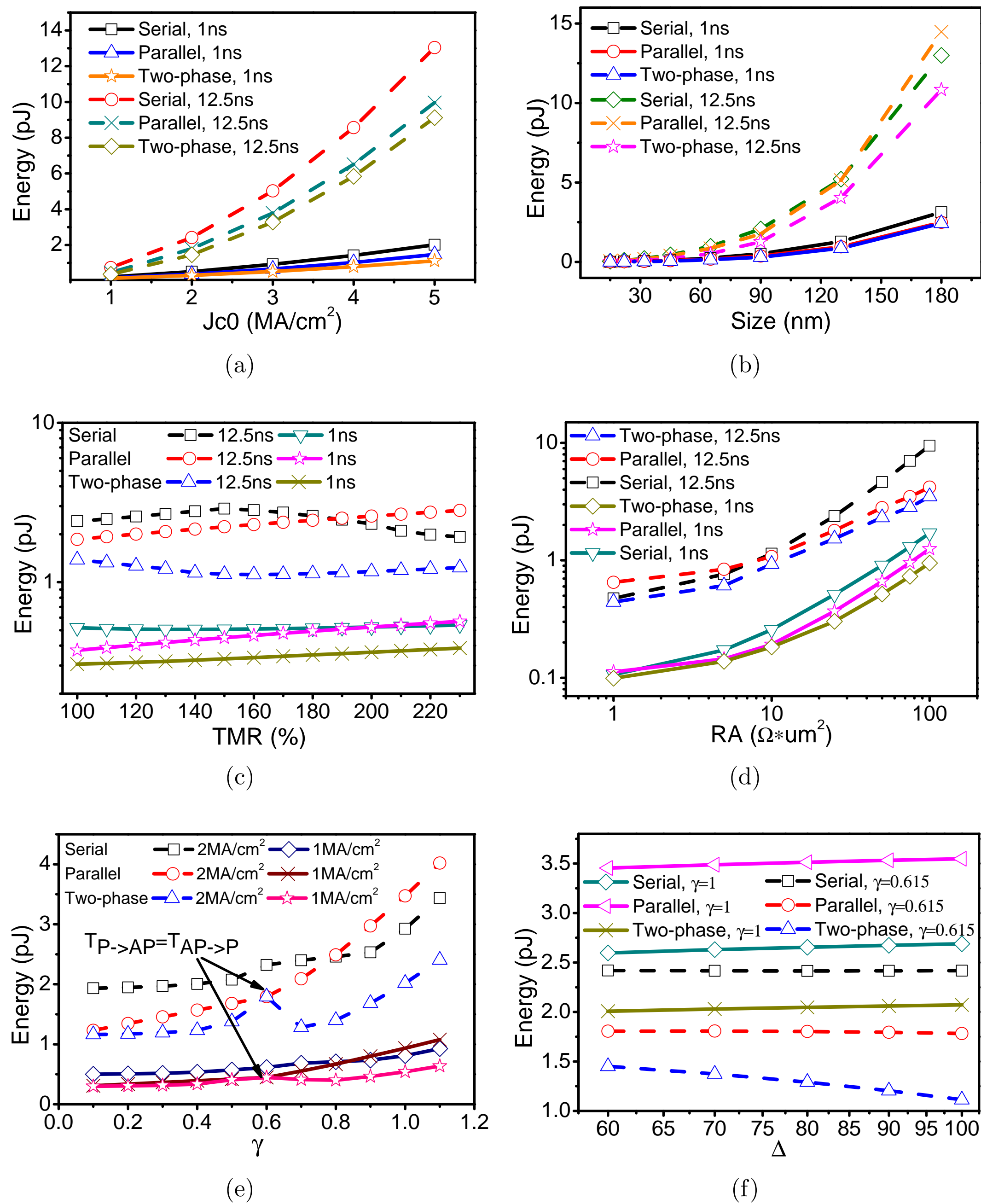


Figure 2.11: The required nvFF saving energy for the three write write approaches vs. (a) $J_{c0}^{P\to AP}$, (b) size of the MTJ cells, (c) TMR, (d) RA, (e) γ, and (f) thermal stability Δ.

nvFF saving energy is much smaller than $J_{c0}^{P\rightarrow AP}$ and the MTJ cell size. The nvFF saving energy of the two-phase write approach achieves its minimum at TMR=150% and increases when TMR>150%. In contrast, the saving energy of the serial write approach reaches its peak at TMR=150% and decreases when TMR>150%. The saving energy of the serial write approach will increase when dominated by $AP \rightarrow P$ switching. The energy of the parallel write approach gets higher than the serial write approach when TMR>190%. It can be seen from Fig. 2.11(d), RA and the write energy have an approximate positive linear function. The energy required by the parallel write approach may be higher than the serial write approach at low RA level is because the parasitic resistance is much higher than the MTJ resistance. Fig. 2.11(e) illustrates the write energy of the three write approaches with different γ and $J_{c0}^{P\rightarrow AP}$. The parallel and two-phase write approaches have the same energy when the switching pulses of $P \rightarrow AP$ and $AP \rightarrow P$ are the same. Except this point, the two-phase write approach has lower write energy than the parallel write approach, since the energy of the two-phase write approach is proportional to $T_{P\rightarrow AP} + T_{AP\rightarrow P}$ and the energy of the parallel write approach is determined by the $max(T_{P\rightarrow AP}, T_{AP\rightarrow P})$. The write energy of the serial write approach gets smaller than the parallel write approach when γ>0.8, which is because $I_{P\rightarrow AP}$ is close to $I_{AP\rightarrow P}$ when γ is close to 1. As can be seen from Fig. 2.11(f), the write energy affected by the Δ is much smaller than the other parameters. When γ=0.615, high Δ helps to reduce the saving energy of the two-phase write approach.

In summary, the lower VDD and saving energy could be achieved by reducing the cell size, RA, $J_{c0}^{P\rightarrow AP}$ or γ. Reducing γ may decrease the current sensing margin if voltage sense amplifier is used. If current sense amplifier is used and keep $(1+TMR)*\gamma$>1, the voltage sensing margin may not be affected. Reducing TMR or Δ to achieve low VDD conflicts the MTJ design targets, since high TMR and

Δ are required for high read reliability [143] and long-term data retention [144], respectively.

2.4 Summary

A low power low VDD nvLatch has been proposed based on STT-MRAM technology to achieve zero sleep power consumption. The low VDD, which is able to scale down to $1V$ and below, is achieved by two-phase write approach and complementary write drivers. The two-phase write and low VDD greatly reduce the saving power to only $4.78pJ$, which has more than 38% reduction compared to the conventional nvFF topologies, and allows the system to be powered off when the sleep time is longer than $1ms$. The area of the proposed nvFF is only 1.77 times of the conventional retention CMOS FF, which is only half of the smallest nvFF size among the reported works. The VDD and saving energy could be further reduced by decreasing the MTJ cell size, RA, $J_{c0}^{P \rightarrow AP}$ or γ.

Chapter 3

Localized Array for Zero Sleep Power Systems

This chapter is written mainly based on the papers "A Low Power Localized 2T1R STT-MRAM Array with Pipelined Quad Phase Saving Scheme for Zero Sleep Power Systems" and "Optimization Scheme to Minimize Reference Resistance Distribution of Spin-transfer-torque MRAM".

3.1 Introduction

The use of nvFFs to retain the states of the register during power-off was proposed in [6, 9, 11, 145] to eliminate the standby power in Fig. 3.1(a), thus achieve zero power dissipation, as shown in Fig. 3.1(b). However, they have high peak power when saving states before powering off. Moreover, they may face issues of reliability and significant extra area, since PCM and RRAM may require high program voltage [114–117], and STT-MRAM may suffer from high read error rate due to its low TMR ratio [64, 139]. A possible solution to address the die area and reliability issues is deploying non-volatile computer data storage to retain the information

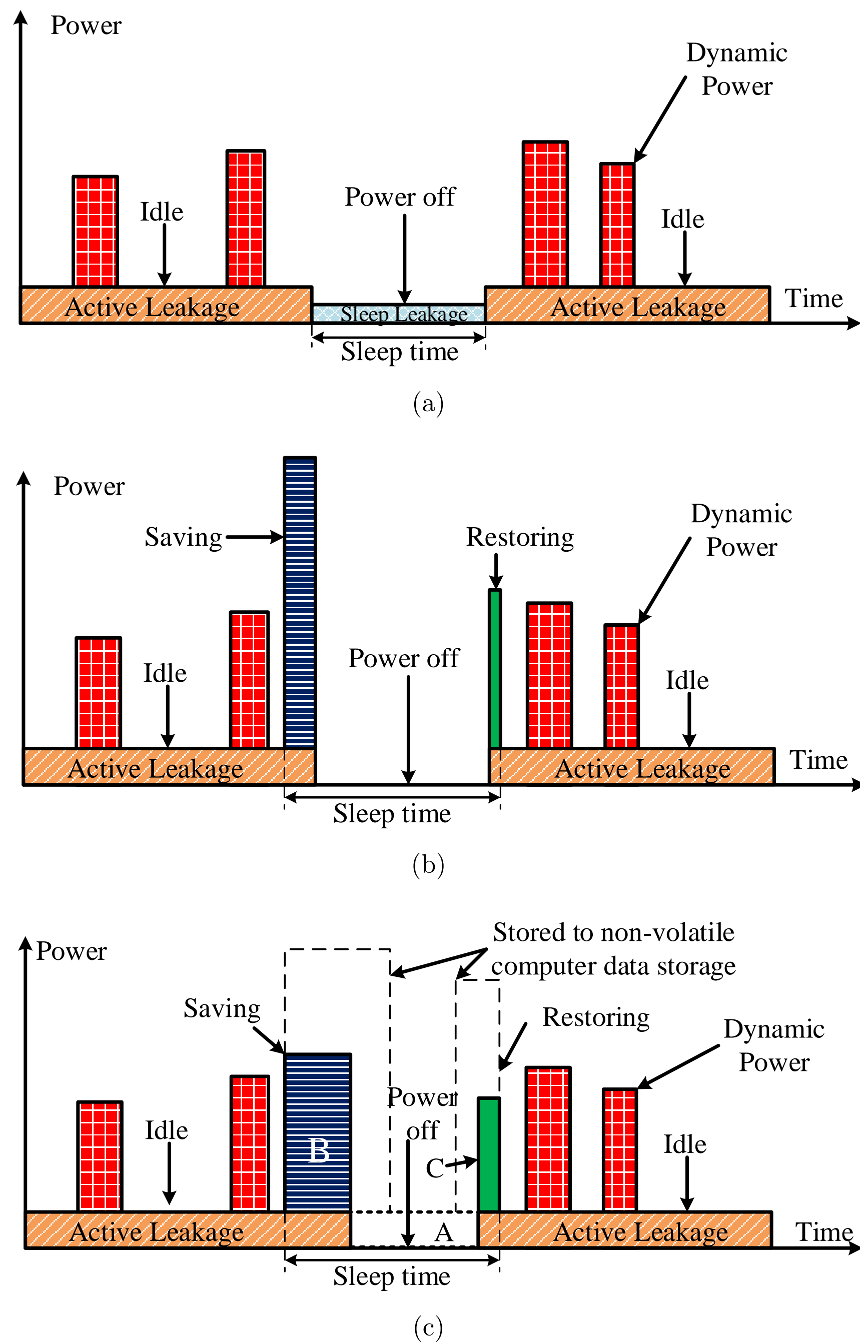


Figure 3.1: Power consumption of (a) CMOS retention registers based approaches, (b) nvFF based approaches, and (c) proposed dedicated NVM array based approach.

of the registers during sleep. However, it requires a processor to support the complicated algorithm for the bus arbitration process, as transferring the information of registers may share the system/data bus and compete the priority with other processes. Moreover, the power to shift data in a long scan chain may dominate the total sleep power. Therefore, the sleep cost may limit the sleep possibility.

This chapter proposes a new direction of the zero standby leakage power dissipation design by storing the states of the registers in a localized NVM array through scan chains. As shown in Fig. 3.1(c), a dedicated local memory block to store the states of the registers may significantly reduce the time and energy for the data transfer than the computer data storage, allowing the system to be powered on/off more frequently. It also converts the high peak power of the nvFF approach to the low power level with longer saving time. The read-before-write and 2σ saving approaches significantly reduce the power consumption of the saving operation. The simulation results show that the whole system only consumes the saving and restoring power, which are less than $1.1pJ$ per bit in total. The BEP, which is defined by the time when the reduced sleep energy equals to the energy required to save and restore the system (the area of A in Fig. 3.1(c) equals to the sum of the area B and C), can be used to evaluate power-off possibilities. Our result shows that the break even point is $22\mu s$ when the leakage power of retention registers is 10% of the total leakage power. In other words, it could boost power consumption reduction when sleep time is longer than $22\mu s$.

3.2 Proposed Scheme

Conventional nvFFs are designed to fully replace the information stored in the NVM cells when powering off. In conventional nvFF based schemes, NVM cells are randomly distributed in the whole very large scale integrated (VLSI) system as shown in Fig. 3.2(a). We propose a localized dedicated NVM array instead

of nvFFs to only store the states of registers during sleep, as illustrated in Fig. 3.2(b). Hence, more techniques (i.e., read-before-write, verify-after-write, ECC, etc.) can be applied to the write operation to improve the reliability and reduce the power consumption. Moreover, write drivers, sense amplifiers and other control blocks could be shared among different NVM cells, which greatly reduces the area overhead. The interface routings between memory array and digital block could be placed above memory array to reduce the area overhead. The estimated routing area overhead per one bit data is

$$A_{routing} = \frac{(W + D)(L_d - L_m)G}{2k} \tag{3.1}$$

where W is the width of the routing metal, D is the space between two routing metals, G is the total number of the registers required to retain their states, k is the number of scan chains, L_d and L_m are the lengths of the digital block and memory block, respectively. Therefore, small $L_d - L_m$ helps reduce the area overhead of the routing.

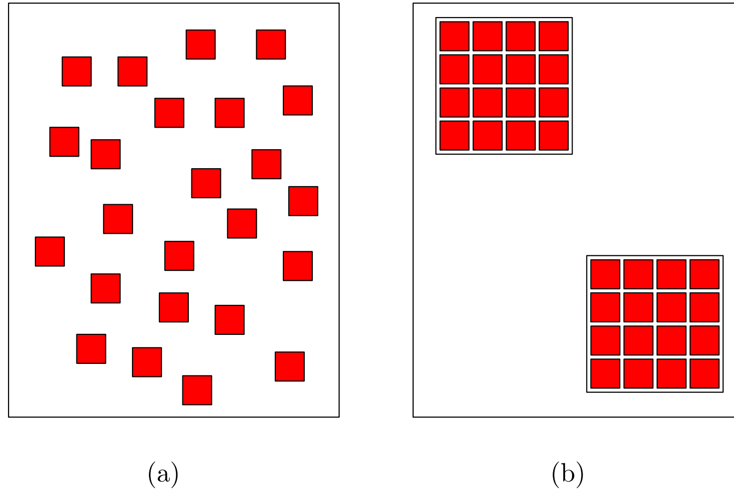


Figure 3.2: (a) MTJ cells are distributed randomly in conventional nvFF schemes; (b) localized NVM arrays in our proposed scheme.

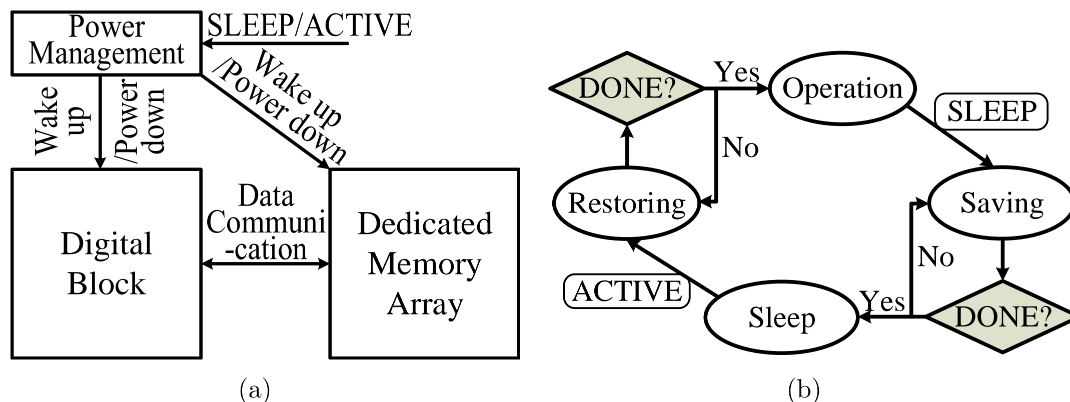


Figure 3.3: (a) Top diagram of the scan based approach to save the states of the registers in the local dedicated NVM array; (b) The four modes of our proposed low power system.

3.2.1 Circuit Architecture

The top level diagram of the proposed scheme is shown in Fig. 3.3(a). $ACTIVE$ and $SLEEP$ are two control signals that determine whether to power on or off the system, respectively. As shown in Fig. 3.3(b), the system has four modes: the restoring mode, the saving mode, the operation mode and the sleep mode. The restoring mode is triggered by asserting $ACTIVE$ signal. Both digital and memory blocks are powered on, and the states of registers stored in the local memory array are loaded to the digital block. The saving mode is triggered by asserting $SLEEP$ signal. The memory block is powered on, and the states of the registers are saved to the memory array.

The detailed system architecture shown in Fig. 3.4 is proposed to write states of the registers to the localized memory array through the scan chain. Since NVM array retains information during sleep, the system could be fully powered off to achieve zero sleep power consumption. Data are written to the dedicated memory array in parallel. k bits parallel bus writing scheme requires k scan chains in the digital block. Each scan chain may have equivalent length. Dummy flip-flops

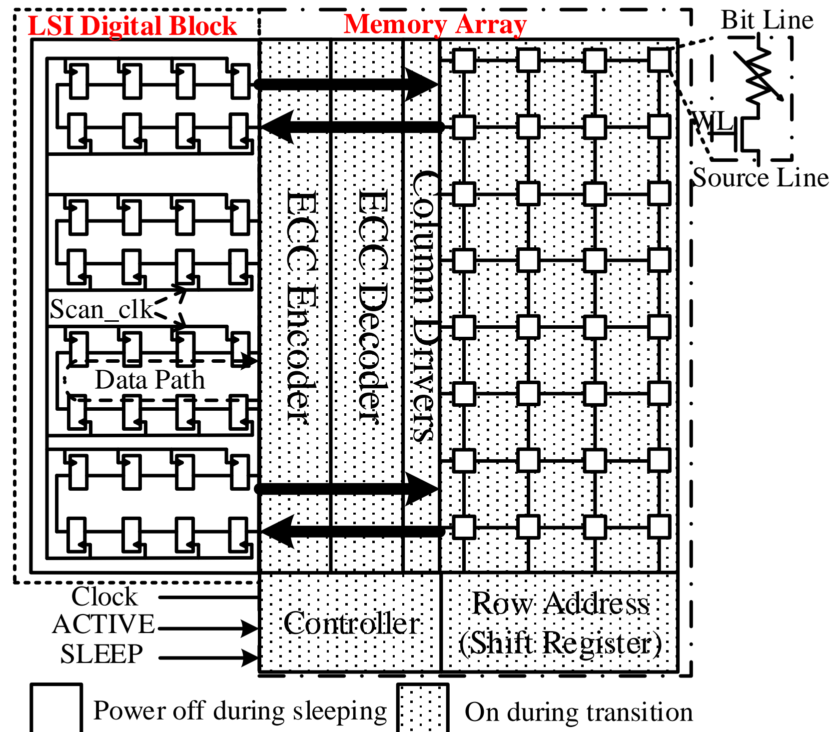


Figure 3.4: Proposed architecture with the localized non-volatile memory array. Left side of the diagram is the LSI block. Right side of the diagram is the NVM array with the memory controller.

may be inserted to equalize each scan chain. The NVM array with the controller is powered on only during the transition periods (saving mode and restoring mode). The sequence of the states to be written in the memory is following first-in-first-out (FIFO) rule. The scan chains are shared for both testing and save/restore purposes, hence no additional area is required in the digital block. The memory array and digital block are suggested to be placed in vicinity to reduce the parasitic capacitance in their interface.

3.2.2 Minimum Sleep Time

The system has to fully write all states to the memory before powering off or fully restore the states to the registers after powering on. Therefore, the system has the minimum sleep time requirement which should be longer than the total time of saving and restoring operations. The total time of saving and restoring operations is

$$t_{retain,total} = (t_{save} + t_{restore})G/k \tag{3.2}$$

where t_{save} and $t_{restore}$ are the equivalent single bit saving and restoring time, respectively. The restoring operation is reading the data from NVM array back to the registers through the same scan chains. The restoring speed is mainly determined by the sensing scheme, clock speed and the length of the scan chain.

The sleep energy cost includes saving energy and restoring energy which are E_{save} and $E_{restore}$, respectively. Therefore, to take the advantage of the sleep power reduction, the BEP time of the proposed scheme is defined as

$$t_{BEP} = \frac{E_{save} + E_{restore}}{\eta P_{FF,leakage}} \tag{3.3}$$

where $P_{FF,leakage}$ is the leakage power of a single scan register in the digital system, and η is the ratio between the power consumption of the selected registers and the total system power consumption of the system. The minimum sleep time requirement should meet the following condition

$$t_{sleep,min} = t_{retain,total} + t_{BEP} \tag{3.4}$$

Therefore, both saving/restoring time and BEP time are important to allow the system to be powered off frequently. The number of scan chains k can be adjusted to allow more registers in the digital system to be simultaneously saved to the memory array. Thus the time required by saving and restoring operations

can be less than t_{BEP}. Large k helps to reduce the saving and restoring time. Moreover, large k also helps to reduce the energy consumed by shifting the scan chain. The energy to shift a scan chain is

$$E_{scan} = \frac{G}{4k}(E_{FF,switch} + 3E_{FF,noswitch})$$

where $E_{FF,switch}$ and $E_{FF,noswitch}$ are the energy of a single scan FF with data switched and without data switched, respectively. The switch possibility of the scan FF is set to 50%. As can be seen from (3.5), the energy consumed by the scan chain to shift one bit is proportional to the length of the scan chain ($\frac{G}{k}$). Therefore, minimizing the length of the scan chain could help to reduce the saving/restoring energy and minimize sleep time. However, there is a tradeoff between the power consumption and area overhead of the localized memory array.

3.3 Localized STT-MRAM Array Design

Since the states of the digital block need to be written into the memory array before powering off and read back from memory array after powering on, small saving/restoring power and fast saving/restoring speed allow the system to be powered on/off frequently. Therefore, the design principles of the localized NVM array in such applications are low energy and high speed of saving and restoring operations. The design of NVM array is based on STT-MRAM, which can switch the phases between the anti-parallel (high resistance R_{AP}) and parallel (low resistance R_P) states. The STT-MRAM is one of the promising resistance-change NVMs, with the advantages of high speed, high density and low power.

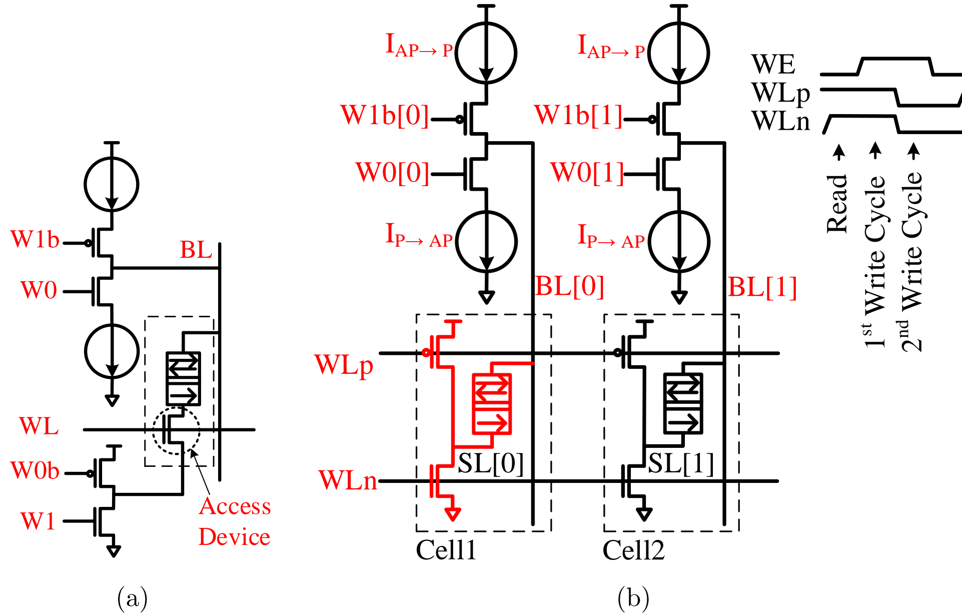


Figure 3.5: (a) The access device in conventional write schemes significantly limit the write current passing through the MTJ. (b) Proposed dual-step-write scheme to achieve low VDD.

3.3.1 Dual-Step-Write for Low VDD

The "source degeneration" issue caused by the access transistor in the conventional 1T1R scheme, as shown in Fig. 3.5(a), significantly limits the current that can pass through. The V_{gs} of the access transistor is reduced from VDD to $VDD - I_W \times R_{MTJ}$, where I_W and R_{MTJ} are the MTJ switching current and resistance, respectively. Therefore, it requires a much higher VDD to provide sufficient write current. From the simulation, the VDD of the 1T1R scheme has to be 60% higher than that of the scheme without access transistor. As a result, the scaling is limited and the power consumption is high.

We propose a complementary access transistor pair as shown in Fig. 3.5(b). The PMOS and NMOS are turned on when switching from AP state to P state and P state to AP state, respectively. Therefore, there is no V_{th} drop in the write paths, thus the "source degeneration" issue is addressed. Moreover, the stacked

transistor in the source line is also removed to help reduce VDD.

Furthermore, we propose a dual-step-write scheme to achieve parallel writing with minimum hardware overhead. For example, cell1 and cell2 in Fig. 3.5(b) are under P to AP switching and AP to P switching, respectively. Therefore, the current directions go through cell1 and cell2 are from SL to BL and from BL to SL, respectively. Hence, PMOS in cell1 and NMOS in cell2 are turned on. As a result, the single state WL is not possible to satisfy the requirement of our proposed scheme. We propose a dual-step-write scheme to allow the data to be written into memory cells in parallel. As shown in Fig. 3.5(b), the dual-step-write is achieved by shifting the address at the half of the WE pulse. NMOS is turned on in the first write step, and PMOS is turned on in the second write step, and vice versa. To switch cell1 from AP state to P state, both $W0b[0]$ and $W1[0]$ are low. When WLn is high, there is no current goes through cell1 since both BL and SL are at the ground. When WL is high, the write current $I_{AP \to P}$ is from SL to BL. It is similar to program cell2. There is a write current $I_{P \to AP}$ from BL to SL in the first write step, and no write current in the second write step.

3.3.2 Read-before-Write for Low Power

Read-before-write scheme (a read cycle is used to sense the data stored in the memory array before a write cycle) is used to reduce the write time and power consumption [146]. The time and power to write one bit data with the read-before-write scheme is

$$t_{rbw} = t_{read} + t_w * S \tag{3.5}$$

$$P_{rbw} = P_{read} + P_w * S \tag{3.6}$$

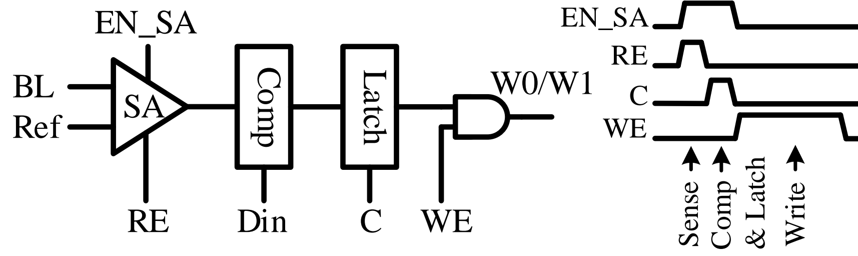


Figure 3.6: The sensing and comparing block diagram for the read-before-write scheme.

where t_r and t_w are the time to read and write one bit data, respectively. P_{read} and P_w are the power to read and write one bit data, respectively. S is the write possibility. It needs longer saving time, but reduces the saving power significantly. P_{read} may be ignored since it is much smaller than P_w. Theoretically, the saving energy could be reduced by around 50% if the probability (S) of the randomized data in the registers being different from those in the NVM array, is about 50%. In practice, more registers may have the same states between two adjacent sleep periods (sleep - power on - sleep), especially when the "on" period is short. Therefore, most of the memory cells only require read operations with dedicated NVM arrays, thus the retention power could be further reduced.

The sensing and comparing scheme is illustrated in Fig. 3.6. The sensing is carried out in the first half clock cycle controlled by the read enable signal RE. The sensed data are compared with the input data, their results are latched in the second half clock cycle controlled by four-phase clock C. The reference circuit used for sensing may use the scheme reported in [66] to reduce the resistance distribution with low sensing power. It will be discussed in Section 3.3.5.

The read-before-write has the advantage of lower saving power, but it also has the disadvantage of longer saving time. To address the disadvantage of the additional read time required by the read-before-write scheme, we propose a read-when-write scheme, which will be discussed in the following section.

Table 3.1: Example of pipelined quad-phase saving scheme. Row clock is used in the table.

Clock	c0	c1	c2	c3
0.5	read	0	0	0
1	w0	read	0	0
1.5	w1	w0	read	0
2	0	w1	w0	read
2.5	read	0	w1	w0

3.3.3 Pipelined Quad-Phase Write Scheme for High Speed

We further propose a pipelined quad-phase write scheme to maximize the write speed. The read-before-write and dual-step-write approaches require at least one cycle for reading and two cycles for writing, thus the time for changing a bit is increased by three times. As shown in Table 3.1, our proposed pipelined quad-phase write scheme has one channel in the read phase, two channels in the write phases (write 0 phase and write 1 phase), and one idle channel. The four channels pipelinedly shift their phases, and each channel has one phase delay. Each channel has k scan chains.

The advantages of our proposed pipelined quad-phase write scheme are: compared to the one channel writing scheme, it not only improves the speed by more than three times, but also reduces the scan chain length by four times ($\frac{G}{4k}$), thus less power will be consumed in the scan chains; compared to the four-channel parallel writing scheme, our proposed scheme reduces the peak power by around two times, and also reduces the hardware cost, i.e., ECC block, read/write control logic, which can be shared for all four channels.

The detailed control diagram of our proposed pipelined quad-phase write scheme is shown in Fig. 3.7. The four parallel channels from scan chains are converted to one series channel as the input of ECC and control blocks. Each channel has k bits data. One ECC block is used to code all four scan chains.

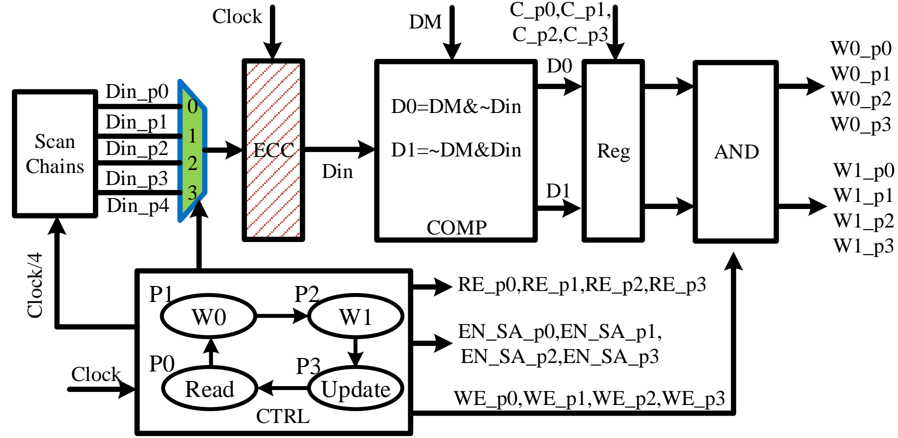


Figure 3.7: Proposed pipelined quad-phase control block diagram.

Scan chains are clocked by $Scan_clk$ and their shifting speed is reduced by four times. The ECC block and the comparison block operate four times faster than scan chains. The comparison block has the following functions to generate write 0 and write 1 pulses,

$$D0 = DM\&\overline{Din} \quad (3.7)$$

$$D1 = \overline{DM}\&Din \quad (3.8)$$

where DM is the data sensed from the dedicated memory array, Din is the encoded data from ECC, $\overline{DM}$ and $\overline{Din}$ are the inverse of DM and Din, respective. $D0$ and $D1$ are the write 0 and write 1 enable signals, respectively. The write enable signals $W0$ and $W1$ are latched by four-phase clocks $C_\{p0, p1, p2, p3\}$, which are generated from the four-phase control block (CTRL). The four-phase control block also generates four-phase read enable signals.

The array block diagram is shown in Fig. 3.8. There are 4 WLs and 1 BL pass through one MTJ cell. Since 1 PMOS and 1 NMOS are used as the access devices, the area is larger than $16F^2$ (i.e., 6T SRAM is $140F^2$ [1], thus the area of 1 PMOS and 1 NMOS is around $47F^2$. Moreover, each access transistor may be

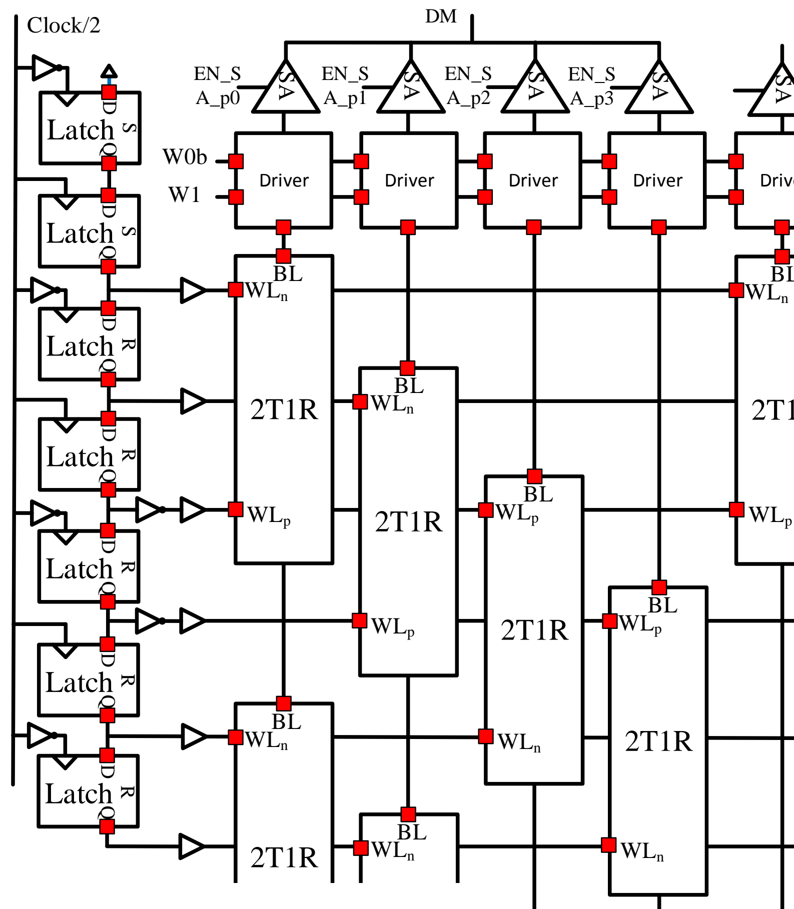


Figure 3.8: The array diagram of our proposed quad-phase writing approach.

larger than a minimum width transistor to pass through enough write current). In case the routing area is much larger than the access device (i.e., diodes are used), an alternative solution is that each channel has its dedicated row address. As a result, there are only 2 WLs and 1 BL pass through the access device.

As shown in Fig. 3.8, a shift latch scheme is used to generate the row address. The quad-phase write scheme also reduces the length of the row address by four times, thus reducing the power consumption of the row address by four times. The first two latches are set to 1 and all others are reset to 0 initially. A high output signal of the last shift register in the row address indicates the end

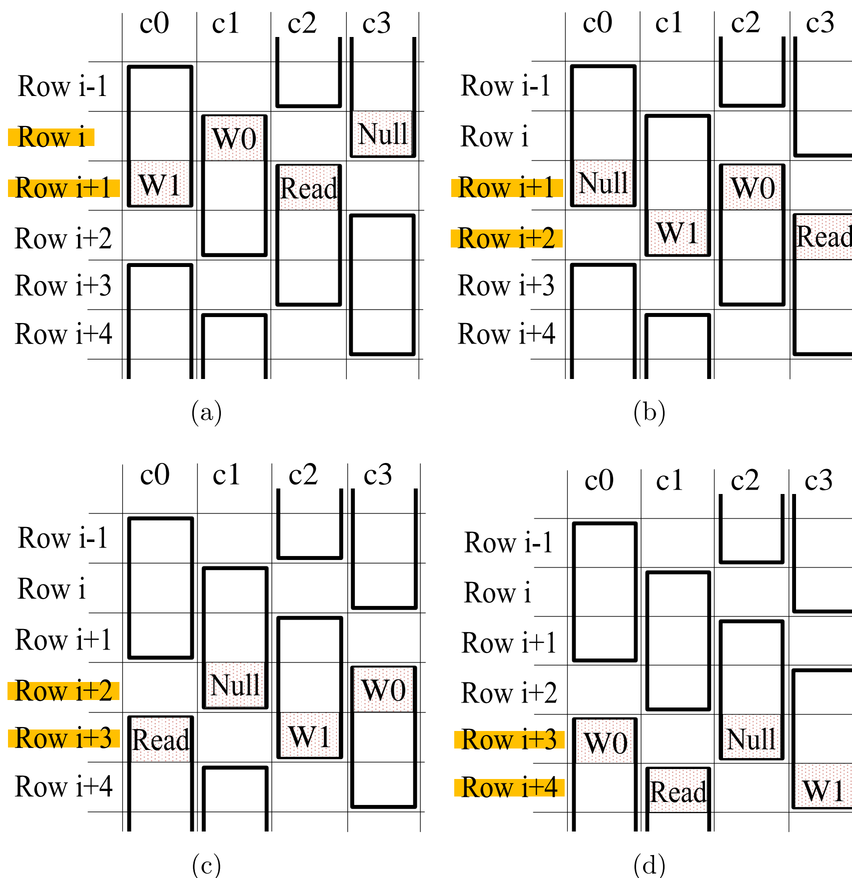


Figure 3.9: Block diagrams of our proposed pipelined scheme in the (a) i^{th}, (b) $(i+1)^{th}$, (c) $(i+2)^{th}$ and (d) $(i+3)^{th}$ system clocks. Each time two rows are active simultaneously. The active row addresses are highlighted in the figures.

of the saving or restoring operations. The clock to shift row address is divided by two from the system clock. Each time two adjacent row addresses are enabled simultaneously. As shown in Fig. 3.8, one address enables two channels, thus four channels are enabled simultaneously. For example, at 1.5 row clock cycle in Table 3.1, c0 and c2 are performing read and w1 operations, respectively, while c1 and c3 are performing w0 and idle operations, respectively.

Fig. 3.9 shows the example of the addressing of our proposed scheme. Each clock cycle (half row clock cycle) moves forward one bit address, and each row

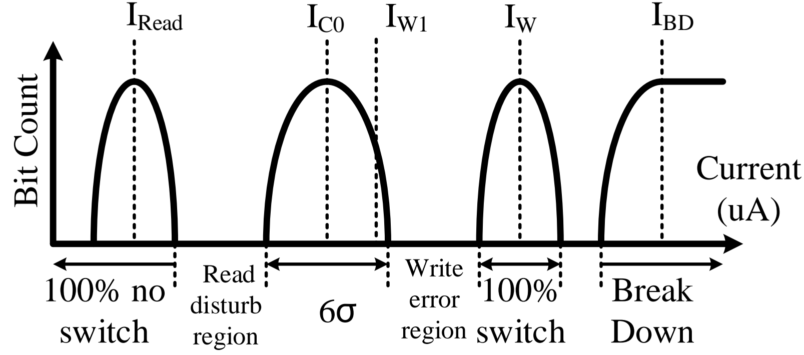


Figure 3.10: Distribution of characteristic currents in STT-MRAM array [10].

address is enabled in one whole row clock cycle. Therefore, two row addresses are enabled simultaneously. The quad phases are shifted every clock cycle. For example, the memory cells in address (Row i+1, c2) are under read, w0, w1 and Null phases at i^{th} to $(i+3)^{th}$ clock cycle, respectively.

3.3.4 2σ Write Scheme for Low Power

The non-uniformity of the material properties and the process imperfections, such as doping density variations and critical dimension variations, translate into cell-to-cell variation of the TMR, Δ, resistance, I_{c0}, and other cell parameters. Memory cell design should accommodate variations of both the MTJ and the accompanying circuit while maintaining the performance requirements. This implies additional constraints on the average MTJ parameters. Fig. 3.10 sketches distributions of the read and write currents in a typical STT-MRAM memory array. The write current should be high enough to achieve low write error rate.

We propose a modified verified-after-write scheme to achieve low saving power. It consists of two "read and write" operations. The first write operation uses reduced write current instead of 6σ write current. From the simulation, 2σ is the best choice. Here 2σ and 6σ mean 2σ and 6σ away from the mean of the

intrinsic switching current, respective. The detailed discussion will be provided later. After that, a read operation is performed to sense the state of the selected MTJ cell in order to determine if the preceding write is successful. The second write operation is only active when necessary. The write power of our proposed scheme is

$$P_w = P_{w1} * A + P_{read} + P_{w2} * (1 - A) \tag{3.9}$$

where $A = \frac{1}{2}[1 + erf(\frac{I_{w1}-I_{c0}}{\sqrt{2\sigma_{c0}}})]$ is switching possibility, I_{w1} is the 2σ write current, I_{c0} is the intrinsic switching current, and P_{read} is the reading power. P_{w1} and P_{w2} are 2σ and 6σ write power, respectively. Therefore, (3.9) can be rewritten as

$$\begin{aligned} P_w &= VDD * (I_{w1} * \frac{1}{2}[1 + erf(\frac{I_{w1} - I_{c0}}{\sqrt{2\sigma_{c0}}})] \\ &+ I_{read} + I_{w2} * (1 - \frac{1}{2}[1 + erf(\frac{I_{w1} - I_{c0}}{\sqrt{2\sigma_{c0}}})]) \end{aligned} \tag{3.10}$$

where I_{read} and I_{w2} are the reading current and 6σ write current, respectively.

It can be seen from Fig. 3.11(a), there is a minimum write power around 2σ away from the mean intrinsic switching current. As shown in Fig. 3.11(b), the power reduction gets higher when the standard deviation of I_{c0} (σ_{c0}) gets wider. The write power reduction is around 2.5% and 22.5% when σ_{c0} is 1% and 10%, respectively. Another benefit is that the switching current gets far away from the breakdown current, which may significantly reduce the breakdown risk, especially when the intrinsic switching current and write current are widely distributed. As shown in Fig. 3.12, 2.3% cells may be fail in the first write. But there is only 1.15% cells need a second write due to read-before-write.

Fig. 3.13 shows the control block diagram of the pipelined quad-phase saving scheme for the 2σ write methodology. There are additional four shift registers delaying the input data, which will be compared with the data saved to the

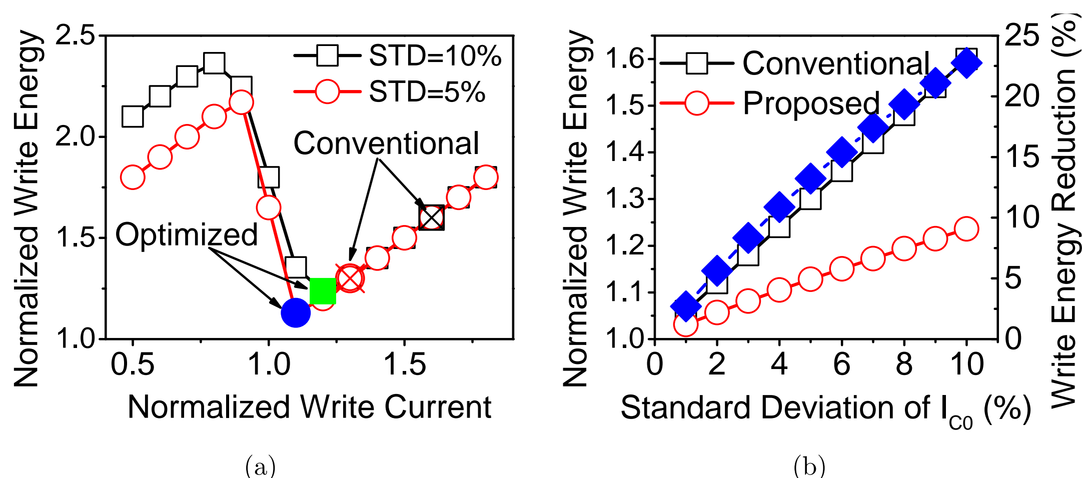


Figure 3.11: (a) The relationship between the first write current amplitude and the total write energy with our proposed write scheme. (b) The relationship between the standard deviation of I_{c0} in percentage and the write energy improvement with our proposed write scheme.

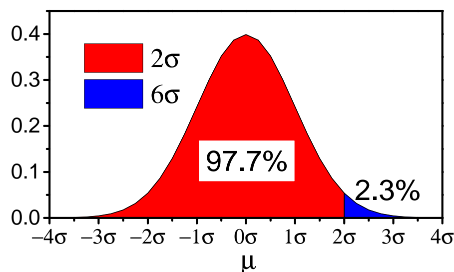


Figure 3.12: The distribution of the 2σ writing.

STT-MRAM array. The CTRL block generates the same quad-phase signals. The output write pulses are alternately switched between 2σ write and 6σ write, which are controlled by the four-phase signal $Scan_clk$.

A simplified array diagram is illustrated in Fig. 3.14. Each four channels are similar to the block diagram provided in Fig. 3.8. The output data are switched between the first write and second write, and controlled by $Scan_clk$. Additional four latches are added in the row shift address to generate the pattern "8'b11001100". The fifth latch is the first bit of the row address.

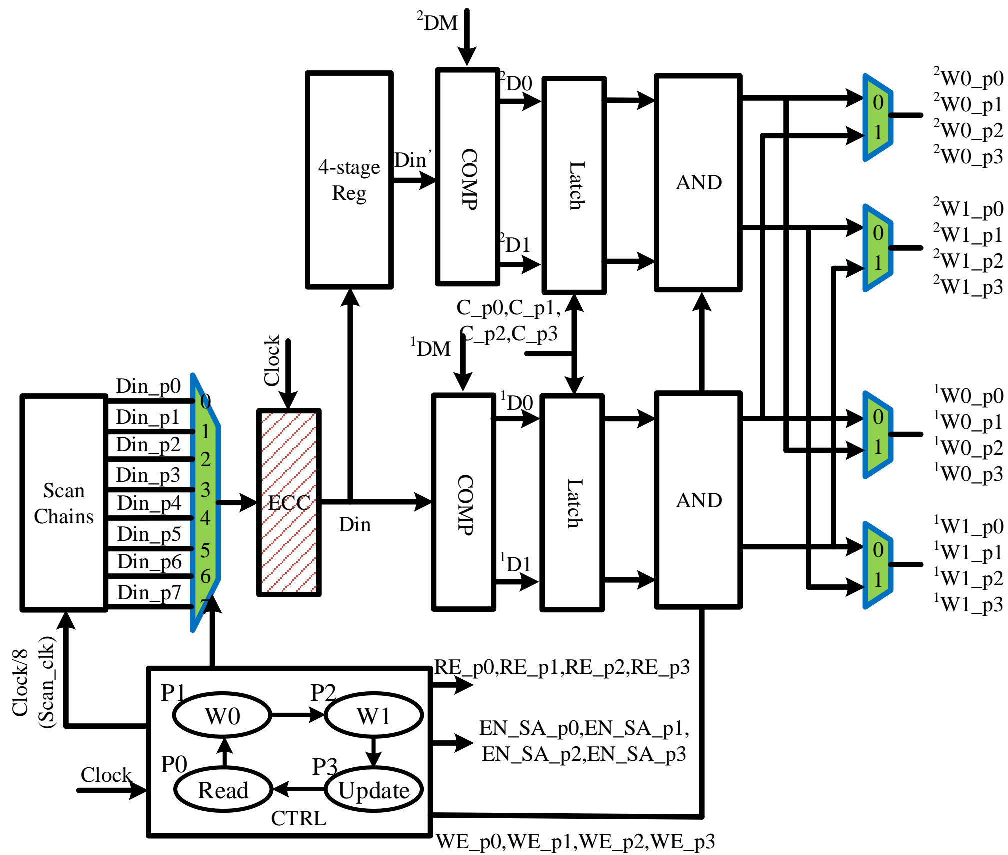


Figure 3.13: Proposed pipelined quad-phase control block diagram for the 2σ saving approach.

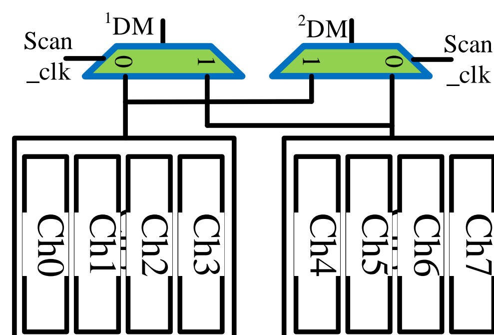


Figure 3.14: The block diagram of 8 memory channels for the 2σ saving approach.

Table 3.2: Example of pipelined quad-phase saving scheme with the 2σ write approach. Row clock is used in the table.

Clock	c0	c1	c2	c3	c4	c5	c6	c7
0.5	$read^1$	0	0	0	0	0	0	0
1	$w0^1$	$read^1$	0	0	0	0	0	0
1.5	$w1^1$	$w0^1$	$read^1$	0	0	0	0	0
2	0	$w1^1$	$w0^1$	$read^1$	0	0	0	0
2.5	$read^2$	0	$w1^1$	$w0^1$	$read^1$	0	0	0
3	$w0^2$	$read^2$	0	$w1^1$	$w0^1$	$read^1$	0	0
3.5	$w1^2$	$w0^2$	$read^2$	0	$w1^1$	$w0^1$	$read^1$	0
4	0	$w1^2$	$w0^2$	$read^2$	0	$w1^1$	$w0^1$	$read^1$
4.5	$read^1$	0	$w1^2$	$w0^2$	$read^2$	0	$w1^1$	$w0^1$

As shown in Table 3.2, our proposed pipelined quad-phase write scheme with 2σ write approach has four channels in 2σ write period ($read^1$, $w1^1$ and $w0^1$) and the other four channels in 6σ write period ($read^2$, $w1^2$ and $w0^2$). The eight channels pipelinedly shift their phases, and each channel has one phase delay. Therefore, the latency from the scan chain to the first data successfully written is 8 clock cycles (4 row clock cycles). At each saving clock cycle, there are 2 read operations and 4 write operations. Only few of the 4 write operations may happen simultaneously due to the 2σ write and read-before-write approaches.

3.3.5 Reference Resistance Generator

STT-MRAM which offers advantages in endurability, scalability, speed and energy consumption over other types of non-volatile memory [147, 148] has attracted increasing research interests. The spin transfer torque (STT) switching technique enables MRAM scalability beyond $90nm$ and leads to simpler memory architecture and manufacturing than conventional MRAM [149,150]. As the process technology shrinks, the write current can be reduced as it is dependent on the size of the MTJ. The scaling down of technology, however, increases the process variation and decreases the supply voltage, which poses great challenges for STT-MRAM

circuit design to maintain the sensing margin. The sensing margin is defined as the voltage difference between the bit line voltage during read operation and the reference of the sense amplifier subtracting the offset voltage and noise. Employing the differential sensing architecture [151] doubles the sensing margin but sacrifices the density of the STT-MRAM array. Further, as its read and write operations share the same current path, STT-MRAM has a known issue of "read disturbance", which is an unintended write occurring during a read operation [152]. Read disturbance occurs when the read current is larger than the critical switching current (I_C) of the write operation. Consequently, the read current is required to be small enough to prevent potential read disturbance for STT-MRAM.

The sensing margin in STT-MRAM can be expressed as $I_{read} \times |R_{MTJ} - R_{ref}|$, where I_{read} is the reading current, R_{MTJ} is the resistance of the MTJ, which could be R_P and R_{AP} for P and AP states of the MTJ, respectively, and R_{ref} is the equivalent resistance of the reference circuit. The requirement of low sensing current, small TMR ratio and distribution of resistance in both high resistance state (AP-state) and low resistance state (P-state) will further reduce the sensing margin.

The conventional design [58] uses two reference cells in parallel per row to generate a reference voltage. If assuming no variance of MTJ reference cell resistance, the total equivalent resistance is $(R_P||R_{AP}) = (1 + TMR)/(2 + TMR)^2 * (R_P + R_{AP})$, where TMR is defined as $TMR = (R_{AP} - R_P)/R_P$. The reference cell resistance, however, follows similar distribution of the resistance of the array cells. Taking such distribution into consideration, the sensing margin will become smaller because both array cell resistance distribution and reference cell resistance distribution will deteriorate the sensing margin. Maximizing the sensing margin will loosen the requirement of the sense amplifier and increase the read reliability. Since R_{MTJ} and TMR are determined by fabrication process and material charac-

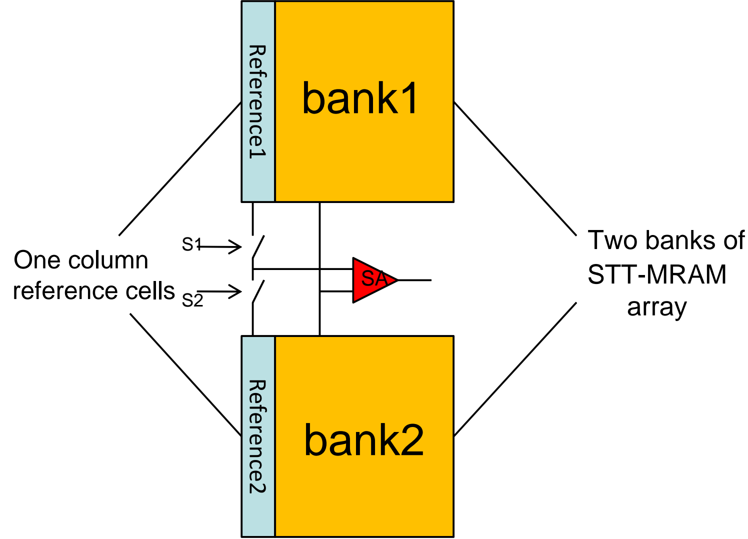


Figure 3.15: Share the reference columns for two adjacent banks, reference1 is from bank1 and put closely to bank1 array, while reference2 is from bank2 and put closely to bank2 array and sense amplifier is shared by two banks of STT-MRAM array.

terization, and I_{read} is constrained by read disturbance consideration, one possible improvement of sensing margin in the circuit design is to reduce the distribution of the resistance of the reference cells. In [139], a merged reference line (MRL) method to reduce the distribution of the reference resistance has been proposed. However, the MRL scheme consumes high power on the reference circuitry during read operation. Since all reference MTJs are in parallel, the equivalent reference resistance is $1/N \times (R_P//R_{AP})$. To make the potential on the reference node in between $I_{read}R_P$ and $I_{read}R_{AP}$, the reading current in the reference path should be N times larger. As a result, 64 reference pairs drew 64 times higher current in the reference circuit than that in [58].

We propose a novel reference circuit architecture that maximizes the sensing margin through averaging the resistance of reference cells from one or two columns of the reference array from one or two memory banks.

The proposed scheme, shown in Fig. 3.15, solves the distribution issue of

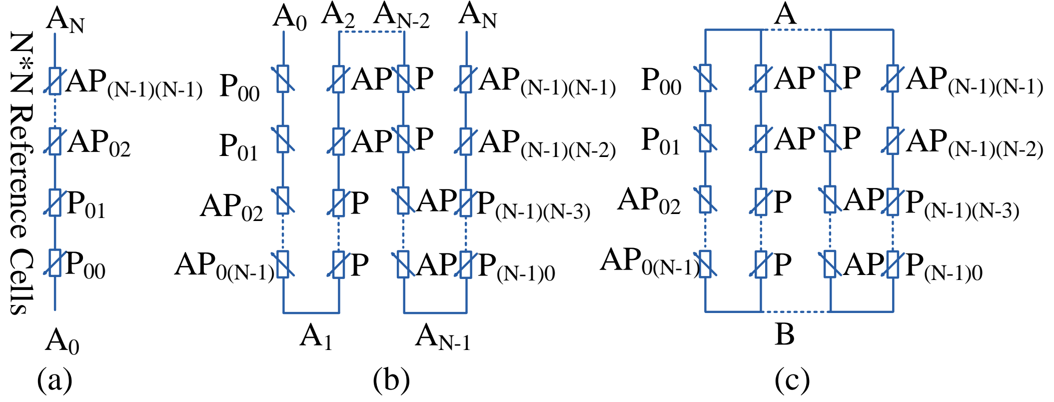


Figure 3.16: Example for concept of reference cell folding. (a) Reference cells connected in series before folding. (b) Folding the whole column of reference cells to a $N \times N$ array. (c) Final construction of the $N \times N$ reference array by connecting the folded points.

the reference resistance, through averaging the resistance of one or two columns of reference cells from one/two banks, as the reference resistance. In Fig. 3.15, two banks of STT-MRAM array of the same dimension, bank1 and bank2, share the sense amplifiers. Each bank has a dedicated column of reference cells, denoted as reference1 and reference2, respectively. The two reference columns are connected to the reference node of the sense amplifier through two switches that are controlled by S1 and S2. If the number of cells in the reference column equals to 2^{2n}, where $n \in [0, 1, 2...)$, the averaged resistance of that column of cells will be used as the equivalent resistance. S1 or S2 will turn off the switch and connect the equivalent resistance to the sense amplifier to sense the cells in bank1 or bank2, respectively. If the number of cells in the reference column equals to 2^{2n-1}, the averaged resistance of both columns will be used as the equivalent resistance by asserting both S1 and S2. The resulting equivalent resistance will be used to sense cells in both bank1 and bank2. Fig. 3.16 shows the detailed concept of this reference averaging scheme. The equivalent $N \times N (N = 2^n)$ reference array is obtained by folding the cells in one column and connecting the folded points $A_0 - A_N$ as

detailed in Fig. 3.16. In each column of the equivalent reference array, half number of cells are programmed at P states and another half are programmed at AP states. Ideally, the equivalent resistance of this reference array is $\frac{1}{2}(R_P + R_{AP})$.

Fig. 3.17 illustrates an implementation of the equivalent $N \times N$ reference circuit when there are 2^{2n} cells in one reference column. Cells in the column are averaged to obtain the equivalent resistance. The linked MTJ cells are alternatively connected to SL_{ref} and BL_{ref} through the write access transistors, and alternatively connected to the sense amplifier and the ground every other N reference cells. The control signals of the write access transistors are generated by the row decoder. To program the selected reference cell, the two connected access transistors are turned on, and all other access transistors are turned off to avoid any unintended write. The reference cells are programmed sequentially before the main STT-MRAM array. Their states are determined by the voltages on BL_{ref} and SL_{ref} during programming. For example, to write data 1 to the reference cell, BL_{ref} and SL_{ref} are connected to the write source and the ground, respectively. The concept in Fig. 3.16 can be achieved by programming the reference cells through the input pattern 0011... at BL_{ref} and SL_{ref}. Other patterns can also be used to program the reference cells to get the desired ratio of P and AP states.

The reference current has the same current amplitude as read current I_{read}, and an equivalent 16×16 reference array produces a better distribution than [139] with 64 reference pairs case. The reference current of the proposed circuits is around half of that in [58], which used 1 pair of reference cells, and around $\frac{1}{128}$ of that in [139] with 64 pairs of reference cells. Since the reference column in the proposed design has been re-arranged to an $N \times N$ array, the current that goes through each cell is only $\frac{1}{N}$ of I_{read}, thus the read disturbance is dramatically reduced [144]. In this design, the resistance model in P and AP states during

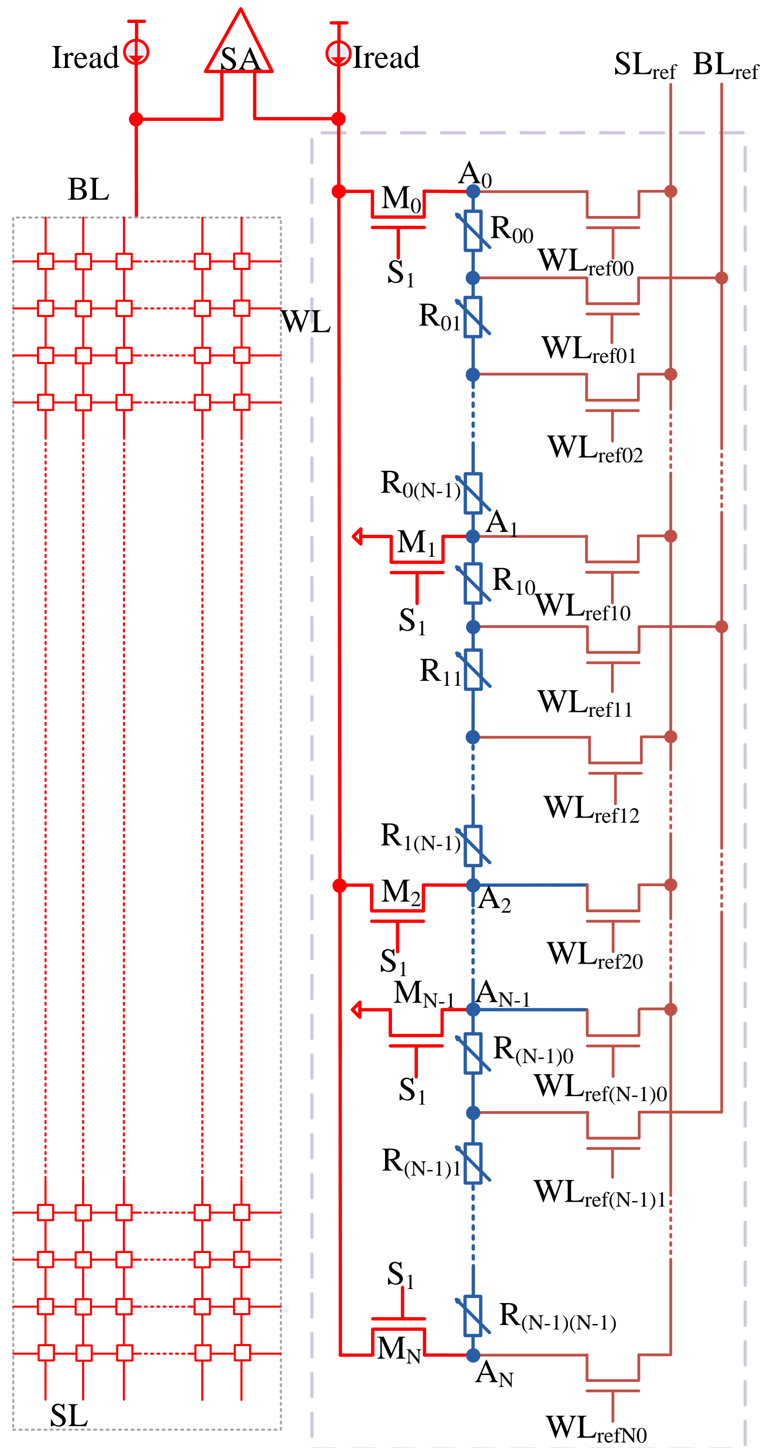


Figure 3.17: A circuit implementation of the equivalent $N \times N$ reference circuit when there are 2^{2n} cells in one reference column in which cells are averaged to obtain the equivalent resistance.

reading is defined as [89, 153],

$$\begin{cases} R_P = R_{0P} \\ R_{AP} = |I_{read}/N| \times K_{AP} + R_{0AP} \end{cases} \tag{3.11}$$

where K_{AP} is the slope of R_{AP}, R_{0P} and R_{0AP} are the zero current resistances.

Therefore, the TMR of the reference cells in $N \times N$ equivalent array during reading is

$$TMR = \frac{I_{read}K_{AP}}{NR_P} + TMR_0 \tag{3.12}$$

where TMR_0 is the TMR of a MTJ at zero read current.

Due to the equivalent environment, each MTJ cell contributes $\frac{1}{N^2}$ resistance. Therefore the total distribution can be derived from the following equation,

$$f(R) = \frac{1}{\sqrt{2\pi\sigma_N^2}} e^{-\frac{(R-R_N)^2}{2\sigma_N^2}} \tag{3.13}$$

where $R_N = \frac{R_P + R_{AP}}{2}$ is the mean of the reference resistance, and σ_N is the standard deviation of R_N, which has the equation,

$$\sigma_N = \frac{\sqrt{\sigma_P^2 + (1+TMR)^2\sigma_{AP}^2}}{(1+0.5 \times TMR)\sqrt{2}N} \tag{3.14}$$

It can be observed from (3.14) that as N increases, the standard deviation of the equivalent resistance can be greatly reduced. Another advantage of the proposed scheme is that even if one or few cells have read disturbance or are not correctly programmed [86], the mean of the reference resistance hardly shifts. Therefore, the circuits to detect failure of the reference cells or reference to neighboring blocks/redudency cells are not necessary. To simplify the analysis, it is discussed here only the case when one AP cell is not programmed. The mean of the $N \times N$ equivalent reference block with one AP cell stuck at P state is

Table 3.3: Description of the 45nm embedded MTJs process.

Device parameters	value
MTJ Size	$65nm * 65nm$
TMR	100%
RA	$13.3\Omega{\cdot}\mu m^2$
$J_0^{AP\rightarrow P}$@6σ	$4e10A/m^2$
$J_0^{P\rightarrow AP}$@6σ	$3e10A/m^2$
Δ	65

$$R_{MEAN} = \frac{N(2+TMR)(N+\frac{N}{2}TMR-TMR)}{2N^2+(N^2-2N+2)TMR}R_P \tag{3.15}$$

The shift of resistance from the mean (R_N)of the reference circuit in percentage is

$$\Delta_{MEAN}\% = \frac{TMR}{N^2+\frac{1}{2}((N-1)^2+1)TMR} \times 100\% \tag{3.16}$$

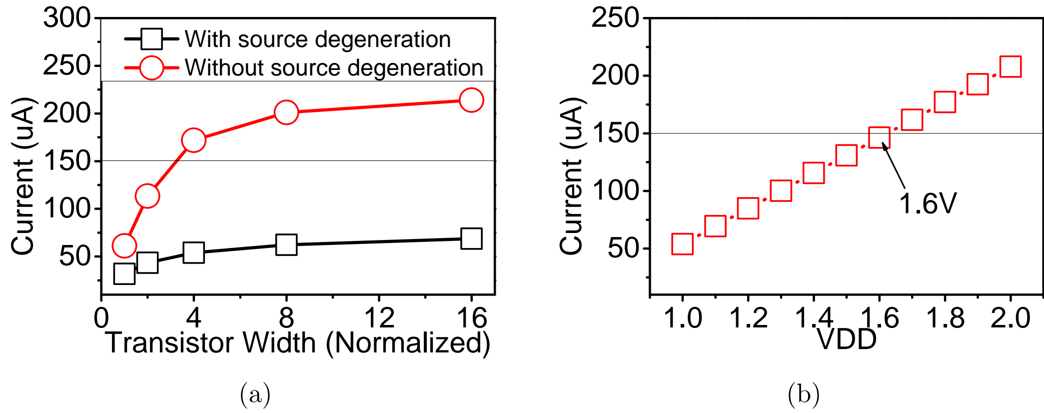
If N is large enough, $\Delta_{MEAN}\% \approx 0$, thus the shift of the mean resistance can be neglected.

3.4 Simulation Results

In this section, we firstly ran spice simulations to show the improvement of the proposed schemes over the conventional nvFF based schemes. Moreover, we also analyze the impact of the scan chain length on the amount of the power reduction. After that, we analyze the impact of the MTJ parameters and equivalent reference array size on the reference resistance generator. The MTJ model in [88,89] is used in this chapter for the simulation. The detailed description of the model has been provided in Section 1.2.1.

3.4.1 Spice Simulation Results of the Proposed Array

A $100MHz$ system clock is used in the simulation. G is set to 2^{13}, k is set to 16 and 32 for our proposed schemes with and without 2σ write schemes, respectively. Thus the length of scan chain is the same for both schemes. The detailed parameters of the MTJ used in the simulation are tabulated in Table 3.3.



(a) (b)

Figure 3.18: (a) The width of the access transistors vs. the write current that can pass through, (b) the VDD of the 1T1R scheme vs. the write current.

The benefit of the 2T1R scheme can be seen from Fig. 3.18. As shown in Fig. 3.18(a), the "source degeneration" effect will significantly limit the write current. Though the width of the access transistor in the 1T1R scheme is increased significantly, the write current is still far smaller than the required value ($150uA$). Our proposed 2T1R scheme can easily reach the $150uA$ write current when the transistor width is increased by 4 times. Fig. 3.18(b) shows that to pass through $150uA$ write current, the VDD of the 1T1R scheme has to be 60% larger.

Fig. 3.19 shows the transition simulation of the saving operation with 2σ write approach. The write enable signal WE is used to generate write 0 enable signal $W0$ and write 1 enable signal $W1$. $Scan_clk$ is used to switch between 2σ write period and 6σ write period. At around $50ns$, a positive $W0$ pulse indicates an AP to P switch is required, since the data in the memory ($Q = 1$) does not

equal to the input data ($Ds = 0$). A second read shows the same Q and Ds that indicates the data is successfully written into the memory cell. Therefore, no further write operation is required, and both $W0$ and $W1$ are low between $80ns$ and $120ns$.

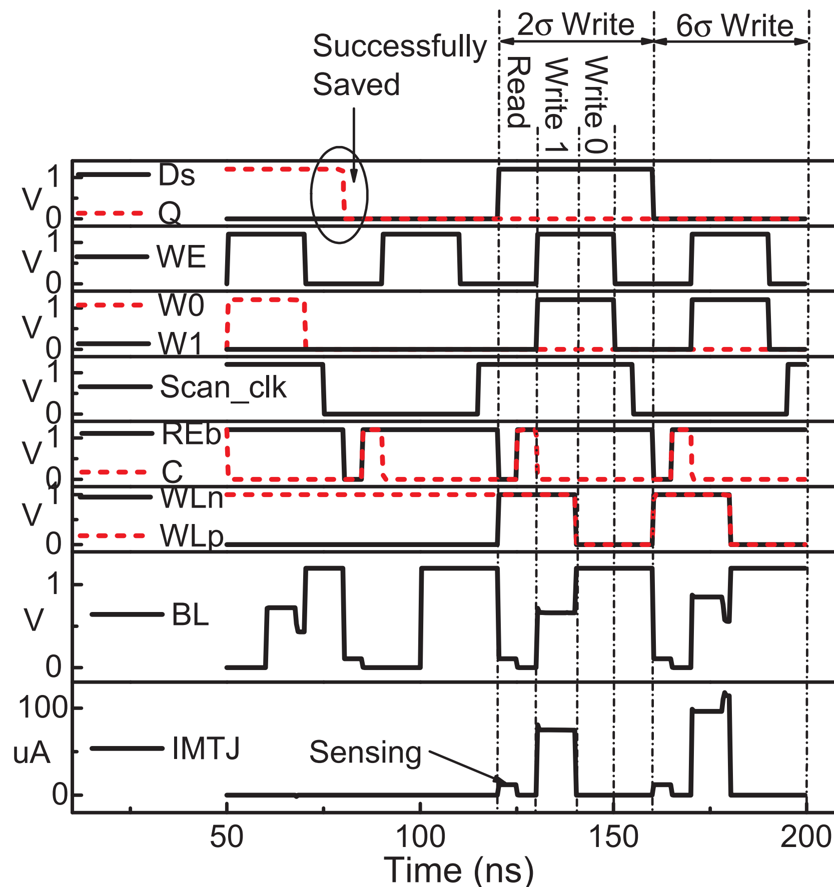


Figure 3.19: The waveform of the read-before-write and verify-after-write functions.

At $120ns$, another data is to be written to the same channel (different row). REb senses the data from the memory array to be compared to the input data. The comparison results are latched by the clock C. When a row is not selected, WLp is high and WLn is low. When the first 2σ write is executed, WLn goes high. In this phase, read and write 1 operations are conducted. When write 1

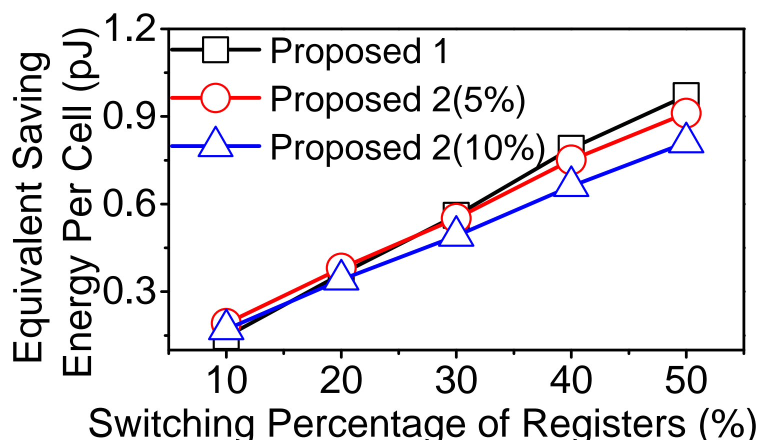


Figure 3.20: The relationship between the power comparison of our proposed two schemes and switching percentage of registers to be saved. 'Proposed 1' and 'Proposed 2' are the scheme without and with 2σ write approach, respectively. In this simulation, the standard deviations of the intrinsic switching current distribution were set to 5% and 10%, and the saving energy of our proposed scheme without 2σ write approach was set to the same for both intrinsic switching current distributions. The scan chain length is set to 64.

operation is finished, both WLn and WLp are pulled to the ground, and the write 0 operation is conducted. A second read operation shows the first write is not successful due to the reduced write current. Therefore, 6σ is performed with a sufficient write current. The current change of signal BL indicates the successful writing of the data.

The benefit of the localized dedicated array is shown in Fig. 3.20. Some registers in the system may have a low possibility to switch their states, i.e., configuration registers, high-order bits of counters, etc. In this simulation, we evaluated the saving energy of our proposed schemes versus different switching percentage of registers. The highest switching percentage of a system is 50%, when all registers are randomly switched. As shown in Fig. 3.20, the saving energy is proportional to the switching percentage of registers. Our proposed scheme 2 (with 2σ write approach) further reduces the saving power when the switching

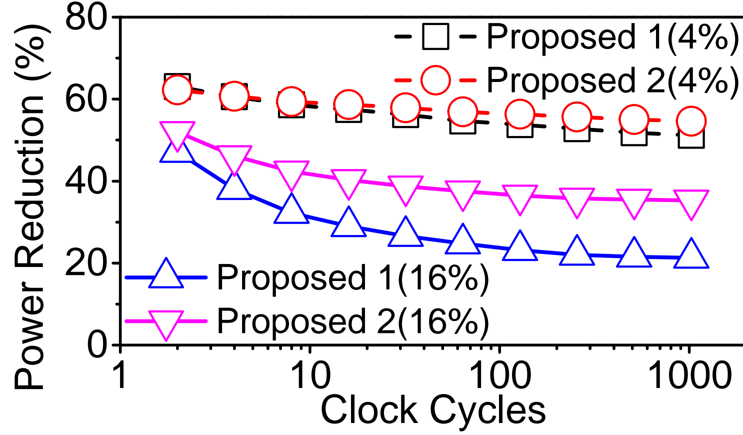


Figure 3.21: The relationship between the power reduction and operation clock cycles. In this simulation, the averaged switching activities of registers were set to 4% and 16%, and the standard deviation of the intrinsic switching current distribution was set to 10%. The scan chain length is set to 64.

percentage is high. We set 6σ switching current the same for all simulations, thus making the write power of our proposed scheme 1 the same for all simulations at different I_{c0} distributions.

The switching percentage may also be affected by the clock cycles of the digital blocks after powering on. Many registers may not switch their states between two adjacent sleep periods, especially when the "on" period is short. We set the mean switching rates of registers to 4% and 16% to evaluate the relationship between clock cycles and the saving power, as shown in Fig. 3.21. The power reduction is compared to the nvFF proposed in [11] after being converted to the single cell saving energy, which consumes $1.3775pJ$ sleep energy with the same MTJ parameters provided in Table 3.3. Fewer 'on' clock cycles between two sleep periods lead to a much higher power reduction. The low switching rate of registers states has higher power reduction. After 1000 cycles, the 16% case is almost saturated (registers switching rate is 50%), the power reduction of the proposed schemes 1 and 2 are 20% and 35%, respectively. In other words, the

proposed schemes 1 and 2 may reduce the sleep power by more than 20% and 35%, respectively. The 4% case needs more clock cycles to be saturated.

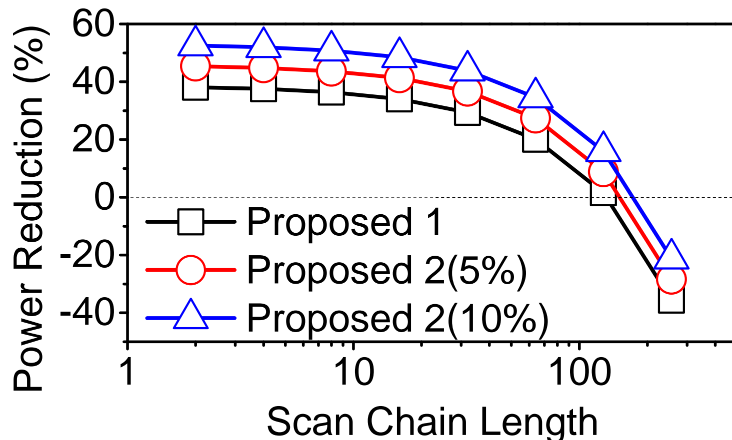


Figure 3.22: The relationship between the power reduction and the scan chain length. In this simulation, the standard deviations of the intrinsic switching current distribution were set to 5% and 10%, and 50% of the registers were switched.

The length of the scan chain may determine the sleep power consumption of our proposed schemes. We evaluated the relationship between the length of the scan chain and the power reduction. As shown in Fig. 3.22, a short scan chain may reduce the power by more than 35%. In contrast, a scan chain longer than 256 increases the power by more than 20%, since shifting a scan chain dominates the sleep power. The sleep power of 'proposed 1', 'proposed 2 (5%) and 'proposed 2 (10%)' schemes can be reduced when the lengths of their scan chains are shorter than 133, 158 and 183, respectively.

Table 3.4.1 tabulates the area comparison among our proposed schemes, conventional nvFFs and the CMOS retention FF. The area of our proposed schemes is much smaller than the nvFF based schemes. If the MUXes used for scan chains are not included as the area overhead, the area could be reduced by more than 50%. Even the transistors of MUXs for scan chains are included, the area reduction is still more than 30%.

Table 3.4: Per cell area overhead comparison among different retention schemes. The data in the '()' have included 6 transistors for scan chains. The number of transistors are estimated based on M=64 and G=8K.

Schemes	Proposed 1	Proposed 2	[9]	[11]	CMOS
Unshared write transistors	2	2	4	4	-
Shared write transistors	4/M	4/M	0	0	-
Other transistors	2.77(8.77)	3.17(9.17)	11	9	9
Total equivalent min width transistors	11.77(17.77)	12.17(18.17)	27	25	9

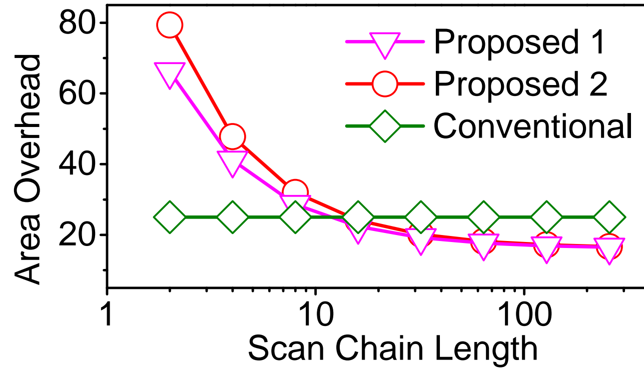


Figure 3.23: Normalized area overhead. The area is normalized to the minimum width transistors.

Fig. 3.23 shows the area overhead of our proposed schemes versus that of the nvFF scheme. Both of our proposed schemes have an area reduction when the scan chain length is longer than 15. Our proposed scheme 2 has slightly higher area overhead than the proposed scheme 1, but the sleep power is further reduced by more than 7%. The scan chain length of 64 may be an optimized solution when considering both area overhead and power reduction.

From the simulation results, it can be observed that the FF has $5nW$ leakage power. The energy used for saving and restoring operations per single bit in the proposed schemes is less than 1.1pJ. From (3.3), the break even time is t_{BEP}=220μs. In conventional designs, the decoupling capacitor and combination-

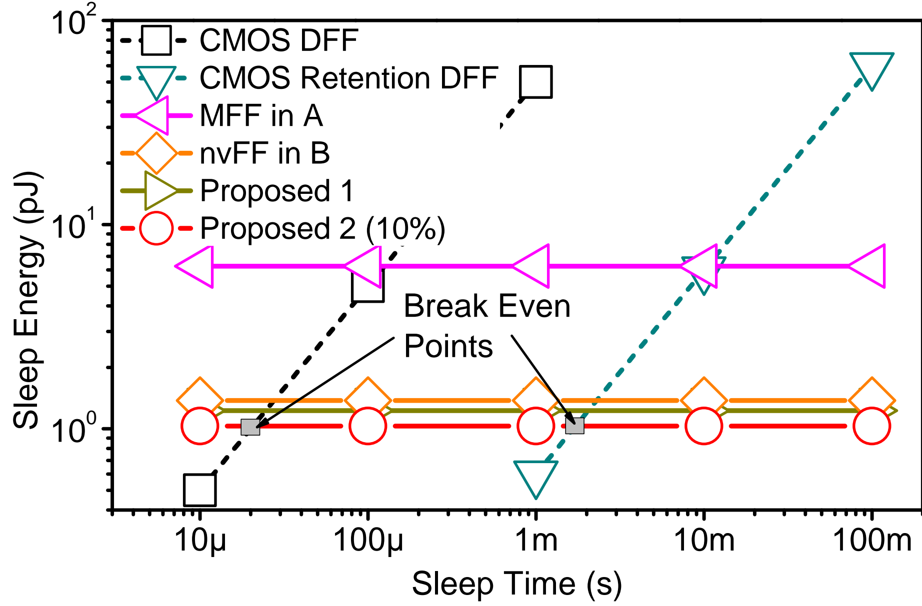


Figure 3.24: The sleep power consumption comparison among conventional structures and our proposed schemes. η is set to 10%. The sleep energy for MFF and nvFF are based on a single cell. A: [9]; B: [11].

al logic also consume the leakage power. Moreover, only a small percentage (i.e., 10%) of the registers need to retain their states. Hence, the equivalent bit leakage is much larger than the leakage of a single DFF. Fig. 3.24 shows the comparison among our proposed schemes, CMOS FF, conventional retention FF, the MFF in [9], and the nvFF taken from [11]. We assume the leakage power consumed by the retention FFs is 10% of the total system leakage power. In such condition, the BEP is less than $22\mu s$ with our proposed schemes. Usually the sleep time of a sensor network or a mobile system is around a few seconds to thousands of seconds. Therefore, the sleep energy could be reduced by more than 99.8% compared to CMOS retention FF based technology. Another conventional scheme is based on the MFF in [9] which required $12.5pJ$ energy for storage. The data is estimated based on $200MHz$, $2.5V$ and $1mA$ write energy for a differential structure, allowing the cell to be successfully programmed. Thus the equivalent write energy for a

Table 3.5: The comparison among non-volatile Flip-flips and proposed schemes. The sleep energy and t_{BEP} are based on M=64. η is set to 10%.

Structures	**Sleep Cost**		t_{BEP}	$t_{sleep,min}$
	Time	Energy		
Proposed 1	$10ns*(G/k+4)$	$1.1pJ*G$	$22\mu s$	$27.16\mu s$
Proposed 2 (5%)	$10ns*(G/k+8)*2$	$1pJ*G$	$20\mu s$	$25.2\mu s$
Proposed 2 (10%)	$10ns*(G/k+8)*2$	$0.9pJ*G$	$18\mu s$	$23.2\mu s$
MFF in [9]	$5ns*2$	$6.25pJ*G$	$125\mu s$	$125\mu s$
nvFF in [11]	$10ns*2$	$1.3775pJ*G$	$26.8\mu s$	$26.8\mu s$

single cell structure is around $6.25pJ$. The design in [11] consumes $1.3775pJ$ sleep energy (after being converted to a single MTJ structure).

The energy and time cost for sleeping among nvFFs and our proposed schemes are compared in Table 3.4.1. Our proposed scheme 1 and 2 reduce the sleep power by more than 20% and 35%, respectively. Though the proposed schemes require more time for saving and restoring operation than nvFFs, the $t_{sleep,min}$ could be smaller than conventional nvFFs. For example, the $t_{sleep,min}$ of the design in [11] is $26.8\mu s$, which is slightly smaller than that of 'proposed 1', but 15% larger than 'proposed 2 (10%)'. To save the states to NVM cells, nvFFs based approaches may have to provide G times more of the write current than proposed one, thus the peak current may be significantly high during saving operation. For example, if there are 8K bits nvFFs with $0.5mA$ write current enter to sleep mode, the saving current is $4A$. Hence a small parasitic resistance may lead to high voltage drop and significant power loss. In comparison, the peak power of our proposed scheme with $M = 64$ is only around $3mA$.

3.4.2 Analysis of the Reference Resistance Generator

The proposed reference scheme was verified by a Python language program with different settings of the MTJ parameters based on $1,000,000$ samples of static data. Fig. 3.25(a) shows the relationship between the standard deviation of different

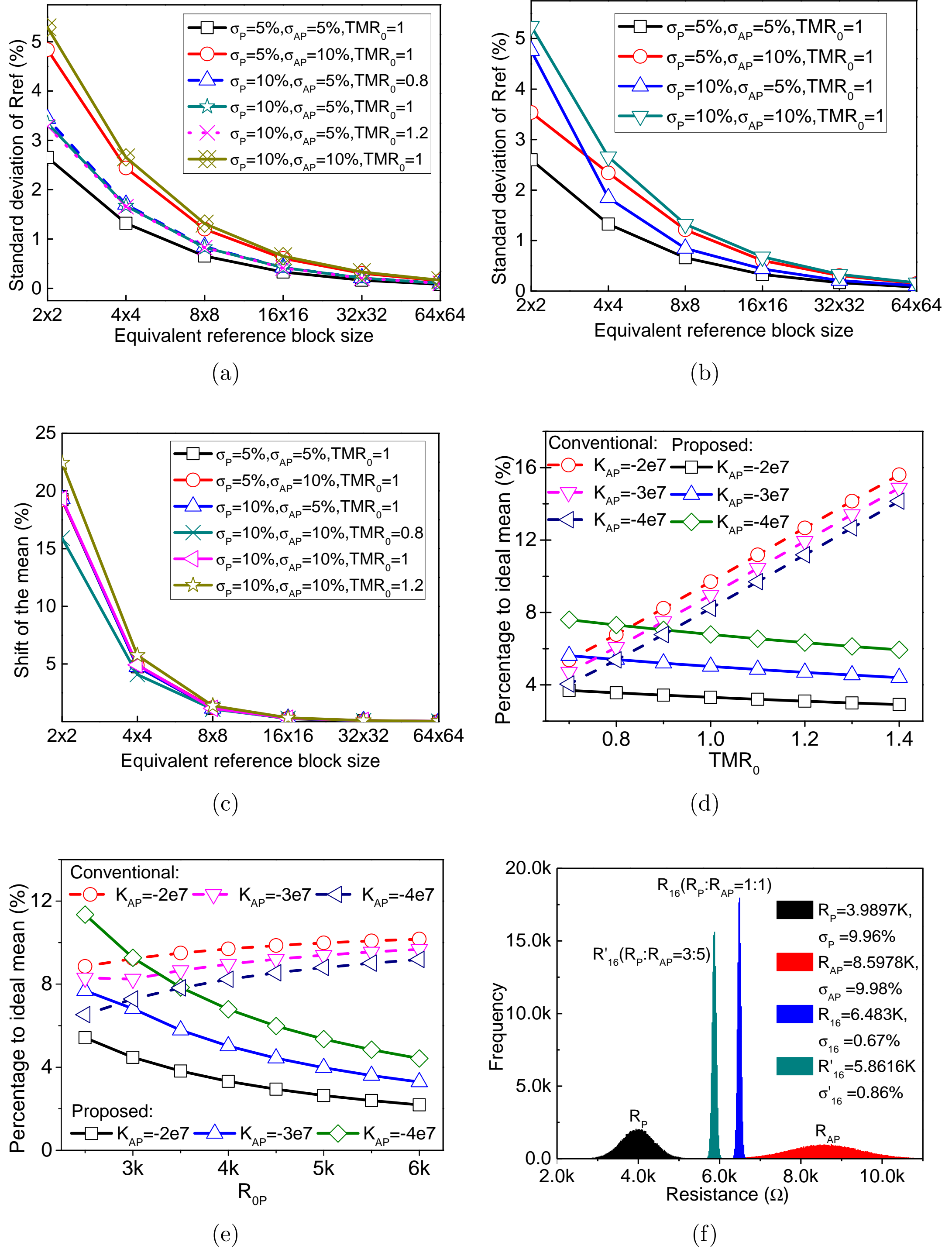


Figure 3.25: Python simulation results for distribution and deviation versus different equivalent reference block size. Distribution of the 16×16 equivalent reference array versus σ_P and σ_{AP} (a) without write failure and (b) with one AP cell stuck to P state; (c) Shift of the mean versus different equivalent reference block size; Deviation from the ideal mean versus (d) TMR ($R_{0P} = 4000$) and (e) R_P ($TMR_0 = 1$) with different slope of R_{AP}, where $I_{read} = 20\mu A$, $N = 16$; (f) Circuits simulation results for equivalent 16×16 reference block size. The standard deviations of both R_P and R_{AP} are set to 10%

equivalent reference array and equivalent reference array size. When the averaged block size increases, the standard deviation of the reference resistance reduces for all cases of different resistance distribution of R_P and R_{AP}, and different TMR. When equivalent reference array size is 16×16, the standard deviations are all smaller than 1% even when both R_P and R_{AP} deviations are set to 10%. The equivalent reference array size of 16×16 or 32×32 could be an optimized choice with the balance of the block size and the standard deviation of reference resistance. It also can be seen from Fig. 3.25(a) that arrays with smaller TMR gets higher distribution. In other words, better TMR may help reference resistance distribution performance.

Fig. 3.25(b) shows the relationship between standard deviation of different equivalent reference array and equivalent reference array size when one AP cell is stuck to P-state. The results show that the reference deviation is very close to the results in Fig. 3.25(a) when the equivalent reference array size is larger than 4×4. Fig. 3.25(c) shows the results when one R_{AP} cell is not programmed, the shifting of mean versus different equivalent reference array size. Higher equivalent reference array size helps to reduce the shift of the mean. The four curves with the same TMR and different deviation are almost overlapped in Fig. 3.25(c), which indicates that the standard deviation has little effect on the mean shift.

Fig. 3.25(d) and 3.25(e) show the deviation from the ideal 50% mean ($\frac{R_P+R_{AP}}{2}@I_{read}$) versus TMR_0 and R_{0P}, respectively, with different slopes of R_{AP}. Clearly, the deviation is much smaller than conventional design especially with large TMR and R_{0P}. The proposed scheme has a better mean with small K_{AP}, large TMR_0 and R_{0P}.

The Monte Carlo spice simulation results of the circuit with 16×16 equivalent reference block size are shown in Fig. 3.25(f). The spice simulation is also based on $1,000,000$ samples of static data. We can see that the reference resis-

tance tends to be very close to the mean with the standard deviation only 0.67%, although the standard deviation for R_P and R_{AP} are both as large as 10%. The mean of the reference resistance could be adjusted by changing the ratio of P and AP states in the serial connections. Therefore, the overlap between R_{AP} (or R_P) and the reference resistance could be minimized. As shown in Fig. 3.25(f), when the ratio of P and AP states is set to $3:5$, the overlap between reference resistance and R_{AP} gets much smaller than in the $1:1$ ratio setting.

3.5 Summary

A localized STT-MRAM array is proposed to retain the states of the registers through scan chains during sleep. In such scheme, power and area are two key improvements. Moreover, the reliability could be improved if the ECC block is added. The sleep energy could be reduced by more than 99.8% compared to the CMOS retention FF approach when sleep time is longer than $1s$. Our proposed schemes have also reduced the sleep energy and area by more than 20% compared to the conventional nvFF based schemes. The scan chain length of 64 may be an optimized solution when considering both area overhead and power reduction. Meanwhile, an optimization scheme based on reference cell folding technique to minimize the reference resistance distribution of STT-MRAM is proposed, discussed and verified in simulations. The proposed circuits substantially reduce the resistance distribution effect and increase the reliability of the readout data. It also reduces the design complexity of the sense amplifier and increases the signal to noise ratio of the data. The proposed optimization scheme refrains the use of high reference current and thus greatly reduces the power consumption of the overall system. The simulation results show that, a block of 16×16 cells for reference averaging provides a good balance of the block size and the reference resistance distribution.

Chapter 4

Non-volatile Switch based FPGA

This chapter is written mainly based on the paper "High Density and High Reliability non-volatile Field Programmable Gate Array (FPGA) with Staked 1D2R RRAM Array".

4.1 Introduction

Several works have been reported in [129–131, 134, 135, 154–156] to integrate R-RAM cells to achieve low power and high performance nvFPGAs. The most straightforward way to integrate NVM in FPGAs is to replace the conventional 6T SRAMs with NVM based new configuration elements, as reported in [129–131]. Despite area efficiency, the designs in [129–131] may suffer from low data retention, since DC biased NVM cells may switch their states during the FPGA operation. Another way is to directly replace both 6T SRAMs and NMOS transistors with the NVM cells in SBs and CBs [129, 134, 135]. A key challenge in this scheme is the interconnect configuration due to the high leakage current in the sneak path. The last solution is to integrate NVM is the non-volatile LUTs (nvLUTs) with crossbar architecture as suggested in [155, 156]. However, such topology cannot

be used for the interconnect, and also has the low read/write reliability limitation due to the high leakage current in the sneak path.

We propose a novel nvFPGA architecture based on the emerging RRAM technologies. With the fully utilization of high resistance ratio, excellent scalability, and high density, RRAM is organized in a 1D2R ('1 diode, 2-RRAM cells') structure. This novel structure is used to replace both SRAMs and NMOS transistors to address the sneak path issue, thus significantly improving the write reliability. Moreover, we propose a complementary look up table (LUT) structure, which greatly reduces the area, delay and power consumption. In our proposed nvFPGA, the diode of '1D2R' is only used during configuration. During normal operation, the diode is not involved and the interconnect become a diode-less crossbar array. By stacking RRAM cells on the top of CMOS circuitries, our proposed nvFPGA architecture can exhibit smaller footprint (78% smaller), higher performance (1.94 times faster), and lower power consumption (40.9% lower). The write reliability is significantly improved by more than $9e7$ times compared to other RRAM-based nvFPGAs.

4.1.1 Baseline 2D FPGA

As shown in Fig. 4.1, a traditional two-dimensional island FPGA architecture taken from [42] is used as the baseline in this chapter. It consists of a number of tiles. Each tile contains one SB, two CBs and one LB, and each LB contains some local routing structures (local interconnect) to route input signals to several basic logic elements (BLE) and also connect the BLEs' outputs to their inputs. LBs connect to the routing channels through CBs. The number of routing tracks to the LB IOs is controlled by an architectural parameter F_c (ratio of routing tracks to the LB input and the channel width W). The global routing structure consists of two-dimensional segmented interconnect channels connected by programmable

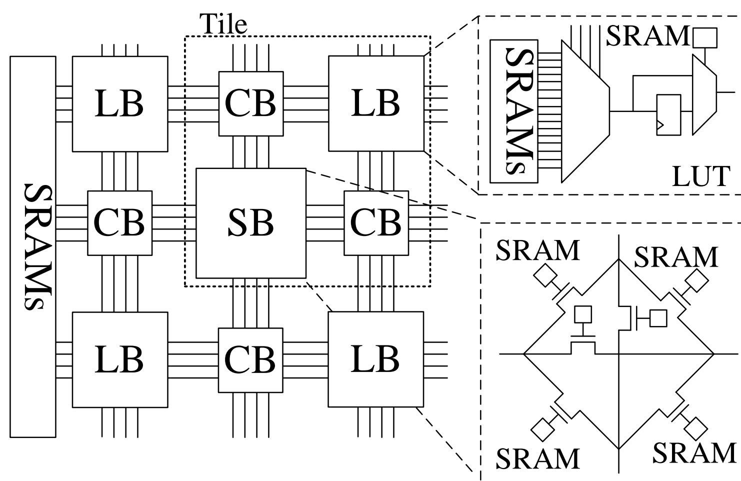
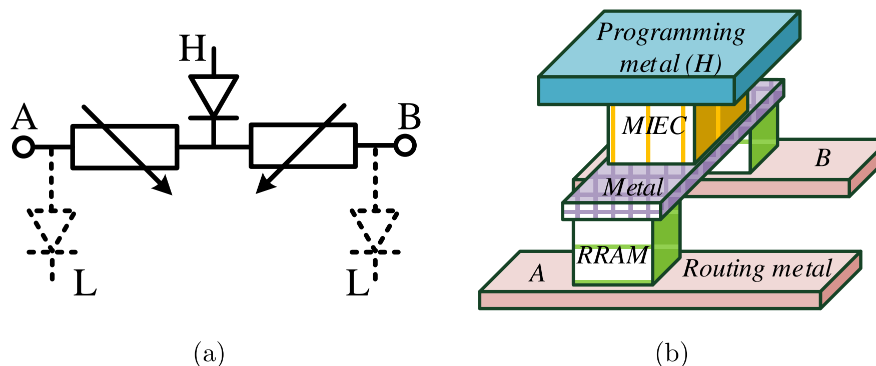


Figure 4.1: A simple island style SRAM-based FPGA layout.

SBs.

4.1.2 Access Device

A significant hurdle to realize the RRAM integration in the FPGA is the sneak path issue which occurs in passive CBs, SBs and local interconnects. In order to avoid the sneak path and achieve the high density, diode is used as the access device because it is back-end of line (BEOL) friendly. Furthermore, it can also provide high driven current and large ON/OFF ratio. IBM has demonstrated a novel diode based on Cu-ion motion in Cu-containing Mixed Ionic Electronic Conduction (MIEC) materials, which supports extremely high current density ($>50MA/cm^2$) and large ON/OFF ratio ($\geq 10^7$) [157]. Stacking RRAM and diode on top of the FPGA CMOS part can significantly reduce the FPGA area and delay, thus greatly improving the FPGA performance.



(a) (b)

Figure 4.2: (a) The proposed non-volatile element to replace the FPGA routing switch and 6T SRAM. Adjacent non-volatile elements connecting to A or B share the same diodes. (b) A 3D schematic of the proposed non-volatile element. Metal line A or B may be routed at different layers depending on the routing direction.

4.2 Proposed Storage Element

In view of above, the access device is indispensable to reduce the sneak path current and improve the reliability, but it cannot be embedded in the FPGA routing lines. Due to the write scheme used in our proposed nvFPGA to eliminate the sneak path, the 'positive set, positive reset' unipolar switching behavior is used in this nvFPGA design. We propose a '1D2R' based non-volatile element to replace both 6T SRAM and FPGA routing switch as shown in Fig. 4.2. It consists of two RRAM cells and one diode. The two RRAM cells are simultaneously programmed to both low or high. In the FPGA operation mode, the diodes are disabled and the two RRAM cells are working as a routing switch in the nvFPGA: when both are at HRS, the switch is turned off due to RRAM's high resistance; when both are at LRS, the switch is turned on to propagate the signal. In the FPGA configuration mode, our proposed '1D2R' based non-volatile element works as a '1D2R' memory cell in a crossbar array.

Additional two diodes at nodes A and B are used instead of the CMOS as reported in [158]. The diode could supply higher current density than CMOS

transistors. More importantly, they can be placed between metals as discussed in Section 4.1.2, to reduce both area and routing complexity. These two diodes are used to program RRAM cells, and they are shared for the adjacent non-volatile elements that connect to A or B. During programming, the node L is pulled down to the ground and the node H is pulled up to V_{set} or V_{reset}, depending on the FPGA configuration information. Since both A and B are pulled to the ground, there is no DC loop to interfere adjacent non-volatile elements during FPGA configuration. In the FPGA operation mode, the diodes are disabled by pulling L and H to VDD and the ground, respectively. The proposed nvFPGA switch structure may double the number of RRAM cells and slightly increase the propagation delay. The slight sacrifices are worthy because the data integrity of the configuration information in RRAM cells can be improved significantly, which is much more important than the speed performance of FPGAs. Moreover, compared to the '1R' scheme, our proposed structure could also reduce the write power and leakage current in the FPGA configuration and normal operation modes, respectively.

A 3D implementation of our proposed non-volatile element is shown in Fig. 4.2(b). The RRAM cells and diode (MIEC material is used in this example) will be stacked between the metals on top of CMOS circuits. All RRAM cells are in the same layer, and their pitch can be as small as $2F$. Therefore, the area of the diode can be at least $3F \times 1F$ to provide sufficient current. The programming metal is the bit line in the crossbar array. The metal line A or B may be routed at different metal layers if they have different routing directions.

4.3 Proposed non-volatile FPGA

In our proposed nvFPGA, there is no CMOS circuitry in SBs and CBs except buffers. We also propose to stack the RRAM on top of CMOS circuitries, which can reduce the area significantly compared to traditional SRAM-based FPGAs.

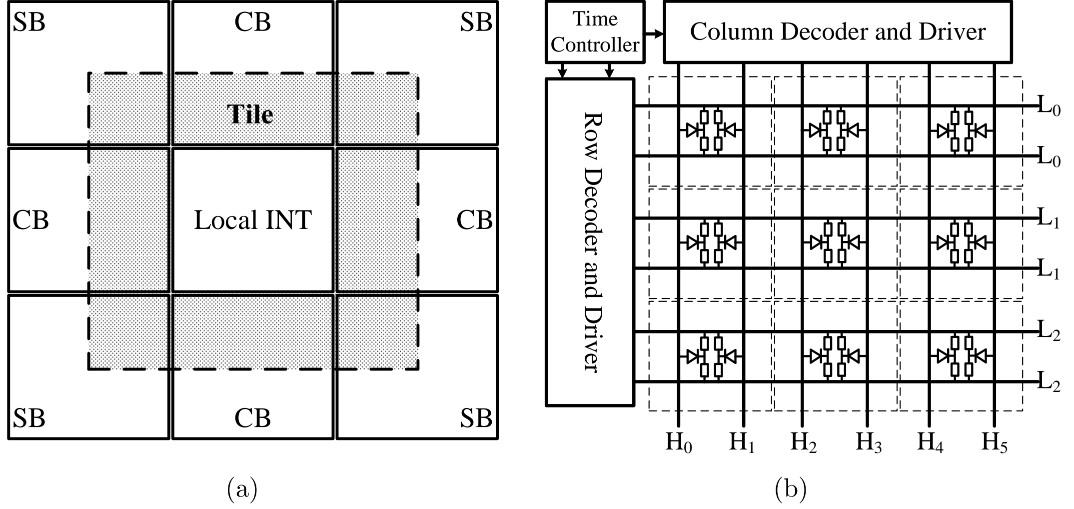


Figure 4.3: (a) Top view structure of the proposed stacking RRAM based nvFGPA, (b) schematic diagram of the memory in our proposed nvFPGA system. The RRAM cells are arranged using '1D2R' crossbar array structure.

A similar island FPGA architecture borrowing from [42] is used in this chapter as shown in Fig. 4.3(a). In our proposed nvFPGA, SBs, CBs and the RRAM part of LBs (Local interconnect, 2-to-1 multiplexer in the BLEs, and RRAM in the LUT) are placed on the top of the CMOS part of LBs and the buffers of CBs and SBs. Therefore, the area is mainly determined by the BLEs and buffers in the interconnect. In such scheme, local interconnect is placed in the center of the tile. Every CB shares the area between two adjacent tiles on the edge, and every SB shares the area among four adjacent tiles at the corner.

The RRAM cells will be arranged as a '1D2R' RRAM crossbar array as shown in Fig. 4.3(b). Each diode connects to one bit line (H_i, where i is the natural number) and two RRAM cells. The other node of the RRAM cell connects to the word line (L_i). Every two word lines are enabled simultaneously to program one diode pair. The RRAM cells are programmed during the FPGA configuration phase.

Our proposed nvFPGA has the FPGA operation mode and the FPGA

configuration mode. The FPGA configuration mode is to program the RRAM cells or write configuration information to the RRAM cells. Unlike the SRAM-based FPGA, our proposed nvFPGA only requires one time configuration. It doesn't need to be reconfigured each time after powering on. Thus the power-on time and energy are significantly reduced. The routing in our proposed nvFPGA is the diode-less crossbar array during FPGA operation that enables high speed, and '1D2R' crossbar array as shown in Fig. 4.3(b) during FPGA configuration that reduces write error rate.

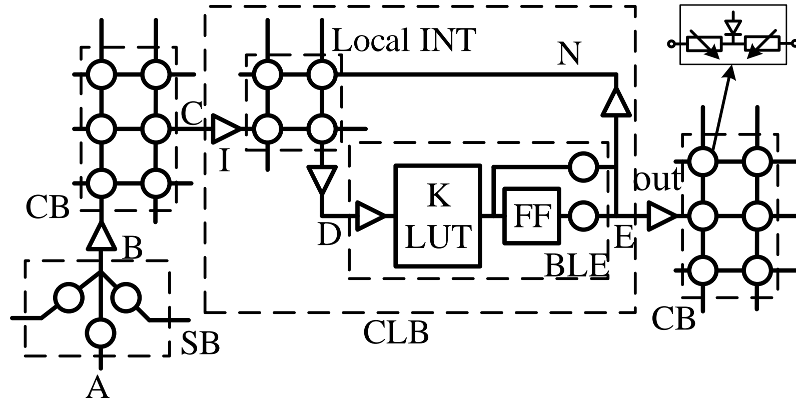


Figure 4.4: The schematic of our proposed '1D2R' based non-volatile FPGA. The crossbar structure is used for both CB and local interconnect.

Fig. 4.4 shows a simplified connection diagram of a tile in the nvFPGA, where I and N represent the number of inputs and clusters in one LB. Each LB has I general inputs, one clock input, and N outputs (where each output corresponds to a BLE. Each BLE consists of one K input look-up table (K-LUT), one FF and a 2-to-1 multiplexer. The BLE inputs can come from either the inputs to the logic block or from the output of other BLEs within the same logic block via a full crossbar array (local interconnect). The main difference between our proposed nvFPGA and the architecture in [42] is that a crossbar structure of the CB and local interconnect is used instead of the multiplexer structure.

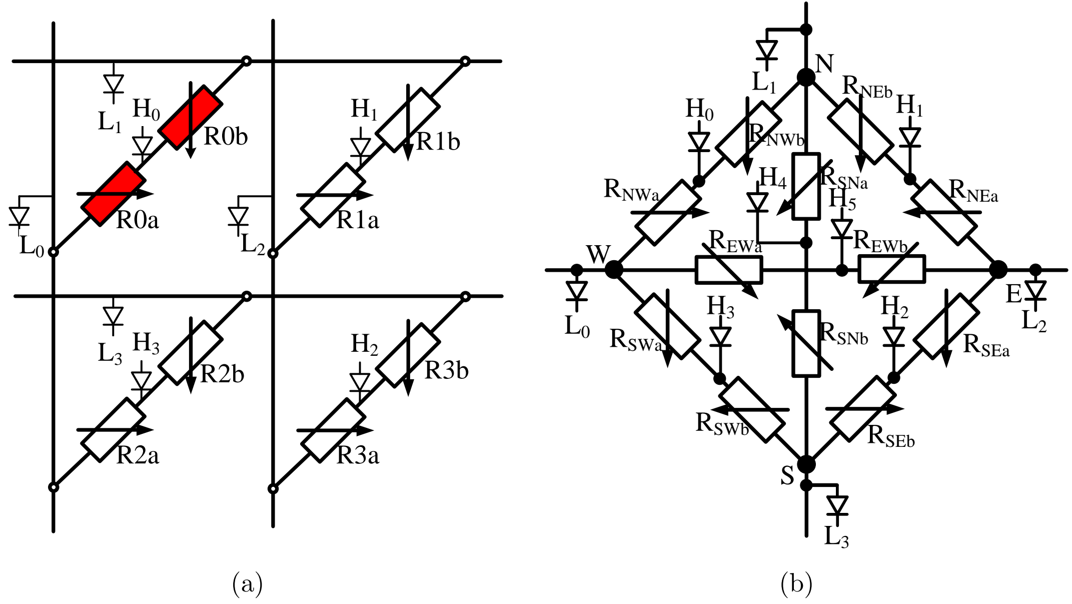


Figure 4.5: The schematic view of '1D2R' based (a) non-volatile crossbar array structure; (b) non-volatile switch point (SP). The non-volatile crossbar array is used in the CB and local interconnect.

The crossbar structure could significantly reduce the delay, since the multiplexer has several transistors in series in the routing path. The detail of each blocks is discussed in the following.

4.3.1 Proposed Crossbar Array and Switch Point

Based on the '1D2R' non-volatile element discussed in Section 4.2, we propose the stacking RRAM based schemes for both non-volatile crossbar array and switch point (SP) as shown in Fig. 4.5(a) and 4.5(b), respectively.

The CBs connect the channel wires to the pins of LBs. There are two major properties that can affect the routing flexibility of a design: 1. the flexibility of the CB, F_c; 2. the CB topology, which is the pattern of switches that make the connection. With the high density benefit of RRAM cells, the crossbar topology, as shown in Fig. 4.5(a), could be used to increase F_c and routing flexibility. In such

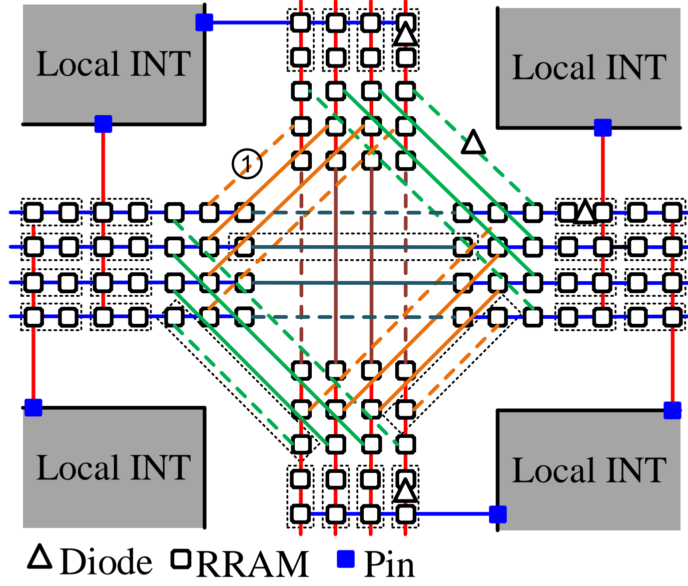


Figure 4.6: The SB and CB structures used in the proposed nvFPGA. The switch box is based on Universal architecture. To simplify, the '1D2R' storage elements show only two RRAM cells in the dash line boxes.

scheme, each logic block pin can be fully connected to the wires in the adjacent channel, and the delay on the switch could also be greatly reduced.

The conventional '1R' approach has the sneak path issue which severely increases the power and degenerates the configuration reliability. To address sneak path limitation, we use '1D2R' structure at each cross point to replace the conventional '1R' structure. To avoid the voltage drop on the FPGA routing, the access device, i.e., diode, are not embedded in the routing wires. Therefore, routing wires and programming wires have different metal layers. The RRAM cells could be removed from some of the cross points to achieve difference F_c parameters. If channel width is W, LB cluster size is N, LB input is I, and the flexibility of the CB is F_c, there is $W(N+I)F_c$ RRAM cells and $W(N+I)F_c+W+N+I$ diodes in one CB. To reduce the diode size, each time only one cross point in the CB is under configuration. Therefore, two word lines (L_i) are pulled to the ground, and

only one bit line (H_i) is pulled up to V_{set} or V_{reset}. For example, to program top left cross point, the two RRAM cells R_{0a} and R_{0b} are under programming. Hence, L_0 and L_1 are at the ground, and H_0 is at V_{set} or V_{reset}. With the minimized diode size, the leakage current of the diode is also minimized when the nvFPGA is in the normal operation phase. However, to reduce the wire area, we connect different H_i to the same bit line. For example, H_1 and H_3 connect to the same bit line. The detail will be discussed in Section 4.4.

The SB has the similar structure as the CB. As shown in Fig. 4.5(b), there are two RRAM cells between every two nodes. Therefore, there are 12 RRAM cells in one SP, and $12W$ RRAM cells in one SB. In the same SP, each RRAM cell pair is programmed sequentially to minimize the diode size as discussed earlier. The RRAM cells in different SPs may be programmed in parallel to reduce the FPGA configuration time.

4.3.2 Proposed Look-Up Table

We propose a novel nvLUT as shown in Fig. 4.7. Our proposed '1D2R' based LUT is using complementary structure where left side RRAM cells and their corresponding right side RRAM cells are programmed to the opposite RRAM states. For example, when the right side RRAM cells with the address '$\overline{A}\ \overline{B}$' are programmed to HRS, the left side RRAM cells with the address 'AB' will be programmed to LRS. In such configuration, the output of the LUT is 0 when the input AB is $2'b11$. The LUT in Fig. 4.7 has only 2 inputs, but it can be extended to 4, 6 and other LUT size. There are 4×2^K RRAM cells and 4×2^K diodes in a K-input LUT. Therefore, there are $2KN(N + I) + 4N \times 2^K + 4N$ RRAM cells in one LB. During the normal FPGA operation phase, the top and bottom lines are connected to VDD and ground, respectively. During the FPGA configuration phase, both top and bottom lines are connected to the word lines. Only two of

the word lines (L_0 and VDD, or L_1 and the ground) are enabled at the same time. The nodes H_i may share the same bit lines to reduce the wire area. For example, H_0 and H_1 connect to the same bit line. Besides the advantage of smaller size and leakage power reduction, the propagation delay is also greatly reduced since there is no V_{th} drop from the storage element to the output.

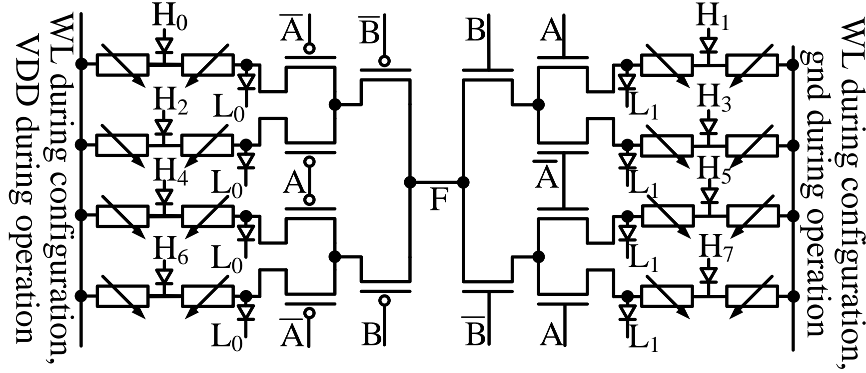


Figure 4.7: Our proposed '1D2R' based non-volatile look-up table. It is an example of a 2-input LUT, and it can be extended to the other LUT size.

4.4 Layout and Area Estimation

4.4.1 Routing of the RRAM cells proposed nvFPGA

The layout of our proposed nvFPGA will be very different from the conventional SRAM-based FPGA layout to achieve the high density. The top level floor plan of our proposed nvFPGA has been discussed in Section 4.3. In this section we provide an RRAM-friendly layout design for both SBs and CBs to fit into the footprint of the CMOS transistors below the RRAM layer.

Currently the most widely used switch box structures are Disjoint [159], Universal [160, 161], HUSB [162, 163] and Wilton [164]. Disjoint is the classical "Xilinx-style" switch block, which is also named as the subset switch block [165].

Similar to the layout in [135], the universal type SB is used for the RRAM-friendly layout design in this chapter. As shown in Fig. 4.6, two RRAM cells are placed at different SB edges. The SB flexibility F_s is set to three for the universal type SB, thus there are three rows/columns of RRAM cells at each edge of the SB. The diodes are placed above the routing metals of the SB to select RRAM cells for programming. We have to pay attention to the connection of the programming wires. As shown in Fig. 4.6, if line ① is pulled up to the write voltage, the other dashed lines should not be enabled to avoid the leakage current. In other words, all dashed lines should be connected to different bit lines. Therefore, there are at least 12 bit lines in one SB.

A fully connected (F_c=1) CB layout is shown in Fig. 4.6. Therefore, each cross point of the CB has two RRAM cells. As can be seen from Fig. 4.8, one of the RRAM cell connects to the metal in x direction, whereas the other one connects to the metal in y direction. The cross section layout of one cross point switch is shown in Fig. 4.8(a), where the metal for channel routing may be placed below the metal for connecting to the pins of the LB. Since the metals in both x and y directions are used for the word lines (L), we use a third direction for the bit lines (H) as illustrated in Fig. 4.8(b). Therefore, each time only one cross point switch is selected if two word lines (one in x direction and one in y direction) and one bit line are enabled. If we want to achieve smallest space between two bit lines, the bit lines should be alternatively routed in the different metal layers. Otherwise, their spaces should be $\sqrt{2}F$.

The area of an RRAM tile is determined by the CB channel width W, feature size F, logic cluster size N and LB inputs I. If the pitch between two channel wires is $2F$ and $F_c = 1$, the minimum area of SB and CB is $(2\sqrt{2}(W+3)F)^2$ and $W(N + I)F^2$, respectively. The SB area is only around 2/9 of the SB area that suggested in [166]. We give a space of $\sqrt{2}F$ to two channel wires, and an

Table 4.1: The number of RRAM cells and the RRAM area partition of each FPGA block.

Blocks	LB	CB	SB
RRAM Cells	$2KN(N+I)+4N\times 2^K+4N$	$2W(N+I)F_c$	$12W$
Area	$(2(2N+2I)F)^2$	$2\sqrt{2}(W+3)F\times 2(2N+2I)F$	$(2\sqrt{2}(W+3)F)^2$

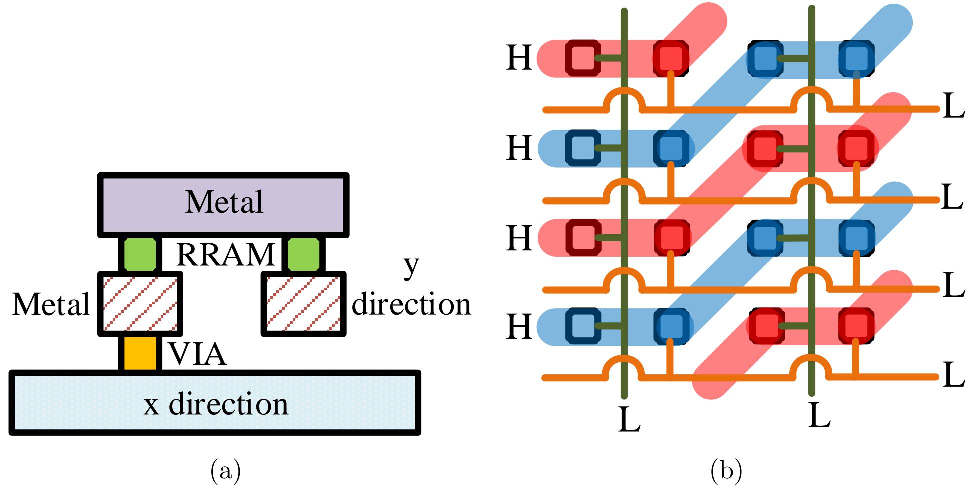


Figure 4.8: (a) The cross-section view of the switch in CB; (b) our proposed crossbar routing architecture to program the RRAM cells.

area for the local interconnect and RRAM cells in BLEs to $(4(N+I)F)^2$. Thus the total area of the RRAM layer and its related routing in our proposed '1D2R' based FPGA tile is $(2(\sqrt{2}(W+3)+2N+2I)F)^2$. The required area and RRAM cells of each FPGA block is tabulated in Table 4.1.

4.4.2 Area Estimation

To compare the relative merits of our proposed '1D2R' based FPGA scheme, and the CMOS-based FPGA scheme, we perform area calculations with a LUT input size K=4, logic cluster size N=10, LB inputs I=22, a fixed routing channel width W=100 and F_c = 0.5. Area breakdown of different components in an FPGA

is based on the architectural model in [42]. The method in [166] was used to estimate the tile area. For the above parameters, we estimate the footprint of a baseline CMOS FPGA tile to be $20149T$. Using a minimum width transistor area of $T = 0.09\mu m^2$ for a $45nm$ transistor [166] gives us a SRAM-based FPGA tile area of $1813.4\mu m^2$. The detailed area of one baseline tile can be partitioned as shown in Fig. 4.9, where the switch and SRAM in the CB and SB occupy around 68% of the total tile area.

By stacking RRAM cells and diodes on the top of the CMOS circuitries, the area of the tile is greatly reduced. Since the complementary LUT structure is used, the input buffer size of the LUT is doubled. Therefore, there are 162 minimum width transistors in one LUT. Moreover, minimum size buffers are used in the interconnect. Hence, the CMOS area of the proposed '1D2R' based nvFPGA tile is 4509 minimum width transistors ($20.14\mu m \times 20.14\mu m$). In contrast, the area of our proposed '1D2R' based FPGA RRAM layer is only $18.87\mu m \times 18.87\mu m$, which is smaller than the CMOS area. The detailed area breakdown of our proposed nvFPGA tile can be partitioned as shown in Fig. 4.9. The percentage of the interconnect and SRAM area reduces from 90.84% in the SRAM-based FPGA tile to 41.16% in our proposed '1D2R' based FPGA tile. The total area of LB switch, LB SRAM, CB switch, CB SRAM, SB switch and SB SRAM occupy 67.85% area in the SRAM-based FPGA tile. The tile area is reduced from $1813.4\mu m^2$ to $405.81\mu m^2$ ($4.47\times$ area reduction).

4.5 Simulation Results

In this section, we first evaluate the write reliability of both diode-less crossbar array and diode-based crossbar array. After that, we provide the spice simulation results based on the schematic in Fig. 4.4, and the LUT performance comparison. Finally, the speed and power of three FPGA schemes are evaluated by the Versatile

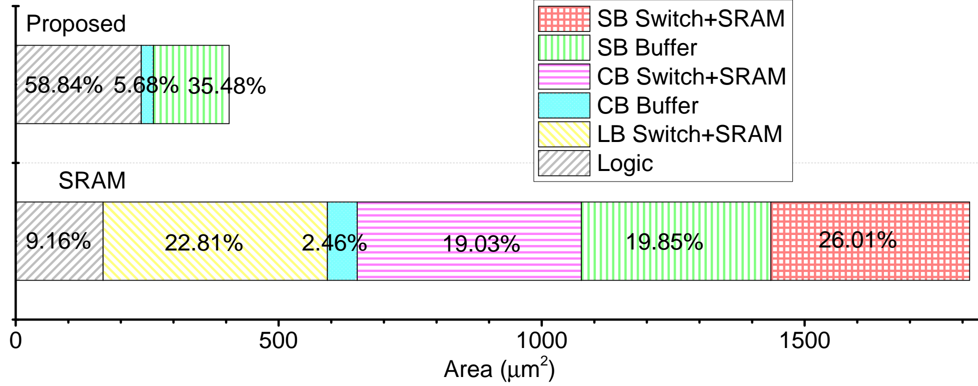


Figure 4.9: Area consumptions of the SRAM-based FPGA tile and our proposed '1D2R' based FPGA tile. The switch and SRAM area in our proposed '1D2R' based scheme is negligible because they are placed on top of the CMOS circuits.

Place and Route (VPR) software [167], and the power model provided in [12, 13].

The RRAM parameters are extracted from the measurement results of the RRAM cells fabricated by the process in [123]. Its low resistance (R_L) and high resistance (R_H) are $10^3\Omega$ and $10^9\Omega$, respectively.

4.5.1 Write Power and Reliability

As shown in Fig. 4.10, a spice model with parasitic resistors in both bit lines (H) and word lines (L) is used to simulate the write voltage distribution, write power and write error rate. In this simulation, copper is used for the bit lines and word lines, and the thickness of the metal is four times of the width of the metal. Therefore, the square sheet resistance is about 0.1Ω and the parasitic resistance between two adjacent cells with $2F$ pitch is 0.2Ω. All unselected RRAM cells are set to LRS (worst case of the leakage current) in this simulation.

It can be seen from Fig. 4.11(a), the write voltage on the selected cell with the $V/2$, $V/3$ and floating schemes drop to 25% when M=128 due to the sneak path leakage current. The diode-based scheme has less than 3% voltage drop on

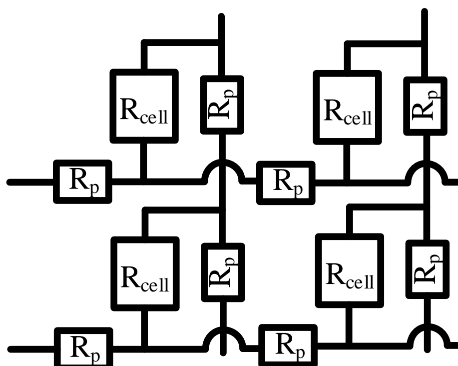


Figure 4.10: A simulation diagram of the diode-less or transistor free crossbar array with parasitic resistance (R_p) in the word lines and bit lines.

the selected RRAM cell, since the leakage current is almost isolated by the 'off' state diodes. The small voltage drop is mainly caused by the IR drop in the H lines and L lines. In the $V/2$, $V/3$ and floating schemes, if all unselected RRAM cells are at HRS, the normalized write voltage on the selected cell is closed to 1. As a result, the write voltage on the selected cell has a very wide distribution $(0.25 - 1)$. Increasing the input driven voltage to improve the write voltage on the selected cell may lead to much higher write energy, breakdown risk and write disturbance in the unselected cells.

To switch a cell, the normalized input write driven current at the selected bit line is shown in Fig. 4.11(b). When $M > 100$, the three diode-less schemes draw more than 100 times more current (caused by the sneak path leakage current) than that of the diode-based scheme. The diode-based scheme has a constant current requirement versus M. Since the write current to switch an RRAM cell is fixed, the total current of the diode-less array will be extremely large. The high write current not only increases the write power, but also requires a large area of the write drivers and wires.

As shown in Fig. 4.11(c), the diode-less schemes spend a very large portion of the write current on the unselected cells. The $V/3$ scheme is even worse since

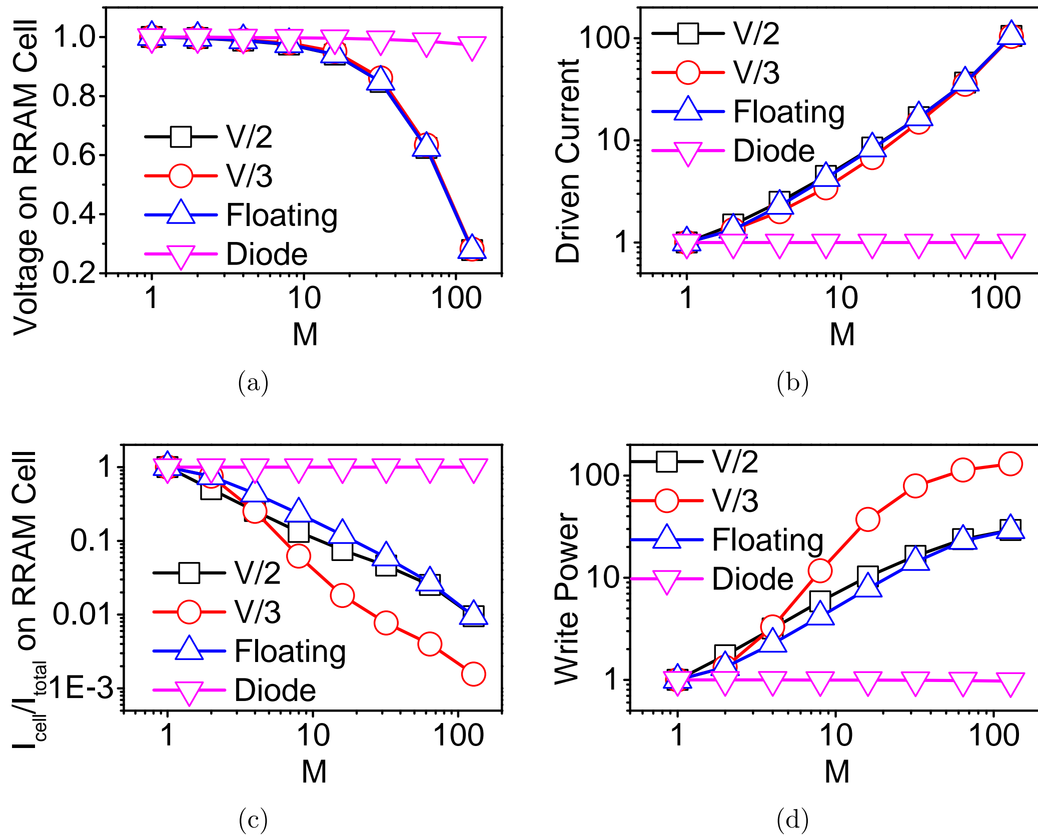


Figure 4.11: (a) The normalized write voltage across the selected RRAM cell; (b) the normalized required current at the input driver of the bit line or word line; (c) the write current analysis of different RRAM array schemes; (d) the normalized total write power. All results are normalized to the one single RRAM cell.

all unselected cells are biased at one third of the write voltage. In comparison, the write current almost all goes to the selected RRAM cell in the diode-based scheme. Fig. 4.11(d) provides the total power consumption with a fixed input write voltage at the bit line. The results show that the write power of the diode-based scheme is constant versus array size. However, the write power is linearly increased in the $V/2$ and floating schemes, and exponentially increased in the $V/3$ scheme.

The diode-less scheme not only requires large area and high write power, but also has an extremely low write reliability. We choose 64×64 array with $V/2$

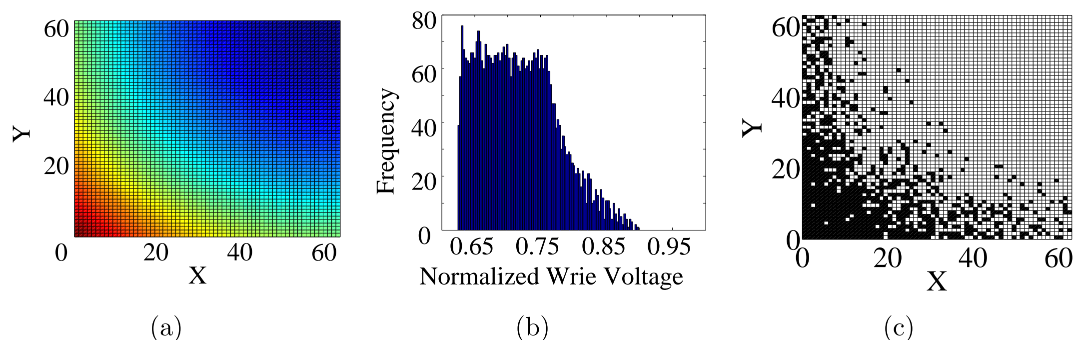


Figure 4.12: (a) The write voltage distribution in a 64 × 64 diode-less crossbar RRAM array due to the parasitic resistance in the word lines and bit lines; (b) the histogram plot of the normalized write voltage distribution in a 64 × 64 diode-less crossbar RRAM array; (c) the programming results in the 64 × 64 diode-less crossbar RRAM array. Black color represents successfully programmed cells and white color represents unprogrammed cells.

write scheme as the baseline to evaluate the write reliability. All unselected RRAM cells are still set to LRS. As shown in Fig. 4.12(a), the voltage drop gets worse from bottom left to top right, since the write drivers are located at the left side and bottom side of the array. Longer metal lines result in much lower voltage across the selected cell. The histogram of Fig. 4.12(a) is illustrated in Fig. 4.12(b). The normalized write voltage across the selected RRAM cell is spread between 0.6 and 1. Most of the voltage on the selected RRAM cells falls into the 0.65−0.75 range. If the unselected RRAM cells have random resistance states, the distribution will be even worse. The write error map is shown in Fig. 4.12(c). Whether an RRAM cell can be successfully programmed is quite randomly in the bottom left region. In the top right region, all RRAM cells are failed to be programmed.

The write error rate is shown in Fig. 4.13. In this simulation, the required switching voltage has a normal distribution with a standard deviation of 5%. The input driven voltage is properly chosen to ensure very low write error rate for the single cell, and very low write disturbance when half biased. For example, since most of the write voltage on the selected RRAM cells falls into the 0.65−0.75

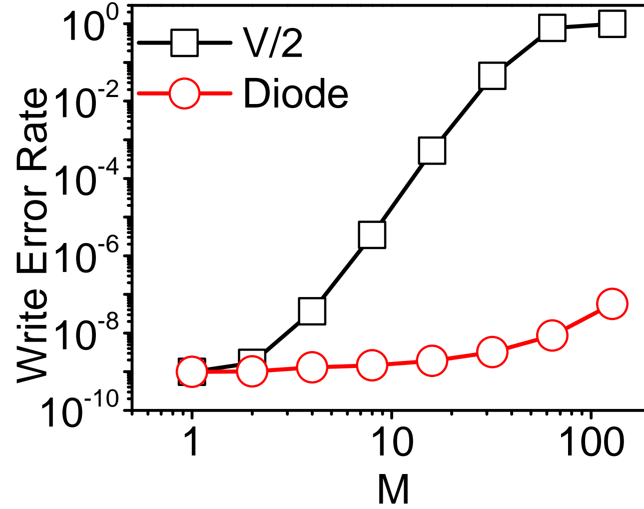


Figure 4.13: The write error rate comparison between $V/2$ write scheme and the scheme using diode as the selector.

Table 4.2: The simulation results of the RC delay among our proposed scheme, the conventional '1R' and SRAM schemes.

Delay (ps)	A→B	B→C	C→D	D→E	E→D	E→out	A→out
Proposed	52.42	41	35.97	145.855	37.95	27.24	302.485
1R	51.645	39.32	34.015	140.165	36.97	25.33	290.475
SRAM	159.525	209.81	164.8	208.33	50.265	33.935	776.395

range, the input driven voltage is set to 1.3× of the mean switching voltage in the 64×64 array. In a small array size, i.e., 2×2, all write schemes have very small write error rate. However, in a larger RRAM array, the diode-less scheme ($V/2$) has a much higher write error rate than the diode-based scheme. Based on a 64×64 array, the write error rate of the diode-less scheme and diode-based scheme are 0.784 and $8.6e{-}9$, respectively. Such high write error rate of the conventional '1R' scheme will make the FPGA function incorrectly.

4.5.2 RC Delay Simulation Results

The RC delay is simulated based on the schematic in Fig. 4.4. One path is enabled from the input of SB (A) to the output of LB (out). RC model is inserted at each node, i.e., an RC delay of the metal in SB, CB, local interconnect, etc. The parasitic resistance and capacitance are estimated based on the area evaluation results in Section 4.4. The space and width of the wires between two channels are set to equal value. The estimated capacitance in the SB, CB, CB to LB and the local interconnect are $2.65fF$, $1.15fF$, $1.2fF$ and $1.15fF$, respectively. The RC delay simulation results will be used in the VPR simulation.

The RC delay simulation results are tabulated in Table 4.2. We assume all RRAM cells are successfully programmed in the '1R' based FPGA. The simulation results show that our proposed scheme has a penalty of only 4% lower speed than the '1R' scheme. The improvement is significant when compared to the SRAM-based scheme. There are four times and two times speed improvement in the interconnect and LB, respectively. The total speed improvement from A to out is around 2.5 times. In the SRAM-based scheme, the delay is mainly caused by the routing, which is 68.8% of the total delay. In contrast, the delay caused by the routing is reduced to 42.8% of the total delay. The improvement of the delay is due to the much shorter routing length and no V_{th} drop on the routing path. The shorter routing reduces parasitic resistance and capacitance, thus reduces both delay and dynamic power.

4.5.3 LUT Comparison

We further evaluate the area, speed and power of our proposed LUT, the '1R' based LUT and the SRAM-based LUT. The '1R' scheme is using the same LUT structure as shown in Fig. 4.7 but replacing all '1D2R' with '1R'. The simulation results are summarized in Table 4.3.

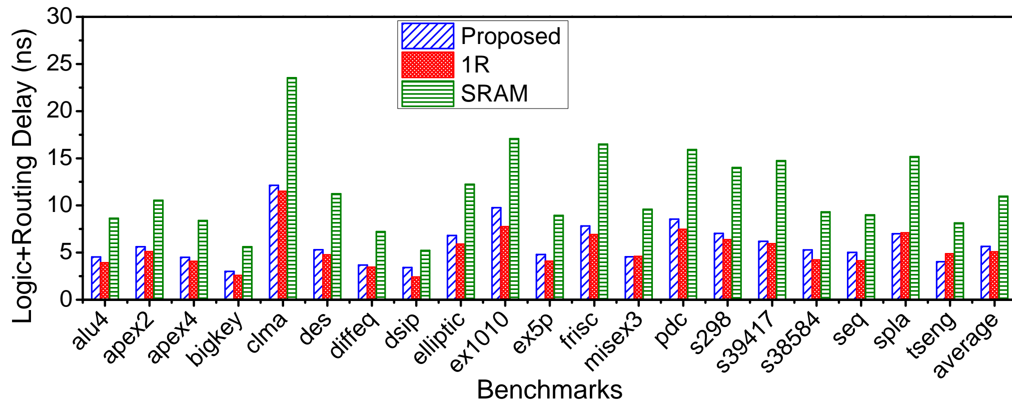
Table 4.3: The speed, power and area comparison among different LUT schemes.

Schemes	Delay	Dynamic Power	Leakage Power	Number of Transistors
Proposed	145.855ps	4.71fJ	2.53nJ	162
1R	140.165ps	4.76fJ	2.861nJ	162
SRAM	208.33ps	6.533fJ	5.61nJ	172

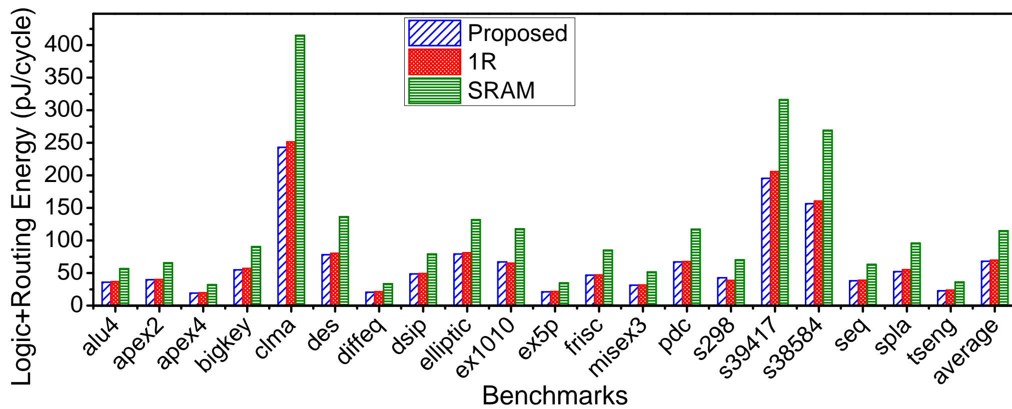
Compared to the SRAM-based LUT, our proposed LUT improves the speed, dynamic power and leakage power by 30%, 28% and 55%, respectively. The speed is improved mainly due to no V_{th} drop in the LUT. The dynamic power is improved due to much narrower short circuit current from VDD to the ground. Because the SRAM-based LUT requires a feedback transistor to pull the output of the multiplexer to VDD. This feedback transistor will fight with the SRAM or the SRAM buffer. The leakage power is improved by replacing the SRAM cells with RRAM cells. Moreover, our proposed scheme also reduces 12% leakage power from the '1R' based scheme, since our proposed scheme has doubled the off-state resistance. The delay of our proposed scheme is slightly higher than the '1R' based scheme, which is due to the on-state resistance is also doubled. The area of '1R' and '1D2R' based LUTs reduces 6% from that of the SRAM-based LUT.

4.5.4 VPR Simulation Results

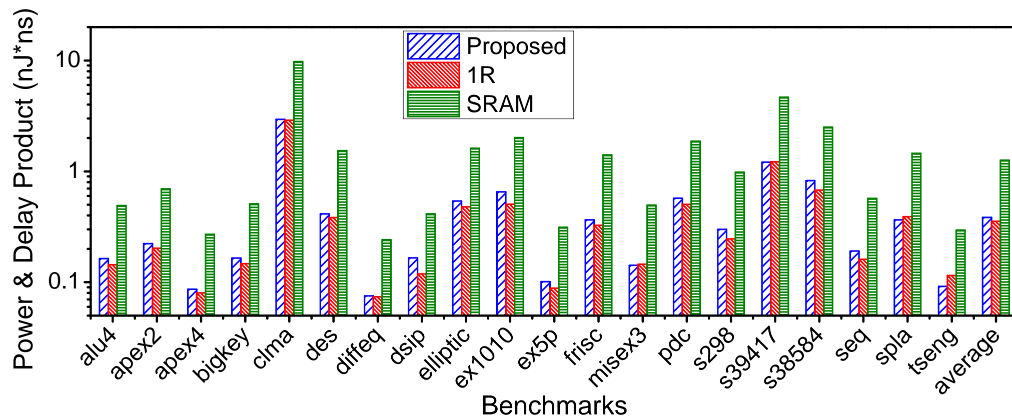
Evaluating our proposed '1D2R' based FPGA scheme is assisted by the VPR software, which is very flexible to compare the newly developed FPGA architecture and many other different FPGA architectures. It provides a behavioral system analysis on different FPGA architectures. We also use the gate-level FPGA power estimator [12,13] to evaluate the power consumption of the proposed '1D2R' based FPGA. The FPGAs used in the VPR simulations are based on the architectures provided in Section 4.4. The RC delays required by the VPR have been evaluated in Section 4.5.2.



(a)



(b)



(c)

Figure 4.14: (a) The delay simulation results; (b) the power simulation results; (c) the power and delay product results. The three schemes are simulated based on 20 MCNC test benches with VPR and the power model in [12, 13].

Fig. 4.14 shows the power and delay simulation results based on 20 Microelectronics Center of North Carolina (MCNC) benchmarks. MCNC benchmark suite is very popular in academic research, and has standardized libraries with representative circuit designs ranging from simple circuits to advanced circuits obtained from industry. Compared to the SRAM-based FPGA, the speed of our proposed '1D2R' based FPGA improves from 1.53× in the 'dsip' benchmark to the 2.38× in the 's39417' benchmark as shown in Fig. 4.14(a). The averaged speed is improved by 1.94 times. As shown in Fig. 4.14(b), the power of our proposed '1D2R' based FPGA reduces from 36.9% in the 'alu4' benchmark to the 45.5% in the 'spla' benchmark. The average power reduction is about 40.9%. As a result, the average power-delay product (PDP) is improved by 3.3 times as shown in Fig. 4.14(c). The delay and dynamic power are greatly reduced due to the much shorter routing length and the improved LUT architecture. Though the switch resistance of our '1D2R' scheme is doubled from the '1R' scheme, there is only 10% downgrade in the speed performance, and 8% of the PDP.

4.6 Summary

In this chapter, we have proposed a '1D2R' based non-volatile storage element, and '1D2R' based nvFPGA architecture. Compared to the SRAM-based FPGA, our proposed '1D2R' scheme has greatly reduced the area and power by 78% and 40.9%, and improved the speed by 1.94 times. Compared to the conventional '1R' based nvFPGA, it has significantly enhanced the write reliability with only 8% performance reduction. The results have shown that the write error rate is as low as $8.6e{-}9$ in a 64×64 crossbar array. The results suggest that our proposed '1D2R' based scheme is a promising solution to achieve low power, high speed and high reliability FPGAs.

Chapter 5

Non-volatile SRAM-based FPGA

The chapter is written mainly based on the paper "A Low Active Leakage and High Reliability Phase Change Memory (PCM) based Non-Volatile FPGA Storage Element".

5.1 Introduction

A few works have been reported to integrate NVM cells into FPGA circuits in [2, 3, 135, 136, 168]. However, those works have various drawbacks that limit their applications in FPGAs. For example, the designs in [135, 136] have a write reliability issue due to sneak paths. [168] in essence is the SRAM-based FPGA. Therefore, it still suffers from long configuration time and high configuration power when powering on. [2] and [3] suffer from high active leakage power (the leakage power during normal operation) and low reliability issues due to high DC voltage (VDD) on NVM cells during the FPGA normal operation. The design in Chapter 4 requires special process of the diode and RRAM cells. High resistance ratio of the RRAM is indispensable to achieve high reliability and low leakage. Therefore, the cost of the nvFPGAs is greatly increased. Moreover, the design cannot be

used in the multi-context FPGAs.

In this chapter, we propose a low active leakage power and high reliability nvSRAM storage element with high loading speed. PCM is used in our nvSRAM, but it is worth noting that our nvSRAM cell can be extended to all resistive NVMs. The process is greatly simplified, thus the cost will be highly reduced. To achieve the low active leakage power and high reliability, PCM cells are only sensed when powering on. In the FPGA operation mode, they are biased at 0V by pulling both nodes of PCM cells to the ground. Therefore, there is no active leakage power in PCM cells, and the retention time can be greatly improved. As a result, our proposed nvSRAM is able to load configuration information within 1ns, achieving fast multi-context switching abilities, and 41.8 pW low active leakage power during FPGA operation. The retention can be longer than 10 years. The FPGA system loading speed and energy are 1ns and $2.54 fJ/cell$, respectively.

The design in Chapter 4 relies on the resistance of the RRAM cells to configure FPGA. Since the high and low resistance of the RRAM has only 6 orders difference, and the resistance value of the RRAM has a much wider distribution than CMOS. Therefore, the variation of the resistance value will significantly affect the active leakage current, timing uncertainty, etc. The NVMs in the proposed nvSRAM is only sensed during power-on period. In other modes, they are turn off. Therefore, the process variation of the NVM will not affect the performance of the FPGA during normal operation.

5.2 Proposed nvSRAM based FPGA

The proposed nvSRAM based FPGA, as shown in Fig. 5.1, has the similar architecture as conventional SRAM-based FPGAs. The only difference is that 6T SRAMs are replaced by PCM based nvSRAMs to configure FPGAs.

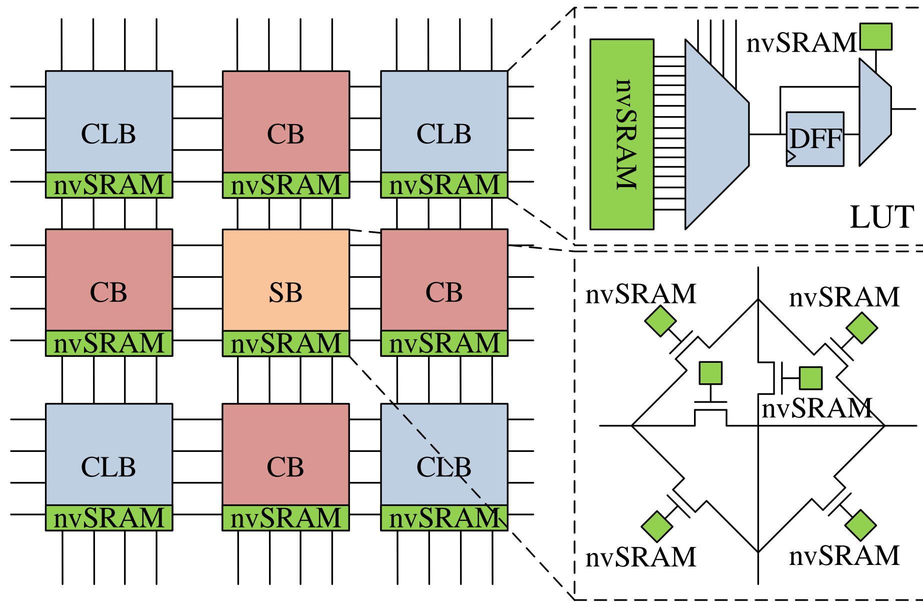


Figure 5.1: The proposed nvSRAM based FPGA Architecture. 6T SRAMs are replaced by our proposed nvSRAMs. SB, CB and CLB are switch block, connection block and configurable logic block, respectively.

5.2.1 Working Modes and Power Advantage

In the proposed nvSRAM based FPGA, we introduced a loading mode in addition to the traditional sleep mode, configuration mode and normal operation mode. The configuration mode and loading mode of the proposed nvSRAM based FPGA are used to write configuration information to PCM cells, and read configuration information from PCM cells to latches, respectively.

The nvSRAM based FPGAs are only programmed once in the configuration mode. Thereafter, the information stored in PCM cells is sensed in the loading mode to configure the logic and routing in FPGAs. There is only one time loading when FPGAs are powered on. The instant power-on and non-volatile abilities of nvSRAMs reduce the sleep power, power-on time and power-on energy, allowing FPGAs to be powered on/off more frequently to reduce the power consumption.

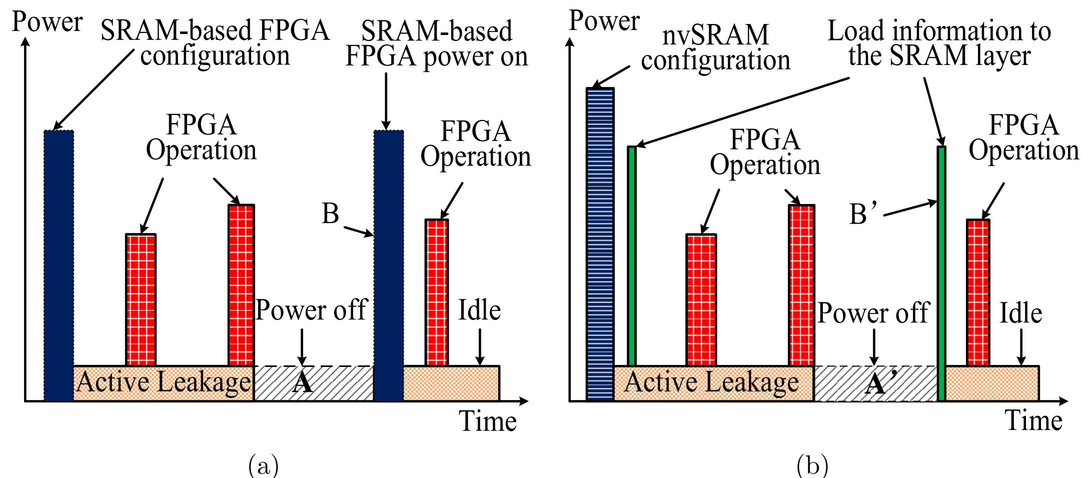


Figure 5.2: The power consumption of the (a) SRAM-based FPGA and (b) our proposed nvSRAM-based FPGA in different operation modes.

Fig. 5.2 explains the power consumption of conventional SRAM-based FPGAs and our nvSRAM-based FPGAs in different modes. As shown in Fig. 5.2(a), SRAM-based FPGAs have high configuration power and long configuration time. Therefore, SRAM-based FPGAs require significant overhead during power on and off. BEP, which is defined by the time when the reduced sleep energy (area A) equals to the energy required to power on the FPGA (area B), can be used to evaluate power-off possibilities. In other words, only when area A is larger than area B, SRAM-based FPGAs benefit from in powering off in terms of power. Another power off condition is that the sleep time between two events has to be longer than the total width of A and B. As shown in Fig. 5.2(b), the smaller area B' of our nvSRAM based FPGA allows area A' to be much smaller to gain power reduction benefit. Therefore, the width of A' is much shorter than that of A, and the width of B' is also much shorter than that of B due to instant power on ability. In other words, our nvSRAM-based FPGAs can be powered off to reduce the FPGA power consumption in a much shorter idle period.

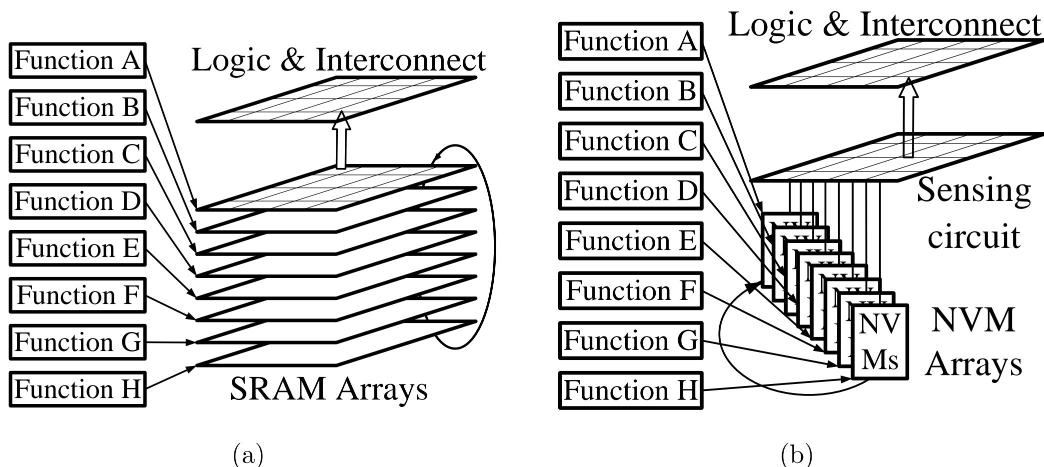


Figure 5.3: (a) Conventional SRAM-based multi-context FPGA; (b) Proposed nvSRAM based multi-context FPGA.

5.2.2 Multi-context FPGA and Area Advantage

One solution to reduce the chip area and power consumption is through run-time reconfiguration (RTR) by increasing the hardware utilization [169]. RTR is the ability to modify or change the functional configuration of the device during operation. It can reduce the hardware components (area) and power consumption by reusing the same FPGA for several functions. As it involves reconfiguration during program execution, fast configuration is very important for RTR. However, the traditional single-context FPGA structure only allows one full-chip configuration to be loaded at a time results in very slow reconfiguration. Therefore, SRAM-based multi-context FPGA has been proposed [170]. A key advantage of the multi-context FPGA over a single-context architecture is that it allows the nanoseconds context switch, whereas the single-context may take milliseconds or more to be reprogrammed [170].

However, due to the volatile nature of the SRAM, SRAM-based multi-context FPGAs still suffer from several fundamental drawbacks, including long configuration loading time (need to reload the configuration from the external

NVM array every time when powering on), excessive active leakage power (have to always power on all context layers), large configuration memory area (large size of SRAM), low standby possibility and etc.

We propose using NVMs to replace SRAMs to form an NVM-based multi-context FPGA. The NVMs are used to store the FPGA configuration information. Fig. 5.3(a) illustrates the N-layer multi-context architecture for conventional SRAM-based multi-context FPGAs. N is set to 8 in this example for illustration, but not limited to 8. It can be seen that there are eight context layers of SRAMs. Each SRAM layer contains the configuration information for a different function. Based on the application, different SRAM layer is selected. The switching among these configuration layers can be achieved during execution. The multiple configuration layers can be combined to emulate a single large function. Fig. 5.3(b) shows the proposed nvSRAM based multi-context FPGA. The main difference is that the eight SRAM layers are replaced by eight NVM layers. Each NVM layer contains different function. It has the same operation scheme as the conventional SRAM-based one. A shared sensing circuit is designed to control the NVM layers. Because the cell size of NVM is only about 3% of that of SRAM [1], the chip area of FPGA could thus be significantly reduced.

5.3 Proposed Storage Element

To reduce the active leakage power and increase the reliability, we follow three design principles. The first principle is to bias PCM cells at $0V$ during the FPGA normal operation. Hence there is no active leakage current on PCM cells, and their states will not be disturbed. The second principle is to quickly load the configuration information from PCM cells to latches with low read power, thus allows the FPGA to be powered on/off more frequently, and switch between contexts much faster. The last principle is to remove the high voltage inside the nvSRAM

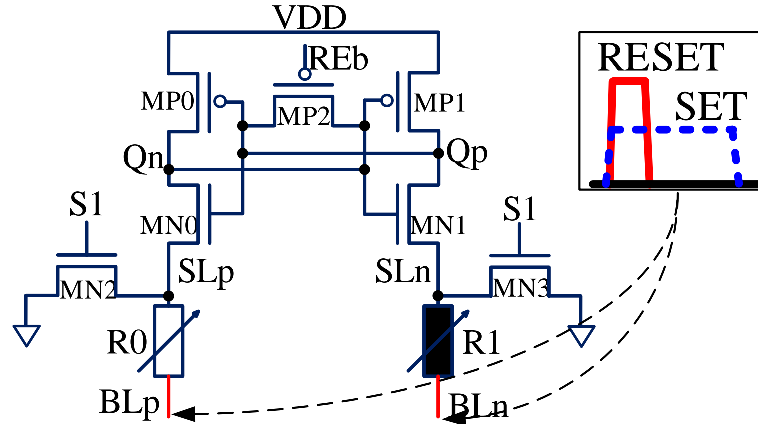


Figure 5.4: The proposed single-context nvSRAM. The signals BL_p and BL_n are shared with other nvSRAMs in the same column.

during PCM cell programming, thus low VDD devices can be used to achieve high density. With these principles, we propose both single-context nvSRAM and multi-context nvSRAM in the following.

5.3.1 Single Context nvSRAM

The proposed PCM based single-context nvSRAM storage element is shown in Fig. 5.4. As discussed in Section 5.2, our proposed nvSRAM has three modes besides the sleep mode, the detailed description of each mode is provided as follows:

a). In the configuration (write) mode, read enable signal (REb) is high to turn off the equalization transistor MP_2, thus the four transistors (MP_0, MP_1, MN_0 and MN_1) formed latch isolates FPGA operation supply voltage (VDD) from nodes SL_p and SL_n. This results in no DC path between VDD and the write voltages (V_{set} and V_{reset}) of the PCM cells. Meanwhile, the control signal S_1 is high to pull nodes SL_p and SL_n to the ground. The nodes BL_p and BL_n are driven by the SET voltage (V_{set}) and RESET voltage (V_{reset}) pulses according to the configuration information. For example, if the configuration information is “0”, R_0 and R_1 are under RESET and SET operations, respectively. It is worth noting

that the high write voltage is not connected to SL_p or SL_n as reported in [171]. This avoids the use of thick oxide transistors in the latch. After configuration, R_0 is at high resistance state (R_H), and R_1 is at low resistance state (R_L). The simplified schematic of the proposed nvSRAM to write the PCM cells is shown in Fig. 5.5(a).

b). In the loading (read) mode, as shown in Fig. 5.5(c), BL_p and BL_n are pulled to the ground, and S_1 is low to disconnect SL_p and SL_n from the ground. Meanwhile, REb is also low to equalize SL_p and SL_n to $VDD - V_{thp} - V_{thn}$, where V_{thp} and V_{thn} are the threshold voltages of PMOS and NMOS transistors, respectively. Due to pre-configured information on R_0 and R_1, the nvSRAM forms two asymmetric current paths. For example, when $R_0 = R_H, R_1 = R_L$, the current on R_1 is much larger than that on R_0. Therefore, the output node Q_p is pulled down, thus pulls up Q_n. The asymmetry of current paths forms a third current path in MP_2 from Q_n to Q_p. Once REb is high, the latch pulls Q_n to VDD and Q_p to the ground.

c). In the FPGA normal operation mode, BL_p and BL_n are still at the ground, and REb is high. Moreover, S_1 is turned on to pull SL_p and SL_n to the ground and thus bias PCM cells at 0V, resulting in zero active leakage power and long retention time. The nvSRAM works like a convectional SRAM to configure the logic and routing in the FPGA. Fig. 5.5(d) shows the simplified SRAM-like schematic of the nvSRAM during the FPGA normal operation mode.

The control logic information of our proposed nvSRAM in different operation modes is tabulated in Table 5.1. The proposed nvSRAM contains 7 transistors, one more than the conventional 6T SRAM. During writing, the drain of transistors MN_2 and MN_3 are pulled to the ground, and the high write voltage is isolated by the PCM cells. As a result, thin oxide transistors can be used in the nvSRAM, leading to significant reduction in nvSRAM size.

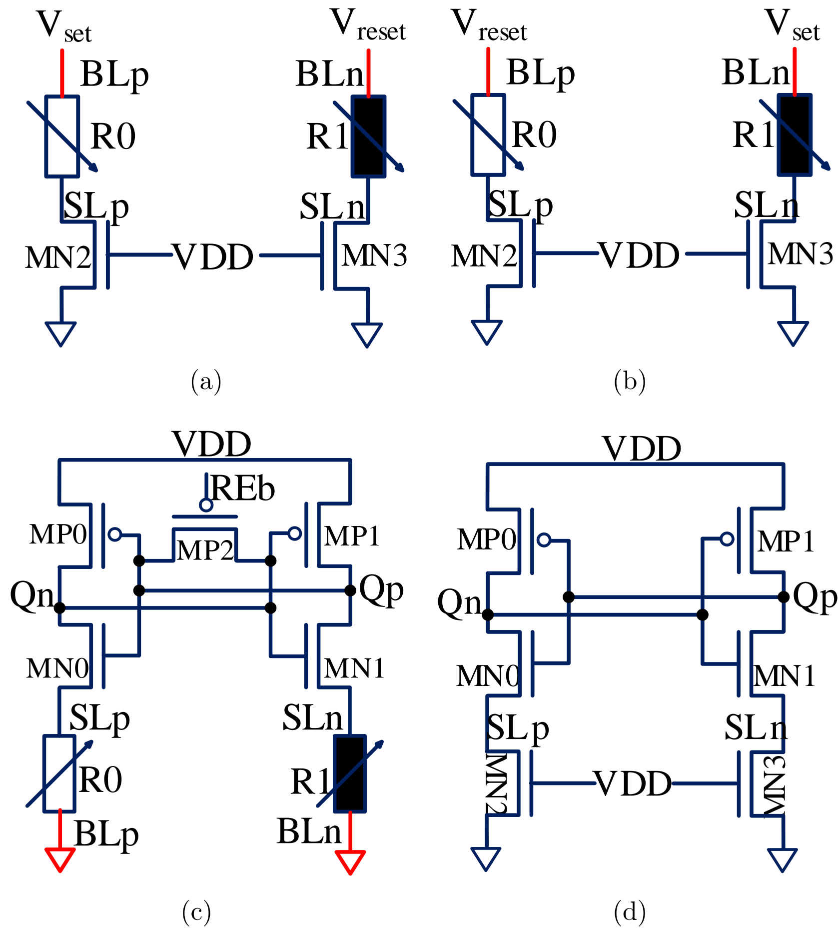


Figure 5.5: The proposed single context in the (a) write mode, (b) read mode, and (d) FPGA execution mode.

5.3.2 Multi-context nvSRAM

We further propose an nvSRAM with multiple layers of programming bits (multi-context nvSRAM), where each layer can be activated at a different time point. Our proposed multi-context nvSRAM shows a great potential in run-time reconfiguration applications, since it only needs less than 1ns to switch between different contexts.

The proposed multi-context nvSRAM, as shown in Fig. 5.6, not only has the non-volatile and instant power-on advantages, but also helps to reduce the

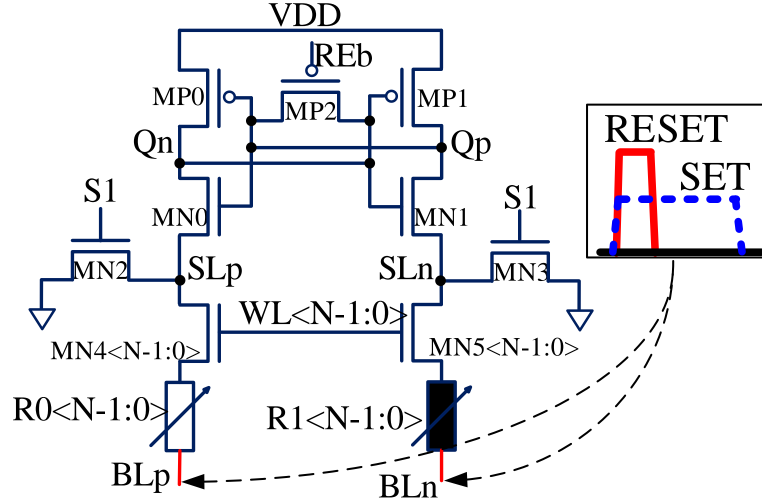


Figure 5.6: The proposed multi-context nvSRAM. The signals BL_p and BL_n are shared with other nvSRAMs in the same column

Table 5.1: The control logic information of our proposed nvSRAM in different operation modes.

Modes	REb	S_1	BL_p	BL_n
Write (1)	1	1	V_{set}	V_{reset}
Write (0)	1	1	V_{reset}	V_{set}
Read	Negative Pulse	0	0	0
Normal operation	1	1	0	0

area by sharing the latch. Compared to the SRAM-based multi-context FPGA, the area, standby power, power-on time and power-on energy could be significantly reduced. In Fig. 5.6, the context select transistor pairs $MN_4<N-1:0>$ and $MN_5<N-1:0>$ are inserted between the latch and PCM cells. The context select transistors are controlled by the context select address $WL<N-1:0>$. The N-context requires N bits context selected address, N pairs of select transistors and N pairs of PCM cells.

The multi-context nvSRAM has four operation modes in addition to the sleep mode: the configuration mode, the loading mode, the multi-context switch mode and the FPGA normal operation mode. These modes are similar to the

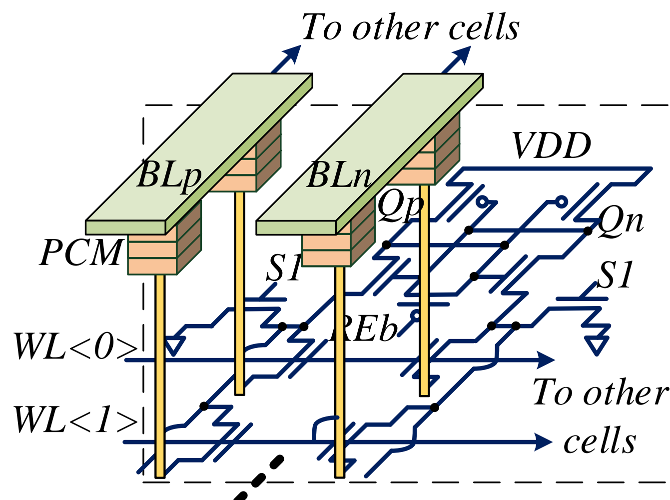


Figure 5.7: A schematic of the nvSRAM 3D integration. The phase change material is deposited in the format of thin-film on the top of the CMOS transistors.

single-context nvSRAM except the context switch mode. The context switching mode is for run-time reconfiguration, which performs almost the same as the read operation. The only difference is that it first changes the context address to the targeted layer before sensing the configuration information from the selected layer to the latch.

A 3D integration schematic of the CMOS circuits and PCM cells is shown in Fig. 5.7. The phase change material is deposited in the format of thin-film on the top of the CMOS circuits, thus no additional area is required for PCM cells. The latch is shared by different context layers, resulting smaller area of the multi-context nvSRAM than the multi-context SRAM. Fig. 5.7 shows an example of 2-context nvSRAM, where all PCM cells are placed in the same layer.

The multi-context nvSRAM also allows dynamic reconfiguration during the FPGA normal operation when required logic function is not pre-configured in PCM cells. The FPGA operation is not interrupted when writing new information to the PCM cells. During dynamic reconfiguration, S_1 is high to pull the nodes SL_p and SL_n to the ground. Therefore, the configuration information is still latched

Table 5.2: The parameters of the PCM used in the simulation.

PCM	Parameter
Technology node	$20nm$
SET/RESET pulse width	$200ns/20ns$
SET/RESET voltage	$1.2V/1.7V$
SET/RESET current	$60\mu A/100\mu A$
Low/High Resistance	$20K\Omega/2M\Omega$

by MP_0, MP_1 and MN_0 to MN_3. Then a normal write operation is performed to the selected PCM cells. The new states of the PCM cells could be sensed at any time when required by the FPGA systems. The FPGA systems are interrupted in a very short time period since the sensing speed is less than $1ns$.

5.4 Simulation Results

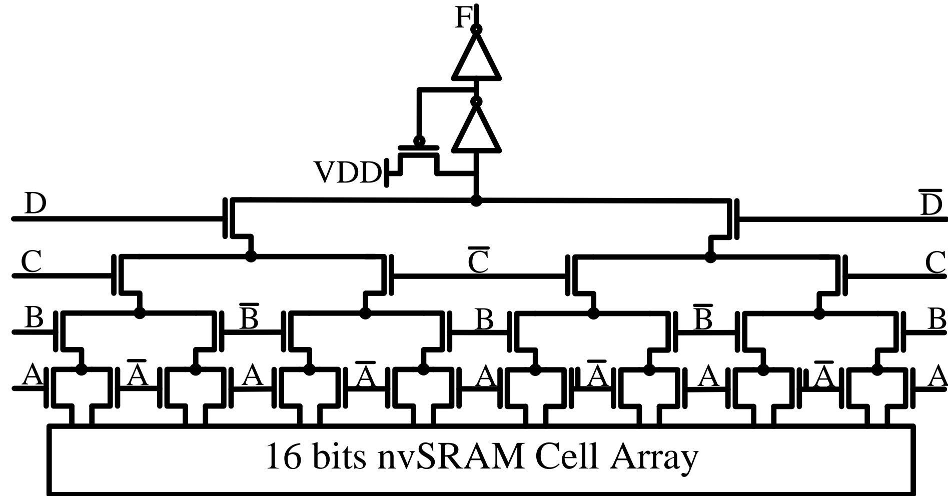


Figure 5.8: The 4-input LUT structure used to evaluate the proposed nvSRAM.

In this section, we first evaluate the power and delay performance of the proposed single-context nvSRAM based 4-input LUT, and another three 4-input LUT architectures. After that, we analyze the retention of PCM cells to be inte-

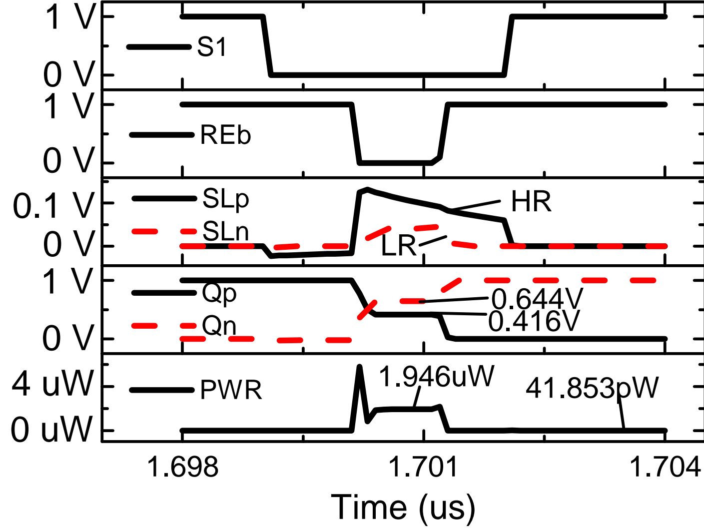


Figure 5.9: The power and delay simulation results of the proposed nvSRAM when loading the states from PCM cells to the latch.

grated in three different schemes. In the second part of this section, we compare the power, delay, loading energy and area among these four multi-context 4-input LUTs.

To evaluate the proposed nvSRAM, test benches were built based on a $45nm$ CMOS process node. GST based PCM is used in our simulation. The model is built by Verilog-A using curve fitting. Our PCM model uses the same resistance value and pulse width as [2]. The high resistance (R_H) and the low resistance (R_L) are $2M\Omega$ and $20K\Omega$, respectively. The SET and $RESET$ pulse widths of the PCM model are $200ns$ and $20ns$, respectively. Our default SET and $RESET$ voltages are $1.2V$ and $1.7V$, respectively. The detailed PCM parameters are tabulated in Table 5.2. We built a read disturbance model according to the data provided by [172] to compare the data retention.

Table 5.3: The results comparison among the SRAM, proposed nvSRAM, [2] and [3].

	This work	[2]	[3]	SRAM
Non-volatile	Yes	Yes	Yes	No
4-input LUT Active Leakage Power	$1.19nW$	$207nW$	$2.15\mu W$	$1.17nW$
4-input LUT Switching Energy	$2.58fJ$	$3fJ$	$2.2fJ$	$2.5fJ$
4-input LUT Pull-down Delay	$280ps$	$310ps$	$316ps$	$270ps$
4-input LUT Pull-up Delay	$250ps$	$220ps$	$186ps$	$220ps$
FPGA Power-on Speed	<1ns (∼300ps)	$90ps$	$90ps$	milliseconds
FPGA Power-on Energy	$2.54fJ/bit$	$2.16fJ/bit$	$3.07fJ/bit$	∼50fJ/bit [173]
Data Retention	>10 years	$250\mu s$	$250\mu s$	Preserved so long as voltage is applied

5.4.1 Single Context Simulation Results

The power and delay simulation results given in Fig. 5.9 shows that our proposed nvSRAM achieves a 41.8pW low active leakage power and a within $1ns$ high sensing speed. The low active leakage power is due to zero bias voltage on PCM cells by pulling SL_p and SL_n to the ground. The reading power of nvSRAM cell is only around $1.95uW$, hence the time and energy consumed by reading are shorter and lower than configuration of the SRAM cell when FPGAs are powered on.

A 4-input LUT in Fig. 5.8 is used to evaluate the performance of the four LUTs based on the proposed nvSRAM, SRAM, and those in [2] and [3]. The LUT in [2] is extended to the same four inputs. The SRAM based LUTs use the same structure as in Fig. 5.8 by replacing nvSRAM cells with 6T SRAMs. The resistance of the pull-down resistor in [3] is set to the logarithmic middle point of R_H and R_L ($200K\Omega$).

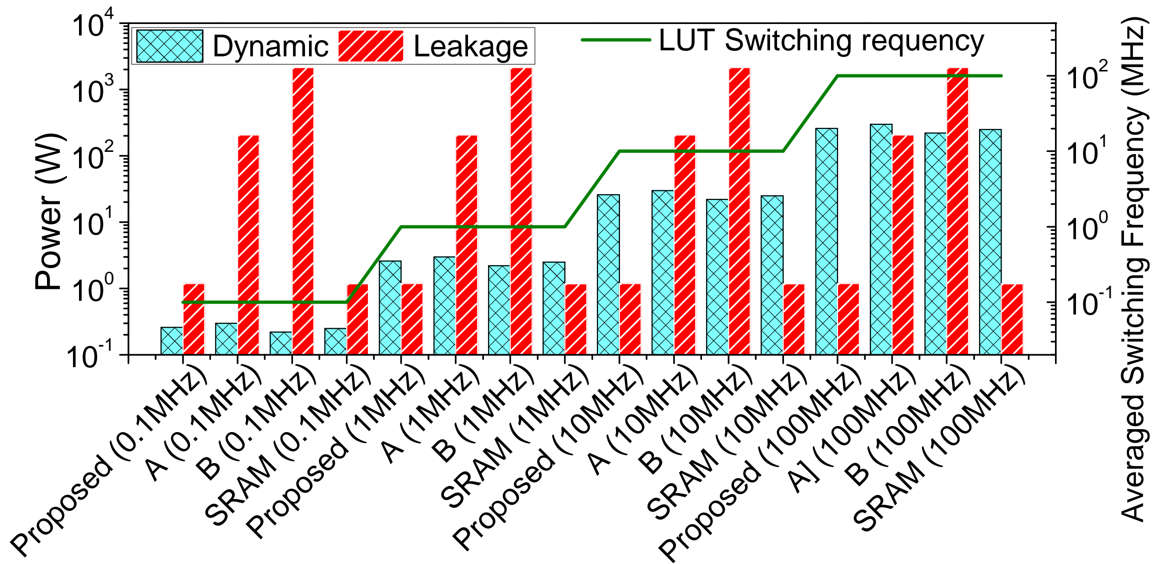
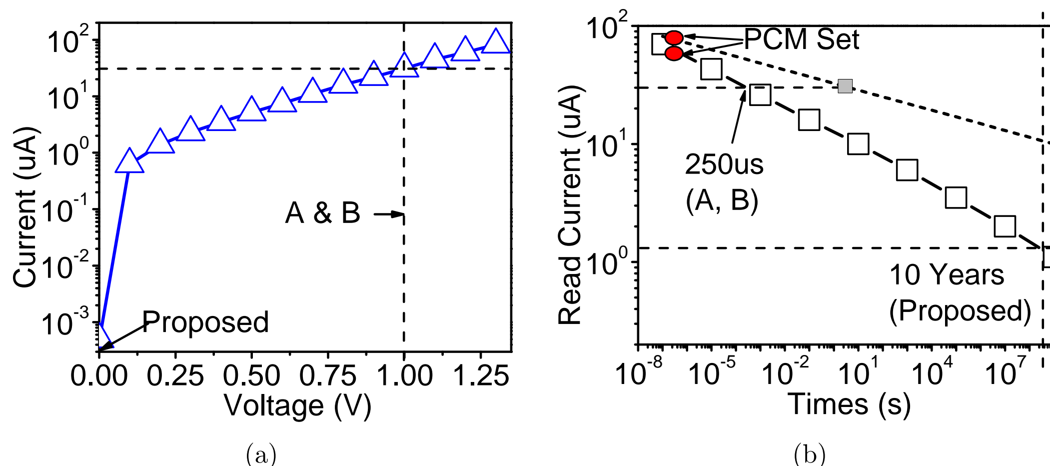


Figure 5.10: The power consumption comparison among different LUT architectures. A: [2]; B: [3].

The power and delay comparison among the four 4-input LUTs is tabulated in Table 5.3. The delay is measured from input A to output F. As shown in Table 5.3, the proposed nvSRAM based 4-input LUT achieves the similar speed performance as the conventional schemes. The $1.19nW$ active leakage power is similar to the SRAM-based LUT, but much smaller than [2] and [3]. The active leakage power of [2] and [3] is about 174 times and 1810 times higher than that of the proposed structure, respectively. Based on the 4-input LUT simulation results, our nvSRAM-based LUT could be powered off to reduce the leakage power when the sleep time is longer than $34.5\mu s$.

As illustrated in Fig. 5.10, the dynamic power and active leakage power of the four LUTs are compared at different operating frequencies. At low frequency ($i.e.$, $0.1MHz$), the active leakage power of [2] and [3] are $2-4$ orders higher than the dynamic power. Only when the averaged switching frequency is higher than $100MHz$, the active leakage power in [2] gets lower than the dynamic power. However, the active leakage power in [3] is still more than 10 times higher than



(a)

(b)

Figure 5.11: (a) IV curve of the PCM cell in the amorphous state. (b) the PCM retention of the designs in [2,3], and our proposed nvSRAM. A: [2]; B: [3].

its dynamic power. In contrast, even at 1MHz low switching frequency, the active leakage power of the LUT with our proposed nvSRAM is already lower than the dynamic power.

The retention time of PCM cells with our proposed nvSRAM, and the circuits in [2] and [3] are evaluated based on the data reported in [172]. As shown in Fig. 5.11, the reading current is exponentially increased with the reading voltage, and the crystallization time of PCM cells is exponentially reduced with reading current increased, which is because of the higher temperature inside PCM cells at higher reading current. Therefore, when the cells are biased at $1V$, the high reading current ($30\mu A$) leads to much shorter data retention time (crystallized in $250\mu s$). In our proposed design, the retention time could be longer than 10 years, since the sensing energy is low and there is no bias current in PCM cells during FPGA normal operations. The results are summarized in Table 5.3. The retention time of PCM may be improved by using different materials (*i.e.*, GeTe) [174,175]. However, the SET voltage/current may be increased due to the different materials. Moreover, the low retention problem may not be fully addressed due to the high

DC biased voltage, *i.e.*, the short-dash line shown in Fig. 5.11(b).

5.4.2 Multi-context Simulation Results

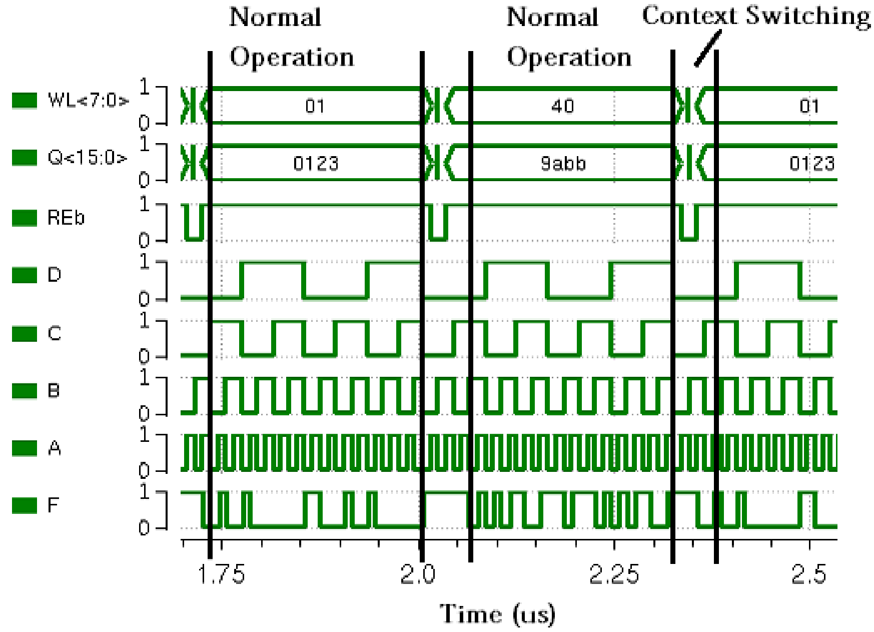
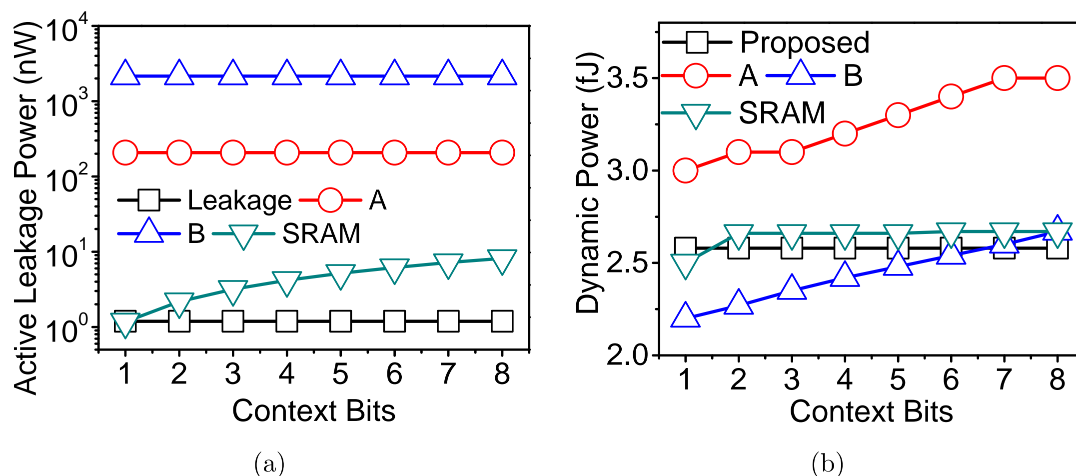


Figure 5.12: The RTR simulation results of the proposed 8-context nvSRAM based 4-input LUT.

The multi-context 4-input LUTs use the same structure as the single-context 4-input LUTs. Fig. 5.12 shows the run time reconfiguration of the 4-input LUT with 8-context nvSRAM. At the first read cycle, the multi-context nvSRAM address $8'h01$ is selected. This address sets the LUT to $16'h0123$ to have the logic function of $F = \bar{A}\bar{B}\bar{C} + A\bar{B}\bar{D}$. When the read operation is finished, the states of the PCM cells ($16'h0123$) are sensed and latched at the output Q<15:0>. The inputs of the LUT are swept from 4'b0000 to 4'b1111, and the sequence of the output signal F is ..1100_0100_1000_0000..., which agrees well with the states of the PCM cells. At around 2us, another read cycle selects $8'h40$ as the context address of the nvSRAM which sets the LUT logic function to $F = AB + A\bar{C} + \bar{B}C + \bar{B}\bar{D}$.

When the read operation is completed, the states of the data $Q<15:0>$ have been changed to $16'h9abb$. The switch between different context could be accomplished in less than 1ns.



(a) (b)

Figure 5.13: the 4-input LUT (a) active leakage power and (b) dynamic power comparison among the 6T SRAM, the designs in [2,3], and the proposed nvSRAM. A: [2]; B: [3].

Fig. 5.13(a) shows the multi-context 4-input LUT leakage power comparison among the 6T SRAM, the designs in [2, 3], and our proposed nvSRAM. Since the designs in [2, 3], and our proposed nvSRAM are using NVM technologies, the unselected context bits could be turned off, thus the active leakage power increases little at the wide span of context bits. However, the SRAM based LUT has to power on the unselected SRAM cells, thus higher context bits LUT draws higher active leakage power. Our nvSRAM based 8-context LUT reduces active power by 8, 174 and 1810 times, respectively, compared to the 8-context LUTs using 6T SRAM, the designs in [2] and [3].

Fig. 5.13(b) shows the 4-input LUT dynamic power comparison among four techniques. The SRAM and our proposed nvSRAM based LUTs have the similar dynamic power due to the same LUT structure is used. The dynamic power of [2,3]

gets higher with larger context bits is due to the parasitic capacitance from the other PCM select transistors.

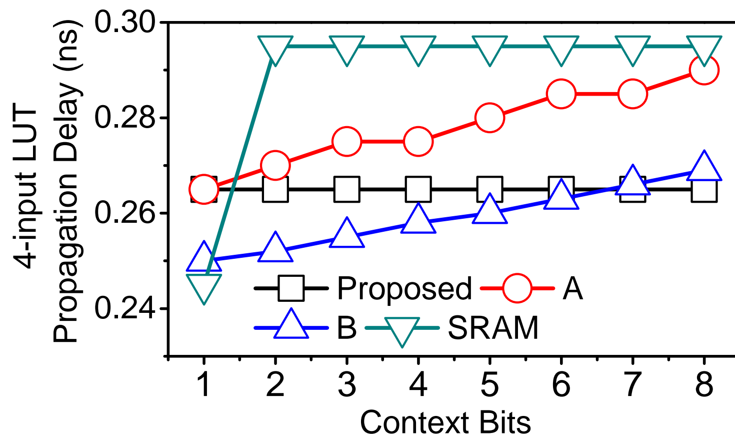


Figure 5.14: The propagation delay comparison among the 6T SRAM, the designs in [2, 3], and the proposed nvSRAM based 4-input LUTs. A: [2]; B: [3].

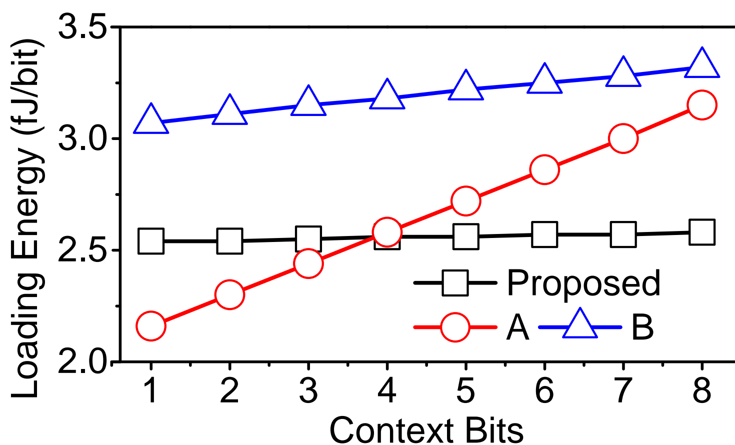


Figure 5.15: 4-input LUT loading power comparison among the 6T SRAM, the designs in [2, 3], and the proposed nvSRAM. A: [2]; B: [3].

Fig. 5.14 shows the propagation delay of four techniques. The additional context select switches are inserted between the multi-context SRAM and LUT switch matrix, resulting a longer delay in the multi-context SRAM based LUT compared to the single-context LUT. The parasitic capacitance of the design

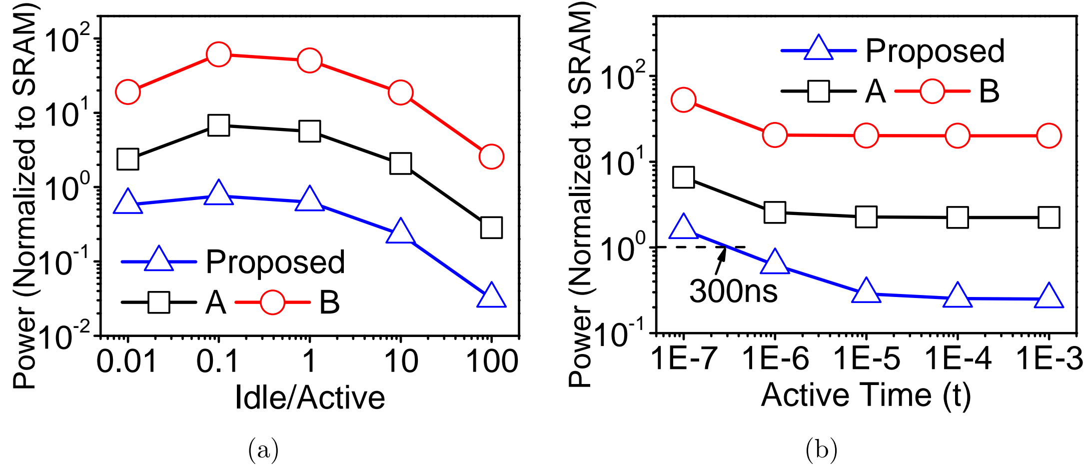


Figure 5.16: 8-context 4-input LUT power comparison among the designs in [2,3], and the proposed nvSRAM. All of the results are normalized to the SRAM based 8-context 4-input LUT under the same conditions. The average LUT switching frequency is set to 10MHz. (a) The power consumption versus the ratio of idle time and active time. The active time is set to $1ms$. (b) The power consumption versus the active time. The ratio of idle time and active time is 0.9. A: [2]; B: [3].

in [2, 3] gets larger at higher context bits, thus the total propagation delay is proportional to the context bits. The speed of our nvSRAM is determined by the latch. Therefore, its propagation delay is not affected by the increase of context bits.

Fig. 5.15 gives the loading energy per information bit comparison between the proposed nvSRAM, and the designs in [2] and [3]. The loading power of our nvSRAM has little dependence on context bits, which are $2.54fJ$ and $2.58fJ$ for the single-context and 8-context, respectively. However, from single-context to 8-context, the loading power increases about 40% and 10%, respectively, in designs of [2] and [3].

We further estimated the power consumption of the 8-context 4-input LUTs involving both idle/sleep time and active time as shown in Fig. 5.16. The SRAM-based LUT is still powered on during idle time, and its results are used as the

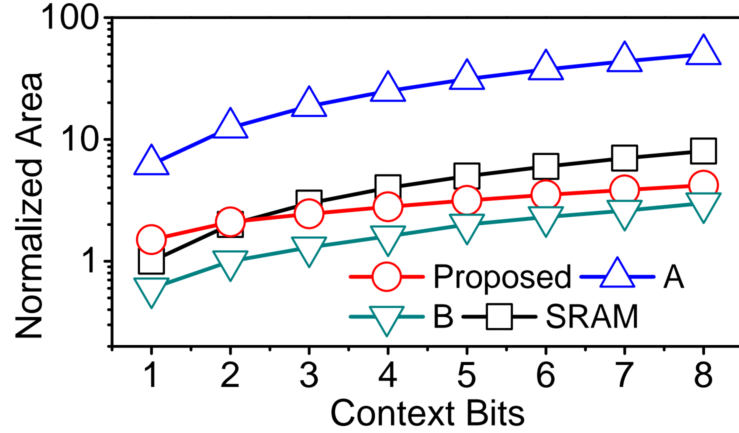


Figure 5.17: Area comparison among the 6T SRAM, the design in [2] and our proposed nvSRAM. The area is normalized to the single context 6T SRAM. A: [2]; B: [3].

baseline to compare the designs in [2, 3], and the proposed nvSRAM. As shown in Fig. 5.16(a), our nvSRAM based LUT has much lower power consumption than the SRAM-based LUT regardless of idle and active ratio. In contrast, the designs of [2] and [3] based LUTs start to outperform SRAM-based LUT in terms of power consumption only if the idle and active ratio is higher than 25 and 300, respectively. This is mainly due to high active leakage power. As shown in Fig. 5.16(b), our nvSRAM-based LUT consumes less power than the SRAM-based LUT when active time is longer than 300ns. Unfortunately, the designs of [2] and [3] based LUTs have more than 2 and 20 times higher power consumption than the SRAM-based LUT, respectively.

The area of the proposed multi-context nvSRAM can be derived from $AREA = AREA_1 + N * AREA_2$, where $AREA_1$ is the area of the latch plus the area of MN_2, MN_3 and the equalization transistor, $AREA_2$ is the area of single memory select pair. $AREA_1$ approximately equals to the area of the single context nvSRAM which is only 0.84um^2 based on the 45nm CMOS process node. The area comparison in Fig. 5.17 is based on the layout and the data provided

in [2], which has been normalized to $45nm$ after dividing it by 4. The cell size of the single context 6T SRAM in Fig. 5.17 is normalized to 1. Because of the thick oxide transistors, the normalized area of the PCM cell in [2] is more than 5 times larger than the proposed nvSRAM. The area of our nvSRAM gets smaller than 6T SRAM when the context bits are larger than 2.

5.5 Summary

In this chapter, we have proposed a PCM based non-volatile SRAM, which greatly reduces the active leakage power, and enhances the reliability of PCM cells by biasing PCM cells at $0V$ during the FPGA normal operation. The results have shown that the 4-input LUT with our nvSRAM has only $1.19nW$ active leakage power while producing $1ns$ fast loading speed. These features allow the system to be powered on/off to reduce the leakage power when standby time is longer than $34.5\mu s$. The analysis also shown that the retention of the PCM cells can be longer than 10 years. The results suggest that our proposed nvSRAM is a promising solution for low power and high reliability FPGAs.

Chapter 6

Conclusions

This dissertation has looked at many facets of using the new NVMs including STT-MRAM, PCM and RRAM in designing low power and high performance circuits.

The new nvFFs and localized NVM array based on STT-MRAM are proposed to retain the states of registers during standby. Both designs are targeting for the low VDD and low write power. The nvFF can be designed as a standard cell to compatible with digital design flow thus the design cycle could be greatly reduced. The localized NVM array could further reduce the power consumption with higher density. The non-volatile storage elements proposed for the nvFPGAs are targeting for the high reliability, high density and low power. Compared to the conventional nvFPGAs, the reliability is significantly improved, while compared to the SRAM-based FPGAs, the FPGA area and power could be greatly reduced.

Chapter 2 proposed a new nvFF to retain the states of registers during standby. Two-phase write approach and complementary write drivers were used in the nvFF, which reduced more than 38% power for the saving operation and also scales VDD down to 1V and below. The proposed nvFF has the closest FF performance as the CMOS retention FF. Moreover, it reduces more than 50% area

when compared to the smallest nvFF in the prior arts.

Chapter 3 proposed a novel NVM based circuit architecture with zero leakage power dissipation to further reduce the sleep power. It stored the states of the registers in the localized STT-MRAM array through scan chains, which had reduced by more than 20% sleep energy than conventional nvFF schemes, and saved by more than 99.8% sleep energy compared to the CMOS retention register based approaches when the sleep time is longer than $1s$. Moreover, the proposed pipelined quad-phase saving scheme maximized the saving speed, while reduced the peak saving current.

Chapter 4 further proposed a novel structure ('1D2R', '1-diode, 2-RRAM cells') to replace the NMOS switch and 6T SRAM. Based on systematic analysis, the proposed nvFPGA reduced the overall area by 78%, improved the speed by 1.94 times, and reduced the operation power by 40.9% compared to the SRAM-based FPGA. Furthermore, compared to other RRAM-based nvFPGAs, this novel structure significantly improved the write reliability by 8 orders magnitude for a 64×64 array with more than 20 times lower write power. This design fully unlocked the true potential of the RRAM-based FPGA and moves a solid step further toward real applications.

Chapter 5 presented a low active leakage power and high reliability PCM based non-volatile SRAM (nvSRAM). The low active leakage power and high reliability were achieved by biasing PCM cells at $0V$ during FPGA operation. Compared to the state-of-the-art, the proposed nvSRAM-based 4-input LUT achieved 174 times reduction in active leakage power and 15000 times increase in retention time. In addition, the proposed nvSRAM-based FPGA system significantly accelerated the loading speed to less than 1ns with $2.54fJ/cell$ loading energy.

In short, this dissertation presented new integration solutions and architectures to address various weaknesses in the conventional resistive NVM based

FFs and FPGAs. It had reduced the power consumption with higher density and performance. Moreover, the reliability was also greatly improved.

Acronyms

ASIC application specific integrated circuit. 3

BE bottom electrode. 7, 8, 12, 31

BEOL back-end of line. 91

BEP break even point. 14, 15, 42, 53, 57, 84, 115

BER bit error rate. 29

BLE basic logic elements. 90, 94, 95, 101

CB connection block. 4, 21, 89–91, 93–100, 102, 108

CMOS complementary metal oxide semiconductor. i, 1, 2, 4, 6, 15, 16, 27, 29, 41, 43–45, 82, 84, 88, 90–94, 99, 101, 102, 113, 122, 124, 132, 134, 135

DIBL drain induced barrier lowering. 2

DRAM dynamic random access memory. 4–7

ECC error correction code. 29, 54, 62, 63

FeRAM ferroelectric RAM. 5

FF flip-flop. 4, 15, 17–20, 25, 28, 29, 38, 41, 43–45, 58, 82–84, 88, 95, 134, 136

FIFO first-in-first-out. 56

FPGA field programmable gate array. i, 3, 4, 6, 16, 17, 20–22, 24, 25, 27, 89–95, 97–99, 101, 102, 107–109, 111–119, 121–123, 125, 127, 133–136

HRS high resistance state. 12, 13, 22, 23, 92, 98, 104

IC integrated circuit. 3

ITRI Industrial Technology Research Institute. 13

LB logic block. 4, 90, 94–97, 100, 102, 108

LRS low resistance state. 12, 13, 22, 23, 92, 98, 103, 106

LSI large scale integrated. 3

LUT look up table. 90, 94, 98, 101, 102, 108, 109, 111, 123–133, 135

MCNC Microelectronics Center of North Carolina. 111

MIEC Mixed Ionic Electronic Conduction. 91, 93

MIM metal insulator metal. 13

MOSFET metal oxide semiconductor field effect transistor. 1–3, 5, 21

MRAM magnetic RAM. 5–7, 12, 13, 29, 70

MRL merged reference line. 72

MTJ magnetic tunnel junction. 7–9, 17–20, 25, 29–34, 36–38, 41, 42, 44, 45, 47, 49, 50, 63, 66, 67, 70, 71, 74, 77, 78, 85

NREL National Renewable Energy Laboratory. 2

nvFF non-volatile flip-flop. i, 4, 15, 17, 18, 25, 28–31, 34–39, 41–45, 47, 49–51, 53, 54, 77, 82–85, 88, 134, 135

nvFPGA non-volatile FPGA. i, 4, 16, 17, 20, 25, 89, 90, 92–95, 98, 99, 102, 111, 112, 134, 135

nvLatch non-volatile latch. 4, 27, 29, 31, 34, 35, 50

nvLUT non-volatile LUT. 89, 98

NVM non-volatile memory. i, 4–7, 11–17, 20, 21, 25, 27–29, 53, 54, 56–58, 61, 85, 89, 112, 113, 117, 129, 134, 135

nvSRAM non-volatile SRAM. 17, 25, 27, 113–115, 117–129, 131–133, 135

PCM phase change memory. 5–7, 11–13, 15–17, 20–22, 25, 51, 113, 114, 117–119, 121–125, 127, 128, 130, 133–135

PDP power-delay product. 111

PG power gating. 3

PMC programmable metallization cell. 6

RA resistance-area product. 29, 45, 47, 49, 50

RRAM resistive random access memory. 5–7, 13, 15–17, 20–24, 51, 89–109, 112, 113, 134, 135

RTR run-time reconfiguration. 116

SB switch block. 4, 21, 89–91, 93, 94, 98–100, 102, 108

SOC System-on-Chip. 17, 28

SP switch point. 96, 98

SRAM static random access memory. i, 3, 4, 7, 16, 17, 20–22, 25, 34, 44, 63, 89, 90, 92, 93, 95, 99, 102, 108, 109, 111–113, 115–117, 119, 121, 122, 125, 126, 129–135

STT spin transfer torque. 70

STT-MRAM spin transfer torque MRAM. i, 7, 8, 15, 20, 22, 29, 43, 50, 51, 58, 66, 68, 70, 71, 73, 74, 88, 134, 135

TE top electrode. 7, 8, 12, 31

TMR tunnel magnetoresistance. 8, 31, 45, 47, 49, 51, 66, 71, 87

VLSI very large scale integrated. 53

VPR Versatile Place and Route. 102, 108, 109

Bibliography

[1] "International technology roadmap for semiconductors - emerging research devices (erd)," *http://www.itrs.net/Links/2011ITRS/Home2011.htm*, 2011.

[2] C. Wen, J. Li, S. Kim, M. Breitwisch, C. Lam, J. Paramesh, and L. Pileggi, "A non-volatile look-up table design using pcm (phase-change memory) cells," in *VLSI Circuits (VLSIC), 2011 Symposium on*, pp. 302–303, IEEE, 2011.

[3] P. Gaillardon, D. Sacchetto, G. Beneventi, M. Ben Jamaa, L. Perniola, F. Clermidy, I. O'Connor, and G. De Micheli, "Design and architectural assessment of 3-d resistive memory technologies in fpgas," *Nanotechnology, IEEE Transactions on*, vol. 12, no. 1, pp. 40–50, 2013.

[4] W. Zhao, E. Belhaire, C. Chappert, F. Jacquet, and P. Mazoyer, "New non-volatile logic based on spin-mtj," *physica status solidi (a)*, vol. 205, no. 6, pp. 1373–1377, 2008.

[5] Y. Shuto, S. Yamamoto, and S. Sugahara, "Nonvolatile static random access memory based on spin-transistor architecture," *Journal of Applied Physics*, vol. 105, no. 7, pp. 07C933–07C933, 2009.

[6] P. Wang, X. Chen, Y. Chen, H. Li, S. Kang, X. Zhu, and W. Wu, "A 1.0 v 45nm nonvolatile magnetic latch design and its robustness analysis," in

Custom Integrated Circuits Conference (CICC), 2011 IEEE, pp. 1–4, IEEE, 2011.

[7] Y. Jung, J. Kim, K. Ryu, J. Kim, S. Kang, and S. Jung, “An mtj-based non-volatile flip-flop for high-performance soc,” *International Journal of Circuit Theory and Applications*, 2012.

[8] T. Endoh, T. Ohsawa, H. Koike, T. Hanyu, and H. Ohno, “Restructuring of memory hierarchy in computing system with spintronics-based technologies,” in *VLSI Technology (VLSIT), 2012 Symposium on*, pp. 89–90, IEEE, 2012.

[9] N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, “Nonvolatile magnetic flip-flop for standby-power-free socs,” *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 8, pp. 2244–2250, 2009.

[10] A. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. Butler, P. Visscher, *et al.*, “Basic principles of stt-mram cell operation in memory arrays,” *Journal of Physics D: Applied Physics*, vol. 46, no. 7, pp. 74001–74020, 2013.

[11] K. Huang and Y. Lian, “A low-power low-vdd nonvolatile latch using spin transfer torque mram,” *Nanotechnology, IEEE Transactions on*, vol. 12, no. 6, pp. 1094–1103, 2013.

[12] K. K. Poon, A. Yan, and S. J. Wilton, “A flexible power model for fpgas,” in *Field-Programmable Logic and Applications: Reconfigurable Computing Is Going Mainstream*, pp. 312–321, Springer, 2002.

[13] K. K. Poon, S. J. Wilton, and A. Yan, “A detailed power model for field-programmable gate arrays,” *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 10, no. 2, pp. 279–302, 2005.

[14] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, “Design of ion-implanted mosfet’s with very small physical dimensions,” *Solid-State Circuits, IEEE Journal of*, vol. 9, no. 5, pp. 256–268, 1974.

[15] S. Wong and C. Salama, “Impact of scaling on mos analog performance,” *Solid-State Circuits, IEEE Journal of*, vol. 18, no. 1, pp. 106–114, 1983.

[16] B. Davari, R. Dennard, and G. Shahidi, “Cmos scaling for high performance and low power-the next ten years,” *Proceedings of the IEEE*, vol. 83, no. 4, pp. 595–606, 1995.

[17] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S. Lo, G. Sai-Halasz, R. Viswanathan, H. Wann, S. Wind, *et al.*, “Cmos scaling into the nanometer regime,” *Proceedings of the IEEE*, vol. 85, no. 4, pp. 486–504, 1997.

[18] K. Kuhn, “Cmos scaling beyond 32nm: challenges and opportunities,” in *Design Automation Conference, 2009. DAC’09. 46th ACM/IEEE*, pp. 310–313, IEEE, 2009.

[19] S. Thompson and S. Parthasarathy, “Moore’s law: the future of si microelectronics,” *Materials Today*, vol. 9, no. 6, pp. 20–25, 2006.

[20] S. Borkar, “Design challenges of technology scaling,” *Micro, IEEE*, vol. 19, no. 4, pp. 23–29, 1999.

[21] V. De and S. Borkar, “Technology and design challenges for low power and high performance,” in *Proceedings of the 1999 international symposium on Low power electronics and design*, pp. 163–168, ACM, 1999.

[22] M. Broussely and G. Archdale, “Li-ion batteries and portable power source prospects for the next 5–10 years,” *Journal of Power Sources*, vol. 136, no. 2, pp. 386–394, 2004.

[23] "Research cell efficiency records," *http://www.nrel.gov/ncpv*, 2013.

[24] F. Fallah and M. Pedram, "Standby and active leakage current control and minimization in cmos vlsi circuits," *IEICE transactions on electronics*, vol. 88, no. 4, pp. 509–519, 2005.

[25] B. Sheu, D. Scharfetter, P. Ko, and M. Jeng, "Bsim: Berkeley short-channel igfet model for mos transistors," *Solid-State Circuits, IEEE Journal of*, vol. 22, no. 4, pp. 558–566, 1987.

[26] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-v power supply high-speed digital circuit technology with multithreshold-voltage cmos," *Solid-State Circuits, IEEE Journal of*, vol. 30, no. 8, pp. 847–854, 1995.

[27] S. Tawfik and V. Kursun, "Low power and high speed multi threshold voltage interface circuits," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 5, pp. 638–645, 2009.

[28] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of reverse body bias for leakage control in scaled dual vt cmos ics," in *Low Power Electronics and Design, International Symposium on, 2001.*, pp. 207–212, IEEE, 2001.

[29] J. Tschanz, S. Narendra, Y. Ye, B. Bloechel, S. Borkar, and V. De, "Dynamic sleep transistor and body bias for active leakage power control of microprocessors," *Solid-State Circuits, IEEE Journal of*, vol. 38, no. 11, pp. 1838–1845, 2003.

[30] S. Narendra, V. De, D. Antoniadis, A. Chandrakasan, and S. Borkar, "Scaling of stack effect and its application for leakage reduction," in *Proceedings*

of the 2001 international symposium on Low power electronics and design, pp. 195–200, ACM, 2001.

[31] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, and P. Bose, "Microarchitectural techniques for power gating of execution units," in *Proceedings of the 2004 international symposium on Low power electronics and design*, pp. 32–37, ACM, 2004.

[32] H. Jiang, M. Marek-Sadowska, and S. Nassif, "Benefits and costs of power-gating technique," in *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, pp. 559–566, IEEE, 2005.

[33] H. Kaul, M. Anders, S. Mathew, S. Hsu, A. Agarwal, R. Krishnamurthy, and S. Borkar, "A 300 mv 494gops/w reconfigurable dual-supply 4-way simd vector processing accelerator in 45 nm cmos," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 1, pp. 95–102, 2010.

[34] S. Henzler, G. Georgakos, M. Eireiner, T. Nirschl, C. Pacha, J. Berthold, and D. Schmitt-Landsiedel, "Dynamic state-retention flip-flop for fine-grained power gating with small design and power overhead," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 7, pp. 1654–1661, 2006.

[35] S. Kim, S. Kosonocky, and D. Knebel, "Understanding and minimizing ground bounce during mode transition of power gating structures," in *Proceedings of the 2003 international symposium on Low power electronics and design*, pp. 22–25, ACM, 2003.

[36] S. Kim, S. Kosonocky, D. Knebel, K. Stawiasz, and M. Papaefthymiou, "A multi-mode power gating structure for low-voltage deep-submicron cmos

ics," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 54, no. 7, pp. 586–590, 2007.

[37] Y. Shin, J. Seomun, K. Choi, and T. Sakurai, "Power gating: Circuits, design methodologies, and best practice for standard-cell vlsi designs," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 15, no. 4, p. 28, 2010.

[38] S. Brown, J. Rose, and Z. Vranesic, "A detailed router for field-programmable gate arrays," in *Computer-Aided Design, 1990. ICCAD-90. Digest of Technical Papers., 1990 IEEE International Conference on*, pp. 382 –385, nov 1990.

[39] P. Chow, S. Seo, J. Rose, K. Chung, G. Páez-Monzón, and I. Rahardja, "The design of an sram-based field-programmable gate array. i. architecture," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 7, no. 2, pp. 191–197, 1999.

[40] I. Kuon, R. Tessier, and J. Rose, "Fpga architecture: Survey and challenges," *Foundations and Trends® in Electronic Design Automation*, vol. 2, no. 2, pp. 135–253, 2008.

[41] T. Lin, W. Zhang, and N. Jha, "Sram-based nature: A dynamically reconfigurable fpga based on 10t low-power srams," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 20, no. 11, pp. 2151–2156, 2012.

[42] V. Betz, J. Rose, and A. Marquardt, *Architecture and CAD for deep-submicron FPGAs.* Kluwer Academic Publishers, 1999.

[43] M. Lin, A. El Gamal, Y. Lu, and S. Wong, "Performance benefits of monolithically stacked 3-d fpga," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 2, pp. 216–229, 2007.

[44] K. Ma, L. Wang, X. Zhou, S. Tan, and J. Tong, "General switch box modeling and optimization for fpga routing architectures," in *Field-Programmable Technology (FPT), 2010 International Conference on*, pp. 320–323, IEEE, 2010.

[45] J. H. Anderson and F. N. Najm, "Active leakage power optimization for fpgas," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 25, no. 3, pp. 423–437, 2006.

[46] E. Morifuji, T. Yoshida, M. Kanda, S. Matsuda, S. Yamada, and F. Matsuoka, "Supply and threshold-voltage trends for scaled logic and sram mosfets," *Electron Devices, IEEE Transactions on*, vol. 53, no. 6, pp. 1427–1432, 2006.

[47] R. Thomas, J. Scott, D. Bose, and R. Katiyar, "Multiferroic thin-film integration onto semiconductor devices," *Journal of Physics: Condensed Matter*, vol. 22, no. 42, p. 423201, 2010.

[48] M. Khatib, P. Hartel, and H. van Dijk, "Energy-efficient streaming using non-volatile memory," *Journal of Signal Processing Systems*, vol. 60, no. 2, pp. 149–168, 2010.

[49] G. Burr, B. Kurdi, J. Scott, C. Lam, K. Gopalakrishnan, and R. Shenoy, "Overview of candidate device technologies for storage-class memory," *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 449–464, 2008.

[50] A. Beck, J. Bednorz, C. Gerber, C. Rossel, and D. Widmer, "Reproducible switching effect in thin oxide films for memory applications," *Applied Physics Letters*, vol. 77, no. 1, pp. 139–141, 2000.

[51] S. Liu, N. Wu, and A. Ignatiev, "Electric-pulse-induced reversible resistance change effect in magnetoresistive films," *Applied Physics Letters*, vol. 76, no. 19, pp. 2749–2751, 2000.

[52] B. Choi, D. Jeong, S. Kim, C. Rohde, S. Choi, J. Oh, H. Kim, C. Hwang, K. Szot, R. Waser, *et al.*, "Resistive switching mechanism of tio2 thin films grown by atomic-layer deposition," *Journal of applied physics*, vol. 98, no. 3, pp. 033715–033715, 2005.

[53] Y. Hosoi, Y. Tamai, T. Ohnishi, K. Ishihara, T. Shibuya, Y. Inoue, S. Yamazaki, T. Nakano, S. Ohnishi, N. Awaya, *et al.*, "High speed unipolar switching resistance ram (rram) technology," in *Electron Devices Meeting, 2006. IEDM'06. International*, pp. 1–4, IEEE, 2006.

[54] R. Waser and M. Aono, "Nanoionics-based resistive switching memories," *Nature materials*, vol. 6, no. 11, pp. 833–840, 2007.

[55] H. Lee, P. Chen, T. Wu, Y. Chen, C. Wang, P. Tzeng, C. Lin, F. Chen, C. Lien, and M. Tsai, "Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust hfo2 based rram," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, IEEE, 2008.

[56] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges," *Advanced Materials*, vol. 21, no. 25-26, pp. 2632–2663, 2009.

[57] D.-H. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, X.-S. Li, G.-S. Park, B. Lee, S. Han, *et al.*, "Atomic structure of conducting nanofilaments in tio2 resistive switching memory," *Nature nanotechnology*, vol. 5, no. 2, pp. 148–153, 2010.

[58] D. Halupka, S. Huda, W. Song, A. Sheikholeslami, K. Tsunoda, C. Yoshida, and M. Aoki, "Negative-resistance read and write schemes for stt-mram in 0.13μm cmos," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 256–257, 2010.

[59] S. Sheu, M. Chang, K. Lin, C. Wu, Y. Chen, P. Chiu, C. Kuo, Y. Yang, P. Chiang, W. Lin, *et al.*, "A 4mb embedded slc resistive-ram macro with 7.2 ns read-write random-access time and 160ns mlc-access capability," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pp. 200–202, IEEE, 2011.

[60] D. S. Jeong, R. Thomas, R. Katiyar, J. Scott, H. Kohlstedt, A. Petraru, and C. S. Hwang, "Emerging memories: resistive switching mechanisms and current status," *Reports on Progress in Physics*, vol. 75, no. 7, p. 076502, 2012.

[61] X. Yang and I. Chen, "Dynamic-load-enabled ultra-low power multiple-state rram devices," *Scientific Reports*, vol. 2, 2012.

[62] R. Fontana and S. Hetzler, "Magnetic memories: Memory hierarchy and processing perspectives," *Journal of applied physics*, vol. 99, no. 8, pp. 08N902–08N902, 2006.

[63] Y. Huai, "Spin-transfer torque mram (stt-mram): Challenges and prospects," *AAPPS Bulletin*, vol. 18, no. 6, pp. 33–40, 2008.

[64] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu, *et al.*, "45nm low power cmos logic compatible embedded stt mram utilizing a reverse-connection 1t/1mtj cell," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, IEEE, 2009.

[65] X. Guo, E. Ipek, and T. Soyata, "Resistive computation: avoiding the power wall with low-leakage, stt-mram based computing," in *ACM SIGARCH Computer Architecture News*, vol. 38, pp. 371–382, ACM, 2010.

[66] K. Huang, N. Ning, and Y. Lian, "Optimization scheme to minimize reference resistance distribution of spin-transfer-torque mram," *Very Large Scale*

Integration (VLSI) Systems, IEEE Transactions on, vol. PP, no. 99, pp. 1–1, 2013.

[67] S. Lai, "Current status of the phase change memory and its future," in *Electron Devices Meeting, 2003. IEDM'03 Technical Digest. IEEE International*, pp. 10–1, IEEE, 2003.

[68] M. Wuttig, "Phase-change materials: Towards a universal memory?," *Nature materials*, vol. 4, no. 4, pp. 265–266, 2005.

[69] M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage," *Nature materials*, vol. 6, no. 11, pp. 824–832, 2007.

[70] S. Raoux, G. Burr, M. Breitwisch, C. Rettner, Y. Chen, R. Shelby, M. Salinga, D. Krebs, S. Chen, H. Lung, *et al.*, "Phase-change random access memory: A scalable technology," *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 465–479, 2008.

[71] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, "Scalable high performance main memory system using phase-change memory technology," in *ACM SIGARCH Computer Architecture News*, vol. 37, pp. 24–33, ACM, 2009.

[72] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "A durable and energy efficient main memory using phase change memory technology," *ACM SIGARCH-Computer Architecture News*, vol. 37, no. 3, p. 14, 2009.

[73] H. Wong, S. Raoux, S. Kim, J. Liang, J. Reifenberg, B. Rajendran, M. Asheghi, and K. Goodson, "Phase change memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[74] R. Simpson, P. Fons, A. Kolobov, T. Fukaya, M. Krbal, T. Yagi, and J. Tominaga, “Interfacial phase-change memory,” *Nature Nanotechnology*, vol. 6, no. 8, pp. 501–505, 2011.

[75] W. J. Yu, S. H. Chae, S. Y. Lee, D. L. Duong, and Y. H. Lee, “Ultra-transparent, flexible single-walled carbon nanotube non-volatile memory device with an oxygen-decorated graphene electrode,” *Advanced Materials*, vol. 23, no. 16, pp. 1889–1893, 2011.

[76] S. Parkin, M. Hayashi, and L. Thomas, “Magnetic domain-wall racetrack memory,” *Science*, vol. 320, no. 5873, pp. 190–194, 2008.

[77] M. Hayashi, L. Thomas, R. Moriya, C. Rettner, and S. S. Parkin, “Current-controlled magnetic domain-wall nanowire shift register,” *Science*, vol. 320, no. 5873, pp. 209–211, 2008.

[78] H. Shiga, D. Takashima, S. Shiratake, K. Hoya, T. Miyakawa, R. Ogiwara, R. Fukuda, R. Takizawa, K. Hatsuda, F. Matsuoka, *et al.*, “A 1.6 gb/s ddr2 128 mb chain feram with scalable octal bitline and sensing schemes,” *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 1, pp. 142–152, 2010.

[79] P. Vettiger, M. Despont, U. Drechsler, U. Durig, W. Haberle, M. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, and G. Binnig, “The millipedemore than thousand tips for future afm storage,” *IBM Journal of Research and Development*, vol. 44, no. 3, pp. 323–340, 2000.

[80] Y. Chen, G. Jung, D. Ohlberg, X. Li, D. Stewart, J. Jeppesen, K. Nielsen, J. Stoddart, and R. Williams, “Nanoscale molecular-switch crossbar circuits,” *Nanotechnology*, vol. 14, no. 4, p. 462, 2003.

[81] S. Yu and H.-S. Wong, “Modeling the switching dynamics of programmable-metallization-cell (pmc) memory and its application as synapse device for a

neuromorphic computation system," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 22–1, IEEE, 2010.

[82] J. Shin and N. Pierce, "Rewritable memory by controllable nanopatterning of dna," *Nano Letters*, vol. 4, no. 5, pp. 905–909, 2004.

[83] C. Hermes, M. Wimmer, S. Menzel, K. Fleck, G. Bruns, M. Salinga, U. Bottger, R. Bruchhaus, T. Schmitz-Kempen, M. Wuttig, *et al.*, "Analysis of transient currents during ultrafast switching of tio2 nanocrossbar devices," *Electron Device Letters, IEEE*, vol. 32, no. 8, pp. 1116–1118, 2011.

[84] J. Yang, M. Zhang, J. Strachan, F. Miao, M. Pickett, R. Kelley, G. Medeiros-Ribeiro, and R. Williams, "High switching endurance in tao memristive devices," *Applied Physics Letters*, vol. 97, no. 23, pp. 232102–232102, 2010.

[85] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, *et al.*, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-ram," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pp. 459–462, IEEE, 2005.

[86] J. Li, P. Ndai, A. Goel, S. Salahuddin, and K. Roy, "Design paradigm for robust spin-torque transfer magnetic ram (stt mram) from circuit/architecture perspective," *IEEE Tran. Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 12, pp. 1710–1723, 2010.

[87] S. Oh, J. Jeong, W. Lim, W. Kim, Y. Kim, H. Shin, J. Lee, Y. Shin, S. Choi, and C. Chung, "On-axis scheme and novel mtj structure for sub-30nm gb density stt-mram," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, pp. 12.6.1 – 12.6.4, 2010.

[88] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L. Wang, and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *J. Physics: Condensed Matter*, vol. 19, p. 165209, 2007.

[89] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Design space and scalability exploration of 1t-1stt mtj memory arrays in the presence of variability and disturbances," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, pp. 1–4, IEEE, 2009.

[90] W. Wernsdorfer, E. Orozco, K. Hasselbach, A. Benoit, B. Barbara, N. Demoncy, A. Loiseau, H. Pascard, and D. Mailly, "Experimental evidence of the néel-brown model of magnetization reversal," *Physical review letters*, vol. 78, no. 9, pp. 1791–1794, 1997.

[91] R. Beach, T. Min, C. Horng, Q. Chen, P. Sherman, S. Le, S. Young, K. Yang, H. Yu, X. Lu, *et al.*, "A statistical study of magnetic tunnel junctions for high-density spin torque transfer-mram (stt-mram)," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, IEEE, 2008.

[92] J. Harms, F. Ebrahimi, X. Yao, and J. Wang, "Spice macromodel of spin-torque-transfer-operated magnetic tunnel junctions," *Electron Devices, IEEE Transactions on*, vol. 57, no. 6, pp. 1425–1430, 2010.

[93] S. R. Ovshinsky, "Reversible electrical switching phenomena in disordered structures," *Physical Review Letters*, vol. 21, no. 20, pp. 1450–1453, 1968.

[94] R. Neale, D. Nelson, and G. Moore, "Nonvolatile and reprogrammable, the read-mostly memory is here," *Electronics*, vol. 43, no. 20, pp. 56–60, 1970.

[95] H. Hamann, M. O'Boyle, Y. Martin, M. Rooks, and H. Wickramasinghe, "Ultra-high-density phase-change storage and memory," *Nature materials*, vol. 5, no. 5, pp. 383–387, 2006.

[96] S. Lee, Y. Jung, and R. Agarwal, "Highly scalable non-volatile and ultra-low-power phase-change nanowire memory," *Nature nanotechnology*, vol. 2, no. 10, pp. 626–630, 2007.

[97] M. Lankhorst, B. Ketelaars, and R. Wolters, "Low-cost and nanoscale non-volatile memory concept for future silicon chips," *Nature materials*, vol. 4, no. 4, pp. 347–352, 2005.

[98] B. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable dram alternative," in *ACM SIGARCH Computer Architecture News*, vol. 37, pp. 2–13, ACM, 2009.

[99] G. Servalli, "A 45nm generation phase change memory technology," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, IEEE, 2009.

[100] I. Kim, S. Cho, D. Im, E. Cho, D. Kim, G. Oh, D. Ahn, S. Park, S. Nam, J. Moon, *et al.*, "High performance pram cell scalable to sub-20nm technology with below 4f2 cell size, extendable to dram applications," in *VLSI Technology (VLSIT), 2010 Symposium on*, pp. 203–204, IEEE, 2010.

[101] Y. Choi, I. Song, M. Park, H. Chung, S. Chang, B. Cho, J. Kim, Y. Oh, D. Kwon, J. Sunwoo, *et al.*, "A 20nm 1.8 v 8gb pram with 40mb/s program bandwidth," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 46–48, IEEE, 2012.

[102] D. Loke, T. Lee, W. Wang, L. Shi, R. Zhao, Y. Yeo, T. Chong, and S. Elliott, "Breaking the speed limits of phase-change memory," *Science*, vol. 336, no. 6088, pp. 1566–1569, 2012.

[103] W. Wang, D. Loke, L. Shi, R. Zhao, H. Yang, L. Law, L. Ng, K. Lim, Y. Yeo, T. Chong, *et al.*, "Enabling universal memory by overcoming the contradictory speed and stability nature of phase-change materials," *Scientific Reports*, vol. 2, 2012.

[104] G. Bruns, P. Merkelbach, C. Schlockermann, M. Salinga, M. Wuttig, T. Happ, J. Philipp, and M. Kund, "Nanosecond switching in gete phase change memory cells," *Applied physics letters*, vol. 95, no. 4, pp. 043108–043108, 2009.

[105] F. Xiong, A. Liao, D. Estrada, and E. Pop, "Low-power switching of phase-change materials with carbon nanotube electrodes," *Science*, vol. 332, no. 6029, pp. 568–570, 2011.

[106] K. Chopra, "Growth of thin metal films under applied electric field," *Applied Physics Letters*, vol. 7, no. 5, pp. 140–142, 1965.

[107] W. Hiatt and T. Hickmott, "Bistable switching in niobium oxide diodes," *Applied Physics Letters*, vol. 6, no. 6, pp. 106–108, 1965.

[108] F. Argall, "Switching phenomena in titanium oxide thin films," *Solid-State Electronics*, vol. 11, no. 5, pp. 535–541, 1968.

[109] T. Hickmott, "Electroluminescence, bistable switching, and dielectric breakdown of nb2o5 diodes," *Journal of Vacuum Science and Technology*, vol. 6, no. 5, pp. 828–833, 1969.

[110] G. Dearnaley, A. Stoneham, and D. Morgan, "Electrical phenomena in amorphous oxide films," *Reports on Progress in Physics*, vol. 33, no. 3, p. 1129, 2002.

[111] R. Muenstermann, T. Menke, R. Dittmann, and R. Waser, "Coexistence of filamentary and homogeneous resistive switching in fe-doped srtio3 thin-film memristive devices," *Advanced Materials*, vol. 22, no. 43, pp. 4819–4822, 2010.

[112] H. Okushi, A. Matsuda, M. Saito, M. Kikuchi, and Y. Hirai, "Polarized (letter 8) memory effects in hetero-systems and non hetero-systems," *Solid State Communications*, vol. 11, no. 1, pp. 283–286, 1972.

[113] J. De Blauwe, "Nanocrystal nonvolatile memory devices," *Nanotechnology, IEEE Transactions on*, vol. 1, no. 1, pp. 72–77, 2002.

[114] F. Bedeschi, R. Bez, C. Boffino, E. Bonizzoni, E. Buda, G. Casagrande, L. Costa, M. Ferraro, R. Gastaldi, O. Khouri, *et al.*, "4-mb mosfet-selected phase-change memory experimental chip," in *Solid-State Circuits Conference, 2004. ESSCIRC 2004. Proceeding of the 30th European*, pp. 207–210, IEEE, 2004.

[115] B. C. Lee, P. Zhou, J. Yang, Y. Zhang, B. Zhao, E. Ipek, O. Mutlu, and D. Burger, "Phase-change technology and the future of main memory," *Micro, IEEE*, vol. 30, no. 1, pp. 143–143, 2010.

[116] W.-Y. Chang, Y.-C. Lai, T.-B. Wu, S.-F. Wang, F. Chen, and M.-J. Tsai, "Unipolar resistive switching characteristics of zno thin films for nonvolatile memory applications," *Applied Physics Letters*, vol. 92, no. 2, pp. 022110–022110, 2008.

[117] C. Cheng, A. Chin, and F. Yeh, "Novel ultra-low power rram with good endurance and retention," in *VLSI Technology (VLSIT), 2010 Symposium on*, pp. 85–86, IEEE, 2010.

[118] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, *et al.*, "Lower-current and fast switching of a perpendicular tmr for high speed and high density spin-transfer-torque mram," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, IEEE, 2008.

[119] Y. Chen, C. Chen, C. Chen, J. Yu, S. Wu, S. Lung, R. Liu, and C. Lu, "An access-transistor-free (0t/1r) non-volatile resistance random access memory (rram) using a novel threshold switching, self-rectifying chalcogenide device," in *Electron Devices Meeting, 2003. IEDM'03 Technical Digest. IEEE International*, pp. 37–4, IEEE, 2003.

[120] Y. Kim, S. Lee, D. Lee, C. Lee, M. Chang, J. Hur, M. Lee, G. Park, C. Kim, U. Chung, *et al.*, "Bi-layered rram with unlimited endurance and extremely uniform switching," in *VLSI Technology (VLSIT), 2011 Symposium on*, pp. 52–53, IEEE, 2011.

[121] M. Chang, C. Wu, C. Kuo, S. Shen, K. Lin, S. Yang, Y. King, C. Lin, and Y. Chih, "A 0.5v 4mb logic-process compatible embedded resistive ram (r-eram) in 65nm cmos using low-voltage current-mode sensing scheme with 45ns random read time," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 434–436, IEEE, 2012.

[122] C. Cheng, A. Chin, and F. Yeh, "Stacked geo/srtiox resistive memory with ultralow resistance currents," *Applied Physics Letters*, vol. 98, p. 052905, 2011.

[123] W. Guan, S. Long, Q. Liu, M. Liu, and W. Wang, "Nonpolar nonvolatile resistive switching in cu doped zro2," *Electron Device Letters, IEEE*, vol. 29, no. 5, pp. 434–437, 2008.

[124] D. Morris, D. Bromberg, J. Zhu, and L. Pileggi, "mlogic: ultra-low voltage non-volatile logic circuits using stt-mtj devices," in *Proceedings of the 49th Annual Design Automation Conference*, pp. 486–491, ACM, 2012.

[125] P. Chiu, M. Chang, C. Wu, C. Chuang, S. Sheu, Y. Chen, and M. Tsai, "Low store energy, low vddmin, 8t2r nonvolatile latch and sram with vertical-stacked resistive memory (memristor) devices for low power mobile applications," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 6, pp. 1483–1496, 2012.

[126] Y.-C. Chen, W. Zhang, and H. Li, "A look up table design with 3d bipolar rrams," in *Design Automation Conference (ASP-DAC), 2012 17th Asia and South Pacific*, pp. 73 –78, 30 2012-feb. 2 2012.

[127] P.-E. Gaillardon, M. Ben-Jamaa, M. Reyboz, G. Beneventi, F. Clermidy, L. Perniola, and I. O'Connor, "Phase-change-memory-based storage elements for configurable logic," in *Field-Programmable Technology (FPT), 2010 International Conference on*, pp. 17 –20, dec. 2010.

[128] W. Zhao, E. Belhaire, C. Chappert, and P. Mazoyer, "Spin transfer torque (stt)-mram–based runtime reconfiguration fpga circuit," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 9, no. 2, p. 14, 2009.

[129] W. Wang, T. Jing, and B. Butcher, "Fpga based on integration of memristors and cmos devices," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 1963–1966, IEEE, 2010.

[130] Y. Chen, W. Wang, H. Li, and W. Zhang, "Non-volatile 3d stacking rram-based fpga," in *Field Programmable Logic and Applications (FPL), 2012 22nd International Conference on*, pp. 367–372, IEEE, 2012.

[131] Y. Liauw, Z. Zhang, W. Kim, A. Gamal, and S. Wong, "Nonvolatile 3d-fpga with monolithically stacked rram-based configuration memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 406–408, IEEE, 2012.

[132] S. Kim, B. Lee, M. Asheghi, G. Hurkx, J. Reifenberg, K. Goodson, and H. Wong, "Thermal disturbance and its impact on reliability of phase-change memory studied by the micro-thermal stage," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 99–103, IEEE, 2010.

[133] Y. Chen, H. Lee, P. Chen, P. Gu, C. Chen, W. Lin, W. Liu, Y. Hsu, S. Sheu, P. Chiang, *et al.*, "Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, IEEE, 2009.

[134] S. Tanachutiwat, M. Liu, and W. Wang, "Fpga based on integration of cmos and rram," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 11, pp. 2023–2032, 2011.

[135] J. Cong and B. Xiao, "mrfpga: A novel fpga architecture with memristor-based reconfiguration," in *Nanoscale Architectures (NANOARCH), 2011 IEEE/ACM International Symposium on*, pp. 1–8, IEEE, 2011.

[136] P.-E. Gaillardon, M. Ben-Jamaa, G. Beneventi, F. Clermidy, and L. Perniola, "Emerging memory technologies for reconfigurable routing in fpga architecture," in *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, pp. 62 –65, dec. 2010.

[137] C. Xu, X. Dong, N. P. Jouppi, and Y. Xie, "Design implications of memristor-based rram cross-point structures," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011*, pp. 1–6, IEEE, 2011.

[138] Y. Zhang, X. Wang, Y. Li, A. Jones, and Y. Chen, "Asymmetry of mtj switching and its implication to stt-ram designs," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012*, pp. 1313–1318, IEEE, 2012.

[139] J. Kim, T. Kim, W. Hao, H. Rao, K. Lee, X. Zhu, X. Li, W. Hsu, S. Kang, N. Matt, *et al.*, "A 45nm 1mb embedded stt-mram with design techniques to minimize read-disturbance," in *VLSI Circuits (VLSIC), 2011 Symposium on*, pp. 296–297, IEEE, 2011.

[140] L. Benini, A. Bogliolo, and G. De Micheli, "A survey of design techniques for system-level dynamic power management," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 299–316, 2000.

[141] L. Benini, G. De Micheli, A. Macii, E. Macii, M. Poncino, and R. Scarsi, "Glitch power minimization by selective gate freezing," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 287–298, 2000.

[142] J. W. McPherson, "Reliability challenges for 45nm and beyond," in *Proceedings of the 43rd annual Design Automation Conference*, pp. 176–181, ACM, 2006.

[143] L. Torres, R. M. Brum, Y. Guillemenet, G. Sassatelli, and L. V. Cargnini, "Evaluation of hybrid mram/cmos cells for reconfigurable computing," in

New Circuits and Systems Conference (NEWCAS), 2013 IEEE 11th International, pp. 1–6, 2013.

[144] K. Ono, T. Kawahara, R. Takemura, K. Miura, M. Yamanouchi, J. Hayakawa, K. Ito, H. Takahashi, H. Matsuoka, S. Ikeda, *et al.*, "Spram with large thermal stability for high immunity to read disturbance and long retention for high-temperature operation," in *Proc. Symp. VLSI Technology*, pp. 228–229, 2009.

[145] S. Yamamoto, Y. Shuto, and S. Sugahara, "Nonvolatile delay flip-flop using spin-transistor architecture with spin transfer torque mtjs for power-gating systems," *Electronics letters*, vol. 47, no. 18, pp. 1027–1029, 2011.

[146] Y. Xie, "Modeling, architecture, and applications for emerging memory technologies," *Design & Test of Computers, IEEE*, vol. 28, no. 1, pp. 44–51, 2011.

[147] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, T. Kajiyama, M. Iwayama, K. Sugiura, S. Ikegawa, *et al.*, "A 64mb mram with clamped-reference and adequate-reference schemes," in *Proc. IEEE. Int. Solid-State Circuits Conf. (ISSCC)*, pp. 258–259, 2010.

[148] Y. Lee, C. Yoshida, K. Tsunoda, S. Umehara, M. Aoki, and T. Sugii, "Highly scalable stt-mram with mtjs of top-pinned structure in 1t/1mtj cell," in *Proc. Symp. VLSI Technology (VLSIT)*, pp. 49–50, IEEE, 2010.

[149] J. Zhu, "Magnetoresistive random access memory: the path to competitiveness and scalability," *Proceedings of the IEEE*, vol. 96, no. 11, pp. 1786–1798, 2008.

[150] J. Katine and E. Fullerton, "Device implications of spin-transfer torques," *Journal of Magnetism and Magnetic Materials*, vol. 320, no. 7, pp. 1217–1226, 2008.

[151] W. Xu, T. Zhang, and Y. Chen, "Design of spin-torque transfer magnetoresistive ram and cam/tcam with high sensing and search speed," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 1, pp. 66–74, 2010.

[152] K. Miura, T. Kawahara, R. Takemura, J. Hayakawa, S. Ikeda, R. Sasaki, H. Takahashi, H. Matsuoka, and H. Ohno, "A novel spram (spin-transfer torque ram) with a synthetic ferrimagnetic free layer for higher immunity to read disturbance and reducing write-current dispersion," in *Proc. Symp. VLSI Technology*, pp. 234–235, 2007.

[153] R. Takemura, T. Kawahara, K. Miura, H. Yamamoto, J. Hayakawa, N. Matsuzaki, K. Ono, M. Yamanouchi, K. Ito, H. Takahashi, *et al.*, "A 32-mb spram with 2t1r memory cell, localized bi-directional write driver and1'/0'dual-array equalized reference scheme," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 4, pp. 869–879, 2010.

[154] M. Liu and W. Wang, "rfga: Cmos-nano hybrid fpga using rram components," in *Nanoscale Architectures, 2008. NANOARCH 2008. IEEE International Symposium on*, pp. 93–98, IEEE, 2008.

[155] Y. Chen, W. Zhang, and H. Li, "A look up table design with 3d bipolar rrams," in *Design Automation Conference (ASP-DAC), 2012 17th Asia and South Pacific*, pp. 73–78, IEEE, 2012.

[156] Y. Chen, H. Li, and W. Zhang, "A novel peripheral circuit for rram-based lut," in *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*, pp. 1811–1814, IEEE, 2012.

[157] K. Gopalakrishnan, R. Shenoy, C. Rettner, K. Virwani, D. Bethune, R. Shelby, G. Burr, A. Kellock, R. King, K. Nguyen, *et al.*, "Highly-scalable novel access device based on mixed ionic electronic conduction (miec) materials for high density phase change memory (pcm) arrays," in *VLSI Technology (VLSIT), 2010 Symposium on*, pp. 205–206, IEEE, 2010.

[158] J. Cong and B. Xiao, "Fpga-rpi: A novel fpga architecture with rram-based programmable interconnects," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.

[159] X. Inc., *The programmable logic data book*. 1994.

[160] Y. Chang, D. Wong, and C. Wong, "Universal switch modules for fpga design," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 1, no. 1, pp. 80–101, 1996.

[161] M. Shyu, G. Wu, Y. Chang, and Y. Chang, "Generic universal switch blocks," *Computers, IEEE Transactions on*, vol. 49, no. 4, pp. 348–359, 2000.

[162] H. Fan, J. Liu, Y. Wu, and C. Cheung, "On optimum switch box designs for 2-d fpgas," in *Proceedings of the 38th annual Design Automation Conference*, pp. 203–208, ACM, 2001.

[163] H. Fan, J. Liu, and Y. Wu, "General models and a reduction design technique for fpga switch box designs," *Computers, IEEE Transactions on*, vol. 52, no. 1, pp. 21–30, 2003.

[164] S. Wilton, *Architectures and algorithms for field-programmable gate arrays with embedded memory*. PhD thesis, Citeseer, 1997.

[165] H. Schmit and V. Chandra, "Fpga switch block layout and evaluation," in *Proceedings of the 2002 ACM/SIGDA tenth international symposium on Field-programmable gate arrays*, pp. 11–18, ACM, 2002.

[166] C. Dong, S. Chilstedt, and D. Chen, "Reconfigurable circuit design with nanomaterials," in *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE'09.*, pp. 442–447, IEEE, 2009.

[167] V. Betz and J. Rose, "Vpr: A new packing, placement and routing tool for fpga research," in *Field-Programmable Logic and Applications*, pp. 213–222, Springer, 1997.

[168] Y. Chen, J. Zhao, and Y. Xie, "3d-nonfar: three-dimensional non-volatile fpga architecture using phase change memory," in *Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design*, pp. 55–60, ACM, 2010.

[169] K. Compton and S. Hauck, "Reconfigurable computing: a survey of systems and software," *ACM Computing Surveys (csuR)*, vol. 34, no. 2, pp. 171–210, 2002.

[170] K. Compton, S. Hauck, and K. Compton, "An introduction to reconfigurable computing," *IEEE Computer*, 2000.

[171] N. Bruchon, L. Torres, G. Sassatelli, and G. Cambon, "New nonvolatile fpga concept using magnetic tunneling junction," in *Emerging VLSI Technologies and Architectures, 2006. IEEE Computer Society Annual Symposium on*, pp. 6–pp, IEEE, 2006.

[172] A. Pirovano, A. Redaelli, F. Pellizzer, F. Ottogalli, M. Tosi, D. Ielmini, A. Lacaita, and R. Bez, "Reliability study of phase-change nonvolatile mem-

ories," *Device and Materials Reliability, IEEE Transactions on*, vol. 4, no. 3, pp. 422–427, 2004.

[173] M. Wieckowski, G. K. Chen, D. Kim, D. Blaauw, and D. Sylvester, "A 128kb high density portless sram using hierarchical bitlines and thyristor sense amplifiers," in *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, pp. 1–4, IEEE, 2011.

[174] L. Perniola, V. Sousa, A. Fantini, E. Arbaoui, A. Bastard, M. Armand, A. Fargeix, C. Jahan, J.-F. Nodin, A. Persico, *et al.*, "Electrical behavior of phase-change memory cells based on gete," *Electron device Letters, IEEE*, vol. 31, no. 5, pp. 488–490, 2010.

[175] G. Betti Beneventi, L. Perniola, V. Sousa, E. Gourvest, S. Maitrejean, J. Bastien, A. Bastard, B. Hyot, A. Fargeix, C. Jahan, *et al.*, "Carbon-doped gete: A promising material for phase-change memories," *Solid-State Electronics*, vol. 65, pp. 197–204, 2011.