UNIVERSITY OF CALIFORNIA
Los Angeles

# Design, Evaluation and Co-optimization of Emerging Devices and Circuits

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Shaodi Wang

2017

ABSTRACT OF THE DISSERTATION

Design, Evaluation and Co-optimization of

Emerging Devices and Circuits

by

Shaodi Wang

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2017

Professor Puneet Gupta, Chair

The continued push for traditional Silicon technology scaling faces the main challenge of non-scaling power density. Exploring alternative power-efficient technologies is essential for sustaining technology development. Many emerging technologies have been proposed as potential replacement for Silicon technology. However, these emerging technologies need rigorous evaluation in the contexts of circuits and systems to identify their value prior to commercial investment. We have developed evaluation frameworks covering emerging Boolean logic devices, memory devices, memory systems, and integration technologies. The evaluation metrics are in terms of delay, power, and reliability. According to the evaluation results, the development of emerging Boolean logic devices is still far from being able to replace Silicon technology, but magnetic random access memory (MRAM) is a promising memory technology showing benefits in performance and energy-efficiency.

As a specific example, we co-optimize MRAM with application circuits and systems. Optimized MRAM write and read design can significantly improve the system performance. We have proposed magnetic tunnel junction (MTJ) based process and temperature variation monitor, which enables variation-aware MRAM write and read optimization. We have also proposed utilizing negative differential resistance (NDR) to enable fast and energy-efficient write and zero-disturbance read for resistive memories including MRAM. In addition, we also design and adapt MRAM technology into low-power stochastic computing system to improve energy-efficiency. To further improve

the stochastic computing system, a promising VC-MTJ based true random stochastic bitstream generator is proposed and utilized.

The dissertation of Shaodi Wang is approved.

Yuan Xie

Subramanian S. Iyer

Kang L. Wang

Puneet Gupta, Committee Chair

University of California, Los Angeles

2017

*I dedicate my PhD dissertation to my wife, my parents, my advisor, and my labmates.*

# Table of Contents

**1 Introduction** . . . . . **1**

1.1 Emerging Technology Introduction . . . . . 2

1.1.1 Emerging Boolean Logic Devices . . . . . 2

1.1.2 Emerging Memories . . . . . 3

1.1.3 Voltage-control Magnetic Tunnel Junction (VC-MTJ) . . . . . 4

1.1.4 Emerging Integration Technologies . . . . . 5

1.1.5 Emerging Memory System Reliability . . . . . 7

1.2 Emerging Technologies Evaluation . . . . . 7

1.3 Design for Efficient MRAM Write and Read . . . . . 8

1.3.1 Optimized MRAM Write and Read Enabled by MTJ-based Variation Monitor . . . . . 9

1.3.2 Negative Differential Resistance-assisted NVM Write and Read . . 9

1.4 Stochastic Non-volatile Computing using MTJ . . . . . 11

**2 PROCEED: A Pareto Optimization-based Circuit-level Evaluator for Emerging Devices** . . . . . **13**

2.1 Chapter Introduction . . . . . 13

2.2 Overview of PROCEED Framework . . . . . 15

2.2.1 Canonical Circuit Construction . . . . . 16

2.2.2 Delay and Power Modeling . . . . . 18

2.2.3 Area Modeling . . . . . 18

2.2.4 Process Variation and Voltage Drop . . . . . 20

2.2.5 Interconnect Load . . . . . 20

2.2.6 Pareto-Based Optimization . . . . . 20

2.2.7 Power Management Modeling . . . . . 21

2.2.8 Activity Factor . . . 22
2.2.9 Multiple $V_{dd}$ and $V_t$ . . . 23
2.2.10 Heterogeneous Integration . . . 23
2.3 Pareto Optimization . . . 24
2.4 Experiment Results . . . 28
2.4.1 Framework Evaluation . . . 29
2.4.2 Impact of Multiple $V_{dd}$, $V_t$, and Gate Sizing . . . 30
2.4.3 Impact of Interconnect Load . . . 31
2.4.4 Impact of Benchmarks on Evaluation SOI vs. TFET . . . 32
2.4.5 Impact of Activity Factor SOI vs. TFET . . . 34
2.4.6 Power Management Modeling . . . 35
2.4.7 Variation-Aware Evaluation . . . 35
2.4.8 Heterogeneous Integration Evaluation . . . 36
2.5 Chapter Conclusion . . . 38

**3 Evaluation of Digital Circuit-Level Variability in Inversion-Mode and Junctionless FinFET Technologies . . . 40**
3.1 Chapter Introduction . . . 40
3.2 Overview of Evaluation Framework . . . 41
3.3 Variability and Device Modeling . . . 42
3.3.1 LER and RDF Modeling . . . 42
3.3.2 Device-Level Variability . . . 44
3.3.3 Device and Variability Model Fitting . . . 45
3.4 Variability Impact on 6T SRAM Memory . . . 47
3.4.1 Baseline Nominal Static Noise Margin . . . 47
3.4.2 Minimum Working $V_{cc}$ ($V_{ccmin}$) . . . 48
3.4.3 SNM Versus Technology . . . 51

3.5 LER Impact on Logic Circuit Variability . . . . . 52
3.5.1 Overview . . . . . 52
3.5.2 Circuit Statistical Timing and Power Analysis . . . . . 52
3.5.3 Circuit Simulation Results . . . . . 53
3.6 Chapter Conclusion . . . . . 55

**4 MEMRES: A Fast Memory System Reliability Simulator . . . . . 57**
4.1 Chapter Introduction . . . . . 57
4.2 MEMRES Software Framework . . . . . 60
4.2.1 Pre-sim Processing . . . . . 60
4.2.2 Monte-Carlo Simulator . . . . . 61
4.2.3 Fault-map and Access-map . . . . . 63
4.2.4 Basic Operations . . . . . 66
4.2.5 Modeling . . . . . 68
4.2.6 Trade-off between Accuracy and Speed . . . . . 76
4.3 Framework Validation . . . . . 77
4.4 A Study of STT-RAM using MEMRES . . . . . 82
4.5 Chapter conclusion . . . . . 88

**5 Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory . . . . . 89**
5.1 Chapter Introduction . . . . . 89
5.2 Modeling and Simulation . . . . . 90
5.3 Scalability . . . . . 95
5.4 MRAM Cell Design and Variation . . . . . 96
5.5 Write Error Rate of MRAM . . . . . 98
5.5.1 Write Error Rate of MTJs without Variation . . . . . 98
5.5.2 Write Error Rate of MRAM Array . . . . . 99

5.6 Circuit-level Evaluation . . . 103
5.6.1 MRAM Write/Read with PWSA Multi-write Design . . . 103
5.6.2 Failure Analysis and Error Correction . . . 108
5.6.3 Latency, Energy, and Area of a 16-MB MRAM Bank . . . 111
5.7 Chapter Conclusion . . . 112

**6 MTJ Variation Monitor for Adaptive MRAM Write and Read . . . 113**
6.1 Chapter Introduction . . . 113
6.2 Write Error and Read Disturbance Rates under Variation . . . 114
6.3 MTJ based Variation Monitor . . . 116
6.3.1 Sensing Principle . . . 116
6.3.2 Circuit Implementation and Simulation . . . 117
6.4 Adaptive Write . . . 121
6.4.1 Adaptive Write Scheme . . . 121
6.4.2 Adaptive Write using Variation Monitor . . . 122
6.5 Adaptive Read . . . 125
6.5.1 Adaptive Sensing Circuit using Multiple Reference Resistance . . . 126
6.6 Chapter Conclusion . . . 129

**7 Negative Differential Resistance-Assisted Resistive NVM for Efficient Read and Write Operations . . . 130**
7.1 Chapter Introduction . . . 130
7.2 Issues of MRAM Write and Read . . . 131
7.2.1 Wasted Power During Write Cycles . . . 131
7.2.2 Read Margin and Read Disturbance Limits . . . 132
7.3 NDR Device Characteristics . . . 132
7.3.1 Tunneling-Based V-NDR Devices . . . 132
7.3.2 CMOS Circuit for V-NDR Generation . . . 133

7.3.3 CMOS Circuit for C-NDR Generation . . . 134
7.4 Behavior of Series-Connected MTJ and V-NDR Devices . . . 135
7.5 Modeling of V-NDR Devices . . . 138
7.5.1 Tunneling V-NDR Modeling . . . 138
7.5.2 Analytical Model of CMOS V-NDR Behavior . . . 139
7.6 V-NDR-assisted MRAM Write and Read . . . 141
7.6.1 STT-RAM Write Energy Reduction . . . 141
7.6.2 STT-RAM Read Assistance using NDR . . . 145
7.6.3 Experimental Validation . . . 148
7.6.4 Array-level Reliability and Performance . . . 151
7.7 V-NDR and C-NDR for MLC ReRAM Programming . . . 158
7.8 Chapter Conclusion . . . 160

**8 Hybrid VC-MTJ/CMOS Non-volatile Stochastic Logic for Efficient Computing . . . 162**
8.1 Chapter Introduction . . . 162
8.2 Overview of SC . . . 163
8.3 VC-MTJ and V-NDR in SC . . . 164
8.3.1 Non-destructive VC-MTJ Read with NDR . . . 165
8.3.2 Deterministic VC-MTJ Write with NDR . . . 166
8.4 VC-MTJ based Operations in Stochastic Computing . . . 167
8.5 Stochastic Bit Stream Generator . . . 170
8.6 Evaluation . . . 171
8.7 Chapter Conclusion . . . 173

**9 Conclusion . . . 174**

**References . . . 176**

# List of Figures


1.1 STT-MTJ and VC-MTJ are switched through current and voltage respectively. . . . 3

1.2 Density and write time of memories. . . . 6

1.3 (a) Original stack of interconnect layers on top of device layers in SOI process (b) New stack with $M_{-1}$ and $V_{-1}$. . . . 6

1.4 (a) Load line for V-NDR-MTJ series circuit (illustrated in inset). Blue and red line correspond to the MTJ HRS and LRS respectively. The stable operating points when the MTJ is in HRS and LRS are indicated by ① and ②, respectively. (b) Diagram of proposed 3T CMOS circuit for generating V-NDR between IN and OUT terminals. . . . 10

1.5 (a) One example of C-NDR comprising of a Schmitt trigger and an NMOS. The Schmitt trigger has two threshold voltages $V_{TH}$ and $V_{TL}$. (b) The I-V of the C-NDR. When VNDR starts from 0, then the current of C-NDR remains low until VNDR reaches $V_{TH}$, the current suddenly rises until VNDR reduces below $V_{TL}$. The Schmitt trigger is supplied with low VCC to reduce leakage. . . . 11

2.1 Overview of PROCEED framework. . . . 15

2.2 Typical logic path depth distribution and logic path delay extracted from a synthesized CortexM0. . . . 16

2.3 (a) Example of simulation block allocation in PROCEED based on logic depth. (b) Circuit schematic used for simulation and optimization. . . . 17

2.4 NAND gate (a) schematic and layouts for (b) CMOS and (c) TFET. . . . 17

2.5 Cell area of CMOS-based and TFET-based NAND gates as a function of gate width. . . . 18

2.6 (a) Cell area and (b) interconnect load as a function of transistor width. In (b), transistor width is the same in Inverter and NAND gate. . . . 19

2.7 Model fitting for simulation blocks delay and power as a function of $V_{dd}$. 21

2.8 Optimizer overview. Adaptive weight is chosen by slope of existing fronts. Based on starting point, meta-modeling is built and gradient descent is used to find potential points. Simulate potential points to get new Pareto points. . . . . 24

2.9 Pareto curves for delay and power as evaluated using a commercial synthesis tool, Model [FHS06], and PROCEED. $V_{dd}$ and $V_t$ are constants and only gate size is a variable. . . . . 29

2.10 45 nm SOI CortexM0 (a) power-minimum working clock period and (b) area-minimum working clock period as tuning parameters are increased. 30

2.11 45 nm TFET and boosted TFET (3X current) Pareto curves of delay and chip area for the CortexM0 design. The red curves show the results using a hypothetical TFET with 3X current boost. . . . . 31

2.12 (a) LDH of MIPS and CortexM0. (b) Power and delay curves and (c) area and delay curves for MIPS and CortexM0 designed with TFET and SOI, respectively. Activity is 1% and one $V_{dd}$, one $V_t$ and two bins are applied. 33

2.13 Activity impact on 45 nm SOI and TFET CortexM0s power-minimum working clock period. . . . . 34

2.14 45 nm SOI and TFET CortexM0 microprocessors with power management. The ratios of average to peak throughput are 10%, 20%, 50% and 100%. Curves with ratios of 100% are designs outputted from Pareto optimizer. 35

2.15 Variation-aware (a) power-delay and (b) area-delay evaluations of 45 nm technologies. Assumed voltage drop is 90% and $V_t$ shift is 50mV. . . . . 36

2.16 (a) Power-delay optimization for 45 nm HGI and non-HGI MIPS and (b) its corresponding design area. The fluctuations in the latter arise because the optimization is carried out for power and delay, not design area. Two sets of $V_{dd}$ and $V_t$ are adjusted during optimization, one for SOI devices and the other for TFETs. . . . . 37

3.1 Overview of the variability evaluation framework used in this paper. The evaluation of (left) 6T SRAM cells and (right) microprocessor circuits are divided into two vertical branches as illustrated. . . . . 41

3.2 Simulated 32-nm IM and junctionless FinFETs with LER and RDF. $H_{fin}$ = 10 nm and $\sigma_{LER}$ = 1 nm are used in the above structures. . . . . 43

3.3 Threshold voltage variation of IM and junctionless FinFETs due to LER (upper row) or RDF (bottom row). Only one source of variability (LER or RDF) is active at a time. Note the scale for JL FinFETs is larger than that for IM FinFETs. . . . . 44

3.4 Matching of baseline FinFET (a) transfer and (b) output curves between TCAD simulation and compact modeling. . . . . 46

3.5 Comparison of $\sigma I_{ON}$ and $\sigma V_{T,sat}$ extracted from 200 samples between TCAD simulations and fitted variability models for (a) JL FinFETs and (b) IM FinFETs show a good fit. . . . . 48

3.6 Nominal SNM as a function of working $V_{cc}$ for high density design JL FinFET 6T SRAM cells. Note that for successive technology nodes, SNM and $V_{cc,min}$ decrease when the other is held fixed. . . . . 49

3.7 $V_{ccmin}$ as a function of technology node and LER amplitude for JL and IM FinFET 6T SRAM. The SNM constraint is 100 mV, and yield is 99%. . . 50

3.8 $V_{ccmin}$ as a function of technology node and LER amplitude for JL and IM FinFET 6T SRAM. The SNM constraint is 50 mV, and yield is 99.9%. . 50

3.9 (a) Nominal clock period and clock period increase (mean shift and variation) and (b) nominal leakage power and leakage power increase (mean shift and variation) due to LER variation ($\sigma_{LER}$ = 0.6 nm) for IM and JL FinFET-based MIPS processors at typical clock speeds. . . . . 53

3.10 (a) Increase in clock period mean and (b) variation of critical clock period as a function of technology node and LER amplitude for JL and IM FinFET circuit benchmark (Cortex M0). . . . . 54

3.11 (a) Increase in leakage power mean and (b) variation of leakage power as a function of technology node and LER amplitude for JL and IM FinFET circuit benchmarks (Cortex M0). . . . . 54

4.1 The limitation of analytical models and FaultSim on memory reliability evaluation. The reliability enhancement techniques in gray boxes can only be evaluated by MEMRES. . . . . 58

4.2 The framework overview of MEMRES. . . . . 60

4.3 Memory reliability simulation. The simulation divides years-long memory lifetime into short intervals. In each interval, events of random fault injection, ECC checks, and memory reliability management are simulated. The simulation terminates when an uncorrectable fault occurs or it reaches the end of preset simulation time. . . . . 61

4.4 Memory architecture and memory fault types. A bank is constructed by columns and rows (several rows are grouped in a mat, which is not shown in the figure). Eight banks are built in a chip (device), and nine x8 chips (8 data chips + 1 ECC chip) or eighteen x4 chips (16 data chips + 2 ECC chips) construct a rank. Several ranks are built in a channel. In a write or read operation, a channel and a rank is firstly selected by a decoder, and then a word is written/read across all chips in the selected rank, e.g., every x8 chip in a rank outputs 8 bits to comprise a 72-bit word (64 data bits + 8 ECC bits), where all chips in a rank share same addresses. A fault type is defined by the component that is affected, e.g., a bank fault indicates multiple faulty rows/columns in the bank. . . . . 62

4.5 Examples of FM/AMs A, B, And C. A is a column FM/AM, B is a 4-bit FM/AM, and C is a single-bit FM/AM. . . . . 64

4.6 The basic operations used in MEMRES: (a) INTERSECT, (b) MERGE, and (c) REMOVE. Cover-Rates of A and B are 0.6 and 0.5 respectively. . 65

4.7 A memory with in-memory SECDED and in-controller ECC. In a memory chip, every eight columns of data bits are protected by one column of ECC bits. The in-memory ECC logic can correct a single-bit error in a 72-bit ECC word (64 data and 8 ECC bits), where a burst length of 8 accesses is required for a x8 chip to have 72 bits together for in-memory SECDED correction. A x8 chip inputs/outputs 8 data bits in an access, and totally eight x8 data chips and one x8 ECC chip input/output 72 bits from/to the memory controller in an access, where data errors in the 72 bits can be corrected/detected by the in-controller ECC. . . . . 70

4.8 An example Mask of a column FM in the memory specified by Table 4.2. A read contains 4 bits from a chip, 18 reads construct a 72-bit word (64 data bits and 8 ECC bits), and a word has a memory physical address. . 78

4.9 Validation of MEMRES with FaultSim and the analytical model [JDB13]. The failure rates for a 4-GB DRAM with SECDED as functions of time are shown. 1x and 4x fault rates are used. MEMRES matches well with the analytical model and FaultSim. . . . . 80

4.10 The failure rate for a 4-GB DRAM with SECDED or SCCDCD as a function of time. The single-bit transient error rate (SBTER, i.e., DRAM only has data-link error as SBT in the validation) of $10^{-14}$ and $10^{-10}$ are used in this validation. . . . . 81

4.11 The memory failure rate breakdown for a 4-GB DRAM operating for 5 years with in-controller SECDED. The data-link BER of $10^{-14}$ and $10^{-10}$ are used in this validation. . . . . 81

4.12 (a) An 8-GB STT-RAM with unbalanced memory access (without interleaving) and balanced memory access (with interleaving). (b) Failure rates (CDF) of STT-RAM with unbalanced and balanced memory access. In-controller SCCDCD and in-memory SECDED are enabled. . . . . 84

4.13 The 5-year failure breakdown and failure rate (CDF) of STT-RAM with enabled rank sparing. The thresholds (percentage of faulty addresses in a rank) to trigger rank repairing are 0.1%, 0.001%, and 0.00001%. The STT-RAM has one spare rank in each channel. In-controller SCCDCD is enabled. . . . . 85

4.14 The 5-year failure breakdown and failure rate (CDF) of STT-RAM with enabled memory page retirement. Different maximum allowed retired pages per channel are tried, including 20, 2000, and 200000. In-controller SCCDCD is enabled. Memory page size is 4kB. . . . . 86

4.15 The 5-year failure breakdown and failure rate (CDF) of STT-RAM with memory mirroring. Different mirrored memory space are simulated including whole memory mirroring (one channel is mirrored to the other one), a half memory mirroring (one rank is mirrored to another one), and a quarter of memory mirroring (a half rank is mirrored to another half). In-controller SCCDCD is enabled. . . . . 86

4.16 (a) Varying fault FIT rate (normalized to constant fault FIT rate) and constant FIT rate vs. time. (b)The failure rate (CDF) of STT-RAM with varying fault FIT rate and constant fault FIT rate (listed in Table. 4.3). In-controller SCCDCD and in-memory SECDED are enabled. . . . . 87

5.1 (a) VC-MTJ is switched by unidirectional voltage pulses. The first two same pulses switch the resistance state of a VC-MTJ from P to AP and then back to P, the third double-width pulse make two switches continuously. (b) STT-MTJ is switched by directional current pulses, and the switching directions depends on the direction of current. . . . . 90

5.2 VCMA-induced precessional switching. When a voltage is applied on the VC-MTJ, the energy barrier separating the two magnetization states of the free layer is reduced so that the magnetization state starts to spin. . 91

5.3 During a write of the STT-MTJ, VCMA may assist the thermal activation to cause unintended switching. This effect can improve the switching probability when the write pulse width is insufficient to switch the STT-MTJ, on the other hand, may lead to switching failure when the write pulse width is sufficiently long. . . . . 92

5.4 Layouts of STT-RAM and MeRAM under 32nm design rules. The area of an STT-RAM cell is twice the area of a MeRAM cell, as an STT-MTJ requires a 3X wider access transistor than a VC-MTJ. Vertical transistor like nanowire may help to reduce area inefficiency [WG14a] . . . . . 96

5.5 WER of the nominal STT-MTJ as a function of pulse width for different perfect current pulses (constant current) and switching directions. . . . . 98

5.6 The WER of the nominal VC-MTJ as a function of pulse width for different perfect voltage pulses and switching directions. A VC-MTJ has an optimal pulse, which leads to the lowest WER. The curve of 1.2V has the lower overall WER than 1.1V and 1.3V, indicating 1.2V is closer to the optimal voltage. . . . . 99

5.7 (a) Write current (voltage) pulse on STT-MTJs (VC-MTJs). The rise and fall time are measured by the time while voltage is rising and falling between 10% and 90% of the peak voltage respectively. Mean of write current on (b) STT-MTJs and (c) VC-MTJs as a function of MTJ resistance. 100

5.8 A Monte-Carlo simulation flow to obtain the WER of 1T1M MRAM array. N is the sample size, and T is the simulation time including a writing time and a waiting time (for the MTJ to settle down, e.g., waiting time is 20ns in the simulations). . . . . 102

5.9 The WER of an STT-RAM under process and temperature variation for different write pulses and switching directions (a: P to AP, b: AP to P). 103

5.10 WER of an MeRAM under process and temperature variation for different write pulses. The WER is averaged over two switching directions. . . . . 103

5.11 Data program flow for the PWSA multi-write design. . . . . 104

5.12 Expected word-write energy and latency for MeRAM and STT-RAM with PWSA multi-write circuit. The word size is 256bits, which are the number of bits being simultaneously written. The WER/bit after multiple writes is minimize below $10^{-23}$. The labels 3ns, 6ns, and 9ns on STT-RAM are single write pulse widths. The pre-read is not mandatory for STT-RAM. The top circled designs are STT-RAMs without pre-read operation. The bottom circled ones are STT-RAMs with pre-read operation, which count the overhead of pre-read but save unnecessary write. . . . . 106

5.13 Impact of word size and temperature on the expected write latency and energy of MRAMs. All bars are normalized within each group to the expected write latency of MRAM at 300K with 64-bit word size. . . . . 107

5.14 Read disturbance rate as a function of read voltage. The read disturbance rate for MeRAM and STT-RAM are extrapolated to the read voltage drop on MTJs (0.48V and 0.15V are respectively for VC-MTJs and STT-MTJs). 110

6.1 (a) The STT-MRAM P-to-AP WER as a function of write pulse width under different $t_{FL}$ and temperature corners. In STT-MRAM, P-to-AP switching is more difficult and dominates write latency. (b) The average AP-to-P and P-to-AP WER of MeRAM as a function of write voltage. . 114

6.2 The STT-MRAM P-to-AP RDR as a function of write pulse width under different $t_{FL}$ and temperature corners. In STT-MRAM, P-to-AP is selected as the read current direction due to less spin polarization efficiency. . . . 115

6.3 The experimentally measured retention time as a function of stress voltage on MTJs. . . . . 116

6.4 The schematic of STT-MRAM and MeRAM based variation monitor. Variation monitoring operations: 1) apply stress voltage/current on MRAM monitor array controlled by stress voltage/current selection circuit; 2) select every MTJ (controlled by MTJ selection circuit) one by one to read and count MTJ switching rate (controlled by sensing and switched MTJ counting circuit). . . . . 117

6.5 (a) Different stress current/voltage in the proposed monitor. (b) Simulated waveforms of read, reset and counting operations. . . . . 119

6.6 Switching rate of (a) STT-MTJ- and (b) VC-MTJ-based variation monitor under different stress current and voltage respectively. The color lines are switching rate for only temperature variation ($10^oC$ interval). The dot lines outline standard deviations ($\sigma$) of thermal activation rate ($\sigma$ is caused by process variation and random thermal activation). . . . . 119

6.7 Adaptive write scheme using the MTJ-based variation monitor or conventional thermal monitors. . . . . 121

6.8 Optimal write pulses for (a) STT-MRAM and (b) MeRAM under different $t_{FL}$ and temperature corners. . . . . 122

6.9 (a) Evaluation flow of adaptive write in MRAM based system. (b) The cross-section structure for thermal simulations. . . . . 123

6.10 The maximum and average write latency in (a) 1MB STT-MRAM L2 and (b) MeRAM L2 from 270K to 370K under different $t_{FL}$ corners with different number of write pulse choices. . . . . 124

6.11 The average/maximum run time of SPEC benchmarks using adaptive write (with three write pulse choices) for (a) one-core processor with single-level 8-MB STT-MRAM cache and (b) single-level 8-MB MeRAM MeRAM cache, a dual-core processor with (c) 1-MB STT-MRAM L2 and 16-MB STTRAM L3, and (d) 1-MB MeRAM L2 and 16-MB MeRAM L3 over temperature corners (270K to 370K). Run time is normalized to the maximum run time for processors without adaptive write (one write pulse choice) for each benchmark. . . . . 124

6.12 Read disturbance rate as a function of voltage drop on P MTJ for a set of temperature and free layer thickness variation corners. The read disturbance rate is obtained from Monte-Carlo simulation with sensing time of 3ns. . . . . 126

6.13 Illustration for using two reference resistance for low and high temperature sensing. . . . . 127

6.14 Sensing circuit used in the adaptive read. The switch of two reference resistors are controlled by temperature sensor. The switch of read voltage is controlled by the proposed monitor with a output . . . . . 128

6.15 (a) Conventional STT-RAM sensing simulation with single read supply voltage and one reference resistance. (b) The proposed STT-RAM sensing simulation with two read voltage and two reference resistance. . . . . 129

7.1 STT-RAM write error rate as a function of write time assuming 0.7 V write voltage extracted from 10 billion Monte Carlo circuit simulations using the methodology of Section 7.6.4. The simulated write circuit includes a MTJ, an access transistor, and the capacitance load of 256 1T1M bit-line. . . 131

7.2 a) Schematic $I - V$ for typical V-NDR device. b) $I_d - V_d$ of analytical TFET model and simulated device data of [LLZ12]. For the TFET model, parameters are $A_{TFET}$ = 1.3E-8 A/um, $B_{TFET}$ = 4E6 eV/cm, $E_g$ = 0.74 eV, $\lambda$ = 6E-7 cm, $A$ = -0.02, $B$ = 0.0456, $C$ = 0.04, $n$ = 0.3, and $D$ = 0.0025.134

7.3 (a) Diagram of proposed 3T CMOS circuit for generating V-NDR between IN and OUT terminals. (b)Experimental $I$-$V$ curves for CMOS V-NDR circuit; current is measured at IN terminal as a function of voltage drop $V_{NDR}$ across the IN and OUT terminals. Different curves correspond to application of $V_{bias}$ between 0.8 and 1.3V. The circuit is constructed using preexisting long channel NMOS devices with gate lengths of 10 $\mu$m, W/L ratios from 4 to 10, and off-currents around 0.1-1 nA and on-currents ranging between 10-100 $\mu$A. . . . . 135

7.4 (a) One example of C-NDR comprising of a Schmitt trigger and an NMOS. The Schmitt trigger has two threshold voltages $V_{TH}$ and $V_{TL}$. (b) The I-V of the C-NDR. When VNDR starts from 0, then the current of C-NDR remains low until VNDR reaches $V_{TH}$, the current suddenly rises until VNDR reduces below $V_{TL}$. The Schmitt trigger is supplied with low $V_{CC}$ to reduce leakage. . . . . 136

7.5 (a) Series connection of MTJ and V-NDR device. Note that each V-NDR device is shared by multiple bit-lines in the proposed design. (b) Load line for V-NDR-MTJ series circuit (illustrated in inset). Blue line corresponds to the MTJ HRS and red line to the MTJ LRS. The stable operating points when the MTJ is in HRS and LRS are indicated by ① and ②, respectively. 136

7.6 (a) $I_{NDR}$ versus $V_{NDR}$ (solid lines) and $I_{MTJ}$ versus $(V_{CC} - V_{NDR})$ (dashed lines) for different $R_{MTJ}$. (b) Current of the series connection of MTJ and V-NDR vs. $R_{MTJ}$. . . . . 138

7.7 Model and SPICE comparison of 3T V-NDR circuit using 45 nm commercial CMOS library for (a) $I$ versus $V_{in}$ and (b) $V_{int}$ versus $V_{in}$. . . . . 140

7.8 Simulated peak current points of V-NDR designs guided by model and their margins for three MRAM applications. The error bars illustrate allowed design margins. . . . . 141

7.9 The proposed V-NDR read and write circuitry designs. Yellow dotted line denotes the STT-RAM array. . . . . 142

7.10 SPICE simulated waveforms for a write-1 termination in the memory design of Fig. 7.9. During write operation, a bit-line is first selected and charged to $V_{CC}$, then a MTJ bit is selected by $WL$, and write current is high or low depending on the MTJ initial state. . . . . 143

7.11 (a) Three early write termination designs using bit-line voltage change sensing [ZZY09], current change sensing [BOE16], and the proposed NDR. (b) Simulated waveforms of MTJ resistance (AP: 5000 Ω, P: 2000 Ω, bit-line voltage, and write current as functions of time. The black line is for the conventional write, and the read dash line is for the early write terminations. 144

7.12 SPICE simulation of read operations using NDR-assisted design in Fig. 7.9. The discharging current ($I_{NDR}$) difference for sensing MTJ states is significantly increased by the high PVR of NDR. A large and constant voltage margin is achieved on the bit-line ($V_{bit-line}$), which is sensed by a constant reference voltage leading to a stable sense amplifier output ($V_{output}$). 147

7.13 Experimental MTJ resistance ($R_H$ = 320 kΩ, $R_L$ = 240 kΩ) as a function of external magnetic field. . . . . 149

7.14 Current through series-connected V-NDR and MTJ as external field is cycled to switch the MTJ from AP to P and back. . . . . 149

7.15 Time-dependent measurement of MTJ-V-NDR current for MTJ initially in AP state and with external magnetic field of -1.34 kG biasing the device in the bistable region. $V_{CC}$ = 0.6 V for this measurement. Note the drastic reduction in current through the circuit around 7.5 ms from 1.2 $\mu$A to 25 nA due to switching of the MTJ from the AP to P state. . . . . 150

7.16 (a) Simulated circuit implementation in CUDA. (b) V-NDR peak current margin variation analysis. . . . . 152

7.17 Simulation results with transistor and MTJ process variations. V-NDR characteristics are varied by scaling diameter for TDs (0.4-0.55 µm) and threshold voltage for TFETs (assuming device width=1.95 µm in write and 0.195 µm in read circuitry).(a) WER and write energy versus nominal V-NDR peak current. Write energy includes bit-line pre-charge, access transistor, and MTJ, but excludes row/column decoders. (b) Read margin versus V-NDR nominal peak current. $C_{BL} = C_{SL}$ = 25 fF in (a) and (b). (c) Read disturbance rate as function of bit-line/source-line load and read margin. High/low margin designs are obtained using different $V_{read}$ (0.35 V/0.25 V for TD, 0.3 V/0.21 V for TFET, and 1.8 V/0.7 V for conventional designs). Read disturbance rates below $10^{-10}$ not detectable within sample size. . . . . 153

7.18 Simulated write energy (normalized to conventional write scheme) and WER (right axis) vs. threshold voltage shift of T3 for 25 fF bit-line load (256 1T-1M cells per bit-line). The WER is extracted from 10 billion Monte-Carlo numerical simulations for V-NDR assisted STT-MRAM write. MTJ device parameters can be found in Table 7.5. . . . . 156

7.19 Simulated STT-MRAM read margin vs. T3 threshold voltage shift. MTJ device parameters can be found in Table 7.5. . . . . 156

7.20 Simulated read disturbance rate vs. bit-line size (load) for read design with and without V-NDR. Larger load leads to more pre-charging/dis-charging current and more read disturbance. . . . . 157

7.21 Using multiple V-NDRs and C-NDRs to program ReRAM cell resistances. In the programming, V-NDR can decrease a cell resistance from high value to a low value that is determined by V-NDR peak current, while C-NDR can program a cell resistance in the reverse direction. Once cell resistance achieves target resistance, both V-NDR and C-NDR can terminate write immediately . . . . . 158

7.22 (a) I-V curves of programming a ReRAM cell resistance using three different sized V-NDR. (b) I-V of programming ReRAM cell resistance using three different sized C-NDR devices (red dashed lines). . . . . 159

7.23 (a) I-V curves of programming a ReRAM cell resistance using three different sized V-NDR. (b) I-V of programming ReRAM cell resistance using three different sized C-NDR devices (red dashed lines). . . . . 160

8.1 Multiplication and addition using unipolar and bipolar encoded SBS. Unipolar coding represents decimal number ranging in [0,1], while bipolar coding is for decimal number in [-1,1]. The SC computations are bit-wise, where corresponding bits in two input SBS are operated using the AND, MUX, or XNOR gates. . . . . 163

8.2 Simulated switching probability (a) and switching error rate (b) as functions of pulse width for different write voltages using an experimentally verified model in [WLE16b, GLL16a]. . . . . 164

8.3 (a) NDR-assisted switching and read. (b) Simulated waveforms of a NDR-assisted read. . . . . 166

8.4 Switching error rate ($1 - P_{switching}$) of NDR-assisted VC-MTJ write from HRS to LRS. The $LRS \rightarrow HRS$ switching rate is $< 10^{-10}$, which is not shown. . . . . 166

8.5 VC-MTJ and V-NDR built SC logic operations. Where a long low-voltage pulse is used to randomize VC-MTJ states, and a short high-voltage pulse is used to switch VC-MTJ states. Every VC-MTJ array stores an SBS. All VC-MTJs in an array are computed simultaneously for throughput and design efficiency purpose. Please note that the XNOR gate directly changes the array Y's data to $\overline{X \oplus Y}$. . . . . 167

8.6 Removing correlation using shuffle operation for $SBS^2$. . . . . 168

8.7 (a) The schematic of a pipe-lined SBS generator with binary fraction input A[n-1:0], B[n-1:0], and C[n-1:0]. $n$ VC-MTJ arrays (every one stores an SBS) are divided into even and odd groups based on the index. At every cycle (10ns), one group is written according to the read-out of the other group. Output SBS is generated every two cycles because of the pipe-line. (b) Simulated waveforms of two-cycle SBS generation. The MTJs in $SBS_0$ are written in the first cycle (*write_even*: high, while the MTJs in $SBS_1$ are written in the second cycle (*write_odd*: low) with the read-out from $SBS_0$. . . . . 168

8.8 An SBS generation example. The input 0.101 (binary) is translated to SBS (01101011). . . . . 170

8.9 (a) Computing energy of 8-tap FIR for fix-point width from 5-bit to 8-bit (32-bit to 256-bit for uni-polar encoded SC). (b) Computing energy of 32-classifier Adaboost with 32-pixel input image for fix-point width from 5-bit to 8-bit (32-bit to 256-bit for bipolar encoded SC). The wire activity is 0.375 for both CMOS binary and SC designs, and 1 for VC-MTJ and V-NDR based SC design. Energy are shown for two categories: energy with SBSG (W/ SBSG) and energy without SBSG (W/O SBSG). . . . . 172

# List of Tables

2.1 Comparison of variables considered in benchmark methodologies in the literature. P-D : power-delay, A-D: area-delay, A-P: area-power, $\mu P$: micro processor, LDH: logic depth histogram, CPI: clock cycles per instruction, SS: subthreshold swing . . . . . 14

3.1 Allowed tuning range of fitted compact model parameters. [1] Parameters in PTM model. . . . . 47

3.2 Nominal SNM and SNM loss from variability for JL FinFET technologies.[1] High density 6T SRAM design. [2] Symmetric N/P design. [3] SNM at $V_{cc}$=0.73 V. [4] SNM with 99% yield constraint: LER variation ($\sigma_{LER}$=0.6 nm ) at $V_{cc}$=0.73 V . . . . . 51

3.3 Circuit benchmarks. . . . . 53

3.4 Average mean shift and standard deviation of timing and leakage for six benchmark circuits. . . . . 55

4.1 Comparison with existing memory fault analysis methods. Run time is measured using single-core on AMD Opteron(tm) Processor 2380. FaultSim has event mode, which uses analytical models to partially replace Monte-Carlo fault injection in regular interval mode to speedup simulations. 59

4.2 Architecture of a 4-GB DRAM DIMM. . . . . 78

4.3 Fault FIT rates per chip and data-link BER for DRAM and STT-RAM. The retention BER (RER) and write BER (WER) are only for STT-RAM. Data-link error, retention error, and write error are single-bit transient errors (SBT). . . . . 79

4.4 The 5-year failure rate for STT-RAMs with different write error rate (WER, per-bit-write failure probability), retention error rate (RER, per-bit-hour failure probability), and ECC designs: 1) in-controller SCCDCD with (W/) in-memory SECDED, 2) in-controller SCCDCD without (W/O) in-memory SECDED. . . . . 83

5.1 Modeling parameters at 300K. . . . . 94

5.2 Design parameters for MTJs and access transistors. The transistors' threshold voltage variation considers the effects of line edge roughness (LER), random dopant fluctuation (RDF), and non-rectangular gate (NRG). Access transistors of MeRAM have larger threshold voltage variation because narrow transistors are affected more by NRG, RDF, and LER. . . . . 97

5.3 Summary of write pulse variation due to transistor process variation at temperature of $300^oC$ and $350^oC$. Mean shift is the percentage change of parameters' mean between high and low MTJ resistance states . . . . . 101

5.4 Energy and delay for operations in the PWSA multi-write circuit at 300K temperature. . . . . 104

5.5 Failure types and FIT for a 16MB memory bank. $10^9$ reads and $10^9$ writes in a bank-hour are assumed. The read disturbance rate is extrapolated from simulations. As a comparison, the FIT of single-bit fault in a 16MB bank is about $2 \cdot 10^{-4}$ [SL12b], and the FIT of DDR bus errors is about 100 [MKS10]. . . . . 108

5.6 Write/read latency/energy for one write/read in a x8 16-MB STT-RAM and MeRAM banks. One write/read operates on 64 bits (72bits in memory banks for in-memory ECC detection and correction) in a row in burst mode. 111

6.1 Comparison between conventional thermal monitors and the proposed variation monitor. The proposed monitor uses 256 MTJs and 10 stress levels 120

7.1 Experimental characteristics of selected V-NDR tunneling devices from literature. Peak current is expressed in terms of per unit width for TFETs and per unit area for TDs. . . . . 133

7.2 NDR and MRAM parameters for three different MRAM read or write design. 141

7.3 Write energy, read margin, and read energy of NDR-assisted designs as extracted from Fig. 7.17. $C_{BL} = 25$ fF; nominal TFET $V_t h = 0.25$ V. Since V-NDR does not affect write-0 operations (*0→0* and *1→0*), conventional designs are used for these cases. Effective PVR is the ratio of circuit current in the 1 and 0 states for chosen $V_{CC}$ and differs for write and read due to different bias. . . . . 154

7.4 Simulation parameters at 300K. . . . . 155

7.5 Variation parameters for STT-MTJs in the simulations of write error rate and read disturbance rate. . . . . 157

8.1 Simulated energy per bit and delay of VC-MTJ-V-NDR based logic operations. Interconnect and fan-out load is considered. . . . . 169

## Acknowledgments

I would like to thank Prof. Puneet Gupta, for his great guidance, encouragement and support in my research and career development.

I would like to thank Prof. Chi On Chui, Dr. Pedram Khalili Amiri, Prof. Kang L. Wang, Prof. Subramanian S. Iyer, and Prof. Yuan Xie for their patience, support, guidance, feedback and cooperation in overcoming numerous obstacles I have faced through my PhD research.

I would like to thank my fellow doctoral students, Andrew Pan, Greg Leung, Hochul Lee, Cecile Grezes, Liangzhen Lai, Wei-che Wang, Yasmine Badr, Mark Gottscho, Abde Ali Kagalwalla, John Lee, Rani S. Ghaida, Saptadeep Pal, Irina Alam, Wei Wu, Min Gao, Juexiao Su, Tianheng Tu, Yunxuan Yu, Xiao Shi, Zhuo Jia, Tianmu Li for their help, cooperation and of course friendship in my 5.5-year UCLA life.

Nevertheless, I am also grateful to my wife, Dr. Xinjie Guo, for her support and collaboration in research and life.

Last but not the least, I would like to thank my parents for their unconditional support.

# Vita

| | |
|---|---|
| 2007–2011 | Bachelor of Science, EECS Department, Peking University, Beijing, China. |
| 2011-2013 | Master of Science, Electrical Engineering Department, University of California, Los Angeles, California |

# Publications

S. Wang, A. Pan, C. Grezes, P. Amiri, K. Wang, C. Chui and P. Gupta, "Leveraging CMOS Negative Differential Resistance for Low Power, High Reliability Magnetic Memory," *Electron Devices, IEEE Transactions on*, (Accepted)

S. Wang, A. Pan, C. Chui, and P. Gupta, "Tunneling Negative Differential Resistance-Assisted STT-RAM for Efficient Read and Write Operations," *Electron Devices, IEEE Transactions on*, 2017

H. Lee, A. Lee, S. Wang, E. Ebrahimi, P. Gupta, P. Khalili, and K. Wang, "A Word Line Pulse Circuit Technique for Reliable Magnetoelectric Random Access Memory," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 2017.

L. Zhu, Y. Badr, S. Wang, S. Iyer, and P. Gupta, "Assessing Benefits of a Buried Interconnect Layer in Digital Designs," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 2017

C. Grezes, H. Lee, A. Lee, S. Wang, F. Ebrahimi, X. Li, K. Wong, J. Katine, B. Ocker, J. Langer, P. Gupta, P. Khalili, and K. Wang, "Write Error Rate and Read Disturbance in Electric-Field-Controlled MRAM," 2017

S. Wang, H. Hu, H. Zheng, and P. Gupta, "MEMRES: A Fast Memory System Reliability Simulator," *Reliability, IEEE Transactions on*, 2016

H. Lee, C. Grezes, S. Wang, E. Ebrahimi, P. Gupta, P. Khalili, and K. Wang, "Source line sensing in magneto-electric random-access memory to reduce read disturbance and improve sensing margin," *IEEE Magnetics Letters*, 2016.

S. Wang, H. Lee, F. Ebrahimi, P. Amiri, K. Wang, and P. Gupta, "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, June 2016

S. Wang, A. Pan, C. Chui, and P. Gupta, "PROCEED: A Pareto Optimization-Based Circuit-Level Evaluator for Emerging Devices," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 2015

G. Leung, S. Wang, A. Pan, P. Gupta, and C. Chui, "An Evaluation Framework for Nanotransfer Printing-Based Feature-Level Heterogeneous Integration in VLSI Circuits," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 2015

S Wang, G. Leung, A. Pan, C. Chui, and P. Gupta, "Evaluation of Digital Circuit-level Variability in Inversion-mode and Junctionless FinFET Technologies,", *Electron Devices, IEEE Transactions on*, 2013

L. Zhang, S. Wang, et. al., "Gate Underlap Design for Short Channel Effects Control in Cylindrical Gate-All-around MOSFETs based on an Analytical Model," *IETE Technical Review*, 2012.

S. Wang, S. Pal, T. Li, A. Pan, C. Grezes, P. Khalili, K. Wang, and P. Gupta, "Hybrid VC-MTJ/CMOS Non-volatile Stochastic Logic for Efficient Computing", in *IEEE Design, Automation & Test in Europe Conference (DATE)*, ***best paper nomination***, March, 2017.

S Wang, H. Lee, C. Grezes, P. Khalili, K. Wang, and P. Gupta, "Adaptive MRAM write through variation monitoring,", in *53rd ACM/IEEE Design Automation Conference (DAC)*, June 2016

S. Wang, A. Pan, C. Chui, and P. Gupta, "PROCEED: A Pareto Optimization-Based Circuit-Level Evaluator for Emerging Devices," in *19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, January, 2014

X. Guo, S. Wang, et. al., "A novel approach to simulate Fin-width Line Edge Roughness effect of FinFET performance," *Proc. IEEE International Conference of Electron Devices and Solid-State Circuits (EDSSC)*, Dec., 2010.

S. Wang, X. Guo, L. Zhang, C. Zhang, F. He, and M. Chan, "Analytical Subthreshold Channel Potential Model of Asymmetric Gate Underlap Gate-all-around MOSFET," in *IEEE International Conference of Electron Devices and Solid-State Circuits (EDSSC)*, Dec., 2010.

S. Wang, et. al., "A potential-based analytic model for monocrystalline silicon thin-film transistors on glass substrates," *Proc. IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, Aug., 2010.

# CHAPTER 1

# Introduction

As traditional silicon (Si) technologies scale ever deeper into nanometer, logic transistors and memory devices are approaching their fundamental limits. One of the main challenge is the emergence of Dark Silicon, which reveals the thermal and power constraints for semiconductor design and drastically limits the benefit brought by technology scaling. Exploring power-efficient emerging technologies is an essential task to continue the technology advancement. Many promising candidate technologies have been proposed. However, their values still require rigorous evaluation. To identify their potential values, technologies must be co-optimized and evaluated with circuit and system applications. My dissertation can be classified into three sections:

- **Emerging technology and circuit evaluation:** evaluating and co-optimizing emerging logic devices, memory devices, integration technologies in the contexts of circuit and system [WLP13, WPC14, WPC15, LWP15b, ZBW16, WHZ15b, WHZ16, WLE16b, GWM10].
- **Emerging memory design and optimization:** developing fast circuit simulation models for emerging memories, and designing circuitry to improve emerging memory reliability and performance [LGW16c, LLW17, WLG16, WPC17, WPG17, GLL16b, WZZ10, WRS17].
- **Non-volatile computing:** designing non-volatile stochastic computing system with emerging memory technologies [WPL17].

## 1.1 Emerging Technology Introduction

With technology scaling, transistor density increases rapidly but supply voltage scales slowly due to the non-scaled threshold voltage. As a result, chip power density grows dramatically, and the elevated temperature degrades transistor electron mobility. A problem called dark silicon emerges, which reveals that the chip power density becomes the bound limiting the performance improvement. Moreover, scaling device to nanoscale faces other challenges like increasing interconnect resistance and routing congestion. Exploring alternative solutions is essential to continue the technology development.

Many emerging technologies have been proposed with potential advantages in energy efficiency. At the device level, emerging logic and memory devices have been proposed to resolve the dark silicon problem, including tunneling field-effect transistor (TFET), III-V FET, Junction-less (JL) FET, spin-transfer torque (STT) magnetic tunnel junction (MTJ), voltage-controlled (VC) MTJ, and charge-trapping memory. At the technology level, the novel heterogeneous integration of different materials is an attractive technology enabling energy-efficient computation. In this section, a brief introduction is given regarding examples of emerging logic and memory devices.

### 1.1.1 Emerging Boolean Logic Devices

#### 1.1.1.1 III-V Compound Transistors

III-V compound transistors comprise of two or more elements from group III (boron, aluminium, gallium, indium) and group V ( nitrogen, phosphorus, arsenic, antimony, bismuth). These transistors have characteristics of faster speed and/or lower capacitance compared with traditional Si transistors. For examples, InGaAs based transistors have high driving capability and low capacitance for its high mobility and conduction band density of states.

#### 1.1.1.2 Tunneling Field-effect Transistor (TFET)

TFET [IR11] has different switching mechanism from traditional MOSFET, even though their structures are similar. The conducting current of TFET is due to quantum tunneling

through an energy barrier. The energy barrier is modulated by its gate voltage, which substantially controls the current density and the TFET switching. TFET is supposed to have lower subthreshold slope (SS) and lower leakage power compared with traditional Si CMOS, allowing TFET to operate at lower supply voltage with lower power consumption.

#### 1.1.1.3 Other Emerging Boolean Logic Devices

Compared with traditional Si FETs, carbon nanotube (CNT) FET has higher electron mobility and high current density, gate-all-around (GAA) FET has better gate-control capability and higher SS, and Junctionless (JL) FET has lower fabrication cost.

### 1.1.2 Emerging Memories

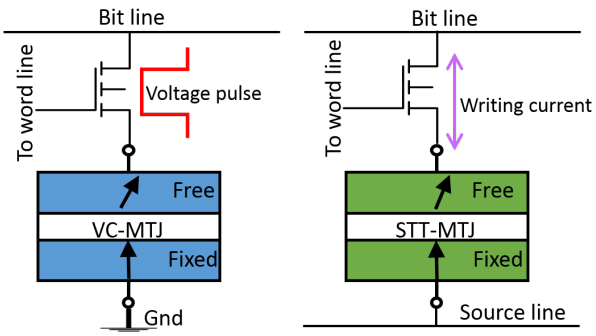



Figure 1.1: STT-MTJ and VC-MTJ are switched through current and voltage respectively.

#### 1.1.2.1 Spin-transfer Torque Magnetic Tunnel Junction (STT-MTJ)

STT-MTJ [Hei01, WHA11] (Fig. 1.1) is a non-volatile storage device that potentially promises the speed and density of dynamic random access memory (DRAM). STT-MTJ has two resistance states (high resistance and low resistance), which are determined by magnetization directions of the free and the fixed layers. A low ("1") and high ("0") resistance are present when magnetic directions are parallel ($P$ state) or anti-parallel ($AP$ state) respectively. The resistance difference is quantified by tunnel magnetoresistance (TMR, defined as $(R_H - R_L)/R_L$), where TMR of 180% has been demonstrated in a 8Mb STT-MRAM chip [SLa16].

The magnetization in free layer is switched by STT current. STT-MTJ can potentially achieve less than 10 ns switching time and 100 fJ/bit switching energy. Therefore, spin-transfer torque RAM (STT-RAM) designed with STT-MTJs is identified as a possible replacement of current memory technologies, such as static RAM (SRAM) Cache [SBL11, SMN11, XSW11, JMX12] and DRAM main memory [KKS13].

### 1.1.3 Voltage-control Magnetic Tunnel Junction (VC-MTJ)

VC-MTJs (Fig. 1.1) [AUA13, AAU12, DAC13, KYI12, SMN12, SNB12, WLH12] with voltage-controlled magnetic anisotropy (VCMA) provides more promising performance than STT-MTJs. This technology allows for precessional switching, a process which provides flipping of the magnetization upon a voltage pulse, irrespective of the initial state. It enables the use of minimum sized access transistors, as well as precessional switching to simultaneously achieve low energy, high density and high speed magnetoelectric random access memory (MeRAM). MeRAM reduces switching energy due to reduced ohmic loss (10 fJ/bit for the VC-MTJs with over 100X higher resistance than STT-MTJs).

Both STT-MTJ and VC-MTJ suffer from the reliability problem of intrinsic switching failure caused by thermal fluctuation exacerbated by process variation [LAS08] and temperature variation. This problem can be quantified by write error rate (WER), which is the average number of switching failures per write. STT-RAM can simply reduce the WER by using high current and long write time. By contrast, MeRAM does not have a trivial solution, because every VC-MTJ has an optimal write pulse giving the lowest WER, and the effect of variation on the optimal pulse is less straight forward.

#### 1.1.3.1 Resistive Random Access Memory (ReRAM)

ReRAM [SSS08] works by changing the resistance through a dielectric layer. There are many mechanisms to change the dielectric resistance, for examples, moving oxygen vacancies. In multi-level-cell (MLC) ReRAM, more than two resistances are utilized to store data. In a ReRAM programming, the resistance change is fast in the beginning and gets slow later (saturated). The programming speed exponentially depends on voltage [STM11, AGH12], where a higher voltage is more efficient in terms of energy and speed

but not the accuracy [AGH12]

#### 1.1.3.2 Phase Change Memory (PCM)

PCM [WRK10] presents low and high resistance states in crystalline and amorphous phases respectively. Generally, quick high-temperature heat and quench can change the crystalline phase to crystalline phase. A relative longer low-temperature heat can gradually switch it to amorphous state. Multiple resistance states can be achieved by creating partial amorphous phase.

#### 1.1.3.3 Flash Memory

Flash memory can be classified to NAND-flash [LHC02] and NOR-flash. NAND-flash has higher density and lower cost, but does not allow random access. NAND-flash is suitable for storage. By contrast, NOR-flash provides random access address but has lower density. NOR-flash is usually used in applications where processors can directly execute code stored in it. Though Flash memory is not an emerging technology, recently, modified NOR-flash has shown high efficiency in emerging analog computation like multiplication and addition, which is suitable to implement emerging applications like deep neural networks [GMG15, GBP17, BGK16].

#### 1.1.3.4 Memory Density and Programming Time

Fig. 1.2 shows density and programming time of emerging memories and existing commercial memories (hard disk drive (HDD), NAND flash, Static RAM (SRAM)). These results are surveyed from published papers.

### 1.1.4 Emerging Integration Technologies

#### 1.1.4.1 Buried Metal Layer

Buried metal layer (M-1) and its contact layer (V-1), lie underneath the device layers. These layers are proposed to reduce intra-cell routing congestion and improve cell pin access. Fig. 1.3a shows a cross-section of the traditional interconnect stack on top of

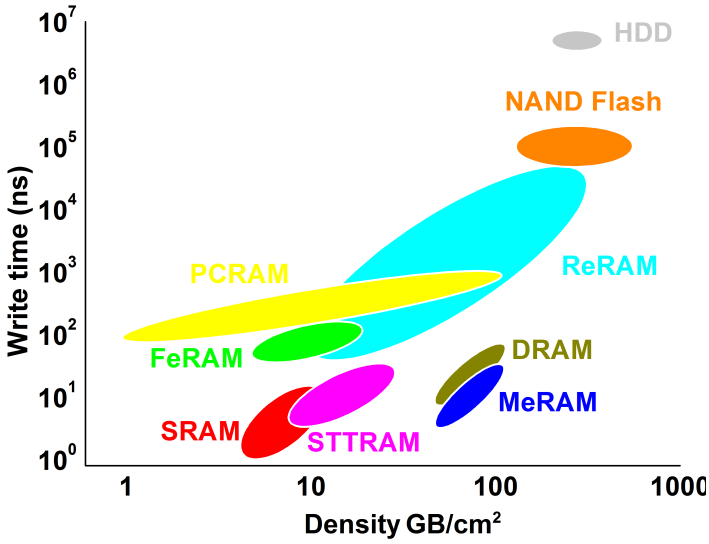



Figure 1.2: Density and write time of memories.

the device layers, and Fig. 1.3b shows the interconnect stack after adding the proposed buried layer under the device layers.

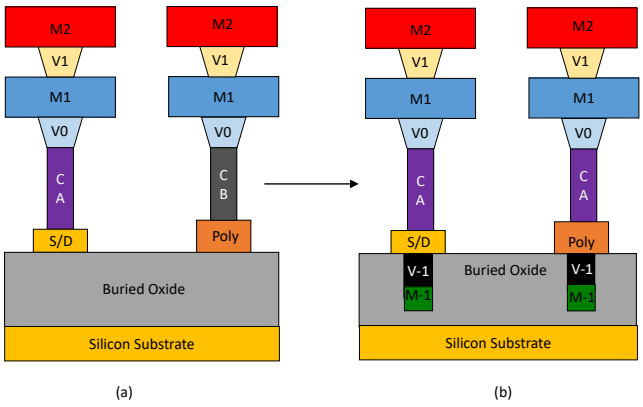



Figure 1.3: (a) Original stack of interconnect layers on top of device layers in SOI process (b) New stack with $M_{-1}$ and $V_{-1}$.

#### 1.1.4.2 Heterogeneous Integration

Heterogeneous integration (HGI) of silicon, germanium, and/or III-V semiconductor devices on a single platform can open alternative pathways to improving the performance and functionality of nanoscale integrated circuits. In contrast with homogeneous (all-Si) designs, HGI combines the advantages of disparate materials to optimize the complex

requirements and tradeoffs faced in circuit design.

### 1.1.5 Emerging Memory System Reliability

Many data-center studies [SL12a, SDB15, MWK15, SPW09] point out that main memory reliability is becoming a crucial problem in computing systems. Although the per-bit memory failure rate improves with technology development, the improvement cannot compensate the growing memory density required by increasing data-heavy applications, and hence worsened memory failure rate has been observed [SPW09]. Moreover, the introduction of emerging non-volatile memories and emerging memory-centric architecture will exacerbate the reliability problem [ZZD12, LAS08]. The consequences of memory failure are frequent system failure recovery and faulty memory replacement, which result in reduced serviceability and increased costs.

## 1.2 Emerging Technologies Evaluation

Though many emerging technologies have been proposed, the best replacement technology for silicon is still unclear. It is risky for semiconductor industry to invest in an emerging technology before the knowledge of its benefit and profit; they need rigorous evaluation to identify the values prior to commercial investment.

A comprehensive technology evaluation should assess the technology in the contexts of applications and circuit designs, which are co-optimized with the evaluated technology to obtain maximum benefits. Partially comparing technology parameters may lead to biased results. For example, the lower leakage current of tunneling FET (TFET) does not always guarantee lower power than Si CMOS, given that to achieve the same speed as Si CMOS, TFET gates should have large size to offset its weak driving capability, resulting in increased interconnect load, chip area and power. Moreover, CMOS with power management techniques can significantly reduce power consumption, which cannot be omitted in comparison with emerging boolean devices.

This dissertation includes several evaluation and optimization frameworks in the contexts of circuit and system for emerging logic and memory devices, and integration technologies:

- PROCEED [WPC15, WPC14] (a Pateto-based circuit-level optimization framework for logic devices), a framework that co-optimizes and evaluates emerging logic devices with circuits (given benchmarks), incorporating gate sizing, power management, and threshold voltage selections.

- An evaluation framework that assesses the impact of process variation in FinFETs on the speed, power, and reliability of microprocessors [WLP13].

- An evaluation framework for STT-MRAM and MeRAM with respect to density, power, speed, scalability, and reliability [WLE16b].

- An evaluation framework for heterogeneous integrated processors [LWP15b]

- An evaluation framework for buried metal layer with respect to chip speed, cost, and density [ZBW16].

- MEMRES (a memory system reliability simulator) [WHZ16, WHZ15b], which simulates memory system reliability (failure rate) considering in-DIMM and in-controller error-correcting code (ECC), memory page retirement, hardware sparing, memory scrubbing, and memory access pattern.

## 1.3 Design for Efficient MRAM Write and Read

According to our evaluation frameworks, MRAM is a promising technology that potentially improves computing system performance due to its fast programming speed, high density, and non-volatility at device-level. Its benefit can be further maximized in an emerging memory-centric architecture for data-driven applications, where most computation is performed near or inside memory. We have proposed several techniques to improve its write and read efficiency. Some techniques can be extended to all resistive non-volatile memories (NVM).

### 1.3.1 Optimized MRAM Write and Read Enabled by MTJ-based Variation Monitor

We propose a variation monitor [WLG16] that senses wafer-level process variation (free layer thickness variation) and monitors dynamic temperature change in MRAM. This variation monitor senses the MTJ free layer energy barrier through monitoring thermal activation rate accelerated by applying small current/voltage pulse across MTJs. The applied pulse exponentially reduces the retention time allowing effective in-situ sensing on small MTJ array. Compared with conventional temperature monitors, this monitor has 20X smaller area, 10X faster speed, and 5X more energy-efficiency with same resolution [WLG16].

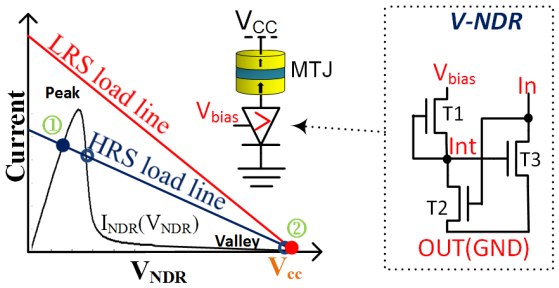
Sensing the variation can not only enhance memory reliability but also can improve MRAM write and read efficiency by dynamically tuning write and read pulse, e.g., MRAM at high temperature switches faster and is vulnerable to read disturbance, which can be written and read with low voltage amplitude. The sensed variation levels can be mapped to an optimized MRAM write and read pulses. Our preliminary results [WLG16] show that with adaptive write, the write latency of STT-RAM and MeRAM cache are reduced by up to 17% and 59% respectively, and application run time is improved by up to 41%.

Additionally, the proposed design can be utilized as magnetic/thermal attack sensor. The lower thermal stability in the monitoring MTJs allows them to react much earlier than normal MRAM array in an attack. The early reaction leaves time for emergency actions to protect data from loss.

### 1.3.2 Negative Differential Resistance-assisted NVM Write and Read

We proposed using negative differential resistance (NDR) to selectively program resistive NVM according to its resistance state [WPC17, WPG17]. This technology reduces programming time and energy. As shown in Fig. 1.4, to realize the function, a voltage-controlled NDR (V-NDR) is connected to a resistive memory cell through a series path. The V-NDR shows either a high or a low resistance when the equivalent resistance of other components in the series connection (mainly from an NVM cell) is lower or higher than a threshold determined by the V-NDR. The high-to-low resistance ratio of V-NDR

can be 10 $\sim$ 1,000 from silicon demonstration, which is drastically higher than those of many NVM, for example, 2$\sim$4 for MRAM. This eases several memory operations. For NVM switched by current, an automatic write termination can be performed to avoid energy waste upon the completion of switching from HRS to low resistance state (LRS) due to the V-NDR resistance change. For NVM with small HRS to LRS ratio, e.g., MRAM, V-NDR can amplify the resistance ratio by 10-100X. In the meantme, V-NDR can minimize read disturbance of MRAM by over $10^9$X given that the disturbance current is cut off by V-NDR. Experimental demonstration has shown 2X write energy reduction and $10^9$X read disturbance reduction.




(a) V-NDR in series with MTJ (b) CMOS built V-NDR

Figure 1.4: (a) Load line for V-NDR-MTJ series circuit (illustrated in inset). Blue and red line correspond to the MTJ HRS and LRS respectively. The stable operating points when the MTJ is in HRS and LRS are indicated by ① and ②, respectively. (b) Diagram of proposed 3T CMOS circuit for generating V-NDR between IN and OUT terminals.

We have also proposed current-controlled NDR (C-NDR) (see Fig. 1.5) as a complementary solution of V-NDR for fast MLC NVM programming. In MLC NVM (ReRAM, CTM, and PCM), more than two resistances are utilized to store data. The conventional scheme for intermediate resistance programming (not the lowest and highest resistance) is multiple write-and-check cycles [PMH15], where a read operation following a write checks the cell resistance against target one and determines additional programming cycles. This scheme takes long programming time and high energy. By utilizing V-NDR and C-NDR, intermediate resistance programming can be completed in one cycle. V-NDR and C-NDR can work together to perform MLC programming automatic termination. V-NDR and C-NDR are used for programming a cell to lower and higher resistance respectively. Multiple V-NDR and C-NDR with different sizes are utilized, where every NDR is sized

according to one target intermediate resistance such that it terminates write current when the programmed cell reaches the target resistance. The simple NDR design combines the functions of read check and write control, dramatically improving MLC programming efficiency.

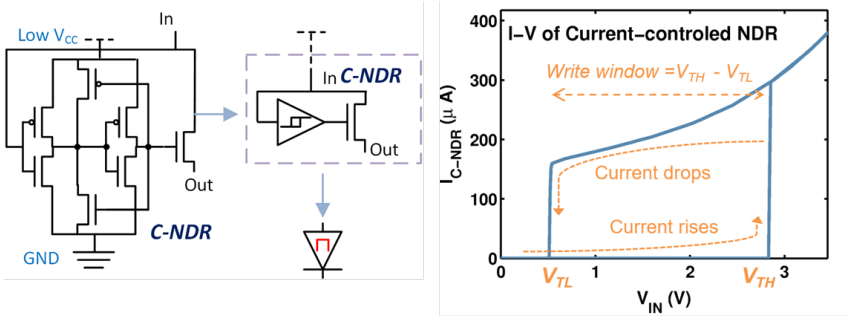



Figure 1.5: (a) One example of C-NDR comprising of a Schmitt trigger and an NMOS. The Schmitt trigger has two threshold voltages $V_{TH}$ and $V_{TL}$. (b) The I-V of the C-NDR. When VNDR starts from 0, then the current of C-NDR remains low until VNDR reaches $V_{TH}$, the current suddenly rises until VNDR reduces below $V_{TL}$. The Schmitt trigger is supplied with low VCC to reduce leakage.

## 1.4 Stochastic Non-volatile Computing using MTJ

The rapid development of Internet of Things (IoT) creates a big potential need for low-power wireless devices. Non-volatile ((NV) computing system designed with NVM, enabling persistent computing and intermediate power gating, potentially holds significant energy advantages. NVM based non-volatile computing is in great demand for IoT. However, simply using NVM as computing memories like register and cache cannot improve computing efficiency, neither the use of NVM as backup storage. By contrast, we have proposed special NV computing methodologies that use NVM as logic elements, which potentially reduce design energy and area.

We have proposed NV stochastic computing (NVSC) [WPL17] designed by VC-MTJ and V-NDR [WLG16]. This computing scheme outperforms conventional CMOS stochastic computing (SC) by over 300% energy efficiency. In addition to the use of NVM for storage and bitstream generation like previous NVM based SC, the proposed NVSC also

computes through VC-MTJ with the assistance of V-NDR. Moreover, the bitstream generation using VC-MTJ is more robust, efficient, and accurate than previous NVM based SC.

Apparently, SC is not a replacement for binary computation. Exporting computation to a SC based accelerator from a binary system will introduce overhead of bitstream conversion and communication. Hence, the SC should be designed only for suitable applications that can dramatically reduce energy consumption. Machine learning and image processing are very good candidates for their massive use of multiplication and addition, of which the SC implementation is simple and efficient. NVSC based Adaboost, as an example, has shown 3-30X more energy-efficiency than that of CMOS binary and SC design [WPL17].

# CHAPTER 2

# PROCEED: A Pareto Optimization-based Circuit-level Evaluator for Emerging Devices

## 2.1 Chapter Introduction

To explore additions or alternatives to CMOS, emerging devices must be assessed within the context of the circuits they might be used to build. Many technology benchmarking methods have been proposed to meet this need [WOW10, WA11, SFK11, FHS06, NY13, SK99, LLA11, KKA08, ARG09, CLD13, CKL09, PN12, SLS14a, SLS14b, SKC14, SLL14]; as summarized in Fig. 2.1, all these methods neglect a number of essential circuit features, any one of which can dramatically alter the results. Because of their variety and complexity, modern devices and circuit designs must be carefully chosen to complement each other before assessing viability; this requires a level of flexibility in the benchmarking process that has not existed until now. Device-circuit assessments must consider several factors to draw realistic conclusions. First of all, any effective power and delay evaluation of modern circuits should cover several orders of magnitude since their operating frequencies range from kHz to GHz. Second, chip area, ignored in all current evaluation methods, should be simultaneously considered because of its impact on manufacturing cost and interconnect length. Third, crucial tuning knobs such as logic gate sizing and supply voltage ($V_{dd}$) or threshold voltage ($V_t$) selection must be optimized for proper use of a particular circuit. Fourth, since circuit performance depends critically on the device operating point, benchmarks should consider the full device current-voltage (I-V) characteristics rather than only simplified metrics like saturation current ($I_{on}$ or off-state leakage ($I_{off}$). Fifth, a given device may not be suitable for all circuit architectures because of variations in logic depth histogram (LDH) patterns, and logical or physical structure. Sixth, as technologies scale down, device variability due to ambient process

Table 2.1: Comparison of variables considered in benchmark methodologies in the literature. P-D : power-delay, A-D: area-delay, A-P: area-power, $\mu P$: micro processor, LDH: logic depth histogram, CPI: clock cycles per instruction, SS: subthreshold swing

| Methodology | | [WOW10] | [WA11] | [SFK11] [FHS06] [NY13] | [SK99] | [LLA11] | [KKA08] | [ARG09] [CLD13] [CKL09] | [PN12] | [SLS14a] [SLS14b] | [SKC14] | [SLL14] | PROCEED |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | | $CV/I$, $CV^2$, $I_{on}/I_{off}$ | P-D curves, $I_{on}/I_{off}$ | P-D Pareto curve | Clock period | SS $I_{on}$ | Energy, Clock period | $CV/I$, $CV^2$ | CPI | P-D Pareto curve, Yield | P-D Pareto curve | P-D Pareto curve | P-D, A-P, A-D Pareto curves |
| Benchmark | | Latch, Inverter chain | $\mu P$ | $\mu P$ | $\mu P$ | Device | Inverter chain | Small logic element | $\mu P$ | $\mu P$ | Small circuits | $\mu P$ | Arbitrary circuit ($\mu P$ here) |
| Power management | | | | | | | | | | | | | ✓ |
| Optimization knobs | $V_{DD}$,$V_t$ | ✓ | ✓ | ✓ | | | ✓ | | $V_{DD}$ | ✓ | | ✓ | ✓ |
| | Size | | | | ✓ | | | | ✓ | | ✓ | ✓ | ✓ |
| | Multiple $V_{DD}$,$V_t$ | | | | ✓ | | | | | | | | ✓ |
| Circuit design parameters | Interconnect | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | LDH | | | | | | | | | Architecture sim | Synthesis | Synthesis Architecture sim | ✓ |
| | Activity | ✓ | ✓ | | ✓ | | ✓ | ✓ | | | | | ✓ |
| Device model | Current | $I_{on}$, $I_{off}$ | $I_{on}$, $I_{half}$ $I_{off}$ | $I_{on}$, $I_{half}$ $I_{off}$ | $I_{on}$, $I_{off}$ | TCAD | Model | Model | $R_{on}$, $R_{off}$ | Lookup table | Lookup table | Lookup table | Compact model |
| | Capacitance | Fixed | Fixed | Full C-V | Fixed | N/A | Fixed | Full C-V | N/A | Lookup table | Lookup table | Lookup table | C-V model |

fluctuations becomes more important and impacts circuit viability. Seventh, the benefits to circuit designs of cooperatively using several device types through heterogeneous integration (HGI) are strongly dependent on the design adaptability and circuit topology, which must be considered in any assessment. All the aforementioned complexities would mandate a complete circuit design flow for performance evaluation, which is nevertheless impractically time consuming. Therefore, an alternative evaluation method must be developed that accounts for the above factors with reasonable computational run time.

In this chapter, we introduce a device evaluation framework, called PaReto Optimization-based Circuit-level Evaluator for Emerging Devices (PROCEED), for fully circuit-aware benchmarking [WPC14, WPC15]. It incorporates typical circuit design flow flexibilities and tunes physically adjustable device and circuit parameters to generate realistic conclusions about the overall performance. We introduce the PROCEED framework and demonstrate its efficacy by using it to evaluate several technology options: traditional silicon-on-insulator (SOI) devices, novel tunneling FETs (TFETs), and HGI combinations of these devices. We outline the methodology behind PROCEED in Section 2.2, and explain details of the Pareto optimization procedure in Section 2.3. We present the results of our PROCEED study on TFET and SOI devices in Section 2.4 and summarize

our conclusions in Section 2.5.

## 2.2 Overview of PROCEED Framework

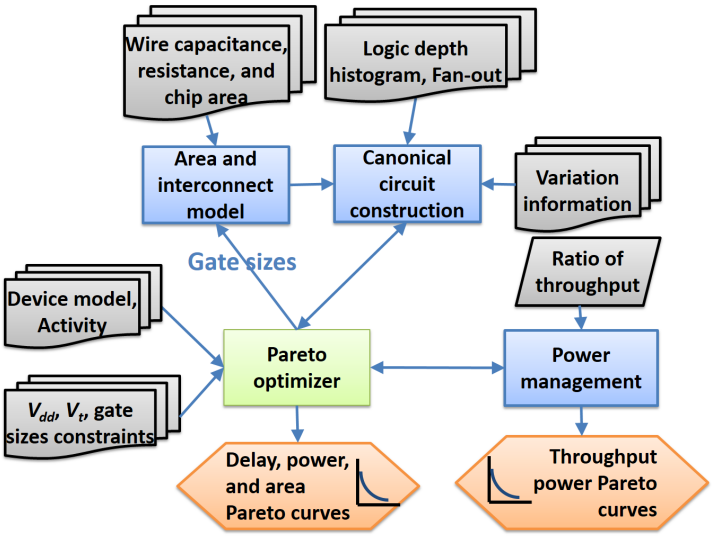



Figure 2.1: Overview of PROCEED framework.

As shown in Fig. 2.1, typical inputs to PROCEED include interconnect information (such as average wire resistance (R) and capacitance (C) and chip size), circuit benchmark design (i.e. design LDH and average fan-out), variability (through $V_{dd}$ drops, $V_t$ shifts, etc.), full device I-V models, and operating activity, as well as optional constraints on $V_{dd}$, $V_t$, chip area, and the ratio of average to peak through-put (i.e., clock cycles per second). Simulation blocks with interconnect loads are generated for canonical circuit construction through a feedback process using input from the Pareto optimizer (through tuning parameters like $V_{dd}$, $V_t$, and gate sizes). Optimized results are generated in the form of the PD Pareto curve. Finally, power management analysis, including dynamic voltage and frequency scaling (DVFS) and power gating, is performed using this Pareto curve. In its present implementation, PROCEED can evaluate an arbitrary logic device candidate as long as it does not cause a dramatic change in circuit topology. For instance, multi-state logic devices fall outside PROCEEDs current scope because of the unconventional circuit architectures in which they operate.

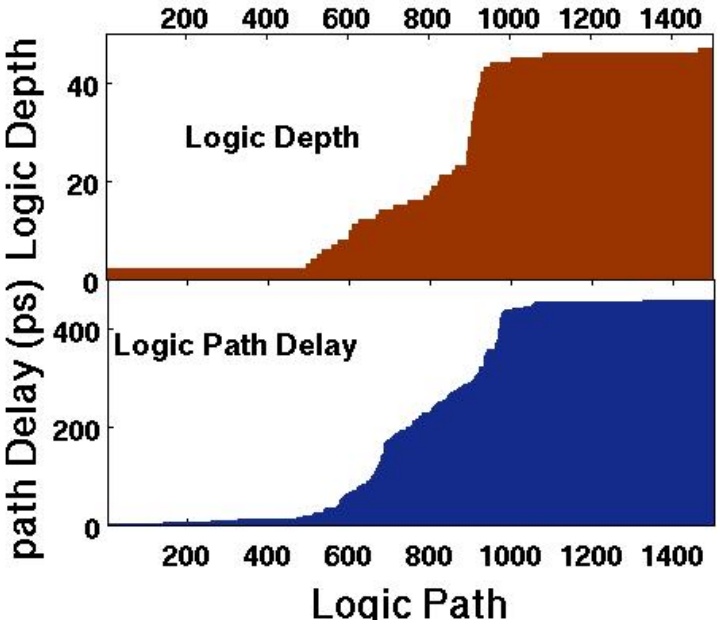



Figure 2.2: Typical logic path depth distribution and logic path delay extracted from a synthesized CortexM0.

### 2.2.1 Canonical Circuit Construction

Full, exact optimization is an impossible job for large digital circuits. Since the goal of PROCEED is to predict the best performance and power tradeoffs for emerging devices, detailed circuit design is not our target and contributes little to evaluation. We therefore utilize only essential design information to maximize performance and determine the optimal $V_{dd}$, $V_t$, and gate sizes for a given power consumption limit. Typical circuit designs contain both short and long logic paths with the path delay roughly proportional to the logic depth, as shown in Fig. 2.2 for a CortexM0 design. Hence we derive the LDH by extracting endpoint slacks from benchmark designs and estimating logic paths.

In Fig. 2.3, we show an example of the simulation blocks used to construct a specific circuit. For simplicity, we first divide logic paths into n bins based on logic depth; in Fig. 2.3(a), for instance, n = 5. A larger number of bins improve accuracy at the expense of computation time. Each bin is modeled by corresponding simulation blocks $S_i$ ($S_1$-$S_5$ in Fig. 2.3(a)), which are in turn made of i gate stages. We use the gate design for $S_i$ to construct logic paths belonging to a given bin i.

The LDH is divided such that the longest path in each bin has the same delay if all these blocks have the same delay. Fig. 2.3 shows an example of this with five evenly

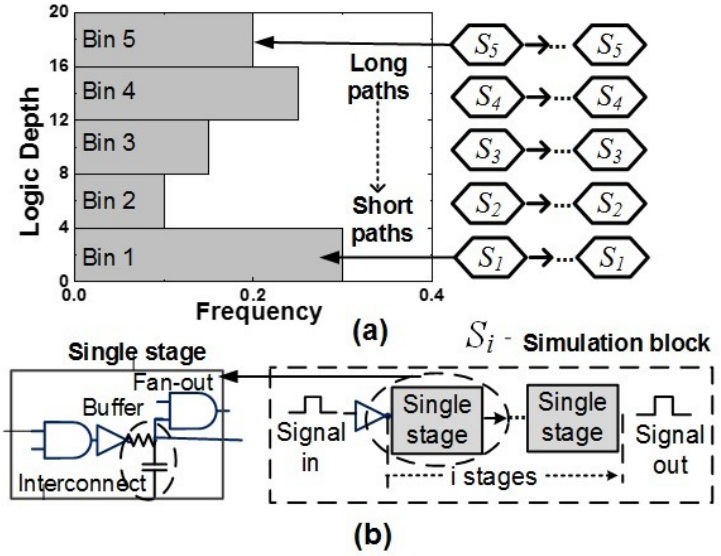



Figure 2.3: (a) Example of simulation block allocation in PROCEED based on logic depth. (b) Circuit schematic used for simulation and optimization.

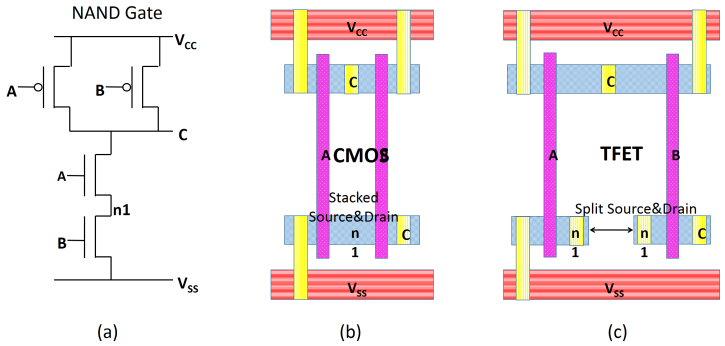



Figure 2.4: NAND gate (a) schematic and layouts for (b) CMOS and (c) TFET.

spaced bins for logic paths from one to twenty stages such that the first bin contains one to four stage paths, the second holds paths with five to eight stages, and so forth. The delay weight $W_D$ is the number of copies of $S_i$ needed to construct the longest path in bin i ($W_D = 4$ in Fig. 2.3). The logic gate and interconnect used for a single stage in the simulation blocks is shown in Fig. 2.3(b). The gate can be NAND, NOR, or something more complicated like XNOR, depending on the average number of transistors per gate in the chosen benchmark. The gate choice can also differ from bin to bin, though in the examples in this chapter we use NAND gates for all bins. An inverter or buffer is inserted after the gate to drive the fan-out (which is a replica of the chosen gate sized to the average fan-out) as well as interconnects, which are represented by the RC elements.

### 2.2.2 Delay and Power Modeling

Delay, power, and area are the three most important gross metrics in the design of digital circuits, but usage constraints lead to tradeoffs between them that must be balanced to maximize the overall efficiency of the design. Hence we use them as evaluation metrics in PROCEED. As described in the beginning of Section 2.2.1, circuit benchmarks are mapped to canonical circuits that are used to estimate these metrics without time-consuming large-scale simulations. The delay and power of the canonical circuits are extracted from SPICE simulations, which are then scaled, summed, and minimized to obtain the corresponding values for the given benchmark. We have verified that the values predicted by this method agree well with those calculated using commercial synthesis tools (as described in Section 2.4.1), demonstrating the high accuracy of PROCEED in emulating realistic circuit behavior.

### 2.2.3 Area Modeling

The area of the gates used in canonical circuit constructions is simulated using UCLADRE [GG12], where they are minimized in accordance with input design rules and gate netlists.

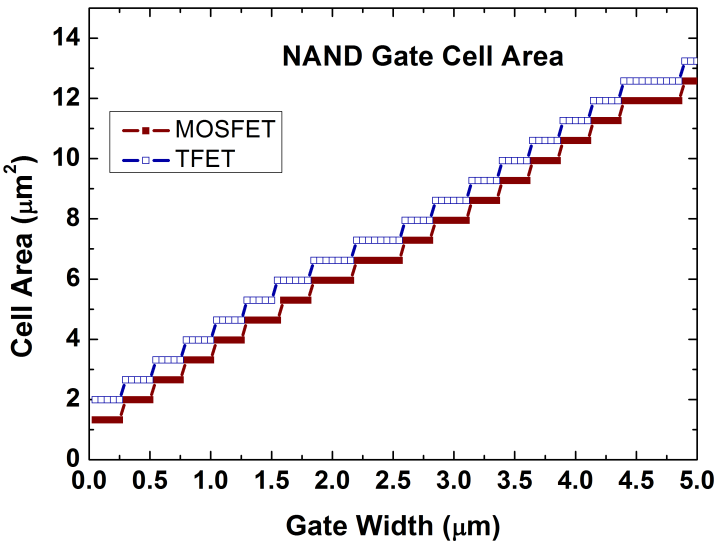



Figure 2.5: Cell area of CMOS-based and TFET-based NAND gates as a function of gate width.

Unlike traditional CMOS devices which have inter-changeable source and drain, some emerging technologies like TFETs [IR11, WG14b, LDN13] have asymmetric structures

where current can only flow in one direction. This asymmetry prohibits stacking of transistors by sharing the source and drain. Fig. 2.4(a) shows a NAND gate logic schematic where adjacent transistors share a source/drain at node n1. Fig. 2.4(b) stacks two NMOS devices to create a compact layout for traditional CMOS technology. However, due to the source/drain asymmetry, a TFET layout for the same circuit must split the stack, leading to additional area overhead as illustrated in Fig. 2.4(c). To account for this effect, we modify UCLADRE such that it generates area-optimal TFET layouts for any input circuit netlist. The cell area of CMOS-based and TFET-based NAND gate as a function of gate width is shown in Fig. 2.6. The additional area overhead of TFET is clearly significant.

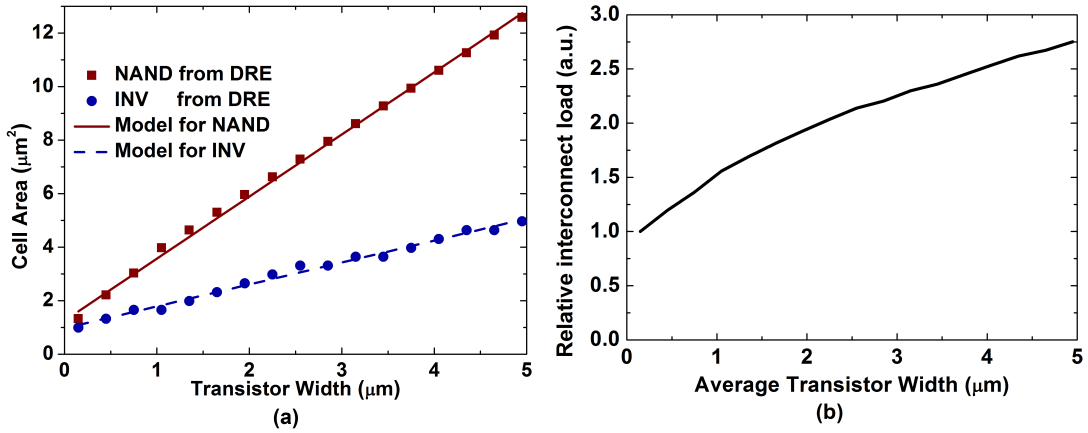



Figure 2.6: (a) Cell area and (b) interconnect load as a function of transistor width. In (b), transistor width is the same in Inverter and NAND gate.

Design rules can differ depending on the technology; for instance, for nanotransfer HGI, additional separations between P wells and N wells may be needed to eliminate overlay errors depending on the material choices for NFETs and PFETs [CSK12]. Similarly, two devices with different design rules will result in different areas even if they are sized to the same gate width and length. For technology evaluations, we calibrate the design rules, sweep gate width in UCLADRE, and fit linear models of gate area to the simulation results. An example of the models accuracy is shown in Fig. 2.6(a). The chip area is calculated using the following procedure: 1) the area of gates in each bin is obtained from the fitted area model. 2) Chip area is calculated as the weighted sum of gate areas in all bins. 3) The weights are decided during canonical circuit construction stage.

### 2.2.4 Process Variation and Voltage Drop

As devices scale to ever smaller technology nodes, device variations due to process and ambient variations are becoming more important and should not be neglected in PD evaluation. In circuit design, slow corner devices are commonly used to estimate the upper bound on minimum working clock period (critical delay) and create a safe design with sufficient delay margin. We define the slow corner as a device with reduced effective $V_{dd}$ and increased $V_t$ due to variability and parasitic effects, and the corresponding voltage shifts are inputted into PROCEED. Separate models for additional variability effects may be incorporated as needed. During circuit optimization, the worst case scenario is considered by calculating delay using the slow corner device and power using the normal device. An illustration of how this may affect the operating point of real devices is given for TFETs and SOI MOSFETs in Section 2.4.7.

### 2.2.5 Interconnect Load

We model interconnect loads using a series RC circuit. We assume R and C are linear with interconnect length, so the load will be proportional to the square root of the chip area [DDM98], and can be dynamically changed based on average gate width. Fig. 2.6(b) shows an example of interconnect load as a function of transistor width, using a combined NAND and INV cell to estimate the cell area. The average RC and extracted gate width are then fed into PROCEED. Even simple considerations of interconnect load will strongly impact the overall evaluation results. In general, gates using devices with low driving ability need to be sized up to achieve the same performance as those with high driving ability. This increases the area of the chip as well as the interconnect loads, which exacerbates the drive demands and requires further gate sizing. PROCEED correctly describes such cases, as quantitatively demonstrated in Section 2.4.3.

### 2.2.6 Pareto-Based Optimization

Following canonical circuit construction, all logic paths are replaced by simulation blocks ($S_i$) which will be optimized. However, these blocks cannot be optimized separately because they usually share a common $V_{dd}$ and $V_t$, complicating the procedure. As a

result, we use a modified form of a general simulation-based Pareto technique to perform the optimization [RKW09], as discussed in more detail in Section 2.3. The simulation target is regarded as a black box with two optimization objectives: any two of design area, power, and critical delay (minimum working clock period).

### 2.2.7 Power Management Modeling

Current technologies usually allow circuits to operate in at least three modes: normal, power saving, and sleep mode. Previous evaluation works only consider the normal mode where devices continuously work at peak performance. PROCEED allows devices to also operate at a second, lower supply $V_{dd2}$ (DVFS) as well as in the off state (power gating). This allows us to evaluate device PD scalability as a function of $V_{dd}$, an important circuit feature which, to the best of our knowledge, has been ignored in all previous evaluations.

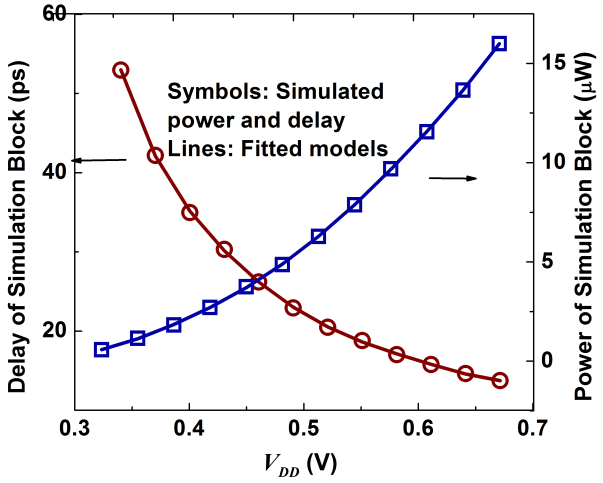



Figure 2.7: Model fitting for simulation blocks delay and power as a function of $V_{dd}$.

The ratio of average to peak throughput is another input for PROCEED. To study power management, we choose all designs from the generated Pareto points which achieve the lowest power and peak throughput. From this, the optimizer selects the best choice for the second power rail and divides the time spent operating at high $V_{dd1}$ (i.e. the original supply) and the new lower $V_{dd2}$. This is done as follows: starting from the optimized design (with maximized peak throughput), we carry out circuit simulations by sweeping voltages lower than the original $V_{dd1}$. The original design may even have multiple $V_{dd}$, in which case different blocks can use different $V_{dd2}$ values. Delay and power models for

every simulation block $S_i$ as functions of $V_{dd}$ are constructed using polynomial functions, as in Fig. 2.7:

$$D_{Si}(V) = \sum_{j=-2}^{5} a_{i,j} V^i, P_{Si}(V) = \sum_{j=-1}^{5} b_{i,j} V^i \tag{2.1}$$

We have tested and found this model to be sufficiently accurate; for instance, in our experiments presented in Section 2.4.6, the relative error of the polynomial fittings is less than 2%. We then optimize for the weighted power sum $f_1 P_1 + f_2 P_2$, subject to

$$D_2 \geq W_D D_{Si}(V_{i2}), P_2 \geq \sum_{i=1}^{n} W_{Pi} P_{Si}(V_{i2})\ i = 1, 2, ..., n \tag{2.2}$$

$$f_1 \cdot 1/D_1 + f_2 \cdot 1/D_2 \geq T_{Ave}, 0 \leq f_1 + f_2 \leq 1$$

Here $D_{1,2}$ and $P_1$ are the design delay and power using $V_{dd1,2}$, $W_D$ and $W_P$ are the delay and power weights mapping from simulation blocks to the design, and f1 and f2 are the fractions of time spent operating with $V_{dd1}$ and $V_{dd2}$, with any remaining time assumed to be spent in the off state. Typically this step is not a feasible convex optimization problem; however, by using the fitted model of (1), an enumeration approach can solve this problem very efficiently with acceptable accuracy. In Section 2.4.6 we give an example of how PROCEEDs power management capabilities are applied in practice.

### 2.2.8 Activity Factor

Activity varies widely with application: in embedded sensing, for instance, factors below 1% are observed in car-park management [BOO06], while those for systems like VigilNet exceed 50% [HVY06]. Activity factor can therefore dramatically change the evaluation results and is included as an input to PROCEED. In circuit simulations, the dynamic and leakage power are separately extracted and the total power is equal to their weighted sum. From this the circuit can be optimized for a known activity factor. This can be of primary importance in determining the usability of a given device, as we experimentally show in Section 2.4.5.

### 2.2.9 Multiple $V_{dd}$ and $V_t$

In modern circuit designs, multiple $V_{dd}$ and $V_t$ values are used, as illustrated in Section 2.4.2. In our scheme, transistors in each simulation block Si must be assigned the same voltages, so to optimize a design with integer m different $V_{dd}$ or $V_t$ biases, the number of simulation blocks must be greater than m. In addition, our optimization is an iterative process whereby Pareto points are updated and improved based on previous iterations. Therefore, if the same $V_{dd}$ or $V_t$ is shared by multiple simulation blocks, this assignment cannot be changed during the optimization. A full optimization for multiple $V_{dd}$ and $V_t$ is implemented by considering designs with all sets of reasonable voltage assignments in parallel. For example, if we have five simulation blocks $S_1$-$S_5$ and two available $V_t$, then for i from 1 to 4, blocks $S_1$ to $S_i$ use the high $V_t$ and $S_{i+1}$ to $S_5$ use the low $V_t$. This comprises the set of useful voltage assignments, since simulation blocks with longer logic paths require higher performance (i.e. lower $V_t$).

### 2.2.10 Heterogeneous Integration

Every emerging device has its own characteristic advantages, such as steep subthreshold slope for TFETs and high mobility and on-current for III-V or CNT FETs. However, any one of these devices cannot fulfill all the disparate requirements of the various macros in future circuit applications. HGI combines several types of devices onto a single chip to maximize performance at the expense of cost and area penalties [CSK12]. In PROCEED we use a quick way to explore the benefits brought by this technology. In general, slow and low-power devices are useful for non-timing critical macros, while high performance devices are suitable for high-speed macros. Furthermore, within single circuit blocks, critical and non-critical paths can be built using different types of devices. In PROCEED, models for all HGI devices are inputted and the delay and power of logic paths built using different devices are modeled accordingly. Since these devices operate in the same circuit and affect the overall performance, PROCEED optimizes the HGI gates in a concurrent fashion. Since different devices are apportioned among the available logic path bins, the granularity of HGI optimization results in this approach is set by the number of bins considered. PROCEED therefore allows us rapidly evaluate combinations of multiple

technologies over a wide range of delay, power and area requirements. As an example of this, we evaluate the potential of circuits implemented using TFET and SOI HGI in Section 2.4.8.

## 2.3 Pareto Optimization

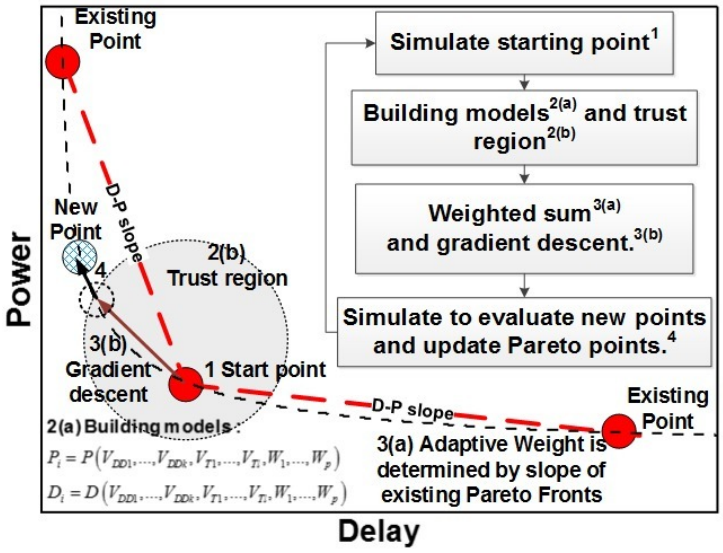



Figure 2.8: Optimizer overview. Adaptive weight is chosen by slope of existing fronts. Based on starting point, meta-modeling is built and gradient descent is used to find potential points. Simulate potential points to get new Pareto points.

PROCEED can simultaneously optimize any two metrics out of delay, power and area while the third is treated as a constraint; for instance, we can perform a Pareto optimization of delay and power with a maximum area constraint. As described in Section 2.2.2, the chosen area model is linear in gate width and hence easier to optimize than delay and power. Therefore, in the remainder of this section, we will describe in detail the Pareto optimization of delay and power with constrained area. Fig. 2.8 presents an overview of our Pareto optimization process. PROCEED treats circuit simulations as a black box and uses models to optimize tuning parameters based on the simulation results. Gradient descent is utilized to find minimal objectives in the trust region. Final simulations are performed on designs outputted by the model-based optimization. The vector of tuning parameters $\boldsymbol{X}$ for optimization is represented as:

$$X = (x_1, x_2, ..., x_m) = (\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_n}) \tag{2.3}$$
$$\boldsymbol{y_i} = (V_{dd,i}, V_{t,i}, W_{i1}, W_{i2}, ..., W_{i,2i}),\ i = 1, ...n$$

where $x_j$ are the variables of $\boldsymbol{X}$, $y_i$ are vectors of the tuning parameter variables for simulation block $S_i$, and $V_{dd}$,i, $V_t$,i, and $W_{ij}$ are the supply, threshold voltages, and the sizes of gates and inverters in each block $S_i$, respectively. The optimization entails the following steps:

(1) Picking a starting point: Each iteration of the optimization process uses a starting set of variables $\boldsymbol{X}_0$ around which to explore. For the first iteration, any reasonable $\boldsymbol{X}_0$ may be inputted. The choice of the initial point can affect runtime but not final accuracy, since bad points will gradually be eliminated by the optimization process and converge to the true answer. Subsequently $\boldsymbol{X}_0$ is determined from already existing Pareto points by computing the Euclidean distance between all neighboring points in delay-power coordinates, as shown in Fig. 2.8. The point with the largest total distance from its two neighbors is chosen as the starting point $\boldsymbol{X}_0$ since it lies in the sparse region, which is usually suboptimal.

(2) Building a local model around $\boldsymbol{X}_0$: To accelerate the optimization process, second-order delay and power models are constructed based on the simulation results. The delay and power models $D_{Si}$ and $P_{Si}$ for each block $S_i$ are calculated separately and then combined to obtain the model for the whole canonical circuit. Compared to simultaneously calculating model parameters for all blocks, this approach reduces the number of simulations, as determined by the size of the Hessian matrix (proportional to the number of variables squared). $D_{Si}$ and $P_{Si}$ are represented by the gradient vector $\boldsymbol{G}$ and Hessian matrix $\boldsymbol{H}$ as

$$D_{Si}(\boldsymbol{y_{i,0}} + \boldsymbol{\Delta y_i}) = D_{Si,0} + \boldsymbol{G_{Di}}^{\boldsymbol{T}} \boldsymbol{\Delta y_i} + \frac{1}{2} \boldsymbol{\Delta y_i^T H_{Di} \Delta y_i} \tag{2.4}$$
$$P_{Si}(\boldsymbol{y_{i,0}} + \boldsymbol{\Delta y_i}) = P_{Si,0} + \boldsymbol{G_{Pi}}^{\boldsymbol{T}} \boldsymbol{\Delta y_i} + \frac{1}{2} \boldsymbol{\Delta y_i}^{\boldsymbol{T}} \boldsymbol{H_{Pi} \Delta y_i}$$

This second-order model is a local estimation near the starting point. To guarantee validity, an adaptive trust region is applied as shown in Fig. 2.8, limiting the model

range inside the region

$$\boldsymbol{X_0} - \boldsymbol{\lambda}(r) < X < \boldsymbol{X_0} + \boldsymbol{\lambda}(r) \tag{2.5}$$

where r is the radius of this trust region and $\lambda$ is the range of the tuning parameters $\boldsymbol{X}$ and is a linear function of r.

(3) Model-based optimization: In this step, four metrics are used in optimization: $D$, $P$, $W_{dl} \cdot D + W_{pl} \cdot P$, and $W_{dr} \cdot D + W_{pr} \cdot P$. Minimization of $D$ and $P$ yields the fastest and lowest power designs in the local region, while the weighted sums of delay and power are used to populate the phase space by finding two Pareto points between the starting point and its neighbors. The optimization also needs to satisfy the constraint from the third metric (e.g. area in this case). Since the problem may not be convex, gradient descent with the logarithmic barrier method [BV04] is used to find these optimal points. The models region of validity lies in the intersection of the trust region and the inputted bounds for the tuning parameters. The objective function is performed as follows:

$$Minimize \;\; W_D D\left(\boldsymbol{X}\right) + W_P P\left(\boldsymbol{X}\right) - t\left(\sum_{j=1}^{m} \log\left|x_j - x_{j,b}\right| - \log\left(-A\left(\boldsymbol{X}\right) + A_{\max}\right)\right) \tag{2.6}$$

$$D(\boldsymbol{X}) = D\left(\boldsymbol{X_0}\right) + \boldsymbol{G_D}(\boldsymbol{X_0})^T\left(\boldsymbol{X} - \boldsymbol{X_0}\right) + \left(\boldsymbol{X} - \boldsymbol{X_0}\right)^T \boldsymbol{H_D}\left(\boldsymbol{X_0}\right)\left(\boldsymbol{X} - \boldsymbol{X_0}\right)$$
$$P(\boldsymbol{X}) = P\left(\boldsymbol{X_0}\right) + \boldsymbol{G_P}(\boldsymbol{X_0})^T\left(\boldsymbol{X} - \boldsymbol{X_0}\right) + \left(\boldsymbol{X} - \boldsymbol{X_0}\right)^T \boldsymbol{H_P}\left(\boldsymbol{X_0}\right)\left(\boldsymbol{X} - \boldsymbol{X_0}\right)$$

where $x_{j,b}$ are the upper and lower bounds for variable $x_j$, $A(\boldsymbol{X})$ and $A_{max}$ are the area model and maximum area constraint respectively, and $D(\boldsymbol{X})$ and $P(\boldsymbol{X})$ are delay and power for the entire design, respectively. The weights for delay and power are defined as follows:

$$W_{dl(r)} = \left(P_{l(r)} - P_0\right) / \sqrt{\left(P_{l(r)} - P_0\right)^2 + \left(D_{l(r)} - D_0\right)^2} \tag{2.7}$$
$$W_{pl(r)} = \left(D_0 - D_{l(r)}\right) / \sqrt{\left(P_{l(r)} - P_0\right)^2 + \left(D_{l(r)} - D_0\right)^2}$$

where ($D_0$, $P_0$) is the starting point and ($D_l$, $P_l$) and ($D_r$, $P_r$) are the left and right neighbor points, respectively. The solid points in Fig. 2.8 are examples of such points. The direction vectors ($W_{dl}$, $W_{pl}$) and ($W_{dr}$, $W_{pr}$) of the weighted sum of objectives are calculated so as to be perpendicular to the connecting lines between the starting point and its neighbors, as illustrated by the dashed line in Fig. 2.8. The weights in the

weighted sum optimization are used to yield the two new optimal points between the starting point and its neighbors. $D$ and $P$ are given by

$$D(\boldsymbol{X}) = W_D \cdot \max\left(\left(D_{S1}(\boldsymbol{y_1}), D_{S2}(\boldsymbol{y_2}), ..., D_{Sn}(\boldsymbol{y_n})\right)\right) \tag{2.8}$$
$$P = \sum_{i=1}^{n} W_i \cdot P_{Si}$$

where $W_D$ is the delay weight discussed in Section 2.2.1 and $W_i$ is the number of $S_i$ used in the canonical circuit construction. Because the maximizing function does not have a continuous derivative, we use higher order norms to estimate the maximum, so the elements of gradient vector and Hessian matrix for delay are derived as follows:

$$D(\boldsymbol{X}) \approx \|\boldsymbol{D}\|_K,\ \boldsymbol{D} = \left(D_{S1}(\boldsymbol{y_1}), D_{S2}(\boldsymbol{y_2}), ..., D_{Sn}(\boldsymbol{y}_n)\right) \tag{2.9}$$
$$G_{D,j}(\boldsymbol{X}) = \frac{\partial D(\boldsymbol{X})}{\partial x_j} \approx \frac{\partial \|\boldsymbol{D}\|_K}{\partial x_j}, \boldsymbol{H_{D,jk}}(\boldsymbol{X}) = \frac{\partial^2 D(\boldsymbol{X})}{\partial x_j \partial x_k} \approx \frac{\partial^2 \|\boldsymbol{D}\|_K}{\partial x_j \partial x_k}$$

where $K$ is the order of the norm. Higher $K$ results in more accurate results (we use $K = 100$ in our simulations). Similarly, the elements of the gradient vector and Hessian matrix for power are given as

$$G_{P,j} = \frac{\partial P(\boldsymbol{X})}{\partial x_j} = \sum_{i=1}^{n} W_i \cdot \frac{\partial P_{Si}(\boldsymbol{y}_i)}{\partial x_j} \tag{2.10}$$
$$H_{P,jk} = \frac{\partial^2 P(\boldsymbol{X})}{\partial x_j \partial x_k} = \sum_{i=1}^{n} W_i \cdot \frac{\partial^2 P_{Si}(\boldsymbol{y}_i)}{\partial x_j \partial x_k}$$

(4) Addition of new Pareto points: To correct for model errors, circuit simulations are performed to evaluate D and $P$for all remaining potential Pareto points found by the optimization. In Fig. 2.8, this process is illustrated by the shift of the hatched point to the dotted circle. Finally, points not on the Pareto frontier (such that at least one other point with both lower delay and power exists) are filtered out.

(5) Iteration termination: For each iteration, when choosing the starting point for each step, the radius of trust region around this point is decreased by a factor of p ($p > 1$).Two termination conditions are applied: 1) existence of a sufficient Pareto point density in the region of interest, defined by the largest gap between any two neighboring points being smaller than a given criterion. This condition is usually used for devices with wide operating regions (i.e. suitable for both high performance and low power applications). 2) Reduction of the radius of trust region below a given criteria. This usually occurs due to limitations on the device operating region or device model discontinuities.

The PROCEED runtime is of the order $O(rm^2) + O(r)$, where r is the resolution constraint (number of points in a unit Pareto curve), m is the total number of tuning parameters, $O(rm^2)$ is the complexity of the simulations for gradient and Hessian matrix calculation, and $O(r)$ is the complexity of simulating potential Pareto points. In our experiments, runtimes are mainly dominated by the resolution constraint; however, for large m, the $O(rm^2)$ term will dominate. The average PROCEED runtime to generate a full Pareto curve over three orders of magnitude in performance is about 4 hours on a single CPU. We use MATLAB in the optimization process and HSPICE for circuit simulations.

## 2.4 Experiment Results

To illustrate PROCEEDs capabilities, we use it to assess SOI and silicon TFET devices at the 45 nm node and compare the evaluation results with existing methods. Because of their interband tunneling conduction principle, TFETs are capable of very low leakage and extremely steep sub-threshold swing, making them well-suited for low voltage operation [IR11]. Currently, however, non-idealities in experimental devices and low on-current limit their performance. We examine the viability of currently achievable TFETs using a device compact model [PC12b, PCC13] calibrated against TCAD simulations and experimental SOI devices [JLP10]. While this does not represent the best possible TFET, which may require a different channel material or device structure, it have the advantages of being experimentally validated and structurally comparable to conventional SOI devices and represents a realistic lower bound.

For these reasons, we emphasize that our results do not constitute a final judgment on the viability or lack thereof of TFETs in future circuits; rather, they represent both a starting point from which to consider possible uses for present experimental (rather than projected) TFETs as well as a platform to demonstrate PROCEEDs unique capabilities. Traditional technologies are represented by 45 nm SOI MOSFETs modeled using commercial characteristics and compact model. Unless otherwise specified, all circuit results are generated with one $V_{dd}$ and two $V_t$. To easily compare devices, we will frequently refer to the Pareto crossover, defined as the (minimum working) clock period (critical

delay) above which the optimized novel device (here, the TFET) consumes less power than the established technology (SOI); lower Pareto crossover means the novel device is more promising for a given case since it has a wider operating range over which it is superior.

### 2.4.1 Framework Evaluation

To validate the PROCEED framework, we use the widely employed evaluation model of [FHS06] (hereafter Model [FHS06]), and a commercial synthesis tool to evaluate the PD Pareto curve for a CortexM0 microprocessor with a commercial 45 nm SOI library and model. The information needed for PROCEED and Model [FHS06] (LDH, average fan-out and interconnect load) is extracted from a synthesized, placed, and routed netlist at a minimum working clock period of 933 ps. Only single constant values of $V_{dd}$ and $V_t$ are used, as Model [FHS06] does not support multiple voltages and the commercial library has only constant $V_{dd}$ and $V_t$.

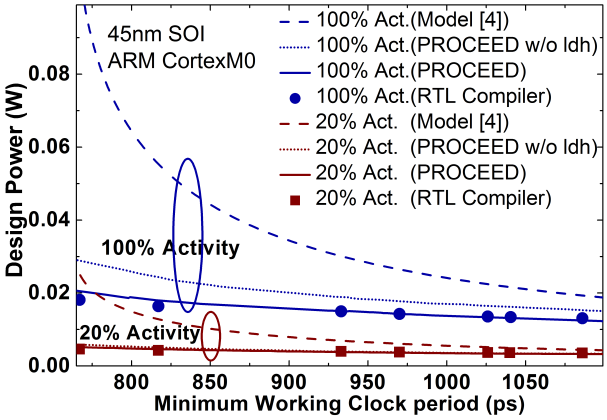



Figure 2.9: Pareto curves for delay and power as evaluated using a commercial synthesis tool, Model [FHS06], and PROCEED. $V_{dd}$ and $V_t$ are constants and only gate size is a variable.

As shown in Fig. 2.9, the PROCEED predictions are in much better agreement with the comprehensive optimized results from the register-transfer level (RTL) compiler compared to Model [FHS06], which is frequently used for device evaluation [WA11, SFK11]. The operating range for comparison is chosen by the synthesis results with the commercial library using one $V_{dd}$ and $V_t$. We note that using the compiler for evaluation purposes is

completely impracticable, since extracting a Pareto curve from kHz to GHz clock frequencies necessitates libraries with $V_{dd}$ and $V_t$ varying from 0.5 V to 1.2 V and 0.1 V to 0.5 V, respectively. However, the generation and optimization of these libraries would consume months of runtime, whereas we completed the same study in hours using PROCEED. Meanwhile, the computationally simple Model [FHS06] takes seconds to complete such Pareto curves but grossly overestimates power for two reasons: (i) the neglect of LDH in assuming all gates have the same (large) size used for the critical path, and (ii) the use of analytical PD models rather than circuit simulations using full device characteristics. The dotted line is the Pareto curve generated by PROCEED while neglecting LDH, illustrating the accuracy improvement contributed by the two aforementioned points. We further note that Model [FHS06] cannot account for adaptability, variability, or multiple $V_{dd}$ and $V_t$ effects. By benchmarking to the RTL results in Fig. 2.9, we observe PROCEED improves accuracy by 3X to 115X compared to the current standard Model [FHS06].

### 2.4.2 Impact of Multiple $V_{dd}$, $V_t$, and Gate Sizing

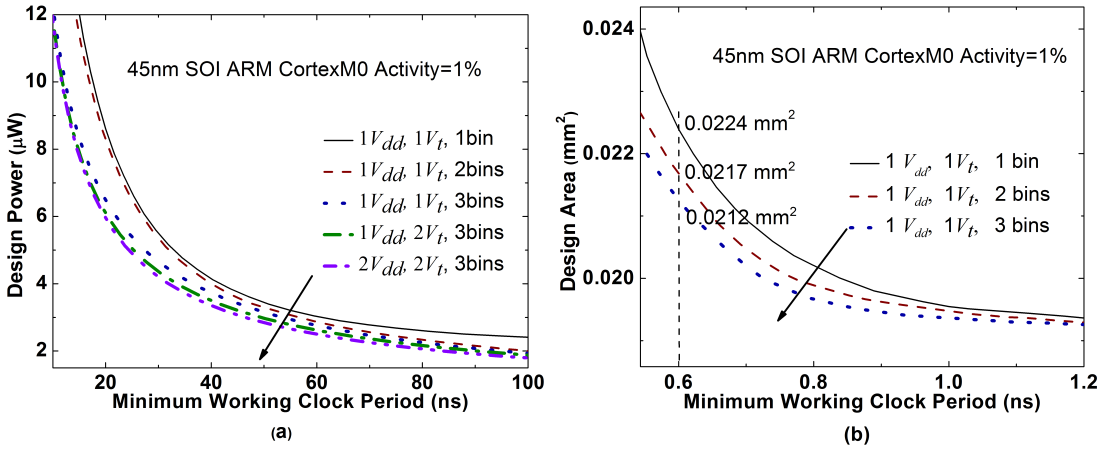



Figure 2.10: 45 nm SOI CortexM0 (a) power-minimum working clock period and (b) area-minimum working clock period as tuning parameters are increased.

Additional tuning parameters create a larger design space for design optimization, as illustrated in Fig. 2.10 for a 45 nm SOI CortexM0 topology. As more LDH bin divisions are introduced, power is increasingly optimized because of a greater range of gate sizes from which to construct the design. Similarly, the introduction of additional $V_{dd}$ rails and

$V_t$ substantially improves power consumption, although the results do not account for the overhead of the voltage shifter used in multiple $V_{dd}$ design. Finer bin division in the LDH also leads to a better optimized DA (delay-area) curve. In PROCEED, the numbers of $V_{dd}$ and $V_t$ do not impact the DA Pareto curve because they are not associated with area calculation, so they automatically converge to their limiting values to achieve the highest possible performance during area and delay co-optimization. Overall, we observe that the evaluated optimal power at a given minimum working clock period may change by over 50% as gate size tuning and multiple $V_{dd}$ and $V_t$ are introduced, demonstrating the necessity of including these effects in any quantitative comparison.

### 2.4.3 Impact of Interconnect Load

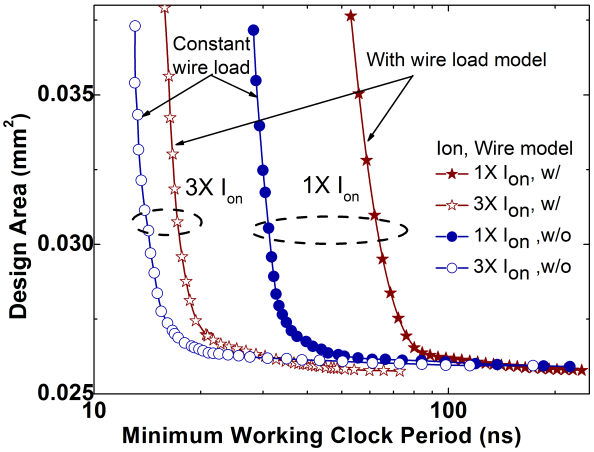



Figure 2.11: 45 nm TFET and boosted TFET (3X current) Pareto curves of delay and chip area for the CortexM0 design. The red curves show the results using a hypothetical TFET with 3X current boost.

As described in Section 2.2.6, PROCEED considers interconnect as a function of gate sizes. Large gates result in large chip area and high interconnect load. For instance, when all gates in a design are sized 1X larger, the delay on interconnect improves less than 1X because interconnect load increases with up sizing gate. The impact of interconnect loads is observed in Fig. 2.11. We compare the area-minimum working clock period tradeoff for the CortexM0 utilizing TFETs with and without the size-scaled interconnect model. Without such model, the interconnect load is set to a constant value independent of gate size, such that sizing leads to an exactly proportional "free" performance boost.

We observe that as the minimum working clock period reduces, the longer interconnects necessitated by gate sizing substantially increase the total area and the assumption of constant interconnect load leads to major inaccuracies. For instance, at a 50 ns minimum working clock period, using the gate size-scaled interconnect model increases total area by 50%.

Since TFETs often suffer from low drive currents and the interconnect model has a larger impact for high performance operation, we also performed the area and delay optimization using a hypothetical device with 3X larger current than the experimental TFETs we have been considering [JLP10]. Note that this and even larger performance boosts should be possible through various device improvements like channel material choice, doping profile, geometry, etc. The benefits of larger current are greater when size scaling of interconnects is considered: the wider shift from the 1X to 3X happens on curves using the interconnect model (which boosts the performance by more than 3X on average) compared to the constant load cases (where current boost simply reduces the delay by the same 3X factor). This is because the reduced gate sizing of the boosted TFET required to reach a given delay also reduces interconnect lengths. A consistent interconnect model that scales consistently with gate sizing can dramatically change chip area and is therefore crucial for evaluating the overall impact of different technologies.

### 2.4.4 Impact of Benchmarks on Evaluation SOI vs. TFET

To show the impact of benchmark selection, we compare the performance of two microprocessors, CortexM0 and MIPS, using SOI and TFET devices and two $V_{dd}$s and two $V_t$s. We choose these benchmarks because, as shown in Fig. 2.12(a), they have a similar number of critical path stages (56 in CortexM0 vs. 62 in MIPS) and total gates (8990 vs. 9248), but the CortexM0 has a more evenly distributed LDH. The power consumption in MIPS is dominated by short paths, which means it will be more accommodating of slow devices compared to the CortexM0. Accordingly, in Fig. 2.12(b), both SOI and TFET achieve better power efficiency in MIPS designs because the second $V_{dd}$ and $V_t$ can be optimized to save power along the short paths. The crossover points where the Pareto curves for different devices intersect define their advantageous operating regions; a device changes from being less power efficient on one side of the crossover to being

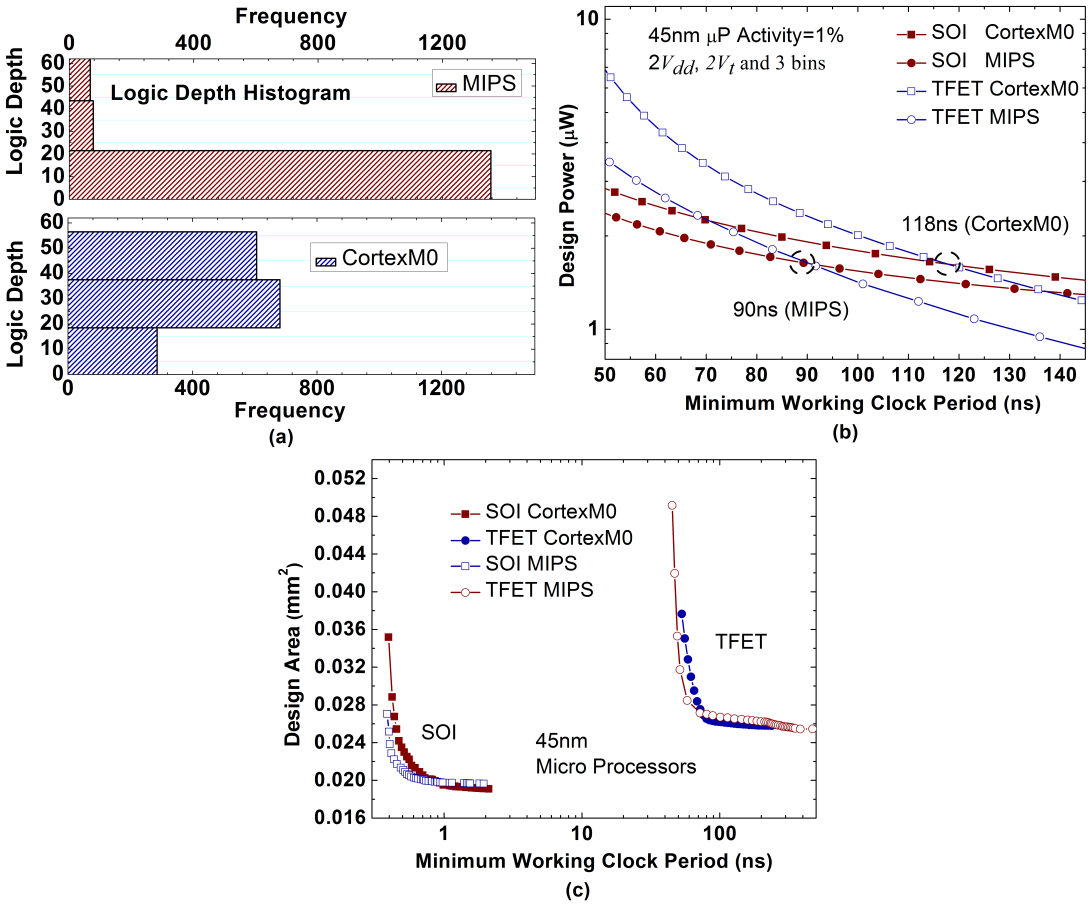



Figure 2.12: (a) LDH of MIPS and CortexM0. (b) Power and delay curves and (c) area and delay curves for MIPS and CortexM0 designed with TFET and SOI, respectively. Activity is 1% and one $V_{dd}$, one $V_t$ and two bins are applied.

more efficient on the other side. If multiple crossovers are found, then the Pareto curve can be divided into several regions (high performance, low power, etc.) such that in each one, there is only a single crossover point. This allows us to demarcate the (possibly multiple) favorable operating ranges for each device. The Pareto crossover occurs at 90 ns and 118 ns for MIPS and CortexM0, respectively, showing that TFETs are more acceptable for applications like MIPS which tolerate slower devices. However, for practical applications, the drive currents for Si TFETs must be increased to reduce sizing and save more dynamic power at high clock rates. As previously mentioned, this may be achieved in practice through a variety of TFET optimization pathways. In Fig. 2.12(c), SOI beats the existing Si TFETs in area and minimum working clock period curves by a wide margin, with no crossover points.

Again, for both SOI and TFETs, better area efficiency is observed in the high per-

formance regime for MIPS on account of the concentration of short paths in its LDH. In the low performance region, all gates are relatively small, so MIPS uses more area than CortexM0 on account of its larger number of gates. By contrast, CortexM0, which contains fewer gates, occupies bigger areas for high performance due to the fact that the longer logic paths in its LDH require larger gates. Previous evaluations, like those in Table 2.1, which ignore LDH, are not able to distinguish between benchmarks in this way. These results show how the choice of circuit topology strongly impacts the suitability of emerging devices.

### 2.4.5 Impact of Activity Factor SOI vs. TFET

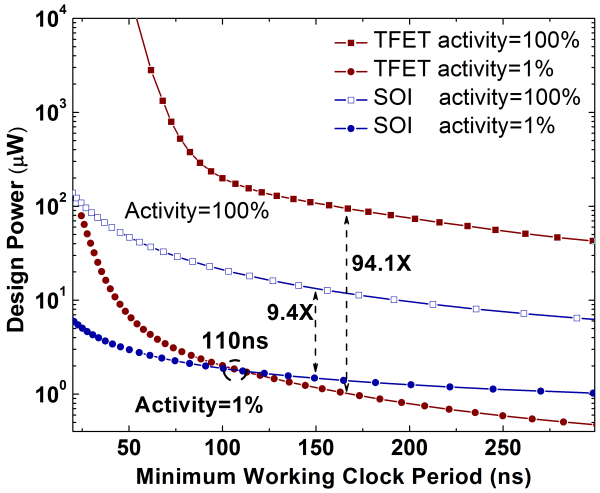



Figure 2.13: Activity impact on 45 nm SOI and TFET CortexM0s power-minimum working clock period.

We next examine how activity factor affects SOI- and TFET-based CortexM0 processors in Fig. 2.13. As activity reduces from 100% to 1%, TFET circuit power scales in lockstep by 94.1X due to low device leakage. However, the corresponding SOI designs only see power reduction of 9.4X because of its higher off-current. We see that TFETs change from being completely impracticable at 100% activity to being superior to SOI beyond the 110 ns minimum working clock period point at 1% activity; thus activity factor, and hence system use contexts, can drastically alter the device evaluation and must be considered

### 2.4.6 Power Management Modeling

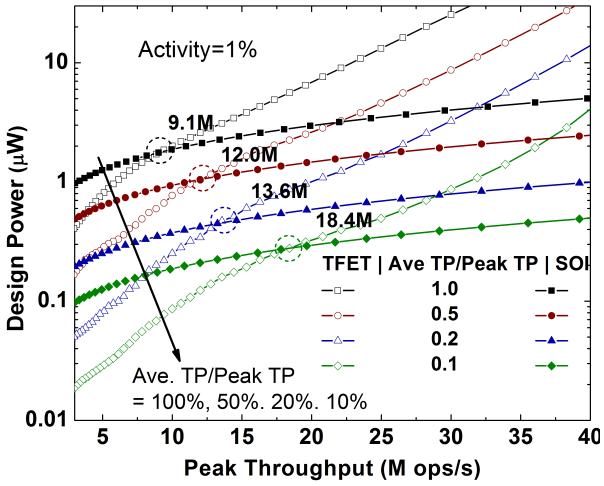



Figure 2.14: 45 nm SOI and TFET CortexM0 microprocessors with power management. The ratios of average to peak throughput are 10%, 20%, 50% and 100%. Curves with ratios of 100% are designs outputted from Pareto optimizer.

The results of the previous subsections make clear that there is no panacea device and that device-circuit evaluation must be done with specific applications and operating windows in mind. DVFS and power gating are crucial ingredients for such usage-mindful evaluation. In Fig. 2.14, we show PROCEED-generated Pareto curves at different ratios of average to peak throughputs for SOI and TFET CortexM0 using DVFS and power gating. Power is reduced by operating at the lower supply rail or turned off by power gating; the achievable power reduction differs with device and operating region. The peak throughput crossover point for TFETs shifts from 9.1M to 18.4M operations per second as the ratio of average to peak throughput reduces from 100% to 10%; the relative performance of TFETs effectively doubles as throughput requirements become less aggressive, emphasizing the importance of incorporating power management into device benchmarking.

### 2.4.7 Variation-Aware Evaluation

To illustrate how variability might impact conclusions drawn using nominal devices, we show in Fig. 2.15 how the SOI and TFET Pareto curves are changed when slow corner devices are used. We define the slow corner as a device with 5% effective voltage reduction

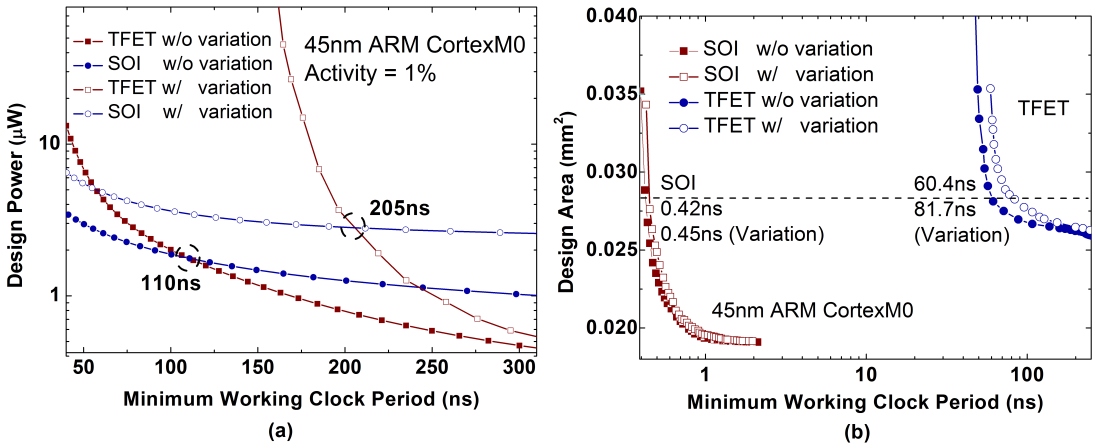



Figure 2.15: Variation-aware (a) power-delay and (b) area-delay evaluations of 45 nm technologies. Assumed voltage drop is 90% and $V_t$ shift is 50mV.

and 50 mV $V_t$ shift; total power is simulated using the nominal device, while delay is evaluated with the slow corner. We observe in Fig. 2.15(a) that the TFET is more vulnerable to variability effects than SOI, as the Pareto crossover of minimum working clock period is shifted from 110 ns to 205 ns when variation is taken into account. In Fig. 2.15 (b), the area and minimum working clock period curve in the presence of variability is shifted more for TFETs compared to SOI. This is due to the TFETs steep subthreshold swing and its low operating voltage, leading a high sensitivity of drive current to voltage [LC13b, LC13a]. This suggests that TFETs need to show substantial nominal device advantages in order to buffer this sensitivity and demonstrates that even a simple consideration of variability is important in device evaluation and selection.

### 2.4.8 Heterogeneous Integration Evaluation

In this section we evaluate three types of technologies: integrated circuits using SOI only, TFETs only, and HGI of the two. From the previous results, the low leakage of TFETs makes them suitable for circuits with low activity or LDH dominated by short paths. HGI offers the chance to merge the strengths of TFETs with the higher performance of SOI to maximize their benefits. TFET is used to build gates in short logic paths, which can save more leakage power without sacrificing performance, because minimum working clock period is decided by long paths where SOI is applied. The optimal DP curves for these three technologies are compared in Fig. 2.16(a). Fig. 2.16(b) shows the corresponding

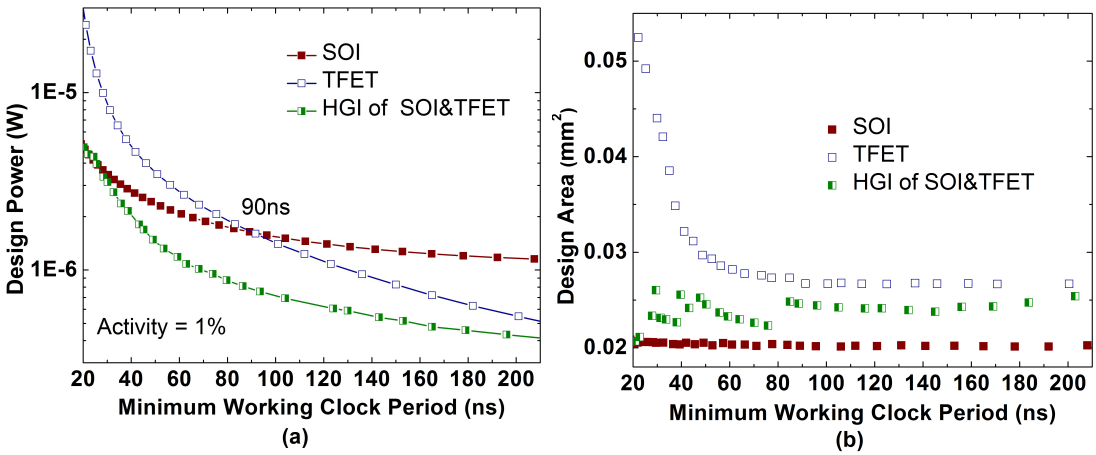



Figure 2.16: (a) Power-delay optimization for 45 nm HGI and non-HGI MIPS and (b) its corresponding design area. The fluctuations in the latter arise because the optimization is carried out for power and delay, not design area. Two sets of $V_{dd}$ and $V_t$ are adjusted during optimization, one for SOI devices and the other for TFETs.

design area and delay for the DP optimized designs for each technology, assuming 45 nm MIPS. Owing to accuracy constraints of the SPICE simulation tool, HGI is evaluated in PROCEED using four bins (which are divided and assigned to either TFET or SOI). In MIPS, gates are mostly distributed along short paths, where devices are mainly idle and leakage power is more significant. The much lower leakage of TFETs gives them a big advantage when designing slow gates, while their performance constraints (due to low current) are mitigated by using SOI for gates along critical paths.

Accordingly, we see in Fig. 2.16(a) that HGI outperforms non-HGI circuits between the minimum working clock periods of 20 ns and 200 ns. In this intermediate region, the respective advantages of TFET and SOI can be combined to give significantly better overall performance. In the (left-most) high performance region, the high drive capabilities of SOI dominate circuit operation, such that optimized HGI designs converge towards the all-SOI counterparts. Similarly, the low leakage of TFETs brings the most benefit for very slow designs lying to the right of the DP curve, so that HGI incorporation of SOI brings negligible benefits. We note that the finite number of bins in our study discretizes the usage of different devices in our HGI designs, limiting the resolution of the latter. For this reason, the HGI results merge with those of non-HGI circuits in the high and low performance limits of Fig. 2.16(a). If the LDH is divided into a larger, near

continuous number of bins, allowing for finer grained designs, the advantages of HGI would become manifest for all operating regions because any small design improvements due to incremental TFET or SOI usage can be evaluated. However, our four bin results and the intuitive arguments above suggest that only small improvements in the very high and low performance regimes would come from HGI for the particular devices under study. Moreover, any performance improvements must be weighed against the accompanying fabrication costs of the more complicated HGI process flow, such that substantial performance enhancements are necessary to justify HGI in practice.

Although we observe obvious advantages for HGI during delay and power optimization, the tradeoff becomes more complicated if the design area is also considered. The areas corresponding to the designs in Fig. 2.16(a) are shown in Fig. 2.16(b). We observe that compared to all-SOI designs, HGI requires more area even when it consumes less power at a given delay. This is because SOI devices have strong driving current and can therefore be sized relatively smaller. For the same reason, HGI designs require less area than all-TFET designs by utilizing some proportion of the smaller SOI gates. By contrast, for very long and short delay periods (corresponding to the leftmost and rightmost regions of Fig. 2.16), HGI optimization leads to all-SOI and all-TFET designs, so the corresponding design area also converges with the non-HGI cases. By quantifying the design area tradeoffs, PROCEED shows that HGI designs receive few benefits in this case because most of the area is consumed by the large number of slow TFET gates.

## 2.5 Chapter Conclusion

The proposed circuit-device co-evaluation framework accounts for circuit topology, adaptability, variability, and use context using efficient Pareto optimization heuristic. Device evaluation ignoring one or more crucial factors like multiple supply and threshold voltages, power management, logic depth, variability, etc., can easily lead to misleading results. For instance, we find that including power management in our evaluation can effectively double the usable operating range for TFETs, and that choice of activity factor can dictate whether TFETs are acceptable at all in a given application. The metrics applied in PROCEED, including delay-power and delay-area tradeoffs, enables a comprehensive

comparison of the benefits and shortcomings of various devices. In addition, we demonstrate how PROCEED enables fast, realistic evaluation of HGI using TFET and SOI technologies as an example. These observations are made possible by the scope of PROCEED scope and its computational efficiency in studying several orders of magnitude in possible device-circuit performance, and demonstrate the capability and flexibility of our new methodology.

# CHAPTER 3

# Evaluation of Digital Circuit-Level Variability in Inversion-Mode and Junctionless FinFET Technologies

## 3.1 Chapter Introduction

AS CMOS technology devices scale ever deeper into the nanometer regime, new transistor designs are being explored to solve the fundamental issues which impede scaling. One innovation, already entering usage, is the inversion-mode (IM) fin field-effect transistor (FinFET), which addresses short channel effects (SCEs) and random dopant fluctuations (RDF) in conventional CMOS. However, all the IM devices still require abrupt and reproducible source/drain junctions, which increase process complexity and face manufacturing limits in the nanometer scale. In response, the junctionless (JL) FET [LAA09], [LFA10] is proposed as a substitute for IM devices; by uniformly doping the entire device and controlling the channel potential purely electrostatically, the JL FET removes these complications of IM FET.

However, all these technologies still face the bane of process variations, which become more important with shrinking feature size, rendering device, and circuit performance increasingly unpredictable. It is well known that FinFET performance suffers from variations due to line edge or line width roughness (LER/LWR). The effect of LER on IM FinFET-based circuits is analyzed [LLG12] with the primary impact being an increase in mean leakage power. However, JL FinFETs are inherently more sensitive to variability, with device-level simulations revealing threshold voltage standard deviations over six times those of IM FinFETs [LC12, LC11]. In contrast to the robustness of IM devices [PLS09a], JL FinFETs are highly sensitive to RDF [LC12], which also impacts their drive

and leakage current, and drain-induced barrier lowering (DIBL). Finally, because of reduction of gate control over the body-centered channel, JL FinFETs show worse SCE compared with IM FinFETs [RCA11]. Therefore, it is crucial to evaluate JL variability at the circuit level to decide if JL transistors are a viable alternative to IM CMOS.

In this chapter, we present the first variability-aware circuit studies [WLP13] of JL FinFETs in multiple technology nodes (32, 21, and 15 nm) and compare the results with IM FinFET circuits, introducing calibrated LER and RDF effects in our simulations. Both large-scale digital circuits (e.g., microprocessors) and six transistor (6T) static random access memory (SRAM) cells are evaluated using an original evaluation framework. Our results indicate that the bottleneck for JL FinFET-based circuits rests in SRAM designs needing much higher $V_{ccmin}$ compared with IM FinFET-based circuits, whereas large-scale microprocessors are robust against stochastic variation regardless of the specific FET implementation.

## 3.2 Overview of Evaluation Framework

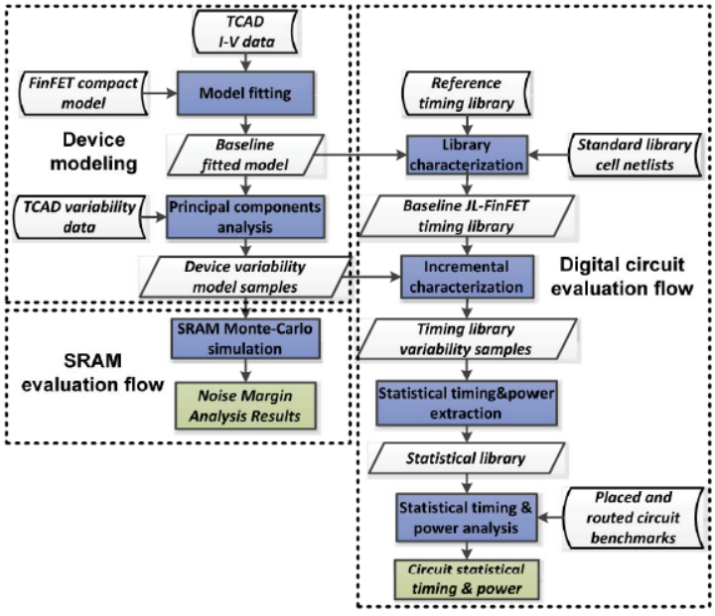



Figure 3.1: Overview of the variability evaluation framework used in this paper. The evaluation of (left) 6T SRAM cells and (right) microprocessor circuits are divided into two vertical branches as illustrated.

The overview of our circuit-level variability evaluation framework is in Fig. 3.1. Tran-

sistor IV characteristics and variability data from device-level technology computer-aided design (TCAD) simulations are used as the starting input for subsequent compact modeling. To create a baseline model, we fit a BSIM model to the predictive technology model (PTM) [refd] to match the TCAD $I_DV_G$ and $I_DV_D$ data. Using the method in [LLG12] to capture the effect of LER/RDF in our compact model, model samples were generated such that their predicted behavior matches the original TCAD simulation results. 6T SRAM cell Monte Carlo simulations are performed by generating individual model samples for each of the six transistors, after which the static noise margins extracted. For logic circuit timing and power analysis, we first create and characterize a baseline timing library from a baseline model and template library. Then, through incremental characterization based on model samples, library samples are generated such that the resulting circuit behavior should correctly reflect the performance impact from LER/RDF. Statistical timing and power information is extracted from these library samples, which are then fed as inputs to a computationally efficient statistical timing and power analysis tool based on [VRK04], [CS05] to evaluate the overall impact of LER/RDF on large-scale digital circuit delay and power consumption. The following sections explain the individual stages of our framework in more detail.

## 3.3 Variability and Device Modeling

### 3.3.1 LER and RDF Modeling

To introduce the effect of LER in our FETs, we first generate 200 random LER patterns with root-mean-square roughness amplitude $\sigma_{LER}$ up to 0.6 nm and correlation length $\lambda$ = 15 nm using the method of Fourier synthesis [AKB03] with a Gaussian autocorrelation function. These values represent typical LER values which may be required by industry heading beyond 32-nm technology, based on the 2011 ITRS [Com11] forecast and experimental data [AKB03]. We fixed the correlation length $\lambda$ = 15 nm as previous studies [BDR07], [PLS09b] have shown that the effect of $\lambda$ diminishes as $\lambda$ ¿ 1520 nm, and some experimental data has shown that current values of $\lambda$ are estimated between 2030 nm [AKB03] and generally reduces with technology, suggesting $\lambda$ = 15 nm as a reasonable estimate for sub-32-nm lithography.

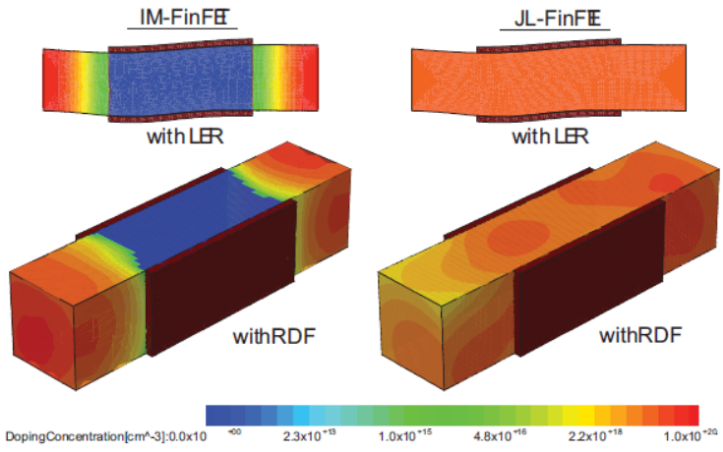



Figure 3.2: Simulated 32-nm IM and junctionless FinFETs with LER and RDF. $H_{fin}$ = 10 nm and $\sigma_{LER}$ = 1 nm are used in the above structures.

The LER patterns are then used as templates to augment the fin sidewalls in our double-gate FinFET structures as shown in Fig. 3.2, thus yielding random performance for individual devices. Here, each FinFET technology was designed according to the ITRS forecast for 32, 21, and 15-nm high-performance logic nodes with specific details provided in [LC11]. Only fin LER along the channel transport direction was considered in this paper for reasons described in [LLG12] and [LC11]. In addition, we assume all line edges to be uncorrelated within individual devices as well as between devices hence the LWR amplitude $\sigma_{LWR} = 2^{1/2}\sigma_{LER}$; this represents the situation of standard resist patterning. The effects of spacer patterning are not explicitly dealt here with the understanding that the device and circuit-level LER impact will likely be minimal [LLG12], even for JL FinFETs.

The impact of RDF was captured using the same approach in [LC12] which randomizes the placement and concentration of ionized dopants based on a Poisson distribution. The locally varying doping concentration is calculated from the long-range part of the Coulomb potential with an appropriate screening length [SMM02]. Because of the high doping concentration and small device volumes in our JL FinFETs, the variability impact of RDF is significant from a device-level perspective. This contrasts with the situation for IM FinFETs where the channel is typically undoped and RDF only exists in the source and drain extensions Fig. 3.2. For JL FinFETs, a nominal doping level of $N_D$ = $2 \times 10^{19} cm^3$ yields optimal performance in terms of ION for a given $I_{OFF}$ ( 100 nA/m)

while satisfying ITRS design specifications.We found that higher doping levels (e.g., $N_D = 3\times10^{19}cm^3$) result in slightly worse nominal performance as well as heightened variability, while lower doping levels (e.g., $N_D = 5 \times 10^{18}cm^3$) result in even higher ION penalties (20%40%), but reduced variability. We also find that any channel doping lower (higher) than roughly $1 \times 10^{19}cm3$ results in accumulation-mode (depletion-mode) behavior for the device geometries considered. With this in mind, the performance versus variability tradeoff for JL technologies may be a critical factor for the optimal design of such devices, and further work will be needed to identify the best strategy for JL FET design (beyond current ITRS guidelines). Unfortunately, such an investigation is beyond the scope of this paper and the remainder of our study will employ FinFETs designs [LC11] which best match the nominal scaling guideline published by the ITRS.

### 3.3.2 Device-Level Variability

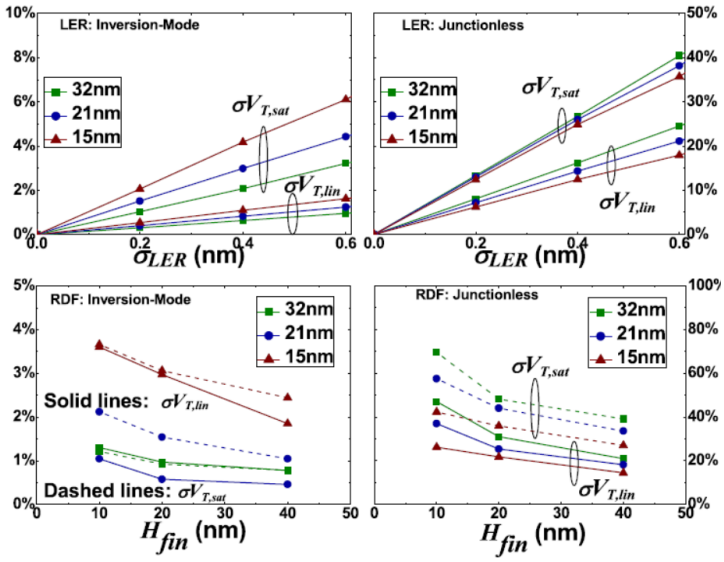



Figure 3.3: Threshold voltage variation of IM and junctionless FinFETs due to LER (upper row) or RDF (bottom row). Only one source of variability (LER or RDF) is active at a time. Note the scale for JL FinFETs is larger than that for IM FinFETs.

We previously quantified the variability impact of LER and RDF for sub-32-nm IM and JL FinFET technologies in [LLG12]C[LC11] using 2-D and 3-D TCAD simulations for LER and RDF, respectively. In those simulations, quantum corrections are modeled using the density gradient approximation, high-field transport with a calibrated hydrodynamic

model [LLG12, NA07], and carrier mobility with doping dependent, surface scattering, and high-field terms. A small subset of our results is shown in Fig. 3.3, comparing the threshold voltage variability of IM and JL FinFETs due to LER and RDF. JL devices (with $N_D = 2 \times 10^{19} cm^3$) exhibit significantly higher variability com-pared with similarly designed IM devices. In fact, some JL devices within $a \pm 3\sigma$ spread may have a negative VT (peak $3\sigma V_{T,sat} > 100\%$) and be permanently on even at zero gate voltage, constituting switching failure; this may occur due to a surplus of dopants inside the channel from RDF or an unusually wide fin from LER. This revelation is due to the different methods by which LER and RDF affect the intrinsic operation of IM versus depletion-mode FETs [LC11]. Similar conclusions are obtained for other performance metrics including $\sigma I_{ON}$, $\sigma I_{OFF}$, $\sigma SS$, and $\sigma DIBL$; data is available in the listed references. With these device-level variability figures, we determine the resulting circuit-level impact in Sections 3.4 and 3.5.

As mentioned in the previous section, we found that JL-FinFETs with lower (higher) doping resulted in less (more) overall variability from LER and RDF. For 32-nm JL-FinFETs with $N_D = 5 \times 10^{18} (3 \times 10^{19}) cm^3$, LER-induced $\sigma V_{T,sat}$ drops (rises) to 12% (60%) at $\sigma_{LER} = 0.6$ nm. Similar changes in JL variability from LER are witnessed for other technology nodes and performance figures as well, suggesting the viability of JL technology will depend on the design strategy employed. A full set of results for RDF-induced variability is not available at this time, but preliminary findings suggest similar trends when the baseline doping is changed.

### 3.3.3 Device and Variability Model Fitting

PTM FinFET models [refd] are fitted to the TCAD-simulated transfer and output characteristics. To match the currents from the 2-D TCAD simulations (in units of A/m) to the 3-D device model, we linearly scale the currents to match single fin transistor characteristics, where we assume $H_{fin}$ to be equal to the feature size in each technology node (e.g., $H_{fin} = 32$ nm for 32-nm FinFETs). Seventeen parameters of the PTM model are chosen as fitting variables according to the PTM and BSIM parameter extraction guide [refd, XDH05], with tuning ranges for each chosen parameter listed in Table 3.1. Our error metric for the fitting procedure is the weighted least square difference between

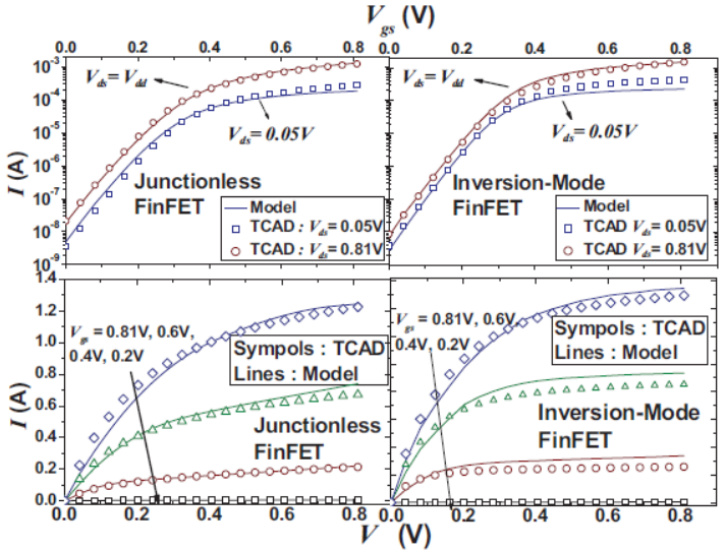



Figure 3.4: Matching of baseline FinFET (a) transfer and (b) output curves between TCAD simulation and compact modeling.

the simulated and model $I_D$-$V_{GS}$ and $I_D$-$V_{DS}$ curves, with random starts and gradient descent methods being applied. Good matching between the compact models against TCAD simulations are obtained, as shown in Fig. 3.4.

With the baseline compact model established, the baseline cell library is characterized using Nangate Open Cell Library [refc] as the template, similar to [LLG12]. Extraction of device-level variability is based on principle component analysis [LLG12], [PML90, TH08]. The model samples are generated [LLG12] hence the resulting device performance variation matches the data from TCAD simulations. The statistical matching results are shown in Fig. 3.5. Standard deviations of ION and VT,sat are calculated from 400 model samples. The maximum error is only 8.2% in $\sigma I_{ON}$ for JL FinFETs, validating our JL Fin-FET circuit model. Unfortunately, when matching $\sigma V_{T,sat}$ for 15-nm IM FinFETs, a maximum error of 25.8% is observed for $\sigma_{LER} = 0.6$ nm; however, since variation has very limited impact on IM FinFETs, we find that this relatively large matching error does not change our conclusions. For both IM and JL FinFETs, $\sigma I_{ON}$ increases with technology scaling whereas $\sigma V_{T,sat}$ increases (decreases) in IM (JL) FinFETs. This unexpected trend for $\sigma V_{T,sat}$ was also reported in [LC12] and [PLS09a], and can be explained by noting that smaller nodes with thinner bodies helps suppress the effects of LER/RDF due to the closer gate-to-channel proximity in JL devices with buried channels [LAA09]. For IM devices with surface channels, the gate-to-channel proximity is relatively insensitive to

Table 3.1: Allowed tuning range of fitted compact model parameters. [1] Parameters in PTM model.

| Param. | Range | Param. | Range | Param. | Range |
|---|---|---|---|---|---|
| nch | 0.1-10x | len | 0.7-1.6x | tox | 0.7-1.6x |
| tsi | 0.5-2x | tbox | 0.5-2x | vth0(f)[1] | ±0.25 V |
| vth0(b)[1] | ±0.25 V | esi[1] | 0.8-1.4x | eox[1] | 0.8-1.4x |
| Lambda | 0.5-2x | N1 | 0.9-1.1x | Vt[1] | ±0.25 V |
| voff1[1] | ±0.1 V | u0 | 0.7-1.6x | eta0 | ±0.1 V |
| dsub | ±0.1 V | rdsw | 0.7-1.6x | | |

the body thickness and, therefore, the effects of LER/RDF are not suppressed at smaller technologies (they are only degraded from SCE).

## 3.4 Variability Impact on 6T SRAM Memory

### 3.4.1 Baseline Nominal Static Noise Margin

As CMOS technology continues to scale down, SRAM design becomes progressively more complicated. To guarantee proper operation, the cell design must meet noise margin requirements that are budgeted for all fluctuation sources, including supply, process, and temperature variations. Increasing variability therefore, strongly degrades performance. For instance, static noise margin (SNM), one of the important metrics for SRAM cell stability, decreases with successive technology generations [CCL08]. Fig. 3.6 shows how nominal SNM changes with supply $V_{cc}$ from 32 to 15 nm for JL FinFET 6T SRAM. With increasing $V_{cc}$, the SNM diverges for different technologies with differences of up to 20 mV at $V_{cc} = 0.9$ V.

In addition to these generic challenges, FinFETs face an additional disadvantage because of their digitized fin structures. Traditionally, device widths are sized to achieve high stability; for example, symmetric (SYM) designs might continuously scale PMOS widths to be larger size than NMOS to equalize the drive current. Realizing this with FinFETs requires parallelizing fins at the cost of cell area, for instance matching three

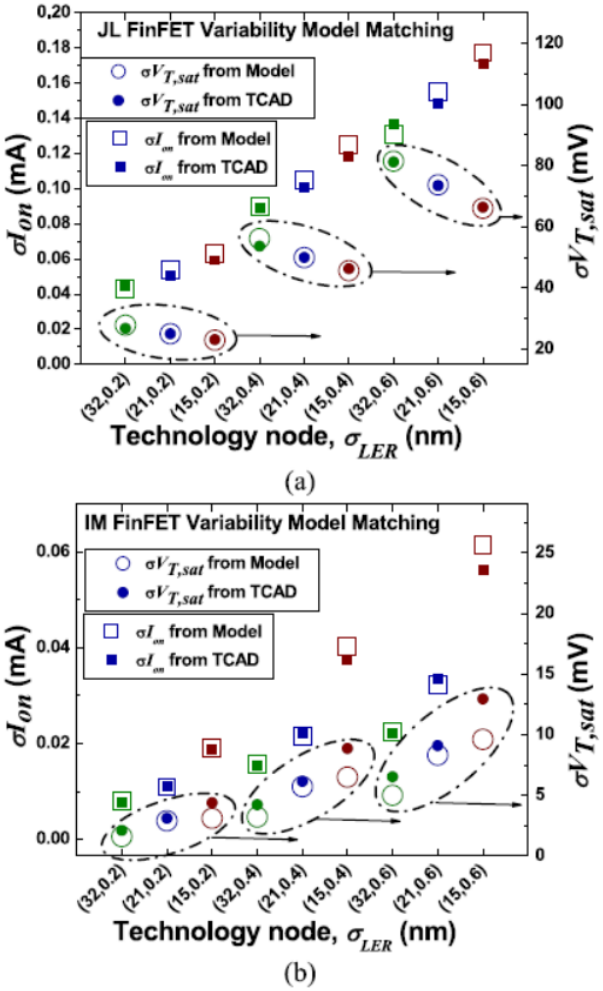



Figure 3.5: Comparison of $\sigma I_{ON}$ and $\sigma V_{T,sat}$ extracted from 200 samples between TCAD simulations and fitted variability models for (a) JL FinFETs and (b) IM FinFETs show a good fit.

PMOS with two NMOS fins; instead, typical designs now use one fin for each gate to maximize density [KWN12, HKA08]. In the following discussion, all SRAM results are generated based on this high density (HD) layout unless otherwise specified.

### 3.4.2 Minimum Working $V_{cc}$ ($V_{ccmin}$)

As cell density increases, power consumption becomes a crucial consideration requiring reduction of $V_{cc}$ to conserve both dynamic and leakage power. The minimum working supply voltage $V_{ccmin}$ is thus an important metric for judging the viability of a cell design. In general, for a fixed SNM, $V_{ccmin}$ increases with scaling. Fig. 3.6 shows for instance how enforcing SNM of 0.2 V causes $V_{ccmin}$ to increase from 0.516 V at the 32-nm node

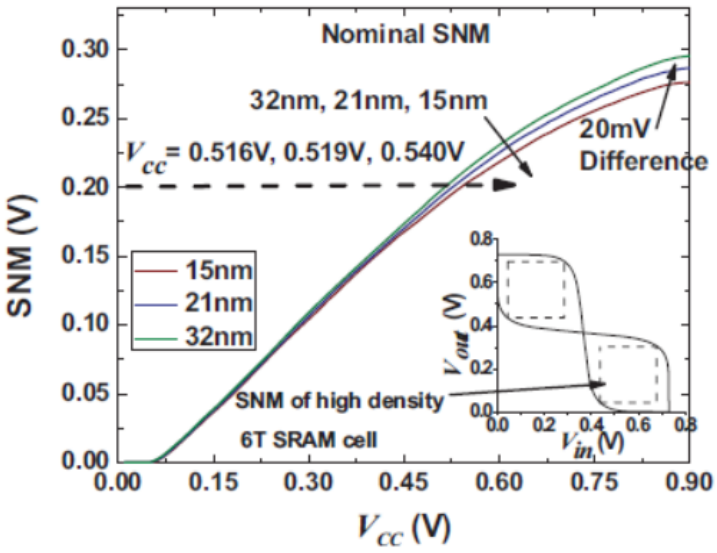



Figure 3.6: Nominal SNM as a function of working $V_{cc}$ for high density design JL FinFET 6T SRAM cells. Note that for successive technology nodes, SNM and $V_{cc,min}$ decrease when the other is held fixed.

to 0.540 V at 15 nm. In addition to SNM, static/dynamic read and write noise margins also affect $V_{ccmin}$; however, considering all such metrics would raise many more design issues outside the scope of this paper. Therefore, we will only consider the effect of SNM on $V_{ccmin}$.

We use Monte Carlo simulations to search for $V_{ccmin}$ underspecified yield and SNM constraints. HSPICE is used for dc simulations of 6T SRAM cells where each individual device is independent and uses a randomly selected device model, as explained in Section 3.3.3. The SNM is measured as the length of the largest square in the butterfly curve, as shown in the inset of Fig. 3.6. A simulated cell with SNM below the given constraint counts as a failed cell. A given supply voltage is said to work for SRAM cells if the number of successful simulations with this $V_{cc}$ reaches the yield requirement (e.g., 99.9% yield requires 9990 successful simulated cells out of 10,000 randomly generated cells). To find the $V_{ccmin}$, we use a binary search (40X faster than exhaustive search). To further improve the runtime of yield analysis, we use the statistical blockade method [SR09] which uses rejection sampling, speeding up the total process by over 10X.

In Fig. 3.7, $V_{ccmin}$ is reported for JL and IM SRAM cells with different technology nodes and LER amplitudes. The improved $V_{ccmin}$ for IM-based SRAM compared with

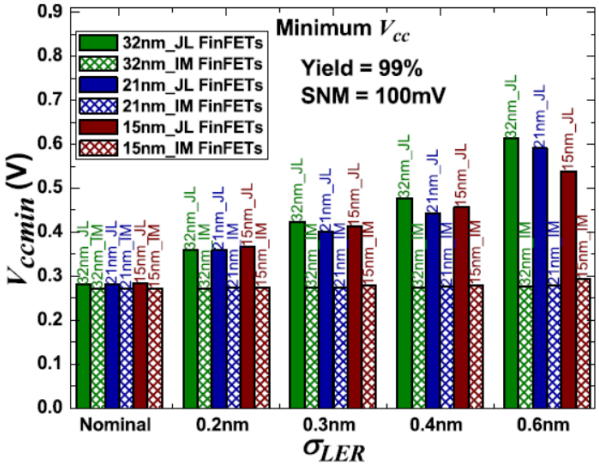



Figure 3.7: $V_{ccmin}$ as a function of technology node and LER amplitude for JL and IM FinFET 6T SRAM. The SNM constraint is 100 mV, and yield is 99%.

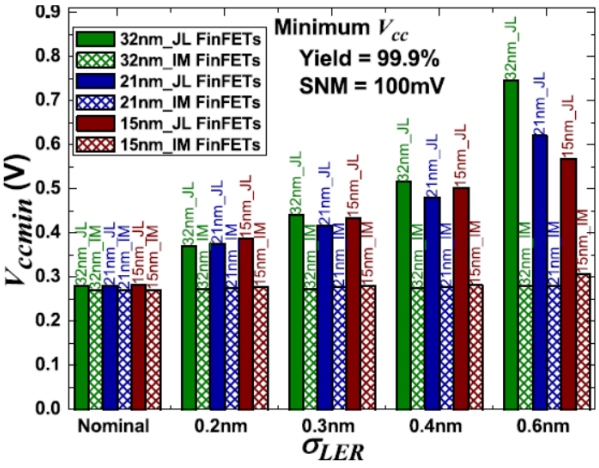



Figure 3.8: $V_{ccmin}$ as a function of technology node and LER amplitude for JL and IM FinFET 6T SRAM. The SNM constraint is 50 mV, and yield is 99.9%.

JL-based SRAM is explained by the fact that IM devices are more robust against LER-induced variability [LC11]. This shows that JL transistors in current technology nodes would not be a good option for memory design. Interestingly for JL technologies, at low LER amplitudes the 32-nm devices perform best, whereas at high LER amplitudes the trend is reversed and the newest generation (15 nm) devices have the lowest $V_{ccmin}$. This trend is more obvious in Fig. 3.8, where the more stringent requirement of 99.9% yield exacerbates the effect of variations on SNM.

This trend can be understood by remembering that $V_{ccmin}$ is dictated by both variability and the nominal SNM. We have already seen that nominal SNM degrades under

Table 3.2: Nominal SNM and SNM loss from variability for JL FinFET technologies.[1] High density 6T SRAM design. [2] Symmetric N/P design. [3] SNM at $V_{cc}$=0.73 V. [4] SNM with 99% yield constraint: LER variation ($\sigma_{LER}$=0.6 nm ) at $V_{cc}$=0.73 V

| | 32nm | | 21nm | | 15nm | |
|---|---|---|---|---|---|---|
| | HD[1] | SYM[2] | HD | SYM | HD | SYM |
| Nominal SNM[3] [V] | 0.264 | 0.268 | 0.26 | 0.262 | 0.251 | 0.252 |
| SNM w/ variation[4] [V] | 0.128 | 0.154 | 0.144 | 0.166 | 0.14 | 0.176 |
| % SNM loss | 51.5% | 42.5% | 44.6% | 36.6% | 44.2% | 32.2% |

size scaling and dominates the trends in Figs. 3.7 and 3.8 at small $\sigma_{LER}$, but JL devices also become less sensitive to variability as technology scales [LC11], allowing the operating conditions to relax. Our largest considered $\sigma_{LER}$ of 0.6 nm is in line with the ITRSprojected $\sigma_{LER}$ requirements of 1, 0.8, and 0.5 nm for the 32, 21, and 15-nm nodes, respectively. Therefore our results hold out hope that for realistic variability levels, JL SRAM technologies will become more competitive if scaling trends continue.

### 3.4.3 SNM Versus Technology

We also explored SYM SRAM designs using three PMOS with two NMOS fins, which can optimize nominal SNM and mitigate the effects of variability due to statistical averaging over the multiple fins. To characterize the impact of variability on the design, we define SNM loss as the percentage difference between the nominal SNM and the variability-affected SNM. Table 3.2 compares SNM loss for JL HD and SYM cells. We find as expected that under scaling and/or use of SYM designs, SNM loss is significantly reduced. On the other hand, the SYM design sacrifices read noise margin and cell area.

To better understand the impact of process variability on JL FinFETs, we also attempted to incorporate both RDF and LER effects in our simulations, assuming the fluctuations to be uncorrelated. This assumption of statistical independence may not be

strictly justified, but forms a best-case scenario for real-world situations. Even under this relaxed assumption, we find that no realistic $V_{ccmin}$ can be realized for 99% yield and 100-mV SNM, reinforcing our conclusion that process variations will be a serious roadblock for JL FinFETs in memory applications.

## 3.5 LER Impact on Logic Circuit Variability

### 3.5.1 Overview

A typical way to analyze the statistical timing and power of circuit benchmarks uses a large number of library samples based on the Monte Carlo method [LLG12]. However, this method is time-consuming and results in round-off errors when synthesizing tool outputs, losing statistical information. To fix these errors, more simulations are needed, with the quantity dependent on the size of the variability impact. In this paper, we use block-based statistical timing and leakage analysis [VRK04], [CS05] to complete this step, drastically improving computational efficiency; in some cases, simulations that would previously require weeks of computation can be reduced to several tens of seconds.

### 3.5.2 Circuit Statistical Timing and Power Analysis

To build the input to the statistical timer, the timing and leakage standard deviation for cells need to be extracted from library samples (we use 200 library samples in this step). We observe that timing variation is highly sensitive to input slew and output load capacitance. Hence, to find accurate timing variation information, a cubic model of delay standard deviation as a function of load capacitance and input slew is fitted to statistical timing information extracted from library samples. This model is found to be accurate enough for the following analyses. Leakage variation is modeled as a lognormal distribution with the standard deviation and mean extracted from the library samples.

The input to the statistical timer includes extracted timing models, extracted leakage lognormal standard deviations, a synthesized and routed circuit benchmark, the baseline library, timing constraints, and SPEF file containing parasitic information. For our benchmarking we select two processors, MIPS [refb] and CortexM0 [refa]. To cover all

Table 3.3: Circuit benchmarks.

| Tech. node | Freq. for CortexM0 [GHz] | | | Freq. for MIPS [GHz] | | |
|---|---|---|---|---|---|---|
| | Fast | Typ | Slow | Fast | Typ | Slow |
| 32 nm | 0.92 | 0.79 | 0.7 | 1.02 | 0.79 | 0.75 |
| 21 nm | 1.47 | 1.3 | 1.12 | 1.61 | 1.44 | 1.09 |
| 15 nm | 2.29 | 2.23 | 1.85 | 3.29 | 3.07 | 2.04 |

working applications, we synthesize them in three operating clock frequencies for fast, typical, and slow speeds as shown in Table 3.3.

### 3.5.3 Circuit Simulation Results

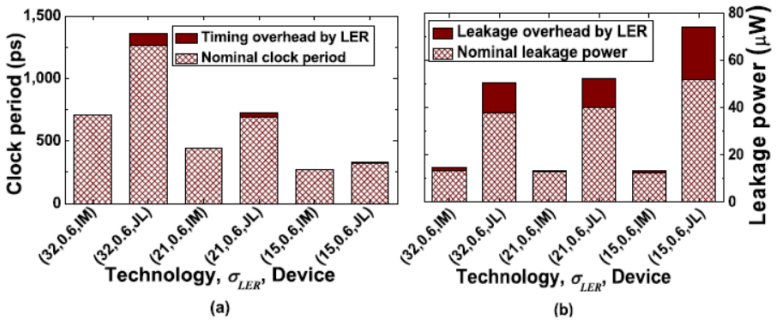



Figure 3.9: (a) Nominal clock period and clock period increase (mean shift and variation) and (b) nominal leakage power and leakage power increase (mean shift and variation) due to LER variation ($\sigma_{LER}$ = 0.6 nm) for IM and JL FinFET-based MIPS processors at typical clock speeds.

Fig. 3.9 shows our results for MIPS designs. The clock period increase due to device variability is calculated as the sum of mean shift and delay uncertainty (3 $\sigma_{clock}$), covering around 99.9% of the possible clock period cases. All uncertainty in our timing results is below 1.20% of nominal delay. The mean clock period shift contributes the most; the highest mean shift is 7.04%. Thus, a delay margin of up to 8.2% may be needed to guarantee sufficient yield in the presence of LER. JL-based processors show a greater improvement in nominal speed with scaling compared with IM-based circuits.

The leakage power is assumed to follow a lognormal distribution. The uncertainty is calculated based on [CS05] at 99.9% yield point of leakage cases. Leakage increase is the

sum of the mean shift and leakage uncertainty. As shown in Fig. 3.9(b), leakage power is severely impacted by LER. Our results show the increase mainly comes from a mean shift, in which the highest observed shift value is 43.02% of the nominal leakage. Leakage uncertainty has a considerable impact, inducing up to 15.57% increase. However, we expect that the leakage uncertainty will be negligible in industrial-scale designs (random leakage variation averages over number of devices in the design). High leakage variations are also predicted by device level simulations, where $\sigma I_{OFF}$ is over 10X nominal leakage for individual JL FinFETs [LC11].

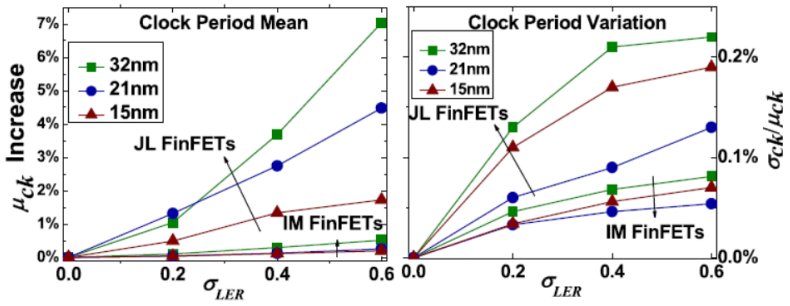



Figure 3.10: (a) Increase in clock period mean and (b) variation of critical clock period as a function of technology node and LER amplitude for JL and IM FinFET circuit benchmark (Cortex M0).

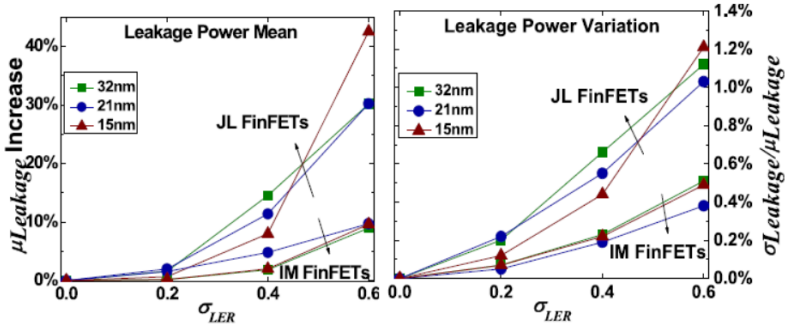



Figure 3.11: (a) Increase in leakage power mean and (b) variation of leakage power as a function of technology node and LER amplitude for JL and IM FinFET circuit benchmarks (Cortex M0).

Figs. 3.10 and 3.11 show the JL-based high speed Cortex-M0 results for clock period mean and leakage mean compared with IM-based processors [LLG12]. JL devices are more severely affected by variability in terms of both mean shift and standard deviation, with circuit clock period mean shift over 10X that of IM FinFETs. Table 3.4 shows the

Table 3.4: Average mean shift and standard deviation of timing and leakage for six benchmark circuits.

| Tech. node | $\sigma_{LER}$ | Timing | | Leakage | |
|---|---|---|---|---|---|
| | [nm] | $\mu_{delay}$ | $\sigma_{delay}$ | $\mu_{leakage}$ | $\sigma_{leakage}$ |
| 32 nm | 0.2 | 1.01% | 0.12% | 1.4% | 0.2% |
| | 0.4 | 2.56% | 0.17% | 12.6% | 0.6% |
| | 0.6 | 4.44% | 0.22% | 26.2% | 1.0% |
| 21 nm | 0.2 | 1.26% | 0.13% | 1.7% | 0.2% |
| | 0.4 | 2.30% | 0.20% | 9.6% | 0.5% |
| | 0.6 | 3.62% | 0.27% | 25.3% | 0.9% |
| 15 nm | 0.2 | 0.70% | 0.17% | 0.6% | 0.1% |
| | 0.4 | 1.32% | 0.25% | 6.8% | 0.4% |
| | 0.6 | 1.60% | 0.28% | 36.8% | 1.1% |

average results from all six circuit benchmarks. For example, at $\sigma_{LER} = 0.6$ nm (near the ITRS predicted LER requirement of 0.5 nm), a 36.8% leakage mean increase is observed at the 15-nm node. However, these impacts are not severe at the logic circuit level.

We have simulated the combined effects of RDF and LER variability, but the huge variations encountered (e.g., normalized $\sigma V_{T,sat} = 70\%$) can lead to statistically significant failure rates in SPICE convergence. Therefore these results are not presented. However, as previously observed [BDM02, BSH04, BKM07], the mean increase of timing variations for circuits is linearly related to the variation of a single logic gate. We can estimate the combined variability to have 3X impact on timing compared with our results considering only LER. For leakage power, a model-based analysis [CS05] using our library extraction results shows the effects of combined variability will have 2X impact on leakage mean compared with the standalone LER variations.

## 3.6 Chapter Conclusion

Device-level TCAD simulation showed that JL FinFETs were more susceptible to process variability (LER and RDF) than IM FinFETs. Fluctuation in threshold voltage reached

up to 40% and 60% due to LER and RDF, respectively. The large-scale digital circuit benchmarks showed LER induces ¡10% mean shift in timing and below 1% standard deviation over the nominal clock period. Leakage power mean shift up to 43% with standard deviation ¡2% (i.e., following lognormal distribution) was observed. The results suggested that large-scale digital circuits will not be affected much by LER-induced variability and that manageable timing and power margins may resolve the issue. However, for memory cells which had fewer transistors, the large degree of device fluctuation resulted in a stronger circuit-level impact. Under the LER target reported by the 2011 ITRS, JL FinFET SRAMs required twice the $V_{ccmin}$ compared with IM FinFET SRAMs. After considering LER and RDF combined variability, JL FinFETs totally fail to produce yields higher than 99%. Fortunately, technology scaling alleviates the effect of LER and RDF variability, with JL FinFET SRAMs at the 15-nm node achieving better noise margin and $V_{ccmin}$ compared with the 32-nm node. On the other hand, IM FinFET SRAMs became more vulnerable going from 32 to 15 nm. This suggested that JL FET technology may eventually become a viable solution in future digital logic generations, especially if circuit-level memory robustness enhancement solutions were considered.

# CHAPTER 4

# MEMRES: A Fast Memory System Reliability Simulator

## 4.1 Chapter Introduction

Evaluation of modern memory system reliability is nontrivial, because memory failure is caused by a large variety of memory fault types [SL12a, SPW09, SDB15], and various sophisticated techniques (see Fig. 4.1) have been proposed to repair the faults. Most previous evaluation studies have relied on analytical models, e.g., [JDB13]. As illustrated in Fig. 4.1, the models are insufficient to handle a variety of memory fault types simultaneously, to capture the effects of reliability enhancement techniques, and to incorporate application influence. Firstly, memory failure is caused by multiple fault types and their interaction, however developing models that include all fault types and interactions is a impractical task, and missing a fault type may result in significant accuracy loss. Secondly, memory failure model is strongly dependent on reliability enhancement techniques; sophisticated error-correcting code (ECC), e.g., double-device data correction (DDDC) [JK13, Wil14], and memory reliability management (see Fig. 4.1) dramatically add to the modeling difficulty. Thirdly, fault rate varies with application and time [SPW09, MWK15, MWK15], but analytical models assume constant fault rate.

Experimentally analyzing memory faults requires a large statistical experiment setup. Field studies [SL12a, SDB15, SPW09, MWK15] have recorded and analyzed memory errors in data centers for over a year. Despite the high cost, conflicting conclusions exist in these studies for lacking the access to finer granularity of memory fault interaction and uncontrolled hardware design variables. In addition, due to the dependence of memory fault on application and memory architecture [MWK15], the conclusion from a field study is difficulty to use to predict the reliability of other computing systems. Therefore,

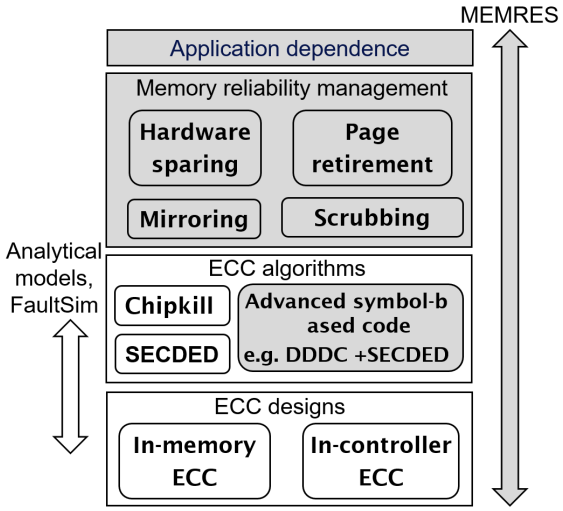



Figure 4.1: The limitation of analytical models and FaultSim on memory reliability evaluation. The reliability enhancement techniques in gray boxes can only be evaluated by MEMRES.

efficient and flexible simulation methodologies that can perform finer analyses are required in memory reliability study.

FaultSim is a recent developed high-speed Monte-Carlo DRAM fault simulator [RN14, NRQ15]. It takes a few hours (seconds in the event mode) to obtain years-long DRAM failure probability. However, FaultSim does not support modern memory system fault simulation due to several missed models, e.g., memory access behavior and memory reliability management.

In this chapter, we introduce MEMRES [WHZ16], an efficient memory reliability simulator that is able to handle the advanced techniques used in state-of-the-art memory systems while minimizing the computation involved in the simulation. It performs long-term (i.e., month-year) memory system reliability simulation. Table 4.1 compares analytical model [JDB13], FaultSim [RN14, NRQ15], and MEMRES. In addition to supporting all features of the analytical model and FaultSim, MEMRES enables simulation with varying fault rate/density and memory access density/distribution, in-memory and in-controller ECCs, and modern reliability enhancement techniques. As examples, MEMRES differently models fault with faulty bit density according to the truth that a fault usually has $< 1\%$ faulty bits in its coverage (e.g., a bank fault only affects $< 1\%$ rows) [SL12a, SPW09], and MEMRES adds the memory access into fault simulation, where

Table 4.1: Comparison with existing memory fault analysis methods. Run time is measured using single-core on AMD Opteron(tm) Processor 2380. FaultSim has event mode, which uses analytical models to partially replace Monte-Carlo fault injection in regular interval mode to speedup simulations.

| | | Analytical model [JDB13] | FaultSim [RN14, NRQ15] | MEMRES |
|---|---|---|---|---|
| ECC | SECDED, Chipkill | ✓ | ✓ | ✓ |
| | Advanced ECCs, e.g., V-ECC [YE10], SWD-ECC [GSD16] | × | ✓ | ✓ |
| | In-Controller ECC | ✓ | ✓ | ✓ |
| | In-Memory ECC | × | × | ✓ |
| Fault model (FIT) | Constant FIT | ✓ | ✓ | ✓ |
| | Temporal variation (Fig. 4.16) | × | × | ✓ |
| | Spacial variation | × | × | ✓ |
| | Data-link and retention errors | × | TSV errors | ✓ |
| Memory access model | Uniform access | ✓ | ✓ | ✓ |
| | Temporal variation | × | × | ✓ |
| | Spacial variation (Fig. 4.12) | × | × | ✓ |
| Memory reliability management | Scrubbing | × | ✓ | ✓ |
| | Hardware sparing (Fig. 4.13) | × | × | ✓ |
| | Page retirement (Fig. 4.14) | × | × | ✓ |
| | Mirroring (Fig. 4.15) | × | × | ✓ |
| Run time | | <1ms | 6.6 hrs (interval) 6.7 secs (event) | 5.2 hrs (1 thread) 50 mins (8 threads) |

a fault is only activated after being accessed. To support finer granularity of memory system fault simulation without involving large run time overhead, we developed statistical models for most of new features in MEMRES. Compared with existing silicon based memories, emerging memories potentially suffer more severe reliability problems. MEMRES is capable of predicting the reliability of emerging memories. To demonstrate this

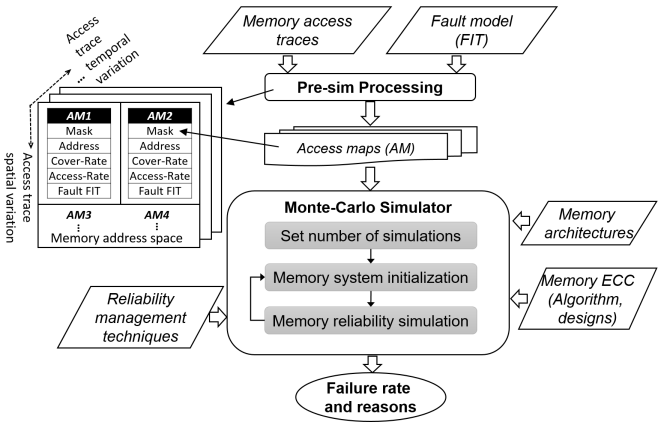



Figure 4.2: The framework overview of MEMRES.

capability, we use STT-RAM as a vehicle to explore optimized designs of different memory reliability enhancement techniques. STT-RAM faces the challenge of write errors and retention errors due to process variation [WLE16a, WLG16, WLP13, KZK15, WPC16]. Adding models of the new memory errors are convenient in MEMRES.

This following sections are organized as follows. Section 4.2 describes data structures, basic operations and models used in MEMRES. Section 4.3 validates MEMRES by the analytical model [JDB13], derived models for transient faults, and FaultSim [RN14, NRQ15]. Section 4.4 evaluates the reliability of STT-RAM designs with different retention time, write time, ECC designs, and algorithms, memory reliability managements and fault models. Section 4.5 concludes our work.

## 4.2 MEMRES Software Framework

Fig. 4.2 illustrates the overview of MEMRES tool flow. MEMRES comprises of two components: pre-sim processing and Monte-Carlo simulator.

### 4.2.1 Pre-sim Processing

The pre-sim processing is a one-time procedure for modeling memory fault and memory access behavior. Simply incorporating memory traces in simulation is impractical due to the fact that the memory traces occupy large memory space (one-second memory traces require several GB storage space) and dramatically affects simulation speed. In

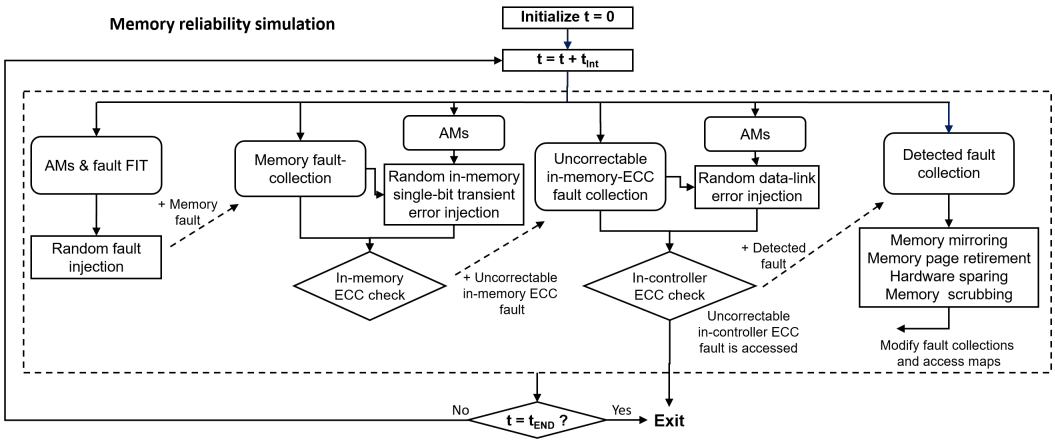



Figure 4.3: Memory reliability simulation. The simulation divides years-long memory lifetime into short intervals. In each interval, events of random fault injection, ECC checks, and memory reliability management are simulated. The simulation terminates when an uncorrectable fault occurs or it reaches the end of preset simulation time.

MEMRES, pre-sim processing extracts memory access density distribution from memory access traces and passes it to Monte-Carlo simulator in the form of Access-map (AM), which is a basic data structure in MEMRES. An AM models the access density on a specified memory address space (e.g., one column, one bank, one channel and etc.) in a specified period. As is shown in Fig. 4.2, memory access spacial variation in a time period (e.g., some space is more intensively accessed) can be captured by multiple AMs, and temporal variation (e.g. memory access density differs from time) can also be described by having more AMs. Hence, access behavior of one application can be captured by a set of AMs. In addition, a server running multiple applications can be described by passing multiple sets of AMs alternatively to simulator. The pre-sim processing also calculates fault failure-in-time (FIT, expected number of failures in one billion device-time) for each AM individually according to memory fault model (e.g., more frequently accessed AM has higher write error rate).

### 4.2.2 Monte-Carlo Simulator

The Monte-Carlo simulator performs a number of memory reliability simulations, and every simulation simulates a memory's lifetime (over years) reliability behavior. These simulations differ from each other due to random event modeling like fault injection, and

hence the memory failure rate and reasons can be statistically extracted from them. The configuration of memory architecture, reliability designs including ECC designs, ECC algorithms, and memory reliability management are inputs of MEMRES. The procedure of one memory's lifetime simulation is shown in Fig. 4.3. The simulation divides memory lifetime into short intervals and then simulates one interval after another. In each interval, four events are simulated including random fault injection, in-memory ECC check, in-controller ECC check, and memory reliability management. In the module of random fault injection, the probability of fault occurrence in one interval is first calculated according to fault FIT rate. Then faults are randomly injected based on the calculated probability, and injected faults are stored in a memory fault-collection. In-memory single-bit transient errors like retention errors and write errors are also injected, which may intersect with the faults in the memory fault-collection. The fault injection models are detailed in Section 4.2.5.1. In the module of in-memory ECC check, injected faults

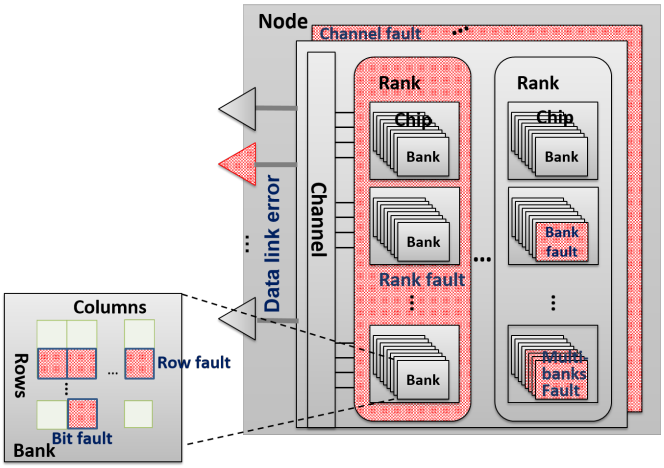


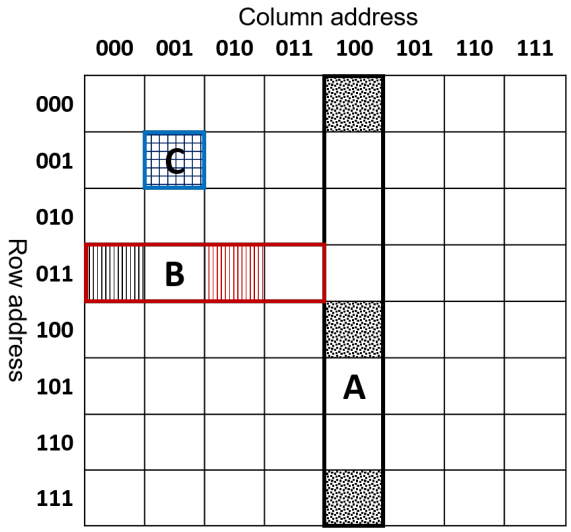

Figure 4.4: Memory architecture and memory fault types. A bank is constructed by columns and rows (several rows are grouped in a mat, which is not shown in the figure). Eight banks are built in a chip (device), and nine x8 chips (8 data chips + 1 ECC chip) or eighteen x4 chips (16 data chips + 2 ECC chips) construct a rank. Several ranks are built in a channel. In a write or read operation, a channel and a rank is firstly selected by a decoder, and then a word is written/read across all chips in the selected rank, e.g., every x8 chip in a rank outputs 8 bits to comprise a 72-bit word (64 data bits + 8 ECC bits), where all chips in a rank share same addresses. A fault type is defined by the component that is affected, e.g., a bank fault indicates multiple faulty rows/columns in the bank.

are checked whether they can be corrected by the in-memory ECC. The uncorrectable in-memory ECC faults are added to an uncorrectable in-memory ECC fault-collection. These faults may intersect with injected data-link errors to produce uncorrectable errors. The faults and data-link errors are checked together against in-controller ECC. Once an uncorrected in-controller ECC fault is accessed by an AM, a memory failure is produced and terminates current simulation. The detailed ECC model is discussed in Section. 4.2.5.2 The detectable faults are added to a detected fault-collection. In the module of memory reliability management, repairing techniques can be triggered by detected faults to physically repair memory devices or to systematically block accessing to the detected faults.

### 4.2.3 Fault-map and Access-map

In the field studies [SL12a, SDB15], detected faults are classified to several fault types, e.g., a row, a column, and a bank. A fault type defines a faulty region where multiple memory errors are produced. Examples of fault types are illustrated in Fig. 4.4. Larger fault types are likely caused by logic circuit failure. For example, a sense amplifier fault may lead to read errors in a column (column fault), or a particle hitting a bank decoder may cause errors in a bank (bank fault). Individually store the affected bits in the simulator data structure requires a huge memory space. Such large fault can be represented by a single data structure [RN14]. This structure is comprised of Mask and Address. Mask specifies the fault size, and Address locates the fault in memory space. This data structure is efficient in terms of computation speed and memory consumption. However, this structure assumed that all addresses in a fault are faulty, which is incorrectly. Most fault types have only $< 1\%$ faulty addresses in their covered memory space [SL12a, SDB15]. In order to accurately model the fault behavior, we add another statistical parameter, Cover-Rate (ranging from 0 to 1), to represent the percentage of faulty addresses in a fault. The Mask, Address, and Cover-Rate comprise a Fault-map (FM), which is the basic data structure in MEMRES to represent faults.

In order to catch application-dependent fault behavior and model system-level reliability enhancement techniques, MEMRES uses Access-map (AM) to model memory access behavior. It has five parameters including Mask, Address, and Cover-Rate, which




Fault/Access Range Representation

| | Mask | | Address | | Cover-Rate |
|---|---|---|---|---|---|
| | Row | Column | Row | Column | |
| A | 111 | 000 | 000 | 100 | 0.375 |
| B | 000 | 011 | 011 | 000 | 0.5 |
| C | 000 | 000 | 001 | 001 | 1 |

Figure 4.5: Examples of FM/AMs A, B, And C. A is a column FM/AM, B is a 4-bit FM/AM, and C is a single-bit FM/AM.

have the same definitions as in FM except that they model accessed memory space not faulty space. In addition, AM also contains another two parameters: Access-Rate representing access rate in an AM (number of read/writes in a hour) and FIT representing fault rate (expected number of faults in a billion device-hour). As is mentioned in Section 4.2.1, memory access behavior changes over time and memory space creating the need of a set of AMs for different simulation intervals and memory space. Finer division in time and space lead to higher modeling accuracy, but result in more AMs, calculations and simulation time. The trade-off between speed and accuracy is analyzed in Section 4.2.6.

Fig. 4.5 shows examples of FM/AMs A, B, and C. Address and Mask are sets of binary bits, which represent a region of device addresses (i.e., physical location in memory). In Fig. 4.5, FM/AM A only occupies the column 100, so its column address is 100. To tell MEMRES that A's three column address bits are valid (i.e., all these the three bits are required to determine A's location, so they are valid), A's column mask should be set to 000, where mask bit 0 means that the corresponding address bit is valid, and mask bit 1 means that the corresponding bit is invalid (masked). As A covers all rows from

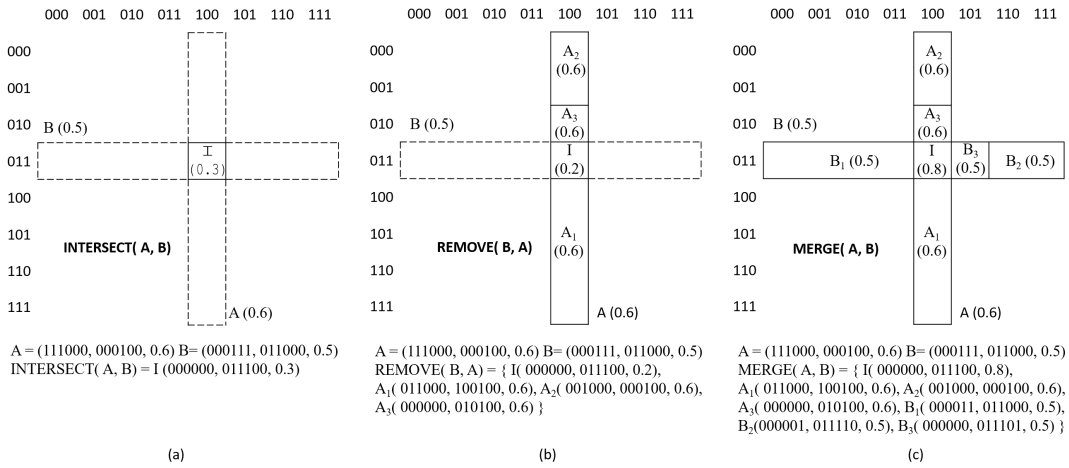



Figure 4.6: The basic operations used in MEMRES: (a) INTERSECT, (b) MERGE, and (c) REMOVE. Cover-Rates of A and B are 0.6 and 0.5 respectively.

row address 000 to 111, row address bits are not necessary to determine A's position, its row mask is 111 to set all corresponding row address bits invalid. For calculation simplicity, address bit is set to 0 when it is invalid. Similarly, B locates in the row 011, hence its row address is 011 and valid, and row mask is 000. As B covers columns 000, 001, 010, and 011, and only the first address bit determines B's position and is valid. Therefore, its column mask and address are 011 and 000 respectively, where the first address bit 0 indicates that B covers the left four columns. FM/AM C is a single bit, and all address bits are valid. The Cover-Rate is the percentage (or probability) of addresses being faulty in a fault. For instance, three out of eight addresses are faulty in A resulting in a Cover-Rate of 0.375.

The number of addresses covered by an FM must be a power of two. Most fault types can be represented by an FM, e.g., the number of addresses in all fault types shown in Fig. 4.4 are powers of 2. A fault with size $S$ not exactly equaling a power of two can be divided into no more than $log_2 S$ parts such that each part is represented by an FM. FM helps MEMRES to save memory space by orders of magnitude compared to traditional simulators.

### 4.2.4 Basic Operations

MEMRES's most operations are constructed by three basic bitwise operations: INTERSECT, MERGE, and REMOVE. They use FM/AM and have the closure property (i.e., both the operands and results of these operations are FM/AMs).

INTERSECT calculates the intersection between two FM/AMs. For examples, when a fault (represented by an FM) is accessed by an AM, the intersection between the FM and AM exists, and when two faults from different chips are accessed simultaneously in a word, their intersection exists. Basically, when two FM/AMs cover some same addresses, they intersect. The bit-wise formula $(A_{Mask} + B_{Mask}) + \overline{A_{Address} \oplus B_{Address}}$ tells whether two FM/AMs intersect; only if the formula outputs all "1"s [RN14], A and B intersect. The intersection I between A and B is obtained using Eqn. (4.1). Fig. 4.6 (a) shows an example of calculating INTERSECT(A, B).

$$I_{Mask} = A_{Mask} \& B_{Mask} \tag{4.1}$$
$$I_{Address} = A_{Address} + B_{Address}$$

REMOVE is used to clear faults (FMs) or to block access (AMs) to certain addresses. For instance, after a rank repairing (i.e. replacing a faulty rank with a spare one) the faults in the rank address are cleared from MEMRES's database. Another example is that when a page is retired, the page address is removed from AMs and will never be accessed. Fig. 4.6 (b) illustrates removing B from A. First, the intersection I between A and B is calculated, where the removing operation is actually only performed on I. Second, the Cover-Rate and Access-rate of I are updated using Eqn. (4.2), which represents the remaining intersection on A after removing B from it (if Access-rate and Cover-rate are zero or negative, I is totally removed from A). Third, Procedure 1 is used to obtain the least FM/AMs to cover the remaining parts of A excluding I. At last, store the remaining parts of A and the updated I together in a collection, which is the output of REMOVE(B,A).

$$I_{Access\text{-}Rate} = A_{Access\text{-}Rate} - B_{Access\text{-}Rate} \tag{4.2}$$
$$I_{Cover\text{-}Rate} = 1 - (1 - A_{Cover\text{-}Rate}) \,/\, (1 - B_{Cover\text{-}Rate})$$

**Procedure 1** Remove FM/AM B from FM/AM A

**Input:** FM/AM $A$ and FM/AM $B$.

**Output:** The collection R of the remaining FM/AM in $A$ after removing $B$

```
1: Set I = INTERSECT(A,B)
2: Set T = A
3: for i from the index of the MSB to the LSB of T_Mask do
4:   if T_Mask(i) == 1 && I_Mask(i) == 0 then
5:     //split T into two halves T0 and T1
6:     Set T_0, T_1 = T
7:     Set T_{0,Mask}(i) = 0 and T_{1,Mask}(i) = 0
8:     Set T_{0,Address}(i) = 0 and T_{1,Address}(i) = 1
9:     if T_0 intersects I then
10:      Add T_1 into the collection R. Set T = T_0
11:    else
12:      Add T_0 into the collection R. Set T = T_1
13:    end if
14:  end if
15:  if T == B then
16:    break loop
17:  end if
18: end for
19: return R
```

MERGE is used to combine two FM/AMs. For example, MEMRES merges and stores all accessed faulty addresses in its data-base so that if some addresses produce multiple errors, these addresses' Access-rates are over 1 after being merged in data-base, and permanent faults are determined according to it, then MEMRES can perform reliability management to clean the detected permanent faults. In Fig. 4.6 (c), a MERGE operation of A and B is illustrated. First, the intersection I between A and B is calculated, and then the Cover-Rate and Access-Rate of I are calculated following Eqn. (4.3). Second, constructing the remaining parts of A and B after excluding I using Procedure 1. Finally,

store I and remaining parts of A and B together in a collection.

$$I_{Access\text{-}Rate} = A_{Access\text{-}Rate} + B_{Access\text{-}Rate} \tag{4.3}$$
$$I_{Cover\text{-}Rate} = 1 - (1 - A_{Cover\text{-}Rate}) \cdot (1 - B_{Cover\text{-}Rate})$$

### 4.2.5 Modeling

The increasing large amount of memory faults hurt the memory reliability significantly [SL12a, SDB15, SPW09, MWK15]. However, compared with intensive data access and massive memory operations, the memory faults and failures are still counted as rare events. Therefore, using traditional simulators to analyze memory reliability involves too many redundant computations like emulating memory operation and simulating fault propagation. In order to improve computation efficiency, MEMRES uses statistical models. In this section, we describe the statistical models for fault injection, ECC detection and correction, and implementation of memory reliability management.

#### 4.2.5.1 Fault Injection

Memory faults are classified into two classes: transient and permanent faults. Intermediate faults can be modeled as permanent faults with a variable to model their existing time, which is not specially handled in MEMRES. Transient faults can be removed by written back after ECC correction, whereas permanent faults cannot be repaired. Field studies show that memory fault rate varies with the time [SL12a, SDB15, SPW09] and locality [MWK15]. However, traditional analytical models can only assume a constant fault rate, which results in inaccuracy. In MEMRES, the fault rate is held in AMs, which dynamically change with time and locality. In addition to the memory fault types studied in [SL12a, SDB15] (e.g., bit/row/column/bank faults and etc.), MEMRES also models single-bit transient errors (data-link errors and write/retention errors of emerging memories), which have high occurring rate and are easier to correct, but may dominate memory failure when they intersect with memory faults.

As an AM is valid through a simulation interval (i.e., AMs can be changed in different intervals), fault FIT rate (injection rate) is assumed to be constant in a simulation interval. Memory fault injection follows continuous Poisson distribution [FIT11], and

MEMRES describes the injection probability of $k$ faults of a fault type in an interval in Eqn. (4.4). It is noticed that the model is calculated once for every fault in an interval due to the different FIT rates of fault types. The injection model is more accurate than FaultSim [RN14], which assumes maximum one fault is injected in an interval.

$$P(k) = \frac{(\lambda \cdot t_{Int})^k \cdot \exp(-\lambda \cdot t_{Int})}{k!} \tag{4.4}$$

where $\lambda$ is the failure rate (FIT rate) of a fault in an AM, and $t_{Int}$ is the simulation interval in the unit of $10^9$ hours.

Data-link has much higher bit error rate (BER) [MKS10, NFL08] than memory faults due to inter-symbol interference, signal reflection, clock jitter, and voltage variation. Data-link error is a transient error and only affects single bit in a 72-bit word (64 data bits and 8 ECC bits), which is easily corrected by ECC. Similarly to data-link error, write/retention errors of emerging memories, e.g. write failure in STT-RAM [WLE16a, KZZ15], are also single-bit transient errors. As the probability of more than one single-bit transient errors simultaneously occurring in a word is extremely low, MEMRES safely assumes that there is at most one single-bit transient error in a ECC word (i.e., the ECC word length depends on algorithms and is commonly 72/144 bits). In a memory system with ECC and memory scrubbing (i.e., scrubbing periodically scans and corrects transient faults in memory), the single-bit transient errors cannot accumulate and may only cause memory system failure when its occurrence intersects with other memory faults in a word (i.e., occurs in the memory address as row/column/bank faults and etc.). Hence, to improve computation and memory consumption efficiency in a memory with ECC and scrubbing, MEMRES only injects single-bit transient errors with accessed memory faults. More specifically, the injection rate depends on the number of memory errors, which are produced when memory faults are accessed. As the data-link and write errors have even probability to occur at every access, these errors follow discrete Poisson distribution, where the probability of injecting $k$ such errors intersected with a memory fault is described in Eqn. (4.5). Differently, the retention errors have even temporal occurring probability, which follows continuous Poisson distribution and is handled in Eqn. (4.4). For a memory without ECC or memory scrubbing, these single-bit transient errors are injected same as single-bit transient faults, which are individually injected according to their FIT rates and can accumulate to create multiple-bit faults. However,

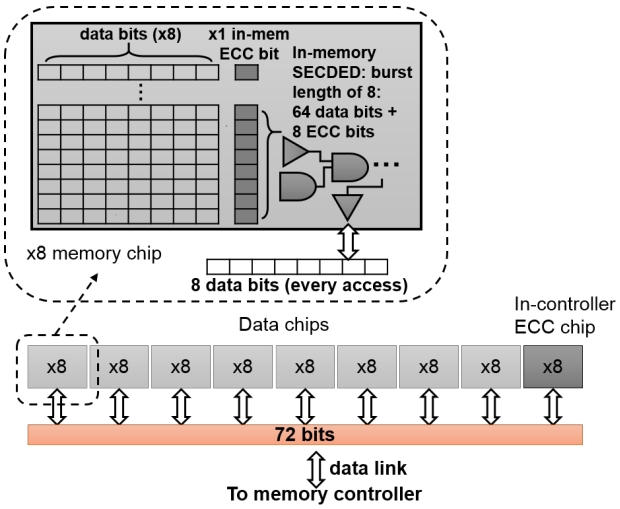



Figure 4.7: A memory with in-memory SECDED and in-controller ECC. In a memory chip, every eight columns of data bits are protected by one column of ECC bits. The in-memory ECC logic can correct a single-bit error in a 72-bit ECC word (64 data and 8 ECC bits), where a burst length of 8 accesses is required for a x8 chip to have 72 bits together for in-memory SECDED correction. A x8 chip inputs/outputs 8 data bits in an access, and totally eight x8 data chips and one x8 ECC chip input/output 72 bits from/to the memory controller in an access, where data errors in the 72 bits can be corrected/detected by the in-controller ECC.

such memory cannot functionally operate for a long period under high volume of single-bit transient faults.

$$P(k) = \frac{(BER \cdot N_A)^k \cdot \exp(-BER \cdot N_A)}{k!} \tag{4.5}$$

Where $BER$ is the bit error rate of single-bit transient errors, and $N_A$ is the expectation of number of accessed bits from memory faults in current simulation interval, which is calculated by Access-Rate and Cover-Rate of AMs and FMs.

#### 4.2.5.2 ECC Algorithms and Designs

Common ECC algorithms are classified into two types: Hamming-based and symbol-based codes. Examples of Hamming-based codes are single-error-correction-double-error-detection (SECDED) [BGM88a, MBR82] and single-chip-correction-double-chip-detection (SCCDCD, which can also be implemented by symbol-based codes) [Del97a, Del97b].

Both of them add 12.5% to redundancy of ECC bits. SECDED adds 8 bits (one x8 chip or two x4 chips) to a 64-bit word (eight x8 chips or 16 x4 chips), and SCCDCD adds 16 bits (four x4 chips) to 128 bits (32 x4 chips), which decodes/encodes two words interleaved across two channels simultaneously. Symbol-based codes are more sophisticated and efficient in redundancy like double-device data correction (DDDC) [JK13, Wil14]. With the same overhead of ECC bits as SCCDCD, DDDC has one sparing x4 chip for failed chip replacement in addition to the function of SCCDCD. An extension of DDDC is that the second failed chip replacement is allowed, after which the ECC algorithm changes from SCCDCD to SECDED. These codes that can correct any errors from single chip are also called Chipkill. In MEMRES, an ECC algorithm is configurable with four parameters: maximum detectable faulty bits, maximum correctable faulty bits, maximum detectable faulty symbols (bits from a chip are a symbol, e.g., a 4-bit symbol for a x4 chip), and maximum correctable faulty symbols. For SECDED, the maximum correctable and detectable faulty bits are one and two respectively. For the extended DDDC, the initial maximum correctable and detectable faulty symbols are one and two respectively, but after replacing two failed chips, these two parameters change to 0, while the maximum correctable and detectable faulty bits change to one and two respectively. The overhead of density, delay, and power is outside the scope of this paper.

MEMRES allows two ECC designs: in-controller ECC and in-memory ECC, which are shown in Fig. 4.7. In-controller ECC designs are commonly used in commercial server-class CPUs, where ECC detection/correction logic circuits locate inside a memory controller and can correct and detect errors from both memories and data links (e.g., double data rate (DDR) buses). An in-memory ECC design locates in a memory chip and corrects errors inside the chip. As an example, the memory in Fig. 4.7 are constructed by x8 chips, each chip inputs/outputs 8 bits in an access, and totally 72 bits from eight data chips and one ECC chip are read/written simultaneously, which comprise a word and are decoded/encoded by an in-controller ECC. In-memory ECC designs require burst mode. In burst mode, memory reads/writes multiple words with continuous physical addresses in one time. In Fig. 4.7. the in-memory SECDED works with burst mode with the burst length of 8. In every chip access, the ECC decodes/encodes 72 bits from 8 continuous chip-words together to perform SECDED. As an in-controller SCCDCD fails

when it faces multiple faulty symbols (chips), an in-memory ECC significantly decreases the probability of such case by correcting symbols inside chips.

The model for memory with only in-controller ECC design was derived in [JDB13]. However, memories with both in-memory and in-controller ECC designs have not been studied. We derived a set of statistical models used in MEMRES's ECC check, which allow for the interaction of in-memory and in-controller ECCs and give the probability that a memory fault or intersection (represented by an FM) fails both in-memory and in-controller ECCs in a simulation interval. Currently, in-memory ECC only considers SECDED, and in-controller ECC allows all Hamming-based and symbol-based algorithms.

As an example, we show a model for the combination of an in-memory SECDED and an in-controller symbol-based code (e.g., SCCDCD). This model describes the probability that after a memory fault FM is injected, the FM fails both in-memory and in-controller ECC in a simulation interval.

Eqn. (4.6) calculates the probability ($P_{correct\_symbol}$) that a symbol (e.g., 4 bits for a x4 chip) covered by an FM is ECC correctable. It includes two cases: 1) there is no faulty bit in the symbol ($P_{0\_faultybit@symbol}$), 2) there is one faulty bit in the symbol, and it is the only one in its burst group of 72 bits ($BL \cdot N_{PFB}$), which is correctable to SECDED ($P_{1\_faultybit@symbol}$).

$$
\begin{aligned}
P_{0\_faultybit@symbol} &= (1 - P_{FB})^{N_{PFB}} \\
P_{1\_faultybit@symbol} &= C_1^{N_{PFB}} \cdot P_{FB} \cdot (1 - P_{FB})^{BL \cdot N_{PFB} - 1} \\
P_{correct\_symbol} &= P_{0\_faultybit@symbol} + P_{1\_faultybit@symbol}
\end{aligned}
\tag{4.6}
$$

Here $BL$ is the burst length (e.g., burst length of 8 is shown in Fig. 4.7). $N_{PFB}$ is the number of possible faulty bits in the symbol, which is determined by the Mask of the FM covering the symbol. $C_1^{N_{PFB}}$ is choosing one faulty bit from $N_{PFB}$ possible faulty bits in the symbol. $P_{FB}$ is the probability that a bit covered by the FM is faulty, which is calculated from Cover-Rate of the FM.

Then we calculate the probability ($P_{correct\_word}$) that a word constructed by possible faulty symbols is ECC correctable in Eqn. (4.7) based on $P_{correct\_symbol}$ and number of

possible faulty symbols $N_{PFS}$ (calculated from Mask) in a word. The appearance of uncorrectable word will cause both memory failure.

$$P_{correct\_word} = \sum_{k=0}^{\min(N_{MFS},N_{PFS})} C_k^{N_{PFS}} \cdot P_{correct\_symbol}{}^{N_{PFS}-k} \cdot \left(1 - P_{correct\_symbol}\right)^k \tag{4.7}$$

Where the $N_{MFS}$ is the maximum correctable symbols of the in-controller ECC, and $k$ is the number of faulty symbols.

The model above excludes the data-link error and write/retention error, which may intersect with memory errors to create uncorrectable errors. The additional occurrence of a write/retention error inside a memory chip may cause failure in the in-memory SECDED which is otherwise able to correct the memory chip, and then the appearance of an additional faulty symbol may further cause failure in the in-controller ECC. As explained in Section4.2.5.1, there is maximum one such error in a symbol, and hence we derive the probability ($P_{wrErr\_fail\_memECC}$) that the occurrence of a write/retention error in an accessed symbol causes an in-memory SECDED failure in Eqn. (4.8). The derivation includes two cases: 1) a memory faulty bit already exists in the accessed symbol before the occurrence of a write/retention error in the symbol, and the write/retention error should not overlap with the faulty bit. 2) a memory faulty bit is not in the accessed symbol but in the same burst group (e.g., the 63 bits excluding the accessed accessed symbol in a burst group, see Fig. 4.7) before the occurrence of a write/retention error in the accessed symbol.

$$P_{wrErr\_fail\_memECC} = (SL-1)/SL \cdot P_{1\_faultybit@symbol} + (BL-1) \cdot N_{PFB} \cdot P_{FB} \cdot (1-P_{FB})^{(BL-1)\cdot N_{PFB}-1} \cdot P_{0\_faultybit@symbol} \tag{4.8}$$

Where $SL$ is the symbol length (number of bits per symbol).

Then the in-controller ECC may fail due to the additional faulty symbol caused by the write/retention error in the case that existing faulty symbols already reach the maximum correction capability $N_{MFS}$ before the write/retention error. This probability $P_{wrErr\_fail\_ctrECC}$ is derived as follows.

$$P_{wrErr\_fail\_ctrECC} = C_{N_{MFS}}^{N_{PFS}} \cdot (1 - P_{correct\_symbol})^{N_{MFS}} \cdot C_1^{N_{PFS}-N_{MFS}} \cdot P_{wrErr\_fail\_memECC} \cdot {P_{correct\_symbol}}^{N_{PFS}-N_{MFS}-1} \tag{4.9}$$

Unlike write/retention errors, which occur and can be corrected inside memory chips, data-link errors occur on data links and are not checked by in-memory ECC. More specifically, when the existing faulty symbols reach the correction capability $N_{MFS}$, a data-link error occurring on any correct symbols will create another faulty symbol to fail in-controller ECC. The probability is derived in Eqn. (4.10).

$$P_{dlErr\_fail\_ctrECC} = (N_S - N_{MFS}) / N_S \cdot C_{N_{MFS}}^{N_{PFS}} \cdot {P_{correct\_symbol}}^{N_{PFS}-N_{MFS}} \cdot (1 - P_{correct\_symbol})^{N_{MFS}} \tag{4.10}$$

Here $N_S$ is the number of symbols in an ECC word (e.g., 32 for SCCDCD with x4 chips).

So far, we have derived the statistical models for probabilities of uncorrectable words caused by a memory-fault, by the interaction of memory-fault and write/retention-error, and by the interaction of memory-fault and data-link-error. These models also work for error detection. For memory faults with faulty bits exceeding the specified ECC detectable capability, ECC can still detect a part of them. For example, some errors with three faulty bits can be detected by SECDED which guarantees to detect faults with two or less faulty bits. A partial detection rate is taken as a constant input to model the detection probability of those faults.

#### 4.2.5.3 Memory Reliability Management

Although ECC designs can correct memory errors, permanent faults can accumulate with time and intersect to produce uncorrectable multiple-bit and multiple-symbol errors. To avoid fault accumulation, modern systems use memory reliability management to deactivate existing faults. These techniques require information of fault location, which can be identified from detected errors. MEMRES has modeled the identification process. One

collection of FMs called fault-collection stores all faulty addresses, which MERGE all existing faults. Errors are produced when fault-collection is accessed by AMs (i.e., AMs intersect with FMs in fault-collection), and then detected by ECC designs. MEMRES MERGE detected errors (also represented by FMs) into a collection called error-log. A high access-rate of an FM in error-log means that the addresses covered by the FM frequently produce errors and are identified as a permanent fault. This can trigger memory reliability management to deactivate it. Modeling of memory reliability management is detailed below:

- Memory scrubbing corrects detected correctable transient faults. Two scrubbing techniques are simulated in MEMRES, 1) if an ECC correctable transient fault (in the form of FM) is accessed by an AM, the accessed faulty addresses are removed from the transient fault using REMOVE; 2) in addition to the on-line correction, the memory scrubbing periodical inspects the memory, and all correctable transient faults are cleared from fault-collection using REMOVE, where the periodically scrubbing cycle is configurable.

- Hardware sparing allows the replacement of failed hardware with sparing hardware. For example, modern memory systems have enabled rank sparing, which uses a spare rank for replacing an in-use rank with detected permanent faults. In MEMRES, the spare devices, number of spare devices, the hardware being protected by this technique, and triggering threshold can be specified in configuration file. When detected permanent faults reach the configurable threshold in a protected hardware, REMOVE is applied to clear faults (in the form of FMs) in the replaced hardware from MEMRES's database (fault-collection and error-log).

- Memory page retirement blocks access to memory pages with permanent faults to avoid fault activation. More specifically, when a permanent fault is detected in a memory physical page, the address mapped to the page is blocked, and the data on the page is moved to other pages. In MEMRES, once a permanent fault (in the form of FM) is detected by error-log, the access (in the form of AMs) to the pages intersecting with the fault are moved to other pages using REMOVE and MERGE. The maximum number of retired pages is configurable.

- Memory mirroring mounts one memory space as a copy of another one, and system reads/writes data from/to both spaces simultaneously so that if data from one space contains errors, system can still obtain correct data from the other memory space. This protection is frequently used for critical data. MEMRES models this technique by having a special ECC check: uncorrectable faults in the mirrored spaces cannot directly cause failure unless two uncorrectable faults from two mirrored space intersect.

### 4.2.6 Trade-off between Accuracy and Speed

In this section, we analyze the accuracy-speed trade-off in fault simulation. In MEMRES, dense grids of memory space (smaller AMs) more accurately models access behavior by holding more precise Cover-Rates, whereas sparse grids (larger AMs) save computation but lead to imprecise Cover-Rates. For example, one bank is intensively accessed but other banks in the same rank are barely accessed, if a large AM is used to represent access to the rank, its Cover-Rate is low after averaging memory access over all banks. Then inaccurate simulation is resulted when a fault occurs in the intensively accessed bank, because the fault is supposed to be quickly activated (accessed) but the low Cover-Rate delays its activation time. Fault activation is a random event and determined by Cover-Rate, hence distribution of the activation time, which is in fact simulated in MEMRES, is crucial to accuracy. The mathematical expectation and variance of fault activation time are listed below in Eqn. (4.11)

$$\begin{aligned} \mu_{ActTime} &= t_{Int} \cdot \frac{(1 - Cover - rate)^{MapSize}}{1 - (1 - Cover - rate)^{MapSize}} \\ \sigma_{ActTime} &= t_{Int} \cdot \frac{(1 - Cover - rate)^{MapSize}}{\left(1 - (1 - Cover - rate)^{MapSize}\right)^2} \end{aligned} \tag{4.11}$$

Here $t_{Int}$ is the simulation interval, *Cover-Rate* is the Cover-Rate product of a fault FM and the AM accessing the FM, and $MapSize$ is the number of addresses in the intersection of the FM and AM. The accuracy is determined by the precision of *Cover-Rate* and $MapSize$. For a high *Cover-Rate*, the sensitivity of accuracy to $MapSize$ is smaller, and vice versa. Therefore if fault injection is dominated by large faults or memory is intensively and uniformly accessed, larger AMs can be used to speed up the simulation without losing accuracy.

Similarly, to prolong simulation intervals can lead to less computation. Nevertheless, longer interval may result in simulation error if too many faults are injected in a long interval. For example, in a system with memory reliability management, a detected permanent fault should be deactivated by repairing techniques before intersecting with other faults occurring later. However, a long simulation interval increases the probability of multiple fault injection, which can intersect to produce non-correctable errors. Therefore, it had better avoid multiple fault injection, and the interval of hours is acceptable according to fault rates reported in [SL12a, SDB15].

The run time and memory consumption complexity of MEMRES are $O\left(\lambda^2 \cdot \log\left(M\right)^2\right)$ and $O\left(\lambda \cdot \log\left(M\right)\right)$ respectively, where $M$ is the size of memory system and $\lambda$ is the failure rate. The size and number of FM/AMs scale with the address length ($\log(M)$) and injection rate ($\lambda$) respectively. Run time of INTERSECT scales with the size of FM/AMs($\log(M)$). The run time of MERGE and REMOVE (described in Procedure 1) is determined by the size of FM/AMs ($\log(M)$) and the run time of INTERSECT ($\log(M)$), which is $\log\left(M\right)^2$. The number of operations is proportion to the square of the number of FM/AMs (proportional to $\lambda^2$). The complexity of total run time, which is the product of the number of operations and the run time of operations, is $O\left(\lambda^2 \cdot \log\left(M\right)^2\right)$. The memory consumption, which is decided by the number and size of FM/AMs, is $O\left(\lambda \cdot \log\left(M\right)\right)$. In experiments, the wall time for the 100,000 5-years simulations (in-controller ECC and memory scrubbing are enabled) for Fig. 4.10 and Fig. 4.11 on Quad-Core AMD Opteron(tm) Processor 2376 with 8 threads is 70 minutes. Its peak memory consumption is 1 GB.

## 4.3 Framework Validation

To substantially validate MEMRES is non-trivial. Large scale experimentation is too expensive and impractical for most research groups given the need to observe failures over an extended period of time on a large scale datacenter. Existing simulators take unacceptable time to complete the validation task, and existing analytical models do not support memory reliability management. In this section, we validate MEMRES's accuracy with FaultSim, the analytical model [JDB13] and derived analytical models.

Table 4.2: Architecture of a 4-GB DRAM DIMM.

| Ranks | Chips | banks | Mats | Rows | Columns | Access-Rate |
|---|---|---|---|---|---|---|
| 2 | 16+2 | 8 | 64 | 512 | 4096 | 1e12/hour |

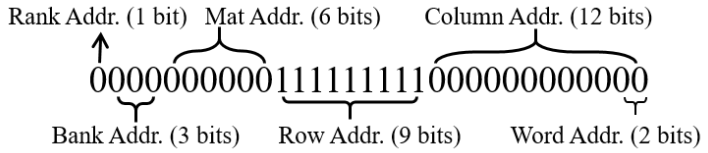



Figure 4.8: An example Mask of a column FM in the memory specified by Table 4.2. A read contains 4 bits from a chip, 18 reads construct a 72-bit word (64 data bits and 8 ECC bits), and a word has a memory physical address.

Because models and FaultSim do not support fault rate distribution and variation over time and address space, memory access behavior, in-memory ECC, and memory reliability management, we disable these features in MEMRES in this section. However, enabling these features strongly affects simulation results, which are illustrated in Section 4.4. In the validation, we assume that fault rate is constant, a large fault affects all covered address space, and a fault is accessed and corrected or causes ECC failure immediately after the fault injection.

We use an 4-GB DRAM based main memory as an example to validate MEMRES. The configuration of the 4-GB DRAM is listed in Table 4.2. In this memory architecture, an example column FM's mask is shown in Fig. 4.8. We use the fault rates reported in [SL12a, SDB15, MKS10, NFL08] for the validation, which is shown in Table 4.3. Because analytical models [JDB13] and FaultSim [RN14, NRQ15] assume all bits in a fault coverage are faulty, we use Cover-Rates of 1 for MEMRES to match their assumption in the validation for all faults (i.e., the Cover-rate listed in the table are used for later case study in Section 4.4).

The predicted failure rates of a 4-GB DRAM by MEMRES, FaultSim (interval and event modes), and the analytical model [JDB13] are drawn in Fig. 4.9 as a function of time. Normal fault rate (Table 4.3) and 4x fault rate (4xFIT) are used. Using the analytical model [JDB13] as the baseline, the maximum mismatch for MEMRES, Fault-

Table 4.3: Fault FIT rates per chip and data-link BER for DRAM and STT-RAM. The retention BER (RER) and write BER (WER) are only for STT-RAM. Data-link error, retention error, and write error are single-bit transient errors (SBT).

| Fault types | Transient FIT | Permanent FIT | Cover-Rate |
|---|---|---|---|
| Single-bit | 0 | 18.6 | 1 |
| Single-word | 1.4 | 0.3 | 1 |
| Single-column | 1.4 | 5.6 | 0.02 |
| Single-row | 0.2 | 8.2 | 0.002 |
| Single-bank | 0.8 | 10 | 0.002 |
| multi-banks | 0.3 | 1.4 | 0.002 |
| single-lane | 0.9 | 2.8 | 0.002 |
| Data-link BER | $10^{-14}$ [NFL08] | | |
| Retention BER/hour (STT-RAM) | 0, $10^{-5}$, $10^{-10}$, $10^{-15}$, (design dependence) | | |
| Write BER (STT-RAM) | 0, $10^{-8}$, $10^{-11}$, $10^{-14}$ (design dependence) | | |

Sim's interval and event modes are 1%, 2%, and 2% respectively. Please note that the analytical model is not the ground truth. The analytical model should overestimate the memory failure rate, because it separately calculates the probability of all critical faults that can cause memory failure and sums them together, but ignores to subtract the case that more than one critical faults exist in the same memory, which is the second order probability. Since MEMRES predicts lower failure probability than the analytical model indicating that MEMRES' error is lower than 1%.

In the following, we validate a DRAM with high frequency transient faults, e.g., DDR bus errors, which are not supported in FaultSim, and the analytical model [JDB13]. These errors strongly affect memory reliability through intersecting with permanent memory faults to cause ECC failure, e.g., ChipKill. To fill the gap, we derive an analytical model for memory failure caused by the intersection between a memory permanent fault and a transient fault (applicable to all transient faults and single-bit transient errors) as a supplemental model to validate MEMRES. Intersection between two transient fault/errors is unlikely given that a transient fault only exists short time in memory with scrubbing, hence we ignore this case.

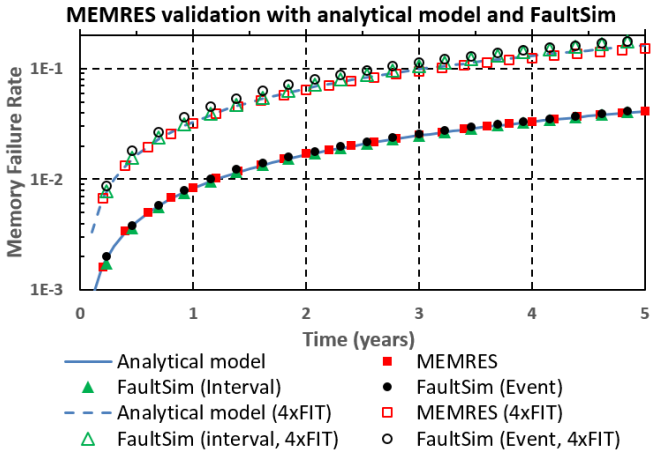



Figure 4.9: Validation of MEMRES with FaultSim and the analytical model [JDB13]. The failure rates for a 4-GB DRAM with SECDED as functions of time are shown. 1x and 4x fault rates are used. MEMRES matches well with the analytical model and FaultSim.

$P_{fail}$ in Eqn. (4.12) is the probability that memory failure is caused by an intersection of one memory fault and one transient fault as a function of time. The model describes that firstly a permanent fault occurs ($P_{perm}(t_1)$) at time $t_1$, then a transient fault occurs ($P_{perm}(t_2)$) within $t - t_1$ after $t_1$, and they intersect with each other to produce uncorrectable errors ($P_{intersect}$), e.g., the two faults occur in different chips but cover overlap addresses to produce multiple-symbol errors to cause Chipkill failure. $P_{intersect}$ depends on fault size and ECC algorithms.

$$P_{fail}(t) = P_{intersect} \cdot \int_0^t P_{perm}(t_1) \cdot \int_0^{t-t_1} P_{tran}(t_2) \tag{4.12}$$

$$\begin{aligned} P_{perm}(t_1) &= \exp(-\lambda_p t_1) \cdot (1 - \exp(-\lambda_p dt_1)) \\ &= \lambda_p \cdot \exp(-\lambda_p t_1)\, dt_1 \\ P_{tran}(t_2) &= \lambda_t \cdot \exp(-\lambda_t t_2)\, dt_2 \end{aligned}$$

Where $\lambda_p$ and $\lambda_t$ are the failure rate of the permanent and transient fault respectively. The analytical form of $P_{fail}$ is shown in Eqn. (4.13).

$$P_{fail}(t) = P_{intersect} \cdot \begin{pmatrix} 1 - \exp(-\lambda_p t) + \lambda_p / (\lambda_t - \lambda_p) \\ \cdot (\exp(-\lambda_t t) - \exp(-\lambda_p t)) \end{pmatrix} \tag{4.13}$$

For the cumulative distribution function (CDF) of memory failure rate shown in Fig. 4.10, MEMRES matches with the analytical model for a 4-GB DRAM with in-controller

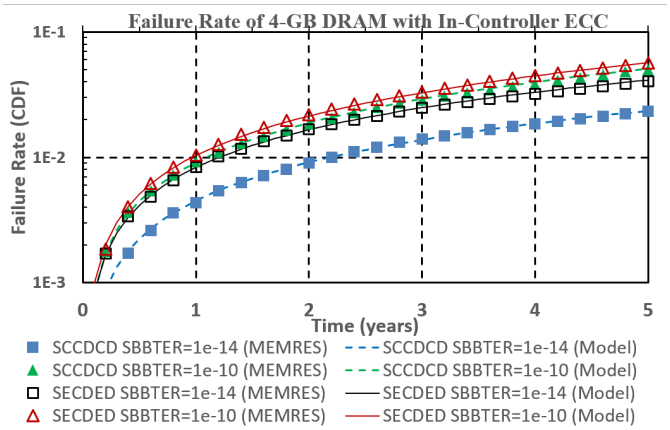



Figure 4.10: The failure rate for a 4-GB DRAM with SECDED or SCCDCD as a function of time. The single-bit transient error rate (SBTER, i.e., DRAM only has data-link error as SBT in the validation) of $10^{-14}$ and $10^{-10}$ are used in this validation.

SECDED and SCCDCD. SCCDCD overall performs better than SECDED because the SCCDCD can correct multiple-bit errors from any single chip, indicating that SCCDCD can correct all individual faults from Table 4.3. However, intersections of multiple faults from different chips and intersection between memory faults and single-bit transient errors (SBT, including data-link error, retention error of STT-RAM and write error of STT-RAM.) can give rise to SCCDCD failure. The failure rate increases dramatically with data-link BER, and when the BER is $10^{-10}$, SCCDCD does not show obvious benefit compared with SECDED .

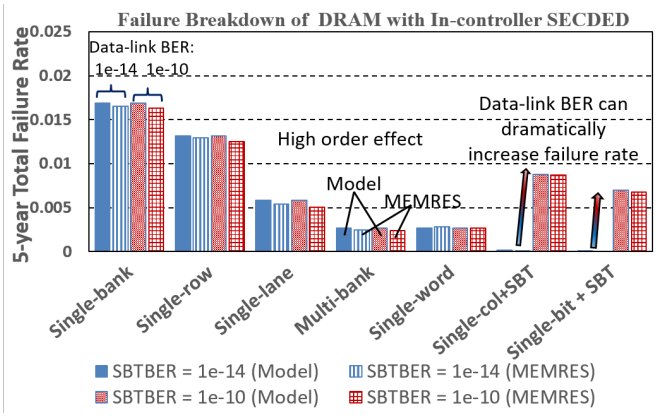



Figure 4.11: The memory failure rate breakdown for a 4-GB DRAM operating for 5 years with in-controller SECDED. The data-link BER of $10^{-14}$ and $10^{-10}$ are used in this validation.

Fig. 4.11 shows the breakdown of failure caused by different fault types. Again, MEMRES matches well with analytical models except for a small difference. MEMRES more accurately shows that all single-fault induced failure rates decrease with increased data-link BER (e.g., single-bank, single-row, single-lane, multi-bank, single-word), while the analytical model gives exact same failure rates at different BER. In reality, as more memory failures are caused by the increased data-link errors interacting with permanent memory faults, failures caused by other memory faults decrease due to the fact that when memory system failure occurs, the system is shut down, and the failed memory is replaced with a new memory without any memory faults. Analytical models calculate the failure probability due to each fault individually because it is very difficult to include the high order effect of failure interactions. Note that, in default MEMRES setup (used in the paper), a memory failure ends simulations, however, MEMRES also models the failed memory replacement similarly to rank replacement (Section 4.2.5.3) where MEMRES allows to continues simulation when a failure occurs and is fixed by hardware replacement.

## 4.4 A Study of STT-RAM using MEMRES

In this section, we perform a case study of analyzing STT-RAM's reliability using MEMRES. Since STT-RAM and traditional memories have similar peripheral circuits (sense amplifier, decoder, etc.), and large memory faults like row, column, bank faults and etc are highly possible caused by peripheral circuit failure, we assume that STT-RAM suffers from the same fault FIT rates as DRAMs (see Table 4.3) except for the single-bit transient fault (not single-bit transient errors) given that STT-RAM is immune to particle-induced faults. In addition to these faults, STT-RAM may suffer from single-bit transient errors including data-link errors, retention errors, and write errors. Their error rates depend on the STT-RAM design, where the error rates can be traded for low energy consumption [WLE16a].

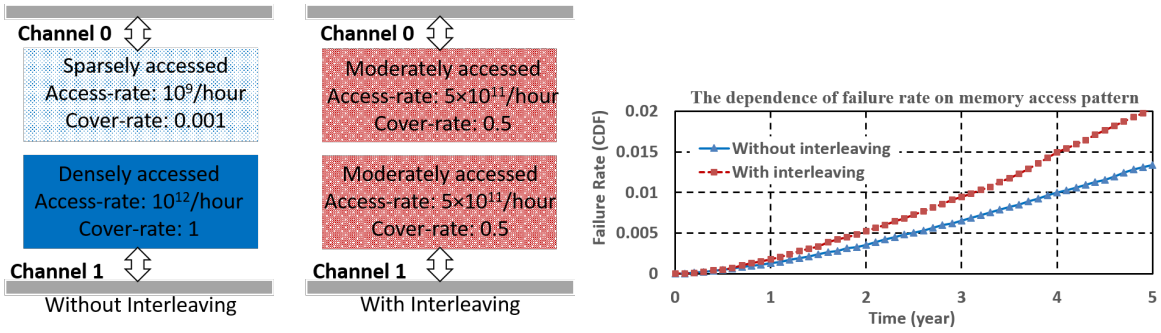
There are trade-offs between performance (write speed and energy) and reliability (retention error and write error) for STT-RAM. The critical current (i.e., proportional to write current) of STT-RAM is approximately proportional to thermal stability [WLE16a], while the thermal stability also determines the retention time, the average time that an

Table 4.4: The 5-year failure rate for STT-RAMs with different write error rate (WER, per-bit-write failure probability), retention error rate (RER, per-bit-hour failure probability), and ECC designs: 1) in-controller SCCDCD with (W/) in-memory SECDED, 2) in-controller SCCDCD without (W/O) in-memory SECDED.

| RER | $10^{-5}$ | | $10^{-10}$ | | $10^{-15}$ | |
|---|---|---|---|---|---|---|
| WER | W/O | W/ | W/O | W/ | W/O | W/ |
| $10^{-8}$ | 0.1365 | 0.0219 | 0.1359 | 0.0219 | 0.1354 | 0.0209 |
| $10^{-11}$ | 0.1336 | 0.0213 | 0.0649 | 0.0213 | 0.0638 | 0.0208 |
| $10^{-14}$ | 0.1317 | 0.0213 | 0.0402 | 0.0212 | 0.0310 | 0.0207 |

STT-RAM cell holds data before false switching during standby state. This indicates that shorter retention time can reduce write energy as well as improve write speed at the risk of retention error [SBL11]. Another trade-off is to reduce write time at the expense of increased write error rate (WER) [WZS09]. With memory reliability enhancement techniques, the reliability requirement of STT-RAM can be relaxed, which leads to faster speed and less power consumption simultaneously. As a case study, we use MEMRES to explore the impact of retention error rate (RER) and WER on STT-RAM with different reliability enhancement techniques. An 8-GB STT-RAM is used for all experiments in this section, which has two channels, and each channel has one dual in-line memory module (DIMM) with configuration in Table 4.2. Fault FIT rates per chip are listed in Table. 4.3.

Table. 4.4 illustrates the 5-year memory failure rate of STT-RAMs with or without in-memory ECC for different write/retention BER. The failure rate of the STT-RAM with both in-memory SECDED and in-controller SCCDCD is significantly lower than the STT-RAM with only in-controller SCCDCD, because in-memory SECDED corrects a single-bit transient error soon after its occurrence, allows chips to output corrected symbols, and hence prevents multi-chip errors to cause in-controller SCCDCD failure. The probability of having multiple single-bit transient errors in a chip read (consequent 4 bits for a x4 chip) is extremely low, which may not happen even once in a data-center for many years given that memory scrubbing is enabled to correct transient errors




(a) (b)

Figure 4.12: (a) An 8-GB STT-RAM with unbalanced memory access (without interleaving) and balanced memory access (with interleaving). (b) Failure rates (CDF) of STT-RAM with unbalanced and balanced memory access. In-controller SCCDCD and in-memory SECDED are enabled.

periodically. Therefore for STT-RAM with in-memory ECC, retention and write errors can be traded for speed and power improvement, but seeking the optimized trade-offs requires the knowledge of in-memory SECDED performance overhead.

As is known, memory interleaving can improve memory throughput by spreading memory access evenly across memory banks and channels. In Fig. 4.12a, one STT-RAM with interleaving has balanced access, while the other one without interleaving has unbalanced access. In this study, all chips have the same memory fault FIT rates (see Table 4.3). More specifically for this experiment, we assume fault occurrence probability (except single-bit transient errors) does not depend on memory access density for the reason that faults are random rare events and caused by process variation or particle induced errors which are not related to access density and hardware wear-out. The field study [SPW09] shows a sub-linear dependence of uncorrected error rate on time indicating that fault FIT rate does not increase with time and hence is not clearly related to memory access and wear-out (i.e., errors increase due to accumulated permanent faults, but fault occurring rate does not increase with time). The Fig. 4.12b illustrates that the STT-RAM without interleaving suffers from lower failure rate given that permanent memory faults in the channel being sparsely accessed is less likely to intersect with data-link errors.

When rank sparing is enabled, detected permanent faulty addresses over a threshold can trigger rank sparing, which replaces the faulty rank with a spare rank. In Fig. 4.13, we

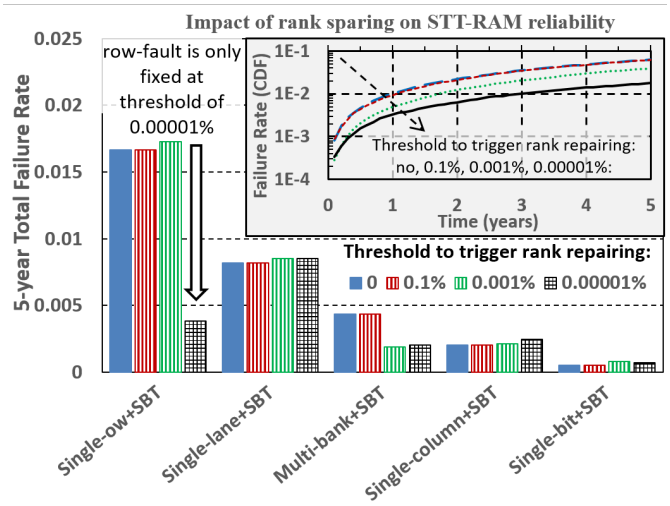



Figure 4.13: The 5-year failure breakdown and failure rate (CDF) of STT-RAM with enabled rank sparing. The thresholds (percentage of faulty addresses in a rank) to trigger rank repairing are 0.1%, 0.001%, and 0.00001%. The STT-RAM has one spare rank in each channel. In-controller SCCDCD is enabled.

simulates an 8-GB STT-RAM with rank sparing and different threshold to trigger rank sparing. As is illustrated, failure rate is not reduced when high threshold is set (e.g., 0.1%), because no individual fault is big enough to trigger rank sparing. With threshold decreasing, overall failure rate decreases given that multi-bank faults and single-row faults can trigger rank sparing at the threshold of 0.001% and 0.00001% respectively. Single-lane fault is not correctable as it affects both ranks in a channel, while a channel only has one spare rank in the experiment setup.

A detected permanent fault can trigger memory page retirement, which removes faulty physical pages from use by the operating system. In Fig. 4.14, we simulate an 8-GB STT-RAM with memory page retirement. As can be seen, more allowed retired pages give rise to less failure rate but more memory space loss. A row-fault is easy to be deactivated by retiring two pages, while a multi-bank fault affects more than 5000 pages and can only be deactivated when a large memory space loss is allowed. However, if rank sparing and memory page retirement are both enabled, they can collaborate to efficiently correct most faults.

System with memory mirroring reads/writes two mirrored space simultaneously, and errors in one space can be corrected by its mirror. In Fig. 4.15, we simulate an 8-GB

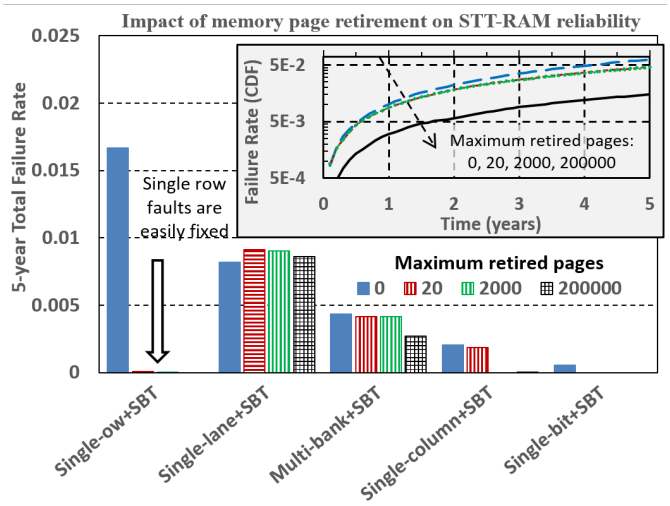



Figure 4.14: The 5-year failure breakdown and failure rate (CDF) of STT-RAM with enabled memory page retirement. Different maximum allowed retired pages per channel are tried, including 20, 2000, and 200000. In-controller SCCDCD is enabled. Memory page size is 4kB.

STT-RAM with memory mirroring, which mounts a memory space as a copy of another memory space. We tried different mirrored memory space including a whole memory space (i.e., a half space is mirrored by the other half), a half memory space, and a quarter

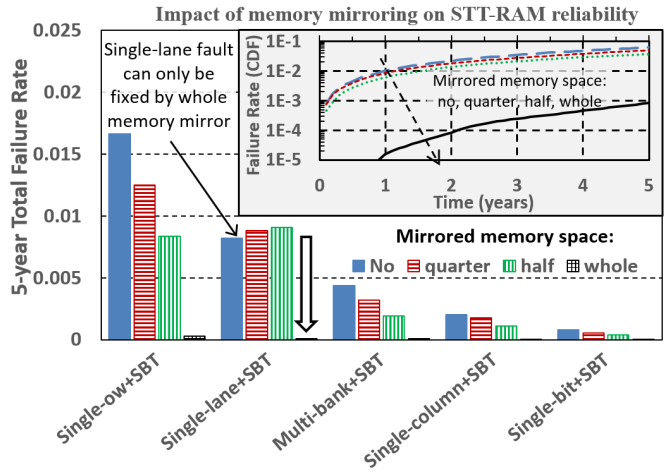
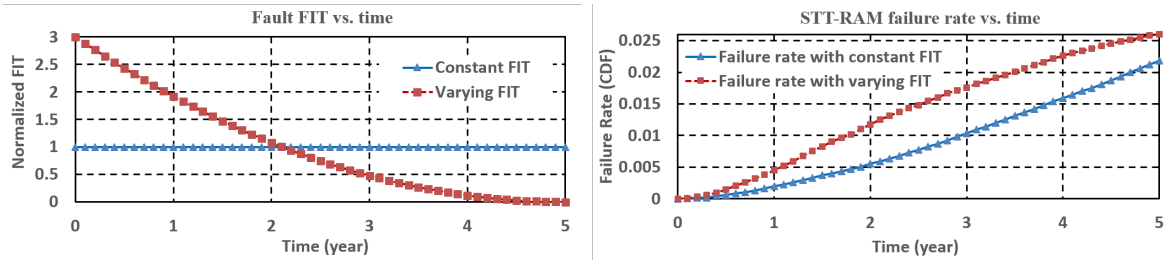



Figure 4.15: The 5-year failure breakdown and failure rate (CDF) of STT-RAM with memory mirroring. Different mirrored memory space are simulated including whole memory mirroring (one channel is mirrored to the other one), a half memory mirroring (one rank is mirrored to another one), and a quarter of memory mirroring (a half rank is mirrored to another half). In-controller SCCDCD is enabled.




(a) (b)

Figure 4.16: (a) Varying fault FIT rate (normalized to constant fault FIT rate) and constant FIT rate vs. time. (b)The failure rate (CDF) of STT-RAM with varying fault FIT rate and constant fault FIT rate (listed in Table. 4.3). In-controller SCCDCD and in-memory SECDED are enabled.

of memory space. As mirrored space increases, the failure rates caused by all fault types decrease as is expected. Larger faults like lane faults require more mirrored space to correct. Memory mirroring significantly increases the system's robustness to memory faults, where failure rate decreases to nearly zero when a whole memory is mirrored, as a trade-off, a half memory space is lost.

Fault FIT rate varies over time and usually decreases with time [SL12a, SDB15], while analytical models always assume constant FIT rate because a varying FIT rate is very difficult to model. However, constant FIT rate assumption may result in inaccuracy. As an example, we use MEMRES to simulate STT-RAMs with constant and varying FIT rats. In Fig. 4.16a, the varying FIT rate is a quadratic function of time, which is normalized to the constant FIT rate such that it gives rise to the same 5-year-fault-occurring probability as the constant FIT rate. Though the varying FIT rate decreases with time, permanent faults continue to produce errors after injection; hence error rates increase still with time, which is not contrary to field studies [SL12a, SDB15, MWK15, SPW09]. Fig. 4.16b shows that the varying FIT rate results in higher STT-RAM failure rate, because its higher FIT rate in the beginning causes faults to occur earlier, which have higher probability to intersect with data-link errors and fail SCCDCD ECC. The failure rate difference demonstrates MEMRES's capability of simulating more realistic situations than analytical models.

## 4.5 Chapter conclusion

The proposed MEMRES framework facilitates a fast and convenient way to assess the reliability of modern memory systems. It can perform memory fault simulation with ECC and memory reliability management. The accuracy of MEMRES is validated by the comparison with the derived analytic model and existing models. Through MEMRES, modern reliability enhancement techniques including ECC designs and memory reliability management can be calibrated to have optimized efficiency for target applications. With additional fault models, we show examples of using MEMRES to optimize the reliability for emerging memory systems.

# CHAPTER 5

# Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory

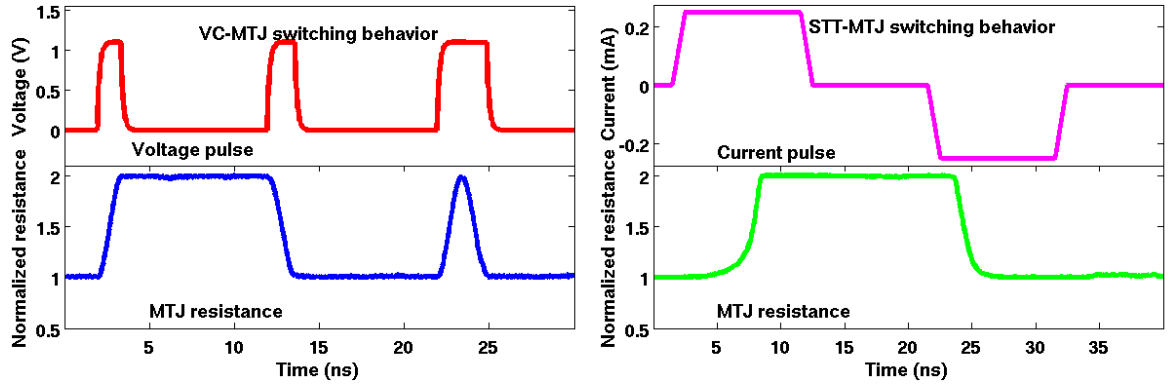
## 5.1 Chapter Introduction

MeRAM requires a comprehensive evaluation, while STT-RAM, which is better known and has similar structure and fabrication process, is an appropriate reference. To accurately compare the reliability of the two technologies, the WER must be precisely captured. The state-of-art method is the LandauLifshitzGilbert (LLG) differential equation based Monte-Carlo simulation (e.g., [WZJ12] for STT-MTJs). However, previous implementations were too slow to be adapted for high-accuracy simulations needed for large memory array. As a result, limited samples were simulated in previous STT-RAM studies [LAS08, ZWC11], which could not address WER below $10^{-4}$. This may lead to inappropriate designs, e.g., the WER of $10^{-8}$ requires 20% more write current than $10^{-4}$ for STT-MTJs. Moreover, the context of circuit-level optimization is also essential given that peripheral circuit can significantly affect memory performance. For instance, MRAM can leverage circuit techniques to mitigate the WER by trading off the speed and power.

In this chapter, we perform the first comprehensive circuit-level comparison between MeRAM and STT-RAM with respect to reliability, power, performance, density, and scalability using a high-speed Monte-Carlo simulator. The chapter is organized as follows. Section 5.2 describes the LLG based model and simulation in detail. Section 5.3 analyzes the scalability of STT-RAM and MeRAM. Section 5.4 designs MRAM cells under process and temperature variation and compares the cell density of two MRAMs with 32nm design rules. Section 5.5 analyzes the WER of the nominal MTJs and MRAMs with process and temperature variation separately. Section 5.6 introduces the PWSA multi-write

design and carries out a circuit-level comparison with respect to write latency, energy and MRAM failure analysis. Section 5.7 concludes the chapter.

## 5.2 Modeling and Simulation




(a) (b)

Figure 5.1: (a) VC-MTJ is switched by unidirectional voltage pulses. The first two same pulses switch the resistance state of a VC-MTJ from P to AP and then back to P, the third double-width pulse make two switches continuously. (b) STT-MTJ is switched by directional current pulses, and the switching directions depends on the direction of current.

Both STT-MTJ and VC-MTJ are resistive memory devices, their resistances are determined by the magnetization directions of two ferromagnetic layers. The direction of one layer is fixed (referred to as reference layer) while the other one can be switched (referred to as free layer). A low resistance is present when magnetic directions in the two layers are parallel (referred as P state); a high resistance is present when the two directions are anti-parallel (referred as AP state). The two states are utilized to store "0" and "1". Tunnel magnetoresistance (TMR, defined as $(R_H - R_L)/R_L$) over 200% has been demonstrated, which means that the high resistance can be over 3X of the low resistance. Based on the magnetization direction of the two layers, MTJs are classified as in-plane and out-of-plane (perpendicular magnetized) devices. Recently, STT-RAM with out-of-plane MTJs is found to have lower write current and less fabrication challenge than in-palne MTJs [SLL11, ZZL12, LRD05, Hua08]. The magnetic anisotropy of out-of-plane MTJs is dominated by the perpendicular magnetic anisotropy (PMA). In this chaper, we

consider the STT-RAM and MeRAM with out-of-plane MTJs.

Although STT-MTJs and VC-MTJs share a similar device structure and data storing mechanism, their switching mechanisms differ as shown in Fig. 5.1a and Fig. 5.1b, e.g., in an STT-MTJ, polarized electrons flowing from the reference layer to the free layer switch the magnetization of the free layer to P state; when electrons flow in the opposite direction, the reflected electrons from the reference layer switch the free layer to AP state.

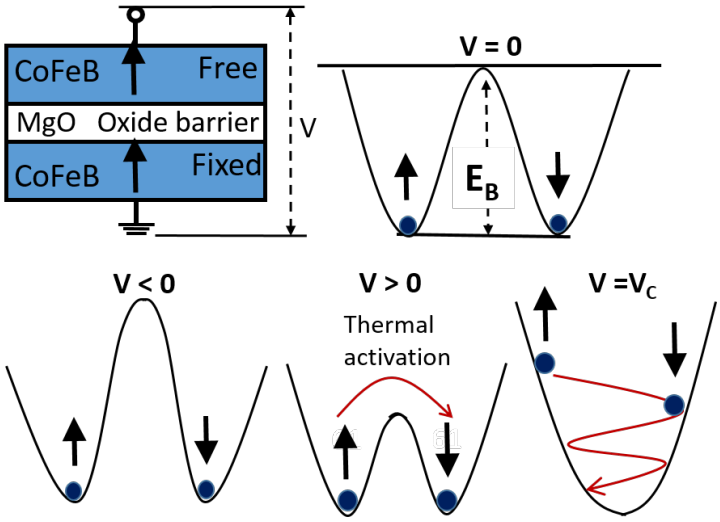



Figure 5.2: VCMA-induced precessional switching. When a voltage is applied on the VC-MTJ, the energy barrier separating the two magnetization states of the free layer is reduced so that the magnetization state starts to spin.

Unlike STT-MTJ, VC-MTJ utilizes an unidirectional voltage pulse to make both switches from AP to P and from P to AP. As is illustrated in Fig. 5.2, the energy barrier Eb separates the two stable states of the free layer magnetization (pointing up and down) when the voltage applied across the VC-MTJ is 0. The energy barrier Eb decreases with the voltage increase due to VCMA effect. When the voltage reaches $V_{Write}$ ($> 0$, see Eqn. 5.8), full $180^o$ switching can be achieved by timing the precessional switching of magnetization.

In the MTJ switching simulation, the magnetization in the free layer during any short interval (e.g., 0.25ps in our setup) is described by an LLG differential equation. The entire switching is captured by iteratively solving the LLG equations in sequence. The WER is then extracted from numerous simulations in a Monte-Carlo approach. Shorter interval and more simulations can improve the accuracy at the expense of time. The LLG

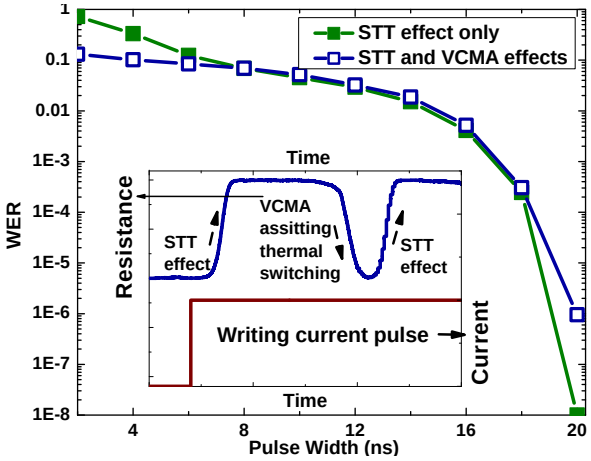



Figure 5.3: During a write of the STT-MTJ, VCMA may assist the thermal activation to cause unintended switching. This effect can improve the switching probability when the write pulse width is insufficient to switch the STT-MTJ, on the other hand, may lead to switching failure when the write pulse width is sufficiently long.

equation (5.1) describes the dynamic behavior of the free layer magnetization vector $\boldsymbol{M}$ in the presence of an external field ($\boldsymbol{H_{Ext}}$), shape anisotropy ($\boldsymbol{H_{Shape}}$), PMA ($H_{PMA}$), and thermal fluctuation ($\boldsymbol{H_{Therm}}$), as follows.

$$\begin{aligned} \frac{d\boldsymbol{M}}{dt} &= -\gamma\left(\boldsymbol{M}\times\boldsymbol{H}\right)+\frac{\alpha}{M_S}\cdot\boldsymbol{M}\times\frac{d\boldsymbol{M}}{dt} \\ &+\gamma\frac{\alpha_J\left(\theta\right)}{M_S}\boldsymbol{M}\times\left(\boldsymbol{M}\times\boldsymbol{p}\right) \\ \boldsymbol{H} &= \boldsymbol{H_{Ext}}+\boldsymbol{H_{Shape}}+H_{PMA}+\boldsymbol{H_{Therm}} \end{aligned} \tag{5.1}$$

Where $\gamma$ is the gyromagnetic ratio, $\boldsymbol{H}$ is the effective magnetic field, $\alpha$ is the intrinsic damping constant, $M_S$ is the saturation magnetization, and $\alpha_J$ is the amplitude of the spin-transfer torque induced by current. $H_{PMA}$ can be reduced by voltage due to the VCMA effect, which is expressed below.

$$\begin{aligned} H_{PMA} &= H_{PMA}\left(0\right)\cdot\left(1-\zeta\cdot V_{MTJ}\right) \\ H_{PMA}\left(0\right) &= 2K/\left(t_{FL}\cdot M_S\right)-M_S \\ \zeta &= \xi/\left(K\cdot t_{MgO}\right) \end{aligned} \tag{5.2}$$

Where $V_{MTJ}$ is the applied voltage, $t_{FL}$ and $t_{MgO}$ are the thickness of the free layer and MgO layer respectively, $K$ is the anistropy constant, $\xi$ is the anisotropy change slope, $\zeta$ is the VCMA factor with the unit of $V^{-1}$. Positive $V_{MTJ}$ causes VCMA effect to re-

duce $H_{PMA}$ as well as the perpendicular magnetization to cause precessional switching [AUA13]. An optimal applied voltage can exactly cancel out the perpendicular magnetization, and then a perfect precessional switching (controlled by the in-plane external magnetic field $H_{Ext}$) starts, during which the magnetization in the free layer rotates. The optimal pulse width equals to the half cycle of the precessional switching [DAC13]. More specifically, the optimal pulse allows the magnetization to rotate exactly $180^o$. The VCMA effect is considered for both STT-MTJ and VC-MTJ, while previous circuit-level STT-RAM studies ignore it. As an example, the impact of VCMA effect on an STT-MTJ is shown in Fig. 5.3: VCMA can change the WER. When a write current is applied, the voltage drop on the STT-MTJ reduces the PMA and increases the chance of thermal activated switching. Hence when the write pulse is not long enough to guarantee a switch, the thermal activated switching assisted by VCMA increases the switching probability, but when the write pulse is long enough, it induces errors.

The $\boldsymbol{H_{Therm}}$ in (5.1) is the thermal fluctuation field and randomly determined as a variable following Normal distribution at each simulation interval (5.3).

$$\boldsymbol{H_{Therm}} = Norm3d(0, \sqrt{\frac{2k_BT}{\gamma M_S t_{FL} A}}) \tag{5.3}$$

Where $A$ is the area of the MTJ, $k_B$ is the Boltzmann constant, and $T$ is the temperature.

Temperature significantly affects the MTJ switching behavior, e.g., the WER of an STT-MTJ can increase from $10^{-8}$ to $10^{-6}$ with temperature rising from 300K to 350K (see Fig. 5.9). Except $\boldsymbol{H_{Therm}}$, other terms in (5.1) also change with temperature as described in (5.4) [AAY14], which are commonly ignored in previous large-scaled MRAM studies [SBL11, SMN11]. In-situ thermal sensors [ZLL13, ZLL15] may help to monitor MRAM temperature and modulate MTJ write schemes.

$$\begin{aligned} M_S(T) =& M_S^* \left(1 - (T/T^*)^{3/2}\right) \\ K(T) =& K^* \cdot (M_S(T)/M_S^*)^{2.18} \\ \xi(T) =& \xi^* \cdot (M_S(T)/M_S^*)^{2.83} \end{aligned} \tag{5.4}$$

Where $T^*$ is 1120K, and $M_S^*$, $K^*$, and $\xi^*$ are corresponding parameters at 1120K.

The spin-transfer torque effect is described in (5.5) [LRD05].

$$\alpha_J(\theta) = \frac{\hbar g(\theta)}{2eM_S t_{FL}} J \tag{5.5}$$

Where $\hbar$ is the reduced Plank constant, $g(\theta)$ is the spin-torque efficiency factor [Slo96], $\theta$ is the angle between the two magnetizations of the free layer and reference layer, and $J$ is the current density through MTJ. $g(\theta)$ can be further expanded at (5.6) [Slo96, ZZL12, FKB05].

$$\begin{aligned} g(\theta) =& g_{Tunnel}(\theta) + g_{SV}(\theta) \\ g_{SV} =& \left[-4 + (1+P_{SV})^3 (3+\cos\theta) / \left(4 \cdot {P_{SV}}^{3/2}\right)\right]^{-1} \\ g_{Tunnel} =& 0.5 \cdot P_{Tunnel} / \left(1 + P_{Tunnel}^2 \cos\theta\right) \end{aligned} \tag{5.6}$$

Where $g_{Tunnel}$ and $g_{SV}$, as functions of $\theta$, are polarization efficiency of tunnel current and spin valve respectively. $P_{Tunnel}$ and $P_{SV}$ are material-dependent polarization factors for the tunnel current and current passing through ferromagnetic layers respectively [FKB05]. These two parameters are not necessarily equal, while we use 0.66 [SCS08] for both of them in this chaper. The required switching current (known as critical current) differs from switching directions due to the difference in polarizing efficiency [HYY05]. The parameters used in the model and simulation are listed in Table 5.1.

Table 5.1: Modeling parameters at 300K.

| $\gamma$ $[m/(A \cdot S)]$ | $M_S$ [A/m] | $\xi$ $[fJ/(V \cdot m)]$ | $\alpha$ |
|---|---|---|---|
| $2.2 \cdot 10^5$ | $1.2 \times 10^6$ | STT: 37 [BMS11, AAY14], VC: 85 [NYT13] | 0.02 |
| $H_{Ext}$ $[A/m]$ | $K$ $[J/m^2]$ | $P_{SV}$, $P_{Tunnel}$ | TMR |
| $1.1 \cdot 10^4$ | $1.068 \cdot 10^{-3}$ | 0.66 [SCS08] | 100% |

Inspired by the massive floating point calculations involved by the LLG equation and highly independent operations in Monte-Carlo simulations, we implement the switching simulator in CUDA, and it completes 100,000 simulations within 2s on NVIDIA Tesla M2070. The model has been validated. The speed improvement comes from highly parallel simulations.

## 5.3 Scalability

In this subsection, we analyze the scalability of STT-RAM and MeRAM regarding retention, write power, area, and fabrication challenges. Retention, as one of the most important metric for memory system [FW08], determines the available data-storing time and thus is a non-scalable parameter [ITR11]. An MTJ with low retention is easy to flip, but high retention increases the write difficulty. Considering the trade-off, an efficient design should have its retention as low as possible but satisfy application requirement. For STT-MTJ and VC-MTJ, the retention time (mean time to false switching during idle state) $\tau$ is an exponential function of thermal stability$\Delta$ [NSM11, RDJ02].

$$\begin{aligned} \tau =& \tau_0 \exp(\Delta) \\ \Delta =& \frac{H_{K,eff} M_S A t_{FL}}{2 k_B T} \end{aligned} \tag{5.7}$$

Where $H_{K,eff}$ is the sum of perpendicular components of $\boldsymbol{H_{Shape}}$, $\boldsymbol{H_{Ext}}$, and $H_{PMA}$. Based on [LRD05, AUA13], we derive the critical current of STT-MTJ and the optimal voltage of VC-MTJ as functions of $\Delta$ and MTJ area $A$ in (5.8).

$$\begin{aligned} I_{STT}(A, \Delta) \approx& \frac{4 k_B T e}{\hbar g} \Delta \propto \frac{\Delta}{g} \\ V_{VC}(A, \Delta) \approx& \frac{2 k_B T \Delta}{\zeta M_S^2 t_{FL} A} \propto \frac{\Delta}{\zeta A} \end{aligned} \tag{5.8}$$

Where $e$ is the elementary charge, $g$ is the spin-torque polarization efficiency, and $\zeta$ is the VCMA factor. From (5.8), the critical current $I_{STT}$ of STT-MTJ does not directly depend on the MTJ dimension given that the thermal stability is constant. But as $g$ increases with decreased $A$ due to the sub-volume excitation for large MTJs with lateral size over 50nm [OIE15], $I_{STT}$ can be slightly reduced by scaling dimension. However, the reduction trend does not continue for small MTJs. The optimal voltage $V_{VC}$ of VC-MTJ is inversely proportional to $A$ indicating that it will increase with dimension scaling. Hence, the key for scaling both technologies is finding materials that provide more $g$ and $\zeta$.

With respect to the memory density, access transistors dominate the area rather than the MTJs (see Fig. 5.4). Because of the non-scaling critical current for small STT-MTJs, access transistors in STT-RAM have to increase the width/length ratio with dimension

scaling down. MeRAM always uses minimum sized transistors and hence promises better scalability in density. Alternatively, MeRAM can be integrated in a much denser cross-bar structure, unlike STT-RAM [DAC13].

In terms of fabrication, both STT-RAM and MeRAM face the challenge scaling MgO thickness. Scaling dimension forces STT-MTJs to reduce MgO thickness, and thin MgO may contain defects, such as pin-holes, which can cause MTJs to fail. Though MeRAM has thicker MgO because of the high resistance of VC-MTJs, increasing write voltage may cause MgO breakdown. Again, these challenges can be overcome by finding better materials with higher $g$ and $\zeta$.

## 5.4 MRAM Cell Design and Variation

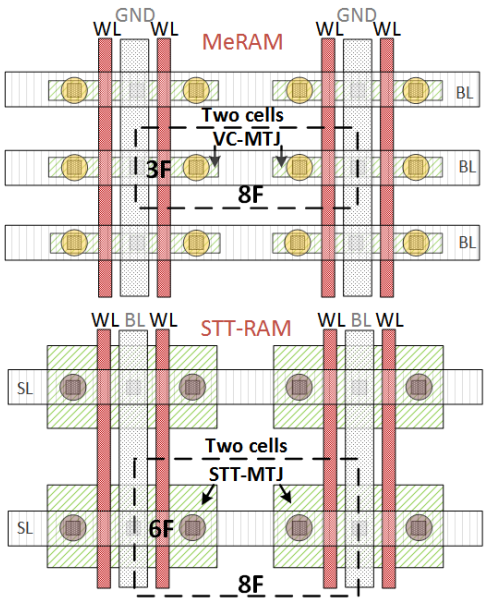



Figure 5.4: Layouts of STT-RAM and MeRAM under 32nm design rules. The area of an STT-RAM cell is twice the area of a MeRAM cell, as an STT-MTJ requires a 3X wider access transistor than a VC-MTJ. Vertical transistor like nanowire may help to reduce area inefficiency [WG14a]

As discussed in Section 5.3, both STT-MTJ and VC-MTJ face scaling problems, and enlarging MTJs exacerbates write difficulty but does not improve thermal stability due to the sub-volume excitation [SRN11]. Considering these factors, we set the diameter of STT-MTJs and VC-MTJs to 60nm (i.e., a circular MTJ structure), which has been

demonstrated [OKM12] for STT-MTJs. Access transistors and peripheral circuit are built with 32nm planar CMOS technology. The layouts of STT-RAM and MeRAM are drawn in Fig. 5.4 under 32nm design rules. The density of MeRAM is twice of STT-RAM for the reason that STT-MTJ needs 3X wider access transistors.

Table 5.2: Design parameters for MTJs and access transistors. The transistors' threshold voltage variation considers the effects of line edge roughness (LER), random dopant fluctuation (RDF), and non-rectangular gate (NRG). Access transistors of MeRAM have larger threshold voltage variation because narrow transistors are affected more by NRG, RDF, and LER.

| Devices | Parameters | Mean | Variation |
|---|---|---|---|
| STT-MTJ | Diameter | 60nm | $\sigma$=1nm [NOL06] |
| | MgO thickness | 0.7nm | $\sigma$=0.001nm [DSS06] |
| | $T_{FL}$ | 1.20nm | $\sigma$=0.003nm [SCW00] |
| | Thermal stability | 71.6 (51.9@350K) | $\sigma$=3.0 (2.3@350K) |
| | Resistance | 1KΩ / 2KΩ | dependence |
| | Cell area | $24F^2$ (F: MTJ diameter) | |
| VC-MTJ | Diameter | 60nm | $\sigma$=1nm [NOL06] |
| | MgO thickness | 1.3nm | $\sigma$=0.001nm [DSS06] |
| | $T_{FL}$ | 1.19nm | $\sigma$=0.003nm [SCW00] |
| | Thermal stability | 73.7 (53.6@350K) | $\sigma$=3.1 (2.3@350K) |
| | Resistance | 100KΩ / 200KΩ | dependence |
| | Cell area | $12F^2$ (F: MTJ diameter) | |
| Access transistor | Length | 30nm | $\sigma$=2.1nm [ITR11] |
| | Width | 200nm(STT)<br>48nm(Me) | $\sigma$=2.1nm [ITR11] |
| | Threshold voltage | 493mV, LER, RDF, NRG [ITR11, YLN08] | $\sigma$=22.6mV (STT)<br>$\sigma$=42.4mV (Me) |

Table 5.2 lists the design parameters of MRAM cells (nominal and variation). In an MRAM cell, an MTJ is connected with an access transistor (1T1M), The MTJ resistance variation due to the MTJ shape variation was identified as a big design concern [CWS00, LLS08, LAS08, ZWC11]. But it is actually a secondary problem of the re-deposition of etched products on the MTJ sidewalls in current plasma-based etching system, where the re-deposition may cause an MTJ failure [PKJ11] and a shape change. Developing selective-etching process is expected to fix both problems by forming volatile compound

during a etch. Less than 4% in size variation has been shown for fabricated 50nm STT-RAM [NOL06]. We pessimistically choose $\sigma = 1nm$ in our simulation where $6\sigma$ is 10% of the MTJ diameter.

The designed MTJs have thermal stability margins of 10 $\sigma$ at 300K and 5 $\sigma$ at 350K for the requirement of 40.3 [ITR11] (i.e., 10 years retention time ). External magnetic field in VC-MTJs assists the precessional switching, but reduces thermal stability, and hence the free layer thickness of VC-MTJs is set thinner to offset the thermal stability loss.

Comparing with the MTJ, CMOS variation has been well analyzed. Major variations in the 32nm planar technology are considered in Table 5.2. FinFET and Tunneling FET technologies [LWP15a, WPC14, WPC16], which is possibly introduced for scaled MRAM, shows slightly smaller impact from process variation [WLP13].

## 5.5 Write Error Rate of MRAM

The reliability problems of MRAMs include retention error, read disturbance, read failure, and write error. We focus on the write error in this section, which is our main contribution, and the other failures are discussed in Section 5.6.2.

### 5.5.1 Write Error Rate of MTJs without Variation

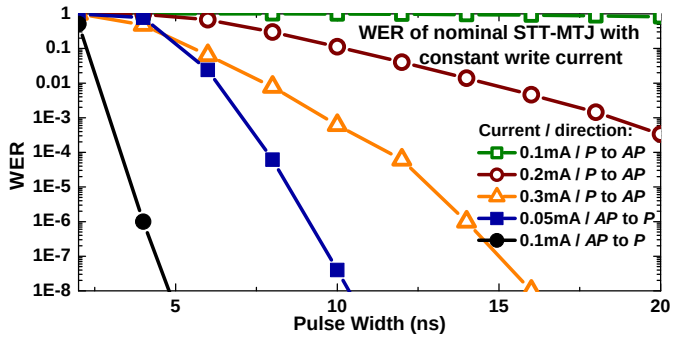



Figure 5.5: WER of the nominal STT-MTJ as a function of pulse width for different perfect current pulses (constant current) and switching directions.

Fig. 5.5 shows the WER of the nominal STT-MTJ. The two switching directions have different WER due to the asymmetric polarization efficiency (5.6). When the STT-MTJ

switches from AP to P, the polarizing current changes from the majority to minority, while the polarizing current changes from the minority to majority in the opposite direction.

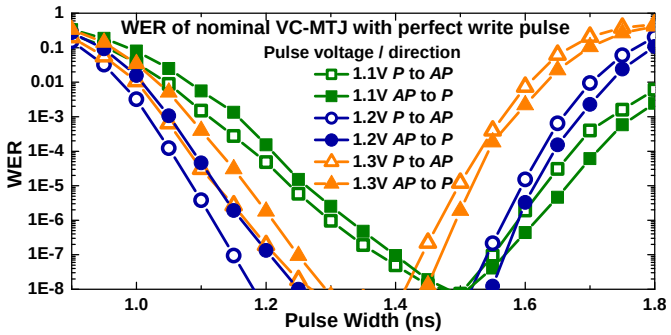



Figure 5.6: The WER of the nominal VC-MTJ as a function of pulse width for different perfect voltage pulses and switching directions. A VC-MTJ has an optimal pulse, which leads to the lowest WER. The curve of 1.2V has the lower overall WER than 1.1V and 1.3V, indicating 1.2V is closer to the optimal voltage.

The WER of the nominal VC-MTJ is shown in Fig. 5.6. The curve of 1.2V is observed to have lower overall WER (for different pulse widths) than 1.1V and 1.3V indicating that it is closer to the optimal voltage. A non-optimal voltage either under-compensate or over-compensate the PMA, resulting in an imperfect precessional switching and thus a higher WER. As can be seen in Fig. 5.6, the low WER region of 1.3V averagely locates left (shorter pulse width) to 1.2V and 1.1V, as 1.3V over-compensates the PMA more to result in a faster precessional switching. Small WER asymmetry is observed for the two switching directions, because the write voltage induces leakage current and corresponding STT effect, which assists the switching from P to AP but resists the switching from AP to P.

### 5.5.2 Write Error Rate of MRAM Array

To estimate the WER of an entire array with temperature and process variations, WER must be simulated for different cells that have varying design parameters.

The variations of access transistors result in variation of write pulse voltage, rise and fall time as is shown in Fig.5.7. We obtain the distribution of the write pulse using Monte-Carlo SPICE simulations. In simulations, an access transistor is connected with a resistor and a capacitor (a lumped model for the MTJ [SGP00]). The parameters of

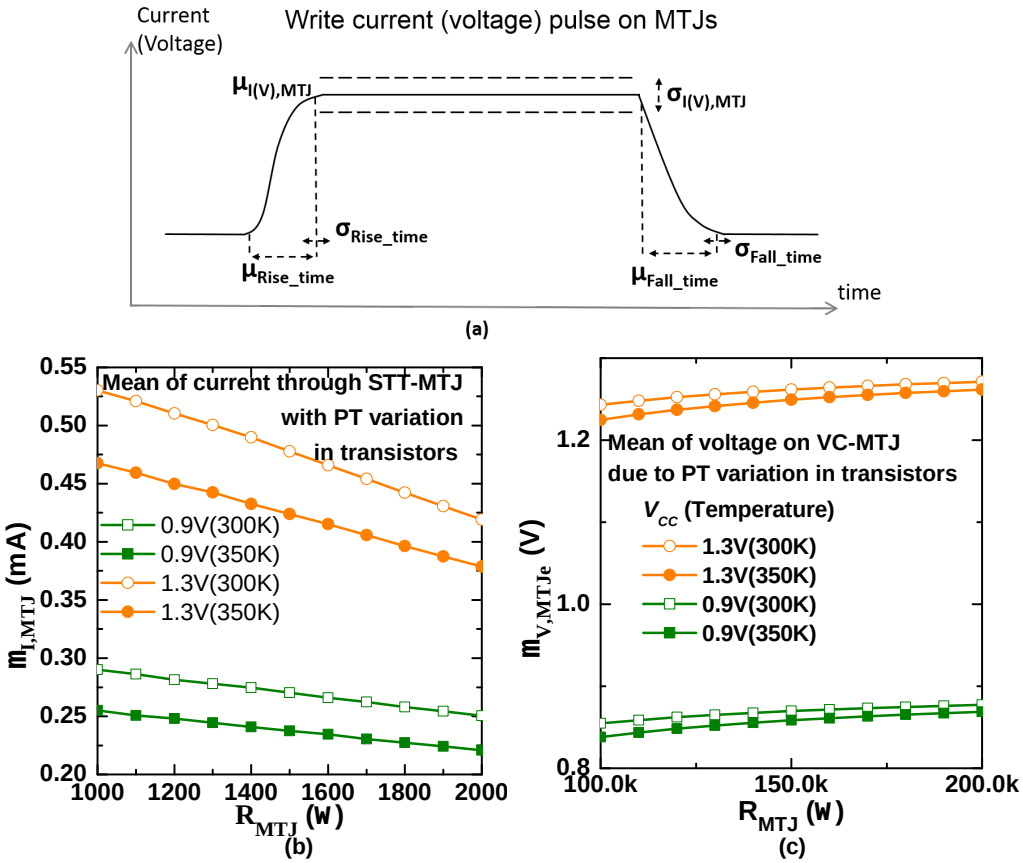



Figure 5.7: (a) Write current (voltage) pulse on STT-MTJs (VC-MTJs). The rise and fall time are measured by the time while voltage is rising and falling between 10% and 90% of the peak voltage respectively. Mean of write current on (b) STT-MTJs and (c) VC-MTJs as a function of MTJ resistance.

access transistors are randomly determined based on Table 5.2. Then the distribution of pulse current (current through STT-MTJs), pulse voltage (voltage on VC-MTJs), pulse rise time, and pulse fall time are statistically extracted from 100,000 simulations for each MTJ resistance state (i.e., resistance changes during switching) and $V_{CC}$ (the supply voltage between 0.9V to 1.3V, which drops over an access transistor and an MTJ in series). The standard deviation ($\sigma$), mean ($\mu$) of pulse current and voltage vary with MTJ resistance. As is shown in Fig. 5.7, the $\mu$ of pulse current changes up to 26.5% with STT-MTJ resistance, whereas the $\mu$ of pulse voltage only changes below 3.5% with VC-MTJ resistance because that the high resistance of VC-MTJs drops more than 95% of the $V_{CC}$. The $\sigma/\mu$ of pulse current in STT-RAM is up to 16%, whereas the $\sigma/\mu$ of pulse voltage in MeRAM is below 1% for the reason that in STT-RAM the pulse current

Table 5.3: Summary of write pulse variation due to transistor process variation at temperature of $300^oC$ and $350^oC$. Mean shift is the percentage change of parameters' mean between high and low MTJ resistance states

| MTJ | Parameters | mean shift | $\sigma/\mu$ |
|---|---|---|---|
| STT-MTJ | $I_{MTJ}$ | $< 26.5\%$ | $< 16\%$ |
| | Rise time | $< 7.0\%$ | $< 10.6\%$ |
| | Fall time | $< 11.5\%$ | $< 14.1\%$ |
| VC-MTJ | $V_{MTJ}$ | $< 3.5\%$ | $< 1.0\%$ |
| | Rise time | $< 3.6\%$ | $< 11.1\%$ |
| | Fall time | $< 88.1\%$ | $< 7.3\%$ |

is mainly controlled by access transistors and thus suffers more impact from transistor variation. The $\sigma$, $\mu$ of pulse rise time are mainly determined by access transistors, which barely do not depend on MTJ resistance. For a given $V_{CC}$, the $\mu$ varies within 7% with resistance, and the $\sigma/\mu$ is around 10%. By contrast, the fall time is mostly determined by the leaking current through MTJs. In STT-RAM the $\mu$ of pulse fall time varies between 61.6ps and 70.6ps with MTJ resistance, while the $\mu$ for MeRAM is much longer and varies between 128ps and 248ps because of the high resistance of VC-MTJs. The $\sigma/\mu$ of pulse fall time due to transistor variation is around 9% (maximum 14.1%) for STT-RAM and 5% (maximum 7.3%) for MeRAM for different $V_{CC}$ and temperatures. We summarize the write pulse variaiton in Table 5.3.

The $\sigma$, $\mu$ of pulse current/voltage, rise/fall time are fitted to polynomial models of MTJ resistance, which are accurate enough as their dependence on resistance is nearly linear. The models are inputs to the CUDA simulator. The Monte-Carlo simulation flow is shown in Fig. 5.8. At the beginning of a simulation, pulse voltage/current, rise/fall time, and MTJ parameters are generated following Normal distribution. During the simulation, pulse voltage, current, and fall time are updated according to the MTJ resistance state.

The WER of a 1T1M STT-RAM is shown in Fig. 5.9. As expected, the WER of STT-RAM decreases monotonically with increasing $V_{CC}$ and pulse width, and the WER

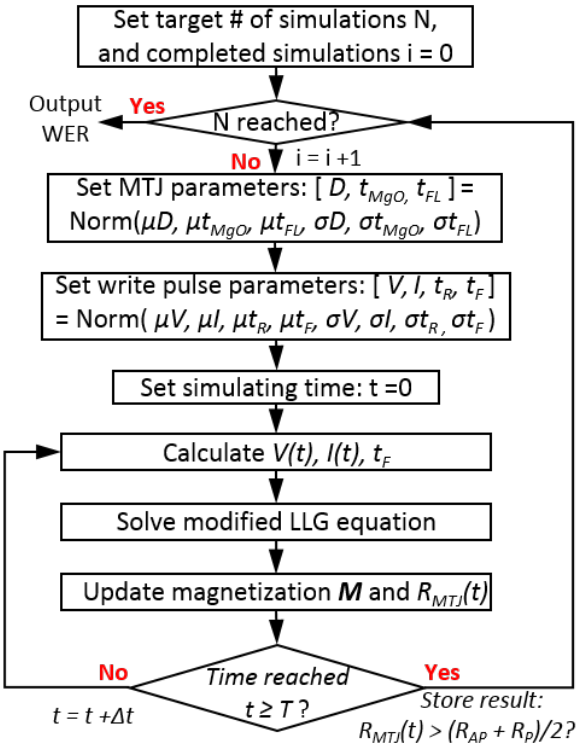



Figure 5.8: A Monte-Carlo simulation flow to obtain the WER of 1T1M MRAM array. N is the sample size, and T is the simulation time including a writing time and a waiting time (for the MTJ to settle down, e.g., waiting time is 20ns in the simulations).

increases with temperature. The switching from P to AP shows higher WER due to the asymmetry in spin-torque polarization efficiency.

The WER of a 1T1M MeRAM is shown in Fig. 5.10. Unlike the results of the nominal VC-MTJ, the MeRAM with process variation cannot achieve WER below $10^{-8}$ because there is no common optimal voltage for all VC-MTJs in the MeRAM. Temperature also has significant impact on the MeRAM: the WER of the the pulse (1.26V/1.42ns), which gives the lowest WER at 300K, increases 1000X at 350K; the voltage that gives the lowest WER changes from 1.26V at 300K to 1.20V at 350K because high temperature leads to low thermal stability and thus low required voltage (5.8); the pulse width giving the lowest WER for a given voltage decreases with temperature for the reason that higher temperature reduces the horizontal demagnetization field, then the external field is less canceled and drives the precessional switching faster. Despite of the high WER, MeRAM shows clear speed advantage.

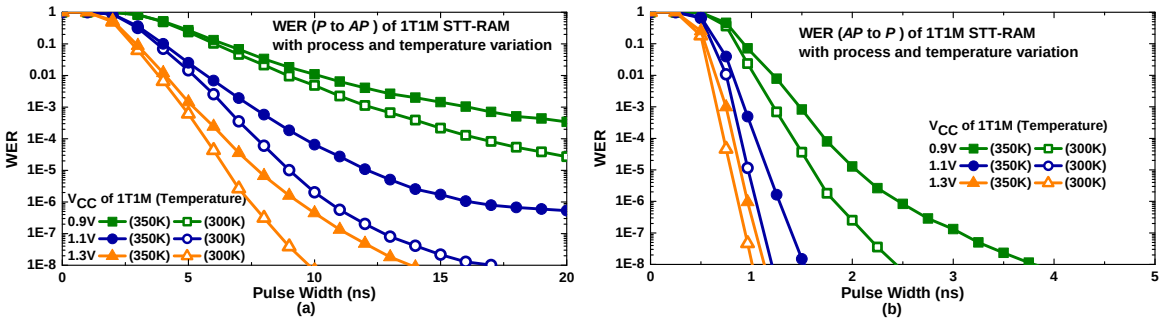



Figure 5.9: The WER of an STT-RAM under process and temperature variation for different write pulses and switching directions (a: P to AP, b: AP to P).

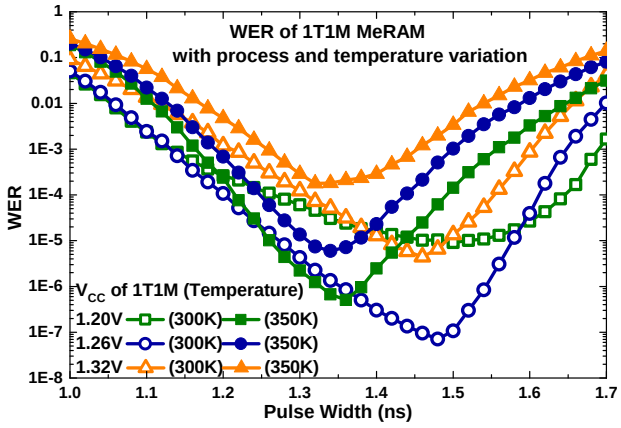



Figure 5.10: WER of an MeRAM under process and temperature variation for different write pulses. The WER is averaged over two switching directions.

## 5.6 Circuit-level Evaluation

### 5.6.1 MRAM Write/Read with PWSA Multi-write Design

As well desired, peripheral circuit can improve the reliability of MRAMs at the expense of speed, power, and area. Memory with multi-write schemes can significantly reduce write errors, e.g., incremental step pulse programming for Flash technology [SSL95]. We utilize PWSA [LAD15] to reduce the WER of MRAMs, where PWSA is a peripheral circuit designed for STT-RAM's and MeRAM's write and read operations. As MeRAM uses an unidirectional pulse to write both "1" and "0", PWSA uses an operation called pre-read to check the stored data prior to a write, and no write is performed if the stored data matches the writing one. In addition to pre-read, PWSA also enables multi-write policy which can perform additional write after a write error. The data flow in Fig. 5.11

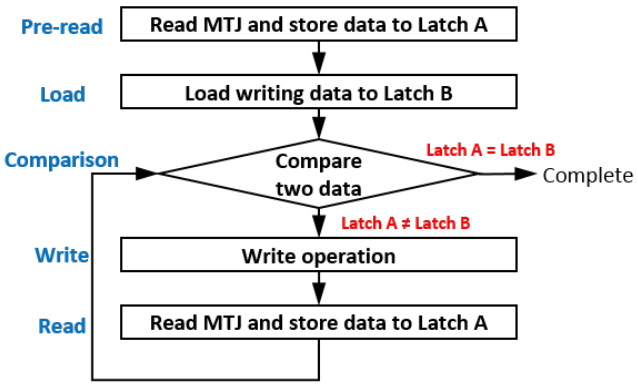



Figure 5.11: Data program flow for the PWSA multi-write design.

illustrates the PWSA multi-write design. It is noticed that read failure has been well analyzed in [LAS08, LLS08, DRT12], which is mainly caused by process variation and is a permanent failure (not like write error). Read failure can be eliminated by chip test at the expense of yield loss. All pre-read, comparison, and read share one sense amplifier and are assumed to be error-free operations in the PWSA multi-write design.

Table 5.4: Energy and delay for operations in the PWSA multi-write circuit at 300K temperature.

| Operations | Energy(fJ) | | Delay (ns) | |
|---|---|---|---|---|
| | STT-RAM | MeRAM | STT-RAM | MeRAM |
| Read (Pre-read) | 54.7 | 91.0 | 1.8 | 2.0 |
| Load | 122.4 | 122.4 | 1.3 | 1.3 |
| Comparison | 15.2 | 15.2 | 0.4 | 0.4 |
| Write (logic) | 691 | 318 | 0.5 | 0.5 |
| Write (MTJ) | 680/ns | 10.1 | $\geq 3$ | 1.42 |

We divide the multi-write flow into four steps: read (also pre-read, including pre-charge and sensing), load, comparison, and write (including control of logic circuit and MTJ switching). To obtain reasonable delay and energy consumption of the peripheral circuit, each sense amplifier is connected to a bit-line of 256 1T1M cells and a ref bit-line. The delay and energy for these steps are extracted from Spectre simulation of PWSA circuit [LAD15] using 32nm PTM HP model [PTM] and are listed in Table 5.4. To guarantee a good pulse shape for the write in MeRAM, a pre-charge operation is

performed to raise the bit-line voltage to $V_{CC}$ prior to turning on access transistors, where the pre-charge takes around 0.15ns. Though STT-MTJs do not have strict requirement on the pulse shape, STT-RAM needs to raise bit-line voltage to bias access transistors to offer required write current, which takes similar delay and consumes more power due to its larger access transistors. Conversely, the read energy of STT-RAM is lower, because the low resistance of STT-MTJ allows lower read voltage than VC-MTJ. The difference of read delay is within 0.2ns. MeRAM shows great advantage in the MTJ switching energy, whereas this energy is a bottleneck for STT-RAM due to the high leakage current caused by the low resistance of STT-MTJs. The switching energy can be reduced by pre-read, though pre-read is not mandatory for STT-RAM (i.e., the STT-MTJ can be directly written with directional current).

We utilize the PWSA multi-write design to achieve reliable STT-RAM and MeRAM with the same acceptable WER and then compare the expected latency and energy of writing a word. The acceptable WER is $< 10^{-23}$ for a cell in one multi-write operation, which is slightly smaller than the soft error rate in DRAM technology [SL12b] and can be handled by error-correction code (ECC) designs. A word is a set of bits to be written in parallel, and the size of word varies from storage devices, e.g., a main memory usually writes 64 bits in parallel. The expected write latency/energy is the sum of products of the delay/energy and probability for all possible scenarios (i.e., different numbers of writes for different number of bits). In our calculations, we assume the memory system to store "1"s and "0"s in balance, which means there is 50% probability that the bit to be written equals to the bit stored in the target MRAM cell. As the multi-write design writes only to the cells that fail in the pre-read check or previous writes, so the expected energy does not count the write energy of the cells that do not need writes (i.e., energy of the pre-read and comparison is always counted). There is a maximum allowed write times per multi-write operation to avoid infinite writes in case that permanently failed cells exist, which is calculated by the number of writes to achieve the acceptable WER of $10^{-23}$, e.g., three for MeRAM at 300K.

We tried multiple write configurations for STT-RAM to explore tradeoffs between write latency and energy: with pre-read, without pre-read, and different write pulse widths for switching from P to AP including 3ns, 6ns, and 9ns. For switching from AP

to P, the pulse width is set to 3ns, which exactly achieves the acceptable WER. The $V_{CC}$ for STT-RAM is chosen to 1.3V, which is the most efficient one in Fig. 5.10a. The write pulse used for MeRAM is the optimized pulse at 300K (1.26V/1.42ns, see Fig. 5.10b). The pre-read is mandatory for MeRAM.

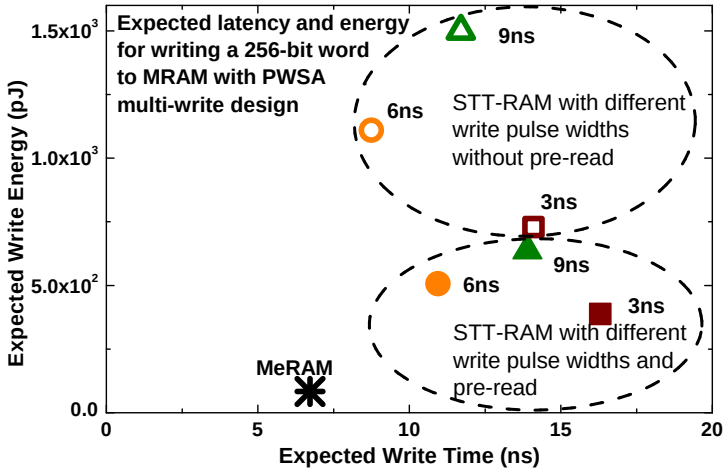



Figure 5.12: Expected word-write energy and latency for MeRAM and STT-RAM with PWSA multi-write circuit. The word size is 256bits, which are the number of bits being simultaneously written. The WER/bit after multiple writes is minimize below $10^{-23}$. The labels 3ns, 6ns, and 9ns on STT-RAM are single write pulse widths. The pre-read is not mandatory for STT-RAM. The top circled designs are STT-RAMs without pre-read operation. The bottom circled ones are STT-RAMs with pre-read operation, which count the overhead of pre-read but save unnecessary write.

Fig. 5.12 shows the expected write energy and latency for STT-RAM and MeRAM with PWSA multi-write design at 300K. Benefiting from the fast and energy-efficient switching of VC-MTJs, MeRAM shows substantial advantages of both speed and energy against STT-RAM. Among configurations for STT-RAM, the one with the shortest pulse width of 3ns and pre-read gives the lowest expected energy, as most cells can pass the comparison check in the pre-read and the first write. Nevertheless, it has the longest expected latency for that the pre-read adds latency and its high WER leads to the most write errors and write iterations. The STT-RAM with 6ns pulse without pre-read shows the fastest speed, as a comparison, the 9ns pulse has lower WER but higher latency and energy, because the benefit of the lower WER does not compensate the overhead brought by its longer pulse. This also indicates that directly using a long pulse in STT-RAM to

guarantee zero WER is not an energy-latency efficient design. The energy-latency Pareto fronts of STT-RAMs are 3ns with pre-read, 6ns without pre-read, and 6ns with pre-read.

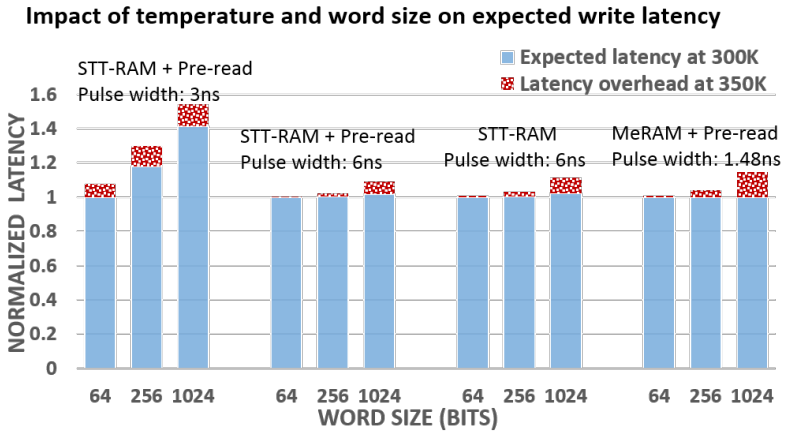



Figure 5.13: Impact of word size and temperature on the expected write latency and energy of MRAMs. All bars are normalized within each group to the expected write latency of MRAM at 300K with 64-bit word size.

Fig. 5.13 shows the impact of word size and temperature on the write latency. A longer word usually leads to more write iterations as more cells are written giving rise to more write errors. The STT-RAM with the 3ns pulse is affected most by the word size due to its high WER. The STT-RAM with 3ns pulse and MeRAM are affected most by temperature, as their WERs increase the most from 300K to 350K (i.e., the WER of a MeRAM cell increases from $9 \times 10^{-8}$ to $2.2 \times 10^{-4}$, and the WER of an STT-RAM cell with 3ns pulse increases from 0.06 to 0.09). Moreover, the temperature induced overhead increases with word size. Illustrated from the comparison of two STT-RAMs with 6ns pulse, pre-read can mitigate the impact of temperature variation for the reason that about 50% writes are saved if pre-read is enabled. For all MRAM designs, maximum 15% latency increase is shown due to temperature variation, whereas energy only shows maximum 2.3% increase at 350K against 300K. Among the energy overhead, a big portion comes from the energy of bit-line charging (i.e., 4% increase in this energy from 300K to 350K). Again, the small energy overhead is because that most cells only need one write.

The delay and power overhead of pre-read and comparison are simulated and listed in Table 5.4. In this section, we analyze the area overhead of PWSA. One PWSA contains 37 transistors and one regular STT-RAM sense amplifier contains 8 transistors [CTC10].

Considering the bit-line size of 256 cells and four bit-lines sharing one sense amplifier, the area overhead is 2.7%. However, the area of design also depends on size of transistors. The transistors of sense amplifier for STT-RAM are much larger than those for MeRAM given that STT-RAM requires larger write current. Indeed the PWSA (37 transistors) for MeRAM occupies 20% less area than the regular sense amplifier (8 transistors) for STT-RAM.

### 5.6.2 Failure Analysis and Error Correction

Table 5.5: Failure types and FIT for a 16MB memory bank. $10^9$ reads and $10^9$ writes in a bank-hour are assumed. The read disturbance rate is extrapolated from simulations. As a comparison, the FIT of single-bit fault in a 16MB bank is about $2 \cdot 10^{-4}$ [SL12b], and the FIT of DDR bus errors is about 100 [MKS10].

| Failures | write errors | retention error | read failure | *read disturbance |
|---|---|---|---|---|
| Types | non-persistent | non-persistent | persistent | non-persistent |
| MeRAM | $< 10^{-5}$ | $< 0.58$ | $< 0.0029$ | $4 \cdot 10^{-6}$ |
| STT-RAM | $< 10^{-5}$ | $< 3.4$ | $< 10^{-15}$ | $3 \cdot 10^{-43}$ |

As is listed in Table 5.5, memory failures in STT-RAM and MeRAM are classified into four types: write errors, retention errors, read failure, and read disturbance. We use failure-in-time (FIT, average number of failures in a billion-device-hours) to represent the error rate for these faulure types.

Write errors have been analyzed in the Sections 5.5 and 5.6.1. The multi-write scheme significantly reduces write error rate. Its FIT for an 16-MB MeRAM bank decreases from over $10^{10}$ without multi-write scheme to below $10^{-5}$ with multi-write scheme.

False switching of MTJs during idle state is called retention error. As is mentioned in Section 5.3, the VC-MTJs and STT-MTJs have been designed with enough margin in thermal stability to minimize the retention error. The FIT of retention error in a 16-MB MRAM is calculated according to Table 5.2 and is listed in Table 5.5.

Read failure due to the MTJ resistance variation [LAS08, LLS08, DRT12] is a per-

sistent memory failure which stay in memory and frequently produces errors. More specifically, large shape variation of an MTJ can lead to significant resistance change which decreases sensing margin and results in read errors. Read errors are produced in all reads to a failed MTJ (with large resistance change) indicating that the multi-write design also creates write errors due to the involved read step. However, the multi-write design does not increase read error rate for the fact that such write errors only occur on failed MTJs, and the failed MTJs are always read out incorrectly. The resistance change due to shape variation is mainly caused by wafer-level process variation [TSC99], which can be minimized by increased TMR or recently developed peripheral circuit designs, e.g., the local-reference reading scheme [KAG07] and the self-reference scheme [CWZ10]. In our experimental setup, the AP and P resistance for STT-RAM (MeRAM) are $2,000\Omega(200,00\Omega)$ and $1,000\Omega(100,000\Omega)$ respectively, and reference resistors are $1,500\Omega(150,000\Omega)$. Reference resistors are fabricated with traditional CMOS process, and their variation are negligible compared to MTJs. The MTJ resistance variation is assumed to follow Gaussian distribution [LAS08, LLS08]. In [DSS06], standard deviation of MTJ resistance is measured as 1.5% of mean resistance from a 4-Mb MRAM array. Accordingly, we calculate its MgO thickness variation in Table 5.2 and estimate the resistance standard deviation of 2.6% for STT-MTJs and VC-MTJs with the diameter of 60nm. By setting 0.05V sensing margin for sense amplifiers to operate functionally (i.e., 0.05V is enough for the limited variation of large sized sense amplifiers), our sensing scheme (using PWSA and 2ns sensing time) can tolerate 17.5% resistance variation (i.e., STT-RAM: 20% in AP and 45% in P, MeRAM: 17.5% in AP and 40% in P). It is noticed that we consider both access transistor variation and MTJ shape variation, but the access transistor variation has negligible impact due to the low sensing voltage (0.2V for STT-RAM and 0.48V for MeRAM), where sensing current is dominated by MTJs. The read failure rate of an MTJ due to resistance variation is $1.75 \cdot 10^{-10}$, which gives rise to 99.22% yield for a 16-MB bank array. Redundancy technique of sparing columns is a common technology for yield improvement. By adding one sparing column (every column has 256 cells) to every mat (contains multiple rows and columns, e.g., a 16-MB bank has 64 mats, and every mat has 256 rows and 8192 columns), the yield of a 16-MB memory bank is improved to 99.994% with 0.01% area overhead.

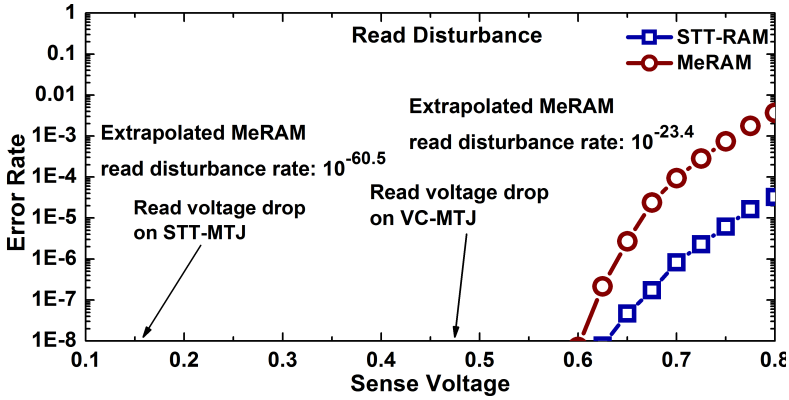



Figure 5.14: Read disturbance rate as a function of read voltage. The read disturbance rate for MeRAM and STT-RAM are extrapolated to the read voltage drop on MTJs (0.48V and 0.15V are respectively for VC-MTJs and STT-MTJs).

The failure rate of read disturbance is close to zero when short read pulse ($< 2ns$) and low read voltage (0.48V on VC-MTJs, and 0.15V on STT-MTJs) are used [ZWC11]. We have simulated the read disturbance of MeRAM and STT-RAM as functions of read voltage with the precision of $10^{-9}$ using the CUDA LLG-based model and extrapolated the read disturbance to our read voltage using polynomial models as shown in Fig. 5.14.

Error-correction-code (ECC) is a common technique to protect memory from memory errors. We use MEMRES (a fast system-level memory reliability simulator) [WHZ15a] to simulate and analyze the impact of MRAM introduced failures on an 8-GB memory. The memory is comprised of 512 16-MB banks and is protected by in-memory ECC (SECDED) (i.e., locates in memory banks) and in-controller SECDED/Chipkill (i.e., locates in memory controller) [BGM88b, Del97a]. MRAM introduced failures (listed in Table 5.5) are included in MEMRES simulations in addition to typical memory logic-circuit induced failures (e.g., bank failure, row failure, column failure, and etc.[WHZ15a]). Based on simulated results, the probability that MRAM introduced failures cause an ECC uncorrectable error in an 8-GB memory is $< 0.0001\%$ for 5-year operating time (i.e., no such error is found in 10,000,000 5-year memory reliability simulations), indicating that traditional ECC designs are strong enough to handle failures in MeRAM and STT-RAM with PWSA multi-write design.

### 5.6.3 Latency, Energy, and Area of a 16-MB MRAM Bank

In order to include the energy and latency of ECC designs, we compare STT-RAM and MeRAM in memory-bank level. With inputs of MTJ cell area (see Table 5.2) and bit-line write/read latency/energy (see Table 5.4), we use NVSIM [DXX12] to obtain the area, energy, and latency of a 16-MB STT-RAM bank and a 16-MB MeRAM bank. In-controller ECC commonly exists in current server-class processors for DRAM error detection and correction. In-memory ECC is a new technology, which correct errors individually in memory banks. We only count the power and latency of in-memory ECC in our STT-RAM and MeRAM comparison, because in-memory ECC can correct all MRAM introduced failures, and in-controller ECC is already used for current memory technologies. The in-memory ECC detection and correction latency are about 0.34ns and 4.4ns respectively [DJK15], and encoding latency is assumed to be 0.3ns (i.e., should be little shorter than detection). The energy of encoding, detection, and correction is below 1 pJ per access, which are ignored compared to other memory components. The area overhead of in-memory ECC is about 12.5%.

Table 5.6: Write/read latency/energy for one write/read in a x8 16-MB STT-RAM and MeRAM banks. One write/read operates on 64 bits (72bits in memory banks for in-memory ECC detection and correction) in a row in burst mode.

| Memory | Write | | Read | | Area |
|---|---|---|---|---|---|
| | latency | energy | latency | energy | |
| MeRAM | 9.4 ns | 271.0 pJ | 5.0 ns | 210.3 pJ | 9.5 $mm^2$ |
| STT-RAM | 17.2 ns | 831.7 pJ | 11.9 ns | 293.2 pJ | 17.0 $mm^2$ |

We summarize the area, latency, and energy of one access to 16-MB STT-RAM and MeRAM banks in Table 5.6 (every access is 64 bits in burst mode). Benefited by the smaller size of VC-MTJs, MeRAM has smaller bank area, shorter interconnect, and smaller sized peripheral circuits, which turn into less energy and shorter latency in the logic operations like row decoding and MUX selection. The bank read latency and energy are dominated by these logic operations, thus MeRAM shows faster read speed and less read energy. For write operation, VC-MTJs' small cell size, shorter write pulse, and less

write energy jointly build the advantages of both write energy and latency.

## 5.7 Chapter Conclusion

We comprehensively compare the two promising non-volatile magnetic memory technologies, STT-RAM and MeRAM, in the circuit context with respect to reliability, energy, speed, area, and scalability. MeRAM has higher WER than STT-RAM under process and temperature variation, but by utilizing a multi-write design, both MRAMs are able to achieve an acceptably low WER. With clear advantages of MTJ switching delay and energy, MeRAM outperforms STT-RAM by 83% in write speed, 67.4% in write energy, 138% in read speed, and 28.2% in read energy. In terms of density, VCMA allows to use minimum sized access transistors, which helps MeRAM to achieve twice the density of STT-RAM at 32nm node, and the density advantage is expected to increase at smaller nodes indicating that MeRAM has better density scalability. With respect to challenge of technology scaling down, simply shrinking dimension does not save energy and introduces fabrication defects for both technologies; more effort should be spent on discovering materials with higher polarization efficiency.

# CHAPTER 6

# MTJ Variation Monitor for Adaptive MRAM Write and Read

## 6.1 Chapter Introduction

Both STT-MRAM and MeRAM face the challenge of high write error rate (WER) due to thermal fluctuation. Increasing write current and time reduces the WER of STT-MRAM at the expense of high write power, large access transistors, and long write latency. For MeRAM, there is no straightforward method to reduce WER. STT-MRAM also suffers from read disturbance, where the read MTJ falsely switches due to thermal activation caused by read current. MeRAM is free from this problem because the read current direction is selected to strengthen VC-MTJ's thermal stability rather than weakening it.

Process and temperature variation further exacerbates the problems [LLS08, WZJ12, WLE16b, EJL14]. Local variations induced MTJ diameter and oxide tunnel barrier thickness changes lead to resistance change or MTJ failure [PKJ11]. Compared with local variation, wafer-level variations, including thickness variation of free layer and oxide tunnel barrier layer, more severely affect MTJ performance [TSC99, SCW00]. The wafer-level free layer thickness variation can dramatically change energy barrier in free layer and thermal stability, especially for out-of-plane MTJs. Temperature variation during operation also affects energy barrier, STT and VCMA effect, and MTJ resistance. Temperature and process variation together can change the energy barrier by 200%, indicating that extreme high write energy is required if STT-MRAM is designed for worst process and temperature corner. Similarly, read disturbance rate (RDR) increases with lower energy barrier, meaning that very low read current is needed to design for the worst corner, resulting in low sensing margin. In contrast, MeRAM does not have read disturbance, but requires precise write voltage tuning to achieve low WER, but the required voltage

varies with energy barrier and hence changes with process and temperature variation. The temperature variation also affects the high-to-low resistance ratio of MTJ, which is quantified by tunnel magnetoresistance (TMR, defined as $(R_H - R_L)/R_L$). TMR changes with temperature dramatically [JZK16], e.g., TMR reduces from ∼230% to ∼150% for temperature from 200K to 300K [DST08], indicating that sensing margin also varies with temperature.

In this chapter, we introduce an MTJ-based variation monitor design [WLG16] utilizing thermal activation and VCMA effect. The monitor enables in-situ process and temperature variation sensing. The monitor achieves remarkable area, power, and latency improvement compared with conventional on-chip thermal monitors. We have proposed an adaptive write scheme which selects optimized write pulse for STT-MRAM and MeRAM to achieve faster write speed based on run-time variation sensing. We have also proposed an adaptive read scheme, which smartly selects sensing voltage and reference resistance to improve sensing margin while maintaining low read disturbance rate.

## 6.2 Write Error and Read Disturbance Rates under Variation

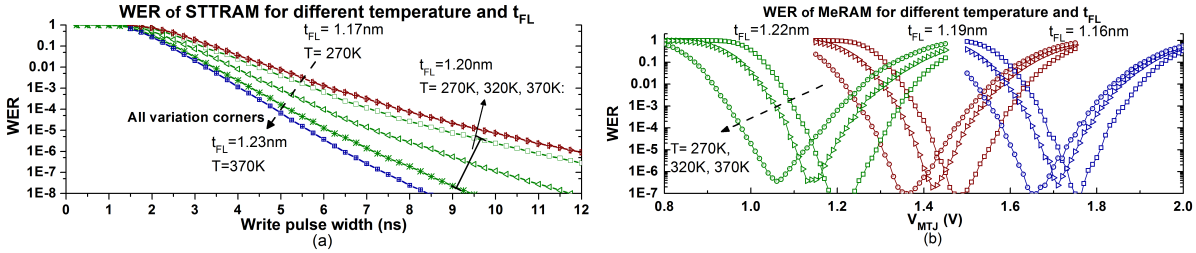



Figure 6.1: (a) The STT-MRAM P-to-AP WER as a function of write pulse width under different $t_{FL}$ and temperature corners. In STT-MRAM, P-to-AP switching is more difficult and dominates write latency. (b) The average AP-to-P and P-to-AP WER of MeRAM as a function of write voltage.

The switching behavior of STT-MRAM and MeRAM are affected by temperature and free layer thickness ($t_{FL}$) [WZJ12, AAY14]. We simulate the WER of STT-MRAM and MeRAM under different $t_{FL}$ and temperature corners using an LLG-based numerical

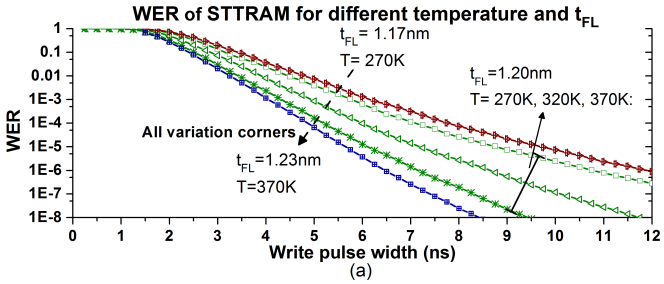



Figure 6.2: The STT-MRAM P-to-AP RDR as a function of write pulse width under different $t_{FL}$ and temperature corners. In STT-MRAM, P-to-AP is selected as the read current direction due to less spin polarization efficiency.

model[1] including temperature dependence, VCMA effect, STT effect, and thermal fluctuation, which has been verified against experimental data in [WLE16b]. The $t_{FL}$ variation are assumed to be within 5% across wafer [SCW00]. The temperature varies from 270K to 370K. Resistance variation (due to MTJ shape change) has limited impact on write behavior (i.e., STT-MTJ has low resistance, and its write current is mainly determined by access transistors, while the high resistance of VC-MTJ drops over 95% supply voltage with negligible variation) and is simply treated as random Gaussian variation in the simulations together with variation of access transistors [WLP13] due to line edge roughness, random doping fluctuation, and non-rectangular gate effect.

The WER of STT-MRAM and MeRAM under different temperature and $t_{FL}$ corners are shown in Fig. 6.1. The variation can shift WER by over 1,000X. The WER of STT-MRAM is mainly affected by temperature, while MeRAM is strongly affected by both $t_{FL}$ and temperature. WER reduction requires to choose appropriate write pulse adaptively for MRAM array according to its temperature and process variation. One conventional solution is exhaustive chip variation test and in-situ temperature monitor [CY11, WMX09, CCP10, APM09] placement in MRAMs.

The RDR of STT-RAM under variation is shown in Fig. 6.2. MeRAM is free from read disturbance because its read uses the reverse direction of write, which strengthens data retention. The variation can shift RDR by over 1,000X. At higher temperature, thermal stability degrades leading to more read disturbance.

[1]Available at http://nanocad.ee.ucla.edu/Main/DownloadForm

## 6.3 MTJ based Variation Monitor

In this section, we propose an MTJ-based variation monitor offering a cheaper solution for in-situ variation monitoring application than exhausted chip testing and expensive conventional thermal monitors. The monitor senses combined temperature and wafer-level $t_{FL}$ variation.

### 6.3.1 Sensing Principle

Monitoring variation through directly measuring WER is expensive, which requires large number of writes and reads. The proposed monitor utilizes thermal activation and VCMA effect to indirectly monitor variation by sensing the thermal activation rate in MTJs under different stress voltage and current.

$$\begin{aligned} t_{R,STT} &= \exp\left(\Delta\left(1 - I_{MTJ}/I_C(\Delta)\right)\right) \\ t_{R,VC} &= \exp\left(\Delta\left(1 - V_{MTJ}/V_C(\Delta)\right)\right) \end{aligned} \tag{6.1}$$

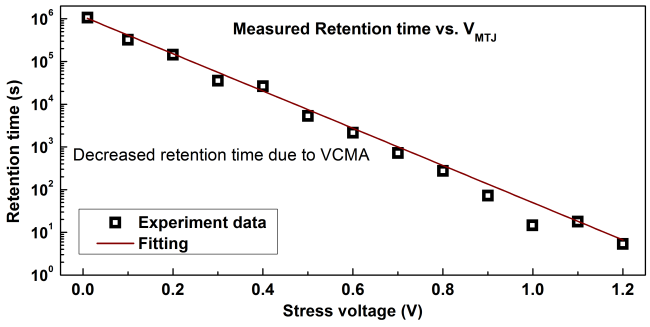



Figure 6.3: The experimentally measured retention time as a function of stress voltage on MTJs.

As described by (6.1) [AUA13, HYO05], the retention time (i.e., the mean of switching time under non-write state) of STT-MTJ ($t_{R,STT}$) and VC-MTJ ($t_{R,VC}$) exponentially depends on thermal stability ($\Delta$, proportional to energy barrier), critical current of STT-MTJs ($I_C(\Delta)$), and critical voltage of VC-MTJs ($V_C(\Delta)$). The write pulse width (determined by ($I_C(\Delta)$ and $\Delta$) and voltage ($V_C(\Delta)$) of STT-MTJs and VC-MTJs also depend on $\Delta$. This indicates that knowing the $t_{R,STT}$ and $t_{R,VC}$ change due to temperature and process variation can predict the MRAM write behavior change. Retention

time of MTJs is too long to be measured directly. Fortunately, as illustrated by the Eqn. (6.1), applying current/voltage on MTJs reduces retention time exponentially giving rise to a possible way of measurement. We utilize this observation in the proposed variation monitor and call such applied voltage/current stress voltage/current for simplicity. This observation is demonstrated in experiment measurement, where retention time decreases exponentially with increasing stress voltage due to VCMA effect in Fig. 6.3.

$$
\begin{aligned}
P_{SW,STT} &= 1 - \exp\left(-t_S/t_{R,STT}\right) \\
P_{SW,VC} &= 1 - 1/2 * \exp\left(-t_S/t_{R,VC}\right)
\end{aligned}
\tag{6.2}
$$

When the retention time reduces to sub-$\mu s$, the MTJ switching rate ($P_{SW}$) due to thermal activation during under stress time ($t_S$ in tens of ns) can be measured as explained in Eqn. (6.2). Then $P_{SW}$ (correlated to $t_{R,STT}$ and $t_{R,VC}$) inherently reflects the ambient variation.

### 6.3.2 Circuit Implementation and Simulation

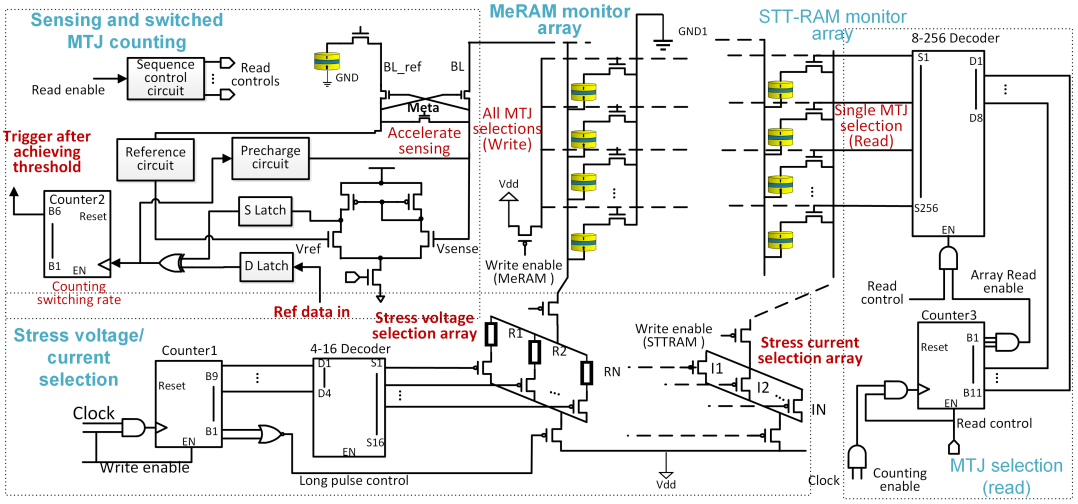



Figure 6.4: The schematic of STT-MRAM and MeRAM based variation monitor. Variation monitoring operations: 1) apply stress voltage/current on MRAM monitor array controlled by stress voltage/current selection circuit; 2) select every MTJ (controlled by MTJ selection circuit) one by one to read and count MTJ switching rate (controlled by sensing and switched MTJ counting circuit).

The principle of the proposed MTJ-based variation monitor is to obtain switching

rate of an MTJ array after a stress operation (applying a stress voltage and current for 20ns). If the switching rate reaches preset threshold after a stress operation, the stress level is output to reflect ambient variation. Otherwise, the monitor continues to try a higher stress level of voltage/current.

The monitor design is shown in Fig. 6.4. In a stress operation, all MTJs in the monitor are in high resistance state initially. The write control circuit applies a stress current (for STT-MRAM) or voltage (for MeRAM) simultaneously on all MTJs in the monitor array for 20ns. The stress current (for 256-MTJ bit-line) ranges from 2.5mA to 10mA, which is precisely controlled by the effective width of transistors in the stress current selection array, where the stress current variation is close to 0 due to the large transistor width guaranteeing monitor accuracy. The stress voltage on VC-MTJs is adjusted by dividing voltage on bit-lines and resistors (vary from 200Ω to 700Ω) in the stress voltage selection array. The stress voltage variation is also close to 0 because the equivalent parallel resistance of all VC-MTJs on a bit-line averages out individual MTJ resistance variation.

After a stress operation, the read control circuit selects each MTJ one by one and reads its state. In the read, the bit-line ($BL$) and reference bit-line ($BL\_ref$) are pre-charged and pulled down by the read MTJ and reference resistor separately. The difference between $Vsense$ and $Vref$ creates an output to S Latch, and a switched MTJ rises S's output from 1 to 0, then the XOR of S Latch and D Latch (output is constantly 1) creates a rise edge, which is counted by Counter2. At last a switched MTJ is reset by a write pulse for future stress operations.

We simulate the monitor design using a 65nm commercial library. The stress pulses are shown in Fig. 6.5 (a). Stress current has $< 0.3\%$ and $< 4.7\%$ variation due to temperature ($27^oC$ to $100^oC$) and oxide thickness variation (9% resistance change) respectively, while stress voltage has $< 1\%$ and $< 2\%$ variation accordingly. In addition, switched MTJs (e.g., 30%) during stress time can cause up to 10% and 2% stress current and voltage change respectively. The low variation demonstrates the proposed monitor accuracy.

Fig. 6.5 (b) shows the simulated waveforms of read, counting, and reset operations. The first and third reads are performed on switched MTJs, where write pulses follow reads to reset MTJs, and the counter increases. The second read is on a non-switched MTJ, and hence no action is taken after the read. If the counted number reaches preset

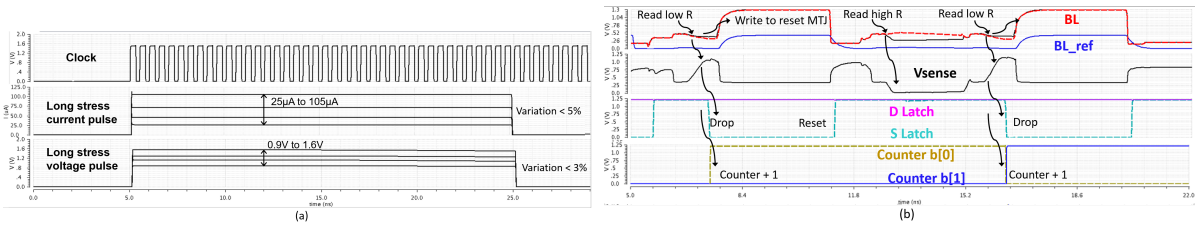



Figure 6.5: (a) Different stress current/voltage in the proposed monitor. (b) Simulated waveforms of read, reset and counting operations.

threshold (e.g., 64 out of 256 MTJs), it sends out a completion signal and outputs the current stress level, which presents the ambient variation level. If the preset threshold is not reached after reading all MTJs, the counter is reset, and a higher stress level is selected in the next variation sensing cycle.

We simulate the switching rate and standard deviation ($\sigma$) of a 256-MTJ variation monitor with different stress levels and variation corners as shown in Fig. 6.6. In these curves, if we select a preset threshold between 10% to 30%, the voltages to reach the threshold under different variation levels ($10^oC$ temperature difference between two consequent curves) can be well differentiated, e.g., the dotted curves show the standard deviation (accuracy of the monitor) is much smaller than curve gaps. Therefore, for a given constant $t_{FL}$, ten stress levels can achieve accuracy of $10^oC$.

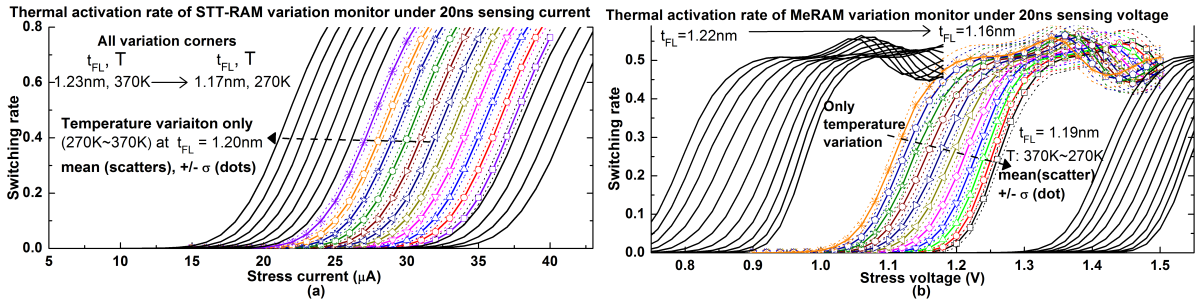



Figure 6.6: Switching rate of (a) STT-MTJ- and (b) VC-MTJ-based variation monitor under different stress current and voltage respectively. The color lines are switching rate for only temperature variation ($10^oC$ interval). The dot lines outline standard deviations ($\sigma$) of thermal activation rate ($\sigma$ is caused by process variation and random thermal activation).

Table 6.1: Comparison between conventional thermal monitors and the proposed variation monitor. The proposed monitor uses 256 MTJs and 10 stress levels

| Monitor | Latency | Accuracy | Energy | Area |
|---|---|---|---|---|
| S1 [CY11] | 0.1ms | $9^oC$ | $0.015\mu J$ | $0.01mm^2$ |
| S2 [WMX09] | 0.2ms | $3^oC$ | $0.24\mu J$ | $0.04mm^2$ |
| S3 [CCP10] | 1ms | $2^oC$ | $0.49\mu J$ | $0.01mm^2$ |
| S4 [APM09] | 100ms | $0.1^oC$ | $13.8\mu J$ | $0.04mm^2$ |
| this(STT) | 1-10$\mu s$ | $10^oC$ | 0.12-1.2$nJ$ | $0.0005mm^2$ |
| this(Me) | 1-10$\mu s$ | $10^oC$ | 0.27-2.7$nJ$ | $0.0005mm^2$ |

Table 6.1 shows the comparison between the proposed variation monitor with conventional thermal monitors. The conventional monitors target on high precision, where long latency and high energy are consumed by analog-to-digital blocks and bipolar sensing transistors. The proposed monitor has less accuracy but faster speed, lower energy/sample, and smaller area. Its accuracy can be improved by using more MTJs to reduce $\sigma$ of curves in Fig. 6.6 as well as using finer grids of stress levels in the monitor, which quadratically increases sensing energy and latency. In addition, finer grids of stress current/voltage require less process variation in circuit, which is also the accuracy limitation. Fortunately, selecting optimal write pulse for STT-MRAM and MeRAM does not require high accuracy (i.e., Section 6.4.1 shows that three stress levels are enough) indicating that the proposed monitor is well suited to the adaptive write selection with the least overhead. The area of the monitor is dominated by the 8-256 decoder (97.1% of total transistors). The area of 8-256 decoder was estimated through synthesize, place and route using commercial 65nm library.

Though the wafer-level resistance variation of STT-MRAM is not considered in the simulation, but it can also be partially monitored because the stress voltage/current shift induced by resistance variation is proportional to write voltage/current shift.

## 6.4 Adaptive Write

### 6.4.1 Adaptive Write Scheme

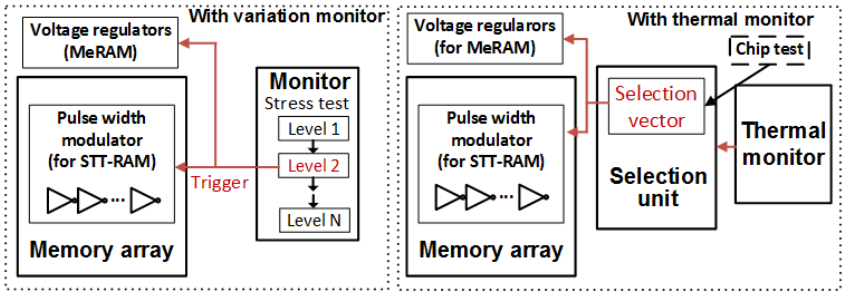



Figure 6.7: Adaptive write scheme using the MTJ-based variation monitor or conventional thermal monitors.

The adaptive write scheme is to dynamically select an optimized pulse width (voltage) for STT-MRAM (MeRAM) out of multiple voltage (current) choices to minimize write latency according to ambient variation. Creating multiple pulse widths uses simple delay circuits, which is shared by multiple bit-lines with negligible overhead. Multiple write pulse voltage requires multiple voltage regulators, and the regulators can be shared by the entire MRAM array. Temperature variation over MRAM array [EJL14] can be captured by placing multiple proposed monitors to monitor local variation. One such monitor only uses one bit-line in MRAM boundary with an area overhead of <0.005% (i.e., adding monitor control circuits in MRAM boundary does not affect MRAM fabrication regularity). The monitor also consumes negligible power (i.e., 2.7nW for one variation sample per second) compared with power of MRAM array (>10 mW).

Schemes to make optimized write pulse selections with and without the proposed variation monitor are shown in Fig. 6.7. With the variation monitor, write pulse is selected according to output variation level. Without the variation monitor, exhaustively memory chip test is required for each chip to obtain and store optimized pulses for different temperature, and a conventional thermal monitor is required to make dynamic pulse selection

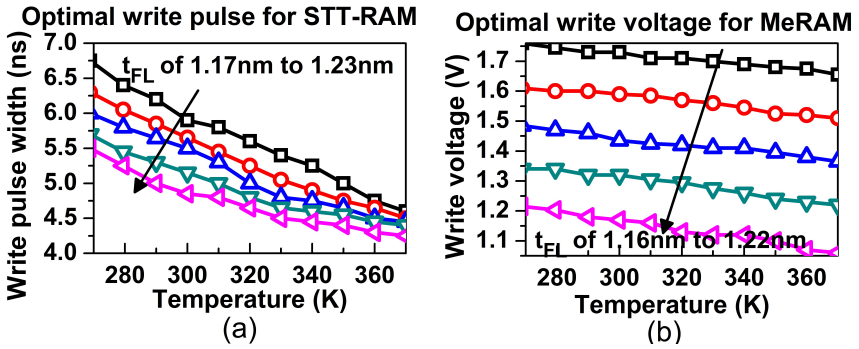



Figure 6.8: Optimal write pulses for (a) STT-MRAM and (b) MeRAM under different $t_{FL}$ and temperature corners.

### 6.4.2 Adaptive Write using Variation Monitor

In this section, we evaluate the write scheme with the proposed variation monitor. The write circuit for MRAM is implemented with read check function [LAD15] which performs a read check following a write (the writing data is pre-stored in D Latch in Fig. 6.4), and a write error gives rise to additional writes until all errors are fixed. With this, WER of 0 is guaranteed for MeRAM and STT-MRAM irrespective of the single write pulse voltage/width. For STT-MRAM, shortening single write pulse reduces latency and energy, as a trade-off, WER of the write and chance of additional writes increase, which add overall latency and energy. Hence, there is an optimal single write pulse achieving minimum expected latency, and it can be found given a WER function of pulse voltage/width. Such optimal pulse can reduce STT-MRAM's expected latency and energy by over 60% compared with conventional write circuit [WLE16b]. The optimal pulse width (voltage) for minimum expected latency (including initial write, read checks, and additional writes) of STT-MRAM (MeRAM) are shown in Fig. 6.8. The pulse width for STT-MRAM spans from 4.25ns to 6.75ns mainly affected by temperature. The voltage range for MeRAM is from 1.05V to 1.75V affected by both temperature and $t_{FL}$.

In the following evaluation, the combined temperature and $t_{FL}$ corners are divided into groups based on the variation monitor's output (stress levels reaching $P_{SW}$ threshold). Each group has an optimized write pulse minimizing the maximum write latency in the group. More write pulse choices (equal to stress levels) result in shorter write latency.

Our evaluation flow is illustrated in Fig. 6.9 (a). We simulate the peripheral circuit

(see Fig. 6.4) with a bit-line size of 256 MTJs using 32nm commercial library and simulate the WER of MTJs with LLG-based numerical model. The bit-line-level write latency varies from 5.5ns to 7.5ns for STT-MRAM and 4 to 10.1ns for MeRAM for all variation corners and number for write pulses (1 to 5). With the inputs of bit-line results, we use NVSIM [DXX12] to obtain latency and energy of MRAM array (cache). In Fig. 6.10, the write latency of L2 Cache with different $t_{FL}$ corners is shown to decrease with increased number of pulse choices, and each point is the maximum or average latency of temperature corners of 270K to 370K. MeRAM's write latency reduction is up to 59%. There is a latency increase for $t_{FL}$ of 1.19nm using from one to two voltage choices, because that 1.19nm $t_{FL}$ corner is closer to optimized voltage when only one write voltage is used (see Fig. 6.1b). The write latency of STT-MRAM is improved by up to 17%. The maximum latency for $t_{FL}$ corner of 1.17nm is not seen improvement because the corner with 1.17nm $t_F$ and 270K is always the worst corner to be optimized in its variation corner group no matter how many choices is adapted. As seen, three choices are efficient enoughfor write latency improment.

We modified gem5 [BBB11] to simulate two cases: 1) an x86 processor with one core and one single-level 8-MB MRAM data cache; 2) an x86 processor with two cores, two 1-Mb MRAM L2, and one 16-MB MRAM L3 caches (L1 uses default SRAM). We modified McPAT [LAS09] to simulate processor power and used Hotspot [HGV06] to simulate

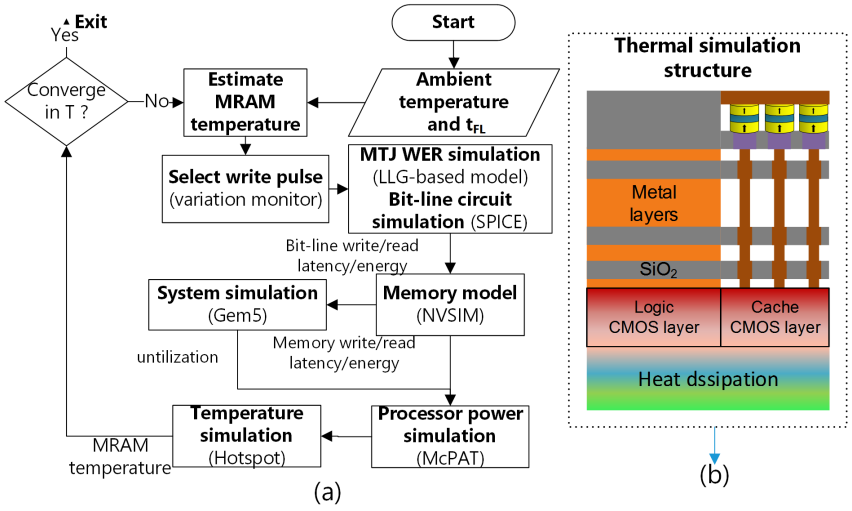



Figure 6.9: (a) Evaluation flow of adaptive write in MRAM based system. (b) The cross-section structure for thermal simulations.

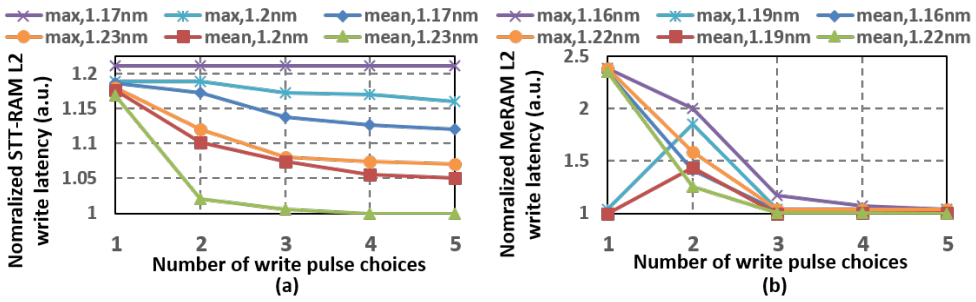



Figure 6.10: The maximum and average write latency in (a) 1MB STT-MRAM L2 and (b) MeRAM L2 from 270K to 370K under different $t_{FL}$ corners with different number of write pulse choices.

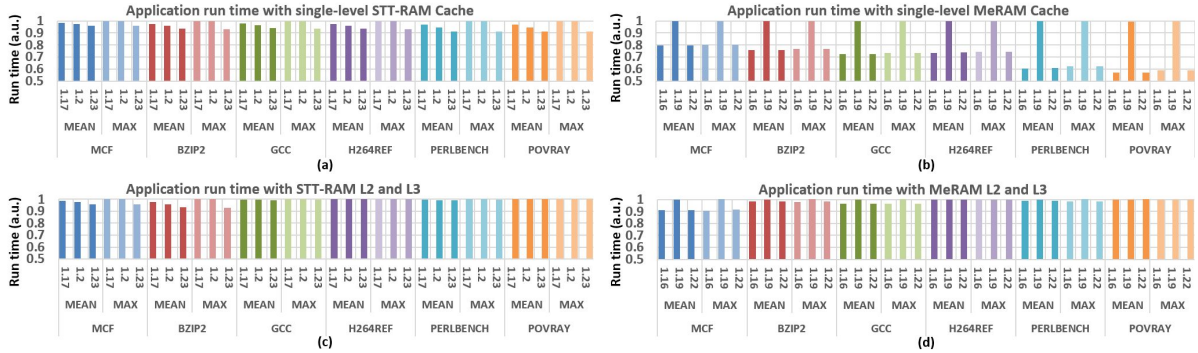



Figure 6.11: The average/maximum run time of SPEC benchmarks using adaptive write (with three write pulse choices) for (a) one-core processor with single-level 8-MB STT-MRAM cache and (b) single-level 8-MB MeRAM MeRAM cache, a dual-core processor with (c) 1-MB STT-MRAM L2 and 16-MB STTRAM L3, and (d) 1-MB MeRAM L2 and 16-MB MeRAM L3 over temperature corners (270K to 370K). Run time is normalized to the maximum run time for processors without adaptive write (one write pulse choice) for each benchmark.

MRAM temperature with the structure shown in Fig. 6.9b.

We simulated one billion instructions of SPEC benchmarks using our evaluation flow. The application run time reduction with adaptive write are shown in Fig. 6.11. The processors with single-level MRAM see noticeable application speedup after using adaptive write, where up to 41% and 9% run time reduction are shown for MeRAM and STT-MRAM respectively. However, the improvement are much less for processors with MRAM L2 and L3 (up to 10% and 2% for MeRAM and STT-MRAM respectively), be-

cause cache write latency improvement is hidden by SRAM L1. This indicates that the adaptive write scheme may be more efficient for embedded applications with single-level MRAM cache. Compared with MeRAM, STT-MRAM write latency improvement is not significant. Actually, the write energy is more crucial issue for high-speed STT-MRAM cache (e.g., write latency within 3 ns), where large write current is required and sensitive to variation. Our future work will evaluate the adaptive write scheme in STT-MRAM energy reduction.

## 6.5 Adaptive Read

In a reliable STT-RAM sensing design, read disturbance rate and sensing error rate should be minimized within error-correcting-code (ECC) [WHZ16] capability. However, there is a tradeoff exposing to them. To increase sensing margin, a high sensing current is required, which adds to more read disturbance. Moreover, these errors are severely affected by process and temperature variation, hence designing for the worst case leads to both a small sensing margin and high read disturbance rate. We propose adaptive read to dynamically increase sensing margin while maintaining read disturbance under control according to temperature and process variation.

Read disturbance rate depends on STT-MTJ thermal stability, which are affected by sensing current, free layer thickness and temperature. The resistance of STT-MTJ also changes with temperature, especially for the the AP resistance as illustrated in Fig. 6.13. The STT-MTJ resistance change can be fitted to a temperature-dependence model as in [DST08]. In the proposed work, we design a STT-MTJ sensing scheme with two reference resistance selections and two read voltage selections. The read voltage is selected by the proposed variation monitor, which detects STT-MTJ energy barrier (thermal stability). STT-MTJs with lower thermal stability suffer from more read disturbance, requiring low read voltage. The reference resistance is selected by a temperature monitor, where the high reference resistance is used at low temperature because of the increased STT-MTJ resistances, while the low reference is used at high temperature.

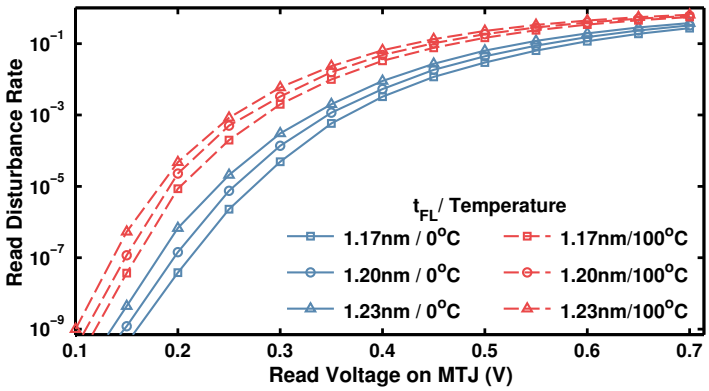



Figure 6.12: Read disturbance rate as a function of voltage drop on P MTJ for a set of temperature and free layer thickness variation corners. The read disturbance rate is obtained from Monte-Carlo simulation with sensing time of 3ns.

### 6.5.1 Adaptive Sensing Circuit using Multiple Reference Resistance

For STT-RAM, the dependence of read disturbance rate on read voltage, process and temperature variation is plotted in Fig. 6.12. As expected, read disturbance rate increases with read voltage for that sensing current lowers thermal stability. Similarly, higher temperature and lower free layer thickness also leads to lower thermal stability and hence higher read disturbance rate. Fortunately, the variation-dependence can be monitored by the proposed monitor (Section 6.3) as for adaptive write in Section 6.4. By contrast, instead of using three stress level for adaptive write, only one threshold stress current level is needed for adaptive read. In the adaptive read scheme, the default read voltage is high to increase sensing current margin. But for the STT-RAM with low thermal stability due to the process variation, the high read voltage may cause more read disturbance than ECC capability (e.g., $10^{-9}$). For this case, the proposed variation monitor is able to detect it with the selected stress current level (see Section. 6.3.2). It controls the read voltage to low to reduce read disturbance rate.

To further improve the sensing margin, adaptive reference resistance is proposed. As is seen in Fig. 6.13, the AP MTJ resistance changes dramatically with temperature, while the P resistance is more stable, resulting in lower TMR at high temperature. Hence, to maximize sensing margin, a low and a high reference resistors can be used at high and low temperature respectively. The resistor selection is controlled by an on-chip temperature monitor. To quantify the temperature dependence of MTJ resistance, we used the model

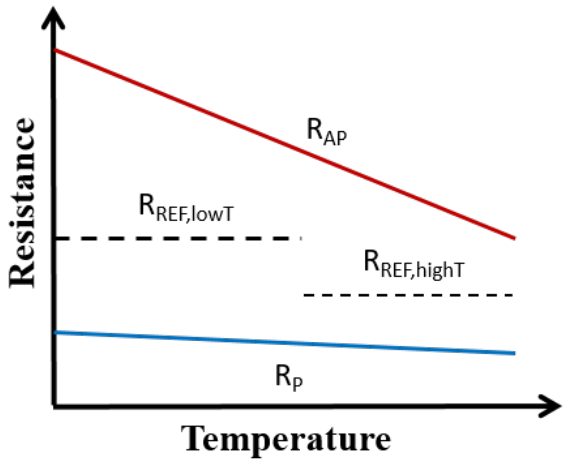



Figure 6.13: Illustration for using two reference resistance for low and high temperature sensing.

[DST08] fitted to experimental data as listed in Equation 6.3 below.

$$R_\gamma(T) = R_\gamma(0)\frac{\sin(CT)}{CT}\left[1 + Q\beta_\gamma \ln\left(\frac{k_B T}{E}\right)\right]^{-1} \tag{6.3}$$

The temperature-dependence is approximately linear. One experiment shows that TMR can change from 192% at 4.2K to 90% at room temperature [IMK06]. The TMR can be further reduced at high temperature, e.g., over 100$^o$C. With the model, a threshold temperature is selected such that the high or low reference resistor is chosen when temperature is below or above the threshold temperature as illustrated in Fig. 6.13.

The proposed sensing circuit is shown in Fig. 6.14. High and low read voltages at the top are selected by "Low $\Delta$" signal output from the proposed variation monitor. Reference resistors at the right are selected by "Low temperature" output from an on-chip temperature monitor. We then design the read voltage and reference resistors with the assumption that MTJ AP resistance changes from 5k (low temperature) to 3.33k (high temperature) and P resistance changes from 2k (low temperature) and 1.8L (high temperature). To minimize the read disturbance of STT-RAM, the sensing current direction is chosen as the direction of P-to-AP switching for the reason that P-to-AP switching is more difficult than AP-to-P switching. Therefore, only P state MTJ is possibly disturbed during sensing. Considering the read disturbance rate $< 10^{-9}$, we use 0.66V for low read voltage and 0.78V for high read voltage, which gives 100 mV and 150 mV voltage drop

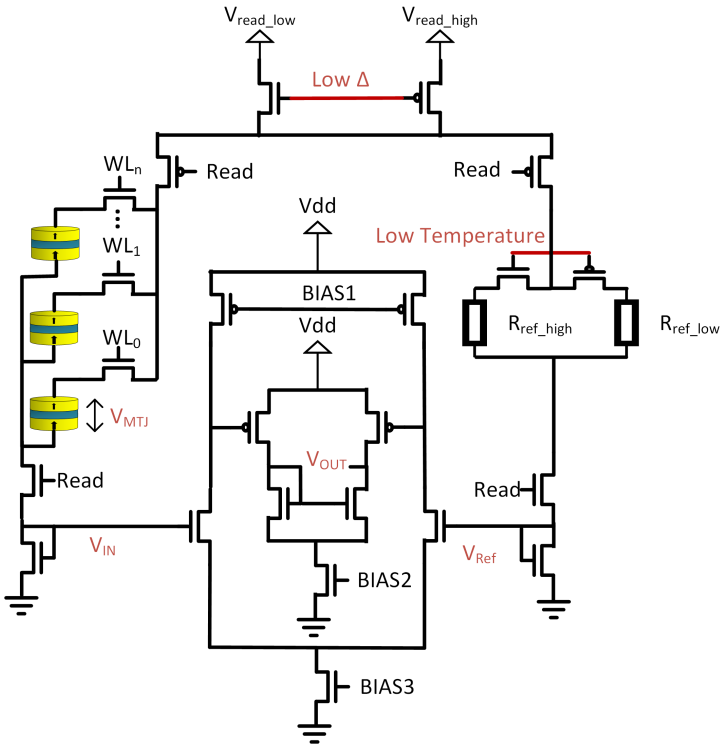
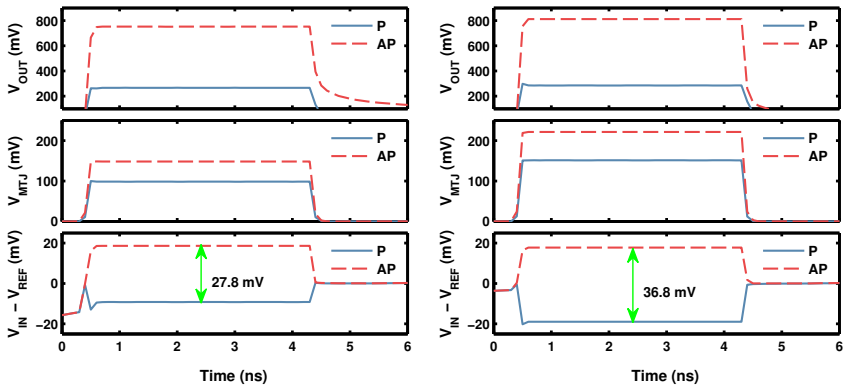



Figure 6.14: Sensing circuit used in the adaptive read. The switch of two reference resistors are controlled by temperature sensor. The switch of read voltage is controlled by the proposed monitor with a output

across P MTJ respectively. Worst 10% resistance variation is assumed, meaning that the reference resistance is designed to sense 90% $R_{AP}$ and 110% $R_P$. To maximize sensing margin, 3.25kΩ and 2.85kΩ are chosen as the high and low reference resistances.

We perform SPICE simulation of the proposed adaptive read design at low temperature, hence the high reference resistance is selected. For normal case, the STT-RAM has high thermal stability so that high read voltage is selected. As a comparison, the conventional read design has only one read voltage and one low reference resistance, which are designed according to the worst variation corner (high temperature and low thermal stability). The sensing waveforms are shown in Fig. 6.15. The sensing noise margin ($V_{in}$ - $V_{ref}$) is significantly increased from 26.8 mV to 37.8 mV using adaptive read.




(a) (b)

Figure 6.15: (a) Conventional STT-RAM sensing simulation with single read supply voltage and one reference resistance. (b) The proposed STT-RAM sensing simulation with two read voltage and two reference resistance.

## 6.6 Chapter Conclusion

We have designed an MTJ-based variation monitor to sense process and temperature variation. At the same accuracy, the variation monitor achieves 20X smaller area, 10X faster speed, and 5X less energy. We have proposed an adaptive write scheme to minimize the write latency of STT-MRAM and MeRAM according to ambient process and temperature variation. The write latency of STT-MRAM and MeRAM cache is reduced up to 17% and 59% respectively, while simulated application run time has 1.7X improvement. We have also proposed adaptive read technique, which can increase the sensing margin by 1.3X.

# CHAPTER 7

# Negative Differential Resistance-Assisted Resistive NVM for Efficient Read and Write Operations

## 7.1 Chapter Introduction

Resistive NVM including conducting-bridge and ReRAM [YW11, SSS08], STT-RAM [TSC99, ZZW12], and PCM [WRK10], are widely considered for both storage class as well as main memory because of their non-volatility, fast speed, and high density. However, these emerging NVM still have challenges to resolve before the adaption into conventional computing system.

STT-RAM faces several key problems: 1) high write currents (up to 100 µA at the 45nm node [ITR11]), 2) low sensing margin [LAS08], which forces trade-offs between read error rate with read time and energy, and 3) susceptibility to read disturbance [KZZ15, KLK14, KZK15], *i.e.*, MTJ false switching during sensing, which unfortunately increases with sensing margin. These limitations are intrinsically due to the low TMR of STT-MTJs. To make matters worse, write time may be further extended by process and temperature variations [WLE16b, LAS08]. Similarly, read errors and read disturbance under variation create more severe reliability concerns in STT-RAM [KZZ15, CAD10, KLK14, LAS08].

For multi-level-cell (MLC) ReRAM and PCM, the programming relies on program-verify (P$V) cycles [PMH15, BFR09], leading to high energy dissipation and long write latency.

In this chapter, we offer an unified solution to simultaneously resolve all of these challenges by utilizing negative differential resistance (NDR) devices within memory read and write peripheral circuitry. Our NDR solutions include two types of devices: voltage-controlled NDR (V-NDR) and current-controlled NDR (C-NDR) devices. V-NDR can be

naturally implemented by TFETs [LLZ12] and tunnel diodes (TDs) [EDS01]. In addition, we also propose a CMOS V-NDR design to improve compatibility with conventional Silicon substrate. For C-NDR, we also provide CMOS implementation. In MRAM, our V-NDR solutions limit redundant write current and allow the sensing current ratio of the high and low MTJ states to be amplified up to the peak-to-valley ratio (PVR) of V-NDR, which is much larger than MTJ resistance ratio. For MLC ReRAM. by utilizing V-NDR and C-NDR, intermediate resistance programming can be simplified from over 10 cycles using P&V to one cycle. Our results show that the proposed designs greatly reduce write energy and read disturbance and increase the sensing margin while simplifying sensing circuitry, enabling truly low power resistive NVM technologies.

## 7.2 Issues of MRAM Write and Read

### 7.2.1 Wasted Power During Write Cycles

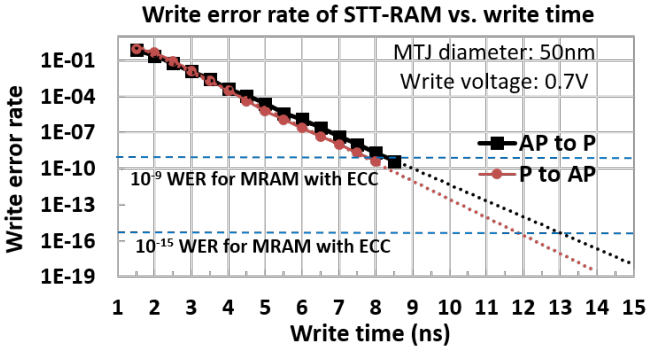



Figure 7.1: STT-RAM write error rate as a function of write time assuming 0.7 V write voltage extracted from 10 billion Monte Carlo circuit simulations using the methodology of Section 7.6.4. The simulated write circuit includes a MTJ, an access transistor, and the capacitance load of 256 1T1M bit-line.

MTJ switching is a stochastic process whose probability increases with the duration of the write current pulse. The write error rate (WER) as a function of write time is shown in Fig. 7.1. Generally, switching from 1 to 0 ($P$ to $AP$) requires a larger write current than 0 to 1 [HYY05]. In our simulation, the switching efficiency ratio for 1-to-0 and 0-to-1 is 0.75 to 1. However, this asymmetry is offset by the different resistances of the two states; hence when the same voltage is used for both switching directions, similar

write times (time to achieve a required WER) are observed in Fig. 7.1. For memories without error-correcting code (ECC), the WER should be below $10^{-18}$, which needs 15 ns write time. In this chapter, we consider designs with ECC, which still require a WER of $10^{-9}$ or better, necessitating at least 9 ns long write pulse [AKW13]. While long write times are necessary to maintain accuracy, over 90% of switching events (i.e., WER $< 0.1$) are completed within the first 3 ns indicating a dramatic waste of energy. In particular, write-1 consumes 1.4X the energy of write-0 because the MTJ stays longer in the low resistance 1 state, leading to higher leakage.

### 7.2.2 Read Margin and Read Disturbance Limits

The low TMR of STT-MTJs limits MRAM read margins and read disturbance, causing reliability problems. In particular it is difficult to simultaneously improve both read margin and read disturbance rate in conventional designs. This is because higher margins require higher read voltages, which increases the read current and can lead to unwanted MTJ switching even if the current is smaller than the critical switching current. For example, to obtain a read margin of 150 mV in a conventional design with 100 fF bit-line load, a WER of over $10^{-8}$ is resulted, exceeding the ECC error tolerance (see Fig. 7.17). Non-uniformity in device characteristics due to process variations can worsen this problem.

## 7.3 NDR Device Characteristics

### 7.3.1 Tunneling-Based V-NDR Devices

To address the challenges faced by MRAM and ReRAM, we introduce V-NDR devices into the read and write circuitry. As illustrated in Fig. 7.2a, V-NDR devices have the property that within a certain bias range (between $V_{peak}$ and $V_{valley}$), the absolute current decreases with increased absolute voltage. The ratio of the maximum and minimum currents ($I_{peak}$ and $I_{valley}$, respectively) within this range is known as the PVR. A variety of two- and three-terminal devices utilizing quantum tunneling such as Esaki diodes, resonant tunneling diodes (RTDs), and reverse-biased TFETs can be used to

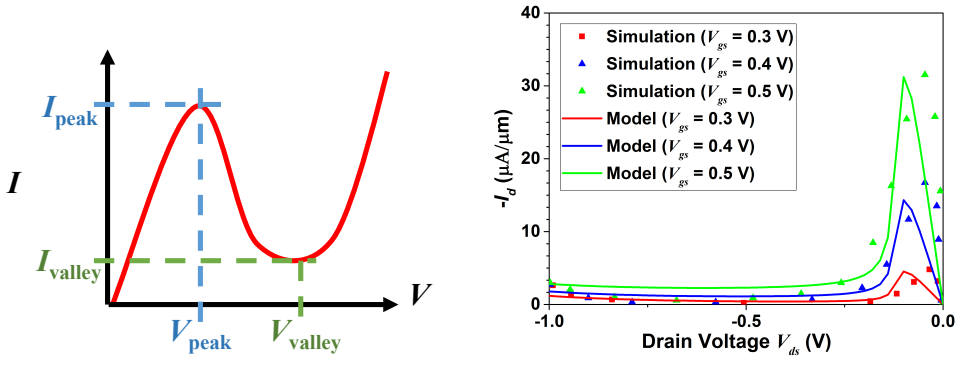
generate this effect. While TDs are relatively mature devices developed specifically to implement V-NDR for various applications, TFETs have the advantage that the PVR can be tuned by gate voltage. In Table 7.1 we summarize the experimental characteristics of some representative TDs and TFETs. RTDs with good performance have already been demonstrated on Si/SiGe [EDS01], which is already a CMOS-compatible platform [Ant]. Many of other best-performing V-NDR devices thus far are based on III-V materials. Non-commercialized technologies like heteroepitaxy [BSS08] and nano-transfer [LWP15b] can integrate III-V MOSFET and FinFET with Si CMOS at the expense of cost increase. The integration of III-V on silicon is already a high priority in industry (for instance, by using III-V MOSFETs or TFETs to supplement or replace silicon transistors in logic), and commercial advances in that direction will also ease integration of V-NDR devices.

Table 7.1: Experimental characteristics of selected V-NDR tunneling devices from literature. Peak current is expressed in terms of per unit width for TFETs and per unit area for TDs.

| **Device** | **Material** | **Substrate** | **Peak Current** | **PVR** |
|---|---|---|---|---|
| TD [EDS01] | Si/SiGe | Si | $50\,\mu A/\mu m^2$ | 6 |
| TD [RPT08] | InGaAs | Si | $2.5\,\mu A/\mu m^2$ | 56 |
| TD [TSL94] | InGaAs/InAlAs | InP | $2\,\mu A/\mu m^2$ | 144 |
| TFET [ZKK13] | InAs/AlSb/GaSb | GaSb | $\leq 230\,\mu A/\mu m$ | $\leq 5.5$ |
| TFET [ZVD14] | InGaAs/InAs | InP | $\leq 4\,\mu A/\mu m$ | $\leq 6.2$ |

### 7.3.2 CMOS Circuit for V-NDR Generation

For integration into commercial memory technology, silicon MOSFETs are preferred over III-V tunnel devices [WPG17, WPC17]. We therefore propose generating V-NDR using three NMOS transistors in the 3T CMOS circuit depicted in Fig. 7.3a. This circuit has three terminals (IN, OUT, and BIAS); T1 and T2 are used to control the gate bias of T3, and the V-NDR current is mainly driven by T3. The gate of T1 is connected to its drain to maintain low leakage, enabling low valley currents, which is the NDR state after write termination. For small $V_{in}$ applied between IN and OUT, the current increases due to the increased voltage drop across T3. However, $V_{int}$ decreases with higher $V_{in}$, such




(a) (b)

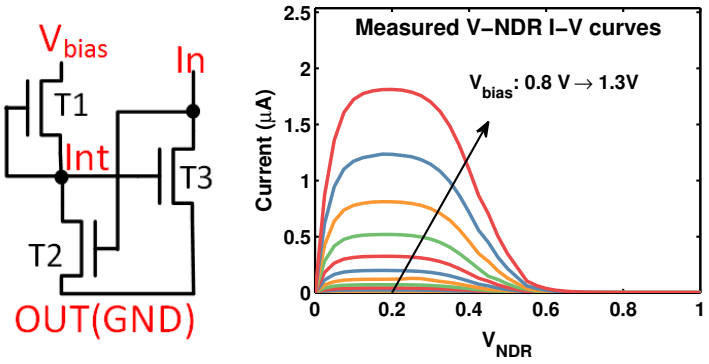
Figure 7.2: a) Schematic $I-V$ for typical V-NDR device. b) $I_d - V_d$ of analytical TFET model and simulated device data of [LLZ12]. For the TFET model, parameters are $A_{TFET}$ = 1.3E-8 A/um, $B_{TFET}$ = 4E6 eV/cm, $E_g$ = 0.74 eV, $\lambda$ = 6E-7 cm, $A$ = -0.02, $B$ = 0.0456, $C$ = 0.04, $n$ = 0.3, and $D$ = 0.0025.

that when $V_{in}$ reaches a certain value $V_{peak}$ (the peak voltage), the output current attains $I_{peak}$ and the T2 transistor turns on and shuts off T3, leading to reduced output current with further increase in $V_{in}$ (the "valley region"). The peak current and PVR can be selected by tuning the $V_{bias}$ applied to T1 to change the T3 operation region. This circuit is therefore capable of low $V_{peak}$ and highly tunable $I_{peak}$ and PVR.

We have demonstrated this structure experimentally by wire bonding NMOS transistors on a single die [WPG17]. As shown in Fig. 7.3b, the peak current can be tuned from the nA to $\mu$A range by changing bias voltage ($V_{bias}$) between 0.8 and 1.3 V while maintaining a peak voltage of 0.25 V and achieving PVR up to 1,000 and greater. Note that the sensitivity of peak current to $V_{bias}$ is determined by the sizes and threshold voltages of the constituent transistors. In our experiment, the sensitivity is relatively high because we used a die whose transistor characteristics had not been specifically optimized for this application. In real implementation, the devices can be designed to limit the $V_{bias}$ tunability and hence reduce possible variability effects.

### 7.3.3 CMOS Circuit for C-NDR Generation

We have also devised a complementary design to V-NDR to mimic C-NDR characteristics. C-NDR is designed using a Schmitt trigger and a transistor as shown in Figure 7.4a. Its




(a) (b)

Figure 7.3: (a) Diagram of proposed 3T CMOS circuit for generating V-NDR between IN and OUT terminals. (b)Experimental $I$-$V$ curves for CMOS V-NDR circuit; current is measured at IN terminal as a function of voltage drop $V_{NDR}$ across the IN and OUT terminals. Different curves correspond to application of $V_{bias}$ between 0.8 and 1.3V. The circuit is constructed using preexisting long channel NMOS devices with gate lengths of 10 $\mu$m, W/L ratios from 4 to 10, and off-currents around 0.1-1 nA and on-currents ranging between 10-100 $\mu$A.

I-V curve is illustrated in Figure 7.4b. The Schmitt trigger has a high threshold voltage and a low threshold voltage, when its input voltage (in) is above the high threshold voltage, the C-NDR turns on, while the C-NDR turns off with input voltage dropping below the low threshold voltage.

## 7.4 Behavior of Series-Connected MTJ and V-NDR Devices

The unique characteristics of V-NDR devices enable them to enhance MRAM performance when integrated into the read and write circuitry. The reasons and conditions for these improvements can be understood by examining the behavior of series-connected MTJ and V-NDR devices, as shown in Fig. 7.5a. For simplicity the MTJ is treated as a resistor whose current is $(V_{cc} - V_{NDR})/R_{MTJ}$, where $R_{MTJ} = R_H$ (0 state) or $R_L$ (1 state) depending on its state. In Fig. 7.5b, we plot the current through the V-NDR device (solid black line) and the MTJ (blue and red lines) as functions of the voltage drop across the V-NDR device $V_{NDR}$. The intersections of the V-NDR and MTJ curves represent the possible steady-state solutions of the circuit: there are three solutions in

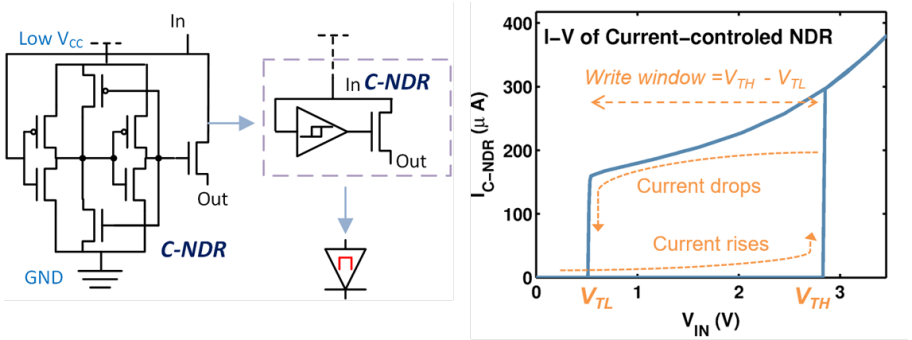



Figure 7.4: (a) One example of C-NDR comprising of a Schmitt trigger and an NMOS. The Schmitt trigger has two threshold voltages $V_{TH}$ and $V_{TL}$. (b) The I-V of the C-NDR. When VNDR starts from 0, then the current of C-NDR remains low until VNDR reaches $V_{TH}$, the current suddenly rises until VNDR reduces below $V_{TL}$. The Schmitt trigger is supplied with low $V_{CC}$ to reduce leakage.

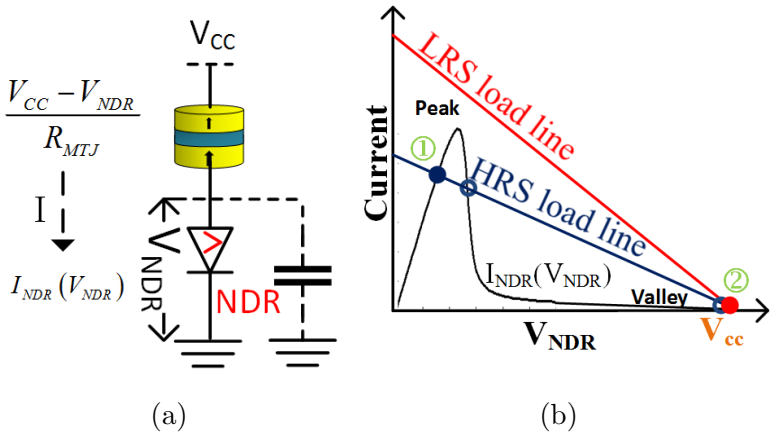



Figure 7.5: (a) Series connection of MTJ and V-NDR device. Note that each V-NDR device is shared by multiple bit-lines in the proposed design. (b) Load line for V-NDR-MTJ series circuit (illustrated in inset). Blue line corresponds to the MTJ HRS and red line to the MTJ LRS. The stable operating points when the MTJ is in HRS and LRS are indicated by ① and ②, respectively.

the 0 state but only one solution in the 1 state. The target operating conditions are that 1) $|-R| < R_L$, where $-R$ is the effective resistance of V-NDR between $V_{peak}$ and $V_{valley}$, and 2) the $I_{peak}$ of V-NDR is greater than the current through $R_H$ but below that of $R_L$ at $V_{Peak}$. Under these conditions, the circuit current in the 1 state is limited by the (minimal) V-NDR valley current.

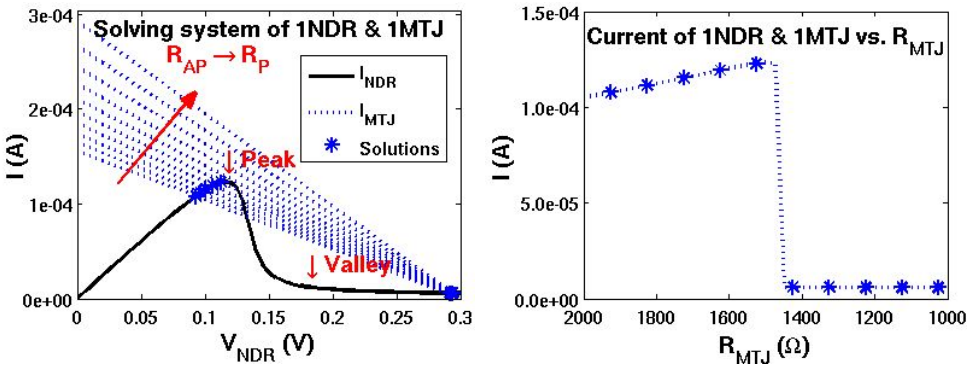
When the externally applied voltage across the circuit increases from 0 to $V_{dd}$, if the MTJ is in its (AP) high resistance state (HRS), the current and voltage drop across the V-NDR stays in the peak region at point ① in Fig. 7.5b. Likewise, if the MTJ is in the (P) low resistance state (LRS), the circuit will stabilize at ② in the V-NDR valley region. This indicates that current flows freely when the MTJ is in the AP state but is blocked when the MTJ is in the P state. If the MTJ switches from AP to P, the current through the circuit drops from ① to ②.

The idea that negative resistance can differentiate between MRAM states has previously been used in a few proposals for read voltage margin improvement or write energy saving [HHS10, UYY14, UYY15]. However, those proposals design the NDR-MTJ circuit for only one solution in both the 0 and 1 states. This requires that the V-NDR curve only intersect the $R_H$ load line once, forcing both MTJ states to have comparatively high current and making the design vulnerable to V-NDR or MTJ device variations. Such nominal designs can improve the voltage margin but not necessarily the current margin between the MTJ states, whereas our concept amplifies both the voltage and current differences since the 1 state is forced into the much lower current V-NDR valley region. Therefore, such proposals do not offer all the operational advantages of our concept, as further discussed below.

In our design, the system should reside in the point ① (below the V-NDR peak voltage) when the MTJ is in the 0 state. This requires that state to be stable; fortunately it can be easily shown using Lyapunov's second method that the point ① and the point ② are asymptotically stable in the sense of being convergent over time, whereas the middle circle is unstable because fluctuations around this point will drive the system towards the stable points.

For the LRS, only one stable solution exists in the current valley as illustrated by the blue dot in Fig. 7.5b.

To illustrate the behavior under switching, we show the V-NDR and MTJ currents for different MTJ resistance under constant $V_{CC}$ in Fig. 7.6a. If the MTJ is initially in the 0 state, the current is close to $I_{peak}$. Upon switching to the 1 state, the MTJ resistance will decrease and the stable solutions approach $I_{peak}$, beyond which the V-NDR current suddenly drops to the valley region. The current vs. resistance of this process is also




(a) (b)

Figure 7.6: (a) $I_{NDR}$ versus $V_{NDR}$ (solid lines) and $I_{MTJ}$ versus $(V_{CC} - V_{NDR})$ (dashed lines) for different $R_{MTJ}$. (b) Current of the series connection of MTJ and V-NDR vs. $R_{MTJ}$.

plotted in Fig. 7.6b. This demonstrates that, given proper choice of peak current, $|-R|$, and $V_{CC}$, the V-NDR device can sense the different MTJ states and switching therein and adjust the current through the circuit accordingly. For sufficiently large PVR, the resulting difference in the series-connected current can be much greater than the ratio of the MTJ resistances. This means that the operating margins are no longer limited by the TMR (typically 0.5-2X) but by the V-NDR PVR (which can be 5-1000X).

## 7.5 Modeling of V-NDR Devices

### 7.5.1 Tunneling V-NDR Modeling

To describe TD characteristics, we adapt the compact model and model parameters in [SDC96], which were chosen to fit the InAs/AlSb/GaSb TD characteristics presented therein. For n-type TFETs, while compact models have been developed to describe the positive drain-source voltage device operation, most simple models neglect the V-NDR characteristics under negative drain-source voltage. In this work, we model TFET V-NDR by fitting device data to the equation

$$I_{drain} = A_{TFET}(V_{gs}, V_{ds}) f_{NDR}(V_{ds}) + I_{diode} \tag{7.1}$$

where $A_{TFET}(V_{gs}, V_{ds})$ is an existing gate- and drain-voltage TFET $I-V$ analytical

model [PC12a], $f_{NDR}(V_{ds})$ is a function describing two-terminal TD current [SDC96], and $I_{diode}$ is a standard expression for intrinsic diode current. By adjusting the model coefficients we can match both the ordinary and V-NDR characteristics of simulated and experimental TFETs. In this chapter, we discuss results for TFETs using the simulated device characteristics presented in [LLZ12], which is compared with our analytical model in Fig. 7.2b.

### 7.5.2 Analytical Model of CMOS V-NDR Behavior

The characteristics of MRAM depend on technology and application, so to ease CMOS V-NDR design for any particular application, we derive a compact model of its current behavior. The essential design task for MRAM usage is to select the appropriate peak current and voltage to lie between the MTJ HRS and LRS load lines as shown in Fig. 7.5b. Typical peak voltages are around 50 to 200 mV ($V_{peak}$), indicating that transistors T1 and T2 (see Fig. 7.5b) are in the subthreshold region. By contrast, T3 operates in the linear region with its gate bias over 400 mV ($V_{int}$ in Fig. 7.5b) when V-NDR is in peak region. The subthreshold current [RMM03] $I_{sub}$ for T1 and T2 can be modeled using

$$I_{sub} = \frac{W I_0}{L} e^{\frac{V_{GS}-V_{th}-\lambda V_S+\eta V_{DS}}{nV_T}} \left(1 - e^{-\frac{V_{DS}}{V_T}}\right) \tag{7.2}$$

where $W$ and $L$ are the transistor width and length, $I_0$ is a constant parameter, $V_{th}$ is the threshold voltage, $V_T$ is the thermal voltage, and $n$, $\lambda$, and $\eta$ are the subthreshold swing, body effect, and drain-induced barrier lowering (DIBL) coefficients, respectively [RMM03].

The gate bias of T3 ($V_{int}$) can be expressed as a function of the T1 and T2 widths $W_1$ and $W_2$ and threshold voltage mismatch ($\Delta V_{th} = V_{th2} - V_{th1}$)

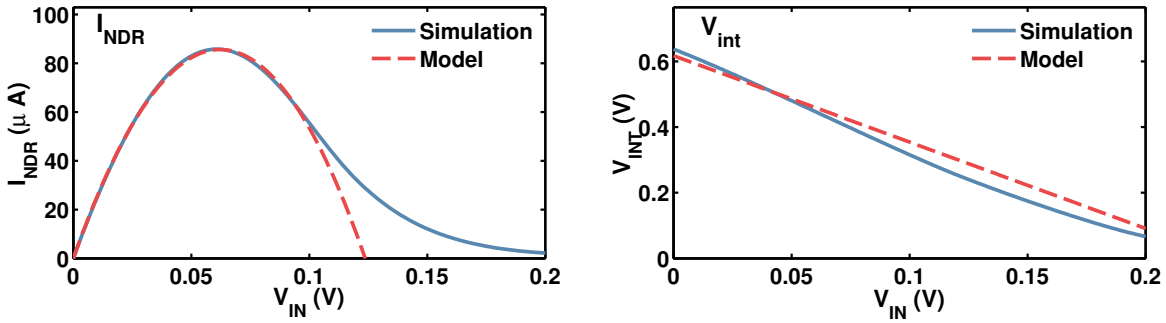
$$V_{\text{int}} = \left(\ln\left(\frac{W_1}{W_2}\right) nV_T + \Delta V_{th} + \eta V_{bias} - V_{in}\right) / \left(\lambda + 2\eta\right). \tag{7.3}$$

The current mostly flows through T3, which is in the linear region [Bre78] around the V-NDR peak. This allows us to estimate the V-NDR peak voltage and current

$$V_{peak} = \frac{\ln\left(\frac{W_1}{W_2}\right) nV_T + \Delta V_{th} - (\lambda + 2\eta)\, V_{th3} + \eta V_{bias}}{2 + \lambda + 2\eta} \tag{7.4}$$

$$I_{peak} = \frac{W_3}{L}\mu C_{ox} \times \frac{\left(\ln\left(\frac{W_1}{W_2}\right) nV_T + \Delta V_{th} - (\lambda + 2\eta)\, V_{th3} + \eta V_{bias}\right)^2}{2\,(2 + \lambda + 2\eta)\,(\lambda + 2\eta)} \tag{7.5}$$

The V-NDR circuit enters the valley region when $V_{in}$ pulls $V_{int}$ to zero and turns off T3, in which case the current is limited by the leakage currents of T1 and T3. Eqns. 7.3-7.4 show the V-NDR characteristics can be designed by adjusting the T1 and T2 size and $V_{th}$ mismatches as well as the terminal voltage $V_{bias}$.




(a) (b)

Figure 7.7: Model and SPICE comparison of 3T V-NDR circuit using 45 nm commercial CMOS library for (a) $I$ versus $V_{in}$ and (b) $V_{int}$ versus $V_{in}$.

This model describes qualitative features of the V-NDR circuit as the comparison with SPICE simulations shows in Fig. 7.7. Good matches are found for $V_{int}$ and $I_{NDR}$ with $V_{in} < 0.1mV$. The current mismatch for high $V_{in}$ in Fig. 7.7a is due to the oversimplified current equation for T3, which only works in linear region and is targeted to predict peak current and voltage.

To validate the model functionality, we use it to guide CMOS V-NDR designs for representative MRAM read and write applications using either spin transfer torque (STT-MRAM) or magnetoelectric memory (MeRAM) technologies. The parameters of MRAM and corresponding V-NDR are specified in Table 7.2. Since data on MTJ resistance of commercial MRAMs is not widely available, we choose illustrative values of 5 kΩ for STT-MRAM P resistance and 50 kΩ for MeRAM, which fall within the reported

Table 7.2: NDR and MRAM parameters for three different MRAM read or write design.

| MRAM | | | | NDR | | |
|---|---|---|---|---|---|---|
| MRAM | Operation | $R_P$ | TMR | $V_{dd}$ | $W_3$ | $V_{bias}$ |
| STT-MRAM | Write | 5 kΩ | 150% | 0.8 V | 1200 nm | 0.85 V |
| STT-MRAM | Read | 5 kΩ | 150% | 0.4 V | 550 nm | 0.8 V |
| MeRAM | Write/Read | 50 kΩ | 150% | 1 V | 210 nm | 0.6 V |

range for these technologies [DRT12, CRK10, GEA16]. Higher TMR gives rise to larger design margin, so we use 150% as a baseline, noting that TMR up to 180% has been demonstrated in commercial STT-MRAM [SLS16]. With these parameters, we design the V-NDR to allocate its peak current point between the $R_{AP}$ and $R_P$ load lines (see Fig. 7.5ba). We also use HSPICE to simulate the peak current point and its margin (the current difference from peak current point to $R_{AP}$ and $R_P$ load lines), which are plotted in Fig. 7.8. All predicted V-NDR peak current points lie close to the center of the simulated design margin, supporting the model.

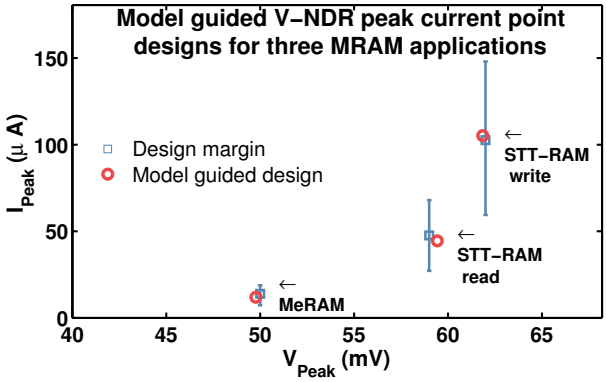



Figure 7.8: Simulated peak current points of V-NDR designs guided by model and their margins for three MRAM applications. The error bars illustrate allowed design margins.

## 7.6 V-NDR-assisted MRAM Write and Read

### 7.6.1 STT-RAM Write Energy Reduction

Having established the scenarios in which V-NDR devices can clamp the current of a serially connected MTJ, we propose using this effect to perform early write current cutoff when STT-MTJs switch to or stay in the 1 state. During writing from the 0 to 1 state,

this configuration can cause write termination by automatically cutting off the write current once the 1 state is attained and the V-NDR device ($V_{NDR}$) enters its valley region. Similarly, if a write-to-1 operation is performed on an MTJ already in the 1 state, the V-NDR forces a very low current during the whole write cycle, saving energy.

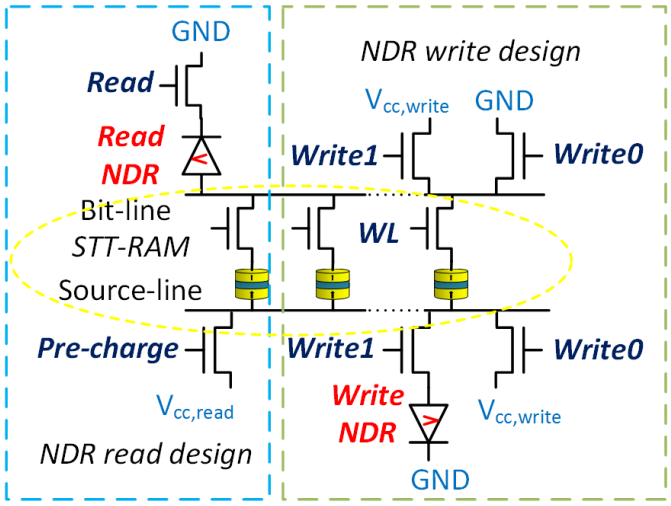



Figure 7.9: The proposed V-NDR read and write circuitry designs. Yellow dotted line denotes the STT-RAM array.

This can be accomplished by integrating V-NDR devices within the memory write circuitry in the manner shown on the right side of Fig. 7.9. An V-NDR device is added to a regular STT-RAM circuit for write assistance. Because only one MTJ cell is activated at a time during a write operation, the required overhead is minimal since, much like the sense amplifier, a single V-NDR device can be shared by multiple MTJ bit-lines. For a write-0 operation, *Write1*, *Read*, and *Pre-charge* are set to GND (0 voltage), and *Write0* is set to $V_{cc}$ to activate a write path without V-NDR (the same as conventional write). For a write-1 operation however, which dissipates the most power, only *Write1* is set to $V_{cc}$ and hence an V-NDR device is connected to the write path and acts to reduce the write energy by curtailing write current after switching.

The advantage of the proposed NDR-based design can be simply understood by com-

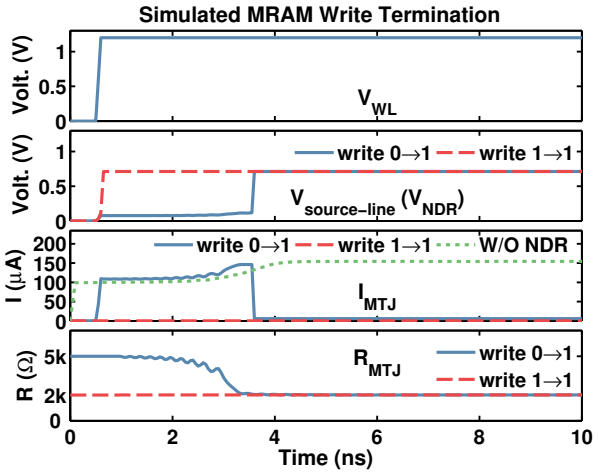


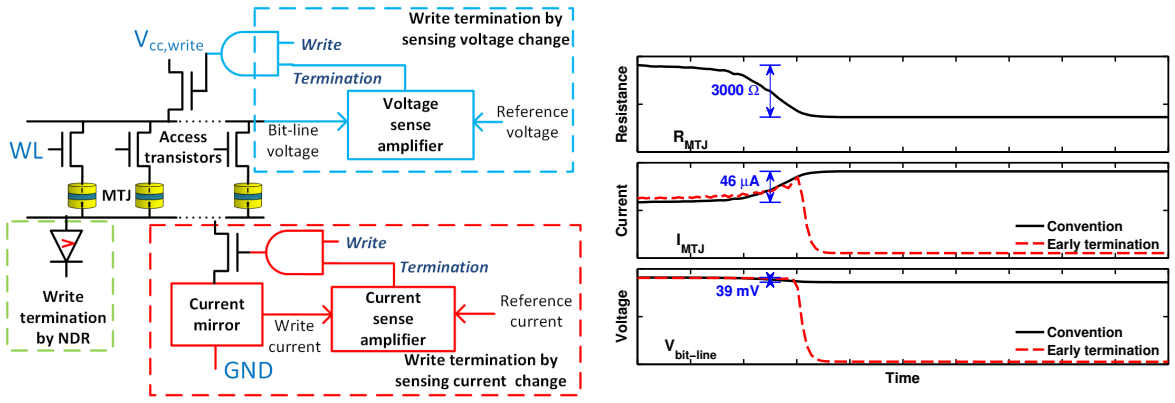

Figure 7.10: SPICE simulated waveforms for a write-1 termination in the memory design of Fig. 7.9. During write operation, a bit-line is first selected and charged to $V_{CC}$, then a MTJ bit is selected by $WL$, and write current is high or low depending on the MTJ initial state.

paring its energy dissipation $E_{NDR}$ with that of conventional MRAM designs ($E_{conv}$):

$$E_{conv} = V_{CC}^2 C_{BL} + V_{CC}(t_{AP} I_{AP} + t_P I_P) \tag{7.6}$$

$$E_{NDR} = (V_{CC} + V_{Peak})^2 C_{BL} + (V_{CC} + V_{Peak})(t_{AP} I_{Peak} + t_P I_{Valley}) \tag{7.7}$$

where $C_{BL}$ is the bitline capacitance, $I_{AP,P}$ are the currents when the MTJ is in the $AP$ or $P$ state, respectively, and $t_{AP,P}$ are the corresponding time intervals when the MTJ is in either state. In the proposed design, the applied voltage is increased slightly by the peak voltage of the V-NDR device (which is typically of order 0.1 V) and the write current in the $AP$ state approaches the V-NDR peak current $I_{Peak}$, but the write current in $P$ state is drastically reduced to the V-NDR valley current $I_{Valley}$. Comparing $E_{conv}$ and $E_{NDR}$, we see that most of the energy savings comes from the reduced consumption in the $P$ state and hinges on the ratio of $I_P$ and $I_{Valley}$, which is bounded by the PVR of the V-NDR device; higher PVR leads to more efficient operation. As the discussion in Section 7.4 shows, one prerequisite for current termination operation in our design is that $V_{NDR}$ at the beginning of the write process is lower than $V_{peak}$, so that current through V-NDR converges to the stable high current solution. This can be guaranteed by pre-discharging the source-line voltage to zero.




(a) (b)

Figure 7.11: (a) Three early write termination designs using bit-line voltage change sensing [ZZY09], current change sensing [BOE16], and the proposed NDR. (b) Simulated waveforms of MTJ resistance (AP: 5000 Ω, P: 2000 Ω, bit-line voltage, and write current as functions of time. The black line is for the conventional write, and the read dash line is for the early write terminations.

To quantify the effect of the proposed design, we simulate the waveforms of bit-line voltage ($V_{bit-line}$), bit selection (WL), voltage drop on V-NDR ($V_{source-line}$), write current ($I_{MTJ}$), and MTJ resistance ($R_{MTJ}$) under early write termination as shown in Fig. 7.10. If the MTJ starts in the 0 state, the V-NDR device voltage and current increase to near $V_{peak}$ and $I_{peak}$, respectively, after the WL goes high and stays there until the MTJ switches, whereupon the V-NDR voltage approaches $V_{CC}$, turning off the write current. If the MTJ is initially in the 1 state, the V-NDR directly goes to $V_{CC}$ after the WL goes high and cuts off the write current.

Our proposed concept has significant advantages over previous attempts at write termination which used additional sense and control circuitry as illustrated in Fig. 7.11a. In [ZZY09] (see blue dashed box in Fig. 7.11a), sensing circuitry is added to sense the voltage change on the bit-line and terminate the write. However, such voltage changes are small (see Fig. 7.11b) in general because 1) the MTJ resistance change is small, and 2) the MTJ-bit selection transistor resistance changes inversely with that of the MTJ (*e.g.*, a MTJ resistance decrease leads to a voltage increase on the transistor and its equivalent transistor resistance), partially canceling bit-line voltage change. The resulting low sensing margin leads to long sensing times and large sensing energy and is susceptible

to process variations. Another write termination design using current change sensing is proposed in [BOE16] (red dashed box in Fig. 7.11a); this requires a current mirror in the write path to copy the current change to sensing circuit, which increases write $V_{CC}$ and adds redundant write energy. Moreover, both voltage and current sensing designs require sense amplifiers and reference voltage/current generation (usually created by writing a reference MTJ in parallel), which add large energy overhead. One design uses different write pulses for the asymmetric write directions to save energy [BEO14], though this method cannot safely write MTJs at the variation corners. Yet another proposal also introduces V-NDR into the MRAM write circuitry to avoid a current increase after MTJ switching to 1 [HHS10]; however, as discussed in Section 7.4, that method cannot fully terminate write current because both MTJ states will still have relatively large current, in contrast to our design where the 1 state current is truly minimized in the V-NDR valley region.

### 7.6.2 STT-RAM Read Assistance using NDR

The sensing margin in STT-RAM is fundamentally bounded by the MTJ TMR ratio and is further reduced in practice by process variations; this has led to many proposals to maximize sensing margin within these limitations [KAG07, CWZ10, EZW14, MGK15, LGW16b]. However, we propose a new read design that boosts the current difference ratio beyond the TMR ratio up to the PVR of the V-NDR device, substantially increasing sensing margins and reducing the sensitivity to process variations.

#### 7.6.2.1 MRAM Sensing Margin Improvement

Similar to our write assistance design we propose a new read design, shown on the left of Fig. 7.9, in which multiple bit-lines share a single V-NDR device, minimizing overhead. During a read cycle, a normal pre-charge operation first charges the bit-line and source-line so that *Pre-charge* and *Read* are set to $V_{cc}$; the charge thus stored is then discharged through the MTJ and V-NDR series connection, where the latter amplifies the current difference between the 1 and 0 MTJ states. In this chapter, read margin is defined as the voltage difference on the bit-line during discharging of the two MTJ states, which is

sensed by a differential sense amplifier [Pro00]. For conventional designs not using NDR, the source-line is used rather than the bit-line. V-NDR read requires the initial voltage of the V-NDR to be below $V_{peak}$ before the pre-charge, which is guaranteed by discharging any remaining charge on the bit-line after each read or write operation. As in the write case, the V-NDR device should be selected such that $I_{peak}$ is larger than the MTJ current in the 0 state but smaller than that of the 1 state. Therefore, as shown in Fig. 7.12, when the 0 state is sensed, the V-NDR device always stays in its low resistance region below the peak and the bit-line cannot be charged up. When the 1 state is sensed, the transient current through the MTJ in the pre-charge stage will exceed $I_{peak}$, pushing the V-NDR device into its high-resistance valley region so that the discharging (sensing) current is cut off.

$$V_{M,conv} \approx \left(1 - \frac{I_{AP}}{I_P}\right) V_{BL} \tag{7.8}$$

$$V_{M,NDR} \approx \left(1 - \frac{I_{Valley}}{I_{Peak}}\right) V_{SL} = \left(1 - \frac{1}{PVR}\right) V_{SL} \tag{7.9}$$

where $V_{SL}$ is the source-line voltage. Whereas the read margin in ordinary designs is limited by the discharging current ratio of the 1 and 0 states (typically around 0.7-0.9), in the V-NDR design $V_{M,NDR}$ is limited by $1 - 1/PVR$ which approaches 1 for well-designed V-NDR devices, dramatically increasing the read margin.

Our design has several key advantages over previous proposals. For instance, [KAG07, CWZ10, EZW14] proposed local-reference and self-reference read designs to improve immunity to process variations, but could not improve the margin beyond the MTJ TMR limits. Others have proposed amplifying the read margin using V-NDR [HHS10, UYY14, UYY15], but because of the limited read current difference in those designs (see Fig. 7.5b), but the achievable margins are smaller and read disturbance is not minimized in contrast to our proposal. Refs. [UHM03, UY03] propose introducing a TD into each MTJ cell to amplify the read margin, but such a requirement is not applicable to current-switched technologies like STT-RAM and greatly increases area overhead.

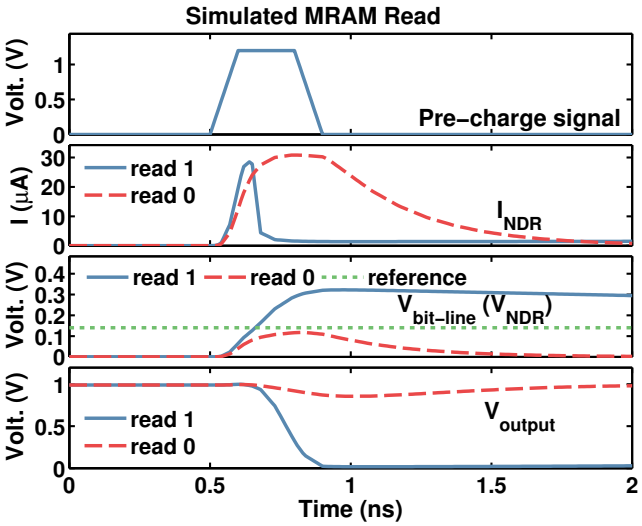



Figure 7.12: SPICE simulation of read operations using NDR-assisted design in Fig. 7.9. The discharging current ($I_{NDR}$) difference for sensing MTJ states is significantly increased by the high PVR of NDR. A large and constant voltage margin is achieved on the bit-line ($V_{bit-line}$), which is sensed by a constant reference voltage leading to a stable sense amplifier output ($V_{output}$).

#### 7.6.2.2 Read Disturbance Minimization

Read disturbance, another primary reliability concern, is caused by false switching of a MTJ via sensing currents through the cell. Any transient current pulse has some possibility of switching a MTJ, but the probability of such switching increases dramatically with the amplitude and duration of the current pulse. To tame this, conventional designs have to reduce read current as well as sensing margin and bit-line size, resulting in larger sensing circuits (large gate sizes and more sensing transistors), longer sensing time, and higher sensing energy. In the proposed NDR-assisted design, the PVR simultaneously reduces read disturbance and improves read margin while introducing little overhead. As Fig. 7.9 shows, an $AP$ state MTJ cannot be falsely switched due to the sensing current direction and read disturbance can occur only during read-1 operation when the sensing current accidentally switches a STT-MTJ from the 1 to 0 state. However, the probability of such disturbances in our design is almost zero since the switching probability scales exponentially with current and duration. We can dramatically lower the sensing current through a 1-state MTJ to the near-0 V-NDR valley current, which is significantly lower than previous works [HHS10, UYY14, UYY15]. The read disturbance rate is simulated

for different bit-line size in Section 7.6.4.

### 7.6.3 Experimental Validation

We validate the proposed applications of CMOS V-NDR in MRAM through experiment and simulation. In 7.6.3.1, we measure the response of a V-NDR circuit in series with a voltage-controlled MTJ (VC-MTJ) as a function of time and applied magnetic field to illustrate how write termination and read margin amplification can occur for a single memory cell. This helps validate the concept and advantages of our approach at the device level (for reading/writing a single cell). Though VC-MTJ is used in the experiment, the demonstrated concept can be extended to other resistive memory devices, like STT-MTJ. However, device and circuit variability play an important role in the actual performance of real MRAM arrays. To prove that the demonstrated cell-level improvements can survive these effects, in section 7.6.4 we use Monte Carlo circuit simulations to show the robustness of the energy savings and accuracy improvements accounting for circuit variation.

#### 7.6.3.1 V-NDR MTJ Experiments

As discussed above, we build a CMOS V-NDR circuit by wiring pre-fabricated MOSFETs on a single die, with the resulting characteristics shown in Fig. 7.3b. We connect this circuit in series with a 50 nm diameter voltage-controlled MTJ [GEA16]. The stack structure of the MTJ consists of bottom electrode/$Co_{20}Fe_{60}B_{20}$(1.1)/MgO (1.4)/$Co_{20}Fe_{60}B_{20}$(1.4)/Ta(0.25)/[Co/P electrode (numbers in parentheses are thicknesses in nm).

The experimental MTJ can be switched between the AP and P states by either tuning the applied voltage (of order 0.4-0.7 V, corresponding to current of 1-2 $\mu$A) or an externally applied out-of-plane magnetic field. A sample MTJ $R-H$ curve is shown in Fig. 7.13. The MTJ is bistable for external fields between -1400 G to -1200 G, where stochastic switching between the P and AP states occurs due to thermal activation; outside the bistable range, the MTJ state can be deterministically controlled. The AP- and P-state resistances are around 320 k$\Omega$ and 240 k$\Omega$ under a voltage bias of 0.1 V.

The external wiring and transistor size limitations make the response time of our

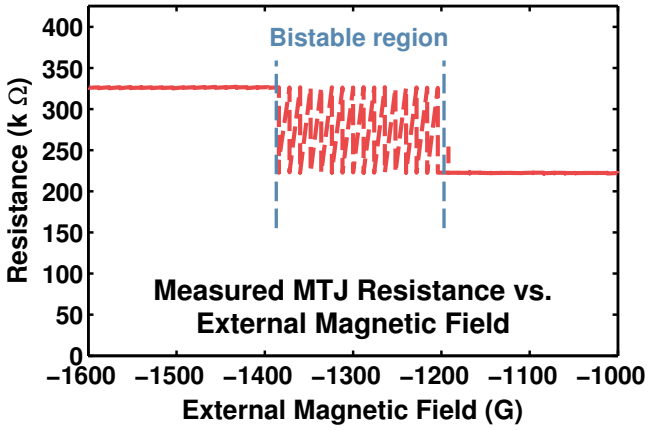



Figure 7.13: Experimental MTJ resistance ($R_H$ = 320 kΩ, $R_L$ = 240 kΩ) as a function of external magnetic field.

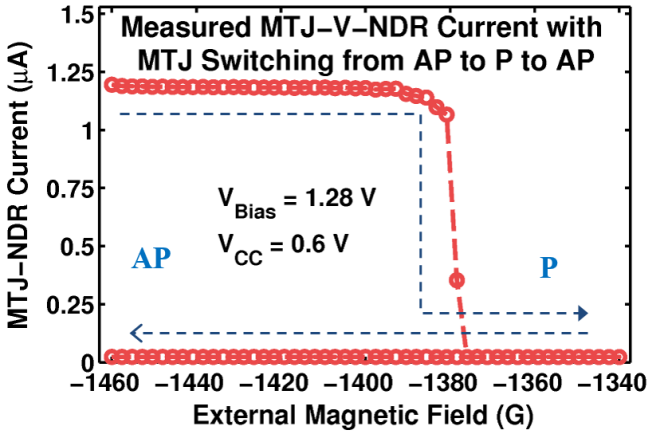



Figure 7.14: Current through series-connected V-NDR and MTJ as external field is cycled to switch the MTJ from AP to P and back.

V-NDR circuits (∼ 1 ms) much slower than the internal switching time of our MTJs (∼ 1 ns), so that at present we cannot observe the switching on the internal time scale of the MTJs. Therefore, in these experiments, rather than relying directly on current-driven switching, we first use the external magnetic field to control the MTJ and avoid MTJ thermal activated switching during the V-NDR response time, as shown in Fig. 7.14. This is not a fundamental constraint since on-chip V-NDR circuits using nanoscale transistors can easily operate at higher speeds (on the order of ps) than MTJ switching times. By sweeping the external field magnitude, we could switch the MTJ from its AP to P state and back and observe the current response of the combined circuit. As seen in Fig. 7.14, a steep decline in current occurs when the MTJ switches from AP to P,

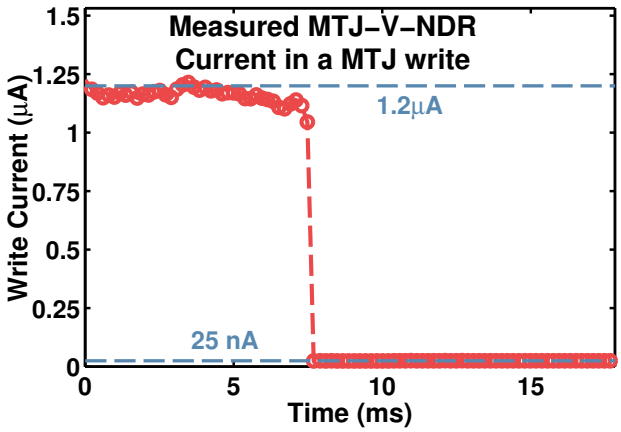



Figure 7.15: Time-dependent measurement of MTJ-V-NDR current for MTJ initially in AP state and with external magnetic field of -1.34 kG biasing the device in the bistable region. $V_{CC} = 0.6$ V for this measurement. Note the drastic reduction in current through the circuit around 7.5 ms from 1.2 $\mu$A to 25 nA due to switching of the MTJ from the AP to P state.

pushing the V-NDR from the peak region ① (see Fig. 7.5b) into the valley region ②; the low current is maintained even after the junction switches back to AP because the valley point ② is a stable state.

To better demonstrate write termination, we next initialize the MTJ in the AP state and then situate it in the bistable region by adjusting the external magnetic field to -1.34 kG. We then measure the current through the circuit (equivalent to the write current) as a function of time. In this regime, the MTJ can switch states due to voltage controlled magnetic anisotropy (VCMA) or spontaneous thermal activation [WLE16b]. The resulting measurement of current versus time is shown in Fig. 7.15. We observe around the 7.5 ms mark that the write current drops dramatically by a factor of 40 (from 1.2 $\mu$A to 25 nA), owing to the switching of the MTJ from the AP to P state at that time. Once this switching occurs, the V-NDR reaches its new stable point in the valley region and the current reaches and stays at a very low level. This mimics how write termination can occur in a real MRAM application when writing from AP to P states. As before, the application of an external magnetic field in this experiment is simply to bias the MTJ in bistable regime and thus lengthen the time scale of the switching so it becomes visible with our instrumentation and circuits. In a practical MRAM array, no external magnetic

field is present and the time scale of switching will be far faster, as noted, but the effect of the V-NDR circuit remains the same.

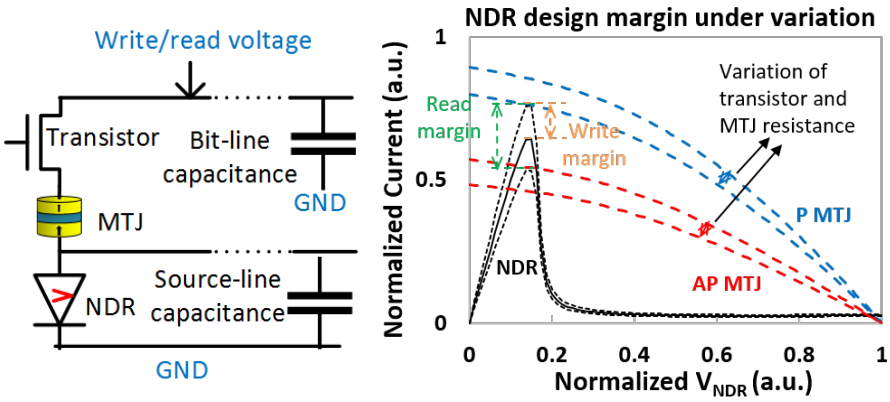
We note that the intrinsic resistance ratio of the MTJ used in this experiment is only 1.33X as shown from the high and low resistance in Fig. 7.13 (=320 kΩ/240 kΩ at 0.1 V). This ratio should further reduce if a higher bias is applied to the device [SLS16]. The effective TMR of 33% is hardly the largest possible experimentally for MTJs; nonetheless, the difference in current between the two states once the V-NDR is introduced is far greater and amplified to in excess of 40X (=1.2 $\mu$A/25 nA in Fig. 7.15 and likewise in Fig. 7.14). This shows how read margin between the AP and P states can be drastically improved by 30X (=40X/1.33X) by using V-NDR during the read process. The improvement in current ratio may actually even more greater than 30X because the effective voltage drop across the MTJ is estimated to be around 0.3-0.4 V in the series connection, and the device TMR is likely to be lower than 33% (measured at 0.1V voltage bias) at high MTJ bias. As this demonstration was performed using preexisting MTJs and CMOS devices whose characteristics were not optimized for usage in V-NDR applications, we expect that even greater quantitative improvements will be possible for a pre-designed and integrated implementation in MRAM.

### 7.6.4 Array-level Reliability and Performance

#### 7.6.4.1 Tunneling-based V-NDR assisted MRAM

We next study the robustness of our proposed design under process variations using comprehensive Monte Carlo circuit simulations. To study variability effects accurately, massive simulations (over 10 billion runs) are needed to detect WER and read disturbance rates down to $10^{-9}$, which are impractical for conventional SPICE methods (indeed, we estimate such a calculation would take decades); instead, we implement our device physical model and a small circuit simulator using CUDA and run them on a Tesla M207 GPU, enabling billions of simulations within a few hours. The simulated circuit is shown in Fig. 7.16a. Process variations for transistors and MTJs are included in the manner of [WLE16b].

We first assess variation-imposed limits on the V-NDR design margin, as illustrated

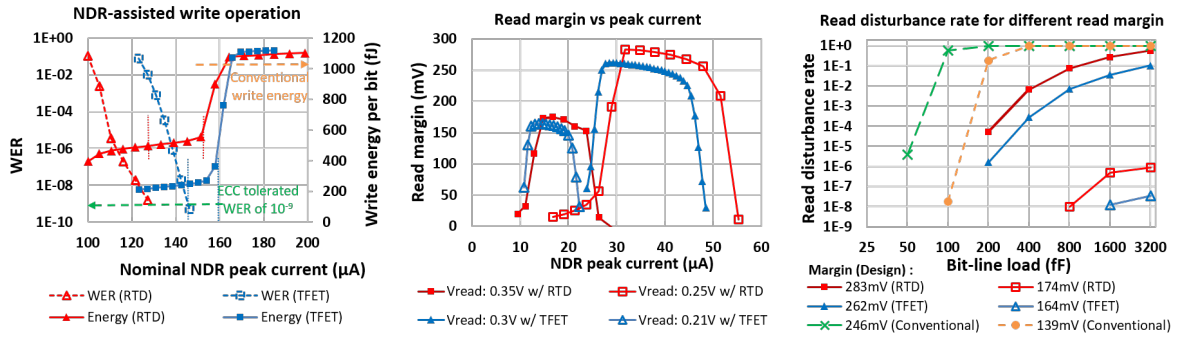



(a) (b)

Figure 7.16: (a) Simulated circuit implementation in CUDA. (b) V-NDR peak current margin variation analysis.

in Fig. 7.16b. Here the blue and red dashed lines show the corner cases for the MTJ high and low resistance states versus $V_{NDR}$. For read applications, the V-NDR $I_{peak}$ must lie between the high and low currents (the top and bottom black dotted lines). When writing 0-to-1, $I_{peak}$ must lie within the top half of the current margin (between the top and middle dotted lines) to guarantee that V-NDR cuts off once the MTJ reaches an intermediate resistance state (between $R_H$ and $R_L$) where it can converge to the 1 state without further assistance from a large switching current. Fluctuations in devices due to process variations like random dopant fluctuations or line edge roughness [DKL13, LC13c, WLP13] can change the threshold voltage of TFETs and affect $I_{Peak}$. We analyze the design tolerance for such process variations through Monte Carlo simulation.

In total we perform over 100 billion Monte-Carlo simulations on the NDR-assisted write process for 0 to 1 switching. In addition to ordinary write errors caused by thermal fluctuations, we find a special error may also occur in the NDR-assisted write process for individual MTJs or transistors with very low resistance due to process variations (*e.g.*, when the intersection of the initial high resistance state approaches the V-NDR peak current too closely in Fig. 7.16b). In such cases, the V-NDR may turn off the write current when variation causes the MTJ current curve past $I_{peak}$ but before the MTJ can switch to its low resistance state. To avoid such write errors, a higher V-NDR peak current is required, increasing write energy. The write energy and WER as functions of V-NDR peak current are shown in Fig. 7.17a. As peak current rises, the WER decreases but




(a) (b) (c)

Figure 7.17: Simulation results with transistor and MTJ process variations. V-NDR characteristics are varied by scaling diameter for TDs (0.4-0.55 µm) and threshold voltage for TFETs (assuming device width=1.95 µm in write and 0.195 µm in read circuitry).(a) WER and write energy versus nominal V-NDR peak current. Write energy includes bit-line pre-charge, access transistor, and MTJ, but excludes row/column decoders. (b) Read margin versus V-NDR nominal peak current. $C_{BL} = C_{SL} = 25$ fF in (a) and (b). (c) Read disturbance rate as function of bit-line/source-line load and read margin. High/low margin designs are obtained using different $V_{read}$ (0.35 V/0.25 V for TD, 0.3 V/0.21 V for TFET, and 1.8 V/0.7 V for conventional designs). Read disturbance rates below $10^{-10}$ not detectable within sample size.

write energy increases. Fortunately, for the standard ECC requirement of WER $< 10^{-9}$, significant energy reductions ($> 50\%$ lower than conventional designs) are realized over a wide range of nominal $I_{peak}$ for both TFET ($161-146 = 15$ µA or 20 mV threshold voltage shift) and TD ($153-126 = 27$ µA) designs; this is the effective design tolerance of V-NDR devices for write circuits. We summarize the write performance results in Table 7.3 where, for fairness, we compare NDR-based and conventional circuits with the same write latency and WER. Comparing the energy usage of write-1 operations (0-to-1 and 1-to-1), we see 76% and 52% energy savings for TFET- and TD-based designs, respectively. Looking at the average write energy (assuming equal usage of the four switching directions), we still see major reductions of 52% and 36% for TFET- and TD-assisted writes, respectively.

The dependence of the read margin (current margin on the source-line) on $I_{peak}$ and read voltage is shown in Fig. 7.17b. Again we observe good design tolerance for V-NDR device variations from the range of $I_{peak}$ for which read margin is large and nearly

Table 7.3: Write energy, read margin, and read energy of NDR-assisted designs as extracted from Fig. 7.17. $C_{BL}$ = 25 fF; nominal TFET $V_th$ = 0.25 V. Since V-NDR does not affect write-0 operations (*0→0* and *1→0*), conventional designs are used for these cases. Effective PVR is the ratio of circuit current in the 1 and 0 states for chosen $V_{CC}$ and differs for write and read due to different bias.

| | | **Conventional** | **TFET** | **TD** |
|---|---|---|---|---|
| Write Energy (fJ) | ***0→1*** | 1040 | 248 | 498 |
| | ***0→0*** | 699 | 699 | 699 |
| | ***1→1*** | 1269 | 61 | 419 |
| | ***1→0*** | 838 | 838 | 838 |
| | Average | 964 | 462 | 613 |
| Write latency (ns) | | 9 | 9 | 9 |
| Read voltage (mV) | | 700 | 210 | 250 |
| Read margin (mV) | | 139 | 164 | 174 |
| Read energy | | 42 | 5.5 | 3.9 |
| NDR design variation tolerance | Peak current shift ($\mu A$), Fig. 7.17 | Write | 15 | 27 |
| | | Read | 8 | 11 |
| | Threshold voltage shift (mV) | Write | 20 | N/A |
| | | Read | 105 | N/A |
| Effective PVR | | Write | 23.6 | 2.64 |
| | | Read | 8.38 | 5.86 |

constant. At low $V_{read}$ (0.21 V for TFET and 0.25 V for TD), a read margin of over 150 mV can be maintained for the TFET design over an 8 µA variation range in peak current (equivalent to 105 mV threshold voltage shift), and for the TD design over an 11 µA variation in $I_{peak}$. The trend continues at higher read voltages as well. The read performance is summarized and compared with a representative conventional read design in Table 7.3; we note that the latter requires a much larger $V_{read}$ = 0.7 V to achieve comparable read margin, consuming much more energy.

Finally, we examine read disturbance rates under V-NDR and conventional designs for different source-line and bit-line loads. Longer bit-lines have larger loads, leading to more charging/discharging current during read. In the read operation, current flows from source-line to bit-line, which may falsely switch the MTJ from 1 to 0. Fig. 7.17c shows simulated read disturbance rates as functions of bit-line/source-line size (load). Compared with conventional designs with similar read margin, NDR-assisted designs enable vastly improved disturbance rates (over 10 million times lower). Moreover, the read disturbance

rates of conventional designs cannot satisfy the ECC requirement ($< 10^{-9}$) for bit-line loads larger than 100 fF. The dramatic read disturbance reductions we observe for the NDR-assisted design are due to its low $V_{read}$ and minimized discharge current (limited by $I_{valley}$).

In the proposed applications, we can use a large V-NDR device (e.g., gate width over 1 $\mu$m for V-NDR write) to provide sufficient peak current and minimize process variation. This does not limit memory density since every V-NDR device can be shared by multiple bit-lines containing thousands of cells. Using results from a device variation analysis on 14nm TFETs [DKL13], we estimate using the variation scaling rule [BS95] that the $6\sigma$ of threshold voltage shift is about 9 $mV$ if a $200nm$ x $1000nm$ TFET-NDR, well within the simulated tolerance level of our design.

The simulation parameters for V-NDR assisted write and read is shown in Table 7.4. The PVR is the effective PVR determined by both V-NDR device and voltage bias.

Table 7.4: Simulation parameters at 300K.

| **STT-RAM** | Diameter | $t_{MgO}$ | $t_{tfl}$ | $R_P$ | $R_{AP}$ | BL |
|---|---|---|---|---|---|---|
| | 50nm | 1.18nm | 1.1nm | 2k $\Omega$ | 5k $\Omega$ | 256 bits |
| **NDR** | Write | | High-margin read | | Low-margin read | |
| | Width | PVR | Width | PVR | Width | PVR |
| TFET | 1.95 $\mu m$ | 23 | 0.39 $\mu m$ | 17 | 0.2 $\mu m$ | 8 |
| RTD | 0.44 $\mu m$ | 2.7 | 0.23 $\mu m$ | 10.5 | 0.16 $\mu m$ | 5.5 |

#### 7.6.4.2 CMOS-built V-NDR assisted MRAM

The write termination and increased read margin characteristics that are experimentally observed can significantly improve the power consumption and reliability of MRAM [WPC17]. At the array level, process variation and write error rate are also main concerns due to the low TMR and stochastic switching behaviour of MTJs. We have proposed that V-NDR can also improve MRAM performance by reducing read disturbance. To quantify these improvements and evaluate their robustness in the presence of device fluctuations, we perform Monte-Carlo simulations of STT-MRAM read and write operations including CMOS V-NDR circuitry and considering both transistor and MTJ process vari-

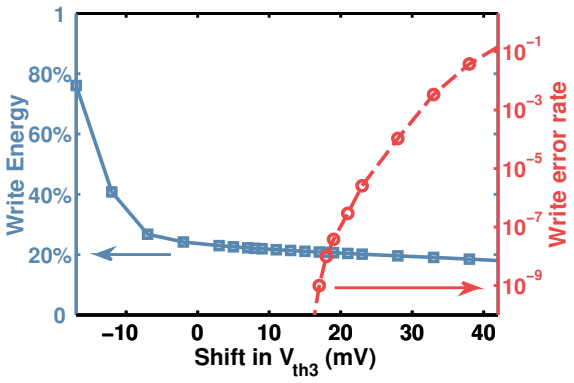



Figure 7.18: Simulated write energy (normalized to conventional write scheme) and WER (right axis) vs. threshold voltage shift of T3 for 25 fF bit-line load (256 1T-1M cells per bit-line). The WER is extracted from 10 billion Monte-Carlo numerical simulations for V-NDR assisted STT-MRAM write. MTJ device parameters can be found in Table 7.5.

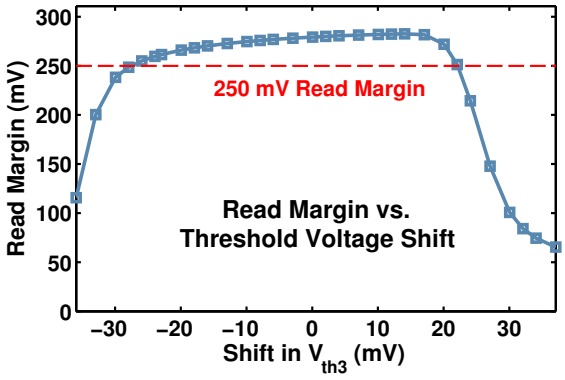



Figure 7.19: Simulated STT-MRAM read margin vs. T3 threshold voltage shift. MTJ device parameters can be found in Table 7.5.

ations [WLE16b] except for Fig. 7.19, where 3-$\sigma$-corner resistance are used for simulating read margin. The MTJ parameters used in our simulations can be found in Table 7.5 [WLE16b].

The dominant source of variability in the V-NDR circuit is the threshold voltage $V_{th3}$ of the T3 transistor in Fig. 7.3a, which affects $I_{peak}$ and PVR. In Fig. 7.18, we observe how the write energy improvement and write error rate (WER) for V-NDR-assisted operation vary with shifts in $V_{th3}$. As $V_{th3}$ increases, $I_{peak}$ reduces according to Eqution 7.4; this causes write termination to occur earlier during switching and reduces write energy but increases WER. For typical system requirements of WER $< 10^{-9}$ [WPC17, WHZ16], $V_{th3}$ can vary over 35 mV while maintaining write energy savings from 20% to 80%. Please

note that the MTJ variation has been included in the WER and RDR simulation, which is shown in details in Table 7.5. Similarly, we examine the impact on read margin in Fig. 7.19 and find that large voltage read margins (source-line voltage difference for reading AP and P) in excess of 250 mV can be sustained over a 60 mV window in $V_{th3}$. In practice, CMOS V-NDR circuits can be designed with large diffusion areas to minimize $V_{th3}$ shifts to nearly zero, providing more than sufficient design margin.

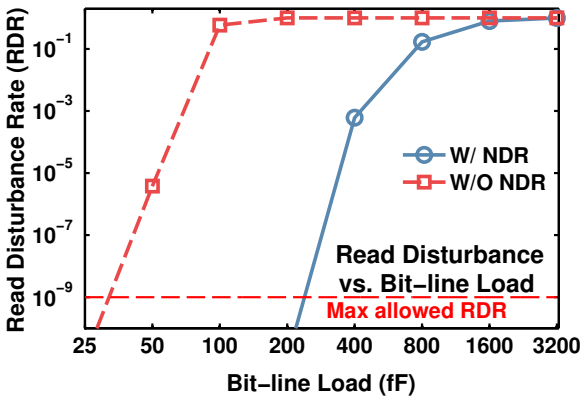



Figure 7.20: Simulated read disturbance rate vs. bit-line size (load) for read design with and without V-NDR. Larger load leads to more pre-charging/dis-charging current and more read disturbance.

Table 7.5: Variation parameters for STT-MTJs in the simulations of write error rate and read disturbance rate.

| Parameters | Mean | Variation |
|---|---|---|
| Diameter | 50nm | $\sigma$=1nm |
| MgO thickness | 0.9nm | $\sigma$=0.003nm |
| $T_{FL}$ | 1.18nm | $\sigma$=0.003nm |
| Resistance | 2KΩ / 5KΩ | dependence |

We also simulated read disturbance rate (RDR) for bit-line size from 25 fF to 400 fF. In STT-MRAM's precharge-and-sense read [LAD15], the precharging/discharging current flowing through STT-MTJ may falsely switch its state. In a V-NDR assisted read, the circuitry is designed such that the precharging/discharging current tries to switch MTJ to AP, and V-NDR is designed to allow peak and valley current for AP and P states respectively. This minimizes the disturbance current (i.e., P-to-AP switching). In the read, larger bit-line load leads to more precharging/discharging current and more read

disturbance. The read disturbance rates of design with and without V-NDR are shown in Fig. 7.20 as a function of load size. V-NDR shows $10^9$ X read disturbance reduction for bit-line load below 200 fF. With $10^{-9}$ error rate as a requirement, V-NDR improves the maximum bit-line load size from 30 fF in conventional read to 250 fF.

## 7.7 V-NDR and C-NDR for MLC ReRAM Programming

In MLC ReRAM, more than two resistances are used to store data. The conventional scheme for intermediate resistance programming (not the lowest and highest resistance) is multiple write-and-check cycles [PMH15], e.g., 20 cycles, where a read operation following a write pulse is to check the cell resistance against target value and determines additional programming cycles. This scheme has long programming time, and repeated charging and discharging wastes a significant amount of energy in a large array.

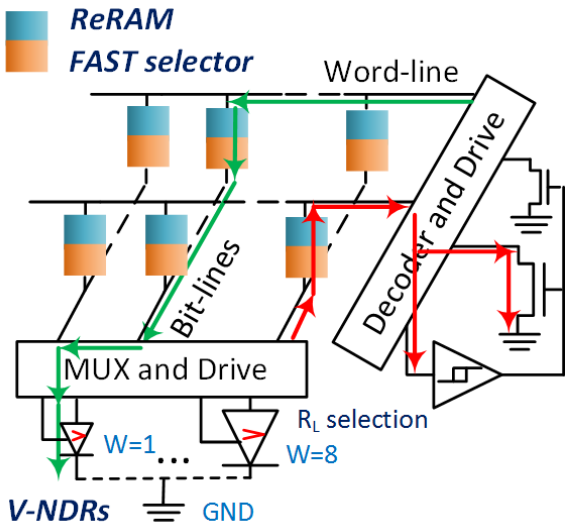
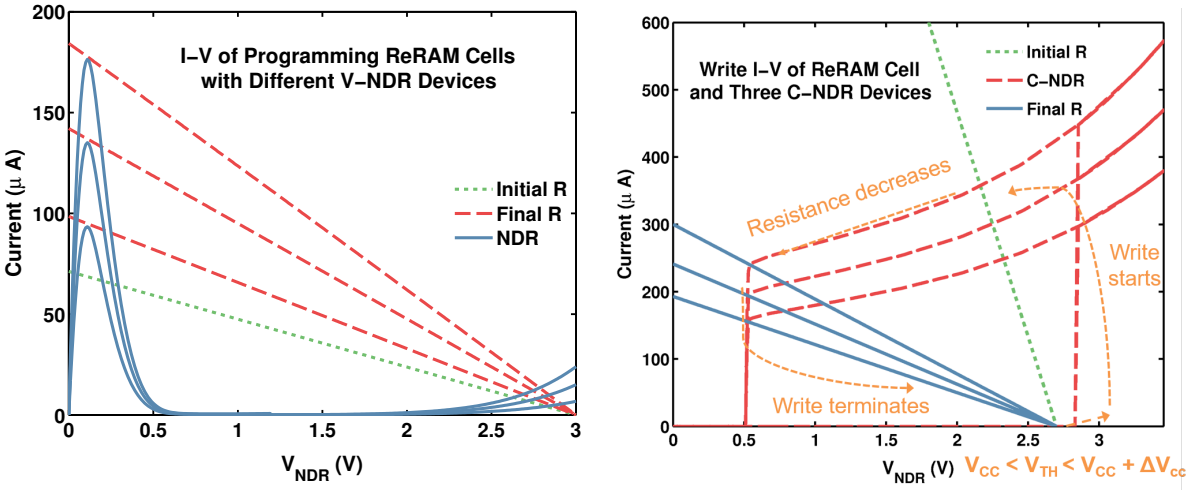



Figure 7.21: Using multiple V-NDRs and C-NDRs to program ReRAM cell resistances. In the programming, V-NDR can decrease a cell resistance from high value to a low value that is determined by V-NDR peak current, while C-NDR can program a cell resistance in the reverse direction. Once cell resistance achieves target resistance, both V-NDR and C-NDR can terminate write immediately

By utilizing V-NDR and C-NDR, intermediate resistance programming can be completed in one cycle. Fig. 7.21 shows an example of using V-NDR and C-NDR for MLC ReRAM with high-selectivity FAST selector [JKN14]. V-NDR and C-NDR are used for

writing a cell to lower and higher resistance respectively. Multiple V-NDR and C-NDR with different sizes are utilized, where every NDR is sized according to one target intermediate resistance such that it terminates write current when the programmed cell reaches the target value. The simple NDR design combines the functions of read check and write control, MLC programming efficiency would be dramatically improved.

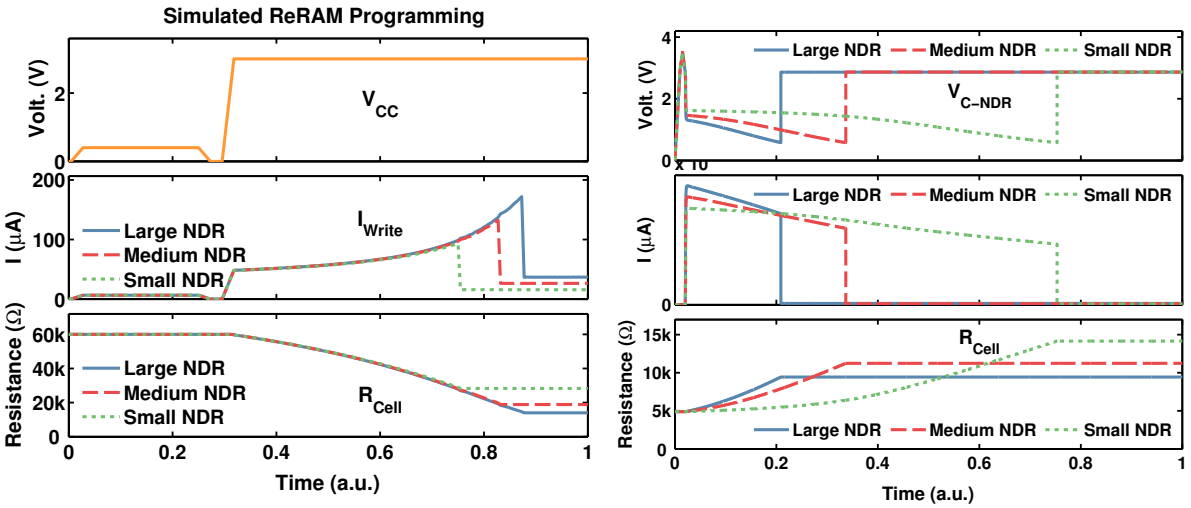



(a) (b)

Figure 7.22: (a) I-V curves of programming a ReRAM cell resistance using three different sized V-NDR. (b) I-V of programming ReRAM cell resistance using three different sized C-NDR devices (red dashed lines).

The I-V curves in Fig. 7.22 illustrate how the target ReRAM intermediate resistance is determined by V-NDR and C-NDR. The V-NDR assisted MLC programming is similar to the write termination of STT-RAM in Section 7.6.1. With write voltage applied on the serially connected ReRAM cell and V-NDR, write current increases as cell resistance reduces, and upon current reaching V-NDR peak, the write current gets terminated. For the LRS-to-HRS programming, C-NDR is utilized and serially connected to an ReRAM cell. The applied voltage on the series connection $V_{cc}$ is slightly below $V_{TH}$. The C-NDR turns on when a small voltage pulse ($\Delta V_{CC}$) is applied on $V_{cc}$ causing $V_{NDR}$ overshooting the $V_{TH}$ (the high threshold voltage of Schmitt trigger) to push C-NDR into high current region. Then, the high write current switches ReRAM to higher resistance, causing $V_{NDR}$ to decrease. Write current terminates when $V_{NDR}$ reaches the low threshold voltage.

To validate the proposed one-cycle programming idea using V-NDR and C-NDR, we




(a)

Figure 7.23: (a) I-V curves of programming a ReRAM cell resistance using three different sized V-NDR. (b) I-V of programming ReRAM cell resistance using three different sized C-NDR devices (red dashed lines).

simulated the MLC programming process described above with ReRAM model [BHS15]. The waveforms are shown in Fig. 7.23. Three different final resistances after programming are determined by the V-NDR and C-NDR sizes.

Please note that ReRAM's current is non-linear to applied voltage (i.e., resistance changes with biased voltage), indicating that the ReRAM cell resistances under read and write voltages are different. There is no variation study on this non-linearity so far, which may limit the resistance programming precision.

## 7.8 Chapter Conclusion

We have proposed using V-NDR and C-NDR to assist MRAM and MLC ReRAM programming and sensing. For MRAM, we propose a novel NDR-assisted write termination design and a NDR-assisted read reliability enhancement design. In a write-to-LRS operation, V-NDR can detect the MTJ switching and cut off current flow to avoid energy wasting. In the read operation, V-NDR can amplify the sensing current and voltage margin, reducing read voltage and read energy. Additionally, read disturbance current can also be reduced by V-NDR. For ReRAM, a one-cycle resistance programming mechanism

is proposed with the assistance of V-NDR and C-NDR. V-NDR and C-NDR are designed for HRS-to-LRS and LRS-to-HRS programming respectively, where target resistance is determined by the size the NDR device.

We have also proposed the CMOS implementation of V-NDR and C-NDR, and have modeled and experimentally demonstrated the CMOS V-NDR device. Experiments with in-house built VC-MTJs have demonstrated write termination and read margin improvement in MRAM, and have shown write current drop ratio of 40X upon VC-MTJ switching, even with unoptimized V-NDR and VC-MTJ devices. Large scale Monte Carlo simulations demonstrate over 50% STT-RAM write power reduction and $10^9$ X reliability improvement of read disturbance in the presence of device variability. Circuit simulation results also demonstrate the one-cycle programming mechanism for ReRAM with V-NDR and C-NDR devices. Though our experiments are conducted using VC-MTJs, STT-MTJs, and ReRAM, the same designs can be applied to read/write circuitry for a wide array of magnetic and nonmagnetic resistive memories.

# CHAPTER 8

# Hybrid VC-MTJ/CMOS Non-volatile Stochastic Logic for Efficient Computing

## 8.1 Chapter Introduction

As a class of accelerators, stochastic computing (SC) [Gai69, QR08, BC01] intrinsically has great advantageous energy-efficiency due to very simple hardware implementation for logic operations such as addition and multiplication. Applications which doesn't rely on precise computation can potentially benefit from the use of SC in terms of energy efficiency, speed and high fault tolerance, e.g., digital signal processing applications, data error correction and neural network [KKY16]. Recently, early evaluation of SC designs using stochastic NVM like memristors [GSZ13, SD12, SMW15] and all spin logics [VVF15], have shown significant improvement in energy efficiency.

However, challenges exist in the adaption of SC designs into modern computing systems. For conventional CMOS SC, the use of Linear-feedback shift register (LFSR) to generate stochastic bit streams (SBS) consumes high energy, offsetting the benefit brought by SC [KKY16]. The recently proposed NVM based SC designs like [GSZ13] introduce additional memory read and write to feed and fetch data from CMOS logic unit, moreover, the endurance limitation and high write voltage are not compatible with on-chip system. The SC designed by all spin logic [VVF15] faces reliability and efficiency concerns on the spin channels, and design challenges of stochastic bit streams generator (SBSG).

In this chapter, we propose a practical NV computing system [WPL17] using stochastic logic built by VC-MTJs [AUA13, SNB12] and CMOS based NDR [WPC17]. Unlike traditional NV logic and SC using NVM as additional data backup, where computation is still in CMOS logic units, this system directly computes data on VC-MTJs, elimi-

nating the need for the memory read and write and communication between NVM and CMOS logic. Thus, the proposed computing architecture is intrinsically fast and energy-efficient. In addition, VC-MTJ and V-NDR enable the design of robust SBSG, which generates truly random SBS with least design challenge. The proposed stochastic logic operations using 60nm in-house built VC-MTJs [GEA16] and CMOS V-NDR have been experimentally demonstrated.

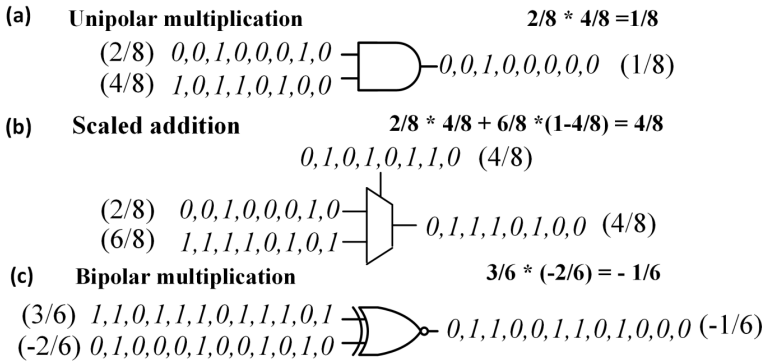



Figure 8.1: Multiplication and addition using unipolar and bipolar encoded SBS. Unipolar coding represents decimal number ranging in [0,1], while bipolar coding is for decimal number in [-1,1]. The SC computations are bit-wise, where corresponding bits in two input SBS are operated using the AND, MUX, or XNOR gates.

This chapter is organized as follows. In Section 8.2, we introduce SC. In Section 8.3, we briefly review the the interaction of V-NDR and VC-MTJ and introduce their functions in SC. In Section 8.2, we describe operations in the VC-MTJ SC design. In Section 8.5, we describe the SBSG built by VC-MTJ and NDR. In Section 8.6, we evaluate the proposed SC with finite impulse response (FIR) filter and Adaboost design [ROM01], and compare the proposed SC with CMOS binary and CMOS SC designs. The chapter is concluded in Section 8.7.

## 8.2 Overview of SC

SC [Gai69, QR08, BC01] uses the fraction of "1"s in an SBS to represent a fraction number. Two common encoding methods are unipolar and bipolar. For a unipolar encoded SBS, the fraction of "1"s is the represented number, e.g., 6 "1"s out of 8 bits is 0.75. By contrast, in a n-bit bipolar encoding, a number with $m$ '1"s represents $m/n-1/2$.

SC computation using SBS is bit-wise. As shown in Fig. 8.1, the multiplication of unipolar SBS is implemented by an AND gate, while that of bipolar SBS is implemented by an XNOR gate. Scaled addition is commonly used in SC instead of normal addition for hardware simplicity, which is implemented by a MUX with an selection SBS (4/8 in Fig. 8.1b) for both unipolar and bipolar encoding..

Due to the bit-wise computing nature, SC is robust to most hardware failures and soft errors, where limited bit false flips are tolerable. Moreover, parallelism of SC is straightforward that can be done by duplicating logic functions. Nevertheless, the hardware resource linearly increases with application precision, which is determined by SBS length. In addition, SC computation inherently has non-deterministic output, e.g., up to $3 \cdot 10^{-4}$ output variation in a 1024-bit XNOR operation. The variation is higher for operands close to 0.5, and hence can be mitigated by avoiding using such numbers [KKY16].

## 8.3 VC-MTJ and V-NDR in SC

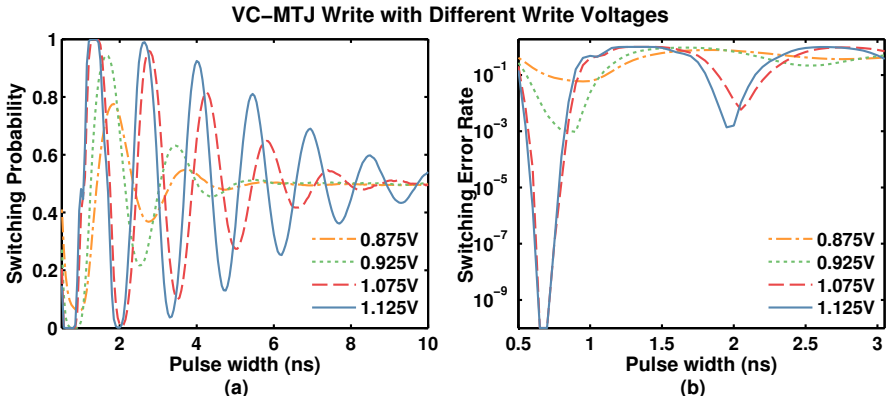



Figure 8.2: Simulated switching probability (a) and switching error rate (b) as functions of pulse width for different write voltages using an experimentally verified model in [WLE16b, GLL16a].

VC-MTJ has been well introduced in Chapter 5. With enough voltage applied across a VC-MTJ, its free layer magnetization keeps flipping quickly between two states. A successful switching alternates the MTJ state by controlling voltage pulse width to the half of switching cycle. A switching error may occur if applied voltage is not sufficient

or the pulse width mismatches the processional switching cycle [WLG16, WLE16b]. Fig. 8.2 illustrates the switching behavior of VC-MTJ. With appropriate write voltage (e.g., 1.075V to 1.125V), a switching can be completed in 700 to 800 ps pulse with error rate $< 10^{-10}$. The switching probability converges to 0.5 with long pulse, where other resistive NVM with stochastic write converge to 1, e.g., memristor and spin-transfer-torque MTJ. The convergence to 0.5 is the key for efficient SBSG design (section 8.5). The convergence is faster with lower voltage (e.g., the 5ns convergence time for 0.875-0.925V) for that damping field is not fully removed.

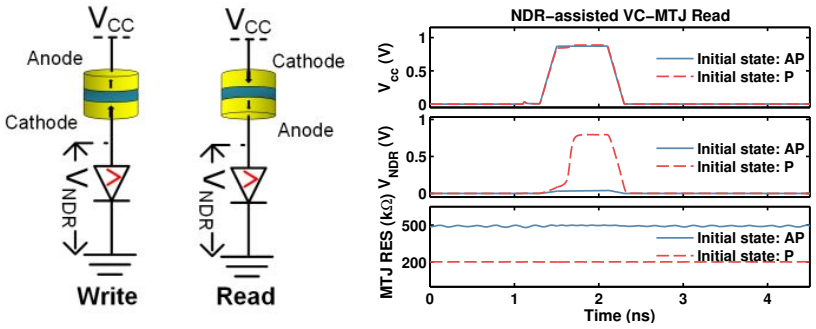
Thanks to the voltage (electric field) induced switching, VC-MTJ is generally designed with thick MgO layer, which leads to high resistance ($> 200k\Omega$) and hence reduces write leakage current and energy. Every VC-MTJ switching consumes about $1fJ$, which results in the lowest energy among existing NVM [AUA13]. Please note that, with low write current, switching effect induced by current (e.g., STT) is minimized, creating symmetric switching for $HRS \rightarrow LRS$ and $LRS \rightarrow HRS$.

A VC-MTJ read signal uses the reversed bias direction of the write (i.e., positive voltage is applied on the cathode), which increases energy barrier and device stability (Section 5.2), leading to non-destructive read [LGW16a].

In the proposed non-volatile stochastic computing (NVSC). V-NDR [WPC17] works as logic elements which directly interact with VC-MTJ based registers. An V-NDR element with a 3T CMOS circuit and corresponding I-V characteristics are introduced in Chapter 7. By placing an V-NDR element in series with a VC-MTJ and choosing $I_{peak}$ between $R_H$ and $R_L$ lines of the VC-MTJ, a high current is obtained when VC-MTJ is in HRS and a close-to-0 current is obtained otherwise. This feature has been experimentally demonstrated in Chapter 7, and it enables logic operations using MTJ in the proposed SC.

### 8.3.1 Non-destructive VC-MTJ Read with NDR

Given that a reversed voltage increases the energy barrier of VC-MTJ and stabilizes the state rather than destroys it, a high voltage ($V_{CC}$) is allowed for read. In a NDR-assisted read, the V-NDR is serially connected with the VC-MTJ as shown in Fig. 8.3a.




(a) (b)

Figure 8.3: (a) NDR-assisted switching and read. (b) Simulated waveforms of a NDR-assisted read.

We simulated the read process in Fig. 8.3b with 0.9V $V_{cc}$, the $V_{NDR}$ has full voltage difference for HRS and LRS.

### 8.3.2 Deterministic VC-MTJ Write with NDR

The VC-MTJ has symmetric switching between two resistance states, which uses same pulse in VC-MTJ based memory. To switch an VC-MTJ deterministically without NDR, a read operation is required firstly, and then the subsequent switching pulse is waived if the MTJ is already in target state [LAD15, WLE16b].

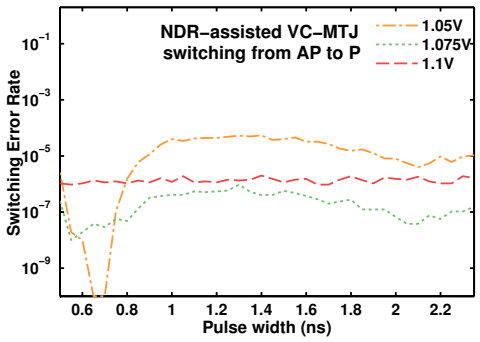



Figure 8.4: Switching error rate $(1-P_{switching})$ of NDR-assisted VC-MTJ write from HRS to LRS. The $LRS \rightarrow HRS$ switching rate is $< 10^{-10}$, which is not shown.

With the assistance of NDR, deterministic write from HRS to LRS is achieved by serially connecting VC-MTJ and V-NDR like Fig. 8.3a. The switching error rate for the deterministic write is illustrated in Fig. 8.4. Compared with Fig. 8.2, V-NDR not only

achieves single-direction switching but also relaxes the need for precise pulse width. In addition, with NDR, the switching rate from LRS to HRS is prohibited to $< 10^{-10}$ (the simulation accuracy), which is lower than the SC natural computing error rate in SC ($\sim$ 0.0001). Please note that the deterministic write has been experimentally demonstrated in Chapter 7.

## 8.4 VC-MTJ based Operations in Stochastic Computing

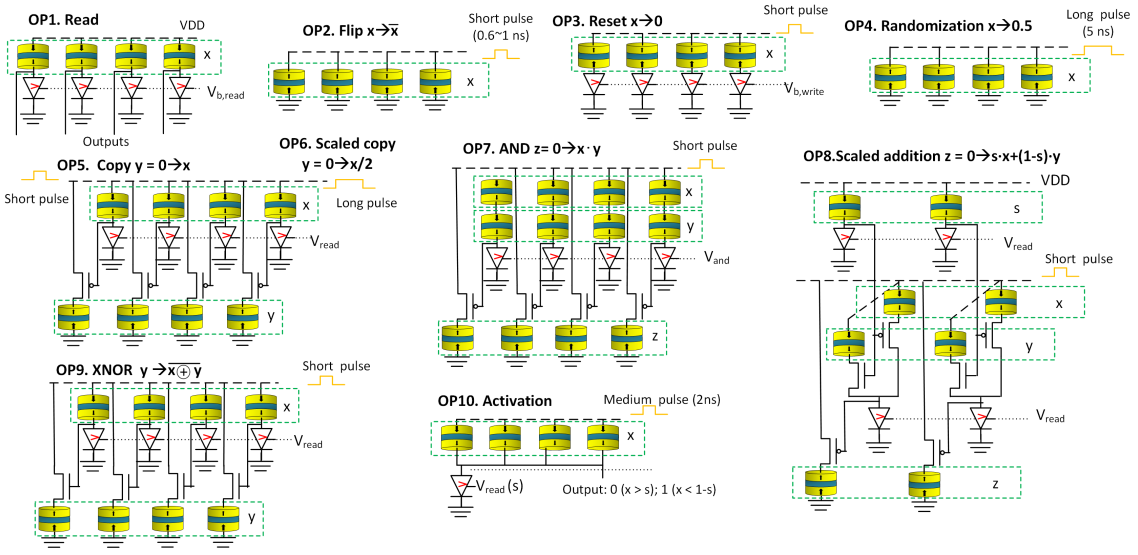



Figure 8.5: VC-MTJ and V-NDR built SC logic operations. Where a long low-voltage pulse is used to randomize VC-MTJ states, and a short high-voltage pulse is used to switch VC-MTJ states. Every VC-MTJ array stores an SBS. All VC-MTJs in an array are computed simultaneously for throughput and design efficiency purpose. Please note that the XNOR gate directly changes the array Y's data to $\overline{X \oplus Y}$.

Addition, subtraction, and multiplication are the basic stochastic logic operations [Gai69, QR08, BC01]. Other operations including division are derived from addition and multiplication, which are usually not as efficient as addition and multiplications.

The VC-MTJ SC uses the same logic operations as CMOS for SC computing, including AND for unipolar coding multiplications, MUX for scaled addition, XNOR for bipolar multiplication (see Fig. 8.1). However, our contributions are MTJ-V-NDR based SC logic gates and registers, where SBS are directly stored and computed in VC-MTJs. We have designed logic operations including AND, MUX, and XNOR, SBS generation, threshold

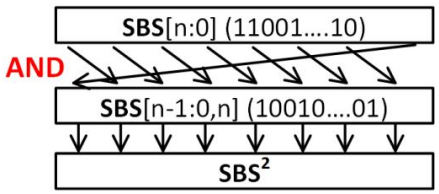
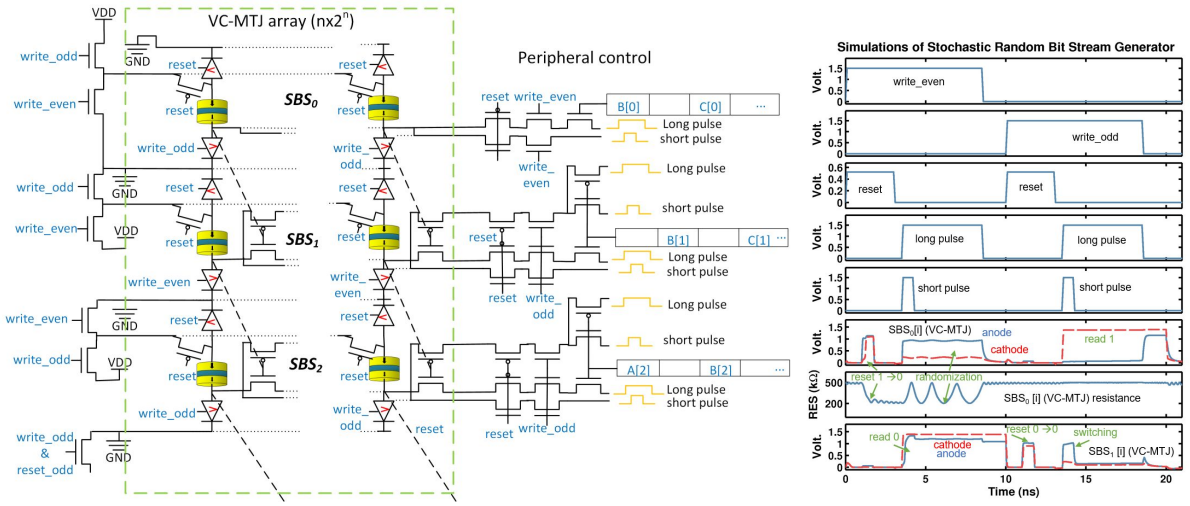



Figure 8.6: Removing correlation using shuffle operation for $SBS^2$.

function (e.g., activation used in machine learning application), and other operations allowing VC-MTJ to move data like dynamic flip flop (DFF) i.e., copy and reset. In this chapter, HRS is recognized as 1 state, while LRS is 0 state for simplicity. The designed logic operations are shown in Fig. 8.5. All operations are based on the experimentally demonstrated VC-MTJ switching [GEA16], NDR-assisted write and read (see Section 8.2).




(a) (b)

Figure 8.7: (a) The schematic of a pipe-lined SBS generator with binary fraction input A[n-1:0], B[n-1:0], and C[n-1:0]. $n$ VC-MTJ arrays (every one stores an SBS) are divided into even and odd groups based on the index. At every cycle (10ns), one group is written according to the read-out of the other group. Output SBS is generated every two cycles because of the pipe-line. (b) Simulated waveforms of two-cycle SBS generation. The MTJs in $SBS_0$ are written in the first cycle (*write_even*: high, while the MTJs in $SBS_1$ are written in the second cycle (*write_odd*: low) with the read-out from $SBS_0$.

OP 1-3 are based on simple write and read as explained in section 8.2. OP4 generates a random bit stream with 50% of "1"s, based on the observation in Fig. 5.2 that a

long write pulse leads to an VC-MTJ switching probability to 50%. The convergence to 50% probability takes about 5ns for write voltage between 0.875 V and 0.925 V. To accomplish OP 5-9, a reset must be performed firstly on the output VC-MTJ array. OP5 (OP6) combines read and flip (randomization) operations, which results in a copy (scaled copy) operation. In OP7, V-NDR is sized or biased to differentiate the highest series resistance combination from two VC-MTJs. Two serially connected 1-state VC-MTJs (in array x and y) allow V-NDR to stay at low voltage bias, thus the PMOS is turned on, and a short pulse passes to switch the corresponding output MTJ to 1. When two read VC-MTJs are in other states, V-NDR is in the valley, drops most voltage, turns off the PMOS, so that output MTJ stays 0 state. This design completes an AND operation. OP8 is a scaled addition with a selection array (array s) selecting corresponding data from array x and y to be copied to z array. OP9 is a V-NDR XNOR gate, the VC-MTJ in array y is flipped for corresponding "0" in x. OP10 is an activation function for machine learning applications. When the number of 1-state VC-MTJs is over a threshold, "1" is output. It needs a large sized V-NDR (e.g., with 2 mA peak current) for judging the parallel resistance of a VC-MTJ array. The proposed logic operations are significantly cheaper than corresponding CMOS ones. One V-NDR contains only three minimal sized transistors, whereas a CMOS DFF has 14-20 transistors.

Table 8.1: Simulated energy per bit and delay of VC-MTJ-V-NDR based logic operations. Interconnect and fan-out load is considered.

| | read | flip | reset | rand | copy |
|---|---|---|---|---|---|
| Energy (fJ) | 1.01 | 2.57 | 1.79 | 16.5 | 3.34 |
| Delay (ns) | 0.7 | 1 | 1 | 5 | 1 |
| | scaled copy | AND | addition | XNOR | activation |
| Energy (fJ) | 15.2 | 5.23 | 3.74 | 3.53 | 1.16 |
| Delay (ns) | 5 | 1.8 | 1.2 | 1 | 5 |

The VC-MTJ and V-NDR based logic operations are simulated with 60nm verified VC-MTJ model [WLE16b, GLL16a] and 45nm SOI library. The energy and delay are listed in Table 8.1.

An SC operation with two correlated SBS would result in an unwanted false output.

Thus eliminating correlations between computed SBS is important to maintain SC accuracy. This can be solved by shuffling the bits of a stochastic number. One example in Fig 8.6 shows how a simple shuffle/shift removes SBS correlation between the input and output. Please note that this correction can be done by simply changing the interconnects, which results in no hardware overhead.

## 8.5 Stochastic Bit Stream Generator

SBSG has been recognized as the bottleneck of previous SC works [KKY16, VVF15]. CMOS LFSR based generator consumes high energy, while other stochastic memory based generators, e.g., memristor [GSZ13] and STt-MTJ [BCW15], suffer from challenges of creating precise bias voltage for accurate switching probability. Utilizing the operations in Section 8.4, we propose a practical true SBSG with VC-MTJ and NDR. This SBSG does not have precision limitation and design challenge.

Binary input: IN[2:0] = ***.101*** (5/8, decimal)

| Step | Input | | Operations | Decimal value | $SBS_0$ | $SBS_1$ | $SBS_2$ |
|---|---|---|---|---|---|---|---|
| 1 | | | Reset $SBS_0$: $SBS_0[i]$ = 0 | $SBS_0$ = 0 | 00000000 | | |
| 2 | IN[0] | ***1*** | **Randomization**: $SBS_0[i]$=random | $SBS_0$ = IN[0]*1/2 | 01101001 | | |
| 3 | | | Reset $SBS_1$: $SBS_1[i]$=0 | $SBS_1$ = 0 | 01101001 | 00000000 | |
| 4 | IN[1] | ***0*** | **Scaled copy**: if $SBS_0[i]$=1, then $SBS_1[i]$=random | $SBS_1 = SBS_0/2 =$ $IN[0]*(1/2)^2 + IN[1]*(1/2)$ | 01101001 | 01001000 | |
| 5 | | | Reset $SBS_2$ | $SBS_2$ = 0 | | 01001000 | 00000000 |
| 6 | IN[2] | ***1*** | **Copy_and_rand**: if $SBS_1[i]$=1, then $SBS_2[i]$=1, else $SBS_2[i]$=random | $SBS_2 = (1-SBS_1)/2 + SBS_1$ $=IN[0]*(1/2)^3 +$ $IN[1]*(1/2)^2 + IN[2]*(1/2)$ | | 01001000 | 01101011 SBS output |

Figure 8.8: An SBS generation example. The input 0.101 (binary) is translated to SBS (01101011).

As explained in Section 8.2 and 8.4, when a long pulse is applied on VC-MTJ, its switches to 1 with 50% probability. We utilize this feature to translate n-bit binary fraction floating number IN[n-1:0] to SBS[$2^n$-1:0]. One example is illustrated in Fig. 8.8. The process starts from the least significant bit IN[0]. If IN[0] is 1, the first array $SBS_0$ is randomized to 0.5, otherwise to 0. Then upon the next bit, if IN[1] is 0, an *scaled_copy* (i.e., OP6 in Fig. 8.5, where the "1"s have 50% to be copied to the next SBS) is performed from $SBS_0$ to $SBS_1$, and otherwise a *copy_and_rand* is performed (i.e., the "1"s in $SBS_0$ are copied to $SBS_1$, the remaining "0"s in $SBS_1$ are then randomized). The *scaled_copy*

obtains half of the origin SBS number, while the *copy_and_rand* obtains half of the origin SBS number plus 0.5. In other words, the previous MTJ array is half copied to the current array, whether a 0.5 is added depends on corresponding bit in IN. This process continues and obtains output at step n (e.g., n=8 for a 256-bit SBS) The generation can be pipe-lined such that an SBS is generated at every clock cycle. In addition, there is no correlation between consequent SBS.

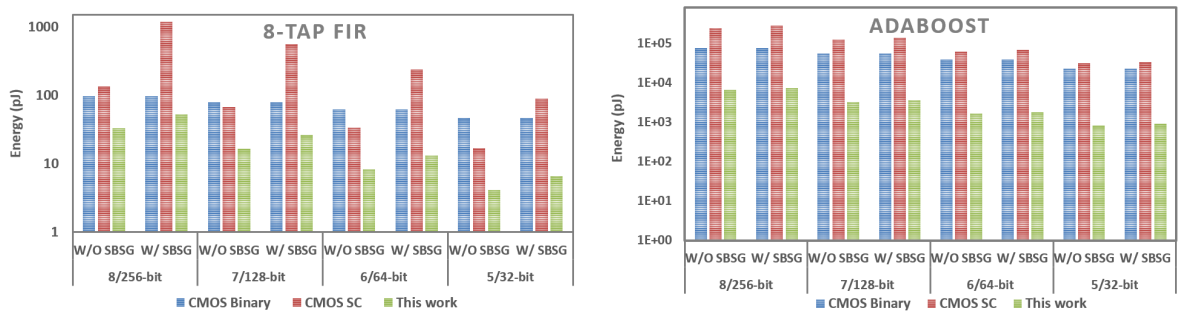
The schematic of an pipe-lined SBS generator is shown in Fig. 8.7a. Every VC-MTJ array stores an SBS. The VC-MTJ arrays are divided into even and odd groups according to the index. At every cycle, write operations are performed in one group using the data read from the other group and input binary fraction number $A, B, C$. The *write_even* and *write_odd* take turn to select group to read or write. A reset operation (controlled by *reset* signal) and an SBS array operation (*scaled_copy* and *copy_and_rand*) are performed in every clock cycle. Input $A, B, C$ are shifted in the registers in sequence. Output is generated every two clock cycles.

Two-cycle SPICE simulated waveforms are shown in Fig 8.7b. In the first cycle, *write_even* is high, and an VC-MTJs at $SBS_0$ is firstly reset and then randomized because input B[0] is 1. In the second cycle, *write_odd* is high, and a reset pulse on 0-state VC-MTJ at $SBS_1$ is prohibited by the NDR, and then the VC-MTJ is switched by a *copy* operation since its corresponding VC-MTJ at $SBS_0$ is in 1.

In the proposed SBSG, every bit generation involves $9n$ transistors, whereas the CMOS LFSR generator involves $n$ DFFs (12-20 transistors in a DFF), many computing logic gates, and a n-bit comparator (about $4n$ logic gates). Thanks to the efficiency of VC-MTJ and the generation scheme, the VC-MTJ based SBSG saves 55X energy compared with the synthesized CMOS LFSR based generator (see Section 8.6).

## 8.6 Evaluation

We evaluate the proposed VC-MTJ SC designs for FIR and Adaboost [ROM01] (a machine learning algorithm commonly used for face detection). We synthesize CMOS binary logic designs with fixed-point width from 5-bit to 8-bit and corresponding CMOS SC implementation with SBS width from 32-bit to 256-bit. CMOS LFSR based generators




(a) (b)

Figure 8.9: (a) Computing energy of 8-tap FIR for fix-point width from 5-bit to 8-bit (32-bit to 256-bit for uni-polar encoded SC). (b) Computing energy of 32-classifier Adaboost with 32-pixel input image for fix-point width from 5-bit to 8-bit (32-bit to 256-bit for bipolar encoded SC). The wire activity is 0.375 for both CMOS binary and SC designs, and 1 for VC-MTJ and V-NDR based SC design. Energy are shown for two categories: energy with SBSG (W/ SBSG) and energy without SBSG (W/O SBSG).

are also synthesized for SBS bandwidth of 32 bits to 256 bits. The VC-MTJ and V-NDR based SC are simulated using HSPICE with experimentally verified VC-MTJ model [WLE16b, GLL16a] and 45nm SOI CMOS library. As is illustrated in Fig. 8.9, VC-MTJ SC shows 3X to 25X advantageous energy-efficient compared with CMOS binary designs. SC is more efficient in low-precision designs and less efficient in high-precision designs, because SC design cost linearly scales with precision, while binary design logarithmically scales. The energy-benefit of the proposed SC against CMOS binary designs is better in Adaboost (12-25X) than in FIR (3X to 7X), because Adaboost relatively contains more adders and multipliers. The computation energy (without SBS generation) of CMOS SC is slightly lower than CMOS binary for low-precision applications. However, the advantage disappears when including the energy of the inefficient LFSR based generator, which is also observed in [KKY16]. Please note that, the comparison here is for computing one output from high-activity designs. The energy benefit of the proposed SC is expected to further increase for low-activity applications, where the non-volatility of VC-MTJ allows for immediate power gating with least energy overhead.

## 8.7 Chapter Conclusion

In this chapter, we have introduced a practical NV SC and a truly random SBSG built by VC-MTJ and NDR. The functionality of the NV SC logic gates is based on experimentally demonstrated NDR-assisted VC-MTJ write and read. The proposed SBSG design consumes 55X lower energy than CMOS LFSR based SBSG. For applications including FIR and Adaboost, the proposed SC is 3-25X and 4-37X more energy-efficient compared with CMOS binary and CMOS SC designs respectively.

# CHAPTER 9

# Conclusion

In this dissertation, we have developed several evaluation frameworks for emerging boolean logic devices, memory technologies, and integration technologies in terms of performance and reliability. In these frameworks, to identify the maximum benefits of emerging technologies, emerging technologies are firstly co-optimized with circuit, system, and application benchmarks, and then comprehensive metrics like chip clock frequency, chip power, and system failure rate are used to evaluate them. Evaluation conclusions are drawn: 1) emerging boolean devices still require more technology development to be able to replace Si technology, 2) emerging integration technologies show performance benefits at the expense of cost, 3) memory systems including main memory and cache suffer from more variability and reliability problems than digital circuits, 4) MRAM shows promising performance for main memory and cache designs.

Therefore, we have specially chosen MRAM technologies as examples for optimization in applications including memory and stochastic computing systems.

- **Energy-efficient MRAM write and read designs**: We have proposed novel system-level and circuit-level designs aiming at improving MRAM write/read efficiency. We have proposed MTJ-based variation monitor, which senses MRAM process and temperature variation. With its assistance, variation-aware adaptive write and read are proposed, reducing STT-RAM and MeRAM cache write latency by 17% and 60% respectively, and increasing STT-RAM sensing margin by 1.3X. We have also proposed novel V-NDR and C-NDR designs to assist resistive NVM write and read. The proposed NDR devices locate in the peripheral circuitry and can be shared by a bit-line of memory cells, hence introducing negligible area overhead and memory array change. With the proposed V-NDR, MRAM programming energy is saved over 50%, and read disturbance rate can be minimized by $10^9$X. With the

assistance of C-NDR and V-NDR devices, MLC resistive NVM programming can be simplified from multiple cycles in conventional write-and-verify mechanism to single cycle. To demonstrate the mentioned benefits, we have built V-NDR and experimentally verified these proposed designs with in-house 60nm VC-MTJs.

- **Non-volatile stochastic computing**: We have proposed a practical low-power stochastic computing system using VC-MTJ and V-NDR. In this computing system, data (in the form of stochastic bitstream) are stored and manipulated in VC-MTJs with the assistance of V-NDR. Stochastic bitstream is truly randomly generated using the proposed VC-MTJ based generator. This generator does not have precision bound and is robust to process and temperature variation. Moreover, it consumes 55X lower energy than CMOS LFSR generator. Overall, the proposed stochastic computing has over 3X energy advantage compared with CMOS binary computing and CMOS stochastic computing for applications like FIR and Adaboost. Please note that state-of-art experimental VC-MTJs are qualified for obtaining the predicted performance benefits of the proposed stochastic computing.

# References


[AAU12] JG Alzate, P Khalili Amiri, P Upadhyaya, SS Cherepov, J Zhu, M Lewis, R Dorrance, JA Katine, J Langer, K Galatsis, et al. "Voltage-induced switching of nanoscale magnetic tunnel junctions." In *Proc. IEDM*, pp. 29–5. IEEE, 2012.

[AAY14] Juan G Alzate, Pedram Khalili Amiri, Guoqiang Yu, Pramey Upadhyaya, Jordan A Katine, Juergen Langer, Berthold Ocker, Ilya N Krivorotov, and Kang L Wang. "Temperature dependence of the voltage-controlled perpendicular anisotropy in nanoscale MgO— CoFeB— Ta magnetic tunnel junctions." *Appl. Phys. Lett.*, **104**(11):112410, 2014.

[AGH12] Fabien Alibart, Ligang Gao, Brian D Hoskins, and Dmitri B Strukov. "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm." *Nanotechnology*, **23**(7):075201, 2012.

[AKB03] Asen Asenov, Savas Kaya, and Andrew R Brown. "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness." *Electron Devices, IEEE Transactions on*, **50**(5):1254–1260, 2003.

[AKW13] Dmytro Apalkov, Alexey Khvalkovskiy, Steven Watts, Vladimir Nikitin, Xueti Tang, Daniel Lottis, Kiseok Moon, Xiao Luo, Eugene Chen, Adrian Ong, et al. "Spin-transfer torque magnetic random access memory (STT-MRAM)." *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, **9**(2):13, 2013.

[Ant] S. Anthony. "Beyond Silicon: IBM Unveils Worlds First 7 nm chipWith a Silicon-Germanium Channel and EUV Lithography, IBM Crosses the 10 nm Barrier. Ars Technica." `http://arstechnica.com/gadgets/2015/07/ibm-unveils-industrys-first-7nm-chip-moving-beyond-silicon`. Accessed: 2016-10-04.

[APM09] André L Aita, Michiel AP Pertijs, Kofi AA Makinwa, and Johan H Huijsing. "A CMOS smart temperature sensor with a batch-calibrated inaccuracy of$\pm$0.25 C ($3\sigma$) from- 70 C to 130 C." In *ISSCC*, pp. 342–343. IEEE, 2009.

[ARG09] Charles Augustine, Arijit Raychowdhury, Yunfei Gao, Mark Lundstrom, and Kaushik Roy. "PETE: A device/circuit analysis framework for evaluation and comparison of charge based emerging devices." In *Quality of Electronic Design, 2009. ISQED 2009. Quality Electronic Design*, pp. 80–85. IEEE, 2009.

[AUA13] P Khalili Amiri, P Upadhyaya, JG Alzate, and KL Wang. "Electric-field-induced thermally assisted switching of monodomain magnetic bits." *J. Appl. Phys.*, **113**(1):013912, 2013.

[BBB11] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R Hower, Tushar Krishna, Somayeh Sardashti, et al. "The gem5 simulator." *ACM SIGARCH Computer Architecture News*, **39**(2):1–7, 2011.

[BC01] Bradley D Brown and Howard C Card. "Stochastic neural computation. I. Computational elements." *IEEE Transactions on computers*, **50**(9):891–905, 2001.

[BCW15] Lirida Alves de Barros Naviner, Hao Cai, You Wang, Weisheng Zhao, and Arwa Ben Dhia. "Stochastic computation with Spin Torque Transfer Magnetic Tunnel Junction." In *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*, pp. 1–4. IEEE, 2015.

[BDM02] Keith Bowman, Steven G Duvall, James D Meindl, et al. "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration." *Solid-State Circuits, IEEE Journal of*, **37**(2):183–190, 2002.

[BDR07] Emanuele Baravelli, Abhisek Dixit, Rita Rooyackers, Malgorzata Jurczak, Nicolò Speciale, and Kristin De Meyer. "Impact of line-edge roughness on FinFET matching performance." *Electron Devices, IEEE Transactions on*, **54**(9):2466–2474, 2007.

[BEO14] Rajendra Bishnoi, Mojtaba Ebrahimi, Fabian Oboril, and Mehdi B Tahoori. "Asynchronous asymmetrical write termination (AAWT) for a low power STT-MRAM." In *Proceedings of the conference on Design, Automation & Test in Europe*, p. 180. European Design and Automation Association, 2014.

[BFR09] Ferdinando Bedeschi, Rich Fackenthal, Claudio Resta, Enzo Michele Donze, Meenatchi Jagasivamani, Egidio Cassiodoro Buda, Fabio Pellizzer, David W Chow, Alessandro Cabrini, Giacomo Matteo Angelo Calvi, et al. "A bipolar-selected phase change memory featuring multi-level cell storage." *IEEE Journal of Solid-State Circuits*, **44**(1):217–227, 2009.

[BGK16] F Merrikh Bayat, X Guo, M Klachko, N Do, K Likharev, and D Strukov. "Model-based high-precision tuning of NOR flash memory cells for analog computing applications." In *Device Research Conference (DRC), 2016 74th Annual*, pp. 1–2. IEEE, 2016.

[BGM88a] Mario Blaum, Rodney Goodman, and Robert McEliece. "The reliability of single-error protected computer memories." *Computers, IEEE Transactions on*, **37**(1):114–119, 1988.

[BGM88b] Mario Blaum, Rodney Goodman, and Robert McEliece. "The reliability of single-error protected computer memories." *Computers, IEEE Transactions on*, **37**(1):114–119, 1988.

[BHS15] F. Merrikh Bayat, B. Hoskins, and D.B. Strukov. "Phenomenological modeling of memristive devices." *Applied Physics A*, **118**(3):779–786, 2015.

[BKM07] Steven M Burns, Mahesh Ketkar, Noel Menezes, Keith Bowman, James W Tschanz, Vivek De, et al. "Comparative analysis of conventional and statistical design techniques." In *Design Automation Conference, 2007. DAC'07. 44th ACM/IEEE*, pp. 238–243. IEEE, 2007.

[BMS11] F Bonell, S Murakami, Y Shiota, T Nozaki, T Shinjo, and Y Suzuki. "Large change in perpendicular magnetic anisotropy induced by an electric field in FePd ultrathin films." *Appl. Phys. Lett.*, **98**(23):232510, 2011.

[BOE16] R. Bishnoi, F. Oboril, M. Ebrahimi, and M. B. Tahoori. "Self-Timed Read and Write Operations in STT-MRAM." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, **24**(5):1783–1793, May 2016.

[BOO06] Jonathan P Benson, Tony O'Donovan, Padraig O'Sullivan, Utz Roedig, Cormac Sreenan, John Barton, Aoife Murphy, and Brendan O'Flynn. "Car-park management using wireless sensor networks." In *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*, pp. 588–595. IEEE, 2006.

[Bre78] John R Brews. "A charge-sheet model of the MOSFET." *Solid-State Electronics*, **21**(2):345–355, 1978.

[BS95] David Burnett and Shih-Wei Sun. "Statistical threshold-voltage variation and its impact on supply-voltage scaling." In *Microelectronic Manufacturing'95*, pp. 83–90. International Society for Optics and Photonics, 1995.

[BSH04] Keith Bowman, Samie B Samaan, Nagib Z Hakim, et al. "Maximum clock frequency distribution model with practical VLSI design considerations." In *Integrated Circuit Design and Technology, 2004. ICICDT'04. International Conference on*, pp. 183–191. IEEE, 2004.

[BSS08] Xin-Yu Bao, Cesare Soci, Darija Susac, Jon Bratvold, David PR Aplin, Wei Wei, Ching-Yang Chen, Shadi A Dayeh, Karen L Kavanagh, and Deli Wang. "Heteroepitaxial growth of vertical GaAs nanowires on Si (111) substrates by metal- organic chemical vapor deposition." *Nano letters*, **8**(11):3755–3760, 2008.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CAD10] E Chen, D Apalkov, Z Diao, A Driskill-Smith, D Druist, D Lottis, V Nikitin, X Tang, S Watts, S Wang, et al. "Advances and future prospects of spin-transfer torque random access memory." *Magnetics, IEEE Transactions on*, **46**(6):1873–1878, 2010.

[CCL08] By Benton H Calhoun, Yu Cao, Xin Li, Ken Mai, Lawrence T Pileggi, Rob Rutenbar, Kenneth L Shepard, et al. "Digital circuit design challenges and opportunities in the era of nanoscale CMOS." *Proceedings of the IEEE*, **96**(2):343–365, 2008.

[CCP10] Poki Chen, Chun-Chi Chen, Yu-Han Peng, Kai-Ming Wang, and Yu-Shin Wang. "A time-domain SAR smart temperature sensor with curvature compensation and a $3\sigma$ inaccuracy of- 0.4 C + 0.6 C over a 0 C to 90 C range." *JSSC*, **45**(3):600–609, 2010.

[CKL09] Geunho Cho, Yong-Bin Kim, and Fabrizio Lombardi. "Assessment of CNT-FET based circuit performance and robustness to PVT variations." In *Circuits and Systems, 2009. MWSCAS'09. 52nd IEEE International Midwest Symposium on*, pp. 1106–1109. IEEE, 2009.

[CLD13] Matthew Cotter, Huichu Liu, Soupayan Datta, and Vijaykrishnan Narayanan. "Evaluation of tunnel FET-based flip-flop designs for low power, high performance applications." In *Quality electronic design (ISQED), 2013 14th international symposium on*, pp. 430–437. IEEE, 2013.

[Com11] International Roadmap Committee et al. "International Technology Roadmap for Semiconductors, 2011 Edition." *Semiconductor Industry Association, http://www. itrs. net/Links/2011ITRS/2011Chapters/2011ExecSum. pdf*, 2011.

[CRK10] Suock Chung, K-M Rho, S-D Kim, H-J Suh, D-J Kim, HJ Kim, SH Lee, J-H Park, H-M Hwang, S-M Hwang, et al. "Fully integrated 54nm STT-RAM with the smallest bit cell dimension for high density memory application." In *Proc. IEDM*, pp. 12–7. IEEE, 2010.

[CS05] Hongliang Chang and Sachin S Sapatnekar. "Full-chip analysis of leakage power under process variations, including spatial correlations." In *Proceedings of the 42nd annual Design Automation Conference*, pp. 523–528. ACM, 2005.

[CSK12] Chi On Chui, Kyeong-Sik Shin, Jorge Kina, Kun-Huan Shih, Pritish Narayanan, and C Andras Moritz. "Heterogeneous integration of epitaxial nanostructures: strategies and application drivers." In *SPIE NanoScience+ Engineering*, pp. 84670R–84670R. International Society for Optics and Photonics, 2012.

[CTC10] Chia-Tsung Cheng, Yu-Chang Tsai, and Kuo-Hsing Cheng. "A high-speed current mode sense amplifier for Spin-Torque Transfer Magnetic Random Access Memory." In *Circuits and Systems (MWSCAS), 2010 53rd IEEE International Midwest Symposium on*, pp. 181–184. IEEE, 2010.

[CWS00] EY Chen, R Whig, JM Slaughter, D Cronk, J Goggin, G Steiner, and S Tehrani. "Comparison of oxidation methods for magnetic tunnel junction material." *J. Appl. Phys.*, **87**(9):6061–6063, 2000.

[CWZ10] Yiran Chen, Xiaobin Wang, Wenzhong Zhu, Wei Xu, Tong Zhang, et al. "A nondestructive self-reference scheme for spin-transfer torque random access memory (STT-RAM)." In *Proc. DATE*, pp. 148–153. IEEE, 2010.

[CY11] Ching-Che Chung and Cheng-Ruei Yang. "An autocalibrated all-digital temperature sensor for on-chip thermal monitoring." *TCS*, **58**(2):105–109, 2011.

[DAC13] Richard Dorrance, Juan G Alzate, Sergiy S Cherepov, Pramey Upadhyaya, Ilya N Krivorotov, Jordan A Katine, Juergen Langer, Kang L Wang, Pedram Khalili Amiri, and Dejan Markovic. "Diode-MTJ Crossbar Memory Cell Using Voltage-Induced Unipolar Switching for High-Density MRAM." *EDL*, **34**(6):753–755, 2013.

[DDM98] Jeffery A Davis, Vivek K De, and James D Meindl. "A stochastic wire-length distribution for gigascale integration (GSI)–part II: Applications to clock frequency, power dissipation, and chip size estimation." *IEEE Transactions on Electron Devices*, **45**(3):590–597, 1998.

[Del97a] Timothy J Dell. "A white paper on the benefits of chipkill-correct ECC for PC server main memory." *IBM Microelectronics Division*, pp. 1–23, 1997.

[Del97b] Timothy J Dell. "A white paper on the benefits of chipkill-correct ECC for PC server main memory." *IBM Microelectronics Division*, pp. 1–23, 1997.

[DJK15] Henry Duwe, Xun Jian, and Rakesh Kumar. "Correction prediction: Reducing error correction latency for on-chip memories." In *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*, pp. 463–475. IEEE, 2015.

[DKL13] Nattapol Damrongplasit, Sung Hwan Kim, and Tsu-Jae King Liu. "Study of random dopant fluctuation induced variability in the raised-ge-source TFET." *IEEE Electron Device Letters*, **34**(2):184–186, 2013.

[DRT12] Richard Dorrance, Fengbo Ren, Yuta Toriyama, Amr Amin Hafez, C-KK Yang, and Dejan Markovic. "Scalability and design-space analysis of a 1T-1MTJ memory cell for STT-RAMs." *TED*, **59**(4):878–887, 2012.

[DSS06] Renu W Dave, Gerald Steiner, JM Slaughter, JJ Sun, B Craigo, S Pietambaram, K Smith, G Grynkewich, M DeHerrera, J Akerman, et al. "MgO-based tunnel junction material for high-speed toggle magnetic random access memory." *Magnetics, IEEE Transactions on*, **42**(8):1935–1939, 2006.

[DST08] Volker Drewello, J Schmalhorst, Andy Thomas, and Günter Reiss. "Evidence for strong magnon contribution to the TMR temperature dependence in MgO based tunnel junctions." *Physical Review B*, **77**(1):014440, 2008.

[DXX12] Xiangyu Dong, Cong Xu, Yuan Xie, and Norman P Jouppi. "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, **31**(7):994–1007, 2012.

[EDS01] K Eberl, R Duschl, OG Schmidt, U Denker, and R Haug. "Si-based resonant inter-and intraband tunneling diodes." *Journal of crystal Growth*, **227**:770–776, 2001.

[EJL14] Y Eckert, Nuwan Jayasena, and G Loh. "Thermal feasibility of die-stacked processing in memory." In *Proceedings of the 2nd Workshop on Near-Data Processing*, 2014.

[EZW14] Enes Eken, Yaojun Zhang, Wujie Wen, Rajiv Joshi, Hai Li, and Yiran Chen. "A New Field-assisted Access Scheme of STT-RAM with Self-reference Capability." In *Proceedings of the 51st Annual Design Automation Conference*, pp. 1–6. ACM, 2014.

[FHS06] David J Frank, Wilfried Haensch, Ghavam Shahidi, and Omer H Dokumaci. "Optimizing CMOS technology for maximum performance." *IBM journal of research and development*, **50**(4.5):419–431, 2006.

[FIT11] "Failure rate." `https://en.wikipedia.org/wiki/Failure_rate`, 2008,2011.

[FKB05] GD Fuchs, IN Krivorotov, PM Braganca, NC Emley, AGF Garcia, DC Ralph, and RA Buhrman. "Adjustable spin torque in magnetic tunnel junctions with two fixed layers." *Appl. Phys. Lett.*, **86**(15):152509, 2005.

[FW08] Richard F Freitas and Winfried W Wilcke. "Storage-class memory: The next storage system technology." *IBM J RES DEV*, **52**(4.5):439–447, 2008.

[Gai69] Brian R Gaines. "Stochastic computing systems." In *Advances in information systems science*, pp. 37–172. Springer, 1969.

[GBP17] X Guo, F Merrikh Bayat, M Prezioso, Y Chen, B Nguyen, N Do, and DB Strukov. "Temperature-Insensitive Analog Vector-by-Matrix Multiplier Based on 55 nm NOR Flash Memory Cells." *CICC'17*, 2017.

[GEA16] C Grezes, F Ebrahimi, JG Alzate, X Cai, JA Katine, J Langer, B Ocker, P Khalili Amiri, and KL Wang. "Ultra-low switching energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product." *Applied Physics Letters*, **108**(1):012403, 2016.

[GG12] Rani S Ghaida and Puneet Gupta. "DRE: A framework for early co-evaluation of design rules, technology choices, and layout methodologies." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, **31**(9):1379–1392, 2012.

[GLL16a] C. Grezes, H. Lee, A. Lee, S. Wang, F. Ebrahimi, X. Li, K. Wong, J. A. Katine, B. Ocker, J. Langer, P. Gupta, P. Khalili, and K. L. Wang. "Write Error Rate and Read Disturbance in Electric-Field-Controlled MRAM." *IEEE Magnetics Letters*, 2016.

[GLL16b] Cecile Grezes, Hochul Lee, Albert Lee, Shaodi Wang, Farbod Ebrahimi, Xiang Li, Kin Wong, Jordan A Katine, Berthold Ocker, Juergen Langer, et al. "Write Error Rate and Read Disturbance in Electric-Field-Controlled MRAM." *IEEE Magnetics Letters*, 2016.

[GMG15] Xinjie Guo, Farnood Merrikh-Bayat, Ligang Gao, Brian D Hoskins, Fabien Alibart, Bernabe Linares-Barranco, Luke Theogarajan, Christof Teuscher, and Dmitri B Strukov. "Modeling and Experimental Demonstration of a Hopfield Network Analog-to-Digital Converter with Hybrid CMOS/Memristor Circuits." *Frontiers in neuroscience*, **9**, 2015.

[GSD16] Mark Gottscho, Clayton Schoeny, Lara Dolecek, and Puneet Gupta. "Software-defined error-correcting codes." In *Dependable Systems and Networks Workshop, 2016 46th Annual IEEE/IFIP International Conference on*, pp. 276–282. IEEE, 2016.

[GSZ13] Siddharth Gaba, Patrick Sheridan, Jiantao Zhou, Shinhyun Choi, and Wei Lu. "Stochastic memristive devices for computing and neuromorphic applications." *Nanoscale*, **5**(13):5872–5878, 2013.

[GWM10] Xinjie Guo, Shaodi Wang, Chenyue Ma, Chenfei Zhang, Xinnan Lin, Wen Wu, Frank He, Wenping Wang, Zhiwei Liu, Wei Zhao, et al. "A novel approach to simulate Fin-width Line Edge Roughness effect of FinFET performance." In

*Electron Devices and Solid-State Circuits (EDSSC), 2010 IEEE International Conference of*, pp. 1–4. IEEE, 2010.

[Hei01] C Heide. "Spin currents in magnetic films." *Phys. Rev. Lett.*, **87**(19):197201, 2001.

[HGV06] Wei Huang, Shougata Ghosh, Siva Velusamy, Karthik Sankaranarayanan, Kevin Skadron, and Mircea R Stan. "HotSpot: A compact thermal modeling methodology for early-stage VLSI design." *TVLSI*, **14**(5):501–513, 2006.

[HHS10] David Halupka, Safeen Huda, Wanjuan Song, Ali Sheikholeslami, Koji Tsunoda, Chikako Yoshida, and Masaki Aoki. "Negative-resistance read and write schemes for STT-MRAM in 0.13$\mu$m CMOS." In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pp. 256–257. IEEE, 2010.

[HKA08] BS Haran, A Kumar, L Adam, J Chang, V Basker, S Kanakasabapathy, D Horak, S Fan, J Chen, J Faltermeier, et al. "22 nm technology compatible fully functional 0.1 $\mu$m 2 6T-SRAM cell." In *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4. IEEE, 2008.

[Hua08] Yiming Huai. "Spin-transfer torque MRAM (STT-MRAM): Challenges and prospects." *AAPPS Bulletin*, **18**(6):33–40, 2008.

[HVY06] Tian He, Pascal Vicaire, Ting Yan, Liqian Luo, Lin Gu, Gang Zhou, Radu Stoleru, Qing Cao, John Stankovic, Tarek Abdelzaher, et al. "Achieving real-time target tracking usingwireless sensor networks." In *Real-Time and Embedded Technology and Applications Symposium, 2006. Proceedings of the 12th IEEE*, pp. 37–48. IEEE, 2006.

[HYO05] Y Higo, K Yamane, K Ohba, H Narisawa, K Bessho, M Hosomi, and H Kano. "Thermal activation effect on spin transfer switching in magnetic tunnel junctions." *Appl. Phys. Lett.*, **87**(8):082502–082502, 2005.

[HYY05] 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 *Proc. IEDM*, pp. 459–462. IEEE, 2005.

[IMK06] T Ishikawa, T Marukame, H Kijima, K-I Matsuda, T Uemura, M Arita, and M Yamamoto. "Spin-dependent tunneling characteristics of fully epitaxial magnetic tunneling junctions with a full-Heusler alloy Co 2 Mn Si thin film and a MgO tunnel barrier." *Applied physics letters*, **89**(19):192505, 2006.

[IR11] Adrian M Ionescu and Heike Riel. "Tunnel field-effect transistors as energy-efficient electronic switches." *Nature*, **479**(7373):329–337, 2011.

[ITR11] "ITRS." `http://www.itrs.net/about.html`, 2008,2011.

[JDB13] Xun Jian, N. Debardeleben, S. Blanchard, V. Sridharan, and R. Kumar. "Analyzing Reliability of Memory Sub-systems with Double-Chipkill Detect/Correct." In *Dependable Computing (PRDC), 2013 IEEE 19th Pacific Rim International Symposium on*, pp. 88–97, Dec 2013.

[JK13] Xun Jian and Ravindra Kumar. "Adaptive reliability chipkill correct (ARCC)." In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*, pp. 270–281. IEEE, 2013.

[JKN14] Sung Hyun Jo, Tanmay Kumar, Sundar Narayanan, Wei D Lu, and Hagop Nazarian. "3D-stackable crossbar resistive memory based on field assisted superlinear threshold (FAST) selector." In *Electron Devices Meeting (IEDM), 2014 IEEE International*, pp. 6–7. IEEE, 2014.

[JLP10] Kanghoon Jeon, Wei-Yip Loh, Pratik Patel, Chang Yong Kang, Jungwoo Oh, Anupama Bowonder, Chanro Park, CS Park, Casey Smith, Prashant Majhi, et al. "Si tunnel transistors with a novel silicided source and 46mV/dec swing." In *VLSI technology (VLSIT), 2010 symposium on*, pp. 121–122. IEEE, 2010.

[JMX12] Adwait Jog, Asit K Mishra, Cong Xu, Yuan Xie, Vijaykrishnan Narayanan, Ravishankar Iyer, and Chita R Das. "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs." In *Proc. DAC*, pp. 243–252. ACM, 2012.

[JZK16] Yanfeng Jiang, Yisong Zhang, Angeline Klemm, and Jian-Ping Wang. "Fast Spintronic Thermal Sensor for IC Power Driver Cooling Down." In *Proc. IEDM*, 2016.

[KAG07] UK Klostermann, M Angerbauer, U Griming, F Kreupl, M Ruhrig, F Dahmani, M Kund, and G Muller. "A perpendicular spin torque switching based MRAM for the 28 nm technology node." In *Proc. IEDM*, pp. 187–190. IEEE, 2007.

[KKA08] Hei Kam, Tsu-Jae King-Liu, Elad Alon, and Mark Horowitz. "Circuit-level requirements for MOSFET-replacement devices." In *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–1. IEEE, 2008.

[KKS13] Emre Kultursay, Mahmut Kandemir, Anand Sivasubramaniam, and Onur Mutlu. "Evaluating STT-RAM as an energy-efficient main memory alternative." In *ISPASS*, pp. 256–267. IEEE, 2013.

[KKY16] Kyounghoon Kim, Jungki Kim, Joonsang Yu, Jungwoo Seo, Jongeun Lee, and Kiyoung Choi. "Dynamic energy-accuracy trade-off using stochastic computing in deep neural networks." In *Proceedings of the 53rd Annual Design Automation Conference*, p. 124. ACM, 2016.

[KLK14] Wang Kang, Zheng Li, Jacques-Olivier Klein, Yuanqing Chen, Youguang Zhang, Dafiné Ravelosona, Claude Chappert, and Weisheng Zhao. "Variation-tolerant and disturbance-free sensing circuit for deep nanometer STT-MRAM." *IEEE Transactions on Nanotechnology*, **13**(6):1088–1092, 2014.

[KWN12] Eric Karl, Yih Wang, Yong-Gee Ng, Zheng Guo, Fatih Hamzaoglu, Uddalak Bhattacharya, Kevin Zhang, Kaizad Mistry, and Mark Bohr. "A 4.6 GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active V MIN-enhancing assist circuitry." In *Solid-State Circuits Conference*

*Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 230–232. IEEE, 2012.

[KYI12] S Kanai, M Yamanouchi, S Ikeda, Y Nakatani, F Matsukura, and H Ohno. "Electric field-induced magnetization reversal in a perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction." *Appl. Phys. Lett.*, **101**(12):122403, 2012.

[KZK15] Wang Kang, Liuyang Zhang, Jacques-Olivier Klein, Youguang Zhang, Dafiné Ravelosona, and Weisheng Zhao. "Reconfigurable codesign of STT-MRAM under process variations in deeply scaled technology." *IEEE TED*, **62**(6):1769–1777, 2015.

[KZZ15] Wang Kang, Liuyang Zhang, Weisheng Zhao, Jacques-Olivier Klein, Youguang Zhang, Dafiné Ravelosona, and Claude Chappert. "Yield and reliability improvement techniques for emerging nonvolatile STT-MRAM." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, **5**(1):28–39, 2015.

[LAA09] Chi-Woo Lee, Aryan Afzalian, Nima Dehdashti Akhavan, Ran Yan, Isabelle Ferain, and Jean-Pierre Colinge. "Junctionless multigate field-effect transistor." *Applied Physics Letters*, **94**(5):053511, 2009.

[LAD15] H. Lee, J.G. Alzate, R. Dorrance, X.Q. Cai, D. Markovic, P. Khalili Amiri, and K.L. wang. "Design of a Fast and Low-Power Sense Amplifier and Writing Circuit for High-Speed MRAM." *TMAG*, **51**(5):1–7, May 2015.

[LAS08] Jing Li, Charles Augustine, Sayeef Salahuddin, and Kaushik Roy. "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement." In *Proc. DAC*, pp. 278–283. ACM/IEEE, 2008.

[LAS09] Sheng Li, Jung Ho Ahn, Richard D Strong, Jay B Brockman, Dean M Tullsen, and Norman P Jouppi. "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures." In *MICRO*, pp. 469–480. IEEE, 2009.

[LC11] Greg Leung and Chi On Chui. "Variability of inversion-mode and junctionless FinFETs due to line edge roughness." *Electron Device Letters, IEEE*, **32**(11):1489–1491, 2011.

[LC12] Greg Leung and Chi On Chui. "Variability impact of random dopant fluctuation on nanoscale junctionless FinFETs." *Electron Device Letters, IEEE*, **33**(6):767–769, 2012.

[LC13a] Greg Leung and Chi On Chui. "Interactions between line edge roughness and random dopant fluctuation in nonplanar field-effect transistor variability." *Electron Devices, IEEE Transactions on*, **60**(10):3277–3284, 2013.

[LC13b] Greg Leung and Chi On Chui. "Stochastic variability in silicon double-gate lateral tunnel field-effect transistors." *Electron Devices, IEEE Transactions on*, **60**(1):84–91, 2013.

[LC13c] Greg Leung and Chi On Chui. "Stochastic variability in silicon double-gate lateral tunnel field-effect transistors." *IEEE Transactions on Electron Devices*, **60**(1):84–91, 2013.

[LDN13] Huichu Liu, Suman Datta, and Vijaykrishnan Narayanan. "Steep switching tunnel FET: A promise to extend the energy efficient roadmap for post-CMOS digital and analog/RF applications." In *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 145–150. IEEE Press, 2013.

[LFA10] Chi-Woo Lee, Isabelle Ferain, Aryan Afzalian, Ran Yan, Nima Dehdashti Akhavan, Pedram Razavi, and Jean-Pierre Colinge. "Performance estimation of junctionless multigate transistors." *Solid-State Electronics*, **54**(2):97–103, 2010.

[LGW16a] H. Lee, C. Grzes, S. Wang, F. Ebrahimi, P. Gupta, P. K. Amiri, and K. L. Wang. "Source Line Sensing in Magneto-Electric Random-Access Memory to Reduce Read Disturbance and Improve Sensing Margin." *IEEE Magnetics Letters*, **7**:1–5, 2016.

[LGW16b] H. Lee, C. Grzes, S. Wang, F. Ebrahimi, P. Gupta, P. K. Amiri, and K. L. Wang. "Source Line Sensing in Magneto-Electric Random-Access Memory to Reduce Read Disturbance and Improve Sensing Margin." *IEEE Magnetics Letters*, **7**:1–5, 2016.

[LGW16c] Hochul Lee, Cécile Grèzes, Shaodi Wang, Farbod Ebrahimi, Puneet Gupta, Pedram Khalili Amiri, and Kang L Wang. "Source line sensing in magneto-electric random-access memory to reduce read disturbance and improve sensing margin." *IEEE Magnetics Letters*, **7**:1–5, 2016.

[LHC02] Jae-Duk Lee, Sung-Hoi Hur, and Jung-Dal Choi. "Effects of floating-gate interference on NAND flash memory cell operation." *IEEE Electron Device Letters*, **23**(5):264–266, 2002.

[LLA11] Mathieu Luisier, Mark Lundstrom, Dimitri Antoniadis, Jeffrey Bokor, et al. "Ultimate device scaling: Intrinsic performance comparisons of carbon-based, InGaAs, and Si field-effect transistors for 5 nm gate length." In *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 11–2. IEEE, 2011.

[LLG12] Greg Leung, Liangzhen Lai, Puneet Gupta, and Chi On Chui. "Device-and circuit-level variability caused by line edge roughness for sub-32-nm FinFET technologies." *Electron Devices, IEEE Transactions on*, **59**(8):2057–2063, 2012.

[LLS08] Jing Li, Haixin Liu, Sayeef Salahuddin, and Kaushik Roy. "Variation-tolerant Spin-Torque Transfer (STT) MRAM array for yield enhancement." In *Proc. CICC*, pp. 193–196. IEEE, 2008.

[LLW17] Hochul Lee, Albert Lee, Shaodi Wang, Farbod Ebrahimi, Puneet Gupta, Pedram Khalili Amiri, and Kang L Wang. "A Word Line Pulse Circuit Technique for Reliable Magnetoelectric Random Access Memory." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2017.

[LLZ12] Rui Li, Yeqing Lu, Guangle Zhou, Qingmin Liu, Soo Doo Chae, T. Vasen, Wan Sik Hwang, Qin Zhang, P. Fay, T. Kosel, M. Wistey, Huili Xing, and A. Seabaugh. "AlGaSb/InAs Tunnel Field-Effect Transistor With On-Current of 78 $\mu$A/$\mu$m at 0.5 V." *Electron Device Letters, IEEE*, **33**(3):363–365, March 2012.

[LRD05] KJ Lee, Olivier Redon, and Bernard Dieny. "Analytical investigation of spin-transfer dynamics using a perpendicular-to-plane polarizer." *Appl. Phys. Lett.*, **86**(2):022505, 2005.

[LWP15a] G. Leung, S. Wang, A. Pan, P. Gupta, and C.O. Chui. "An Evaluation Framework for Nanotransfer Printing-Based Feature-Level Heterogeneous Integration in VLSI Circuits." *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, **PP**(99):1–13, 2015.

[LWP15b] Greg Leung, Shaodi Wang, Andrew Pan, Puneet Gupta, and Chi On Chui. "An Evaluation Framework for Nanotransfer Printing-Based Feature-Level Heterogeneous Integration in VLSI Circuits." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, **24**(5):1858–1870, 2015.

[MBR82] WF Mikhail, RW Bartoldus, and RA Rutledge. "The reliability of memory with single-error correction." *IEEE Transactions on Computers*, **31**(6):560–564, 1982.

[MGK15] Seyedhamidreza Motaman, Swaroop Ghosh, and Jaydeep P Kulkarni. "A novel slope detection technique for robust STTRAM sensing." In *Proc. ISLPED*, pp. 7–12. IEEE, 2015.

[MKS10] Noriyuki Miura, Kazutaka Kasuga, Mitsuko Saito, and Tadahiro Kuroda. "An 8Tb/s 1pJ/b 0.8 mm2/Tb/s QDR Inductive-Coupling Interface Between 65nm CMOS GPU and 0.1 $\mu$m DRAM." In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pp. 436–437. IEEE, 2010.

[MWK15] Justin Meza, Qiang Wu, Sanjeev Kumar, and Onur Mutlu. "Revisiting memory errors in large-scale production data centers: Analysis and modeling of new trends from the field." In *Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on*, pp. 415–426. IEEE, 2015.

[NA07] OM Nayfeh and DA Antoniadis. *Calibrated hydrodynamic simulation of deeply-scaled well-tempered nanowire field effect transistors*. Springer, 2007.

[NFL08] Nhat Nguyen, Yohan Frans, Brian Leibowitz, Simon Li, Reza Navid, Marko Aleksic, Fred Lee, Fredy Quan, Jared Zerbe, Rich Perego, et al. "A 16-Gb/s differential I/O cell with 380fs RJ in an emulated 40nm DRAM process." In *VLSI Circuits, 2008 IEEE Symposium on*, pp. 128–129. IEEE, 2008.

[NOL06] KT Nam, SC Oh, Y Lee, JH Jeong, IG Baek, EK Yim, JS Zhao, SO Park, HS Kim, U Chung, et al. "Switching properties in spin transper torque MRAM with sub-5Onm MTJ size." In *Proc. NVMTS*, pp. 49–51. IEEE, 2006.

[NRQ15] Prashant J Nair, David A Roberts, and Moinuddin K Qureshi. "FaultSim: A Fast, Configurable Memory-Reliability Simulator for Conventional and 3D-Stacked Systems." *ACM Transactions on Architecture and Code Optimization (TACO)*, **12**(4):44, 2015.

[NSM11] Anurag Nigam, Clinton W Smullen IV, Vidyabhushan Mohan, Eugene Chen, Sudhanva Gurumurthi, and Mircea R Stan. "Delivering on the promise of universal memory for spin-transfer torque RAM (STT-RAM)." In *Proc. ISLPED*, pp. 121–126. IEEE, 2011.

[NY13] Dmitri E Nikonov, Ian Young, et al. "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking." *Proceedings of the IEEE*, **101**(12):2498–2533, 2013.

[NYT13] Takayuki Nozaki, Kay Yakushiji, Shingo Tamaru, Masaki Sekine, Rie Matsumoto, Makoto Konoto, Hitoshi Kubota, Akio Fukushima, and Shinji Yuasa. "Voltage-Induced Magnetic Anisotropy Changes in an Ultrathin FeB Layer Sandwiched between Two MgO Layers." *Applied Physics Express*, **6**(7):073005, 2013.

[OIE15] Satoshi Ohuchida, Kenchi Ito, and Tetsuo Endoh. "Impact of sub-volume excitation on improving overdrive delay product of sub-40 nm perpendicular magnetic tunnel junctions in adiabatic regime and its beyond." *Japanese J. Appl. Phys.*, **54**(4S):04DD05, 2015.

[OKM12] T Ohsawa, H Koike, S Miura, H Honjo, K Tokutome, S Ikeda, T Hanyu, H Ohno, and T Endoh. "1Mb 4T-2MTJ nonvolatile STT-RAM for embedded memories using 32b fine-grained power gating technique with 1.0 ns/200ps wake-up/power-off times." In *Proc. VLSIC*, pp. 46–47. IEEE, 2012.

[PC12a] A. Pan and Chi On Chui. "A Quasi-Analytical Model for Double-Gate Tunneling Field-Effect Transistors." *Electron Device Letters, IEEE*, **33**(10):1468–1470, Oct 2012.

[PC12b] Andrew Pan and ON CHUI CHI. "A quasi-analytical model for double-gate tunneling field-effect transistors." *IEEE electron device letters*, **33**(10):1468–1470, 2012.

[PCC13] Andrew Pan, Songtao Chen, and Chi On Chui. "Electrostatic modeling and insights regarding multigate lateral tunneling transistors." *Electron Devices, IEEE Transactions on*, **60**(9):2712–2720, 2013.

[PKJ11] Jong-Yoon Park, Se-Koo Kang, Min-Hwan Jeon, Myung S Jhon, and Geun-Young Yeom. "Etching of CoFeB Using CO/ NH3 in an Inductively Coupled Plasma Etching System." *J. Electrochem. Soc*, **158**(1):H1–H4, 2011.

[PLS09a] Kedar Patel, Tsu-Jae King Liu, and Costas J Spanos. "Gate line edge roughness model for estimation of FinFET performance variability." *Electron Devices, IEEE Transactions on*, **56**(12):3055–3063, 2009.

[PLS09b] Kedar Patel, Tsu-Jae King Liu, and Costas J Spanos. "Gate line edge roughness model for estimation of FinFET performance variability." *Electron Devices, IEEE Transactions on*, **56**(12):3055–3063, 2009.

[PMH15] Mirko Prezioso, Farnood Merrikh-Bayat, Brian Hoskins, Gina Adam, Konstantin K Likharev, and Dmitri B Strukov. "Training and Operation of an Integrated Neuromorphic Network Based on Metal-Oxide Memristors." *Nature*, **521**:61–64, 2015.

[PML90] JA Power, A Mathewson, and WA Lane. "MOSFET statistical parameter extraction using multivariate statistics." In *Microelectronic Test Structures, 1991. ICMTS 1991. Proceedings of the 1991 International Conference on*, pp. 209–214. IEEE, 1990.

[PN12] Chenyun Pan and Azad Naeemi. "System-level optimization and benchmarking of graphene PN junction logic system based on empirical CPI model." In *IC Design & Technology (ICICDT), 2012 IEEE International Conference on*, pp. 1–5. IEEE, 2012.

[Pro00] Robert J Proebsting. "Differential sense amplifier circuit.", November 28 2000. US Patent 6,154,064.

[PTM] "PTM model." `http://ptm.asu.edu/`.

[QR08] Weikang Qian and Marc D Riedel. "The synthesis of robust polynomial arithmetic with stochastic logic." In *Proc. DAC. 45th ACM/IEEE*, pp. 648–653. IEEE, 2008.

[RCA11] R Rios, A Cappellani, M Armstrong, A Budrevich, H Gomez, R Pai, N Rahhal-Orabi, and K Kuhn. "Comparison of junctionless and conventional trigate transistors with down to 26 nm." *Electron Device Letters, IEEE*, **32**(9):1170–1172, 2011.

[RDJ02] ND Rizzo, M DeHerrera, J Janesky, B Engel, J Slaughter, and S Tehrani. "Thermally activated magnetization reversal in submicron magnetic tunnel junctions for magnetoresistive random access memory." *Appl. Phys. Lett.*, **80**(13):2335–2337, 2002.

[refa] "CortexM0." `http://www.arm.com/products/processors/cortex-m/cortex-m0.php`. Accessed: 2017-04-11.

[refb] "MIPS." `http://opencores.org`. Accessed: 2017-04-11.

[refc] "NanGate FreePDK45 Generic Open Cell Library." `http://www.si2.org/openeda.si2.org/projects/nangatelib`. Accessed: 2017-04-11.

[refd] "Predictive technology model." `http://ptm.asu.edu/`. Accessed: 2017-04-11.

[RKW09] Jong-hyun Ryu, Sujin Kim, and Hong Wan. "Pareto front approximation with adaptive weighted sum method in multiobjective simulation optimization." In *Winter Simulation Conference*, pp. 623–633. Winter Simulation Conference, 2009.

[RMM03] Kaushik Roy, Saibal Mukhopadhyay, and Hamid Mahmoodi-Meimand. "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits." *Proceedings of the IEEE*, **91**(2):305–327, 2003.

[RN14] D. Roberts and P. Nair. "FAULTSIM: A fast, configurable memory-resilience simulator." Technical report, The Memory Forum: In conjunction with ISCA-41, 2014.

[ROM01] Gunnar Rätsch, Takashi Onoda, and K-R Müller. "Soft margins for AdaBoost." *Machine learning*, **42**(3):287–320, 2001.

[RPT08] SL Rommel, D Pawlik, P Thomas, M Barth, K Johnson, SK Kurinec, A Seabaugh, Z Cheng, JZ Li, J-S Park, et al. "Record PVCR GaAs-based tunnel diodes fabricated on Si substrates using aspect ratio trapping." In *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4. IEEE, 2008.

[SBL11] Zhenyu Sun, Xiuyuan Bi, Hai Helen Li, Weng-Fai Wong, Zhong-Liang Ong, Xiaochun Zhu, and Wenqing Wu. "Multi retention level STT-RAM cache designs with a dynamic refresh scheme." In *Proc. MICRO*, pp. 329–338. ACM, 2011.

[SCS08] Jack C Sankey, Yong-Tao Cui, Jonathan Z Sun, John C Slonczewski, Robert A Buhrman, and Daniel C Ralph. "Measurement of the spin-transfer-torque vector in magnetic tunnel junctions." *Nature Physics*, **4**(1):67–71, 2008.

[SCW00] JM Slaughter, EY Chen, R Whig, BN Engel, J Janesky, and S Tehrani. "Magnetic tunnel junction materials for electronic applications." *JOM(USA)*, **52**(6):11, 2000.

[SD12] Alexander Stotland and Massimiliano Di Ventra. "Stochastic memory: memory enhancement due to noise." *Physical Review E*, **85**(1):011116, 2012.

[SDB15] Vilas Sridharan, Nathan DeBardeleben, Sean Blanchard, Kurt B Ferreira, Jon Stearley, John Shalf, and Sudhanva Gurumurthi. "Memory Errors in Modern Systems: The Good, The Bad, and The Ugly." In *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 297–310. ACM, 2015.

[SDC96] J.N. Schulman, H.J. De Los Santos, and D.H. Chow. "Physics-based RTD current-voltage equation." *Electron Device Letters, IEEE*, **17**(5):220–222, May 1996.

[SFK11] PM Solomon, DJ Frank, and SO Koswatta. "Compact model and performance estimation for tunneling nanowire FET." In *69th Device Research Conference*, 2011.

[SGP00] Roy Scheuerlein, William Gallagher, Stuart Parkin, Alex Lee, Sam Ray, Ray Robertazzi, and William Reohr. "A 10 ns read and write non-volatile memory array using a magnetic tunnel junction and FET switch in each cell." In *Proc. ISSCC*, pp. 128–129. IEEE, 2000.

[SK99] Dennis Sylvester and Kurt Keutzer. "System-level performance modeling with BACPAC–Berkeley advanced chip performance calculator." In *Proc. SLIP*, pp. 109–114. Citeseer, 1999.

[SKC14] Karthik Swaminathan, Moon Seok Kim, Nandhini Chandramoorthy, Behnam Sedighi, Robert Perricone, Jack Sampson, and Vijaykrishnan Narayanan. "Modeling steep slope devices: From circuits to architectures." In *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pp. 1–6. IEEE, 2014.

[SL12a] V. Sridharan and D. Liberty. "A study of DRAM failures in the field." In *High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for*, pp. 1–11, Nov 2012.

[SL12b] Vilas Sridharan and Dean Liberty. "A study of DRAM failures in the field." In *Proc. SC*, pp. 1–11. IEEE, 2012.

[SLa16] Y. J. Song, J. H. Lee, and et. al. "Highly Functional and Reliable 8Mb STT-MRAM Embedded in 28nm Logic." In *Proc. IEDM*, 2016.

[SLL11] R Sbiaa, SYH Lua, R Law, H Meng, R Lye, and HK Tan. "Reduction of switching current by spin transfer torque effect in perpendicular anisotropy magnetoresistive devices." *J. Appl. Phys.*, **109**(7):07C707, 2011.

[SLL14] Karthik Swaminathan, Huichu Liu, Xueqing Li, Moon Seok Kim, Jack Sampson, and Vijaykrishnan Narayanan. "Steep slope devices: Enabling new architectural paradigms." In *Proceedings of the 51st Annual Design Automation Conference*, pp. 1–6. ACM, 2014.

[Slo96] John C Slonczewski. "Current-driven excitation of magnetic multilayers." *J. Magn. Magn. Mater.*, **159**(1):L1–L7, 1996.

[SLS14a] Karthik Swaminathan, Huichu Liu, Jack Sampson, and Vijaykrishnan Narayanan. "An examination of the architecture and system-level tradeoffs of employing steep slope devices in 3D CMPs." In *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*, pp. 241–252. IEEE, 2014.

[SLS14b] Karthik Swaminathan, Huichu Liu, Jack Sampson, and Vijaykrishnan Narayanan. "An examination of the architecture and system-level tradeoffs of employing steep slope devices in 3D CMPs." In *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*, pp. 241–252. IEEE, 2014.

[SLS16] YJ Song, JH Lee, HC Shin, KH Lee, K Suh, JR Kang, SS Pyo, HT Jung, SH Hwang, GH Koh, et al. "Highly functional and reliable 8Mb STT-MRAM embedded in 28nm logic." In *Electron Devices Meeting (IEDM), 2016 IEEE International*, pp. 27–2. IEEE, 2016.

[SMM02] Nobuyuki Sano, Kazuya Matsuzawa, Mikio Mukai, and Noriaki Nakayama. "On discrete random dopant modeling in drift-diffusion simulations: physical meaning ofatomistic'dopants." *Microelectronics Reliability*, **42**(2):189–199, 2002.

[SMN11] Clinton W Smullen, Vidyabhushan Mohan, Anurag Nigam, Sudhanva Gurumurthi, and Mircea R Stan. "Relaxing non-volatility for fast and energy-efficient STT-RAM caches." In *HPCA*, pp. 50–61. IEEE, 2011.

[SMN12] Yoichi Shiota, Shinji Miwa, Takayuki Nozaki, Frédéric Bonell, Norikazu Mizuochi, Teruya Shinjo, Hitoshi Kubota, Shinji Yuasa, and Yoshishige Suzuki. "Pulse voltage-induced dynamic magnetization switching in magnetic tunneling junctions with high resistance-area product." *Appl. Phys. Lett.*, **101**(10):102406, 2012.

[SMW15] Stefan Slesazeck, Hannes Mähne, Helge Wylezich, Andre Wachowiak, Janaki Radhakrishnan, Alon Ascoli, Ronald Tetzlaff, and Thomas Mikolajick. "Physical model of threshold switching in NbO 2 based memristors." *RSC Advances*, **5**(124):102318–102322, 2015.

[SNB12] Yoichi Shiota, Takayuki Nozaki, Frédéric Bonell, Shinichi Murakami, Teruya Shinjo, and Yoshishige Suzuki. "Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses." *Nature materials*, **11**(1):39–43, 2012.

[SPW09] Bianca Schroeder, Eduardo Pinheiro, and Wolf-Dietrich Weber. "DRAM errors in the wild: a large-scale field study." In *ACM SIGMETRICS Performance Evaluation Review*, volume 37, pp. 193–204. ACM, 2009.

[SR09] Amith Singhee, Rob Rutenbar, et al. "Statistical blockade: very fast statistical simulation and modeling of rare circuit events and its application to memory design." *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, **28**(8):1176–1189, 2009.

[SRN11] JZ Sun, RP Robertazzi, J Nowak, PL Trouilloud, G Hu, DW Abraham, MC Gaidis, SL Brown, EJ OSullivan, WJ Gallagher, et al. "Effect of subvolume excitation and spin-torque efficiency on magnetic switching." *Phys. Rev. B*, **84**(6):064413, 2011.

[SSL95] Kang-Deog Suh, Byung-Hoom Suh, Young-Ho Lim, Jin-Ki Kim, Young-Joon Choi, Yong-Nam Koh, Sung-Soo Lee, Suk-Chon Suk-Chon, Byung-Soon Choi, Jin-Sun Yum, et al. "A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme." *Solid-State Circuits, IEEE Journal of*, **30**(11):1149–1156, 1995.

[SSS08] Dmitri B Strukov, Gregory S Snider, Duncan R Stewart, and R Stanley Williams. "The missing memristor found." *nature*, **453**(7191):80–83, 2008.

[STM11] John Paul Strachan, Antonio C Torrezan, Gilberto Medeiros-Ribeiro, and R Stanley Williams. "Measuring the switching dynamics and energy efficiency of tantalum oxide memristors." *Nanotechnology*, **22**(50):505402, 2011.

[TH08] Kiyoshi Takeuchi and Masami Hane. "Statistical compact model parameter extraction by direct fitting to variations." *Electron Devices, IEEE Transactions on*, **55**(6):1487–1493, 2008.

[TSC99] Said Tehrani, JM Slaughter, E Chen, M Durlam, J Shi, and M DeHerren. "Progress and outlook for MRAM technology." *TMAG*, **35**(5):2814–2819, 1999.

[TSL94] H.H. Tsai, Y.K. Su, H.H. Lin, R.-L. Wang, and T.L. Lee. "P-N double quantum well resonant interband tunneling diode with peak-to-valley current ratio of 144 at room temperature." *Electron Device Letters, IEEE*, **15**(9):357–359, Sept 1994.

[UHM03] Tetsuya Uemura, Satoshi Honma, Takao Marukame, and Masafumi Yamamoto. "Large enhancement of tunneling magnetoresistance ratio in magnetic tunnel junction connected in series with tunnel diode." *Japanese journal of applied physics*, **43**(1A):L44, 2003.

[UY03] Tetsuya Uemura and Masafumi Yamamoto. "Proposal of four-valued MRAM based on MTJ/RTD structure." In *Multiple-Valued Logic, 2003. Proceedings. 33rd International Symposium on*, pp. 273–278. IEEE, 2003.

[UYY14] Yohei Umeki, Koji Yanagida, Shusuke Yoshimoto, Shintaro Izumi, Masahiko Yoshimoto, Hiroshi Kawaguchi, Koji Tsunoda, and Toshihiro Sugii. "STT-MRAM Operating at 0.38 V Using Negative-Resistance Sense Amplifier." *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **97**(12):2411–2417, 2014.

[UYY15] Yohei Umeki, Koji Yanagida, Shusuke Yoshimoto, Shintaro Izumi, Masahiko Yoshimoto, Hiroshi Kawaguchi, Koji Tsunoda, and Toshihiro Sugii. "A negative-resistance sense amplifier for low-voltage operating STT-MRAM." In *Proc. ASPDAC*, pp. 8–9. IEEE, 2015.

[VRK04] Chandramouli Visweswariah, Kaushik Ravindran, Kerim Kalafala, Steven G Walker, and Sambasivan Narayan. "First-order incremental block-based statistical timing analysis." In *Proceedings of the 41st annual Design Automation Conference*, pp. 331–336. ACM, 2004.

[VVF15] Rangharajan Venkatesan, Swagath Venkataramani, Xuanyao Fong, Kaushik Roy, and Anand Raghunathan. "Spintastic: spin-based stochastic logic for energy-efficient computing." In *Proc. DATE*, pp. 1575–1578. IEEE, 2015.

[WA11] Lan Wei and Dimitri Antoniadis. "CMOS device design and optimization from a perspective of circuit-level energy-delay optimization." In *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 15–3. IEEE, 2011.

[WG14a] Wei-Che Wang and Puneet Gupta. "Efficient layout generation and evaluation of vertical channel devices." In *Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design*, pp. 550–556. IEEE Press, 2014.

[WG14b] Wei-Che Wang and Puneet Gupta. "Efficient layout generation and evaluation of vertical channel devices." In *Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design*, pp. 550–556. IEEE Press, 2014.

[WHA11] DC Worledge, G Hu, David W Abraham, JZ Sun, PL Trouilloud, J Nowak, S Brown, MC Gaidis, EJ OSullivan, and RP Robertazzi. "Spin torque switching of perpendicular Ta— CoFeB— MgO-based magnetic tunnel junctions." *Appl. Phys. Lett.*, **98**(2):022501–022501, 2011.

[WHZ15a] Shaodi Wang, Henry Hu, Hongzhong Zheng, and Puneet Gupta. "MEMRES: A Fast Memory System Reliability Simulator." In *SELSE: the 12th Workshop on Silicon Errors in Logic - System Effects*. IEEE, 2015.

[WHZ15b] Shaodi Wang, Henry Chaohong Hu, Hongzhong Zheng, and Puneet Gupta. "MEMRES: A Fast Memory System Reliability Simulator." In *The 11th IEEE Workshop on Silicon Errors in Logic System Effects (SELSE)*, 2015.

[WHZ16] Shaodi Wang, Henry (Chaohong) Hu, Hongzhong Zheng, and Puneet Gupta. "MEMRES: A Fast Memory System Reliability Simulator." *IEEE Transactions on Reliability*, **65**(4):1783–1797, 2016.

[Wil14] Thomas Willhalm. "Independent Channel vs. Lockstep Mode - Drive your Memory Faster or Safer." *Intel*, 2014.

[WLE16a] S. Wang, H. Lee, F. Ebrahimi, P. K. Amiri, K. L. Wang, and P. Gupta. "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, **6**(2):134–145, June 2016.

[WLE16b] Shaodi Wang, Hochul Lee, Farbod Ebrahimi, P. Khalili Amiri, Kang L. Wang, and Puneet Gupta. "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, **6**(2):134–145, 2016.

[WLG16] Shaodi Wang, Hochul Lee, Cecile Grezes, Pedram Khalili, Kang L Wang, and Puneet Gupta. "MTJ variation monitor-assisted adaptive MRAM write." In *53rd Annual Design Automation Conference (DAC)*, p. 169. ACM, 2016.

[WLH12] Wei-Gang Wang, Mingen Li, Stephen Hageman, and CL Chien. "Electric-field-assisted switching in magnetic tunnel junctions." *Nature materials*, **11**(1):64–68, 2012.

[WLP13] Shaodi Wang, Greg Leung, Andrew Pan, Chi On Chui, and Puneet Gupta. "Evaluation of digital circuit-level variability in inversion-mode and junctionless FinFET technologies." *TED*, **60**(7):2186–2193, 2013.

[WMX09] Kyoungho Woo, Scott Meninger, Thucydides Xanthopoulos, Ethan Crain, Dongwan Ha, and Donhee Ham. "Dual-DLL-based CMOS all-digital temperature sensor for microprocessor thermal monitoring." In *ISSCC*, pp. 68–69. IEEE, 2009.

[WOW10] Lan Wei, Saeroonter Oh, and HS Philip Wong. "Performance benchmarks for Si, III–V, TFET, and carbon nanotube FET-re-thinking the technology assessment methodology for complementary logic applications." In *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 16–2. IEEE, 2010.

[WPC14] Shaodi Wang, Andrew Pan, Chi On Chui, and Puneet Gupta. "PROCEED: A pareto optimization-based circuit-level evaluator for emerging devices." In *Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific*, pp. 818–824. IEEE, 2014.

[WPC15] Shaodi Wang, Andrew Pan, Chi On Chui, and Puneet Gupta. "PROCEED: A Pareto Optimization-Based Circuit-Level Evaluator for Emerging Devices." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, **24**(1):192–205, 2015.

[WPC16] S. Wang, A. Pan, C. O. Chui, and P. Gupta. "PROCEED: A Pareto Optimization-Based Circuit-Level Evaluator for Emerging Devices." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, **24**(1):192–205, Jan 2016.

[WPC17] Shaodi Wang, Andrew Pan, Chi On Chui, and Puneet Gupta. "Tunneling Negative Differential Resistance-Assisted STT-RAM for Efficient Read and Write Operations." *IEEE Transactions on Electron Devices*, **64**(1):121–129, 2017.

[WPG17] Shaodi Wang, Andrew Pan, Cecile Grezes, Pedram Khalili Amiri, Kang L. Wang, Chi On Chui, and Puneet Gupta. "Leveraging CMOS Negative Differential Resistance for Low Power, High Reliability Magnetic Memory." *IEEE Transactions on Electron Devices (accepted)*, 2017.

[WPL17] Shaodi Wang, Saptadeep Pal, Tianmu Li, Andrew Pan, Cecile Grezes, Pedram Khalili Amiri, Kang L. Wang, and Puneet Gupta. "Hybrid VC-MTJ/CMOS Non-volatile Stochastic Logic for Efficient Computing." In *Design, Automation & Test in Europe Conference (DATE)*. IEEE, 2017.

[WRK10] H-S Philip Wong, Simone Raoux, SangBum Kim, Jiale Liang, John P Reifenberg, Bipin Rajendran, Mehdi Asheghi, and Kenneth E Goodson. "Phase change memory." *Proceedings of the IEEE*, **98**(12):2201–2227, 2010.

[WRS17] Shaodi Wang, Glen Rosendale, Lucian Shifren, Carlos Araujo, and Puneet Gupta. "CERAM Modeling." *In preparation*, 2017.

[WZJ12] Peiyuan Wang, Wei Zhang, Rajiv Joshi, Rouwaida Kanj, and Yiran Chen. "A thermal and process variation aware MTJ switching model and its applications in soft error analysis." In *Proc. ICCAD*, pp. 720–727. IEEE, 2012.

[WZS09] Xiaobin Wang, Wenzhong Zhu, Markus Siegert, and Dimitar Dimitrov. "Spin torque induced magnetization switching variations." *Magnetics, IEEE Transactions on*, **45**(4):2038–2041, 2009.

[WZZ10] Shaodi Wang, Lining Zhang, Jian Zhang, Wenping Wang, Wen Wu, Xukai Zhang, Zhiwei Liu, Wei Bian, Frank He, and Mansun Chan. "A potential-based analytic model for monocrystalline silicon thin-film transistors on glass substrates." In *Solid-State and Integrated Circuit Technology (ICSICT), 2010 10th IEEE International Conference on*, pp. 1880–1882. IEEE, 2010.

[XDH05] Xuemei Xi, Mohan Dunga, Jin He, Weidong Liu, Kanyu M Cao, Xiaodong Jin, Jeff J Ou, Mansun Chan, Ali M Niknejad, and Chenming Hu. "BSIM4 manual." *UC Berkeley Device Group*, 2005.

[XSW11] Wei Xu, Hongbin Sun, Xiaobin Wang, Yiran Chen, and Tong Zhang. "Design of last-level on-chip cache using spin-torque transfer RAM (STT RAM)." *TVLSI*, **19**(3):483–493, 2011.

[YE10] Doe Hyun Yoon and Mattan Erez. "Virtualized and flexible ECC for main memory." In *ACM SIGARCH Computer Architecture News*, volume 38, pp. 397–408. ACM, 2010.

[YLN08] Yun Ye, Frank Liu, Sani Nassif, and Yu Cao. "Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness." In *Proc. DAC*, pp. 900–905. IEEE, 2008.

[YW11] Shimeng Yu and H-S Philip Wong. "Compact modeling of conducting-bridge random-access memory (CBRAM)." *IEEE Transactions on Electron devices*, **58**(5):1352–1360, 2011.

[ZBW16] Liheng Zhu, Yasmine Badr, Shaodi Wang, Subramanian Iyer, and Puneet Gupta. "Assessing Benefits of a Buried Interconnect Layer in Digital Designs." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2016.

[ZKK13] Yuping Zeng, Chien-I Kuo, Rehan Kapadia, Ching-Yi Hsu, Ali Javey, and Chenming Hu. "Two-dimensional to three-dimensional tunneling in InAs/AlSb/GaSb quantum well heterojunctions." *Journal of Applied Physics*, **114**(2), 2013.

[ZLL13] Xiaoyang Zhang, Lin Liu, Wenxuan Liang, Xingde Li, and Huikai Xie. "An electrothermal/electrostatic dual driven MEMS scanner with large in-plane and out-of-plane displacement." In *Optical MEMS and Nanophotonics (OMN), 2013 International Conference on*, pp. 13–14. IEEE, 2013.

[ZLL15] Xiaoyang Zhang, Boxiao Li, Xingde Li, and Huikai Xie. "A robust, fast electrothermal micromirror with symmetric bimorph actuators made of copper/tungsten." In *Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS), 2015 Transducers-2015 18th International Conference on*, pp. 912–915. IEEE, 2015.

[ZVD14] Xin Zhao, A. Vardi, and J.A. Del Alamo. "InGaAs/InAs heterojunction vertical nanowire tunnel fets fabricated by a top-down approach." In *Proc. International Electron Devices Meeting*, pp. 25.5.1–25.5.4. IEEE, Dec 2014.

[ZWC11] Yaojun Zhang, Xiaobin Wang, and Yiran Chen. "STT-RAM cell design optimization for persistent and non-persistent error rate reduction: a statistical design view." In *Proc. ICCAD*, pp. 471–477. IEEE, 2011.

[ZZD12] WS Zhao, Yue Zhang, Thibaut Devolder, Jacques-Olivier Klein, Dafine Ravelosona, Claude Chappert, and Pascale Mazoyer. "Failure and reliability analysis of STT-MRAM." *Microelectronics Reliability*, **52**(9):1848–1852, 2012.

[ZZL12] Yue Zhang, Weisheng Zhao, Yahya Lakys, J-O Klein, Joo-Von Kim, Dafiné Ravelosona, and Claude Chappert. "Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions." *TED*, **59**(3):819–826, 2012.

[ZZW12] Yaojun Zhang, Lu Zhang, Wujie Wen, Guangyu Sun, and Yiran Chen. "Multi-level cell STT-RAM: Is it realistic or just a dream?" In *Proceedings of the International Conference on Computer-Aided Design*, pp. 526–532. ACM, 2012.

[ZZY09] Ping Zhou, Bo Zhao, Jun Yang, and Youtao Zhang. "Energy reduction for STT-RAM using early write termination." In *ICCAD*, pp. 264–268. IEEE, 2009.