Design Techniques for Dense Embedded Memory in Advanced CMOS Technologies

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Ki Chul Chun

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chris H. Kim, Adviser

February 2012

© Ki Chul Chun 2012

Acknowledgements

This dissertation has been accomplished with the advice of Professor Chris H. Kim, collaboration with colleagues, financial support from Samsung Electronics, and dedication of my family and wife.

First, I would like to thank my academic adviser, Professor Chris H. Kim, for guiding me with exploratory research items and providing me with friendly environments for doing research as well as enjoying life in MN. His endless enthusiasm for academic explorations has always inspired me; consequently, we achieved many outstanding research outcomes.

Next, I am grateful to my final defense committee; Professors Ramesh Harjani, Sachin Sapatnekar, and Antonia Zhai for reviewing this dissertation and providing valuable discussions during your busy schedules.

I owe a great deal of gratitude to Samsung Electronics for supporting my entire PhD study through the Samsung PhD scholarship. I started my career as a VLSI designer at Samsung Electronics in 2000. I am grateful to all the members of Samsung DRAM Design Team for making my understanding of memory circuit design strong.

The time that I have got together with all the members of VLSI Research Lab at the University of Minnesota would last for the last of my life. To previous members; I thank Jie Gu, Tony T. Kim, and John Keane for your many helps during my settlement. To current members; I am grateful for lots of insightful discussions with you. I have been happy with you; Dong Jiao, Wei Zhang, Pulkit Jain, Xiaofei Wang, Seunghwan Song,

Ayan Paul, Bongjin Kim, Jongyeon Kim, Wonho Choi, Weichao D. Xu, Renukprasad Hiremath, and Dan Liu.

My PhD study has never been lonely due to many good Korean friends. I thank Jaehyup Kim, Youngil Kim, Boram Lee, Jaesang Oh, Hweerin Sohn, and Sungmin Sohn for sharing your valuable times with me.

I would also like to thank my parents and my younger brother. They have always been supporting me and encouraging me with their best wishes.

Finally, I would like to thank my wife, Sun Young Hwang. She has dedicated herself astoundingly to support my PhD study. I also thank our daughter; Siyeon Chun to be such a lovely kid. Thank you for everything.

Abstract

On-die cache memory is a key component in advanced processors since it can boost micro-architectural level performance at a moderate power penalty. Demand for denser memories only going to increase as the number of cores in a microprocessor goes up with technology scaling. A commensurate increase in the amount of cache memory is needed to fully utilize the larger and more powerful processing units. 6T SRAMs have been the embedded memory of choice for modern microprocessors due to their logic compatibility, high speed, and refresh-free operation. However, the relatively large cell size and conflicting requirements for read and write make aggressive scaling of 6T SRAMs challenging in sub-22 nm.

In this dissertation, circuit techniques and simulation methodologies are presented to demonstrate the potential of alternative options such as gain cell eDRAMs and spin-torque-transfer magnetic RAMs (STT-MRAMs) for high density embedded memories.

Three unique test chip designs are presented to enhance the retention time and access speed of gain cell eDRAMs. Proposed bit-cells utilize preferential boostings, beneficial couplings, and aggregated cell leakages for expanding signal window between data '1' and '0'. The design space of power-delay product can be further enhanced with various assist schemes that harness the innate properties of gain cell eDRAMs. Experimental results from the test chips demonstrate that the proposed gain cell eDRAMs achieve

overall faster system performances and lower static power dissipations than SRAMs in a generic 65 nm low-power (LP) CMOS process.

A magnetic tunnel junction (MTJ) scaling scenario and an efficient HSPICE simulation methodology are proposed for exploring the scalability of STT-MRAMs under variation effects from 65 nm to 8 nm. A constant J_{C0}•RA/VDD scaling method is adopted to achieve optimal read and write performances of STT-MRAMs and thermal stabilities for a 10 year retention are achieved by adjusting free layer thicknesses as well as projecting crystalline anisotropy improvements. Studies based on the proposed methodology show that in-plane STT-MRAM will outperform SRAM from 15 nm node, while its perpendicular counterpart requires further innovations in MTJ material properties in order to overcome the poor write performance from 22 nm node.

Table of Contents

List of Figures viii

1 Introduction 1

1.1 Embedded Memories in Multi-Core Microprocessor 2

1.2 Alternative Memory Technologies 4

1.3 Summary of Dissertation Contributions 6

2 A 3T Embedded DRAM Utilizing Preferential Boosting for Low Voltage On-Die Caches 8

2.1 Basic Operation of a Conventional 3T EDRAM 8

2.2 Boosted 3T EDRAM Design 10

2.2.1 Boosted 3T Gain Cell 11

2.2.2 Regulated Bit-Line Write Scheme 15

2.2.3 PVT-Tracking Read Reference Bias 16

2.2.4 Architecture and Operation of a 32 kb Sub-Array 17

2.3 Statistical Simulation Results for 6T SRAM and 3T EDRAM Arrays 19

2.3.1 Read and Write Performance 20

2.3.2 Static Power Consumption 24

2.4 Test Chip Implementation and Measurements 27

2.5 Conclusions 32

3 An Asymmetric 2T EDRAM for High Speed On-die Caches............34

3.1 Retention Characteristics of Conventional Gain Cells 34

3.2 Asymmetric 2T EDRAM Design 36

3.2.1 Asymmetric 2T Gain Cell..36

3.2.2 Pseudo-PMOS Diode Based Current-Mode Sense Amplifier (C-S/A)...........38

3.2.3 Half Swing Write Bit-Line Scheme 43

3.2.4 Stepped Write Word-Line Driver...44

3.2.5 Sense Amplifier and Write-Back Circuit Design..46

3.3 Comparison Between SRAM and Gain Cell EDRAM 48

3.3.1 Macro Layout Comparison 48

3.3.2 Macro Performance Comparison 50

3.4 Test Chip Implementation and Measurements 52

3.5 Conclusions 57

4 A Logic-Compatible 2T1C EDRAM for Enhanced Reliability 59

4.1 Boosted Supply Level vs. EDRAM Performance 59

4.2 2T1C EDRAM with No Boosted Supplies 60

4.2.1 2T1C Gain Cell 60

4.2.2 Decoupled 7T SRAM Repair Cell with Shared Control 65

4.2.3 Cell Storage Monitor...69

4.3 2T1C EDRAM Test Chip Measurements 71

4.4 Conclusions 78

5 A Scalability Exploration of STT-MRAMs Considering Variation Effects 80

5.1 Introduction to STT-MRAM 80

5.2 STT-MTJ Scaling Roadmap 82

5.2.1 In-Plane STT-MTJ Scaling Scenario 84

5.2.2 Perpendicular STT-MTJ Scaling Scenario 85

5.2.3 STT-MTJ Scaling Trend 86

5.3 STT-MRAM HSPICE Simulation Methodology 88

5.3.1 STT Switching and MTJ Macromodel 88

5.3.2 Transistor Scaling Trend 90

5.3.3 Sub-Array Architecture and Variation Sources 91

5.4 Comparison Between SRAM and STT-MRAM 94

5.4.1 Macro Performance 94

5.4.2 In-Plane STT-MRAM vs. 6T SRAM 95

5.4.3 In-Plane STT-MRAM vs. Perpendicular STT-MRAM 99

5.5 Conclusions 100

6 Conclusion 102

Bibliography 106

List of Figures

Fig. 1.1: High-end microprocessors with high density on-die L3 caches based on 6T SRAM and 1T1C eDRAM. 2

Fig. 1.2: (a) Trend of number of cores and corresponding on-die L3 cache densities in Intel's and IBM's high-end microprocessors. (b) Bit-cell size scaling trend of 6T SRAMs and 1T1C eDRAMs. 4

Fig. 2.1: (a) Conventional 3T PMOS eDRAM gain cell circuit diagram. (b) Signal voltages in each operating mode. 8

Fig. 2.2: Monte Carlo simulation results of storage node voltage during data hold mode. 10

Fig. 2.3: (a) Proposed boosted 3T PMOS eDRAM gain cell. (b) Preferential RWL coupling effects of the proposed cell. (c) Simulation results of the storage node preferential boosting effects. (d) Signal voltage conditions for each operating mode. 12

Fig. 2.4: (a) Hybrid bit-line current/voltage sense amplifier (S/A) with read port, write port, and write-back circuits. (b) Read and write-back timing diagram of the proposed S/A. 14

Fig. 2.5: (a) Storage node disturbance problem when writing data '1' to a cell sharing the same WBL. (b) Simulation results showing steady-state storage node voltage in case of no refresh. (c) Proposed regulated bit-line write bias generator based on replica cells...15

Fig. 2.6: PVT-tracking and die-to-die adjustable read reference bias (VDUM) generator. 17
Fig. 2.7: (a) Simulation results of the proposed VDUM generator tracking temperature and process variations. (b) Simulation results showing the dependency of VDUM on VDD. 17
Fig. 2.8: A 32 kb array structure of the proposed eDRAM including (a) boosted 3T gain cell, (b) hybrid current/voltage S/A, (c) regulated bit-line write scheme, and (d) PVT-tracking read reference scheme. 18
Fig. 2.9: Read and write-back simulation waveform with a 2 ns random cycle time. 19
Fig. 2.10: Simulation setup for 1 M Monte Carlo simulations. 20
Fig. 2.11: Read performance comparisons between 6T SRAM and 3T eDRAM obtained from 2^{20} Monte-Carlo iterations. Results are equivalent to the distribution of a 1 Mb macro array. 6T SRAM has the shortest bitline delay attributed to the differential swing nature and large drive current (361.7ps @ 6σ) followed by the proposed 3T eDRAM (607.4ps @6σ) and the conventional 3T eDRAM (944.5ps @6σ). 21
Fig. 2.12: Comparison of various logic-compatible embedded memory cell layouts using a 65nm logic design rule (the authors did not have access to the dense bitcell design rule but for area comparison purposes, the logic design rule is generally acceptable). The outer box represents the cell boundary. Signal names, wire tracks, and device names are marked for the boosted 3T and conventional 3T cells. 22

Fig. 2.13: Latency comparisons between SRAM and 3T eDRAM for 1 Mb and 16 Mb cache sizes. Gain cells have a shorter interconnect delay due to the smaller cell size making their performance favorable in larger arrays. 23

Fig. 2.14: Write delay distributions of 1Mb arrays using 6T SRAM and 3T eDRAM. 24

Fig. 2.15: Leakage components of a (a) 6T SRAM, a (b) conventional 3T eDRAM and the (c) proposed 3T eDRAM. (d) Bias conditions and normalized cell leakages of SRAM and 3T eDRAM in active and sleep modes..............25

Fig. 2.16: Static power comparisons between a 1 Mb SRAM and a 2 Mb 3T eDRAM. Leakage power of the peripheral circuit is assumed to be negligible. 26

Fig. 2.17: Comparison of logic-compatible embedded memories. 27

Fig. 2.18: Microphotograph of the 65 nm eDRAM test chip and feature summary. 28

Fig. 2.19: (a) Measured regulated bit-line write bias (VWR) level. (b) Storage node voltage measurement results under data '1' disturbance conditions..............29

Fig. 2.20: (a) Measured retention time statistics. Due to limitations in the test setup, only 32 cells were measured from each sub-array. The measured cells were located evenly across the memory array. (b) Measured storage node voltage in the proposed boosted 3T cell and the conventional 3T cell. The cell voltage was indirectly and noninvasively measured by sweeping the reference cell node voltage. 30

Fig. 2.21: Measured storage node voltages at (a) 85°C and (b) 25°C. (c) Measured PVT-tracking read reference (VDUM) level at different supply voltages. 32

Fig. 3.1: (a) Leakage components of a 3T NMOS gain cell during data hold mode. (b) Monte-Carlo simulation results of storage node voltage during data hold mode showing 1 Mb macro retention characteristics. 35

Fig. 3.2: Circuit diagrams and retention characteristics of (a) a previous asymmetric 3T gain cell [16] and (b) the proposed asymmetric 2T gain cell. 38

Fig. 3.3: Illustration of limiting read margin by adjacent cells holding high state in a 2T eDRAM...39

Fig. 3.4: (a) Simulated RBL sensing waveform when all adjacent cells hold a data '0'. (b) All adjacent cells hold a data '1' indicating a data '1' read failure. The shaded regions denote the ΔV_{RBL}=100mV window between the accessed RBL and the reference. 39

Fig. 3.5: (a) NP series-stacked C-S/A [39]. (b) Hybrid C-S/A [40]...............................41

Fig. 3.6: Proposed pseudo-PMOS diode based C-S/A to overcome the issue of limited RBL voltage swing in a 2T eDRAM with improved voltage headroom and better impedance matching. 41

Fig. 3.7: (a) Simulated input resistance (ΔV_{RBL}/ ΔI_{IN}) vs. VDD. (b) Comparison of RBL sensing delay under PVT variations and mismatches in the C-S/A pairs. 43

Fig. 3.8: (a) Simulated waveforms of the WBL charging delay. (b) Simulated storage voltage distributions of a conventional GND pre-discharge (full swing WBL) and the proposed half-VDD pre-charge (half swing WBL) schemes. 44

Fig. 3.9: Proposed stepped WWL driver. (a) Schematic. (b) Timing diagram. (c) Simulated boosted current consumptions and WWL waveforms during transition. 46

Fig. 3.10: (a) Circuit diagram of the proposed Sense Amplifier (S/A) with read port, write port, and write-back circuits. (b) Two-stage read and write-back timing diagram. 47

Fig. 3.11: Simulated waveforms of back-to-back read and write-back operations for a 1.5 ns cycle time. 48

Fig. 3.12: Comparison of bit-cell and 128 kb sub-array layout between 6T SRAM and 2T eDRAM. 49

Fig. 3.13: RBL sensing delay distributions of SRAM and gain cell eDRAMs each with a 1 Mb macro density. 50

Fig. 3.14: Performance comparison of 1 Mb macros using SRAM and gain cell eDRAMs. (a) Latency. (b) Random cycle. 51

Fig. 3.15: Performance comparison of 1 Mb macros using SRAM and gain cell eDRAMs. (a) Latency. (b) Random cycle. 52

Fig. 3.16: A 192 kb test array architecture with 192 cells-per-WL and 512 cells-per-BL. 53

Fig. 3.17: (a) Microphotograph of the 65nm eDRAM test chip. (b) Chip feature summary. 53

Fig. 3.18: Measured storage node voltage at different retention times at (a) 85ºC and (b) 25ºC. 54

Fig. 3.19: Measured retention time distribution vs. boosted high supply (VPP) level. 55

Fig. 3.20: (a) Measured random cycle time vs. retention time. (b) Measured VDD shmoo of random cycle time and corresponding retention time. 56

Fig. 3.21: Static power comparison between 6T SRAM and the proposed 2T eDRAM with varying random cycle time at 85 ºC and 25 ºC. 57

Fig. 4.1: Impact of boosted supply level on 2T eDRAM performance [35]. (a) Boosted high supply (VPP) level vs. retention time (measured). (b) Boosted low supply (VBB) level vs. data '0' write time (simulated). 60

Fig. 4.2: Proposed 2T1C gain cell based on thin oxide devices with no boosted supplies. (a) Schematic. (b) Signal conditions for each operating modes. 61

Fig. 4.3: Timing diagram of the proposed 2T1C cell for read and write-back operations. 62

Fig. 4.4: Simulated waveforms of read and write-back operations for (a) a conventional 2T eDRAM and (b) the proposed 2T1C eDRAM. 63

Fig. 4.5: Comparison of retention characteristics between (a) a conventional 2T eDRAM with boosted supplies and (b) the proposed 2T1C eDRAM with no boosted supplies. 64

Fig. 4.6: Schematic diagram of a 64 kb 2T1C gain cell eDRAM macro with no boosted supplies. 64

Fig. 4.7: Shared coupling signal (PCOU). (a) Bit-cell schematic and layouts. (b) Simulated waveforms show negligible disturbance in an unselected cell. 65

Fig. 4.8: Proposed decoupled 7T SRAM repair cell shares BL and WL signals with the 2T1C cell. 66

Fig. 4.9: Signal-to-noise margin (SNM) of a 6T SRAM and the proposed 7T SRAM. (a) Read SNM. (b) Write SNM. 67

Fig. 4.10: Bit-cell comparison(6T SRAM, 3T, 2T, 2T1C cells): All bit-cells were drawn in a generic 65 nm LP process. 68

Fig. 4.11: Proposed storage voltage monitor for adaptive refresh control. 70

Fig. 4.12: Block diagram of the adaptive refresh control. 71

Fig. 4.13: (a) Microphotograph of the 65 nm eDRAM test chip. (b) Chip feature summary. 72

Fig. 4.14: Measured retention time distribution. 73

Fig. 4.15: Measured retention bit-map of 2T1C and decoupled 7T arrays. 74

Fig. 4.16: (a) Measured retention time distribution of the 2T1C array. (b) Effectiveness of various repair schemes. 74

Fig. 4.17: (a) Measured retention time distribution of data ‘1’ and ‘0’ at 25 ºC and 85 ºC. (b) Measurement storage voltage with varying temperature and retention time. 75

Fig. 4.18: Comparison of static current between SRAM (with power-gating) and 2T1C eDRAM (with adaptive refresh control)..76

Fig. 4.19: Measured VDD shmoo. (a) Random cycle time and retention time of the 2T1C eDRAM. (b) Static power dissipations of a 6T SRAM and the 2T1C eDRAM...76

Fig. 4.20: Comparison between our design and several embedded memory options. 77

Fig. 5.1: (a) Magnetic tunnel junction (MTJ) stack and its corresponding circuit schematic as a two-terminal device with varying resistance. (b) Resistance vs. write current (R-I) hysteresis curve. (c) Illustration of spin torque transfer (STT) switching's. 81

Fig. 5.2: (a) STT-MRAM bit-cell schematic. (b) Signal voltages for each operating mode. 82

Fig. 5.3: STT-MTJ scaling scenario based on dimensional adjustment and/or material innovation in order to maintain non-volatility and achieve optimal RD and WR operations. 83

Fig. 5.4: In-plane and perpendicular STT-MTJ scaling trends based on Fig. 5.3............86

Fig. 5.5: Vertical structure and SEM image of fabricated STT-MTJ (left) and summary of measured MTJ parameters (right)..87

Fig. 5.6: STT-MTJ characterization array layout and test chip feature summary. 88

Fig. 5.7: MTJ macromodel fitting results using MTJ data from our characterization array and [23]..89

Fig. 5.8: High performance (HP) transistor scaling trend based on ITRS......................91

Fig. 5.9: 128 kb sub-array architectures of (a) SRAM and (b) STT-MRAM. 92

Fig. 5.10: Simulation set-up for evaluating SRAM and STT-MRAM variability. 93

Fig. 5.11: Simulated read disturb rate with varying J_{RD}/J_{C0}..93

Fig. 5.12: Latency comparison between several embedded memory options with (a) 1 Mb and (b) 64 Mb densities. 95

Fig. 5.13: In-plane STT-MRAM scaling trends: Write time. (a) Absolute values. (b) Normalized to SRAM. 96

Fig. 5.14: Write time distributions of SRAM and in-plane STT-MRAM (P-AP) for a 1 Mb macro density at 15 nm node. 97

Fig. 5.15: In-plane STT-MRAM scaling trends: Read sensing delay. (a) Absolute values. (b) Normalized to SRAM..98

Fig. 5.16: Read sensing delay distributions of SRAM and in-plane STT-MRAM for a 1 Mb macro density at 15 nm node. 98

Fig. 5.17: J_{C0} and RA scaling trends of in-plane and perpendicular STT-MTJs. 99

Fig. 5.18: Perpendicular STT-MRAM scaling trends. (a) Sensing delay comparison with SRAM. (b) Write time comparison with SRAM................................100

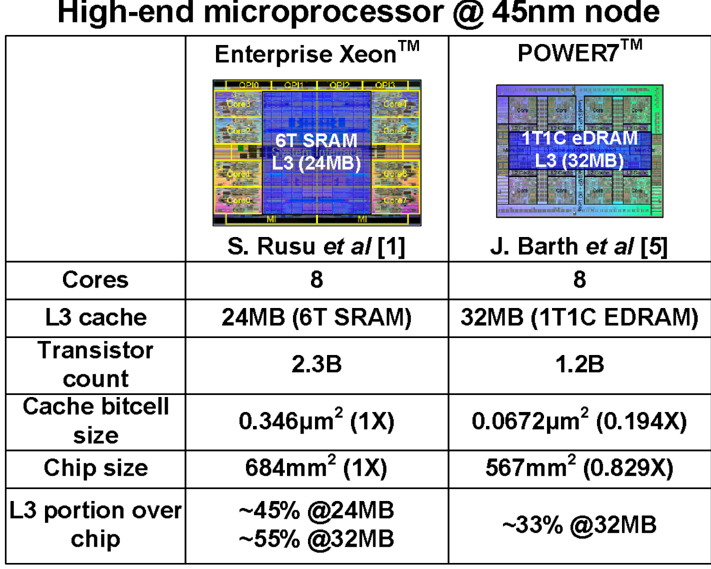
Chapter 1

Introduction

Multi-core processors exploit microarchitecture-level parallelism to deliver higher computing performance while curbing chip power dissipation. The number of cores per socket has increased at a pace of two per year for high end enterprise processors [1]. There needs to be a commensurate increase in the amount of on-die embedded memory in order to utilize the multi-core architecture fully with a larger appetite for data [1]-[3]. As a result, in the past decade, the die area devoted to cache memory has grown to approximately 50% in state-of-the-art processors (Fig. 1.1). For example, Intel's 8-core enterprise Xeon™ processor has a 24 MB Last Level Cache (LLC) [1] based on SRAM cells while IBM's POWER7™ processor has a 32 MB L3 cache built in an embedded DRAM (eDRAM) technology [4], [5]. The need for robust high-density embedded memories is projected to grow as designers continue to seek power-conscious ways to improve multi-core chip performance.

However, delivering dense embedded memories with higher performance for computing systems have faced with a unique scaling challenge compared to logic circuits. With technology scaling, transistor parameters and parasitic loadings are innovated in

such a way that enhances the performance of the core that is the most important building block in microprocessors. On the other side of the technology scaling, embedded memories, which are the other essential component in computing systems, have suffered from the reduced signal-to-noise margin due to the scaled power supply level for device reliability and power constraint. Furthermore, memory bit-cells implemented using minimally sized devices are susceptible to device mismatches, various noise couplings, and sense amplifier offsets resulting in the difficulty of maintaining the traditional scaling trend in advanced CMOS technologies such as sub-22 nm node.



High-end microprocessor @ 45nm node

	Enterprise Xeon™ (6T SRAM L3 (24MB)) S. Rusu *et al* [1]	POWER7™ (1T1C eDRAM L3 (32MB)) J. Barth *et al* [5]
Cores	8	8
L3 cache	24MB (6T SRAM)	32MB (1T1C EDRAM)
Transistor count	2.3B	1.2B
Cache bitcell size	0.346μm^2 (1X)	0.0672μm^2 (0.194X)
Chip size	684mm^2 (1X)	567mm^2 (0.829X)
L3 portion over chip	~45% @24MB ~55% @32MB	~33% @32MB

Fig. 1.1: High-end microprocessors with high density on-die L3 caches based on 6T SRAM and 1T1C eDRAM.

1.1 Embedded Memories in Multi-Core Microprocessor

6T SRAMs and 1T1C eDRAMs have been the embedded memory of choice. The logic compatible bit-cell, fast differential read, and refresh-free operation make 6T

SRAMs the most viable option for on-die cache memories. 1T1C eDRAMs have features such as small cell size, low cell leakage, and non-ratioed circuit operation. The smaller footprints of 1T1C eDRAMs reduce global interconnect delay and enable faster overall system performance in high density cache memories [4]-[9]. Fig. 1.2 shows the trend of Intel's and IBM's high-end microprocessors based on 6T SRAM cells [10]-[15] and 1T1C DRAM cells [5], [6], respectively. The number of cores per socket has doubled at every process generations with commensurate increases of cache densities, namely 4 MB L3 caches have been added for every additional core (Fig. 1.2(a)). However, the aforementioned trend has deviated from 45 nm due to the difficulty of achieving practical parallel computing performance as the chip design and programming become more complicated with the increased core count. Delivering a dense and stable embedded memory is another constraint limiting the continuance of the traditional way to achieve higher computing performance while curbing chip power dissipation. The relatively large cell size and conflicting requirements for read and write at low operating voltages make aggressive scaling of 6T SRAMs challenging in scaled CMOS technologies. The bit-cell scaling trend of the 6T SRAMs has deviated from 45 nm due to the above-mentioned technological difficulties, while the 1T1C eDRAMs have followed the 0.5X bit-cell scaling trend up to 45 nm (Fig. 1.2(b)). However, the noise margin of 1T1C eDRAMs is reduced substantially at low voltages as the read operation is based on the charge sharing principle, and difficulties in scaling the trench capacitor and the additional process steps involved in manufacturing the thick oxide (T_{OX}) access devices are currently limiting the wide spread adoption of 1T1C technology. As a matter of fact,

only a limited company has been able to deliver stable 1T1C eDRAMs for high density on-die caches [4]-[6].

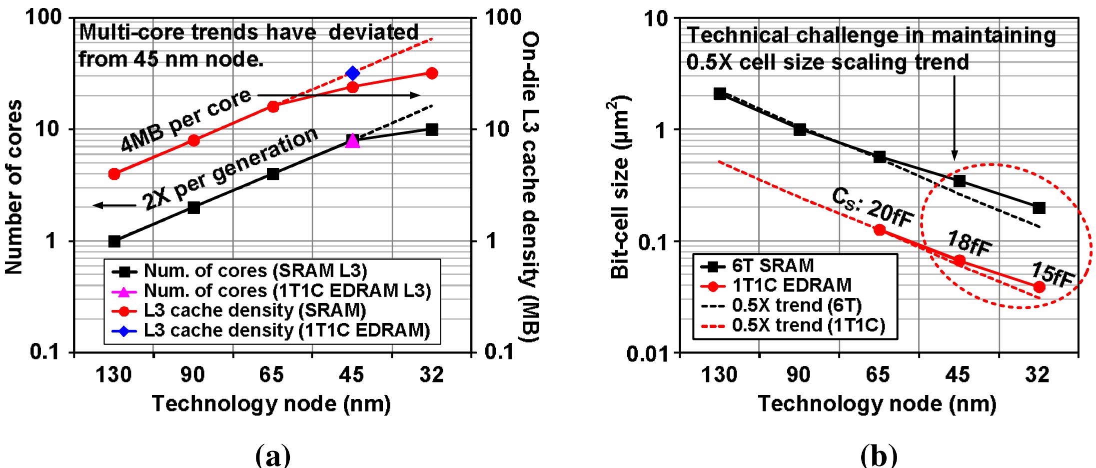


Fig. 1.2: (a) Trend of number of cores and corresponding on-die L3 cache densities in Intel's and IBM's high-end microprocessors. (b) Bit-cell size scaling trend of 6T SRAMs and 1T1C eDRAMs.

1.2 Alternative Memory Technologies

Gain cell eDRAMs and spin-torque-transfer magnetic RAMs (STT-MRAMs) are gaining popularity in the research community due to their compact bit-cells and excellent scalabilities. Gain cells are implemented using logic devices allowing them to be built in a standard CMOS process with minimal adjustments. The cell can be implemented using three transistors, or even two transistors when used with delicate read control circuits, achieving a roughly 2X higher bit-cell density than SRAM as recently demonstrated by several industrial designs [16]-[19]. Furthermore, gain cells can have smaller cell leakage current than SRAMs in sleep mode due to the fewer number of devices and the negative-Vgs biasing condition. Therefore, the static power dissipation of gain cell

eDRAM including both leakage and refresh components can be smaller than that of SRAMs and similar to that of 1T1C eDRAMs. The cell write margin is better than SRAMs since there is no contention between the access device and the cross-coupled latch in a gain cell. Despite these favorable attributes, conventional gain cells suffer from short retention times due to the small storage capacitor and leakage currents that vary exponentially under Process-Voltage-Temperature (PVT) variations [16]-[19]. A shorter retention time leads to higher refresh power dissipation and/or smaller read current. The former is a result of the frequent refresh operation while the latter is due to the fast loss of cell voltage. Frequent refresh operation also reduces memory availability resulting in degradation in overall system performance. Therefore, attaining practical retention time and improving random access speed remain as key challenges in gain cell eDRAM designs. In this dissertation, three unique test chip designs are presented to enhance the retention time and random access speed of gain cell eDRAMs for achieving overall faster system performances and lower power dissipations than SRAMs and 1T1C eDRAMs.

As the second part of this dissertation, STT-MRAMs are investigated to evaluate their potential as an alternative for high density on-die caches. An STT-MRAM bit-cell consists of an access transistor and a magnetic tunnel junction (MTJ). The simple structure makes the bit-cell size of STT-MRAM comparable to that of an eDRAM in a memory specific process. The MTJ device has a free magnetic layer and a pinned magnetic layer which are separated by a thin insulator layer. Depending on the direction of the write (WR) current, magnetization of the two layers can be set to a parallel state (P: low resistance, data '0') or an anti-parallel state (AP: high resistance, data'1') using spin

polarized current. Read (RD) operation is accomplished by sensing the resistance difference between the two states using voltage or current Sense Amplifier (S/A).

The state reversal happens only to the selected bit-cell when the current flowing into the MTJ is larger than its threshold current (write threshold current; I_{C0}). The STT switching originates from the exchange of angular momentum between a spin-polarized current and the magnetization of the free layer. The localized spin-injection within a bit-cell enables the excellent write selectivity with no high oxide field involved nor high temperature required during the process, which are the most critical scaling challenges in currently popular non-volatile memories such as FLASHs and phase-change RAMs (PCRAMs or PRAMs). Most interestingly, the I_{C0} decreases exponentially with technology scaling as the critical current density (J_{C0}) remains constant due to the STT switching phenomenon when there is no thermal stability constraint. Therefore, the slow write time (T_{WR}) which is several nanoseconds or even larger, is projected to be relieved with technology scaling [20]-[22]. Despite the recent advances in STT-MRAM fabrication and circuit techniques [23]-[32], it is still unclear whether this emerging memory technology can achieve higher overall performance than conventional SRAMs or eDRAMs in future technology nodes in the presence of variation effects. In this dissertation, I explore the scalability and variability of STT-MRAM by comparing its performance with 6T SRAM from 65 nm to 8 nm process nodes.

1.3 Summary of Dissertation Contributions

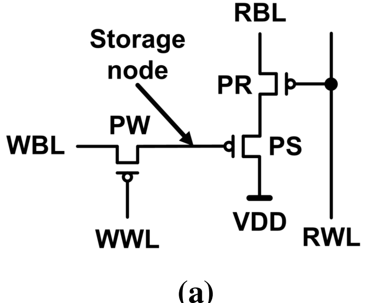
The remainder of this dissertation will explore the benefits of various circuit techniques and HSPICE simulation methodologies that we have proposed to demonstrate the potential of gain cell eDRAMs and STT-MRAMs as alternative options for high density embedded memories. Three unique test chip designs will be presented to enhance the retention time and read speed of gain cell eDRAMs, enabling faster overall system performances and lower static power dissipations than SRAMs and eDRAMs [33]-[37]. An MTJ scaling scenario, an efficient HSPICE simulation methodology, and a characterization macro are described for exploring the scalability of STT-MRAMs under variation effects from 65 nm to 8 nm, demonstrating that the alternative option can outperform SRAMs and eDRAMs in advanced CMOS technologies [38].

Chapter 2

A 3T Embedded DRAM Utilizing Preferential Boosting for Low Voltage On-Die Caches

2.1 Basic Operation of a Conventional 3T EDRAM

To aid the understanding of our proposed techniques, in this section, the basic operation of a conventional 3T gain cell eDRAM is described. Fig. 2.1(a) shows the cell schematic and Fig. 2.1(b) summarizes the signal conditions for each operating mode.



(a)

	Data hold	Read 0/1	Write 0/1
RWL	VDD	0V	VDD
RBL	0V	200mV/0V	0V
WWL	VDD	VDD	-500mV
WBL	0V	0V	0V/VDD

(b)

Fig. 2.1: (a) Conventional 3T PMOS eDRAM gain cell circuit diagram. (b) Signal voltages in each operating mode.

PMOS devices are chosen over NMOS devices because they have significantly less gate tunneling leakage current, which extends the data retention time [19]. This

preference may not hold in the future where high-*k* gate dielectrics become prevalent. The operating principle of an NMOS cell is identical to that of a PMOS cell with the only difference being the signal polarities. In the 3T PMOS cell, PW denotes the write access device, PS denotes the cell storage device, and PR denotes the read access device. In write (or write-back) mode, the Write Bit-Line (WBL) data is written into the storage node through PW. Similar to a 1T1C eDRAM cell, the Write World-Line (WWL) is negatively over-driven so that a 0 V can be written into the cell without the threshold voltage loss. In read mode, the pre-discharged Read Bit-Line (RBL) voltage is pulled up only when the voltage stored in the gate of PS is low. In case the storage voltage is high, PS is off so RBL remains at the pre-discharged level. Cell data can be determined by comparing the RBL voltage with a reference RBL, whose level is between the data '1' and data '0' RBL levels, using a sense amplifier.

During hold mode, PW and PR are turned off and the storage node is left floating. The sub-threshold, gate, and junction leakages in the surrounding devices make the floating voltage change with time as shown in Fig. 2.2. Since the storage node is surrounded by high voltages in the PMOS cell, the retention time of data '0' is much shorter than data '1'. Similarly, the retention time of data '1' becomes critical in an NMOS cell where the surrounding signal voltages are 0 V during hold mode. The data retention time is directly related to the aggregated leakage currents flowing into the storage node. In the presence of process variation, each cell in a memory array will have different retention characteristics so the cell with the shortest retention time (after

applying any redundancy schemes to remove bad cells) will determine the refresh rate of the entire eDRAM array.

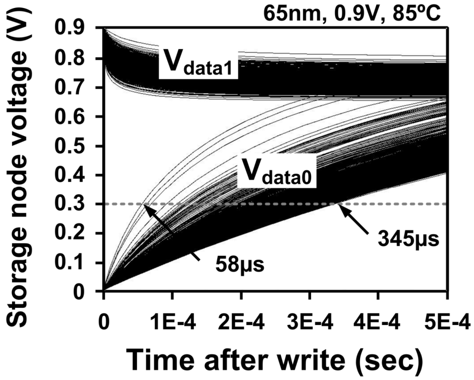


Fig. 2.2: Monte Carlo simulation results of storage node voltage during data hold mode.

Fig. 2.2 shows the simulation results of cell retention time variation. This plot was obtained by running Monte-Carlo simulations in HSPICE with 1024 iterations, which gives a cell-to-cell variation equivalent to a 1 kb array. Results indicate that the time it takes for the data '0' voltage to rise to a specific voltage (0.3 V in this simulation to guarantee a 0.3V gate over-drive voltage in the storage transistor which has a V_{TP} of 0.3 V) ranges from 58 μs to 345 μs at a 0.9 V supply voltage and 85 °C temperature. Poor retention characteristics of tail cells result in a large refresh current and decreased read performance. Therefore, increasing the cell retention time is the foremost challenge in low voltage gain cell eDRAMs.

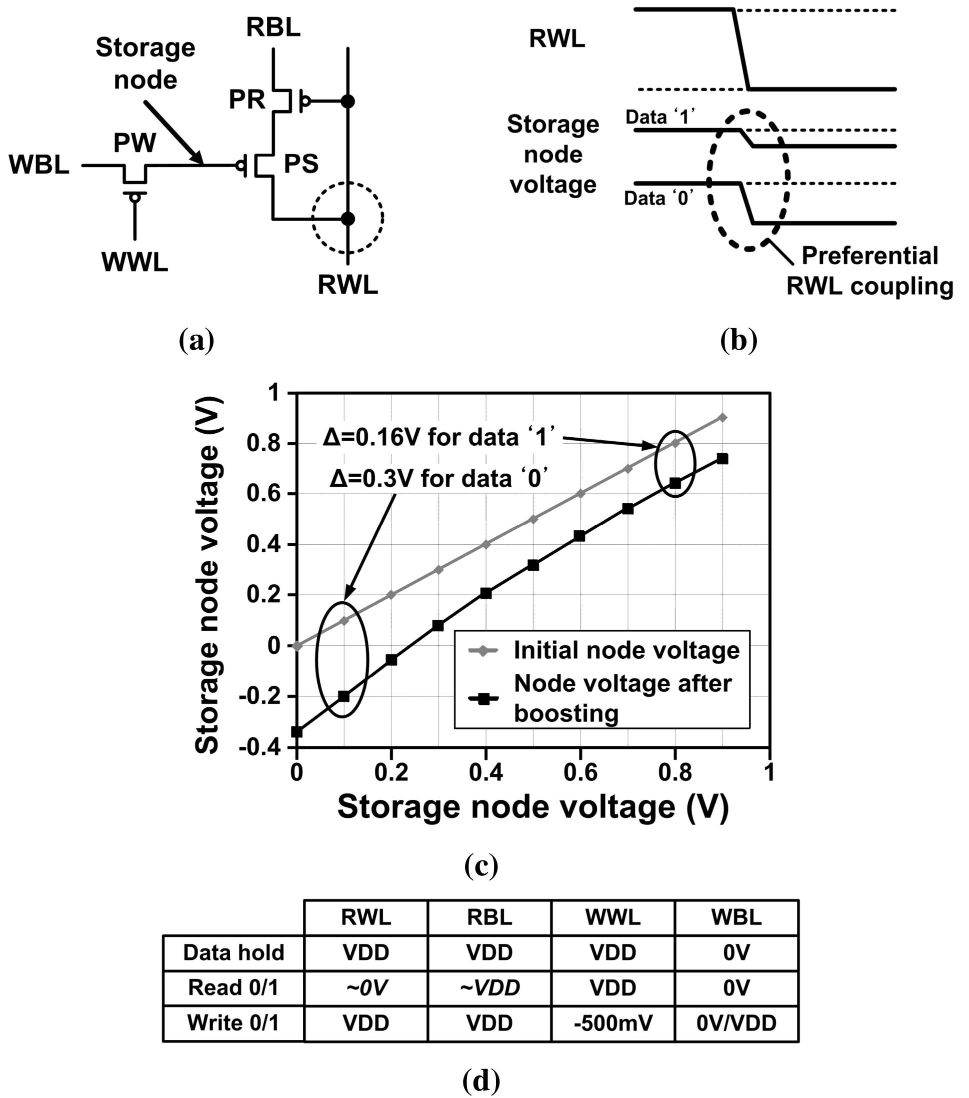
2.2 Boosted 3T EDRAM Design

In this section, three circuit techniques are presented for improving the eDRAM data retention time and ensuring robust circuit operation under PVT variations.

2.2.1 Boosted 3T Gain Cell

The retention time and read speed of eDRAMs are highly dependent upon the storage node voltage at the time when the cell is accessed. Even a small signal loss can cause severe speed degradation at low operating voltages. Fig. 2.3(a) shows the proposed 3T PMOS gain cell which can preferentially boost the storage voltage via capacitive coupling. Unlike the conventional design in Fig. 2.1(a), the drain of the storage device PS is connected to the RWL signal instead of the supply voltage. For read operation, RBL is first precharged to VDD and then the RWL switches from VDD to 0V. The resultant bitline signal is detected by a sense amplifier. The central idea of the proposed cell is to preferentially boost the storage node voltage using the RWL signal for improving the cell's data retention capability. For example, consider the case when the storage node voltage is low (e.g. 0V). This will make the gate-to-RWL coupling capacitance larger compared to when the storage node voltage is high (e.g. VDD). PS in inversion mode makes the entire oxide capacitance act as the coupling capacitance whereas PS in weak-inversion mode, the significantly smaller depletion capacitance acts as the coupling capacitance. Since a lower storage voltage has a larger coupling capacitance, it is coupled down more than a higher storage voltage when the RWL switches from high to low as illustrated in Fig. 2.3(b). This preferential boosting action amplifies the signal difference during read which allows the storage node voltage to decay further before it needs to be refreshed. This translates into a longer effective data retention time. A similar concept was proposed by Luk *et al.*, where a 3T1D cell was used to boost the cell voltage [18]. However, this cell structure requires an additional

diode device which increases the cell area as well as the gate tunneling leakage. It also has a limited signal amplification effect since the storage device acts as a parasitic capacitor limiting the amount of coupling that can be achieved. The proposed boosted 3T gain cell can provide a stronger coupling effect with only three transistors, increasing data retention time, enhancing the RBL margin and improving read performance.



	RWL	RBL	WWL	WBL
Data hold	VDD	VDD	VDD	0V
Read 0/1	*~0V*	*~VDD*	VDD	0V
Write 0/1	VDD	VDD	-500mV	0V/VDD

Fig. 2.3: (a) Proposed boosted 3T PMOS eDRAM gain cell. (b) Preferential RWL coupling effects of the proposed cell. (c) Simulation results of the storage node preferential boosting effects. (d) Signal voltage conditions for each operating mode.

Simulation results in Fig. 2.3(c) verify that the data '0' voltage is amplified by 0.3 V while the data '1' voltage is coupled down by only 0.16 V. In addition to the amplification effect, the proposed cell can provide a ~2X larger current than conventional 3T gain cells since the boosted voltage provides a higher gate overdrive for PS. Fig. 2.3(d) summarizes the signal conditions for each operating mode for the proposed gain cell. It should be pointed out that the higher drive current is only observed when the RBL level is high, as the read current quickly diminishes as the RBL voltage drops due to the V_{TP} loss in the PMOS read device. To utilize the boosted read current of the proposed 3T cell, we employ a hybrid current/voltage sense amplification technique that keeps the RBL level close to VDD during the read operation [39], [40].

Fig. 2.4 shows the schematic and timing diagram of the bit-line Sense Amplifier (S/A) consisting of a hybrid current/voltage S/A, read port, write port and drivers for write-back. During read, the RBL signals to the current S/A are amplified and converted to voltage signals through a cross-coupled PMOS pair and a NMOS resistor pair while a load PMOS pair keeps the RBL swing small. After transferring the input differential current, the cross-coupled PMOS pair, in tandem with the cross-coupled NMOS pair, acts as a voltage S/A which generates a full CMOS swing signal. Dedicated timing control circuits are implemented for the equalizer to ensure stable current S/A operation as shown in Fig. 2.4(b). The write-back operation automatically follows the read cycle to refresh the cell data.

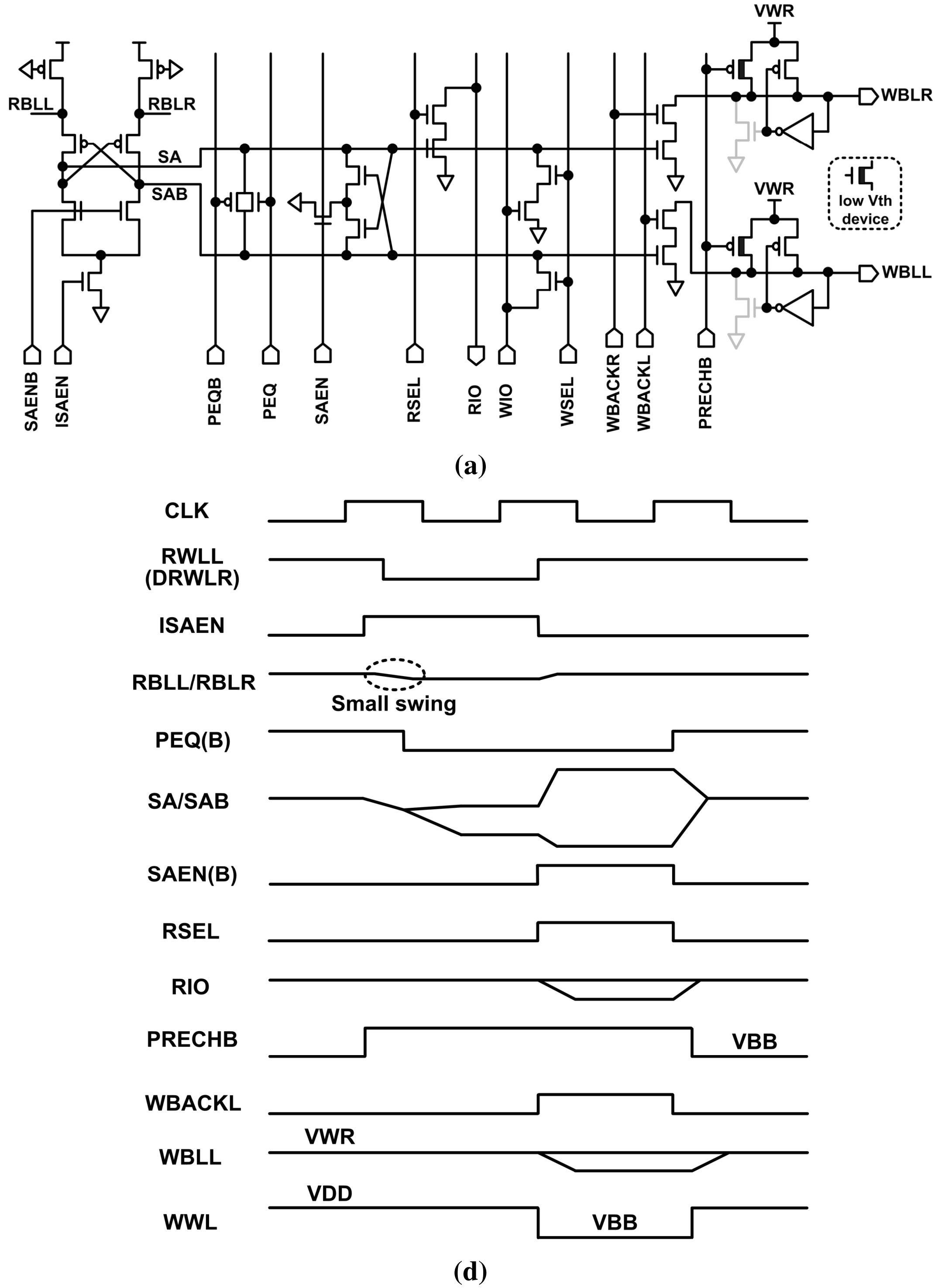


Fig. 2.4: (a) Hybrid bit-line current/voltage sense amplifier (S/A) with read port, write port, and write-back circuits. (b) Read and write-back timing diagram of the proposed S/A.

2.2.2 Regulated Bit-Line Write Scheme

When the WBL is driven to data ‘1’, the data ‘0’ levels in the unselected cells on the same WBL are pulled up by the sub-threshold leakage through the write access PMOS devices as shown in Fig. 2.5(a). Most DRAM designs use a boosted supply for the WWL to prevent the signal loss in the unselected cells by asserting a negative Vgs in the write access devices. However, this method incurs area and power penalty due to the large charge pump capacitors and poor pumping efficiency at low voltages. In this work, we propose a regulated bit-line write scheme which can eliminate the data ‘1’ disturbance issue without having to generate an additional boosted supply.

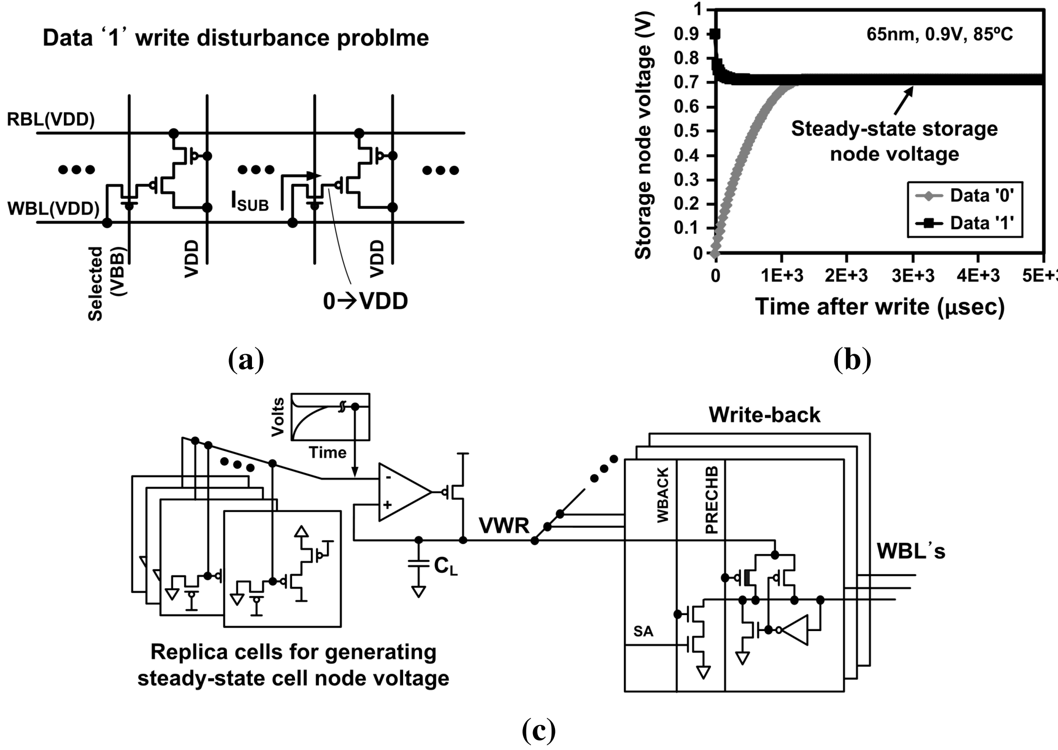


Fig. 2.5: (a) Storage node disturbance problem when writing data ‘1’ to a cell sharing the same WBL. (b) Simulation results showing steady-state storage node voltage in case of no refresh. (c) Proposed regulated bit-line write bias generator based on replica cells.

Without a refresh, the storage node voltage eventually converges to a steady-state level close to VDD regardless of the initial cell voltage as shown in Fig. 2.5(b). In our design, we use this steady-state voltage level for writing data '1', as it will produce a negative Vgs in all the unselected cells without impacting the retention time of the selected cell. Note that the retention time is determined by the data '0' cell voltage rather than the data '1' voltage in a PMOS gain cell. A steady-state storage node voltage monitor shown in Fig. 2.5(c) is implemented with replica cells biased in hold mode, followed by a voltage down converter to drive the large WBL load. The speed loss due to the regulated bit-line write voltage (VWR) is prevented by pre-charging the WBL to VWR using the negative supply VBB as the gate signal, which is readily available on-chip for the WWL under-drive.

2.2.3 PVT-Tracking Read Reference Bias

An optimal bias voltage (VDUM) is applied to the reference dummy cells to maximize the read operating margin. VDUM must be carefully chosen as it affects both the data retention time and the read speed; a higher VDUM level improves the data retention time at a read speed penalty. Fig. 2.6 shows the proposed PVT-tracking and die-to-die adjustable read reference bias generator to cope with PVT variations. The negative feedback circuit tracks the desired cell read reference current (I_{REF} in the figure). Fig 2.7 shows simulation results of the proposed VDUM level under PVT variations. Unlike previous designs which use a fixed VDUM level or a simple averaging scheme [19], our circuit can achieve the target retention time without sacrificing read speed by adaptively lowering the VDUM level at low leakage PVT conditions as shown in Fig. 2.7.

For example, at lower temperatures or in slow corner dies, the excess retention time is traded off for faster read speed by lowering the VDUM level. Similarly, at low supply voltages, the VDUM level is shifted down since the reduced leakage make the storage node voltage lower compared to at high supply voltages for the same retention time. Binary weighted read path replica branches are implemented to precisely adjust the VDUM level according to the retention characteristics and read performance of each chip.

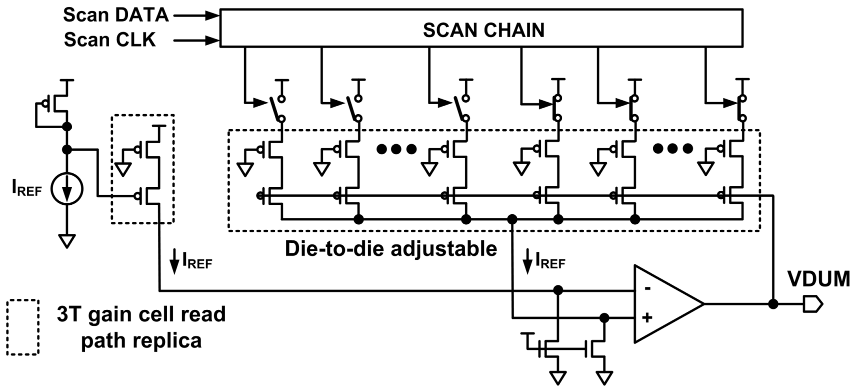


Fig. 2.6: PVT-tracking and die-to-die adjustable read reference bias (VDUM) generator.

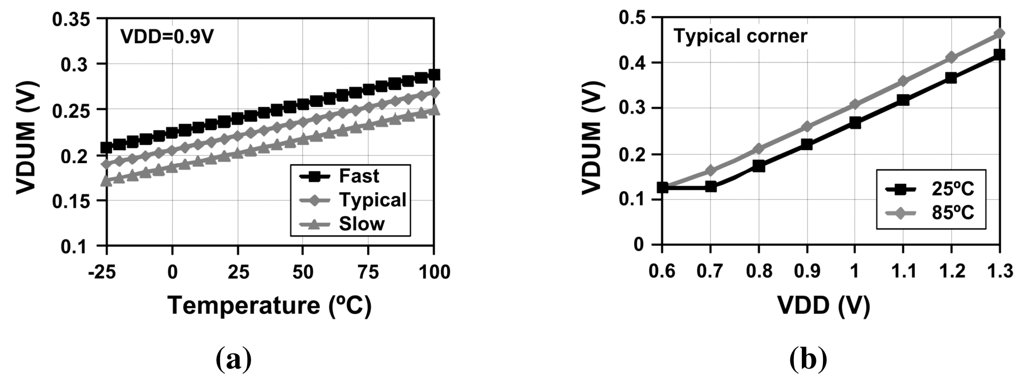


Fig. 2.7: (a) Simulation results of the proposed VDUM generator tracking temperature and process variations. (b) Simulation results showing the dependency of VDUM on VDD.

2.2.4 Architecture and Operation of a 32 kb Sub-Array

A detailed circuit diagram of the 32 kb boosted 3T array is shown in Fig. 2.8. The array has 128 cells per WL and 128 cells per split BL, which share a common BL S/A located at the center of the array. The proposed VDUM bias is connected to the dummy cells placed at both edges of the array, and the VWR bias is connected to the write-back circuitry of the BL S/A. The RWL pull-down keepers are located at the top row of the array to keep the ground noise of the activated RWL as small as possible. HSPICE simulations indicate a 66 mV RWL ground noise at 0.9 V, 85 ºC when all cells connected to the same RWL contain data '0' which corresponds to the worst case scenario.

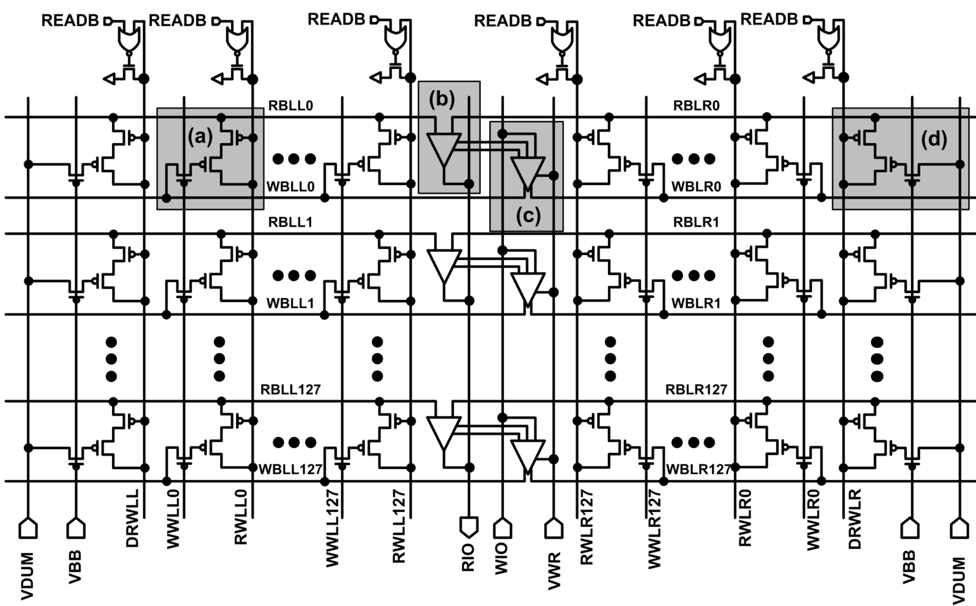


Fig. 2.8: A 32 kb array structure of the proposed eDRAM including (a) boosted 3T gain cell, (b) hybrid current/voltage S/A, (c) regulated bit-line write scheme, and (d) PVT-tracking read reference scheme.

Fig. 2.9 shows simulation waveforms of read and write-back operations with a 2 ns random cycle time. A two-stage full pipeline structure was implemented to control read and write-back operations. At the first clock cycle, RWL is selected, and this amplifies

the cell node by preferential coupling. When the current S/A control signal (ISAEN) is enabled, the current S/A amplifies its input signals to analog voltage signals with RBL held close to VDD. After achieving a recognizable voltage difference, the voltage S/A control signal (SAEN) is enabled. At the second clock cycle, read-out and write-back operations are followed. After write-back, discharged WBLs are pre-charged using the negative supply VBB control signal (PRECHB).

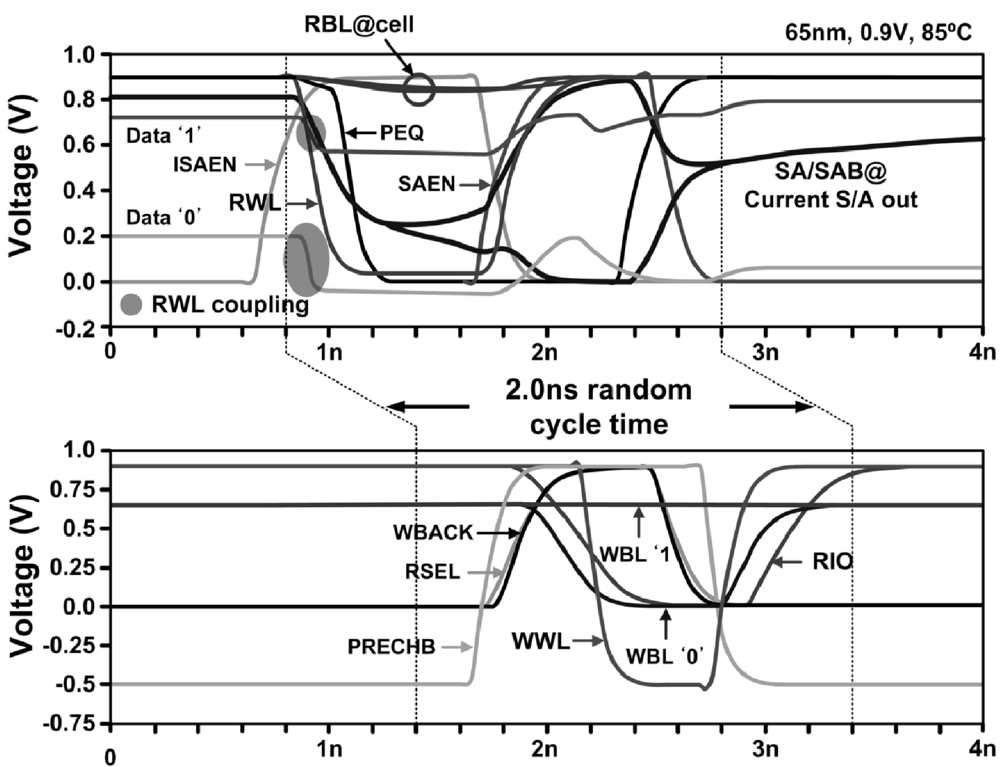


Fig. 2.9: Read and write-back simulation waveform with a 2 ns random cycle time.

2.3 Statistical Simulation Results for 6T SRAM and 3T EDRAM Arrays

This section presents Monte-Carlo simulation results on megabit density SRAM and eDRAM arrays to estimate their speed and power in a practical scenario [41]. An operating voltage of 0.9 V was chosen (nominal operating voltage of the 65 nm process used is 1.2 V) so that cell failures exist in the small 32 kb unit test array. Fig. 2.10 summarizes the simulation setup for the Monte-Carlo iterations including assumptions on the mismatch and voltage variations.

		CONV 3T eDRAM	Proposed 3T eDRAM	6T SRAM
0.9V, 85°C, 1M Monte Carlo full array simulation w/ 1.2V, 65nm, LP CMOS process				
Read operation	Cell node voltage	@100μs with voltage distribution under T_{OX} and V_{TH} variations		N/A
	Reference bias	Adaptive VDUM with 10% variations		N/A
	Cell	Device mismatches		
	Dummy cell	Dummy cell averaging scheme [19]	4X upsized device mismatches	N/A
	Current S/A	N/A	S/A pair mismatches	N/A
Write operation	Boosted supply	-0.5V with 10% variations		N/A
	Cell	Device mismatches		

*Mismatches (i.e. sigma of Vt) are based on process parameters provided in the design kit adjusted by individual device dimensions using the inverse square root relationship

Fig. 2.10: Simulation setup for 1 M Monte Carlo simulations.

2.3.1 Read and Write Performance

Fig. 2.11 shows read bitline delay distributions with average and 6-sigma point delays annotated for the following three memory arrays; a 1Mb SRAM, a 2 Mb conventional 3T, and a 2 Mb boosted 3T. Simulation results were obtained from 2^{20} Monte-Carlo

iterations. The peripheral circuit delay, which is a function of the unit sub-array size, and the global interconnect delay, which is a function of the total cache area, are identical for the three simulated arrays since we selected an SRAM with half the number of cells as the eDRAMs. Recall that an SRAM bitcell is about twice the area of an eDRAM bitcell. The single-ended sensing nature and the gradual loss in the storage node voltage of the conventional 3T eDRAM result in a 6-sigma read bit-line delay that is 2.6 times longer than a 6T SRAM as shown in Fig. 2.11. The proposed 3T eDRAM with preferential amplification effect partially makes up for this performance shortfall, improving the bit-line sensing speed by 36% compared with the conventional 3T eDRAM. Although 6T SRAMs still have a 40% faster sensing delay than the proposed circuit, we will see later that their performance becomes worse than eDRAMs for large cache sizes due to the longer global interconnect delay.

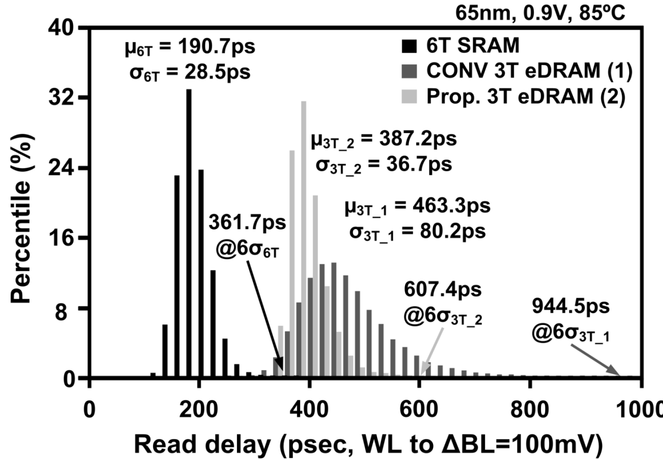


Fig. 2.11: Read performance comparisons between 6T SRAM and 3T eDRAM obtained from 2^{20} Monte-Carlo iterations. Results are equivalent to the distribution of a 1 Mb macro array. 6T SRAM has the shortest bitline delay attributed to the differential swing nature and large drive current (361.7ps @ 6σ) followed by the proposed 3T eDRAM (607.4ps @6σ) and the conventional 3T eDRAM (944.5ps @6σ).

Fig. 2.12 shows detailed cell layouts of various logic-compatible embedded memory cells drawn using a standard 65nm logic design rule. The dense bitcell design rules were not available to the authors but for area comparison purposes, using a logic design rule is generally sufficient. The four signal wire lines and the three transistors of the conventional boosted 3T gain cells are marked in Fig. 2.12. The proposed boosted 3T gain cell is 47% smaller than a 6T SRAM cell.

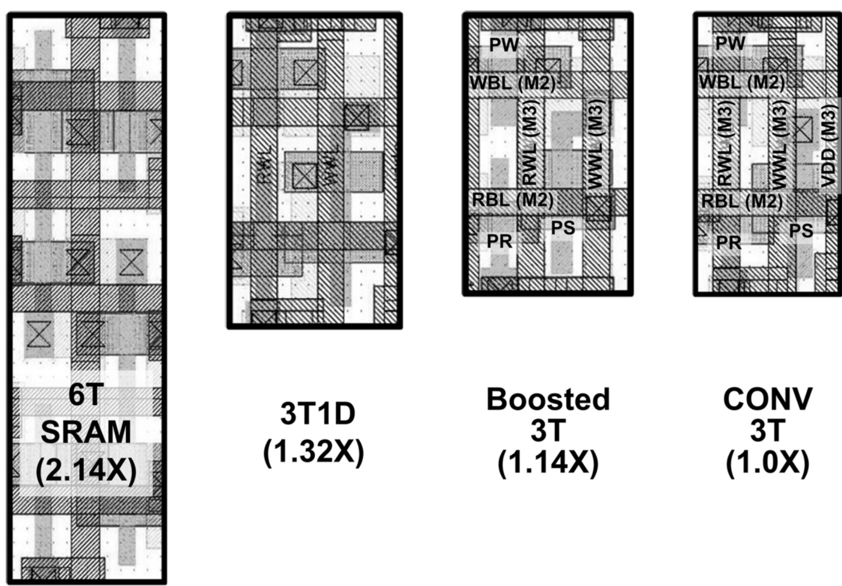


Fig. 2.12: Comparison of various logic-compatible embedded memory cell layouts using a 65nm logic design rule (the authors did not have access to the dense bitcell design rule but for area comparison purposes, the logic design rule is generally acceptable). The outer box represents the cell boundary. Signal names, wire tracks, and device names are marked for the boosted 3T and conventional 3T cells.

Fig. 2.13 shows latency comparison results between a 6T SRAM array and the boosted 3T eDRAM array for two different cache sizes. The latency of a cache shown in Fig. 2.13 consists of the bit-line sensing time (6-sigma value from Fig. 2.11), the peripheral circuit delay, and the global interconnect delay. The boosted 3T eDRAM

achieves faster access times for cache sizes greater than 16 Mb (or 2 MB) owing the shorter interconnect delay made possible by the smaller bitcell.

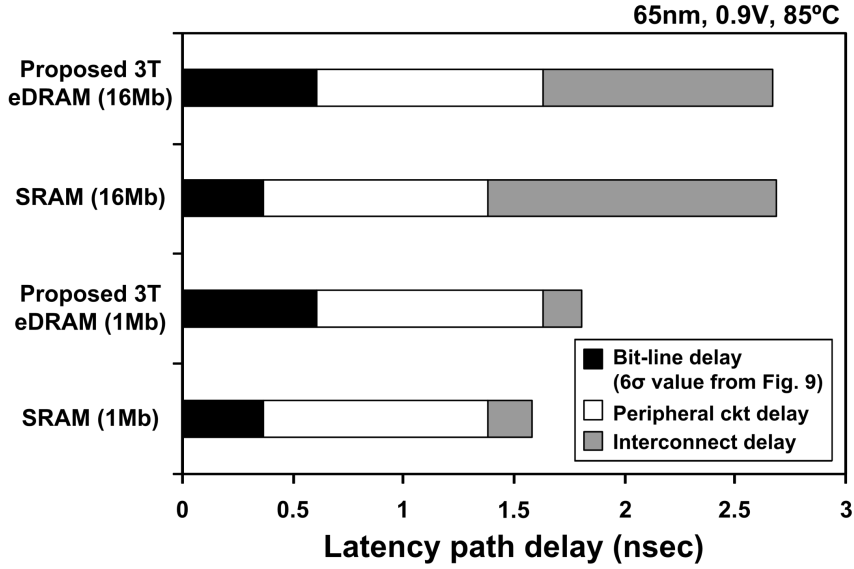


Fig. 2.13: Latency comparisons between SRAM and 3T eDRAM for 1 Mb and 16 Mb cache sizes. Gain cells have a shorter interconnect delay due to the smaller cell size making their performance favorable in larger arrays.

Fig. 2.14 shows the 1 Mb write delay distributions of a 6T SRAM array and the proposed 3T eDRAM array. Here, the write delay is defined as the WL signal to the time when the cell node reaches 95% of the full voltage swing. The write speed of the gain cell is faster than the 6T SRAM since the latter is based on a ratioed operation. Note that the WWL of the gain cell must be sufficiently negative in order for the PMOS write devices to pass a good data '0' level. For a WWL under-drive voltage of -0.5V, the 1Mb Monte-Carlo simulations show a write speedup of 17 % (6-sigma point) for the boosted 3T eDRAM.

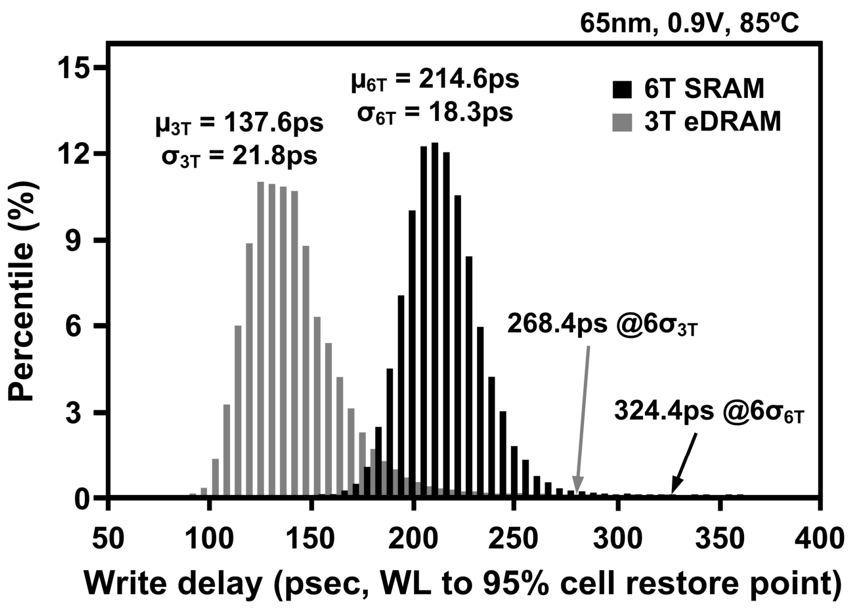
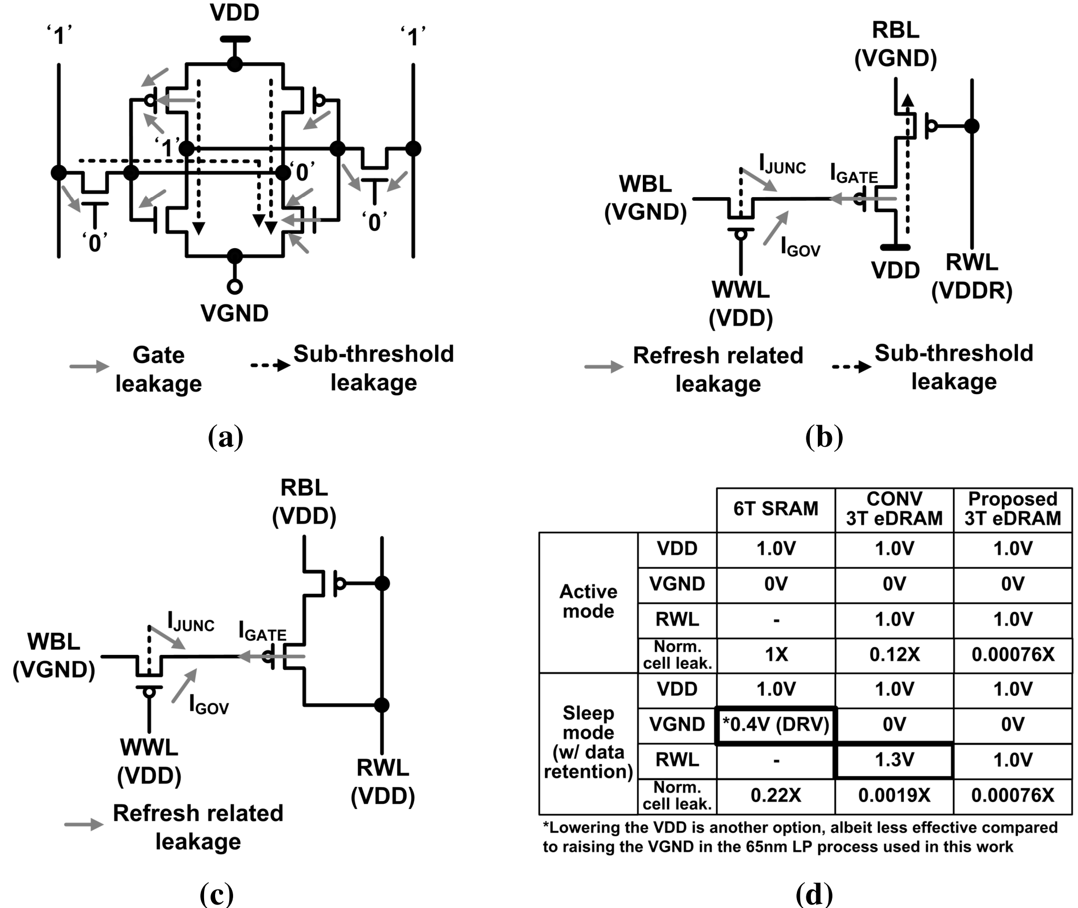


Fig. 2.14: Write delay distributions of 1Mb arrays using 6T SRAM and 3T eDRAM.

2.3.2 Static Power Consumption

Static power consumption of an eDRAM system consists of two main components: (i) the leakage current of the cell itself and (ii) the refresh power to keep the data "alive". The refresh operation is a dummy read followed by a write-back cycle which simply reinforces the cell data. Hence, the refresh power is inversely proportional to refresh period. The data '0' storage node voltage should be kept sufficiently low so that the PMOS read device can provide enough drive current that meets the target read speed. This criterion determines the refresh period as pointed out in section 2.1. Fig. 2.15 (a), (b), and (c) illustrate the leakage components in the three memory cells. Due to the higher number of devices per cell, there are more leakage paths from the supply to the ground in a 6T SRAM cell than in the 3T eDRAM cells. Since the leakage current through the storage node has to be extremely small in an eDRAM cell for it to be viable (e.g., >100 μs retention time), the main cell leakage component is through the read access

device. In other words, the refresh related leakages shown in Fig. 15 (b) and (c) are much smaller than the leakage current through the read access device.



		6T SRAM	CONV 3T eDRAM	Proposed 3T eDRAM
Active mode	VDD	1.0V	1.0V	1.0V
	VGND	0V	0V	0V
	RWL	-	1.0V	1.0V
	Norm. cell leak.	1X	0.12X	0.00076X
Sleep mode (w/ data retention)	VDD	1.0V	1.0V	1.0V
	VGND	*0.4V (DRV)	0V	0V
	RWL	-	1.3V	1.0V
	Norm. cell leak.	0.22X	0.0019X	0.00076X

*Lowering the VDD is another option, albeit less effective compared to raising the VGND in the 65nm LP process used in this work

Fig. 2.15: Leakage components of a (a) 6T SRAM, a (b) conventional 3T eDRAM and the (c) proposed 3T eDRAM. (d) Bias conditions and normalized cell leakages of SRAM and 3T eDRAM in active and sleep modes.

Fig. 2.16 compares the static power consumption of a 1 Mb 6T SRAM array and a 2 Mb 3T eDRAM array with a 100 μs refresh period. HSPICE simulations were performed using a 65 nm low-leakage CMOS process at 1.0 V, 85 ºC. Again, the number of cells of the 3T eDRAM array was chosen to be twice that of the SRAM array to account for the ~50% smaller cell size. Note that the eDRAM's higher density makes up for its longer

latency improving the overall architectural performance [4], [5]. Simulation results show that the static power of a 2Mb conventional 3T eDRAM array is similar to that of a 1 Mb SRAM during active mode. The refresh current consists of the RBL and WBL switching currents for the dummy read and write-back operations, as well as the refresh control power in the peripheral circuits. The refresh power constitutes 75% of the total eDRAM static power for a 100 μs refresh period.

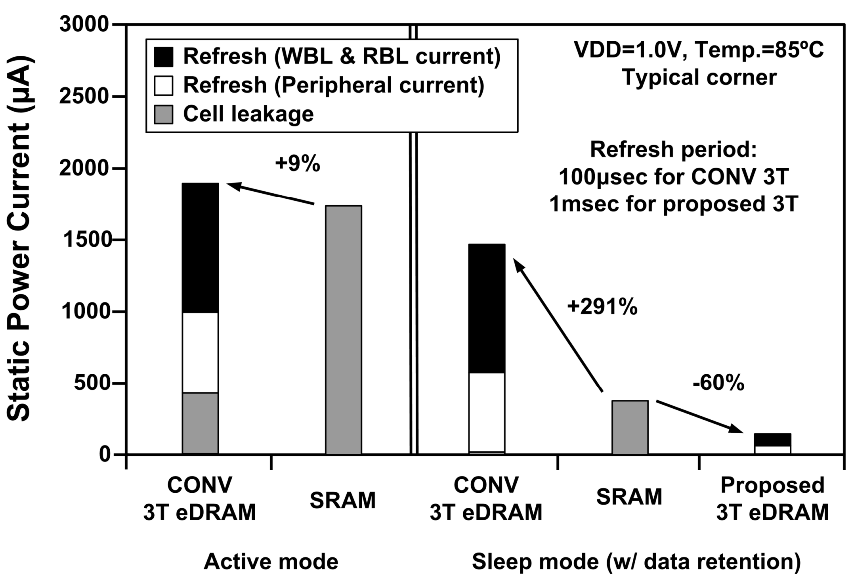


Fig. 2.16: Static power comparisons between a 1 Mb SRAM and a 2 Mb 3T eDRAM. Leakage power of the peripheral circuit is assumed to be negligible.

Most embedded memories are now equipped with sleep mode capability, so it is important to compare the sleep mode power between SRAM and the proposed eDRAM. When power gating and wordline overdrive techniques shown in Fig. 2.15(d) are applied, the cell leakage component is reduced in both the SRAM and the eDRAM arrays [12], [42]. Since refresh power is not affected by these sleep techniques, the eDRAM's total static power becomes 3X larger compared to the SRAM's even with an additional boosted high supply for the RWL to suppress the read path sub-threshold leakage as

shown in Fig. 2.15(b). Our proposed 3T eDRAM cell significantly reduces the refresh power component as it has a 10X longer retention time without any extra boosted supply. This makes the static power of the proposed eDRAM 53% less than that of a power gated SRAM, as shown in Fig. 2.16.

Fig. 2.17 summarizes simulation and layout results of various logic-compatible embedded memory cells.

65nm, 0.9V, 85°C

	CONV 3T (2T [19])	3T1D [18]	Proposed 3T [33]	6T SRAM [12]
*Cell schematic	WBL, RBL, WWL, RWL	WBL, RBL, WWL, RWL	WBL, RBL, WWL, RWL	WL, WL, BL, BLB
Features	Small size	Partial storage node amplification	Full storage node amplification	Fast
Issues	Short retention time	Additional device	RWL noise	Large size, low noise margin
**Cell size (ratio)	0.54x1.02= 0.551μm^2 (1.0X)	0.64x1.14= 0.73μm^2 (1.32X)	0.615x1.02= 0.627μm^2 (1.14X)	0.575x2.05= 1.178μm^2 (2.14X)
***RWL-BL delay (Δ=100mV)	945ps	794ps	607ps	362ps
***WWL-Cell 95% restore delay	-	-	268ps	324ps
Latency (simulated)	-	-	1.81ns @1Mb 2.67ns @16Mb	1.58ns @1Mb 2.69ns @16Mb
Retention time	110μs (measured)	200μs (simulated)	1.25ms (measured)	-
Static power	Large due to short retention time	Medium	Small	Large due to transistor leakage

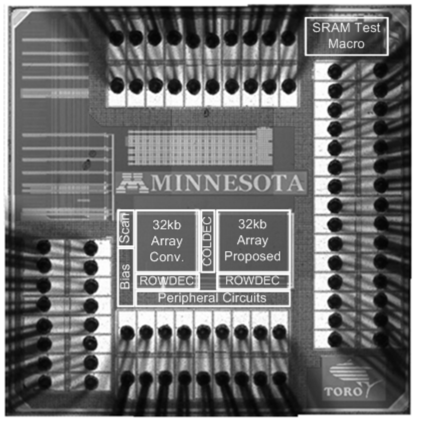
* PMOS cells for low I_{gate}, ** 65nm logic design rule, *** Monte-carlo 6σ simulation results

Fig. 2.17: Comparison of logic-compatible embedded memories.

2.4 Test Chip Implementation and Measurements

A proof-of-concept 64 kb eDRAM test chip was built in a 1.2 V, 65 nm low-leakage logic CMOS process to demonstrate the proposed circuit techniques. In order to fully verify the proposed techniques against the existing ones, each sub-array has a different combination of cell structure (boosted 3T vs. conventional 3T), reference scheme

(proposed PVT-tracking vs. cell averaging [19]), and write scheme (conventional vs. regulated bit-line write). Fig. 2.18 shows the chip microphotograph and feature summary of the 64 kb eDRAM test chip fabricated in a 1.2 V, 65 nm low-leakage logic CMOS process.



Process	65nm LP CMOS (7-MET)
Circuit dimension	250.7x129.9μm^2
Cell size	53% of 6T SRAM
Retention time	> 1.25ms @ 0.9V, 85ºC
Operating voltage	1.2V - 0.7V
Preferential boosting @ 1.0ms	0.27V for data '0' 0.14V for data '1'
Cycle time	2.0ns @ 0.9V (simulation)
Refresh power	91.3μW per Mb @ 1V, 85ºC, 1.0msec refresh rate

Fig. 2.18: Microphotograph of the 65 nm eDRAM test chip and feature summary.

Fig. 2.19(a) shows the measured VWR levels at different supply voltages. The data '1' voltage (i.e., VWR) is high enough to keep the storage transistor off: the PMOS threshold voltage (V_{TP}) of this process is 0.315 V at 85 ºC and the measured VWR level is slightly lower than VDD-V_{TP}. The unselected cells undergoing the data '1' disturbance situation are not affected since a sufficient amount of negative Vgs is applied to the write access transistor. The VWR level is determined by the balance between the sub-threshold, gate, and junction leakage components. In most cases, sub-threshold leakage is the dominant factor in determining the VWR level. At high temperature and high VDD conditions however, the junction and gate leakage components have a stronger affect on the VWR level than the sub-threshold leakage component resulting in higher level over 1.1 V as shown in Fig. 2.19(a).

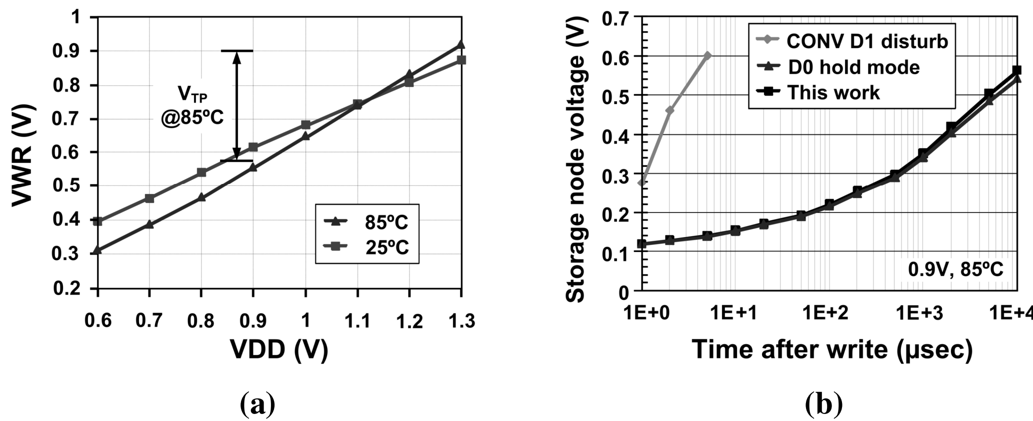


Fig. 2.19: (a) Measured regulated bit-line write bias (VWR) level. (b) Storage node voltage measurement results under data '1' disturbance conditions.

By externally adjusting the VDUM voltage, we can indirectly and noninvasively measure the storage node voltage at different data retention times. For example, read failure will happen for data '0' if the VDUM level is lower than the storage node voltage so the storage voltage can be measured by sweeping the VDUM voltage and measuring the failure point. It is worth mentioning that the storage node voltage measured using this method include effects such as process variation or transient noise (e.g. coupling noise or supply noise) providing us with an "effective" cell node voltage. Fig. 2.19(b) shows the measurement results of the storage node voltage of the proposed regulated write scheme compared with the conventional 3T gain cell under the data '1' disturbance condition. The data retention characteristics of the data '1' disturbance case and the data hold mode case are virtually identical when using the proposed regulated bit-line write scheme.

Fig. 2.20(a) shows the data retention characteristics of the conventional 3T and the proposed boosted 3T from the same test chip, including the cell-to-cell retention time variation. The retention time was for a read speed (i.e., RWL enable to voltage S/A

enable interval) of 1.0 ns at 0.9 V and 85 ºC. This translates into a 2.0 ns cycle time. The proposed boosted 3T design achieves a data retention time of 1.25 ms at 0.9 V, 85 ºC, which is a 10X improvement over the conventional 3T cell measured from the same silicon die. Note that due to limitation in the test setup, only 32 cells were measured from each sub-array. As a point of reference, the target retention time of a 2T gain cell eDRAM was 10 μs in [19] and the measured retention time of a 1T1C eDRAM was 40 μs in [6].

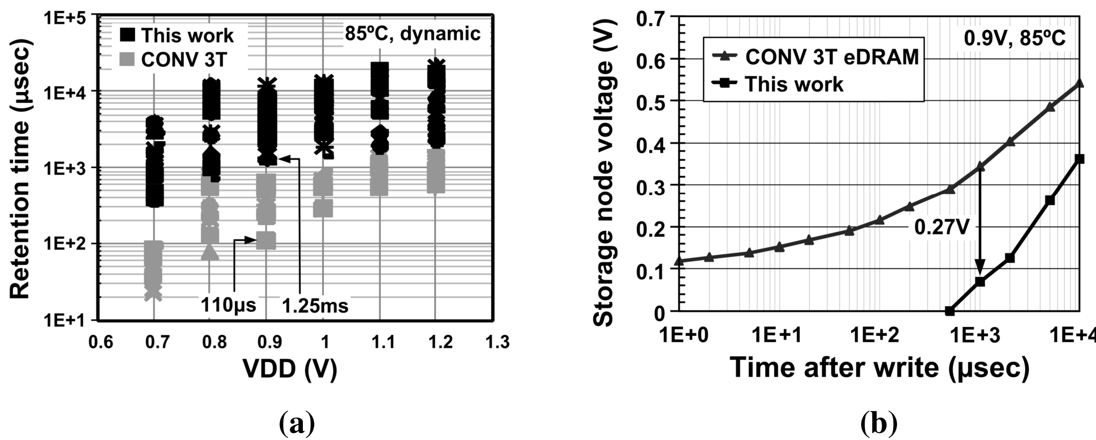


Fig. 2.20: (a) Measured retention time statistics. Due to limitations in the test setup, only 32 cells were measured from each sub-array. The measured cells were located evenly across the memory array. (b) Measured storage node voltage in the proposed boosted 3T cell and the conventional 3T cell. The cell voltage was indirectly and noninvasively measured by sweeping the reference cell node voltage.

Similar to Fig. 2.19(b), Fig. 2.20(b) shows the measured storage node voltage of the proposed boosted 3T and the conventional 3T gain cell. Due to threshold voltage variations between the read devices and the WWL coupling effect after the write-back, the data '0' voltage of the conventional 3T started at around 0.1 V. Read failures start to occur when the cell voltage is higher than around 0.2 V for the conventional 3T. The

amount of cell node boosting of the proposed cell was 0.27 V after a 1.0 ms of hold time. The preferential boosting effect can be clearly observed in the measured data as the difference between the two curves diminishes at longer hold times. Note that the VDUM level could not be lowered below 0V in the test chip, so although a large negative cell voltage is expected at short retention times, we were only able to measure the positive cell voltages as shown in Fig. 2.20(b). This is sufficient as we are more interested in measuring the positive storage node voltage region which is when the memory operation starts to fail.

Figs. 2.21(a) and (b) show the measured storage node voltage of data '1' and data '0' enabling a 2.0 ns random cycle time at 0.9 V, for high (85 ºC) and room (25 ºC) temperature corners, respectively. Optimal VDUM levels to achieve longer retention time with fixed read speed were 0.2 V for high temperature and 0.14 V for room temperature. Fig. 2.21(c) shows the measured VDUM level at high and room temperature corners for various supply voltages. VDUM level change across a temperature range of 25 ºC to 85 ºC and a supply voltage range of 0.8V to 1.3V was 50 mV. The 50 mV voltage difference is approximately the threshold voltage difference between the two temperature conditions.

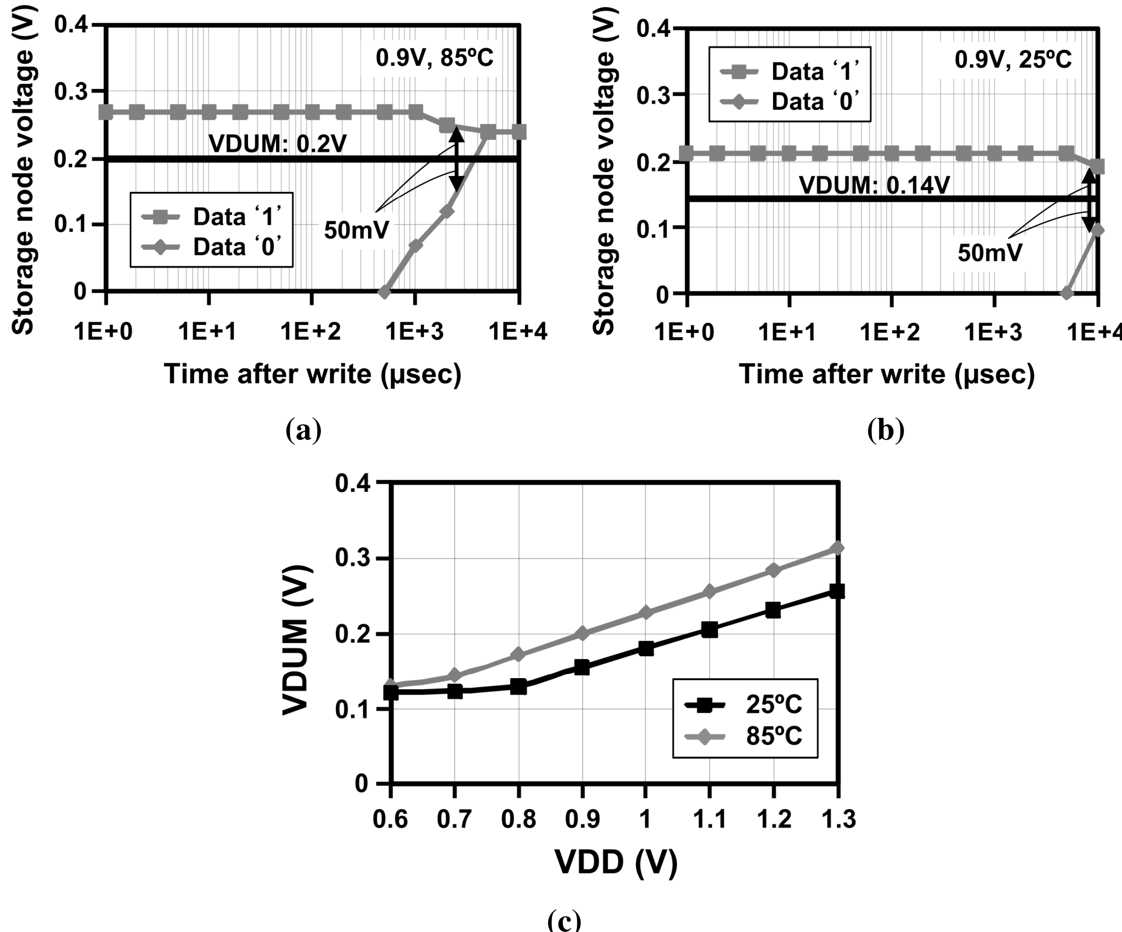


Fig. 2.21: Measured storage node voltages at (a) 85°C and (b) 25°C. (c) Measured PVT-tracking read reference (VDUM) level at different supply voltages.

2.5 Conclusions

Circuit techniques have been presented for increasing the data retention time and enhancing the performance of gain cell eDRAMs. The proposed boosted 3T eDRAM cell preferentially boosts the cell voltage to obtain high performance and low static power dissipation, with a layout penalty of only 14% compared to a conventional 3T cell. The proposed regulated bit-line write scheme can eliminate the data '1' write disturbance problem without introducing another boosted supply for WWL. The measurement results

show the 1.25 ms data retention time with 2 ns random cycle time at 0.9 V, 85 ºC, which is a 10X improvement compared to a conventional 3T gain cell measured from the same silicon die. The measured static power dissipation from a 64 kb test chip with the proposed schemes was 91.3 μW per Mb at 1.0 V, 85 ºC, and 1.0 ms refresh period, which is about 50% smaller compared with a power gated SRAM with half the number of cells.

Chapter 3

An Asymmetric 2T EDRAM for High Speed On-die Caches

3.1 Retention Characteristics of Conventional Gain Cells

In the second test chip, read paths implemented only with NMOS's are explored to achieve higher performance than PMOS based gain cells described in Chapter 2. To aid the understanding of our proposed techniques, I first describe the detailed retention characteristics of conventional gain cells having a NMOS read path.

In the 3T NMOS cell shown in Fig. 3.1(a), PW denotes the write access device, PS the storage device, and PR the read access device. Unlike 6T SRAMs or 1T1C eDRAMs, gain cells have a decoupled read and write structure – Read Word-Line (RWL) and Read Bit-Line (RBL) are used for read access and Write Word-Line (WWL) and Write Bit-Line (WBL) are used for write access. This attribute leads to improved read and write margins and flexibility in the bit-cell design - for example, the read and write paths can be optimized separately allowing gain cells to scale favorably in future technology nodes. In data retention mode, PW and PR are turned off and the storage node is left floating.

The sub-threshold, gate, and junction leakages in the surrounding devices cause the floating voltage to change with time as shown in Fig. 3.1(b). Since the storage node is surrounded by many low supplies in an NMOS only cell, the retention time of data '1' is much shorter than that of data '0'. To make matters worse, the data '1' (not data '0') voltage level is critical for the read access speed as the read port also uses an NMOS. The data retention time depends on the aggregated leakage current flowing into the storage node. Fig. 3.1(b) shows the cell retention time variations obtained by running 2^{20} Monte-Carlo simulations in HSPICE, which represents the cell-to-cell variation of a 1 Mb memory macro. In this analysis, we define retention time as the time it takes for the cell node voltage to reach a level corresponding to a target RBL delay of 500 ps.

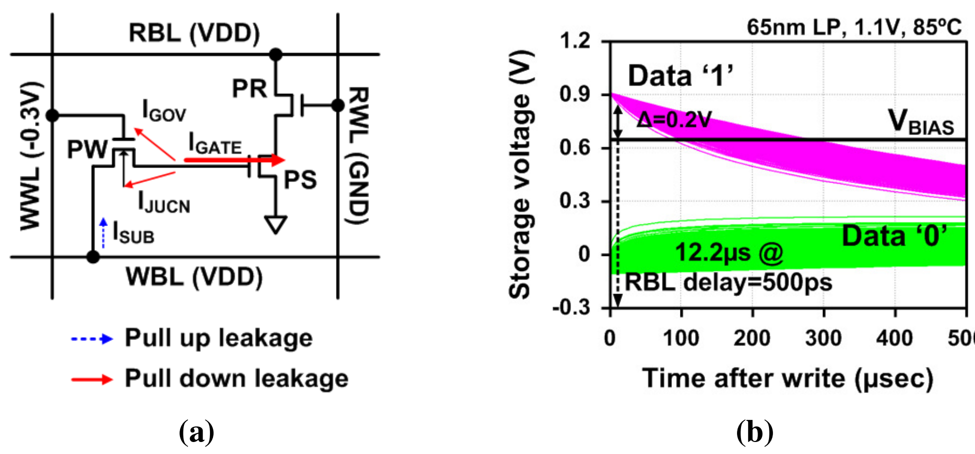


Fig. 3.1: (a) Leakage components of a 3T NMOS gain cell during data hold mode. (b) Monte-Carlo simulation results of storage node voltage during data hold mode showing 1 Mb macro retention characteristics.

The read reference bias level is set as 0.65 V and the data '1' voltage should be higher than this reference voltage by at least 0.2 V to achieve the same read margins as the data '0' case. Results based on our criterion indicate that the retention time of data '1' varies

from 12.2 μs to 54.1 μs mainly due to the gate leakage through the inverted channel of the NMOS storage device, while the non-critical data '0' voltage shows a very stable retention characteristic. Note that the WWL coupling after write-back operation results in lower initial storage levels than VDD and GND in case of data '1' and data '0', respectively. This further degrades the retention time of data '1' when a gain cell is implemented only with NMOS devices. The central idea of this work is to maximize the retention time and performance by using a new bit cell that balances the retention characteristics of data '0' and '1'.

3.2 Asymmetric 2T EDRAM Design

3.2.1 Asymmetric 2T Gain Cell

PMOS only gain cells were used in recent designs for improving retention time as they have 1-2 orders of magnitude lower gate leakage compared to their NMOS counterpart [19], [33]. However, the pull-up leakage currents of the PMOS devices surrounding the storage node have a negative impact especially on the data '0' level which determines the current through the PMOS read device. In addition, the poor channel mobility of PMOS devices limits the read performance. The new 2T gain cell structure proposed in this work achieves a long retention time without sacrificing read speed by using an NMOS read device driven by RWL for high drive current and a PMOS write device to keep the speed critical data '1' voltage close to VDD [35], [36]. Fig. 3.2 shows the proposed 2T cell and a previous Asymmetric 3T Cell (ATC) which was chosen for comparison because it also contains both NMOS and PMOS devices, albeit the

structure and operating principle are considerably different [16]. In the previous ATC cell, a PMOS device was used for the write access transistor to extend the cell retention time by compensating the NMOS gate leakage with the PMOS gate overlap and junction leakages. However, the leakage compensation effect of this cell is poor under PVT variations because the gate leakage through the inverted channel of the NMOS storage device is dominant for data '1' as shown in Fig. 3.2(a). In the proposed cell shown in Fig. 3.2(b), the read access transistor is replaced by the RWL signal whose pre-charge level is VDD. The storage transistor is nominally off making its gate leakage negligible. Since there is no sub-threshold leakage through the read path, a low Vth transistor can be utilized to further improve read speed. The proposed current sensing scheme described in the next section limits the RBL voltage swing to about 100 mV which eliminates problems associated with the pull-up leakage from the data '1' cells on the same RBL.

Fig. 3.2 (right) shows the simulated retention characteristics of a 1 Mb macro. The WWL coupling after write-back operation boosts data '1' level by 110 mV in a PMOS write device. However, the previous 3T ATC still suffers from a poor data '1' retention time due to the large gate leakage of storage device. The proposed asymmetric 2T gain cell improves worst case retention time by 3.4X while at the same time achieving a 45% shorter RBL delay compared to the previous 3T ATC. An additional benefit of the proposed 2T asymmetric cell is the balanced P and N diffusion densities which makes it more ideal to address Design-For-Manufacturability (DFM) concerns in extremely scaled technologies.

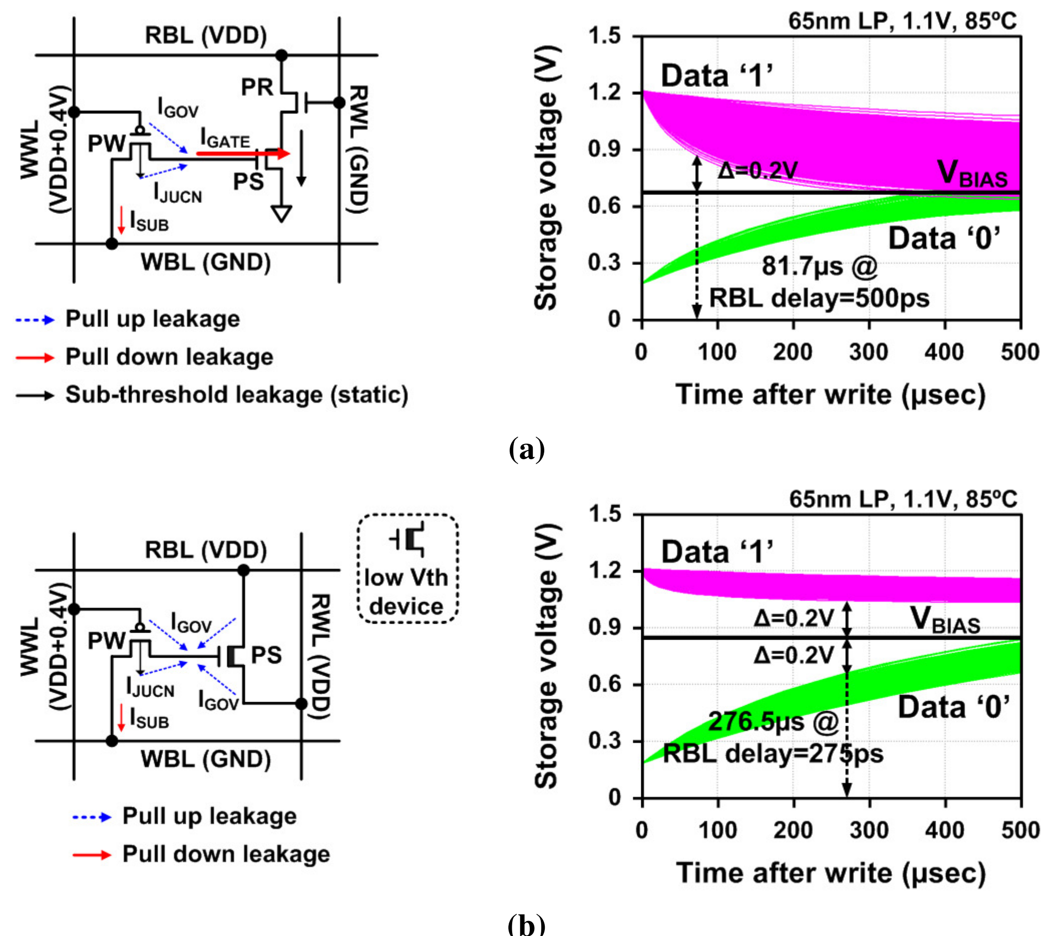


Fig. 3.2: Circuit diagrams and retention characteristics of (a) a previous asymmetric 3T gain cell [16] and (b) the proposed asymmetric 2T gain cell.

3.2.2 Pseudo-PMOS Diode Based Current-Mode Sense Amplifier (C-S/A)

Unlike in 3T cell designs, the RBL of 2T cells must have a limited swing to prevent the leakage current of the unselected cells from causing a read failure as illustrated in Fig. 3.3. However, a small voltage swing means that the read sensing margin is poor. The proposed asymmetric 2T gain cell worsens this situation since it utilizes a low Vth read device to achieve faster read speed by keeping the speed critical data '1' voltage close to VDD. Simulation results in Fig. 3.4 show a read failure in the worst case when all

unselected cells on the same RBL hold a strong data '1' at a high temperature and fast process corner condition.

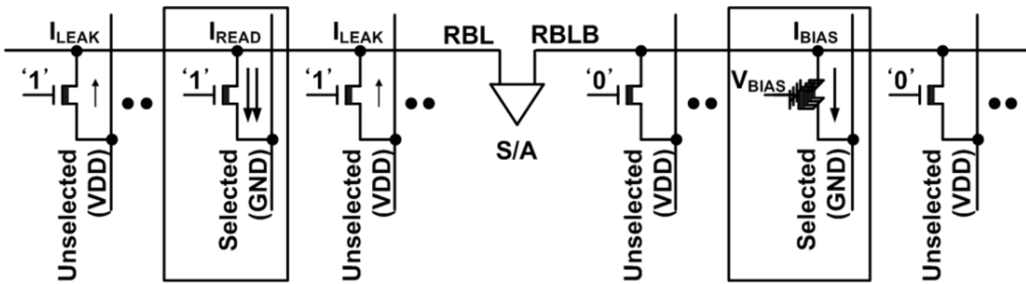


Fig. 3.3: Illustration of limiting read margin by adjacent cells holding high state in a 2T eDRAM.

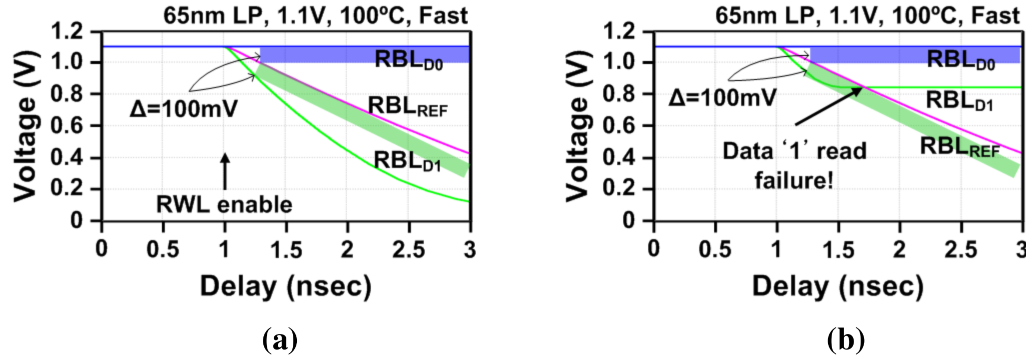


Fig. 3.4: (a) Simulated RBL sensing waveform when all adjacent cells hold a data '0'. (b) All adjacent cells hold a data '1' indicating a data '1' read failure. The shaded regions denote the ΔV_{RBL}=100mV window between the accessed RBL and the reference.

To overcome this problem, a Current-mode Sense Amplifier (C-S/A) is employed in our design to hold the RBL voltage close to VDD while sensing, allowing a large number of low Vth cells to be connected to a single RBL. The most common C-S/A shown in Fig. 3.5(a) consists of a PMOS load (P0), a cross-coupled PMOS latch (P1) and an NMOS diode (N1) pair [39]. The PMOS load pair provides currents to the cells and the C-S/A so that RBL can remain close to VDD during read operation. The cross-coupled

PMOS latch pair has a negative input impedance and amplifies the input currents. The NMOS diode pair has a positive input impedance and stabilizes the output voltages. The total input impedance of the C-S/A can be expressed as

$$R_{IN} = \frac{g_{m,N1} - g_{m,P1}}{g_{m,N1} g_{m,P1}} \quad (1)$$

indicating that a good matching between the PMOS latch and the NMOS diode pairs is required for a low input impedance. However, in the presence of P/N skew and PVT variations, matching the two impedances becomes difficult. Moreover, this conventional C-S/A suffers from a limited voltage headroom due to the stacked devices between VDD and GND.

An improved circuit shown in Fig. 3.5(b) consists of two folded PMOS diode pairs (P2 and P3), an NMOS current source (N2), and a cross-coupled PMOS latch pair (P1). N2 is biased using a separate voltage so the voltage headroom is increased by approximately 1xVth. Note that the conventional NMOS diode pair (N1) turns on only at a high supply voltage condition to improve the stability of this C-S/A [40]. Despite these advantages, the large number of devices in this circuit makes it impractical for DRAM circuits where every BL should have a dedicated S/A for a row-by-row refresh operation. This results in a large BL-S/A layout overhead in addition to impedance mismatch issues under PVT variations. The input resistance of this hybrid C-S/A is given as

$$R_{IN} = \frac{g_{m,N1} + g_{m,P2} + g_{m,P3} - g_{m,P1}}{(g_{m,P1} + g_{m,P2})(g_{m,N1} + g_{m,P3})}. \quad (2)$$

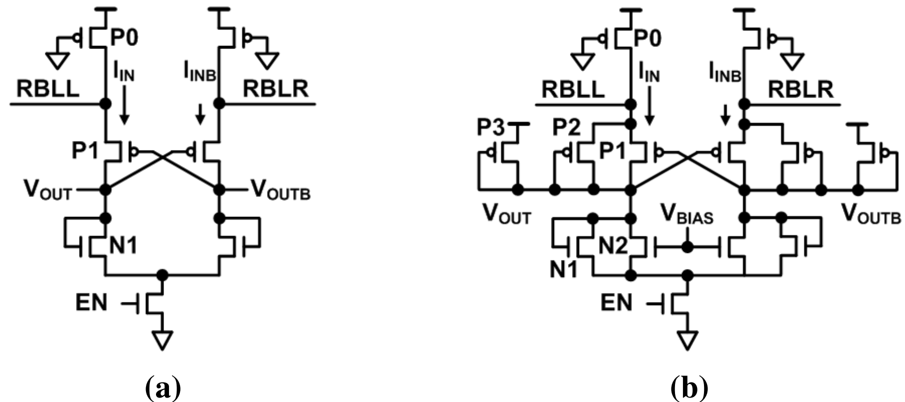


Fig. 3.5: (a) NP series-stacked C-S/A [39]. (b) Hybrid C-S/A [40].

The proposed C-S/A shown in Fig. 3.6 consists of a cross-coupled PMOS latch (P1) and a pseudo-PMOS diode (P2) driven by the negative supply VBB which is readily available on the chip for WWL under-driving. Recall that a negative WWL is needed for a PMOS device to write a data '0' into the cell without a threshold voltage loss. Similar to the conventional C-S/A, the input resistance of the proposed C-S/A can be expressed as:

$$R_{IN} = \frac{g_{m,P2} - g_{m,P1}}{g_{m,P1} g_{m,P2}}. \quad (3)$$

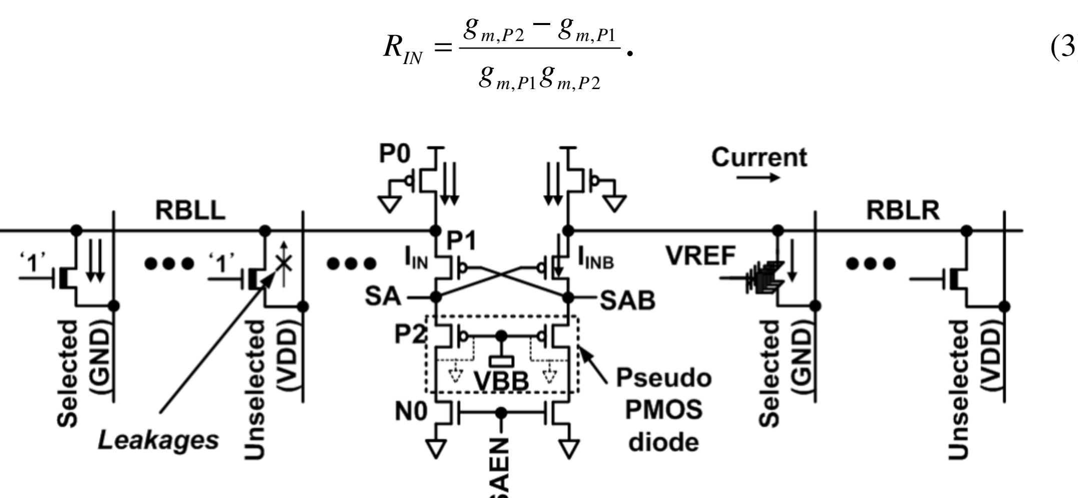
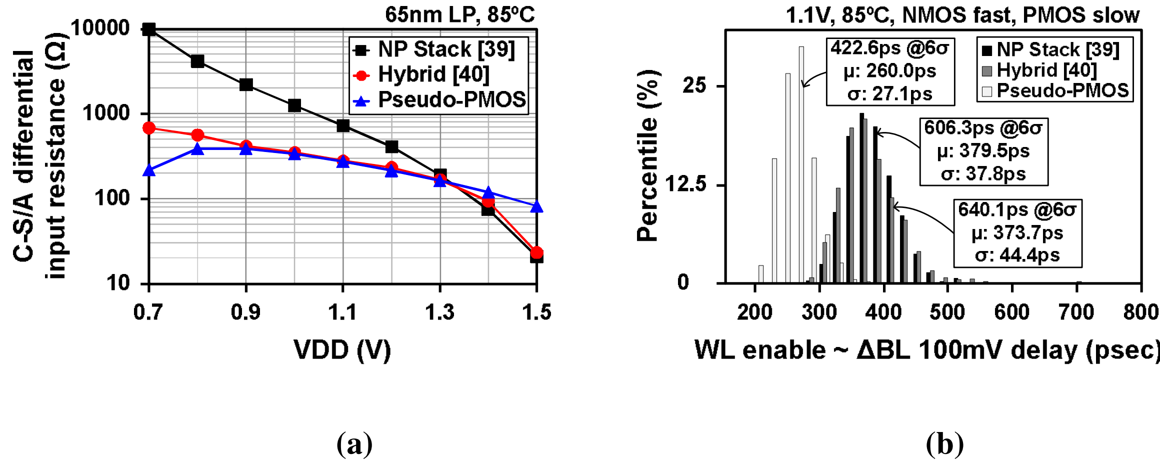


Fig. 3.6: Proposed pseudo-PMOS diode based C-S/A to overcome the issue of limited RBL voltage swing in a 2T eDRAM with improved voltage headroom and better impedance matching.

Fig. 3.7(a) shows simulated differential input resistances of the three C-S/A's at different VDDs. For this comparison, the C-S/A pairs were designed to have a minimum input resistance at the high VDD corner to ensure good stability [40]. The previous NP stack structure suffers from large input resistance at low operating voltage conditions leading to a considerable signal loss for the current sensing scheme. When this C-S/A operates in the sub-threshold region, the transconductances of the two pairs decrease. The denominator of (1) is the product of the two transconductances, while the numerator is the sum. This results in a rapid increase in input resistance at lower supply voltages as shown in Fig. 3.7(a). Input resistance of the previous hybrid C-S/A and the proposed pseudo-PMOS C-S/A show a stable response down to 0.9 V and 0.7 V, respectively. The maximum input resistance allowed in this design is 500 Ω which corresponds to a 10% signal loss during current sensing. Unlike the previous hybrid C-S/A, the improvement of low voltage margin in the proposed design depends on the voltage difference between the VBB (-0.5 V) and the threshold voltage (-0.315 V). Fig. 3.7(b) shows simulation results of RBL sensing delay for the NP stack, hybrid, and proposed C-S/A's. Each distribution represents the delay variation of the proposed gain cells from a 1 Mb macro with a refresh period of 100 μs. These Monte-Carlo results include cell leakage variations as well as device variations in the read path and C-S/A pairs. Although the hybrid C-S/A has a smaller input resistance than the NP stack C-S/A at 1.1 V, ensuring good matching between the large number of device pairs is difficult and results in a poor overall performance. The proposed C-S/A utilizing a pseudo-PMOS diode enhances the

RBL sensing delay by 30.3% (6-sigma point) due to the improved impedance matching and better low VDD margin.



(a) **(b)**

Fig. 3.7: (a) Simulated input resistance ($\Delta V_{RBL}/\ \Delta I_{IN}$) vs. VDD. (b) Comparison of RBL sensing delay under PVT variations and mismatches in the C-S/A pairs.

3.2.3 Half Swing Write Bit-Line Scheme

With the improved read bit-line sensing speed and increased number of cells per BL, WBL switching speed becomes the performance bottleneck. Similar to the half-VDD pre-charge technique employed in standard 1T1C DRAMs, a half swing WBL scheme can be applied to gain cell eDRAMs. By using a half swing WBL scheme with a tri-state buffer, the write speed is improved by 33% and the average WBL charging current is reduced by 25% without affecting the retention characteristics of the proposed 2T cell.

Fig. 3.8(a) shows simulated waveforms for a conventional GND pre-discharge scheme (full swing) and the half-VDD pre-charge scheme (half swing) indicating a 33% improvement in WBL charging speed. Retention characteristics of the GND pre-discharge scheme and the half-VDD pre-charge scheme are similar as shown in Fig.

3.8(b) since the sub-threshold leakage through the write device during data hold mode is negligible owing to the WWL over drive (VDD+α, where α=0.3 V in this design). Moreover, sub-threshold leakage through the write device can be effectively cut off during the data '1' write-back operation of a cell sharing the same WBL. The half swing WBL scheme is implemented as a part of the write-back circuit as shown in Fig. 3.10.

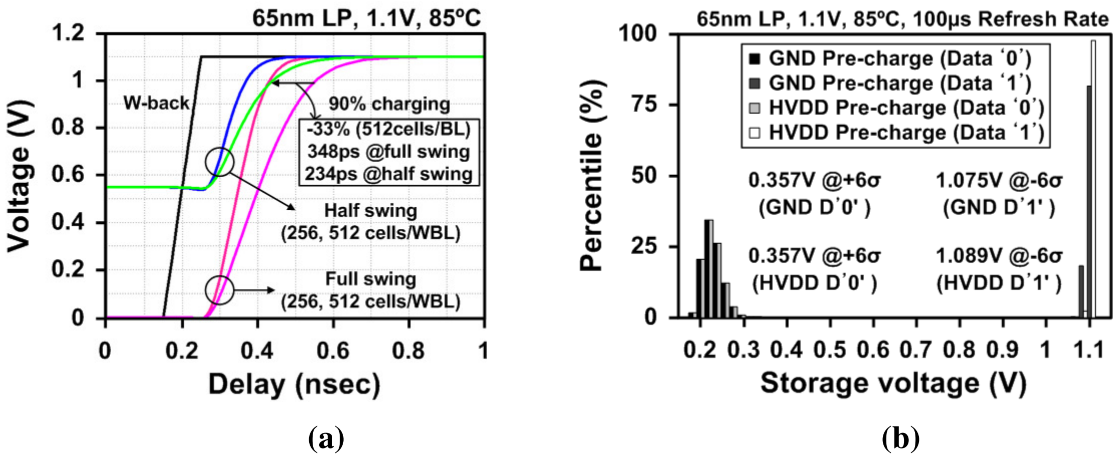


Fig. 3.8: (a) Simulated waveforms of the WBL charging delay. (b) Simulated storage voltage distributions of a conventional GND pre-discharge (full swing WBL) and the proposed half-VDD pre-charge (half swing WBL) schemes.

3.2.4 Stepped Write Word-Line Driver

DRAMs require a positive boosted voltage (VPP) to suppress the sub-threshold leakage in the write access device as well as a negative boosted voltage (VBB) to write data into the cell without a Vth drop (PMOS write device case). In order to reduce the power and area overhead of charge pumps during fast chip operation, we adopted a stepped WWL control scheme which minimizes the current drawn from the boosted VPP and VBB voltages by utilizing the main VDD and GND supplies for most of the WWL transition. The proposed WWL scheme consists of a nominal VDD/GND driver

including tri-state control circuits, a boosted VPP/VBB driver with an inverted signal, and a reset device as shown in Fig. 3.9(a). Before the cell access, PUB and PDN nodes in Fig. 3.9(a) are set to VPP and VBB, respectively. This deactivates the VDD/GND driver by cutting off the short circuit current path from VPP to VDD and from GND to VBB. The RSET signal is switched to VPP ensuring that all WWL's are pre-charged to the desired VPP level. Except during the initialization phase, the RSET signal stays at VBB. At the beginning of the write-back operation, decoded address signals and a short pulsed signal of PDNGND enable the GND pull-down path in Fig. 3.9(a). This drives the selected WWL towards GND. As the selected WWL is discharged, WWLB switches and enables the VBB pull-down path which drives the WWL to VBB. The pulse duration has to be carefully controlled to guarantee proper circuit operation while saving the WWL switching power. If the pulse duration is too short, the VBB pull-down path will not be enabled whereas if it is too long, there will be short circuit current between VBB and GND. In this design, we chose a pulse duration of 375 ps which gave sufficient timing margin at a slight increase in the current drawn from the boosted supply. The operating principle of the opposite high-to-low WWL transition is similar to what we described above and the waveforms are shown in Fig. 3.9(b).

Fig. 3.9(c) shows the simulated waveforms of the current consumption and WWL transition for the conventional and proposed schemes. With a stepped WWL control scheme, 67% of the boosted supply current and 4.3% of the total chip area can be saved with two additional peripheral control signals and four more transistors in the WWL control circuit compared to conventional two-stage level shifters. Note that during a step

transition of WWL, the effective pulse width is decreased. Nevertheless, a WWL pulse width of 406 ps can be achieved at a 1.5 ns cycle time which is significantly longer than the required pulse width of 210 ps. Further details on the macro level timing will be given in section 3.3.

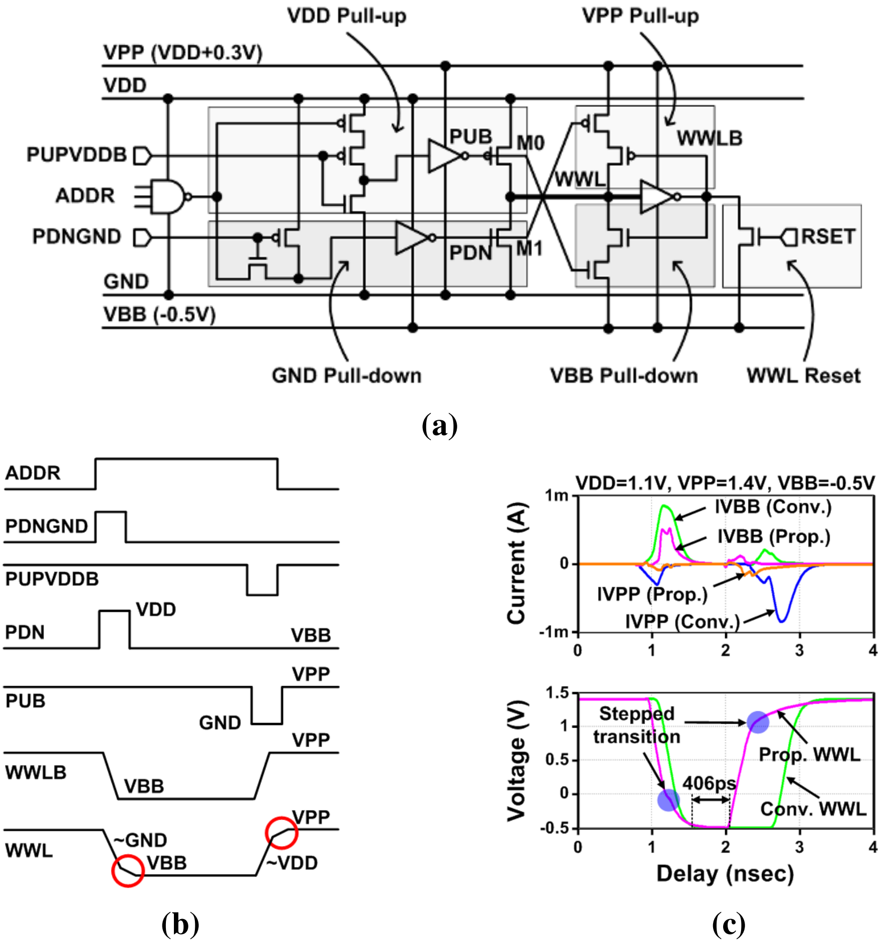


Fig. 3.9: Proposed stepped WWL driver. (a) Schematic. (b) Timing diagram. (c) Simulated boosted current consumptions and WWL waveforms during transition.

3.2.5 Sense Amplifier and Write-Back Circuit Design

Fig. 3.10 shows the complete schematic and timing diagram of the proposed S/A, read port, write-back, and write port. A two-stage full pipeline structure was

implemented to control the read and write-back operations. In the first clock cycle, the RWL is selected. When the C-S/A control signal (ISAEN) is enabled, the C-S/A amplifies the input signals to analog voltage signals while the RBL held close to VDD. Once a recognizable voltage difference is developed, the voltage S/A control signal (VSAEN) is fired. In the second clock cycle, read-out and write-back operations follow. After the write-back, WBLs are pre-charged back to half-VDD using the boosted supply VPP control signal (PRECHL/R). A stepped PRECH control scheme can be also adopted to further minimize the current drawn from the boosted supply VPP.

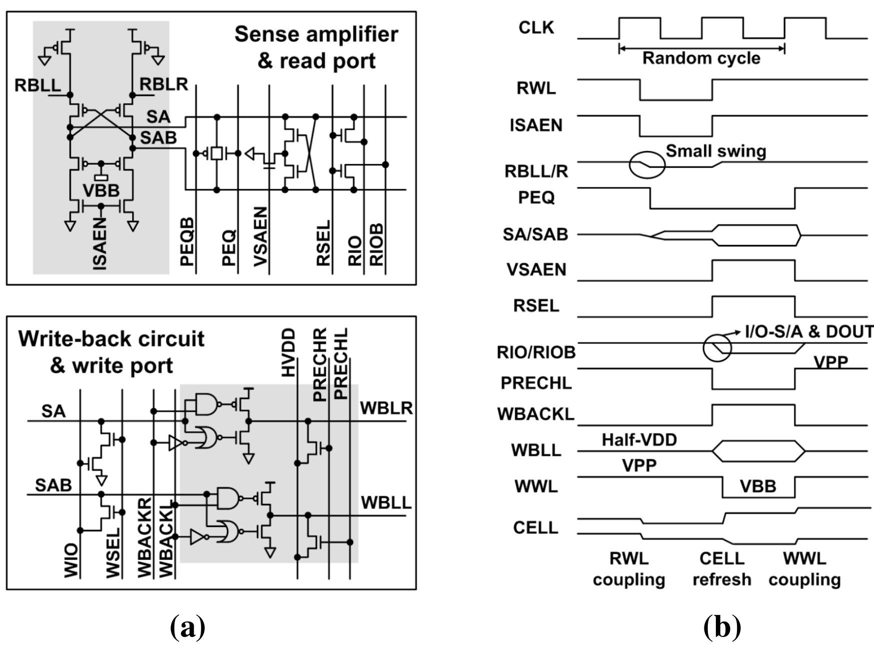


Fig. 3.10: (a) Circuit diagram of the proposed Sense Amplifier (S/A) with read port, write port, and write-back circuits. (b) Two-stage read and write-back timing diagram.

Fig. 3.11 shows post-layout simulation waveforms of the proposed 2T eDRAM. This includes the proposed asymmetric 2T gain cell, the pseudo-PMOS diode based C-S/A, a half-swing WBL scheme, and a stepped WWL driver. The memory array with 192 cells-

per-WL and 512-cells-per-BL can operate at a random cycle time of 1.5 ns for a test sequence of data '0' read and write-back followed by data '1' read and write-back.

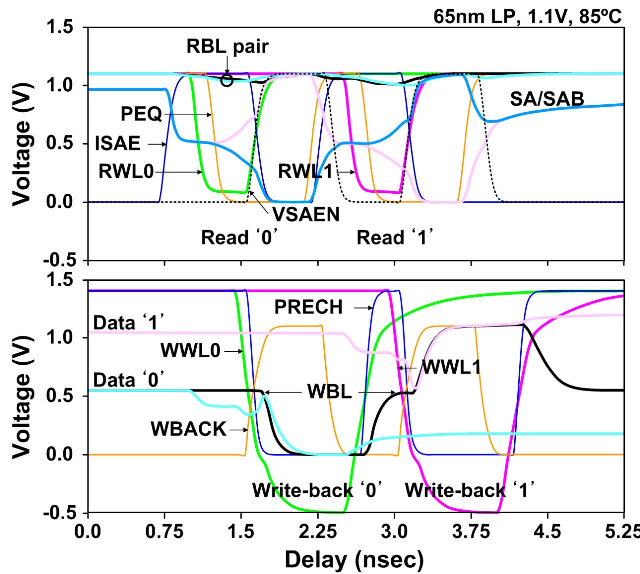


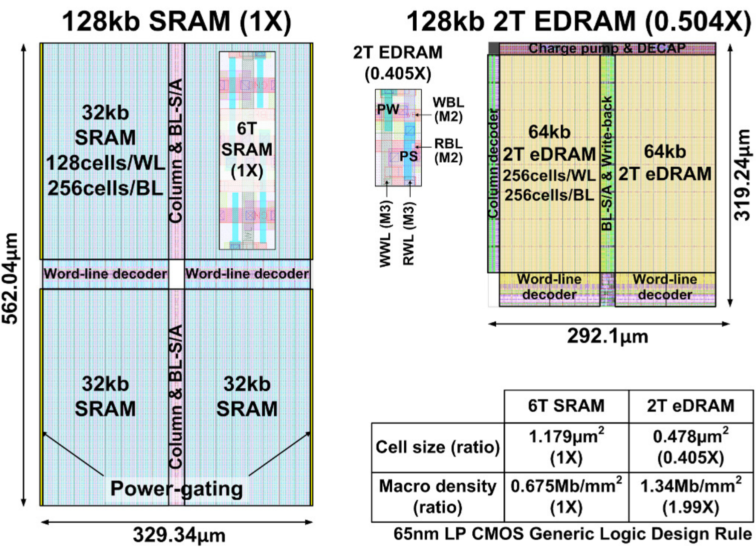
Fig. 3.11: Simulated waveforms of back-to-back read and write-back operations for a 1.5 ns cycle time.

3.3 Comparison Between SRAM and Gain Cell EDRAM

In order to demonstrate the advantages of the proposed 2T eDRAM over conventional 3T eDRAM or 6T SRAM, this section presents macro level layout and performance comparisons. Static power comparisons are detailed in Section 3.4. Extensive Monte-Carlo simulations were performed on megabit density SRAM and eDRAM arrays to estimate their performance in a practical scenario [35], [41]. Our analysis includes process variation in the memory cells and the C-S/A as well as realistic fluctuations for the reference biases and boosted supplies.

3.3.1 Macro Layout Comparison

Fig. 3.12 shows the bit-cell and 128 kb sub-array layouts of a 6T SRAM and the proposed 2T eDRAM in a generic 65 nm LP CMOS process. Dense bit-cell design rules were not available to the authors but for area comparison purposes, using a logic design rule is a generally accepted practice [35]. The 6T SRAM used for the comparison has the following transistor dimensions: W_{PU}=W_{min}, W_{PD}=2xW_{min}, and W_{ACCESS}=W_{min}, with all devices using a minimum channel length. This is the most general sizing scheme and extensive Monte Carlo simulations were performed to verify good read and write margins. The bit cell area of the proposed 2T gain cell is 59.5% smaller (or 2.47X denser) than that of a 6T SRAM resulting in a 49.6% smaller area for a 128 kb sub-array. It is worth mentioning that layout of the 128 kb 2T eDRAM sub-array includes a BL-S/A and write-back driver in each BL, full RWL and WWL decoders, and charge pumps for generating boosted high and low supplies. The unit 128 kb sub-array can be tiled to build a larger memory macro.



	6T SRAM	2T eDRAM
Cell size (ratio)	1.179μm^2 (1X)	0.478μm^2 (0.405X)
Macro density (ratio)	0.675Mb/mm^2 (1X)	1.34Mb/mm^2 (1.99X)

Fig. 3.12: Comparison of bit-cell and 128 kb sub-array layout between 6T SRAM and 2T eDRAM.

3.3.2 Macro Performance Comparison

Fig. 3.13 shows read bit-line delay distributions for the following four memory arrays; a 1 Mb SRAM with 256 cells-per-BL, a 1 Mb conventional 3T eDRAM with 256 cells-per-BL, and a 1 Mb proposed 2T eDRAM with 256 and 512 cells-per-BL. The single-ended sensing nature and the gradual loss in the storage node voltage of the conventional 3T eDRAM result in a 6-sigma read bit-line delay that is 1.9 times longer than a 6T SRAM as shown in Fig. 3.13. The proposed 2T eDRAM makes up for this performance shortfall, achieving a bit-line sensing speed comparable to that of a 6T SRAM with 256 cells-per-BL. For an array with 512 cells-per-BL, the proposed 2T eDRAM shows only a 4% longer RBL sensing delay than a 6T SRAM that has half the number of cells-per BL. The performance improvement is attributed to the following three factors: excellent data '1' retention, low Vth device in the decoupled read path, and the proposed C-S/A which makes the read speed more or less independent of the RBL capacitance.

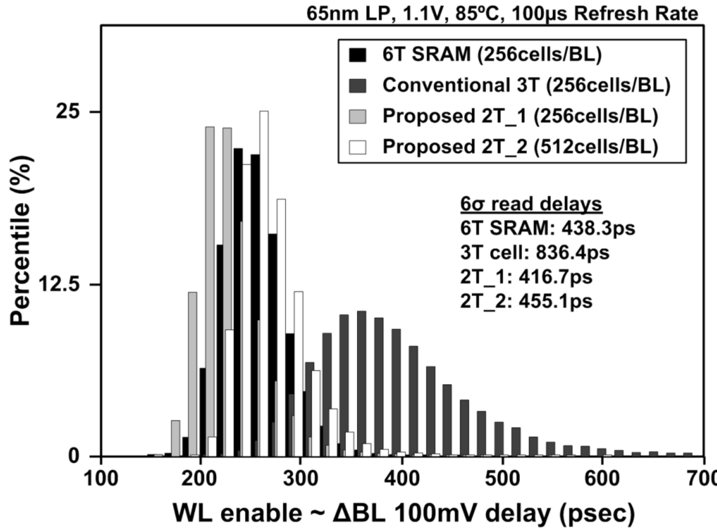


Fig. 3.13: RBL sensing delay distributions of SRAM and gain cell eDRAMs each with a 1 Mb macro density.

For cache sizes of 1 Mb or larger, the proposed 2T eDRAM achieves a faster access time owing to the shorter global interconnect delay made possible by the smaller bit-cell size as shown Fig. 3.14(a). Therefore, a 512 cells-per-BL architecture was chosen for this 2T eDRAM design in order to verify our proposed schemes under extreme cases and to reduce the array layout overhead stemming from the complicated BL-S/A and write-back circuits. Embedded DRAMs require a write-back operation after the read operation to restore the cell data. This results in a 66.5% slower random cycle time for a conventional 3T eDRAM compared to a 6T SRAM as shown in Fig. 3.14(b). The proposed 2T eDRAM improves the random cycle time by 31.6% compared to a conventional 3T eDRAM that has half the number of cells per BL.

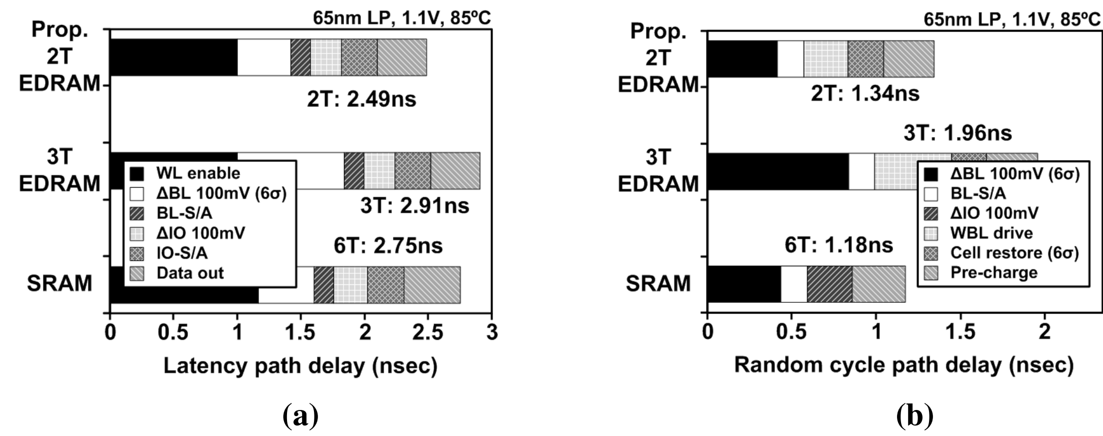


Fig. 3.14: Performance comparison of 1 Mb macros using SRAM and gain cell eDRAMs. (a) Latency. (b) Random cycle.

Fig. 3.15 shows the 1Mb write delay distributions of a 6T SRAM array and the proposed 2T eDRAM array. Here, the write delay is defined as the WL activation to the time when the cell node reaches 95% of the full voltage swing. The write speed of the gain cell is faster than the 6T SRAM since the latter is based on a ratioed operation. For the speed critical data ‘1’ case, the proposed 2T eDRAM achieves an 11.5X faster write-

back (6-sigma point performance) compared to the 6T SRAM as shown in Fig. 3.15. Note that the WWL of the gain cell must be sufficiently negative in order for the PMOS write devices to pass a good data '0' level. For a WWL under-drive voltage of -0.5 V, the 1 Mb Monte-Carlo simulations show a write speedup of 35% (6-sigma point) for data '0'.

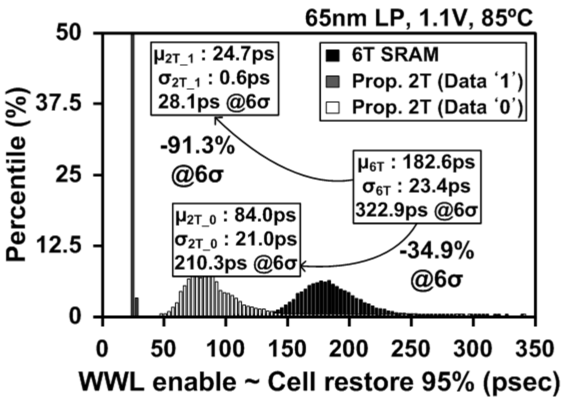


Fig. 3.15: Performance comparison of 1 Mb macros using SRAM and gain cell eDRAMs. (a) Latency. (b) Random cycle.

3.4 Test Chip Implementation and Measurements

A 192 kb eDRAM test chip was implemented in a 1.2 V, 65 nm Low-Power (LP) logic CMOS process to demonstrate the proposed circuit techniques. The detailed array architecture is shown in Fig. 3.16 consisting of two 96 kb blocks sharing BL-S/A and write-back circuits located at the center of the array. The dummy memory cells in each block are 4X larger than the regular cells to minimize random device mismatch. RWL pull-down drivers are inserted every 64 WL's in order to minimize the RWL ground noise during read access.

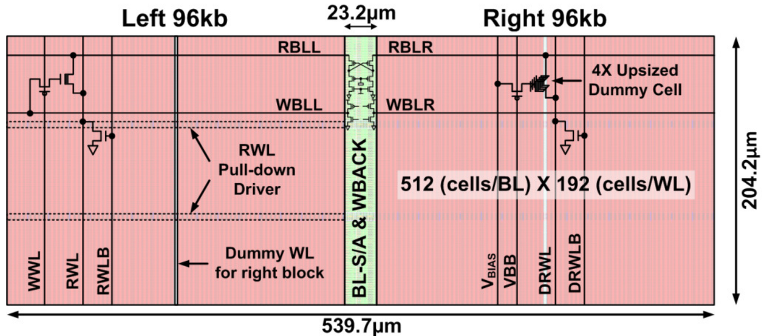
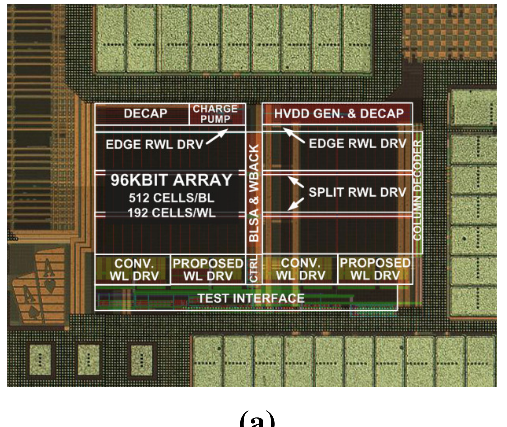


Fig. 3.16: A 192 kb test array architecture with 192 cells-per-WL and 512 cells-per-BL.

Fig. 3.17 shows the chip microphotograph and a feature summary table of the 192 kb eDRAM test chip. For a 99.9% bit yield at 1.1 V and 85 ºC, our design achieves a random cycle frequency of 667 MHz and 500 MHz using a refresh period of 110 µs and 1200 µs, respectively. By increasing the VPP level from 1.5 V to 1.6 V, a 100 µs retention time can be achieved under a 99.99% bit yield condition. To put this into perspective, the target retention time of a previous 2T gain cell eDRAM design was 10 µs [19] while the measured retention time of a commercial 1T1C eDRAM was 40 µs at 105 ºC with a 99.99% bit yield [6] each with a random cycle of 500 MHz.



Process	65nm LP CMOS
Circuit dimension	555.8x297.8µm^2
Array size	192kb (192 WLs, 2x512 BLs)
Cell size	41% of 6T SRAM
* Retention time @ 1.1V, 85ºC	110µs @ 667MHz 1200µs @ 500MHz
Latency	1.39ns @ 1.2V 1.65ns @ 1.1V
** Refresh power @ 1.2V, 85ºC	1.16mW/Mb @ 667MHz 108.84µW/Mb @ 500MHz

* 99.9% bit yield condition
** 110µsec refresh rate for 667MHz
1200µsec refresh rate for 500MHz

(a) (b)

Fig. 3.17: (a) Microphotograph of the 65nm eDRAM test chip. (b) Chip feature summary.

By externally adjusting the read reference voltage (VDUM), we can indirectly and noninvasively measure the storage node voltage at different data retention times [34]. For example, read failure will happen for data '1' if the VDUM level is higher than the storage node voltage so the storage voltage can be measured by sweeping the VDUM voltage and measuring the point of failure. It is worth mentioning that the storage node voltage measured using this method includes effects such as process variation or transient noise (e.g. coupling noise or supply noise) providing us with the "effective" cell node voltage. The measured storage node voltage of the proposed 2T eDRAM in Fig. 3.18 shows that retention times even longer than 1 ms can be achieved.

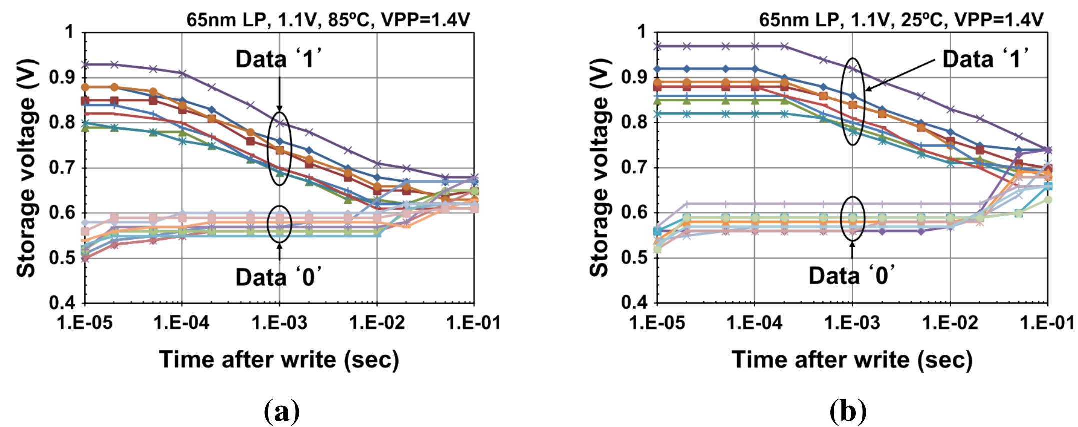


Fig. 3.18: Measured storage node voltage at different retention times at (a) 85°C and (b) 25°C.

Adjusting the VPP level modulates the gate overlap and gate-induced drain leakages and hence allows us to achieve an optimal retention time with the consideration of both data '1' and data '0' cases as shown in Fig. 3.19. This dependency can be further exploited for post-fabrication trimming to cope with die-to-die variations.

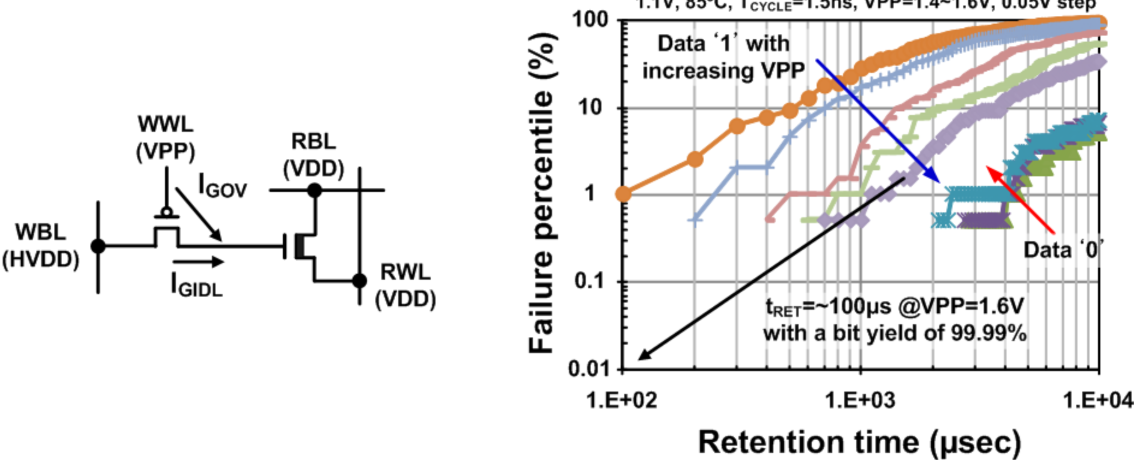


Fig. 3.19: Measured retention time distribution vs. boosted high supply (VPP) level.

The retention time of a 2T eDRAM can be extended at the expense of a longer random cycle time as shown in Fig. 3.20(a). We can utilize this trade-off to enhance access speed, and at the same time minimize refresh power dissipation of the 2T eDRAM. During memory access, the S/A enable signal was triggered as early as possible after the RWL activation to achieve a high random cycle frequency as high as 667 MHz. Moreover, a delayed S/A enable signal extends the retention time resulting in significant refresh power savings. The measured refresh power at a random cycle of 667 MHz and 500 MHz were 1.16 mW/Mb and 109 µW/Mb, respectively at 1.1V and 85°C. The flexibility in the cycle time offers further opportunities to reduce refresh power depending on the system level workload and frequency requirements. For a 1 Mb macro with 1024 WL's, only 1.40% of the total operating time is spent on refresh for a 1.5 ns random cycle and a 110 µs refresh period. The refresh overhead reduces to 0.17% for a 2.0 ns random cycle and a 1200 µs refresh period. The measured VDD shmoo of cycle

time and the corresponding retention time in Fig. 3.20(b) shows a wide operating voltage range from 1.4 V down to 0.8 V.

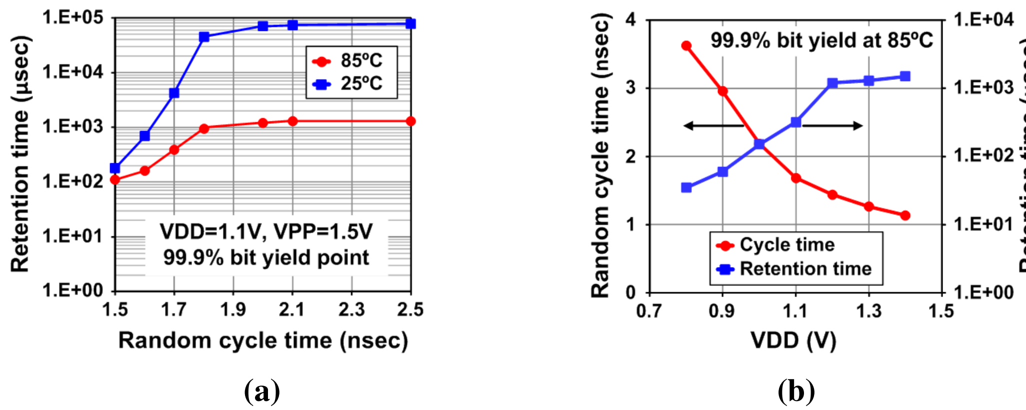


Fig. 3.20: (a) Measured random cycle time vs. retention time. (b) Measured VDD shmoo of random cycle time and corresponding retention time.

Fig. 3.21 shows the static current consumption of a 6T SRAM and the proposed 2T eDRAM for different random cycle times. We assume a power-gated SRAM with a data retention voltage of 0.6V. Supply voltage of the 2T eDRAM is assumed to be 1.1V during hold mode. For very short random cycles (e.g. 1.5 ns), the static current of the proposed 2T eDRAM is much larger than that of the 6T SRAM due the frequent refresh operation required to maintain a good cell node voltage. However, for longer random cycle times, the RBL sensing margin of the 2T eDRAM improves significantly which increases the retention time. For a 2.0 ns random cycle time, the proposed 2T eDRAM has an 81% and 91% smaller static current consumption than a power-gated SRAM [42] at 85 °C and 25 °C, respectively.

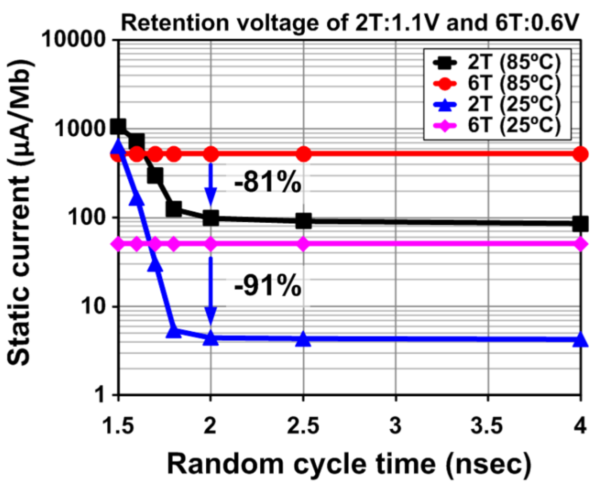


Fig. 3.21: Static power comparison between 6T SRAM and the proposed 2T eDRAM with varying random cycle time at 85 ºC and 25 ºC.

The retention time of the proposed 2T eDRAM cannot be improved further for cycle times longer than 2.0 ns cycle time as shown in Fig. 3.20(a). The maximum achievable retention time is set by the data window shown in Fig. 3.18 and the variability in the bit-cells and BL-S/A's. The random cycle and retention time of eDRAMs are highly dependent on the number of cells-per-BL. The proposed 2T eDRAM has 16 times more cells on the same RBL than previous 1T1C eDRAMs [6], [7] and 4 times more cells than a previous 2T PMOS eDRAM [19]. The measured random cycle time with 512 cells-per-BL was 1.5 ns (667 MHz) which is a 33.4% improvement compared to previous eDRAM designs while achieving a retention time similar to 1T1C eDRAMs. For a random cycle of 500 MHz, the measured retention time is >120X longer than a previous 2T PMOS eDRAM and around 12X longer than a 1T1C eDRAM.

3.5 Conclusions

Several circuit techniques have been presented for improving data retention time and enhancing performance of gain cell eDRAMs for high speed and high density on-die caches. The proposed asymmetric 2T gain cell keeps the critical data '1' level close to VDD to improve memory performance and reduce static power dissipation. The proposed pseudo-PMOS diode based C-S/A eliminates the RBL leakage, provides better impedance matching, and offers more voltage headroom than previous designs. The half swing WBL scheme with a tri-state buffer achieves a 33% faster write speed and a 25% smaller WBL charging current without affecting the retention characteristics. Finally, a stepped WWL control scheme reduces the current drawn from the boosted supply by 67% which results in a 4.3% reduction in memory array area due to the smaller charge pump circuit and decoupling capacitors. Measurement results show a 667 MHz random cycle using a 110 µs refresh period for a 99.9% bit yield at 1.1 V, 85 ºC. The static power dissipation including refresh currents and cell leakages was 109 µW/Mb at 500 MHz, 1.1 V, 85 ºC which is 81% smaller than a power gated SRAM under a data retention voltage of 0.6 V.

Chapter 4

A Logic-Compatible 2T1C EDRAM for Enhanced Reliability

4.1 Boosted Supply Level vs. EDRAM Performance

DRAMs typically require two boosted supplies: a boosted high voltage (VPP) to suppress the subthreshold leakage (assuming a PMOS write device) and a boosted low voltage (VBB) to prevent V_{TH} drop during write. Fig.4.1 illustrates how the boosted supply level affects the performance of a 2T gain cell eDRAM. Here we consider an asymmetric 2T cell described in Chapter 3 [35] with a PMOS write device and an NMOS read device, although a similar analysis can be made for other types of gain cells. Write Word-Line (WWL) is biased at VPP during data retention mode in order to suppress the subthreshold leakages flowing into unselected cells. The subthreshold leakage is worst when writing '1' to another cell on the same the Write Bit-Line (WBL). The VPP level modulates the gate overlap and gate-induced drain leakages and therefore the optimal retention time can be achieved by considering the retention times of both data '1' and data '0' as shown in Fig. 4.1(a). The VBB level on the other hand affects the data '0'

restore time during write. The simulated data ‘0’ restore time dependency on VBB level in Fig. 4.1(b) indicates that VBB should be -0.4 V or below to ensure a practical write time. The above analysis shows that a WWL voltage swing from -0.4V to 1.6V is required for optimal memory cell operation which is 67% higher than the nominal supply level of 1.2 V in this 65 nm LP CMOS process. Boosted high and low supplies can only be used with special devices with a thicker T_{OX} to avoid voltage overstress. Alternatively, I/O devices can be considered; however, this will increase the bit-cell area considerably and in turn degrade the macro performance. Layout comparison between several embedded memory bit cells are presented in Section 4.2.

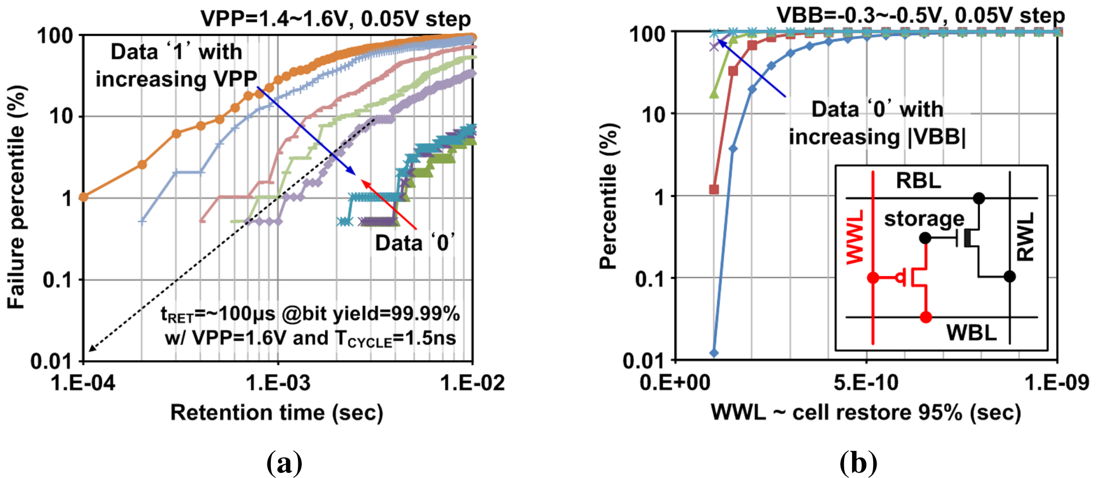
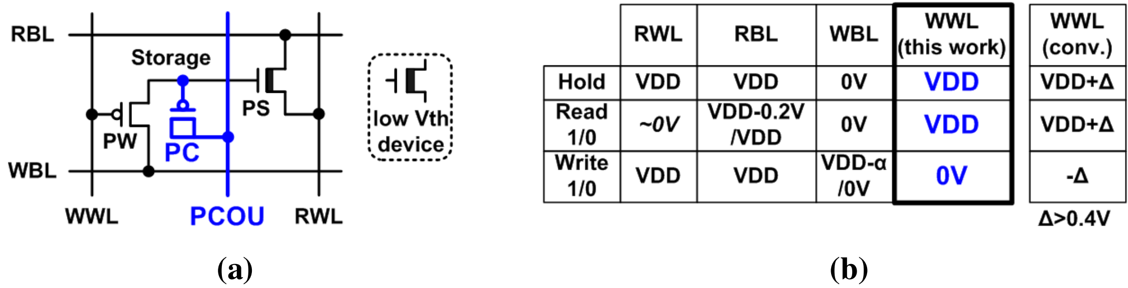


Fig. 4.1: Impact of boosted supply level on 2T eDRAM performance [35]. (a) Boosted high supply (VPP) level vs. retention time (measured). (b) Boosted low supply (VBB) level vs. data ‘0’ write time (simulated).

4.2 2T1C EDRAM with No Boosted Supplies

4.2.1 2T1C Gain Cell

In order to realize a truly logic-compatible eDRAM with a competitive bit-cell size and higher macro level performance, we propose a 2T1C gain cell that can be implemented with regular thin oxide devices. The new cell structure consists of an asymmetric 2T cell [35] and a separate coupling MOS capacitor controlled by a control signal. Fig. 4.2 shows the proposed bit-cell schematic along with the signal conditions for each operating modes. It's important to note that none of the voltage levels exceed the nominal VDD. The bit-cell may look similar to the previous 3T1D cell that has an additional gated-diode controlled by Read Word-Line (RWL) in order to enhance speed and retention time by signal amplification [18]. However, the structure and operating principle of the 2T1C cell are considerably different from prior work.



	RWL	RBL	WBL	WWL (this work)	WWL (conv.)
Hold	VDD	VDD	0V	VDD	VDD+Δ
Read 1/0	~0V	VDD-0.2V /VDD	0V	VDD	VDD+Δ
Write 1/0	VDD	VDD	VDD-α /0V	0V	-Δ

Fig. 4.2: Proposed 2T1C gain cell based on thin oxide devices with no boosted supplies. (a) Schematic. (b) Signal conditions for each operating modes.

The timing diagram shown in Fig. 4.3 illustrates the operation principle of the proposed cell. The capacitor control signal (PCOU) is pre-discharged to 0 V during hold mode introducing only a small amount of gate-overlap leakage through the coupling device (PC). At the beginning of the read access when the RWL is activated, PCOU is also switched to VDD. This couples up both data '1' and '0' storage voltages. The higher voltage levels increase the drive current for the NMOS read access device (PS) enhancing

the read performance. After the Sense Amplifier (S/A) samples the Read Bit-Line (RBL) data, a write-back operation follows which drives the WWL to 0 V instead of the usual negative boosted supply. Using a data '1' WBL voltage that is slightly lower than VDD (i.e. VDD-α in Fig. 4.3), the subthreshold leakage in the unselected cell can be effectively cut off without using a boosted high supply for WWL [33]. Data '1' can be easily written back to the cell with a PMOS write device (PW). However, without a boosted negative supply, data '0' will not be fully restored due to the V_{TH} drop in PW. To resolve this issue, PCOU is switched to 0V immediately after write back. This couples down the data '0' voltage while the data '1' voltage is not affected since PW remains on when WBL is high. Finally, WWL is switched back to its precharge level of VDD and this slightly couples up both data '1' and '0' voltages through the gate-overlap capacitance, fully restoring the cell storage levels.

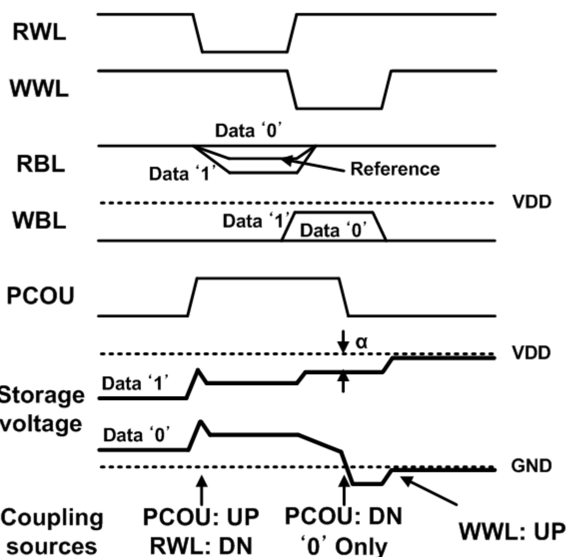


Fig. 4.3: Timing diagram of the proposed 2T1C cell for read and write-back operations.

Fig. 4.4 shows the simulated waveforms of read and write-back operations. The proposed 2T1C with no boosted supplies achieves a similar data '1' voltage level during read and a similar data '0' level after write-back operations compared to the asymmetric 2T with boosted supplies.

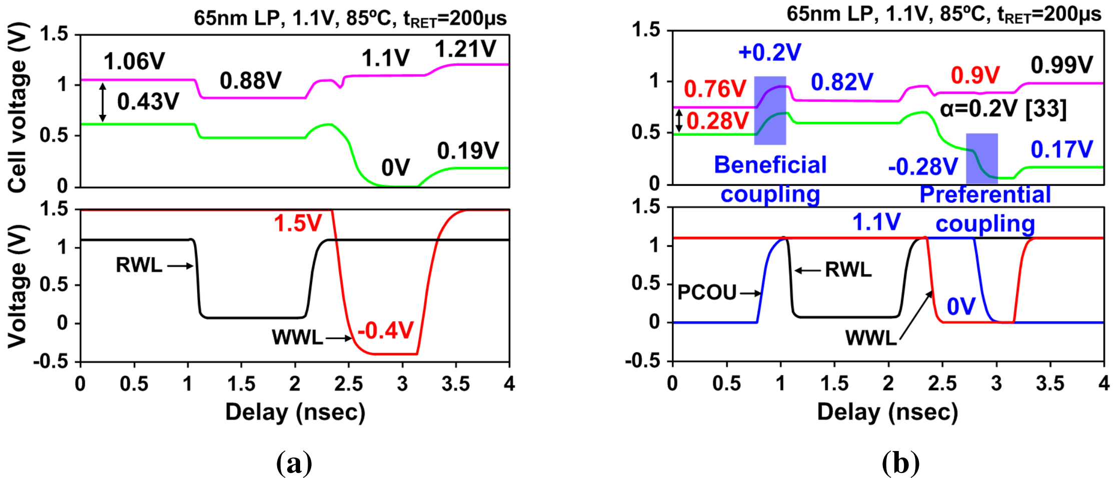


Fig. 4.4: Simulated waveforms of read and write-back operations for (a) a conventional 2T eDRAM and (b) the proposed 2T1C eDRAM.

The initial voltage levels and data windows between '1' and '0' in Fig. 4.4 are based on retention simulations for a 1 Mb macro (Fig. 4.5). The data window at 200 μs for the 2T1C eDRAM is 150 mV smaller than that of a 2T eDRAM with boosted high and low supplies. A narrower data window reduces the margin between data '1' and '0' resulting in worse retention time and increased static power due to the frequent refresh operation. To cope with this issue, we propose two circuit techniques: (i) a single-ended 7T SRAM for weak gain cell repair and (ii) a storage voltage monitor for adaptive refresh by tracking the retention characteristics under PVT variations. Fig. 4.6 shows the schematic diagram of a 64 kb 2T1C eDRAM macro including the 7T SRAM repair cells (details in

Section 4.2.2) and the storage voltage monitor (details in Section 4.2.3) that are seamlessly integrated into the array.

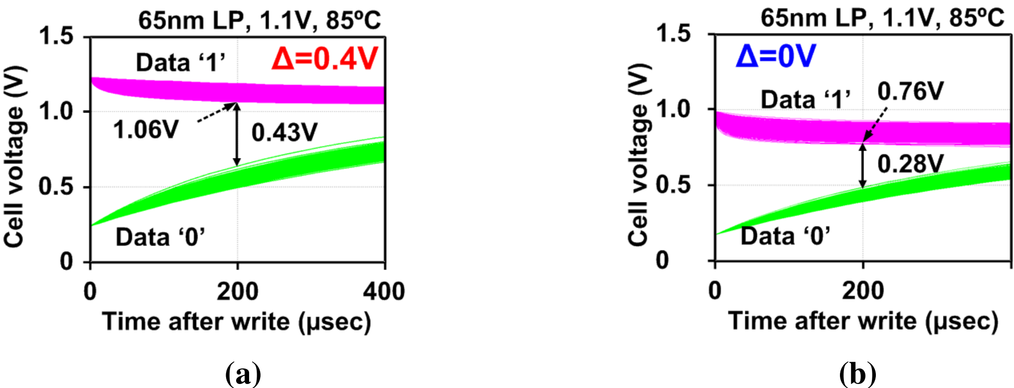


Fig. 4.5: Comparison of retention characteristics between (a) a conventional 2T eDRAM with boosted supplies and (b) the proposed 2T1C eDRAM with no boosted supplies.

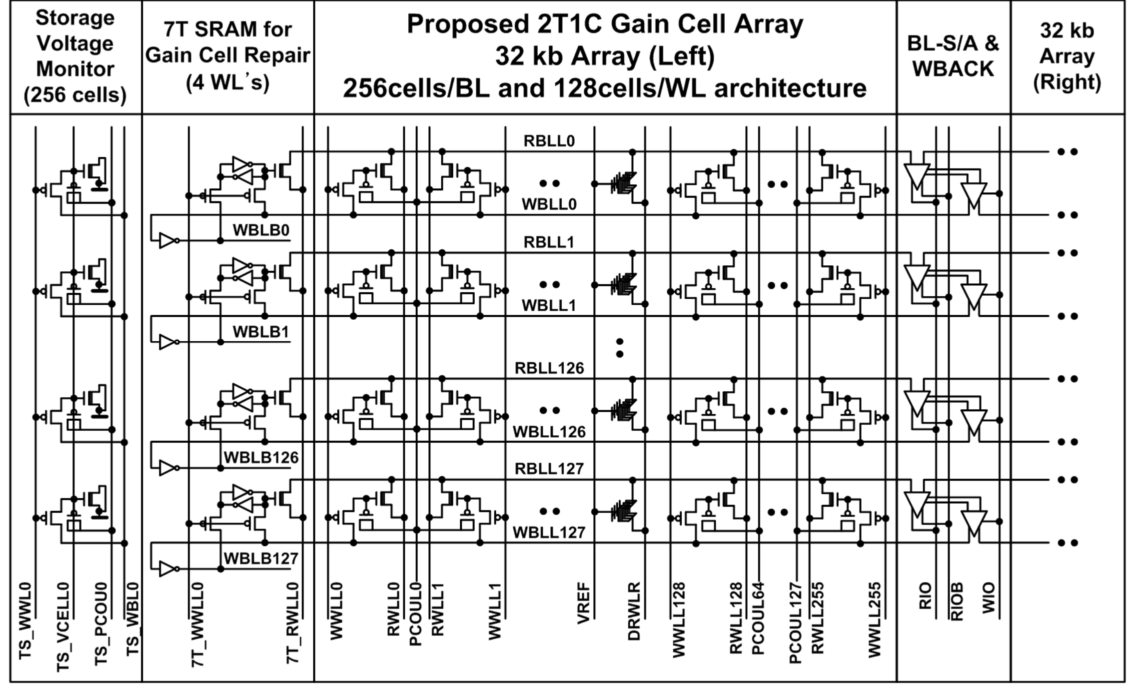


Fig. 4.6: Schematic diagram of a 64 kb 2T1C gain cell eDRAM macro with no boosted supplies.

In the array, two adjacent WL's share a single PCOU signal in order to minimize the cell size overhead. Since gain cells have a non-destructive read, the shared PCOU has no effects on the retention time of the unselected cells when the WL is activated. The shared PCOU reduces the bit-cell size by 21% compared to the separated layout shown in Fig. 4.7(a). Simulated waveforms in Fig. 4.7(b) confirm that the signal loss due to the redundant PCOU activation is negligible. Note that the storage capacitance of a gain cell is very small (<1 fF), so the additional power consumption due to the shared PCOU is also insignificant.

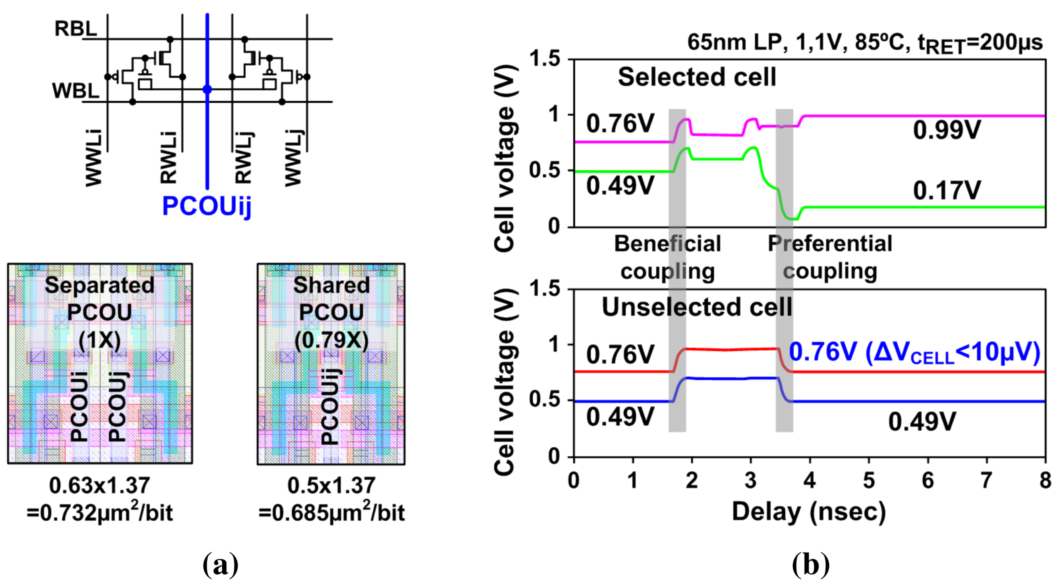


Fig. 4.7: Shared coupling signal (PCOU). (a) Bit-cell schematic and layouts. (b) Simulated waveforms show negligible disturbance in an unselected cell.

4.2.2 Decoupled 7T SRAM Repair Cell with Shared Control

Outlier cells having poor retention times are usually repaired using the same type of cell as the main array. However, gain cells have a very small storage capacitance, so the probability of having a failure cell in a redundant row or column is also high compared to

a 1T1C DRAM. The proposed 2T1C eDRAM has a narrower data window due to the reduced WWL voltage swing aggravating this situation. In order to improve the retention time of the 2T1C eDRAM, we devise a single-ended decoupled 7T SRAM based repair scheme. The proposed 7T SRAM consists of a decoupled read by replicating the 2T1C gain cell and a differential write using a locally generated complementary WBL signal (WBLB) as shown in Fig. 4.8. The pitch matched 7T SRAM cell shares control signals (i.e. RBL, WBL, WWL, RWL) with the main 2T1C array minimizing the area overhead. Note that WBLB is generated by an inverter inside the 7T SRAM cell while WBL is connected to every cell in the bitline direction as shown in Fig. 4.6. Therefore, the local differential write minimizes power dissipation incurred by the additional signal switching during memory access.

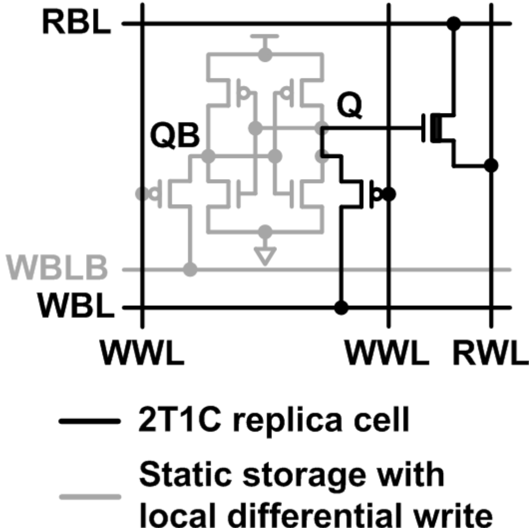


Fig. 4.8: Proposed decoupled 7T SRAM repair cell shares BL and WL signals with the 2T1C cell.

Fig. 4.9 shows the comparison of signal-to-noise margin (SNM) between the proposed 7T SRAM and a conventional 6T SRAM. The decoupled read structure of the

7T SRAM improves the read SNM by 113% than a 6T SRAM, and the write SNM of the 7T SRAM having a lower WBL voltage can be made comparable to that of a 6T SRAM by sizing optimization. As explained in Section 4.2.1, the data '1' WBL voltage is lower than VDD by 0.2 V in order to suppress the subthreshold leakages flowing into unselected cells during write in the absence of a boosted WWL voltage [33].

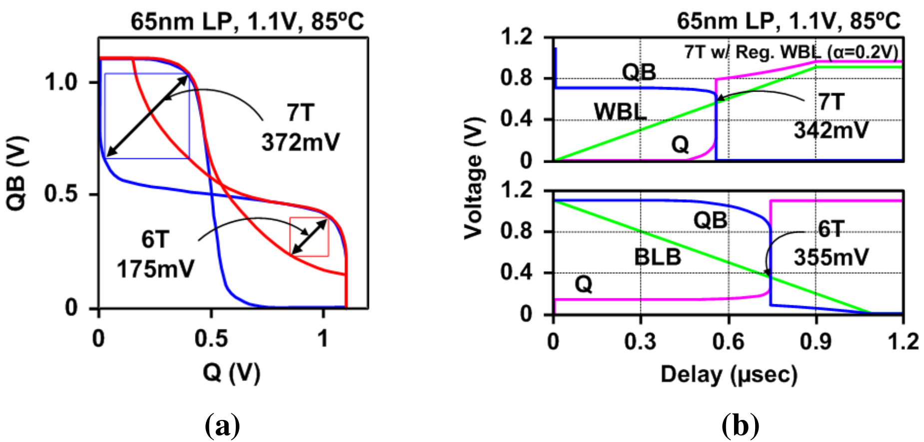
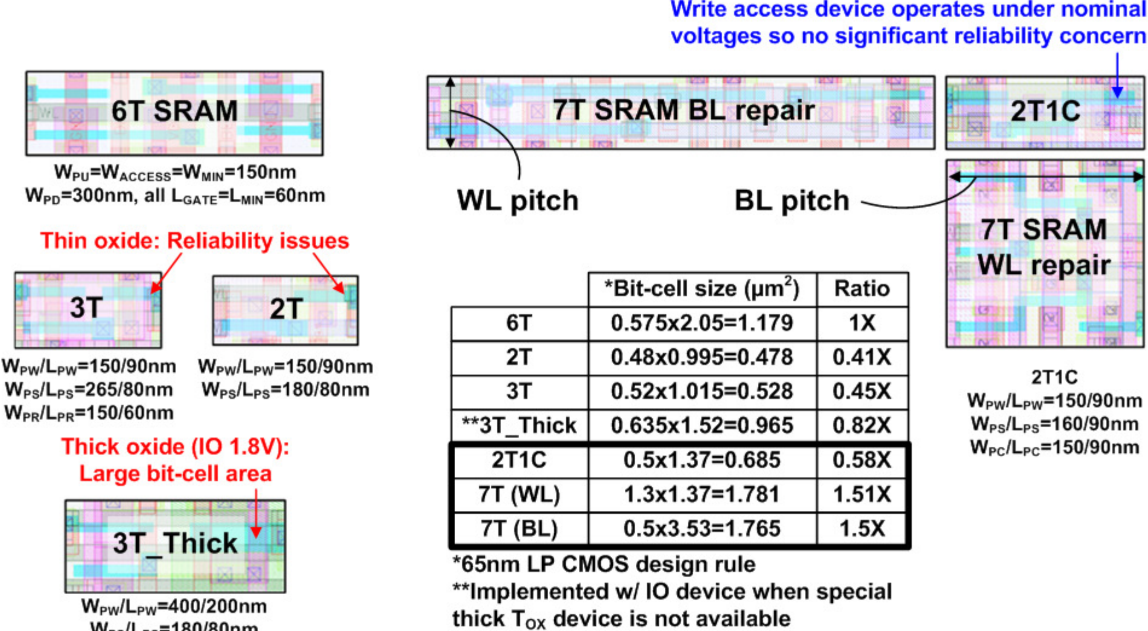


Fig. 4.9: Signal-to-noise margin (SNM) of a 6T SRAM and the proposed 7T SRAM. (a) Read SNM. (b) Write SNM.

Fig. 4.10 shows the transistor dimensions and layouts of the following memory cells: 6T SRAM, 2T gain cell, 3T gain cells using thin and thick T_{OX} devices, 7T SRAM, and 2T1C gain cell. All bit-cells were designed and drawn in a generic 65 nm LP process. Dense bit-cell design rules were not available to the authors but for area comparison purposes, using a logic design rule is a generally accepted practice. The 2T and 3T gain cells are 2.4X and 2.2X denser than a 6T SRAM, respectively. Similar cell area ratios have been reported in industry designs based on dense design rules; for example, the 2T gain cell in [19] is 2.1X denser than the 6T SRAM in [12], both implemented in Intel's 65 nm process. However, the density advantage of gain cell over 6T SRAM claimed in

prior literature is misleading since the boosted supply voltages will cause oxide reliability concerns. One way to get around this problem is to use 1.8 V I/O devices in which case the bit cell area density improvement compared to SRAM is reduced to around 1.2X. Since the array efficiency of gain cell eDRAMs is typically lower than SRAM due to charge pumps and the complex peripheral circuitry (e.g. RWL and WWL decoders for the decoupled bit-cell access, S/A with write-back circuits in each RBL and WBL [35]), gain cells no longer have an area advantage at the macro level when implemented using I/O devices. Conversely, the proposed 2T1C gain cell implemented using regular thin T_{OX} devices is 1.7X denser than a 6T SRAM without having an oxide reliability concerns.



	*Bit-cell size (μm^2)	Ratio
6T	0.575x2.05=1.179	1X
2T	0.48x0.995=0.478	0.41X
3T	0.52x1.015=0.528	0.45X
**3T_Thick	0.635x1.52=0.965	0.82X
2T1C	0.5x1.37=0.685	0.58X
7T (WL)	1.3x1.37=1.781	1.51X
7T (BL)	0.5x3.53=1.765	1.5X

Fig. 4.10: Bit-cell comparison(6T SRAM, 3T, 2T, 2T1C cells): All bit-cells were drawn in a generic 65 nm LP process.

For a 1 Mb macro including all peripheral circuitry, a 2T1C eDRAM is still 1.6X denser than a 6T SRAM array making it a viable alternative to conventional SRAM for last level caches

4.2.3 Cell Storage Monitor

Retention time of commodity DRAMs varies exponentially with temperature since it is highly sensitive to the junction and subthreshold leakages. Therefore, DRAM products have on-chip temperature sensors to control the refresh period adaptively according to the chip operating temperature [40]. Similarly, retention time of gain cells is also dependent on operating temperature since the storage node voltage changes according to the junction, subthreshold and gate leakages. However, the gate leakage has a weaker dependency on temperature and the various coupling effects illustrated in Fig. 4.4 makes a simple temperature sensor based refresh control ineffective for gain cell designs. To overcome this problem, we propose a gain cell based temperature sensor that directly measures the storage node voltage using a cell access pattern generator and 2T1C replica cells. Fig. 4.11 shows the proposed storage voltage monitor and its timing diagram. The SCAN signal triggers the cell access pattern generator (PG) that provides control signals (WBL, WWL, PCOU, and RWL) to the 2T1C replica cells. The repetitive access patterns have the same timing as the main array in order to track storage node voltages under a realistic memory access condition. The operating clock frequency of the PG generated by the VCO-1 indicates the current retention time setting. The merged storage node voltage of the 256 replica cells is captured by the sample-and-hold circuit. The buffered storage voltage using a unity gain amplifier is temporarily stored in MOS

capacitors implemented with thick T_{OX} devices whose gate leakage is negligible. The final storage node voltage is utilized to adaptively control the refresh rate of the 2T1C gain cell eDRAM. In this design, the measured storage voltage is translated in the form of frequency for convenient off-chip measurement, and the corresponding storage voltage can be found out using a calibration procedure. To remove any systematic error that may have been introduced while merging the 256 cells, the calibration step is needed to obtain the relationship between the measured storage voltage and the actual retention characteristic.

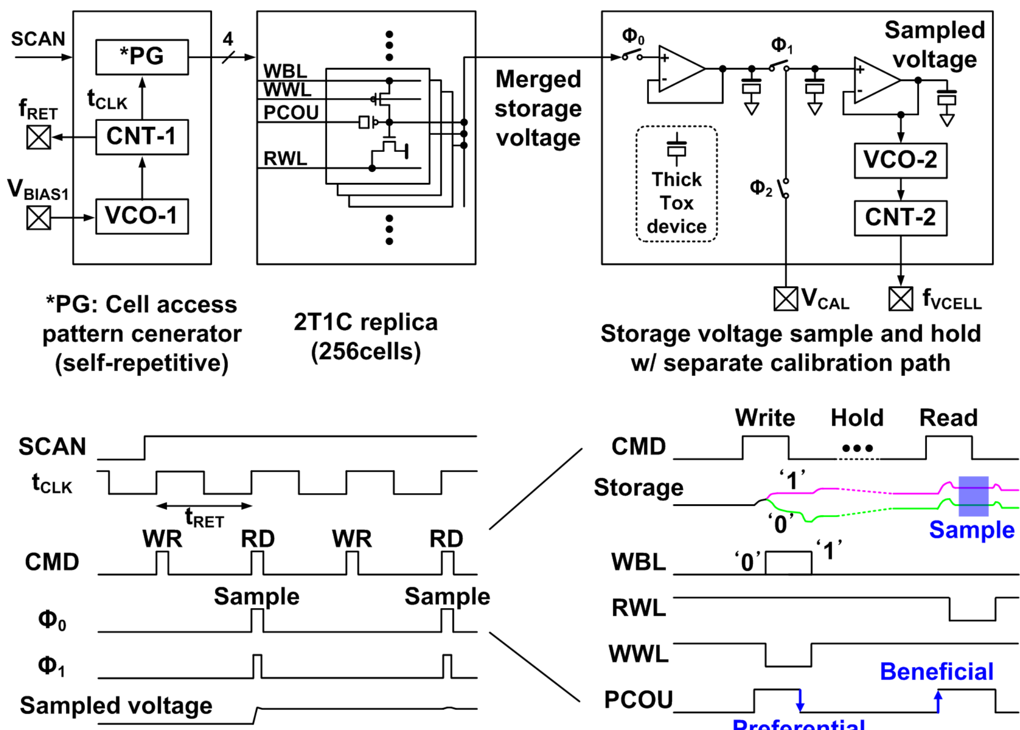


Fig. 4.11: Proposed storage voltage monitor for adaptive refresh control.

In real systems, operating temperature of cache memories is strongly related with the activity of the nearby cores [1]. Therefore, on-chip thermal sensors readily available

across the microprocessor can be utilized to control the storage voltage monitor. For example, the thermal sensor can trigger the monitor when there is a predetermined temperature change such as 10 ºC in the core area. The measured storage voltage is then sampled and the retention information is sent to the refresh rate control and event scheduler as shown in Fig 4.12. Repetitively sampling the storage node voltage stream lines the overall operation and removes any residual voltages in the sample-hold capacitors built using thick T_{OX} devices. During normal chip operation, two consecutive samples are enough to for a stable captured storage voltage. For a retention time of 500 μs, the average power dissipation of the monitor circuit is less than 1% of the total operating power dissipation as thermal conduction has a very long time constant in the order of hundreds of milliseconds [43]. In our design, the current consumption of the monitor circuit is 849 μA.

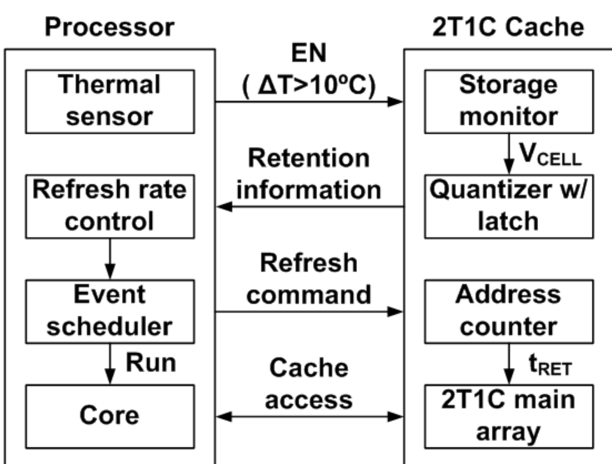
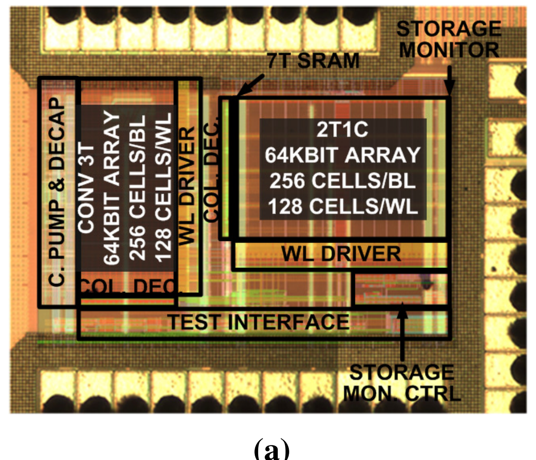


Fig. 4.12: Block diagram of the adaptive refresh control.

4.3 2T1C EDRAM Test Chip Measurements

A 128 kb test macro implemented in a 1.2V, 65 nm low-power logic CMOS process comprises a conventional 3T array and the proposed 2T1C array for performance comparison. Fig. 4.13 shows the chip microphotograph and key features of the 65 nm eDRAM test chip. Our design achieves a 1.4 ns (=714 MHz) random cycle and a 500 μs retention time (after a single-BL repair scheme) at 1.1 V and 85 ºC without using a boosted supply.



(a)

Process	65nm LP CMOS
Ckt dimension	556x345μm^2
Array size	2x64kbits (Conv. 3T & Prop. 2T1C)
Cell size	58% of 6T SRAM
Retention time	500μs @ 1.1V, 85ºC
Random cycle time	1.40ns (714MHz) @ 1.1V
VMIN	0.7V @ 10μs retention
Refresh power	*161.8μW per Mb (0.28X of **6T SRAM)

*@ 1.1V, 85ºC, 500μsec refresh rate
**@ Retention voltage of 0.6V

(b)

Fig. 4.13: (a) Microphotograph of the 65 nm eDRAM test chip. (b) Chip feature summary.

Fig. 4.14 shows the measured retention time distribution of a three eDRAM implementations: conventional 3T, the previous 2T [35], and the proposed 2T1C. The amount of boosting (Δ) above VDD and below GND is 0.5 V for the 2T and 3T eDRAMs whereas the 2T1C operates under a nominal power supply level. The single-ended sensing nature and the small storage capacitance of conventional 3T eDRAMs result in the poor retention characteristics. The asymmetric 2T eDRAM achieves a 400 μs retention time for a 99.9% bit yield condition at 1.1 V and 85 ºC. The retention time for a 99.99% bit yield is estimated to be 80 μs at 105 ºC which is 2x longer than that reported

for a commercial 1T1C eDRAM under the same yield and temperature condition [6]. Therefore, it is fair to say that an asymmetric 2T eDRAM has a retention time that is comparable to real product eDRAMs. However, the need for a special thick T_{OX} device would limit the wide spread adoption of asymmetric 2T eDRAMs, especially for fabless companies. The proposed 2T1C eDRAM achieves similar performance as previous designs but without any boosted supplies.

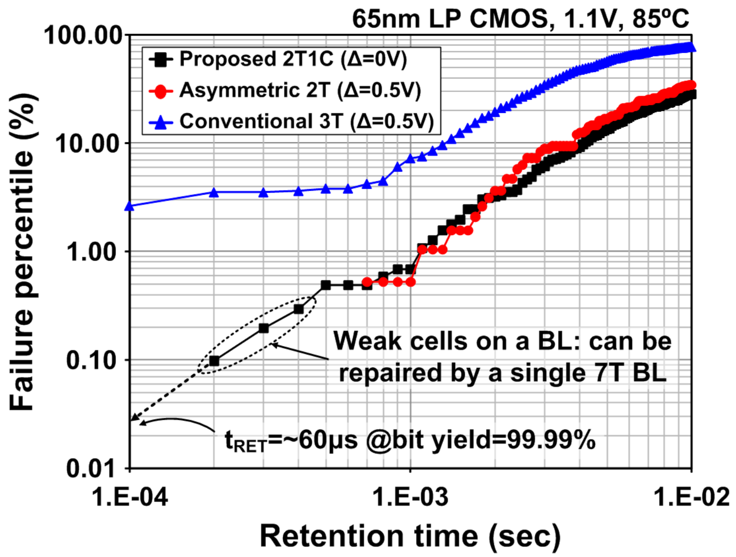


Fig. 4.14: Measured retention time distribution.

Single-ended sensing methods usually exhibit more BL failures than WL failures since variation in the dummy reference cells and the BL-S/A offset impacts the read margin of the entire BL. A decoupled 7T SRAM array was implemented to evaluate the effectiveness of a repair scheme under variation effects in the dummy cell and BL-S/A. The measured retention bit-map of a 1 kb 2T1C sub-array shows weak bit-lines as well as randomly located weak cells (Fig. 4.15). The proposed 7T SRAM sharing the same BL-S/A shows better stability compared to a 2T1C cell under the same operating condition.

Based on the measured retention time distribution of a 2T1C array in Fig. 4.16(a), we can estimate the effectiveness of various repair schemes. A single BL repair scheme using a redundant 2T1C bitline will fail to meet a target retention time of 500 μs with a probability of 6.25%. On the other hand, a single BL repair scheme based on a 7T SRAM for an array with 128 BLs can improve the retention time by 150% (200 μs to 500 μs) while the array overhead is 1.23% as shown in Fig. 4.16(b).

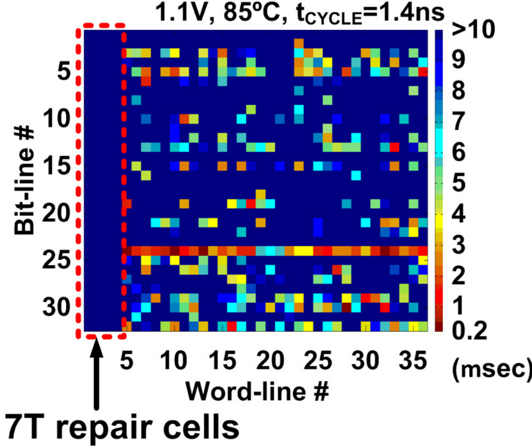
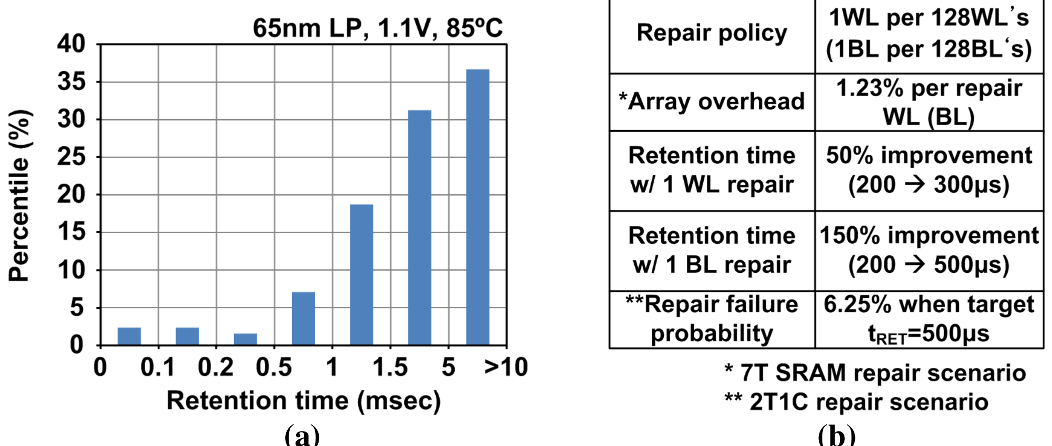


Fig. 4.15: Measured retention bit-map of 2T1C and decoupled 7T arrays.



Repair policy	1WL per 128WL's (1BL per 128BL's)
*Array overhead	1.23% per repair WL (BL)
Retention time w/ 1 WL repair	50% improvement (200 → 300μs)
Retention time w/ 1 BL repair	150% improvement (200 → 500μs)
**Repair failure probability	6.25% when target t_{RET}=500μs

* 7T SRAM repair scenario
** 2T1C repair scenario

(b)

Fig. 4.16: (a) Measured retention time distribution of the 2T1C array. (b) Effectiveness of various repair schemes.

Fig. 4.17(a) shows the measured retention characteristics of the 2T1C eDRAM at 25 ºC and 85 ºC, respectively indicating a 5X retention time difference in the tail cells between the two temperatures. This implies that a significant reduction in refresh power dissipation can be achieved at lower temperatures by adjusting the refresh rate accordingly. Storage voltages were measured using the proposed monitor scheme at various temperatures and retention times as shown in Fig. 4.17(b). The measured storage voltage includes all coupling effects during memory access as well as the change in leakage currents at different temperatures.

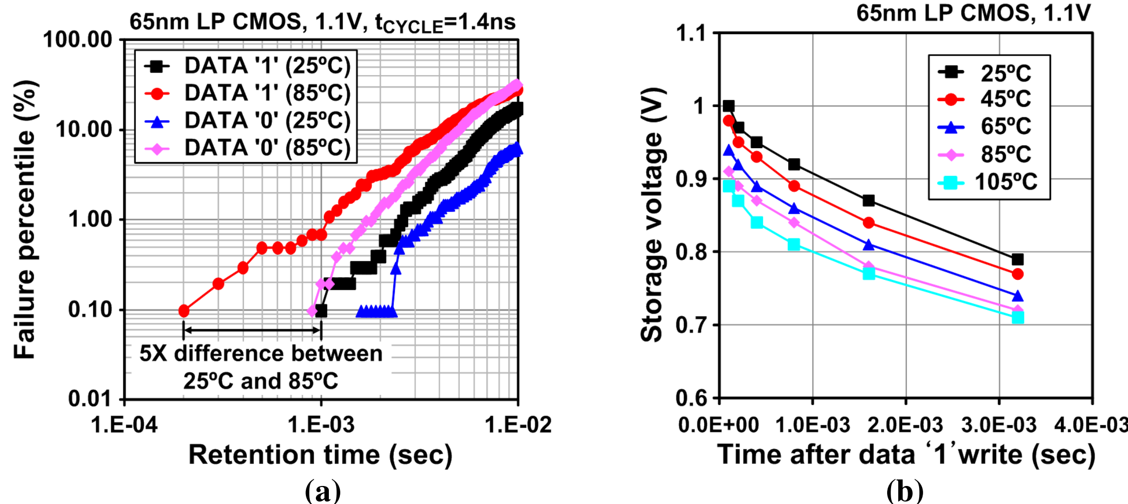


Fig. 4.17: (a) Measured retention time distribution of data '1' and '0' at 25 ºC and 85 ºC. (b) Measurement storage voltage with varying temperature and retention time.

Fig. 4.18 shows the static current comparison between a 1Mb SRAM in power down mode and proposed 1Mb 2T1C eDRAM. The data retention voltages of the 6T SRAM and the 2T1C eDRAM are 0.6 V and 1.1 V, respectively. The static current of the 6T SRAM decreases exponentially at lower operating temperatures. Similarly, the static current of 2T1C eDRAM can be reduced exponentially by adjusting the refresh rate using

the proposed storage voltage monitor. The static current of the proposed 2T1C eDRAM is 72% and 83% smaller than that of 6T SRAM at 85 ºC and 105 ºC, respectively. Without an adaptive refresh control, the 2T1C eDRAM has larger static power dissipation than the 6T SRAM at lower operating temperatures such as below 65 ºC in our tests.

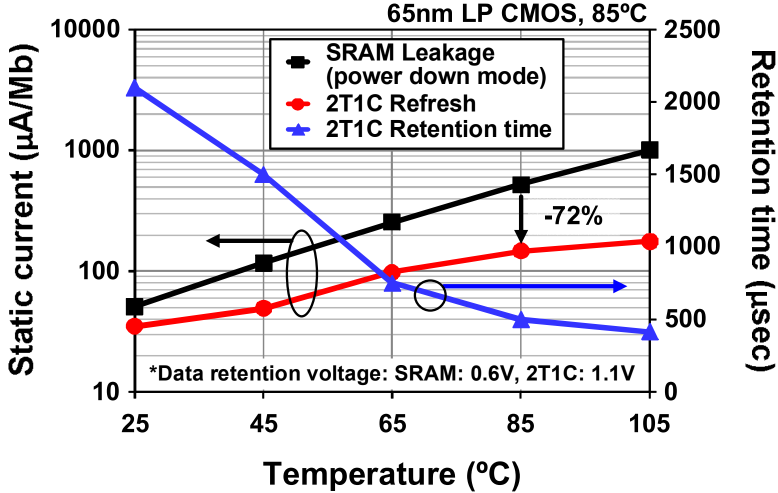


Fig. 4.18: Comparison of static current between SRAM (with power-gating) and 2T1C eDRAM (with adaptive refresh control).

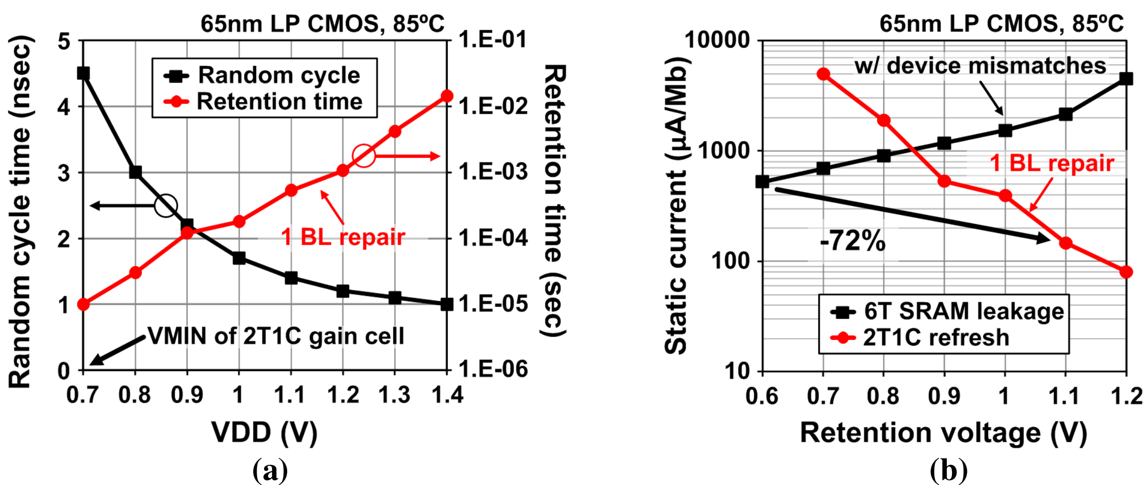


Fig. 4.19: Measured VDD shmoo. (a) Random cycle time and retention time of the 2T1C eDRAM. (b) Static power dissipations of a 6T SRAM and the 2T1C eDRAM.

The measured VDD shmoo of random cycle time, retention time and the corresponding static power dissipations in Fig. 4.19 shows a wide operating voltage range from 1.4 V down to 0.8 V. One unique aspect of eDRAMs (both 1T1C and gain cell) that may not be very obvious to SRAM designers is that a lower operating voltage does not necessarily result in a lower static power consumption as shown in Fig. 4.19 (b). This is contrary to SRAMs where the static power goes down at lower supply voltages making V_{MIN} the chief design parameter. Static power in eDRAM is dominated by refresh power which can be lower at higher supply voltages due to the robust cell retention characteristics. So the operating voltage for eDRAMs should be chosen not based on the functional V_{MIN} of the memory, but based on the lowest power consumption point which tends to be higher than an SRAM V_{MIN}.

65nm CMOS	6T SRAM [12]	1T1C eDRAM [6]	2T eDRAM [35]	This work [37]
Cell Schematic	WL, WL, BL, BLB	BL, WL, VP	RBL, WWL, RWL, WBL	RBL, WBL, WWL, PCOU, RWL
Process	Logic-compatible	Logic compatible +2 (FEOL)+3 (Cap)	Logic compatible +2 (FEOL)	Logic compatible
Boosted supplies	No required	Required (High & Low)	Required (High & Low)	No required
(1)Reported cell size (ratio)	$135F^2$ (1X)	$30F^2$ (0.22X)	$65F^2$ (0.48X) [19]	NA
(2)Redrawn cell size (ratio)	0.575x2.05= 1.179μm² (1X)	0.45x0.545= 0.245μm² (0.21X)	0.48x0.995= 0.478μm² (0.41X)	0.50x1.37= 0.685μm² (0.58X)
(2)Redrawn 1Mb macro (ratio)	1.377x1.124= 1.548mm² (1X)	0.632x0.739= 0.467mm² (0.30X)	1.168x0.638= 0.746mm² (0.48X)	1.191x0.807= 0.961mm² (0.62X)
Data storage	Latch (Static)	Capacitor (20fF)	MOS gate (<1fF)	MOS gate (<1fF)
Cell access	(+) Differential read (-) Ratioed operation	(-) Destructive read (-) Refresh	(+) Decoupled read and write, (-) Refresh	(+) Decoupled read and write, (-) Refresh
Random cycle	(2)1GHz	500MHz	(2)667MHz	(2)700MHz
Retention time @99.9% yield	NA	NA	400μs @85ºC (*Meas.*)	500μs @85ºC (*Meas.*)
Retention time @99.99% yield		40μs @105ºC (*Meas.*)	80μs @105ºC (*Est.*)	50μs @105ºC (*Est.*)
Static power	1X	0.2X	0.19X @500MHz	0.28X @700MHz

(1)All designs are in 65nm, (2)Based on the same 65nm low power CMOS process

Fig. 4.20: Comparison between our design and several embedded memory options.

Fig. 4.20 compares the proposed 2T1C eDRAM with several other embedded memory options in the same 65 nm LP process. The measured random cycle time of the 2T1C eDRAM is 40% faster than that of a 1T1C eDRAM while achieving a similar retention time at 105 ºC under a 99.99% bit yield condition. The 1T1C eDRAM has replaced 6T SRAMs in IBM's POWER7™ microprocessor [6]. Although the cycle time of the 1T1C eDRAM is 2X longer than that of 6T SRAMs, the smaller memory footprint and shorter global interconnect delay leads to a high overall cache performance. Bit-cell size and random cycle time of the proposed 2T1C eDRAM stands between those of 6T SRAM and 1T1C eDRAM, and the read and write paths can be optimized separately allowing gain cells to scale favorably in future technology nodes. Our experimental results show that gain cell based eDRAMs can be a strong contender for future embedded memories.

4.4 Conclusions

Several circuit techniques have been presented for enabling a truly logic-compatible gain cell eDRAM with a competitive bit-cell size and improved memory performance. The proposed 2T1C gain cell utilizes a beneficial coupling that enhances read margin and a preferential boosting that improves write margin. This unique feature allows us to achieve robust DRAM operation without any boosted supplies. A decoupled 7T SRAM was seamlessly integrated as part of the array by sharing control signals with the main 2T1C array. The retention time of the 2T1C eDRAM was improved by 2.5X using the 7T SRAM based repair scheme while the repair failure rate was 6.25% when using

redundant 2T1C cells. The array overhead of the 7T SRAM repair is 1.23% for a single redundant BL for every 128 BL's. The storage voltage monitor tracks the retention characteristics of the 2T1C gain cell under PVT variations while capturing realistic coupling effects during memory access. Measurement results show a 714 MHz random cycle using a 500 µs refresh period for a 1 BL repair scheme at 1.1 V, 85 ºC. The static power dissipation including refresh currents and cell leakages was 161.8 µA/Mb at 1.1 V and 85 ºC which is 72% lower than that of a power gated SRAM with a data retention voltage of 0.6 V.

Chapter 5

A Scalability Exploration of STT-MRAMs Considering Variation Effects

5.1 Introduction to STT-MRAM

Spin-torque-transfer magnetic RAMs (STT-MRAMs) are gaining popularity in the research community due to their compact bit-cell structure, excellent scalability and non-volatility [23]-[32]. An STT-MRAM bit-cell consists of an access transistor and a magnetic tunnel junction (MTJ). The MTJ device has a free magnetic layer and a pinned magnetic layer which are separated by a thin insulator layer as shown in Fig. 5.1(a). The simple structure makes the bit-cell size of STT-MRAMs comparable to that of 1T1C eDRAMs in a memory specific process. The MTJ is schematized as a two-terminal device with varying resistance. The typical relationship between MTJ resistance and write (WR) current (R-I hysteresis curve) is shown in Fig. 5.1(b). Depending on the direction of the WR current, magnetization of the two layers can be set to a parallel state (P: low resistance, data '0') or an anti-parallel state (AP: high resistance, data'1') using

spin polarized current as illustrated in Fig. 5.1(c). Read (RD) operation is accomplished by sensing the resistance difference between the two states using voltage or current Sense Amplifier (S/A).

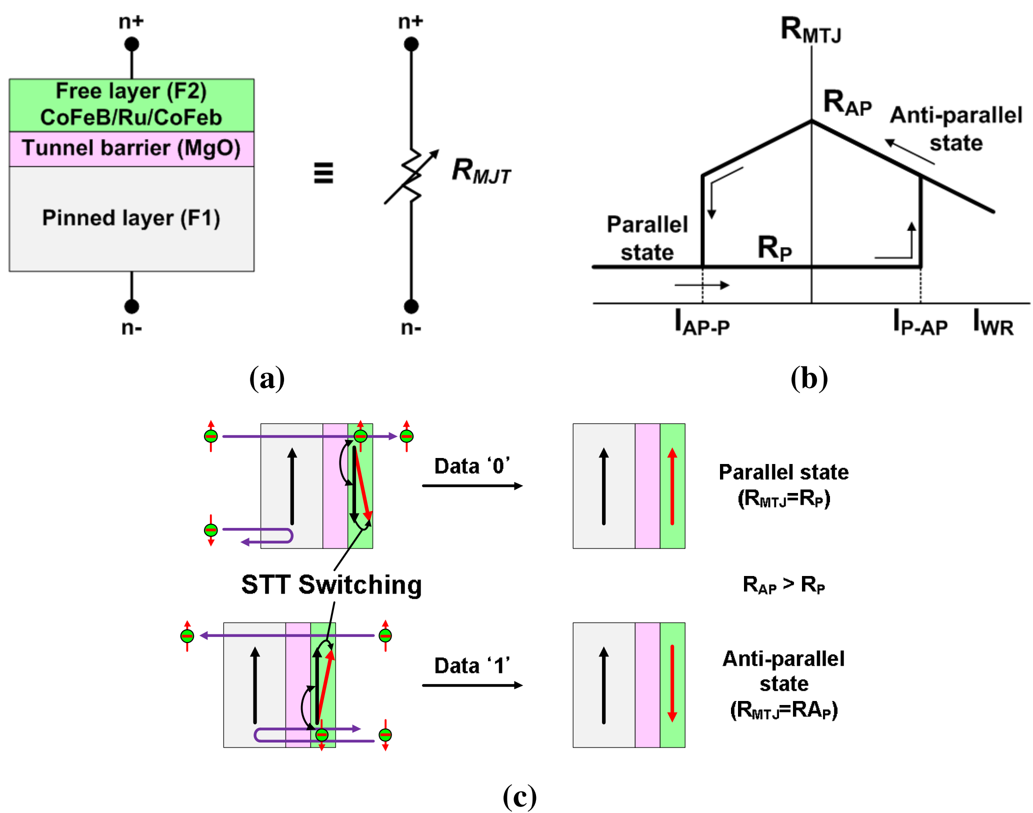
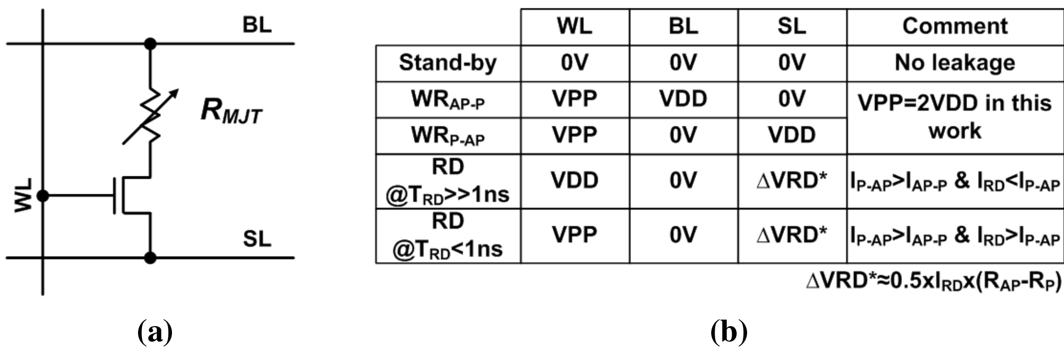


Fig. 5.1: (a) Magnetic tunnel junction (MTJ) stack and its corresponding circuit schematic as a two-terminal device with varying resistance. (b) Resistance vs. write current (R-I) hysteresis curve. (c) Illustration of spin torque transfer (STT) switching's.

The state reversal happens only to the selected bit-cell when the current flowing into the MTJ is larger than its threshold current (write threshold current; I_{C0}). The STT switching originates from the exchange of angular momentum between a spin-polarized current and the magnetization of the free layer. The localized spin-injection within a bit-cell enables the excellent write selectivity with no high oxide field involved nor high

temperature required during the switching, which are the most critical scaling challenges in currently practical non-volatile memories such as FLASHs and phase-change RAMs (PCRAMs or PRAMs). Most interestingly, the I_{C0} decreases exponentially with technology scaling as the critical current density (J_{C0}) remains constant due to the STT switching phenomenon when there is no thermal stability constraint. Therefore, the slow write time (T_{WR}) which is several nanoseconds or even larger, is projected to be improved with technology scaling. Fig. 5.2 shows the STT-MRAM bit-cell schematic and signal voltage conditions for each operating mode.



	WL	BL	SL	Comment
Stand-by	0V	0V	0V	No leakage
$WR_{AP\text{-}P}$	VPP	VDD	0V	VPP=2VDD in this work
$WR_{P\text{-}AP}$	VPP	0V	VDD	
RD @T_{RD}>>1ns	VDD	0V	ΔVRD*	$I_{P\text{-}AP} > I_{AP\text{-}P}$ & $I_{RD} < I_{P\text{-}AP}$
RD @T_{RD}<1ns	VPP	0V	ΔVRD*	$I_{P\text{-}AP} > I_{AP\text{-}P}$ & $I_{RD} > I_{P\text{-}AP}$

ΔVRD*≈0.5xI_{RD}x(R_{AP}-R_P)

(a) (b)

Fig. 5.2: (a) STT-MRAM bit-cell schematic. (b) Signal voltages for each operating mode.

Despite the recent advances in STT-MRAM fabrication and circuit techniques [23]-[32], it is still unclear whether this emerging memory technology can achieve higher overall performance than conventional SRAMs or eDRAMs in future technology nodes in the presence of variation effects. In this work, I explore the scalability and variability of STT-MRAM by comparing its performance with 6T SRAM from 65 nm to 8 nm process nodes.

5.2 STT-MTJ Scaling Roadmap

Fig. 5.3 shows the proposed scaling methodology for both in-plane and perpendicular STT-MTJs. The lateral dimension of the in-plane MTJ is fixed at 2FxF, where F is the half-pitch for a given process node, while the diameter of the perpendicular MTJ is fixed at F. This enables the smallest bit-cell size for standalone as well as embedded memories. Although the bit-cell size of embedded memories can be larger than that of standalone counterparts, I stick to the aforementioned MTJ dimensions in order to evaluate the scalability of STT-MRAMs at the worst condition. With technology scaling, we should maintain signal to noise margins (SNMs) for SRAMs and C_S/C_{BL} ratio with practical retention time for eDRAMs in order to achieve stable memory operations. Similarly, the J_{C0}•RA (J_{C0}: critical current density in MA/cm^2 and RA: resistance area product in Ω•μm^2) design space and the thermal stability (Δ) of STT-MRAMs needs to be maintained for optimal RD and WR performances as well as non-volatile properties.

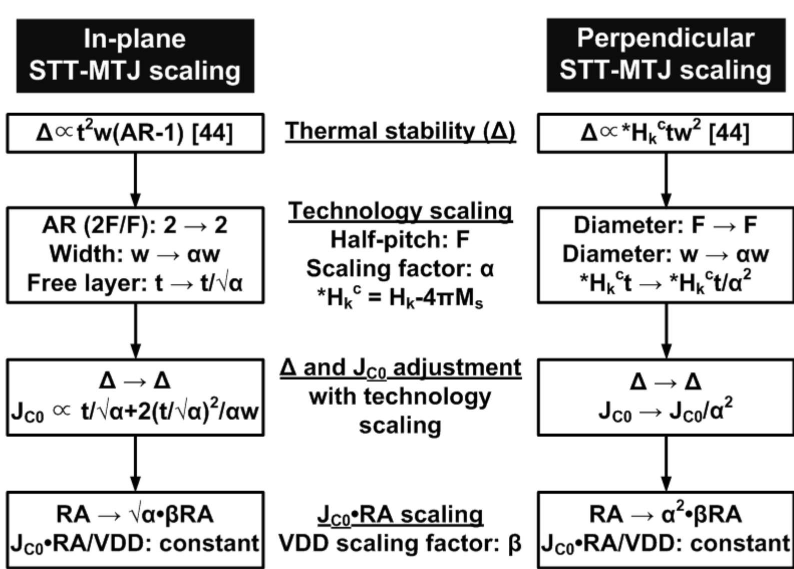


Fig. 5.3: STT-MTJ scaling scenario based on dimensional adjustment and/or material innovation in order to maintain non-volatility and achieve optimal RD and WR operations.

5.2.1 In-Plane STT-MTJ Scaling Scenario

The thermal stability of in-plane STT-MTJs is mainly affected by the elongated shape of the free layer. The fundamental equation is expressed as

$$\Delta_{in-plane} = \frac{E}{k_B T} = \frac{K_U V}{k_B T} = \frac{H_K M_S V}{2 k_B T} \tag{1}$$

where K_U is the uniaxial anisotropy energy density, V is the volume of the free layer, T is the operating temperature, H_K is the anisotropy, and M_S is the saturation magnetization. The in-plane shape anisotropy H_K can be approximated as

$$H_K \approx \frac{8\pi M_S t(AR-1)}{wAR} \tag{2}$$

where w, AR, and t are width, aspect ratio, and thickness of the free layer, respectively [44]. Then, the thermal stability of in-plane STT-MTJs can be simplified as

$$\Delta_{in-plane} \propto t^2 w(AR-1) \tag{3}$$

As technology scales, w is shrunk to αw, where α is a scaling factor. This requires the increase of t by $1/\sqrt{\alpha}$ times in order to maintain thermal stability when AR remains at a constant. The thickened free layer increases J_{C0}, and it is expressed as

$$J_{C0} = \frac{2ea M_S t(H_K + H_{ext} + 2\pi M_S)}{\hbar \eta} \tag{4}$$

where e is the electron charge, a is the damping constant, $\hbar$ is the reduced Planck's constant, H_{ext} is the external field, and η is the spin transfer efficiency [45]. If we combine (2) with (4) while assuming that H_{ext}=0, then J_{C0} can be expressed as

$$J_{C0} \propto t(1+\frac{2t}{w}) \xrightarrow{scaling} J_{C0}' \propto \frac{t}{\sqrt{\alpha}}(1+\frac{2t/\sqrt{\alpha}}{\alpha w}) \tag{5}$$

This indicates that J_{C0} scales by ~$1/\alpha^{0.5}$ and write threshold current (I_{C0}) by ~$\alpha^{1.5}$ when $t \ll w$, where t=1 nm and w=65 nm at 65 nm process node in this work. In the J_{C0}•RA design space, a lower I_{C0} improves the write time of STT-MRAMs at the expense of read margin [46]. In order to ensure the optimal tradeoff between the voltage headroom during WR and the sensing margin during RD, a constant J_{C0}•RA/VDD scaling scenario is adopted. The J_{C0}•RA/VDD value was chosen as 0.25 at 65 nm process node determined by empirical methods, and this value remains as constant throughout the scalability analysis. Note that, in reality, the J_{C0}•RA/VDD value should be further adjusted in each technology for optimal RD and WR performances. The aforementioned STT-MTJ scaling scenario is summarized in Fig. 5.3 (left).

5.2.2 Perpendicular STT-MTJ Scaling Scenario

Compared to in-plane STT-MTJs, perpendicular counterparts have no shape anisotropy. The thermal stability of perpendicular STT-MTJs is directly proportional to bulk crystalline anisotropy (H_K^C) and the volume of the free layer that is an analogy with the cell capacitance of 1T1C DRAMs, where its capacitance is proportional to dielectric constant and area. The thermal stability of perpendicular STT-MTJs is given by

$$\Delta_{perpendicular} = \frac{K_U V}{k_B T} = \frac{H_K{}^C M_S V}{2k_B T} = \frac{H_K{}^C M_S \frac{\pi}{4} w^2 t}{2k_B T} \propto H_K{}^C w^2 t \tag{6}$$

where the bulk crystalline anisotropy $H_K^C = H_K - 4\pi M_S$ for perpendicular STT-MTJs [44]. With technology scaling, $H_K^C \bullet t$ should be adjusted to $H_K^C \bullet t / \alpha^2$ in order to maintain the thermal stability since w is shrunk to αw. Consequently, the J_{C0} of perpendicular STT-MTJs scales by $1/\alpha^2$ and I_{C0} remains as constant. Similar to in-plane case, RA values were adjusted such that $J_{C0} \bullet RA/VDD$ to be constant as shown in Fig. 5.3 (right).

5.2.3 STT-MTJ Scaling Trend

Year		2007	2010	2012	2015	2018	2021	2024
Technology node (nm)		65	45	32	22	15	11	8
VDD: Supply voltage (V)		1.2	1.1	1	0.9	0.85	0.8	0.75
On-chip cache memory size (MByte)		16	24	32	48	64	96	128
Number of cores		4	6	8	12	16	24	32
Δ: Thermal stability (for 10 yrs retention)		72	73	74	74	75	75	76
*t: Free layer thickness	In-plane	1.00	1.21	1.44	1.74	2.11	2.48	2.91
*$H_k^c \bullet t$: Anisotropy and t	Perpendicular	1.00	2.11	4.19	8.93	19.32	36.24	68.91
$J_{C0_P\text{-}AP}$: Critical current density (MA/cm^2)	In-plane	3.00	3.70	4.55	5.86	7.87	10.45	14.65
	Perpendicular	1.50	3.16	6.28	13.40	28.97	54.35	103.37
$I_{C0_P\text{-}AP}$: Threshold write current (μA)	In-plane	253.5	149.9	93.3	56.7	35.4	25.3	18.7
	Perpendicular	49.8	50.2	50.5	50.9	51.2	51.7	52.0
$J_{C0} \bullet RA/VDD$ (MTJ voltage headroom)		0.25 when TMR=150%						
$R_{AP}A$: Resistance area product (Ω•μm^2)	In-plane	10.00	7.43	5.49	3.84	2.70	1.91	1.28
	Perpendicular	20.00	8.71	3.98	1.68	0.73	0.37	0.18
J_{RD}/J_{C0}: Read current density		1.50 (based on our analysis in Fig. 11 as well as [51])						

*t and $H_k^c \bullet t$ are normalized to 65nm technology node.

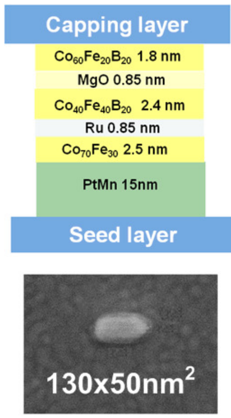
Fig. 5.4: In-plane and perpendicular STT-MTJ scaling trends based on Fig. 5.3.

Fig. 5.4 shows the in-plane and perpendicular STT-MTJ scaling trends based on the proposed scaling scenario shown in Fig. 5.3. The number of cores is 4 at 65 nm process node and has been doubled at every two process generations based on Intel's server processor trends with the commensurate increase of on-chip memory densities such that each core has a dedicated 4 MB L3 cache [1]-[3]. The required thermal stabilities for 10 year retention were calculated based on cache densities and allowable chip failure rates.

The chip failure rate (F_{chip}) can be estimated by expanding the cell reversal probability as follows:

$$F_{chip} = 1 - \exp[-m \frac{t}{\tau_0} \exp\{-\frac{E}{k_B T}(1 - \frac{I_{cell}}{I_{C0}})\}] \tag{7}$$

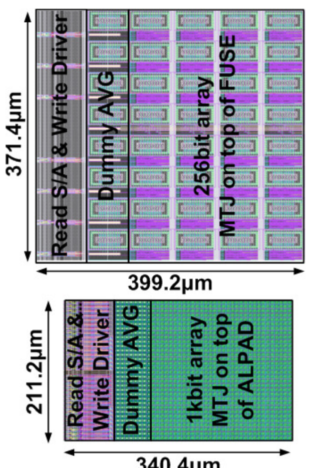
where m is the cache density, t is the 10 years, and I_{cell} is the cell current [29]. The allowable chip failure rate is determined by the capability of ECC and/or repair schemes. For example, at 22 nm process node with 48 MB density, the allowable F_{chip}=7.947e-7 with a repair scheme having a redundant WL and BL per every 64 WL's and 64 BL's, respectively, and the corresponding required thermal stability is 74. The J_{C0} and RA values at 65 nm process node were calibrated and refined based on our MTJ test devices as well as the previously reported data in [23], [24]. The fabricated MTJ structure, SEM image, and summary of measured data are shown in Fig. 5.5 [47].



MTJ dimension (nm^2)		130x50
Thermal stability		29
JC0 (MA/cm^2)	P-AP	6.81
	AP-P	2.57
RA (Ω•μm^2)		12.86
TMR (%)		126
Ultra fast switching (ps)	P-AP	560
	AP-P	586
Optimal write energy (pJ/bit)	P-AP	0.706 @T_{WR}=0.68ns
	AP-P	0.286 @T_{WR}=1.54ns

Fig. 5.5: Vertical structure and SEM image of fabricated STT-MTJ (left) and summary of measured MTJ parameters (right).

A STT-MTJ characterization macro was implemented in a 130 nm CMOS process. Fig. 5.6 shows the array layout and feature summary of the STT-MRAM test chip.



Process	130nm 1P8M CMOS
Chip dimension	3000x1518μm^2
MTJ process after CMOS	1kbit Al pad array, 256bit fuse array
WR pulse width	Sub 1ns ~ DC
Sensing method	I-applying & V-sensing, V-applying & I-sensing
Reference cell	Dummy cell averaging with MTJ replica cells

Fig. 5.6: STT-MTJ characterization array layout and test chip feature summary.

5.3 STT-MRAM HSPICE Simulation Methodology

In order to compare the performances of STT-MRAMs with 6T SRAMs considering variation effects while technology scales, this section presents the proposed simulation methodology including an accurate MTJ macromodel, transistor parameters, sub-array architectures, and variation sources.

5.3.1 STT Switching and MTJ Macromodel

For efficient Monte Carlo simulations, an accurate MTJ macromodel capturing key MTJ properties such as hysteresis, TMR dependency on bias voltage, and the relationship between I_{WR} and T_{WR} was adopted [48], [49]. Fig. 5.7 shows examples of the MTJ macromodel fitted well with experimental data from [23] as well as our MTJ test devices in the characterization macro indicating the write time dependency on the write current. There are three distinct STT switching modes; thermal activation, dynamic reversal and precessional switching [22]. For fast switching in nanosecond regime, the required

switching current density is several times greater than the critical current density (J_{C0}). The required switching current density is estimated as follow:

$$J_{C0}(\tau) \propto J_{C0} + [\frac{\ln(\pi/2\theta)}{\tau}] \tag{8}$$

where τ is the pulse width of switching current and θ is the initial angle between the magnetization vector of the free layer and the easy axis.

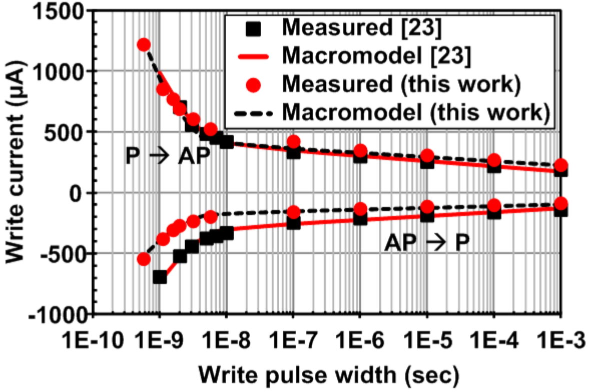


Fig. 5.7: MTJ macromodel fitting results using MTJ data from our characterization array and [23].

At a finite temperature, thermal agitation plays an important role in reducing the switching current at long write pulses (>10 ns). In this slow thermal activated switching regime, the switching current is dependent on the write current density and the thermal stability factor of the free layer.

$$J_C(\tau) = J_{C0}[1 - \frac{k_B T}{E}\ln(\frac{\tau}{\tau_0})] \tag{9}$$

where τ_0 ~1 ns is the inverse of the attempt frequency. The critical current density (J_{C0}) can be determined by extrapolating the experimentally obtained switching current density

(J_C) at $\tau=\tau_0$. The J_{C0} is a good measure of STT performance in a nano-magnetic device and corresponds to J_C value at switching times ranged from roughly 5 to 10 ns for room temperature operation. The dynamic switching occurs at intermediate write pulses within a small range of 3 to 10 ns, which corresponds to the operating speed of currently practical STT-MRAMs [28]-[32].

5.3.2 Transistor Scaling Trend

As for simulating access devices and peripheral circuitries, transistor parameters available on ITRS [50] were utilized from 65 nm down to 8 nm. Based on high performance logic technology requirements of ITRS, we reproduced core NMOS parameters using the ITRS provided MASTAR tool that has been extensively used to predict electrical characteristics of advanced CMOS transistors. The resulting I_{dsat}'s of the core NMOS transistors were linearly extrapolated in order to make them have a gradual improvement. This prevents the performance trends of SRAMs and STT-MRAMs from being distorted by any abrupt change in transistor parameters. The V_{thsat}'s of the core PMOS transistors are the same with the NMOS counterparts while the I_{dsat}'s of PMOS were determined based on ITRS $I_{on,n}/I_{on,p}$ ratios. A boosted wordline scheme limited to 2VDD was adopted for a reliable WR operation with a commensurate increase in T_{OX} for oxide reliability and a 1.2X longer gate length (L_{gate}) for variability. The transistor parameters of this special thick oxide device were also extracted using the MASTAR tool. Further increase in the boosted voltage will result in oxide reliability issues and difficulties in generating a boosted level stably as witnessed in DRAM designs.

The I_{dsat} and V_{thsat} trends of the core thin and the special thick T_{OX} devices are shown in Fig. 5.8 with indicating the innovations of CMOS technologies.

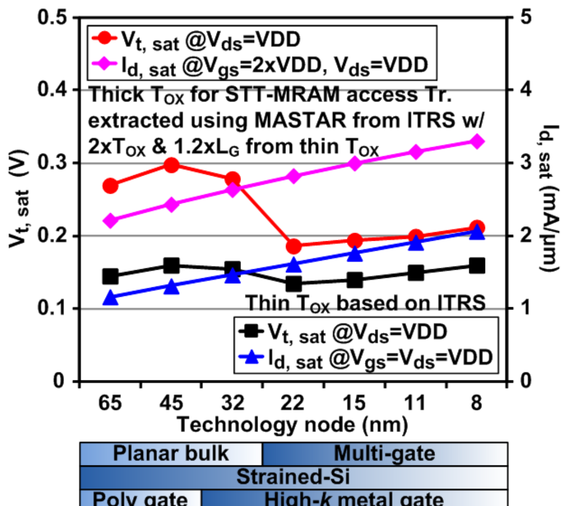


Fig. 5.8: High performance (HP) transistor scaling trend based on ITRS.

5.3.3 Sub-Array Architecture and Variation Sources

6T SRAMs used for the comparison have the following transistor dimensions: $W_{PU}=W_{min}$, $W_{PD}=2xW_{min}$, and $W_{ACCESS}=W_{min}$, with all devices using a minimum channel length. This is the most general sizing scheme and extensive Monte Carlo simulations were performed to verify good read and write margins. The width of the STT-MRAM access device (W_{TX}) is chosen as 12F based on a 2T1MTJ style cell layout [29]. This makes the cell size of the STT-MRAM comparable to that of an eDRAM in a memory specific process or 3X denser when compared to an SRAM cell in a generic logic process.

Fig. 5.9 shows the 128 kb sub-array architectures of 6T SRAM and STT-MRAM that has been extensively used in this work for performance evaluations. The unit 128 kb sub-array can be tiled to build a larger memory macro. The layout dimension denoted in the

figure shows that STT-MRAM is roughly 3X denser that 6T SRAM including all control circuitries in a 65 nm low power generic logic process. The SRAM array includes an assist schemes to achieve good RD and WR margins. The power supply level of SRAM array (V_{SRAM}) is dynamically switched in a column-based manner [13]. Namely, the V_{SRAM} is controlled to be a 0.1 V higher than wordline voltage level (V_{WL}) in order to improve SNM during RD. On the contrary, V_{WL} is boosted by a 0.1 V than V_{SRAM} for better WR margin. Similarly, dummy cell averaging with disturb-free reference [32] and localized write driver techniques [29] are implemented in the STT-MRAM array for optimal RD and WR performances.

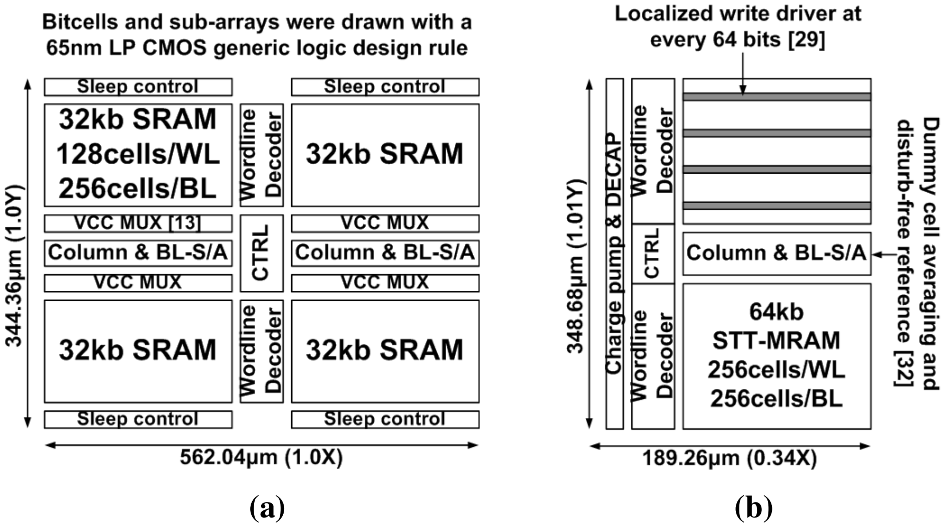


Fig. 5.9: 128 kb sub-array architectures of (a) SRAM and (b) STT-MRAM.

Variation sources present in practical industry designs have been included in our analysis as described in Fig. 5.10 [41], [52], [53]. This ranges from process variation in the memory cells and the S/A to realistic fluctuations for the resistances, capacitances, reference biases and supply levels. Here, a gradual scaling of σ_{Vt} and C_{BL} is again

assumed to prevent the performance scaling trends from being distorted with any abrupt change in the transistor parameters and the parasitic capacitances although this can happen in real situations.

	6T SRAM	STT-MRAM
Power supply noise	Nominal level -10% is assumed	
Bit-cell	Device mismatches	
Parasitic capacitance (C_{BL})	σ/μ=5%: each μ are calculated based on sub-array size	
Resistance area product	-	σ/μ=5%
Sense Amplifier (S/A)	Voltage S/A pair mismatches	I-applying and V-sensing method (AP direction read) + Voltage S/A : I_{REF} σ/μ=2.5%, S/A pair mismatches
Reference cell	-	Reference cell averaging scheme with MTJ replica cells
Write threshold current	-	σ/μ=5%

* Mismatches are based on inverse square root relationship of devices' areas.
* σ_{Vt}/F is assumed to be constant with technology scaling.
* $\mu(C_{BL})$ is assumed to be scaled proportional to scaling factor.

Fig. 5.10: Simulation set-up for evaluating SRAM and STT-MRAM variability.

Fig. 5.11 shows simulation results of J_{RD}/J_{C0} vs. read disturb rate at T_{RD}=2 ns with the proposed scaling scenario and simulation methodology. This indicates that read disturb worsens for $J_{RD}/J_{C0}>2$. The simulation results coincide with the previous MTJ physics-based analysis in [51] showing that $J_{RD}/J_{C0}<2$ is required for disturb-free RD operation with a duty cycle (T_{RD}/T_{CYCLE}) of 50%. So we chose a J_{RD}/J_{C0} of 1.5 in this work.

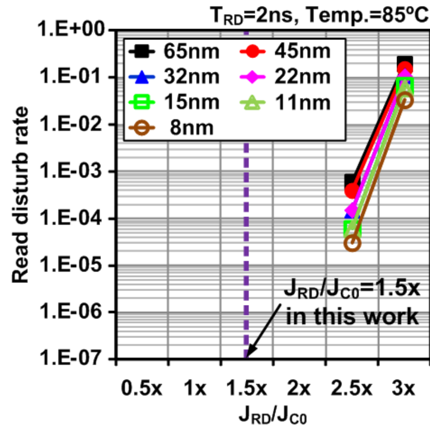


Fig. 5.11: Simulated read disturb rate with varying J_{RD}/J_{C0}.

5.4 Comparison Between SRAM and STT-MRAM

In order to demonstrate the potential of STT-MRAM as an alternative for large density on-chip memories, this section presents macro level performance comparisons with 6T SRAM considering variation effects with technology scaling. Extensive Monte-Carlo simulations were performed on megabit density SRAM and STT-MRAM arrays to estimate their performance in a practical scenario [41], [52], [53].

5.4.1 Macro Performance

Despite a longer access time, dense memories such as eDRAMs or STT-MRAMs are preferred for L2 or L3 caches for their smaller memory footprint and shorter global interconnect delay. Since STT-MRAMs have a 3-5X bit-cell density advantage over SRAMs, their system level performance is expected to be higher even with a longer access time. Fig. 5.12 shows the latency comparison between several embedded memory options. We assume a practical scenario that a denser memory has a longer access time. 1T1C eDRAMs in [4]-[6] have 5X denser bit-cells and roughly 5X longer sensing delays than SRAMs. Similarly, we assumed that the 3X denser STT-MRAM in a generic logic process have a 3X longer sensing delay. When a cache density is small, a memory access time is dominant in determining system performance. Simulated cache latencies with 1 Mb densities in Fig. 5.12(a) show that shorter sensing delays enable faster system performances. With the increase of memory density, the interconnect with repeater delay becomes significant, eventually overcoming the longer sensing delay of denser memories as shown in Fig. 5.12(b).

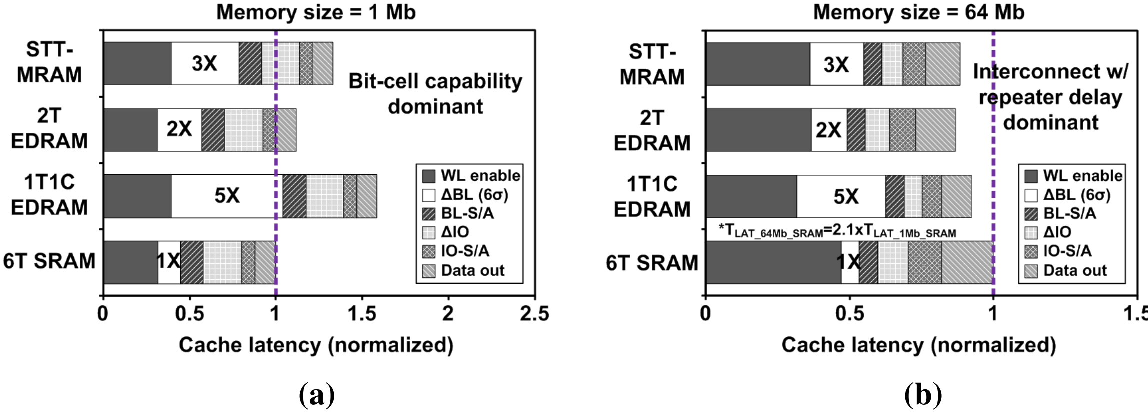


Fig. 5.12: Latency comparison between several embedded memory options with (a) 1 Mb and (b) 64 Mb densities.

This concept is well appreciated in 1T1C eDRAMs [4]-[6] and can be also applied to STT-MRAMs. As a rule of thumb, 3-5X denser STT-MRAMs having 3-5X longer memory access times can outperform SRAMs in large arrays such as 64 Mb or more. We chose the 3X longer memory access time of STT-MRAMs than SRAMs as an iso-latency criterion, which is more than enough since cache densities in Fig. 5.4 start from 16 MB (128 Mb) and have been doubled for every two process generations.

5.4.2 In-Plane STT-MRAM vs. 6T SRAM

Fig. 5.13 shows the 6σ WR performance comparison between SRAM and in-plane STT-MRAM, absolute values in (a) and normalized values in (b), respectively. Here, the write time (T_{WR}) is defined as the WL activation to the time when the cell node flips for SRAMs, and the MTJ switching time for STT-MRAMs, respectively. With technology scaling, the T_{WR} of 6T SRAM degrades due to the reduced supply voltage level and the ratioed operation even with the write assist scheme [13]. On the contrary, I_{C0} of in-plane STT-MRAMs scales by roughly $\alpha^{1.5}$ enabling continuous improvement of the MTJ

switching time. However, this is not enough for STT-MRAMs to outperform SRAMs in write latency even up to 8 nm process node. If the supply voltage of STT-MRAMs can be increased by 0.3V, STT-MRAMs can outperform SRAMs from 15 nm when following the constant J_{C0}•RA/VDD scaling scenario. Unlike in STT-MRAMs where the standby power is zero, increasing the supply voltage is difficult in SRAMs as it will directly impact the leakage power consumption.

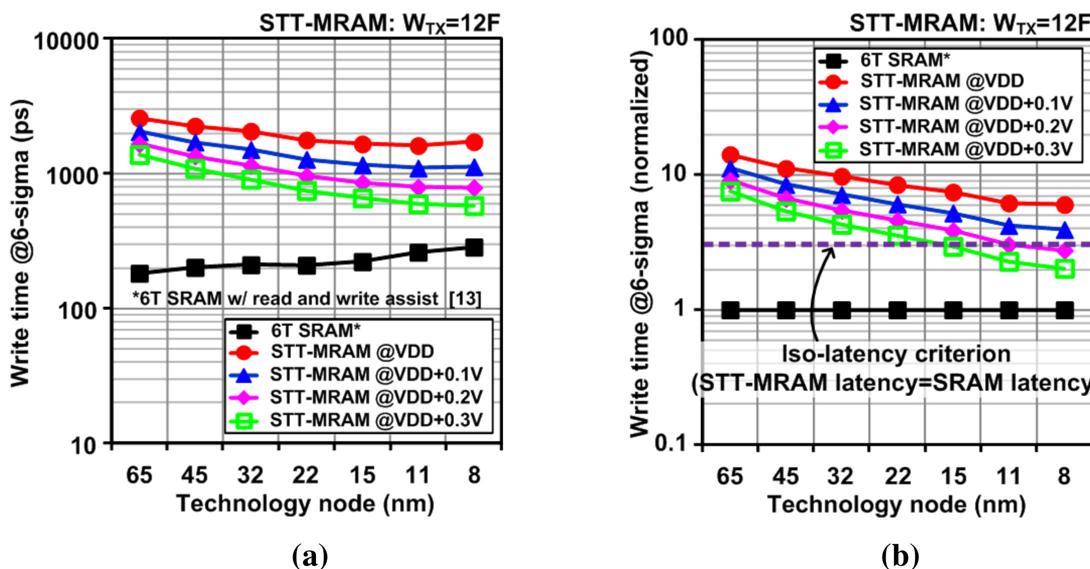


Fig. 5.13: In-plane STT-MRAM scaling trends: Write time. (a) Absolute values. (b) Normalized to SRAM.

Fig 5.14 shows the detailed T_{WR} distributions of SRAM and in-plane STT-MRAM for a 1 Mb macro density in 15 nm process node obtained by running 2^{20} Monte-Carlo simulations in HSPICE, which represents the cell-to-cell variation of a 1 Mb memory macro.

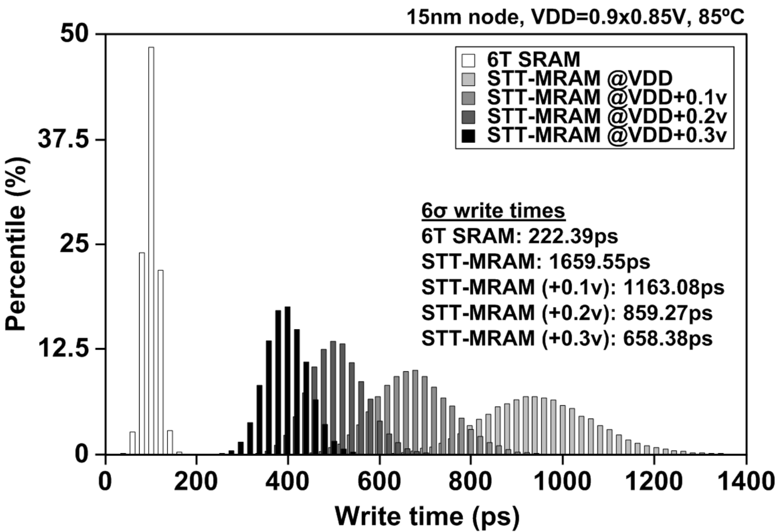


Fig. 5.14: Write time distributions of SRAM and in-plane STT-MRAM (P-AP) for a 1 Mb macro density at 15 nm node.

Fig. 5.15 shows the 6σ RD sensing delay comparison between SRAM and in-plane STT-MRAM, absolute values in (a) and normalized values in (b), respectively. Both SRAM and STT-MRAM arrays have a 1 Mb macro density and a 256 cells-per-BL architecture. Here, the RD sensing delay (T_{RD}) is defined as the WL activation to the time when ΔBL reaches 50 mV for SRAMs, 25 mV for STT-MRAMs, respectively. Due to the single-ended sensing nature and the small TMR, it is not practical for STT-MRAMs to have the same BL voltage difference with SRAMs. This requires more robust S/As such as delicate pre-amplifiers and complicated reference schemes [29]-[32] in STT-MRAM design, resulting in the increase of S/A layout area. The layout dimension of 128 kb sub-arrays in Fig. 5.9 counts this situation. Based on the assumption, the 6σ T_{RD} comparison in Fig. 5.15 indicates that a TMR greater than 200% is required in order for STT-MRAMs to be advantageous over SRAMs. Note that the ITRS predicted value 150% for the next decade. Fig. 5.16 shows the T_{RD} distributions of SRAM and in-

plane STT-MRAM for 1 Mb densities at 15 nm node. Even with the reduced BL voltage difference of 25 mV, STT-MRAM suffers from read failure due to the small TMR. This requirement can be relaxed by increasing J_{C0}•RA since the fast write performance in Fig. 5.13 can be traded off for better read margin.

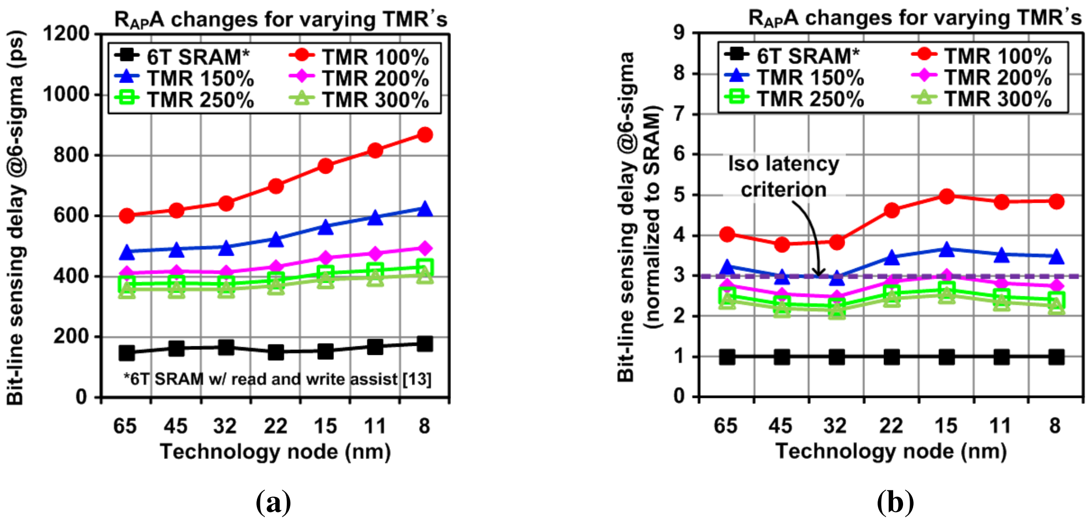


Fig. 5.15: In-plane STT-MRAM scaling trends: Read sensing delay. (a) Absolute values. (b) Normalized to SRAM.

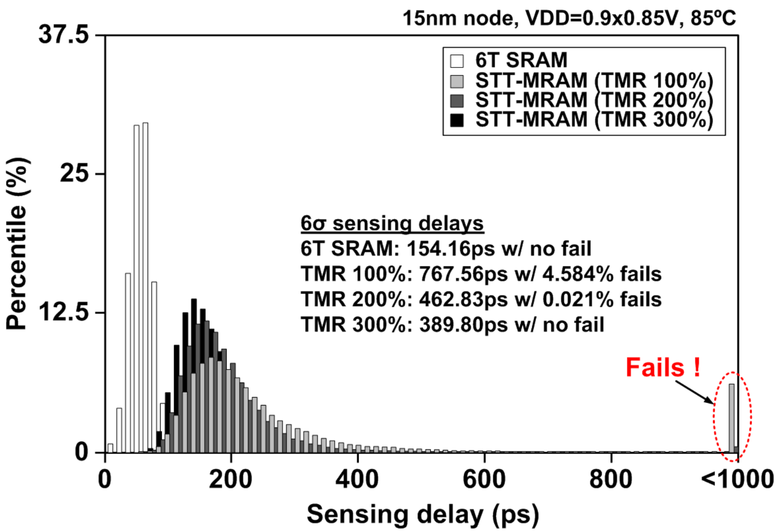


Fig. 5.16: Read sensing delay distributions of SRAM and in-plane STT-MRAM for a 1 Mb macro density at 15 nm node.

5.4.3 In-Plane STT-MRAM vs. Perpendicular STT-MRAM

Fig. 5.17 shows the scaling trends of J_{C0} and RA both for in-plane and perpendicular STT-MTJs. Due to the different origin of magnetic anisotropy, the scaling trend of a perpendicular MTJ is drastically different from its in-plane counterpart, namely J_{C0} scales by $1/\alpha^2$, resulting in the constant I_{C0}, where α is a scaling factor [44]. It is commonly accepted in the research community that the I_{C0} of perpendicular STT-MTJs is smaller than that of in-plane counterpart while maintaining the required thermal stability. This is valid until 22 nm process node based on our projection shown in Fig. 5.4. The I_{C0} of in-plane STT-MRAM is scaled by roughly $\alpha^{1.5}$ while perpendicular one remains constant, so the aforementioned trend will be reversed in scaled technologies; from 15 nm in this work.

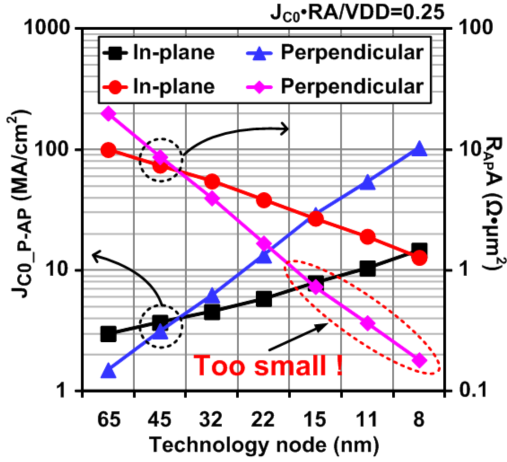


Fig. 5.17: J_{C0} and RA scaling trends of in-plane and perpendicular STT-MTJs.

Under the constant J_{C0}•RA/VDD scaling scenario, the RD margin of perpendicular STT-MRAMs would be similar to that of in-plane counterparts as shown in Fig. 5.18 (a). Fig. 18(b) shows the T_{WR} scaling trends of perpendicular STT-MRAMs. Even with the

constant J_{C0}•RA/VDD scaling scenario, perpendicular STT-MRAMs show very poor write performance. As technology scales, the current drivability of STT-MRAM access transistor decrease since the device width scales by α as well as the scaling of supply voltage level even with process innovations. This limits the fast MTJ switching of perpendicular STT-MRAM under the constant I_{C0} scaling conditions in scaled technologies. Note that RA of a perpendicular MTJ needs to be scaled exponentially in order to maintain a constant J_{C0}•RA/VDD as shown in Fig. 5.17. Under this scenario, the required RA at 8 nm node is expected to be less than 0.2 Ω•μm^2 which will cause severe reliability issues in the thin insulator as well as imposing limits on the TMR value. In the J_{C0}•RA design space, there should be a significant reduction in J_{C0} can be achieved through MTJ device innovations so that the poor write performance of perpendicular STT-MRAMAs can be traded off with the excellent RD margin in Fig. 5.18(a).

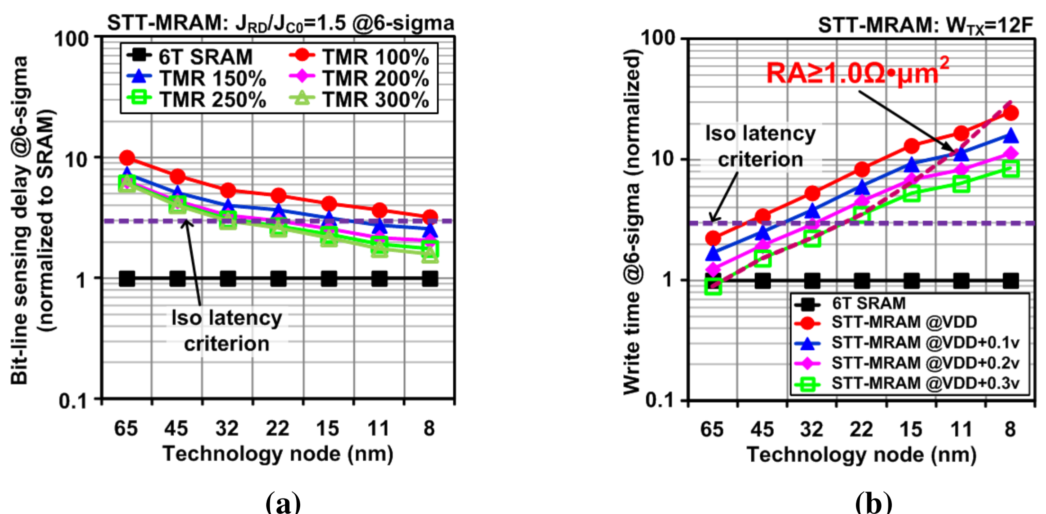


Fig. 5.18: Perpendicular STT-MRAM scaling trends. (a) Sensing delay comparison with SRAM. (b) Write time comparison with SRAM.

5.5 Conclusions

The scalability and variability of in-plane and perpendicular MTJ based STT-MRAMs have been explored considering MTJ properties as well as CMOS circuits for demonstrating the potential as an alternative of high density caches. The proposed STT-MTJ scaling scenario is based on dimensional adjustments and MTJ material innovations in order to maintain the thermal stability. The proposed constant J_{C0}•RA/VDD scaling method provides optimal RD and WR performances. The proposed simulation methodology includes efficient MTJ macromodel, transistor parameters for access devices and peripheral circuitries, state-of-the-art sub-array architectures, and variation sources present in practical industry designs. Our studies based on the proposed methodology shows that the in-plane STT-MRAM is a promising alternative for future high density cache memories by outperforming SRAMs from 15 nm process node when the TMR can be achieved greater than 200% as well as the write supply voltage level is raised by 0.3 V with boosted write WL scheme having a 2VDD level. Perpendicular STT-MRAMs suffer from the poor write performance due to the difficulty of I_{C0} scaling while maintaining thermal stability in scaled technologies from 22 nm process node. Unless there is a significant reduction in J_{C0} through MTJ device innovations, write performance is expected to become the main bottleneck for perpendicular MTJs in future technology nodes.

Chapter 6

Conclusion

A dense embedded memory is one of the most important components in modern microprocessors as a larger cache improves micro-architectural performance with only a modest increase in CV^2f power. The embedded memory options of 6T SRAMs and 1T1C eDRAMs have faced with severe scaling challenges in advanced CMOS technologies. In this dissertation, I presented circuit techniques and simulation methodologies to demonstrate the potential of gain cell eDRAMs and spin-torque-transfer magnetic RAMs (STT-MRAMs) as alternative options for high density embedded memories. Three unique test chip designs that we have been implemented in a generic 65 nm low-leakage CMOS process were presented to enhance the retention time and read speed of gain cell eDRAMs, enabling faster overall system performances and lower static power dissipations than SRAMs and eDRAMs. The scalability of STT-MRAMs under variation effects from 65 nm to 8 nm process nodes were explored with our proposed scaling scenario and simulation methodology, demonstrating that the alternative option can outperform SRAMs in advanced CMOS technologies.

In Chapter 2, a 3T gain cell eDRAM with the proposed boosted 3T gain cell, regulated bit-line write scheme, and adaptive and die-to-die adjustable read reference was presented for enabling a low voltage operation. Measurement results from a 64 kb eDRAM test chip implemented in the 65 nm LP process whose nominal supply level is 1.2 V show a 1.25 ms data retention time with a 2 ns random cycle time at 0.9 V, 85 ºC, and a 91.3 μW per Mb static power dissipation at 1.0 V, 85 ºC. The measured retention time is a 10X improvement over a conventional design measured from the same silicon die and the static power dissipation is about 75% smaller than a power-gated SRAM with a 0.6 V retention voltage.

A high performance 2T gain cell eDRAM with a dense bit-cell was presented in Chapter 3. The benefits of the proposed asymmetric 2T gain cell, pseudo-PMOS diode based current sense amplifier, half-swing WBL scheme, and stepped WWL driver were demonstrated using a 192 kb eDRAM test chip. Measurement results from the test chip show a 667 MHz random cycle frequency with a 512 cells-per-BL architecture and a 400 μs retention time with a 99.9% bit yield condition at 1.1 V and 85 ºC, and an estimated retention time is a 80 μs with a 99.99% bit yield condition at 1.1 V and 105 ºC. To put this perspective, recently published 1T1C eDRAMs show a 500 MHz random cycle time and a 40 μs retention time with a 99.99% bit yield condition and 32 cells-per-local BL architecture at 1.0 V and 105 ºC. The latency of the 2T eDRAM is 9.5% faster than a 6T SRAM each with a 1 Mb density, and the static power dissipation of the 2T eDRAM is 81% smaller than a power-gated 6T SRAM. As the cache density increases, the system performance of the 2T eDRAM can be further enhanced than a 6T SRAM while

maintaining the static power advantage due to the smaller footprint. In the 65nm LP process, the 2T gain cell is 59.5% smaller than a 6T SRAM.

In Chapter 4, the proposed 2T1C gain cell, single-ended 7T SRAM repair, and gain cell storage monitor enabled a truly logic-compatible gain cell eDRAM with no boosted supplies. A 128 kb eDRAM test chip shows a random access frequency of 714 MHz and a static power dissipation of 161.8 μW per Mb with a 500 μs refresh rate at 1.1 V and 85 °C which are comparable to the previous 2T eDRAM operating with boosted supplies. Although the bit-cell size of the 2T1C is 43% larger than that of the 2T, still a 1.72X denser than a 6T SRAM.

Unlike 6T SRAMs or 1T1C eDRAMs, gain cells have a decoupled read and write structure leading to improved read and write margins and flexibility in the bit-cell design - for example, the read and write paths can be optimized separately allowing gain cells to scale favorably in future technology nodes. With the proposed three designs, this dissertation demonstrates that gain cell eDRAMs are a promising alternative for high density embedded memories.

Finally, in Chapter 5, the scalability and variability of in-plane and perpendicular MTJ based STT-MRAMs are explored by comparing its performances with 6T SRAM for high density on-die memories with the proposed scaling scenario and the simulation methodology. Our studies based on the proposed methodology show that in-plane STT-MRAM will outperform SRAM from 15 nm node, while its perpendicular counterpart

requires further innovations in MTJ material properties in order to overcome the poor write performance from 22 nm node.

Bibliography

[1] S. Rusu, S. Tam, H. Muljono, J. Stinson, D. Ayers, J. Chang, R. Varada, M. Ratta, S. Kottapalli, and S. Vora, "A 45 nm 8-Core Enterprise Xeon® Processor," *IEEE Jour. Solid-State Circuits*, vol. 45, no. 1, pp. 7-14, January 2010.

[2] S. Rusu, S. Tam, H. Muljono, D. Ayers, J. Chang, B. Cherkauer, J. Stinson, J. Benoit, R. Varada, J. Leung, R. D. Limaye, and S. Vora, "A 65-nm Dual-Core Multithreaded Xeon® Processor With 16-MB L3 Cache," *IEEE Jour. Solid-State Circuits*, vol. 42, no. 1, pp. 17-25, January 2007.

[3] R. J. Riedlinger, R. Bhatia, L. Biro, B. Bowhill, E. Fetzer, P. Gronowski, and T. Grutkowski, "A 32nm 3.1 Billion Transistor 12-Wide-Issue Itanium® Processor for Mission-Critical Servers," *IEEE Int. Solid-State Circuits Conf.*, pp. 84-85, 2011.

[4] R. Kalla, B. Sinharoy, W. J. Starke, and M. Floyd, "POWER7: IBM's Next-Generation Server Processor," *IEEE Micro*, vol. 30, issue 2, pp. 7-15, March/April 2010.

[5] J. Barth, D. Plass, E. Nelson, C. Hwang, G. Fredeman, M. Sperling, A. Mathews, T. Kirihata, W. R. Reohr, K. Nair, and N. Cao, "A 45 nm SOI Embedded DRAM

Macro for the POWER™ Processor 32 MByte On-Chip L3 Cache," *IEEE Jour. Solid-State Circuits*, vol. 46, no. 1, pp. 64-75, January 2011.

[6] J. Barth, W. R. Reohr, P. Parries, G. Fredeman, J. Golz, S. E. Schuster, R. E. Matick, H. Hunter, C. C. Tanner, III, J. Harig, H. Kim, B. A. Khan, J. Griesemer, R. P. Havreluk, K. Yanagisawa, T. Kirihata, and S. S. Iyer, "A 500 MHz Random Cycle, 1.5 ns Latency, SOI Embedded DRAM Macro Featuring a Three-Transistor Micro Sense Amplifier," *IEEE Jour. Solid-State Circuits*, vol. 43, no. 1, pp. 86-95, January 2008.

[7] S. Romanovsky, A. Katoch, A. Achyuthan, C. O'Connell, S. Natarajan, C. Huang, C.-Y. Wu, M.-J. Wang, C. J. Wang, P. Chen, and R. Hsieh, "A 500MHz Random-Access Embedded 1Mb DRAM Macro in Bulk CMOS," *IEEE Int. Solid-State Circuits Conf.*, pp. 270-271, 2009.

[8] P. J. Klim, J. Barth, W. R. Reohr, D. Dick, G. Fredeman, G. Koch, H. M. Le, A. Khargonekar, P. Wilcox, J. Golz, J. B. Kuang, A. Mathews, J. C. Law, T. Luong, H. C. Ngo, R. Freese, H. C. Hunter, E. Nelson, P. Parries, T. Kirihata, and S. S. Iyer, "A 1 MB Cache Subsystem Prototype With 1.8 ns Embedded DRAMs in 45 nm SOI CMOS," *IEEE Jour. Solid-State Circuits*, vol. 44, no. 4, pp. 1216-1226, April 2009.

[9] R. E. Matick, and S. E. Schuster, "Logic-Based EDRAM: Origins and Rationale for Use," *IBM Jour. Res. & Dev.*, vol. 49, no. 1, pp. 145-165, January 2005.

[10] S. Tyagi, M. Alavi, R. Bigwood, T. Bramblett, J. Brandenburg, W. Chen, B. Crew, M. Hussein, P. Jacob, C. Kenyon, C. Lo, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen,

L. Rumaner, R. Schweinfurth, S. Sivakumar, M. Stettler, S.Thompson, B. Tufts, J. Xu, S. Yang and M. Bohr, "A 130 nm Generation Logic Technology Featuring 70 nm Transistors, Dual Vt Transistors and 6 layers of Cu Interconnects," *IEEE Electron Devices Meeting*, pp. 567-570, 2000.

[11] S. Thompson, N. Anand, M. Armstrong, C. Auth, B. Arcot, M. Alavi, P. Bai, J. Bielefeld, R. Bigwood, J. Brandenburg, M. Buehler, S. Cea, V. Chikarmane, C. Choi, R. Frankovic, T. Ghani, G. Glass, W. Han, T. Hoffmann, M. Hussein, P. Jacob, A. Jain, C. Jan, S. Joshi, C. Kenyon, J. Klaus, S. Klopcic, J. Luce, Z. Ma, B. Mcintyre, K. Misty, A. Murthy, P. Nguyen, H. Pearson, T. Sandford, R. Schweinfurth, R. Shaheed, S. Sivakumar, M. Taylor, B. Tufts, C. Wallace, P. Wang, C. Weber, and M. Bohr, "A 90 nm Logic Technology Featuring 50 nm Strained Silicon Channel Transistors, 7 layers of Cu Interconnects, Low k ILD, and 1 μm^2 SRAM Cell," *IEEE Electron Devices Meeting*, pp. 61-64, 2002.

[12] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "SRAM Design on 65-nm CMOS Technology With Dynamic Sleep Transistor for Leakage Reduction," *IEEE Jour. Solid-State Circuits*, vol. 40, no. 4, pp. 895-901, April 2005.

[13] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and Mark Bohr, "A 3-GHz 70-Mb SRAM in 65-nm CMOS Technology With Integrated Column-Based Dynamic Power Supply," *IEEE Jour. Solid-State Circuits*, vol. 41, no. 1, pp. 146-151, January 2006.

[14] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, Y.-G. Ng, A. Pavlov, K. Smits, and M. Bohr, "A 153Mb-SRAM Design with Dynamic Stability Enhancement and Leakage Reduction in 45nm High-k Metal-Gate CMOS Technology," *IEEE Int. Solid-State Circuits Conf.*, pp. 376-377, 2008.

[15] P. Packan, S. Akbar, M. Armstrong, D. Bergstrom, M. Brazier, H. Deshpande, K. Dev, G. Ding, T. Ghani, O. Golonzka, W. Han, J. He, R. Heussner, R. James, J. Jopling, C. Kenyon, S.-H. Lee, M. Liu, S. Lodha, B. Mattis, A. Murthy, L. Neiberg, J. Neirynck, S. Pae, C. Parker, L. Pipes, J. Sebastian, J. Seiple, B. Sell, A. Sharma, S. Sivakumar, B. Song, A. St. Amour, K. Tone, T. Troeger, C. Weber, K. Zhang, Y. Luo, and S. Natarajan, "High Performance 32nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistors," *IEEE Electron Devices Meeting*, pp. 659-662, 2009.

[16] M. Ichihashi, H. Toda, Y. Itoh, and K. Ishibashi, "0.5V Asymmetric Three-Tr. Cell (ATC) DRAM Using 90nm Generic CMOS Logic Process," *IEEE VLSI Circuits Symp.*, pp. 366-369, 2005.

[17] W. K. Luk, and R. H. Dennard, "2T1D Memory Cell with Voltage Gain," *IEEE VLSI Circuits Symp.*, pp. 184-187, 2004.

[18] W. K. Luk, J. Cai, R. H. Dennard, M. J. Immediato, and S. V. Kosonocky, "A 3-Transistor DRAM Cell with Gated Diode for Enhanced Speed and Retention Time," *IEEE VLSI Circuits Symp.*, pp. 184-185, 2006.

[19] D. Somasekhar, Y. Ye, P. Aseron, S.-L. Lu, M. M. Khellah, J. Howard, Greg Ruhl, T. Karnik, S. Borkar, V. K. De, and A. Keshavarzi, "2 GHz 2 Mb 2T Gain

Cell Memory Macro with 128 GBytes/sec Bandwidth in a 65 nm Logic Process Technology," *IEEE Jour. Solid-State Circuits*, vol. 44, no. 1, pp. 174-185, January 2009.

[20] E. B. Myers, D. C. Ralph, J. A. Katine, R. N. Louie, and R. A. Buhrman, "Current-Induced Switching of Domains in Magnetic Multilayer Devices," *Science*, vol. 285, no. 5429, pp. 867–870, August 1999.

[21] J. Z. Sun, "Spin Angular Momentum Transfer in Current-Perpendicular Nanomagnetic Junctions," *IBM Jour. Res. & Dev.*, vol. 50, no. 1, pp. 81-100, January 2006.

[22] Y. Huai, "Spin-Transfer Torque MRAM (STT-MRAM) Challenges and Prospects," *AAPPS Bulletin*, vol. 18, no. 6, pp. 33-40, December 2008.

[23] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. KanoM, "A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin-RAM," *IEEE Electron Devices Meeting*, pp. 459-462, 2005.

[24] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, S. Takahashi, M. Nakayama, N. Shimomura, H. Aikawa, S. Ikegawa, S. Yuasa, K. Yakushiji, H. Kubota, A. Fukushima, M. Oogane, T. Miyazaki, and K. Ando, "Lower-current and Fast switching of A Perpendicular TMR for High Speed and High density Spin-Transfer-Torque MRAM," *IEEE Electron Devices Meeting*, pp. 309-312, 2008.

[25] C. J. Lin, S. H. Kang, Y. J. Wang, K. Lee, X. Zhu, W. C. Chen, X. Li, W. N. Hsu, Y. C. Kao, M. T. Liu, W. C. Chen, Y. C. Lin, M. Nowak, N. Yu, and L. Tran, "45nm Low Power CMOS Logic Compatible Embedded STT MRAM Utilizing a Reverse-Connection 1T/1MTJ Cell," *IEEE Electron Devices Meeting*, pp. 279-282, 2009.

[26] T. Ishigaki, T. Kawahara, R. Takemura, K. Ono, K. Ito, H. Matsuoka, and H. Ohno, "A Multi-Level-Cell Spin-Transfer Torque Memory with Series-Stacked Magnetotunnel Junctions," *IEEE VLSI Technology Symp.*, pp. 47-48, 2010.

[27] Y. Lee, C. Yoshida, K. Tsunoda, S. Umehara, M. Aoki, and T. Sugii, "Highly Scalable STT-MRAM with MTJs of Top-pinned Structure in 1T/1MTJ Cell," *IEEE VLSI Technology Symp.*, pp. 49-50, 2010.

[28] T. Kawahara, R. Takemura, K. Miura, J. Hayakawa, S. Ikeda, Y. M. Lee, R. Sasaki, Y. Goto, K. Ito, T. Meguro, F. Matsukura, H. Takahashi, H. Matsuoka, and H. Ohno, "2 Mb SPRAM (SPin-Transfer Torque RAM) With Bit-by-Bit Bi-Directional Current Write and Parallelizing-Direction Current Read," *IEEE Jour. Solid-State Circuits*, vol. 43, no. 1, pp. 109-120, January 2008.

[29] R. Takemura, T. Kawahara, K. Miura, H. Yamamoto, J. Hayakawa, N. Matsuzaki, K. Ono, M. Yamanouchi, K. Ito, H. Takahashi, S. Ikeda, H. Hasegawa, H. Matsuoka, and H. Ohno, "A 32-Mb SPRAM With 2T1R Memory Cell, Localized Bi-Directional Write Driver and '1'/'0' Dual-Array Equalized Reference Scheme," *IEEE Jour. Solid-State Circuits*, vol. 45, no. 4, pp. 869-879, April 2010.

[30] D. Halupka, S. Huda, W. Song, A. Sheikholeslami, K. Tsunoda, C. Yoshida, and M. Aoki, "Negative-Resistance Read and Write Schemes for STT-MRAM in 0.13μm CMOS," *IEEE Int. Solid-State Circuits Conf.*, pp. 256-257, 2010.

[31] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, T. Kajiyama, M. Iwayama, K. Sugiura, S. Ikegawa, T. Kishi, T. Kai, M. Amano, N. Shimomura, H. Yoda, and Y. Watanabe, "A 64Mb MRAM with Clamped-Reference and Adequate-Reference Schemes," *IEEE Int. Solid-State Circuits Conf.*, pp. 258-259, 2010.

[32] J. Kim, T. Kim, W. Hao, H. M. Rao, K. Lee, X. Zhu, X. Li, W. Hsu, S. H. Kang, N. Matt, and N. Yu, "A 45nm 1Mb Embedded STT-MRAM with design techniques to minimize read-disturbance," *IEEE VLSI Circuits Symp.*, pp. 296-297, 2011.

[33] K. Chun, P. Jain, J. Lee, and C. H. Kim, "A Sub-0.9V Logic-compatible Embedded DRAM with Boosted 3T Gain Cell, Regulated Bit-line Write Scheme and PVT-tracking Read Reference Bias," *IEEE VLSI Circuits Symp.*, pp. 134-135, 2009.

[34] K. Chun, P. Jain, J. Lee, and C. H. Kim, "A 3T Gain Cell Embedded DRAM Utilizing Preferential Boosting for High Density and Low Power On-Die Caches," *IEEE Jour. Solid-State Circuits*, vol. 46, no. 6, pp. 1495-1505, June 2011.

[35] K. Chun, P. Jain, T. Kim, and C. H. Kim, "A 1.1V, 667MHz Random Cycle, Asymmetric 2T Gain Cell Embedded DRAM with a 99.9 Percentile Retention Time of 110μsec," *IEEE VLSI Circuits Symp.*, pp. 191-192, 2010.

[36] K. Chun, P. Jain, T. Kim, and C. H. Kim, "A 667 MHz Logic-Compatible embedded DRAM Featuring an Asymmetric 2T Gain Cell for High Speed On-Die Caches," *IEEE Jour. Solid-State Circuits*, vol. 47, no. 2, pp. 547-559, February 2012.

[37] K. Chun, W Zhang, P. Jain, and C. H. Kim, "A 700MHz 2T1C Embedded DRAM Macro in a Generic Logic Process with No Boosted Supplies," *IEEE Int. Solid-State Circuits Conf.*, pp. 506-507, 2011.

[38] K. Chun, H. Zhao, J. D. Harms, Tony T. Kim, J.-P. Wang, and C. H. Kim, "A Technology Roadmap for In-plane and Perpendicular MTJ Based STT-MRAMs Considering Variability Effects," *IEEE VLSI Circuits Symp.*, pp. 1-2, 2012 (submitted).

[39] E. Seevinck, P. J. van Beers, and H. Ontrop, "Current-Mode Techniques for High-Speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAM's," *IEEE Jour. Solid-State Circuits*, vol. 26, no. 4, pp. 525-536, April 1991.

[40] J. Sim, H. Yoon, K. Chun, H. Lee, S. Hong, K. Lee, J. Yoo, D. Seo, and S. Cho, "A 1.8-V 128-Mb Mobile DRAM With Double Boosting Pump, Hybrid Current Sense Amplifier, and Dual-Referenced Adjustment Scheme for Temperature Sensor," *IEEE Jour. Solid-State Circuits*, vol. 38, no. 4, pp. 631-640, April 2003.

[41] K. Agarwal, and S. Nassif, "The Impact of Random Device Variation on SRAM Cell Stability in Sub-90-nm CMOS Technologies," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 16, no. 1, pp. 86-97, January 2008.

[42] C. H. Kim, J. Kim, I. Chang, and K. Roy, "PVT-Aware Leakage Reduction for On-Die Caches With Improved Read Stability," *IEEE Jour. Solid-State Circuits*, vol. 41, no. 1, pp. 170-178, January 2006.

[43] F. J. M.-Martinez, E. K. Ardestani and J. Renau, "Characterizing Processor Thermal Behavior," *ACM ASPLOS*, pp. 193-204, 2010.

[44] D. Apalkov, S. Watts, A. D.-Smith, E. Chen, Z. Diao, and V. Nikitin, "Comparison of Scaling of In-Plane and Perpendicular Spin Transfer Switching Technologies by Micromagnetic Simulation," *IEEE Trans. Magnetics*, vol. 46, no. 6, pp. 2240-2243, June 2010.

[45] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *Jour. Phys.: Condensed Matter*, vol. 19, no. 16, pp. 165209-1-165209-13, April 2007.

[46] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Design Space and Scalability Exploration of 1T-1STT MTJ Memory Arrays in the Presence of Variability and Disturbances," *IEEE Electron Devices Meeting*, pp. 707-710, 2009.

[47] H. Zhao, A. Lyle, Y. Zhang, P. K. Amiri, G. Rowlands, Z. Zeng, J. Katine, H. Jiang, K. Galatsis, K. L. Wang, I. N. Krivorotov, and J.-P. Wang, "Low Writing Energy and Sub Nanosecond Spin Torque Transfer Switching of In-plane Magnetic Tunnel Junction for Spin Torque Transfer Random Access Memory," *Jour. Appl. Phys.*, vol. 109, issue. 7, pp. 07C720-1-07C720-3, March 2011.

[48] J. D. Harms, F. Ebrahimi, X. Yao, and J.-P. Wang, "SPICE Macromodel of Spin-Torque-Transfer-Operated Magnetic Tunnel Junctions," *IEEE Trans. Electron Devices*, vol. 57, no. 6, pp. 1425-1430, June 2010.

[49] S. Lee, S. Lee, H. Shin, and D. Kim, "Advanced HSPICE macromodel for magnetic tunnel junctions," *Jpn. Jour. Appl. Phys.*, vol. 44, no. 4B, pp. 2696–2700, April 2005.

[50] International Technology Roadmap for Semiconductors (2009). [Online]. Available: http://www.itrs.net

[51] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. K. De, "Modeling and Analysis of Read (RD) Disturb in IT-ISTT MTJ Memory Bits," *IEEE Device Research Conf.*, pp. 43-44, 2010.

[52] R. Beach, T. Min, C. Horng, Q. Chen, P. Sherman, S. Le, S. Young, K. Yang, H. Yu, X. Lu, W. Kula, T. Zhong, R. Xiao, A. Zhong, G. Liu, J. Kan, J. Yuan, J. Chen, R. Tong, J. Chien, T. Torng, D. Tang, P. Wang, M. Chen, S. Assefa, M. Qazi, J. DeBrosse, M. Gaidis, S. Kanakasabapathy, Y. Lu, J. Nowak, E. O'Sullivan, T. Maffitt, J.Z. Sun, and W.J. Gallagher, "A Statistical Study of

Magnetic Tunnel Junctions for High-Density Spin Torque Transfer-MRAM (STT-MRAM)," *IEEE Trans. Electron Devices*, pp. 306-308, 2008.

[53] D. C. Worledge, G. Hu, P. L. Trouilloud, D. W. Abraham, S. Brown, M. C. Gaidis, J. Nowak, E. J. O'Sullivan, R. P. Robertazzi, J. Z. Sun, and W. J. Gallagher, "Switching distributions and write reliability of perpendicular spin torque MRAM," *IEEE Trans. Electron Devices*, pp. 296-299, 2010.